Springer Complexity

Springer Complexity is a publication program, cutting across all traditional disciplines of sciences as well as engineering, economics, medicine, psychology and computer sciences, which is aimed at researchers, students and practitioners working in the field of complex systems. Complex Systems are systems that comprise many interacting parts with the ability to generate a new quality of macroscopic collective behavior through self-organization, e.g., the spontaneous formation of temporal, spatial or functional structures. This recognition, that the collective behavior of the whole system cannot be simply inferred from the understanding of the behavior of the individual components, has led to various new concepts and sophisticated tools of complexity. The main concepts and tools – with sometimes overlapping contents and methodologies – are the theories of self-organization, complex systems, synergetics, dynamical systems, turbulence, catastrophes, instabilities, nonlinearity, stochastic processes, chaos, neural networks, cellular automata, adaptive systems, and genetic algorithms.

The topics treated within Springer Complexity are as diverse as lasers or fluids in physics, machine cutting phenomena of workpieces or electric circuits with feedback in engineering, growth of crystals or pattern formation in chemistry, morphogenesis in biology, brain function in neurology, behavior of stock exchange rates in economics, or the formation of public opinion in sociology. All these seemingly quite different kinds of structure formation have a number of important features and underlying structures in common. These deep structural similarities can be exploited to transfer analytical methods and understanding from one field to another. The Springer Complexity program therefore seeks to foster cross-fertilization between the disciplines and a dialogue between theoreticians and experimentalists for a deeper understanding of the general structure and behavior of complex systems.

The program consists of individual books, books series such as “Springer Series in Synergetics”, “Institute of Nonlinear Science”, “Physics of Neural Networks”, and “Understanding Complex Systems”, as well as various journals.
Understanding Complex Systems

Series Editor

J.A. Scott Kelso
Florida Atlantic University
Center for Complex Systems
Glades Road 777
Boca Raton, FL 33431-0991, USA

Understanding Complex Systems

Future scientific and technological developments in many fields will necessarily depend upon coming to grips with complex systems. Such systems are complex in both their composition (typically many different kinds of components interacting with each other and their environments on multiple levels) and in the rich diversity of behavior of which they are capable. The Springer Series in Understanding Complex Systems series (UCS) promotes new strategies and paradigms for understanding and realizing applications of complex systems research in a wide variety of fields and endeavors. UCS is explicitly transdisciplinary. It has three main goals: First, to elaborate the concepts, methods and tools of self-organizing dynamical systems at all levels of description and in all scientific fields, especially newly emerging areas within the Life, Social, Behavioral, Economic, Neuro- and Cognitive Sciences (and derivatives thereof); second, to encourage novel applications of these ideas in various fields of Engineering and Computation such as robotics, nanotechnology and informatics, third, to provide a single forum within which commonalities and differences in the workings of complex systems may be discerned, hence leading to deeper insight and understanding. UCS will publish monographs and selected edited contributions from specialized conferences and workshops aimed at communicating new findings to a large multidisciplinary audience.
Preface

This monograph is devoted to a new approach to an old field of scientific investigation, freeway traffic research. Freeway traffic is an extremely complex spatiotemporal nonlinear dynamic process. For this reason, it is not surprising that empirical traffic pattern features have only recently been sufficiently understood. Such empirical features are in serious conflict with almost all earlier theoretical and model results. Consequently, the author introduced a new traffic flow theory called “three-phase traffic theory,” which can explain these empirical spatiotemporal traffic patterns. The main focus of this book is a consideration of empirical spatiotemporal traffic pattern features, their engineering applications, and explanations based on the three-phase traffic theory.

The book consists of four parts. In Part I, empirical studies of traffic flow patterns, earlier traffic flow theories, and mathematical models are briefly reviewed. Three-phase traffic theory is considered as well. This theory is a qualitative theory. Main ideas and results of the three-phase traffic flow theory will be introduced and explained without complex mathematical models. This should be suitable for a very broad audience of practical engineers, physicists, and other readers who may not necessarily be specialists in traffic flow problems, and who may not necessarily have worked in the field of spatiotemporal pattern formation.

In Part II, empirical spatiotemporal traffic pattern features are considered. A microscopic three-phase traffic theory of these patterns and results of an application of the pattern features to engineering applications are presented in Part III and Part IV, respectively.

I am very grateful to Herman Haken for the opportunity to write this book. I am also very grateful to my colleagues at DaimlerChrysler AG, Peter Häußer mann, Harald Brunini, Ralf Guido Herrtwich, and Matthias Schulze for their support. I thank my colleagues and friends Hani Mahmassani, Dietrich Wolf, and Michael Schreckenberg for their support in the first publications of my three-phase traffic theory. I would also like to thank the coauthors of our joint publications, Peter Konhäuser, Martin Schilke, Hubert Rehborn, Sergey Klenov, Dietrich Wolf, Matthias Herrmann, Malte Rödiger, Heribert Kirschfink, Mario Aleksić, and Andreas Haug for their very fruitful cooperation. In particular, I thank Sergey Klenov, Hubert Rehborn, Mario Aleksić,
Ines Maiwald-Hiller, Andreas Haug, and James Banks for their suggestions and help in the preparation of this book. I would like to thank the Hessen (Germany) Ministry of Roads and Traffic for help in the preparation of the empirical data. I acknowledge funding by the German Ministry of Education (BMBF) within projects SANDY and DAISY. I would like to thank Pravin Varaiya and his colleagues for access to traffic data of the PeMS (Freeway Performance Measurement System) database in the USA. I also thank my wife, Tatiana Kerner, for her help and understanding.

Stuttgart,  
August 2004

Boris Kerner
Contents

1 Introduction ................................................................. 1

Part I Historical Overview
and Three-Phase Traffic Theory

2 Spatiotemporal Pattern Formation
in Freeway Traffic ....................................................... 13
  2.1 Introduction ......................................................... 13
  2.2 Traffic and Synergetics ............................................. 14
  2.3 Free and Congested Traffic ......................................... 15
     2.3.1 Local Measurements of Traffic Variables .................. 15
     2.3.2 Examples of Freeway Infrastructures
          and Detector Arrangements .................................. 17
     2.3.3 Free Traffic Flow ............................................. 18
     2.3.4 Congested Traffic ............................................. 21
     2.3.5 Empirical Fundamental Diagram ............................. 22
     2.3.6 Complex Local Dynamics of Congested Traffic ............ 24
  2.4 Main Empirical Features of Spatiotemporal
     Congested Patterns ................................................... 27
     2.4.1 Three Traffic Phases ........................................... 27
     2.4.2 Characteristic Parameters of Wide Moving Jams .......... 28
     2.4.3 Spontaneous Breakdown Phenomenon
          (Spontaneous F→S Transition) ............................... 32
     2.4.4 Induced Breakdown Phenomenon .............................. 34
     2.4.5 Synchronized Flow Patterns ................................. 35
     2.4.6 Catch Effect .................................................... 37
     2.4.7 Moving Jam Emergence in Synchronized Flow:
          General Pattern ............................................... 41
     2.4.8 Expanded Congested Patterns ................................ 46
     2.4.9 Foreign Wide Moving Jams ................................... 51
     2.4.10 Reproducible and Predictable Congested Patterns ....... 53
     2.4.11 Methodology for Empirical Congested Pattern Study .... 57
  2.5 Conclusions. Fundamental Empirical Features
     of Spatiotemporal Congested Patterns ............................. 58
3 Overview of Freeway Traffic Theories and Models: Fundamental Diagram Approach .......................... 63
  3.1 Introduction: Hypothesis About Theoretical Fundamental Diagram ......................... 63
  3.2 Achievements of Fundamental Diagram Approach to Traffic Flow Modeling and Theory .................. 64
    3.2.1 Conservation of Vehicle Number on Road and Front Velocity ...................... 66
    3.2.2 The Lighthill–Whitham–Richards Model and Shock Wave Theory .................. 67
    3.2.3 Collective Flow Concept and Probability of Passing ............................ 68
    3.2.4 Scenarios for Moving Jam Emergence ........................................ 69
    3.2.5 Wide Moving Jam Characteristics ........................................... 69
    3.2.6 Flow Rate in Wide Moving Jam Outflow, the Line $J$ ...................................... 71
    3.2.7 Metastable States of Free Flow with Respect to Moving Jam Emergence ............... 73
  3.3 Drawbacks of Fundamental Diagram Approach in Describing of Spatiotemporal Congested Freeway Patterns ........................................... 78
    3.3.1 Shock Wave Theory ........................................................... 78
    3.3.2 Models and Theories of Moving Jam Emergence in Free Flow .......................... 80
    3.3.3 Models and Theories with Variety of Vehicle and Driver Characteristics .................. 83
    3.3.4 Application of Classical Queuing Theories to Freeway Congested Traffic Patterns ............... 84
  3.4 Conclusions ........................................................................ 85

4 Basis of Three-Phase Traffic Theory ................................. 87
  4.1 Introduction and Remarks on Three-Phase Traffic Theory ........................................ 87
  4.2 Definition of Traffic Phases in Congested Traffic Based on Empirical Data ................... 88
    4.2.1 Objective Criteria for Traffic Phases in Congested Traffic............................ 88
    4.2.2 Explanation of Terms “Synchronized Flow” and “Wide Moving Jam” .................. 90
    4.2.3 Mean Vehicle Trajectories .................................................. 91
    4.2.4 Flow Rate in Synchronized Flow ........................................... 91
    4.2.5 Empirical Line $J$ .............................................................. 93
    4.2.6 Propagation of Two Wide Moving Jams ................................... 94
  4.3 Fundamental Hypothesis of Three-Phase Traffic Theory ............................................. 95
5 Breakdown Phenomenon (F→S Transition) in Three-Phase Traffic Theory ............................. 105
  5.1 Introduction ........................................... 105
  5.2 Breakdown Phenomenon on Homogeneous Road .......... 106
    5.2.1 Speed Breakdown at Limit Point of Free Flow ...... 106
    5.2.2 Critical Local Perturbation for Speed Breakdown .. 108
    5.2.3 Probability for Breakdown Phenomenon .......... 110
    5.2.4 Threshold Flow Rate and Density, Metastability, and Nucleation Effects .......... 111
    5.2.5 Z-Shaped Speed–Density and Passing Probability Characteristics .......... 114
    5.2.6 Physics of Breakdown Phenomenon: Competition Between Over-Acceleration and Speed Adaptation ................. 119
    5.2.7 Physics of Threshold Point in Free Flow .......... 120
    5.2.8 Moving Synchronized Flow Pattern .......... 122
  5.3 Breakdown Phenomenon at Freeway Bottlenecks .......... 123
    5.3.1 Deterministic Local Perturbation .......... 123
    5.3.2 Deterministic F→S Transition .......... 128
    5.3.3 Physics of Deterministic Speed Breakdown at Bottleneck .......... 133
    5.3.4 Influence of Random Perturbations .......... 133
    5.3.5 Z-Characteristic for Speed Breakdown at Bottleneck 136
    5.3.6 Physics of Speed Breakdown at Bottleneck .. 139
    5.3.7 Time Delay of Speed Breakdown .......... 140
  5.4 Conclusions ............................................ 142

6 Moving Jam Emergence in Three-Phase Traffic Theory ... 145
  6.1 Introduction .......................................... 145
  6.2 Wide Moving Jam Emergence in Free Flow .......... 147
  6.3 Wide Moving Jam Emergence in Synchronized Flow .......... 150
    6.3.1 Hypothesis for Moving Jam Emergence in Synchronized Flow .......... 150
    6.3.2 Features of Metastable Synchronized Flow States .......... 155
    6.3.3 Stable High Density Synchronized Flow States .......... 156
6.4 Double Z-Shaped Traffic Flow Characteristics .............. 158
  6.4.1 Z-Characteristic for S→J Transition ................... 158
  6.4.2 Cascade of Two Phase Transitions
      (F→S→J Transitions) ................................... 161
  6.4.3 Wide Moving Jam Emergence
      Within Initial Moving Synchronized Flow Pattern ... 167

6.5 Moving Jam Emergence
  in Synchronized Flow at Bottlenecks ..................... 169
  6.5.1 Why Moving Jams Do not Emerge
      in Free Flow at Bottlenecks .......................... 169
  6.5.2 Z-Characteristic for S→J Transition at Bottlenecks . 171
  6.5.3 Physics of Moving Jam Emergence
      in Synchronized Flow .................................. 172
  6.5.4 Double Z-Characteristic and F→S→J Transitions
      at Bottlenecks ...................................... 176

6.6 Conclusions ............................................ 178

7 Congested Patterns at Freeway Bottlenecks
   in Three-Phase Traffic Theory ............................ 179
  7.1 Introduction ........................................... 179
  7.2 Two Main Types
      of Spatiotemporal Congested Patterns .................. 179
  7.3 Simplified Diagram of Congested Patterns
      at Isolated Bottlenecks ................................ 180
  7.4 Synchronized Flow Patterns ............................. 183
      7.4.1 Influence of Fluctuations on Limit Point
      for Free Flow at Bottlenecks .......................... 183
      7.4.2 Moving Synchronized Flow Pattern
      Emergence at Bottlenecks .............................. 185
      7.4.3 Pinning of Downstream Front
      of Synchronized Flow at Bottlenecks .................. 189
      7.4.4 Transformation Between Widening
      and Localized Synchronized Flow Patterns ............. 192
  7.5 General Patterns ....................................... 194
      7.5.1 Spatiotemporal Structure of General Patterns .... 194
      7.5.2 Dissolving General Pattern
      and Pattern Transformation ........................... 198
  7.6 Physics of General Patterns ............................. 200
      7.6.1 Region of Wide Moving Jams ....................... 200
      7.6.2 Narrow Moving Jam Emergence in Pinch Region .... 204
      7.6.3 Moving Jam Suppression Effect ..................... 209
      7.6.4 Width of Pinch Region ............................ 210
      7.6.5 Wide Moving Jam Propagation Through Bottlenecks 211
  7.7 Conclusions ............................................ 213
8 Freeway Capacity
in Three-Phase Traffic Theory ............................ 217
  8.1 Introduction ........................................... 217
  8.2 Homogeneous Road ..................................... 217
  8.3 Freeway Capacity in Free Flow at Bottlenecks .......... 218
    8.3.1 Definition of Freeway Capacity ................... 218
    8.3.2 Probability for Speed Breakdown at Bottlenecks .... 220
    8.3.3 Threshold Boundary for Speed Breakdown .......... 223
    8.3.4 Features of Freeway Capacity at Bottlenecks ...... 226
  8.4 Z-Characteristic and Probability
      for Speed Breakdown ................................... 228
  8.5 Congested Pattern Capacity at Bottlenecks ............. 232
  8.6 Main Behavioral Assumptions
      of Three-Phase Traffic Theory ........................... 234
  8.7 Conclusions ............................................ 237

Part II Empirical Spatiotemporal Congested
Traffic Patterns

9 Empirical Congested Patterns
at Isolated Bottlenecks ................................... 241
  9.1 Introduction .......................................... 241
  9.2 Effectual Bottlenecks and Effective Locations
      of Bottlenecks ........................................... 242
    9.2.1 Effectual Bottlenecks on Freeway A5-South ........ 244
    9.2.2 Effectual Bottlenecks on Freeway A5-North ........ 247
    9.2.3 Isolated Effectual Bottleneck ..................... 247
  9.3 Empirical Synchronized Flow Patterns .................... 250
    9.3.1 Widening Synchronized Flow Pattern ............. 250
    9.3.2 Localized Synchronized Flow Pattern .............. 255
    9.3.3 Moving Synchronized Flow Pattern ............... 256
  9.4 Empirical General Patterns ................................... 259
    9.4.1 Empirical General Pattern of Type (1) .......... 259
    9.4.2 Empirical General Pattern of Type (2) .......... 262
    9.4.3 Dependence of Effective Location
        of Bottleneck on Time .................................. 264
  9.5 Conclusions ............................................ 268

10 Empirical Breakdown Phenomenon:
  Phase Transition from Free Flow
to Synchronized Flow ...................................... 269
  10.1 Introduction .......................................... 269
  10.2 Spontaneous Breakdown Phenomenon
      (Spontaneous F→S Transition) at On-Ramp Bottlenecks ...... 270
10.3 Probability for $F \rightarrow S$ Transition .......................... 274
  10.3.1 Empirical and Theoretical Definitions of Freeway Capacities at Bottlenecks .......... 275
  10.3.2 Pre-Discharge Flow Rate ........................ 278
10.4 Induced Speed Breakdown at On-Ramp Bottlenecks ............................. 281
  10.4.1 $F \rightarrow S$ Transition Induced by Wide Moving Jam
       Propagation Through Effectual Bottleneck .......... 282
  10.4.2 Induced Speed Breakdown at Bottlenecks
       Caused by Synchronized Flow Propagation .......... 282
10.5 Breakdown Phenomenon at Off-Ramp Bottlenecks ........................ 285
10.6 Breakdown Phenomenon Away from Bottlenecks ........................ 289
10.7 Some Empirical Features of Synchronized Flow ......................... 294
  10.7.1 Complex Behavior in Flow–Density Plane ............... 294
  10.7.2 Three Types of Synchronized Flow .................. 296
  10.7.3 Overlapping States of Free Flow
       and Synchronized Flow in Density ................. 299
  10.7.4 Analysis of Individual Vehicle Speeds ............... 302
10.8 Conclusions ............................................ 302

11 Empirical Features of Wide Moving Jam Propagation ........................ 305
11.1 Introduction ........................................... 305
11.2 Characteristic Parameters of Wide Moving Jams ..................... 305
  11.2.1 Empirical Determination of Line $J$ .................. 306
  11.2.2 Dependence of Characteristic Jam Parameters
       on Traffic Conditions ................................ 310
  11.2.3 Propagation of Wide Moving Jams
       Through Synchronized Flow ........................... 311
  11.2.4 Moving Blanks Within Wide Moving Jams ............... 314
11.3 Features of Foreign Wide Moving Jams .............................. 316
11.4 Conclusions ............................................ 318

12 Empirical Features of Moving Jam Emergence ................................ 321
12.1 Introduction ........................................... 321
12.2 Pinch Effect in Synchronized Flow ................................ 321
  12.2.1 Narrow Moving Jam Emergence ....................... 323
  12.2.2 Wide Moving Jam Emergence ($S \rightarrow J$ Transition) .... 328
  12.2.3 Correlation of Characteristics for Pinch Region
       and Wide Moving Jams ............................... 332
  12.2.4 Frequency of Narrow Moving Jam Emergence .......... 332
  12.2.5 Saturation and Dynamic Features of Pinch Effect .... 334
  12.2.6 Spatial Dependence of Speed Correlation Function .... 335
12.2.7 Effect of Wide Moving Jam Emergence in Pinch Region of General Pattern ............... 337
12.3 Strong and Weak Congestion ........................................... 337
12.4 Moving Jam Emergence in Synchronized Flow Away from Bottlenecks ......................... 340
12.5 Pattern Formation at Off-Ramp Bottlenecks ............................................ 343
12.6 Induced F→J Transition .................................................. 344
12.7 Conclusions ................................................................. 348

13 Empirical Pattern Evolution and Transformation at Isolated Bottlenecks ........................................... 349
13.1 Introduction ................................................................. 349
13.2 Evolution of General Patterns
at On-Ramp Bottlenecks .................................................... 350
13.2.1 Transformation of General Pattern into Synchronized Flow Pattern ......................... 350
13.2.2 Alternation of Free Flow and Synchronized Flow in Congested Patterns ................... 350
13.2.3 Hysteresis Effects Due to Pattern Formation and Dissolution ................................ 352
13.3 Transformations of Congested Patterns Under Weak Congestion ................................... 354
13.4 Discharge Flow Rate and Capacity Drop ..................................... 357
13.5 Conclusions ................................................................. 363

14 Empirical Complex Pattern Formation Caused by Peculiarities of Freeway Infrastructure .... 365
14.1 Introduction ................................................................. 365
14.2 Expanded Congested Pattern .............................................. 366
14.2.1 Common Features ....................................................... 366
14.2.2 Example of Expanded Congested Pattern .................................. 367
14.3 Dissolution of Moving Jams at Bottlenecks ............................................. 369
14.3.1 Dynamics of Wide Moving Jam Outflow .................................... 369
14.3.2 Localized Synchronized Flow Patterns Resulting from Moving Jam Dissolution .......... 371
14.4 Conclusions ................................................................. 372

15 Dependence of Empirical Fundamental Diagram on Congested Pattern Features ................. 373
15.1 Introduction ................................................................. 373
15.1.1 Empirical Fundamental Diagram and Steady State Model Solutions ....................... 373
15.1.2 Two Branches of Empirical Fundamental Diagram ........................................... 374
15.1.3 Line $J$ and Wide Moving Jam Outflow .............................................. 375
15.2 Empirical Fundamental Diagram
and Line $J$ .................................................. 378
15.2.1 Asymptotic Behavior of Empirical
Fundamental Diagrams ................................. 378
15.2.2 Influence of Different Vehicle Characteristics
on Fundamental Diagrams ............................ 383
15.3 Dependence of Empirical Fundamental Diagram
on Congested Pattern Type ............................ 385
15.4 Explanation of Reversed-λ, Inverted-V, and Inverted-U Empirical Fundamental Diagrams .......................... 392
15.5 Conclusions ........................................... 394

Part III Microscopic Three-Phase Traffic Theory

16 Microscopic Traffic Flow Models
for Spatiotemporal Congested Patterns ................. 399
16.1 Introduction ........................................... 399
16.2 Cellular Automata Approach to Three-Phase Traffic Theory ........................... 401
16.2.1 General Rules of Vehicle Motion .................. 401
16.2.2 Synchronization Distance ......................... 402
16.2.3 Steady States ..................................... 403
16.2.4 Fluctuations of Acceleration and Deceleration
in Cellular Automata Models .......................... 405
16.2.5 Boundary Conditions and Model of On-Ramp .... 407
16.2.6 Summary of Model Equations and Parameters ...... 408
16.3 Continuum in Space Model Approach
to Three-Phase Traffic Theory .......................... 408
16.3.1 Vehicle Motion Rules .............................. 408
16.3.2 Speed Adaptation Effect Within Synchronization
Distance .................................................. 409
16.3.3 Motion State Model for Random Acceleration
and Deceleration ........................................... 411
16.3.4 Safe Speed ......................................... 413
16.3.5 2D Region of Steady States ......................... 414
16.3.6 Physics of Driver Time Delays ...................... 415
16.3.7 Over-Acceleration and Over-Deceleration Effects .... 419
16.3.8 Lane Changing Rules .............................. 420
16.3.9 Boundary Conditions and Models of Bottlenecks .. 421
16.3.10 Summary of Model Equations and Parameters .... 425
16.4 Conclusions ........................................... 431

17 Microscopic Theory of Phase Transitions
in Freeway Traffic ........................................... 433
17.1 Introduction ........................................... 433
17.2 Microscopic Theory of Breakdown Phenomenon
   (F→S Transition) ........................................ 434
   17.2.1 Homogeneous Road ................................ 434
   17.2.2 Breakdown Phenomenon at On-Ramp Bottlenecks ... 438
17.3 Moving Jam Emergence and Double Z-Shaped
   Characteristics of Traffic Flow .......................... 442
   17.3.1 F→J Transition on Homogeneous Road .......... 442
   17.3.2 S→J Transition on Homogeneous Road .......... 443
   17.3.3 Moving Jam Emergence in Synchronized
   Flow Upstream of Bottlenecks .......................... 445
17.4 Conclusions ........................................... 448

18 Congested Patterns at Isolated Bottlenecks ............. 449
   18.1 Introduction ........................................ 449
   18.2 Diagram of Congested Patterns
      at Isolated On-Ramp Bottlenecks ..................... 450
      18.2.1 Synchronized Flow Patterns .................... 450
      18.2.2 Single Vehicle Characteristics in Synchronized Flow . 454
      18.2.3 Maximum Freeway Capacities
      and Limit Point in Diagram .......................... 458
      18.2.4 Pinch Effect in General Patterns ............... 458
      18.2.5 Peculiarities of General Patterns .............. 462
   18.3 Weak and Strong Congestion
      in General Patterns .................................. 464
      18.3.1 Criteria for Strong and Weak Congestion ........ 464
      18.3.2 Strong Congestion Features ..................... 467
   18.4 Evolution of Congested Patterns
      at On-Ramp Bottlenecks ................................ 469
   18.5 Hysteresis and Nucleation Effects by Pattern
      Formation at On-Ramp Bottlenecks .................... 471
      18.5.1 Threshold Boundary for Synchronized Flow Patterns 471
      18.5.2 Threshold Boundary for General Patterns .......... 475
      18.5.3 Overlap of Different Metastable Regions
      and Multiple Pattern Excitation ..................... 476
   18.6 Strong Congestion at Merge Bottlenecks ............... 477
      18.6.1 Comparison of General Patterns at Merge Bottleneck
      and at On-Ramp Bottleneck ........................... 477
      18.6.2 Diagram of Congested Patterns .................. 478
   18.7 Weak Congestion at Off-Ramp Bottlenecks .............. 480
      18.7.1 Diagram of Congested Patterns ................. 480
      18.7.2 Comparison of Pattern Features
      at Various Bottlenecks ............................... 480
   18.8 Congested Pattern Capacity
      at On-Ramp Bottlenecks ............................... 483
18.8.1 Transformations of Congested Patterns
at On-Ramp Bottlenecks ........................ 483
18.8.2 Temporal Evolution of Discharge Flow Rate .... 486
18.8.3 Dependence of Congested Pattern Capacity
on On-Ramp Inflow ............................. 490
18.9 Conclusions ............................................ 492

19 Complex Congested Pattern Interaction
and Transformation ...................................... 495
19.1 Introduction ........................................... 495
19.2 Catch Effect and Induced Congested
Pattern Formation ...................................... 496
19.2.1 Induced Pattern Emergence ...................... 496
19.3 Complex Congested Patterns
and Pattern Interaction .................................. 498
19.3.1 Foreign Wide Moving Jams ...................... 498
19.3.2 Expanded Congested Patterns .................... 501
19.4 Intensification of Downstream Congestion
Due to Upstream Congestion ............................ 504
19.5 Conclusions ............................................ 507

20 Spatiotemporal Patterns in Heterogeneous Traffic Flow
.. 509
20.1 Introduction ........................................... 509
20.2 Microscopic Two-Lane Model for Heterogeneous Traffic
Flow with Various Driver Behavioral Characteristics
and Vehicle Parameters ............................... 510
20.2.1 Single-Lane Model ............................... 510
20.2.2 Two-Lane Model ............................... 512
20.2.3 Boundary, Initial Conditions, and Model
of Bottleneck ................................... 514
20.2.4 Simulation Parameters .......................... 515
20.3 Patterns in Heterogeneous Traffic Flow
with Different Driver Behavioral Characteristics .......... 515
20.3.1 Vehicle Separation Effect in Free Flow .......... 515
20.3.2 Onset of Congestion in Free Flow
on Homogeneous Road .............................. 516
20.3.3 Lane Asymmetric Emergence
of Moving Synchronized Flow Patterns ............... 519
20.3.4 Congested Patterns at On-Ramp Bottlenecks ...... 519
20.3.5 Wide Moving Jam Propagation ............................ 525
20.4 Patterns in Heterogeneous Traffic Flow
with Different Vehicle Parameters ........................ 528
20.4.1 Peculiarity of Wide Moving Jam Propagation .... 530
20.4.2 Partial Destroying of Speed Synchronization .... 533
20.4.3 Extension of Free Flow Recovering and Vehicle Separation ............................................ 533
20.5 Weak Heterogeneous Flow ............................................ 535
  20.5.1 Spontaneous Onset of Congestion Away from Bottlenecks ........................................... 535
  20.5.2 Lane Asymmetric Free Flow Distributions ........................................... 537
20.6 Characteristics of Congested Pattern Propagation in Heterogeneous Traffic Flow .................... 538
  20.6.1 Velocity of Downstream Jam Front ........................................... 538
  20.6.2 Flow Rate in Jam Outflow ........................................... 540
  20.6.3 Velocity of Downstream Front of Moving Synchronized Flow Patterns ........................................... 541
20.7 Conclusions ............................................ 542

Part IV Engineering Applications

21 ASDA and FOTO Models of Spatiotemporal Pattern Dynamics based on Local Traffic Flow Measurements .......................... 547
  21.1 Introduction ........................................... 547
  21.2 Identification of Traffic Phases ........................................... 548
  21.3 Determination of Traffic Phases with FOTO Model .......................... 550
    21.3.1 Fuzzy Rules for FOTO Model ........................................... 551
  21.4 Tracking Moving Jams with ASDA: Simplified Discussion ........................................... 554
    21.4.1 Tracking Synchronized Flow with FOTO Model ........................................... 557
    21.4.2 ASDA-Like Approach to Tracking Synchronized Flow ........................................... 559
    21.4.3 Cumulative Flow Rate Approach to Tracking Synchronized Flow ........................................... 560
  21.5 Conclusions ............................................ 561

22 Spatiotemporal Pattern Recognition, Tracking, and Prediction ........................................... 563
  22.1 Introduction ........................................... 563
  22.2 FOTO and ASDA Application for Congested Pattern Recognition and Tracking .......................... 563
    22.2.1 Validation of FOTO and ASDA Models at Traffic Control Center of German Federal State of Hessen ........................................... 563
    22.2.2 Application of FOTO and ASDA Models on Other Freeways in Germany and USA ........................................... 565
  22.3 Spatiotemporal Pattern Prediction ........................................... 568
    22.3.1 Historical Time Series ........................................... 568
23 Control of Spatiotemporal Congested Patterns

23.1 Introduction ........................................... 591
23.2 Scenarios for Traffic Management and Control .......... 592
23.3 Spatiotemporal Pattern Control
   Through Ramp Metering ..................................... 593
23.3.1 Free Flow Control Approach .......................... 595
23.3.2 Congested Pattern Control Approach ................... 606
23.3.3 Comparison of Free Flow
   and Congested Pattern Control Approaches ............. 610
23.3.4 Comparison of Different Control Rules
   in Congested Pattern Control Approach ................. 614
23.4 Dissolution of Congested Patterns ........................ 617
23.5 Prevention of Induced Congestion ........................ 620
23.6 Influence of Automatic Cruise Control
   on Congested Patterns .................................... 624
23.6.1 Model of Automatic Cruise Control ................. 624
23.6.2 Automatic Cruise Control
   with Quick Dynamic Adaptation ........................ 626
23.6.3 Automatic Cruise Control
   with Slow Dynamic Adaptation .......................... 628
23.7 Conclusions ............................................ 629

24 Conclusion ............................................... 631

A Terms and Definitions ..................................... 633
A.1 Traffic States, Parameters, and Variables ............... 633
A.2 Traffic Phases ........................................... 633
A.3 Phase Transitions ....................................... 634
A.4 Bottleneck Characteristics .................................. 635
A.5 Congested Patterns at Bottlenecks ....................... 636
A.6 Local Perturbations ..................................... 637
A.7 Critical and Threshold Traffic Variables ............... 637
A.8 Some Features of Phase Transitions
   and Traffic State Stability .................................. 638
B ASDA and FOTO Models

for Practical Applications ........................................... 641

B.1 ASDA Model for Several Road Detectors .................... 641
  B.1.1 Extensions of ASDA for On-Ramps, Off-Ramps, and Changing of Number of Freeway Lanes
          Upstream of Moving Jam ................................... 643
  B.1.2 Extensions of ASDA for On-Ramps, Off-Ramps, and Changing of Number of Freeway Lanes
          Downstream of Moving Jam ................................ 645
  B.1.3 FOTO Model for Several Road Detectors ................. 646
  B.1.4 Extended Rules for FOTO Model ......................... 646

B.2 Statistical Evaluation of Different 

Reduced Detector Configurations ................................. 651

References ............................................................ 655

Index ................................................................. 679
Acronyms and Conventions

SP: synchronized flow pattern
MSP: moving SP
WSP: widening SP
LSP: localized SP
ASP: alternating SP
GP: general pattern
DGP: dissolving GP
AGP: alternating GP
EP: expanded congested pattern
FCD: floating car data
TCC: traffic control center
UTA: model for traffic prediction in city networks
ASDA: model for automatic tracking of moving jams
FOTO: model for automatic identification of traffic phases and tracking of synchronized flow
CA model: cellular automata traffic flow model
ALINEA: model for automatic feedback on-ramp metering
ANCONA: model for automatic on-ramp control of congested patterns at freeway bottleneck
ACC: automatic cruise control
x: spatial coordinate in direction of traffic flow
t: time
q: flow rate
ρ: vehicle density
v: vehicle speed
d: vehicle length
v_{g(\cdot)}: velocity of downstream front of wide moving jam
q_{out}: flow rate in traffic flow formed by wide moving jam outflow
q_{out}: flow rate in free flow formed by wide moving jam outflow
\rho_{\text{min}}: density in free flow formed by wide moving jam outflow
Acronyms and Conventions

\[ v_{\text{max}} \] average speed in free flow formed by wide moving jam outflow

\[ \rho_{\text{max}} \] density within wide moving jam (jam density)

\[ v_{\text{min}} \] average speed within wide moving jam

\[ L_{J} \] width (in longitudinal direction) of wide moving jam

\[ \text{F} \rightarrow \text{S} \text{ transition} \] phase transition from free flow to synchronized flow

\[ \text{F} \rightarrow \text{J} \text{ transition} \] phase transition from free flow to wide moving jam

\[ \text{S} \rightarrow \text{J} \text{ transition} \] phase transition from synchronized flow to wide moving jam

\[ \text{S} \rightarrow \text{F} \text{ transition} \] phase transition from synchronized flow to free flow

\[ \text{J} \rightarrow \text{S} \text{ transition} \] phase transition from wide moving jam to synchronized flow

\[ \text{J} \rightarrow \text{F} \text{ transition} \] phase transition from wide moving jam to free flow

\[ \text{F} \rightarrow \text{S} \rightarrow \text{J} \text{ transitions} \] F \rightarrow S transition followed by S \rightarrow J transition

\[ P_{\text{FS}} \] probability for F \rightarrow S transition on hypothetical homogeneous road for given observation time \( T_{\text{ob}} \) and given road length

\[ P_{\text{FS}}^{(B)} \] probability for F \rightarrow S transition at freeway bottleneck for given observation time \( T_{\text{ob}} \)

\[ q_{\text{c}}^{(B)} \] freeway capacity in free flow at freeway bottleneck

\[ q_{\text{max}}^{(B)} \] maximum freeway capacity in free flow at freeway bottleneck relative to \( P_{\text{FS}}^{(B)} = 1 \)

\[ q_{\text{th}}^{(B)} \] minimum freeway capacity in free flow at freeway bottleneck

\[ q_{\text{FS}}^{(B)} \] pre-discharge flow rate

\[ q_{\text{out}}^{(\text{bottle})} \] discharge flow rate from congested pattern at freeway bottleneck

\[ q_{\text{pinch}}^{(\text{pinch})} \] average flow rate in pinch region of GP or EP

\[ q_{\text{lim}}^{(\text{pinch})} \] limiting (minimum) flow rate in pinch region of GP or EP

\[ L_{\text{syn}} \] width of synchronized flow region (in longitudinal direction) in congested pattern

\[ q_{\text{on}} \] flow rate to on-ramp

\[ q_{\text{in}} \] flow rate in free flow on main road upstream of on-ramp bottleneck

\[ q_{\text{sum}} \] flow rate downstream under free flow condition at on-ramp bottleneck

\[ \eta \] percentage of vehicles which want to leave main road via off-ramp

\[ \rho_{\text{max}}^{(\text{free}, \text{emp})} \] maximum density relative to empirical limit point for free flow

\[ q_{\text{max}}^{(\text{free}, \text{emp})} \] maximum flow rate relative to empirical limit point for free flow
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{\text{max}}$</td>
<td>maximum density relative to hypothetical limit point for free flow on homogeneous road</td>
</tr>
<tr>
<td>$q_{\text{max}}$</td>
<td>maximum flow rate relative to hypothetical limit point for free flow on homogeneous road</td>
</tr>
<tr>
<td>$T_{\text{av}}$</td>
<td>averaging time interval for traffic variables</td>
</tr>
<tr>
<td>$T_{\text{ob}}$</td>
<td>time interval for observing traffic flow</td>
</tr>
<tr>
<td>$T_{J}^{(\text{wide})}$</td>
<td>mean time between downstream fronts of wide moving jams</td>
</tr>
<tr>
<td>$\tau_{J}$</td>
<td>mean duration of wide moving jams</td>
</tr>
<tr>
<td>$T_{J}$</td>
<td>mean time between narrow moving jams</td>
</tr>
</tbody>
</table>
Traffic congestion is a fact of life for many car drivers. Every morning millions of drivers around the world sit almost motionless in their vehicles for long periods of time as they try to get to work, and then repeat the experience on their journey home in the evening. The same thing often happens when they are driving to the coast or the airport to go on their holidays. They blame other drivers, increasing volumes of traffic and inevitably, roadworks. So what has any of this got to do with physics?

Well, consider every car an “elementary particle” constrained to move along a one-dimensional trajectory. This particle must also obey certain conditions: for example, it must try to get from A to B and it must not collide with other particles! Could the collective behavior of this complex system be responsible for traffic congestion and the various other features associated with traffic flow? Have these phenomena anything in common with synergetics, i.e., with the phenomena of self-organization and pattern formation that have been discovered in a lot of non-equilibrium physical, chemical, and biological spatial systems in recent years? Indeed, should traffic phenomena be considered part of statistical or nonlinear physics?

Over the past 50 years scientists have developed a wide range of mathematical traffic flow models and traffic flow theories to understand these complex nonlinear traffic phenomena (see the classic papers by Pipes [1], Lighthill and Whitham [2], Richards [3], Herman, Montroll, Potts, Rothery [4], Gazis, Herman, Rothery [5], Kometani, Sasaki [6–8], Newell [9–13], Prigogine [14], Payne [15,16], Gipps [17,18], the books by Gazis [19], Leuzbach [20], May [21], Haight [22], Daganzo [23], Prigogine and Herman [24], Wiedemann [25], Whitham [26], Cremer [27], Newell [28], Steierwald and Lapierre [29], the reviews by Hall et al. [30], Gartner et al. (eds.) [31], Wolf [32], Chowdhury et al. [33], Helbing [34,35], Nagatani [36], Banks [37], Nagel et al. [38], and the conference proceedings [39–46]). Clearly traffic flow models must be based on the real behavior of drivers in traffic, and their solutions should show phenomena observed in real traffic.

Traffic on multilane one-way freeways without light signals (the case that will be considered most in this book) is a complex dynamic process. This process unfolds both in space and time. Therefore, to understand this dynamic process, empirical spatiotemporal traffic pattern features should be known. A
A huge number of publications have been devoted to empirical investigations of spatiotemporal freeway traffic features (e.g., [47–202]).

In particular, it has been found that traffic can be either “free” or “congested.” Congested traffic occurs most at freeway bottlenecks. Just as defects and impurities are important for phase transitions in physical systems, so are freeway bottlenecks in traffic flow. The bottleneck can be a result of roadworks, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, etc. The onset of traffic congestion is accompanied by a sharp drop in average vehicle speed. This speed breakdown is called “the breakdown phenomenon” (e.g., [30,37,58,59,61–79,182]). Usually the downstream front of congestion where vehicles accelerate from congested traffic upstream of the bottleneck to free flow downstream of the bottleneck is fixed at the bottleneck. This means that the flow rate within this downstream front does not depend on a spatial coordinate [21,30,64,71,74]. It has been found that the capacity of a congested bottleneck, i.e., after the breakdown phenomenon at the bottleneck has occurred, is often lower than the capacity in free flow before. This phenomenon is called “capacity drop” [30,64,71,74].

The breakdown phenomenon has a probabilistic nature. At the same freeway bottleneck the speed breakdown in an initial free flow is observed at very different flow rates in different realizations (days) (see papers by Elefteriadou et al. [163], Persaud et al. [61], Lorenz and Elefteriadou [60]).

In congested traffic, the “stop-and-go” phenomenon is observed, i.e., a sequence of different moving traffic jams can appear. Note that a moving jam is a localized structure that moves upstream in traffic flow (Fig. 1.1). Within the moving jam the average vehicle speed is very low (sometimes as low as zero), and the density is very high. The moving jam is spatially restricted by the downstream jam front and the upstream jam front. Within the downstream jam front vehicles accelerate from low speed states within the jam to higher speeds in traffic flow downstream of the moving jam. Within the upstream jam front vehicles must slow down to the speed within the jam. Both jam fronts move upstream. Within the jam fronts the vehicle speed, flow rate, and density vary sharply. Moving jams have been studied empirically by many authors, in particular, in classic works by Edie et al. [80,82], Treiterer et al. [85–87] (Fig. 1.1) and Koshi et al. [88–90].

In congested traffic, a wide spread of empirical points in the flow–density plane and two-regime empirical phase diagrams have been found (e.g., [30,55,56,88]); the effect of synchronization of the vehicle speed between different freeway lanes (e.g., [88,114]), “periodic orbits” and many diverse hysteresis phenomena are observed (e.g., [84–87]). Many other observations of congested traffic have been discussed, in particular regarding “capacity drop” caused by the onset of congestion [64], “catastrophe theory” that should explain the onset of congestion [58], “reverse lambda shape of the fundamental diagram” [88], and many others (for a review see [30,35,37]).
However, it is only recently that a “puzzle” of these and many other spatiotemporal features of congested patterns has been solved and these pattern features adequately understood [203–222]. Consequently, earlier traffic flow theories and models cannot explain and predict many of these empirical spatiotemporal traffic pattern features. When these traffic flow models are used for simulations of freeway control and management strategies, the related simulations cannot predict many of the freeway traffic phenomena that would occur through the use of these strategies.

Therefore, in 1996–1999 the author introduced a concept that is now called “synchronized flow” and the related “three-phase traffic theory” [204–211]. In the three-phase traffic theory, besides the “free flow” phase there are two other phases in congested traffic: “synchronized flow” and “wide moving jam.” Thus, there are three traffic phases in this theory:

1. Free flow  
2. Synchronized flow  
3. Wide moving jam

The empirical (objective) criteria that distinguish between the two traffic phases in congested traffic are related to spatiotemporal features of these different traffic phases [208,212,215]. A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or freeway bottlenecks (Fig. 1.2).

Fig. 1.1. A moving jam: traffic dynamics derived from aerial photography. Taken from Treiterer [87]
Two Phases in Congested Traffic

induced F→S transition

synchronized flow

(a)

wide moving jam

(b)

Fig. 1.2. Two traffic phases in congested traffic. (a) Vehicle speed in space and time. (b) A graph of (a) with the “free flow” phase (white), the “synchronized flow” phase (gray), and the “wide moving jam” phase (black). $B_1$, $B_2$, and $B_3$ are the locations of freeway bottlenecks (see explanation in Sect. 9.2.1). Empirical data from June 23, 1998 measured on a section of the A5-South freeway near Frankfurt, Germany. Taken from [214, 218]
In contrast, the downstream front of the “synchronized flow” phase is often fixed at a freeway bottleneck. Within this front vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow. In an example of these objective criteria (Fig. 1.2), the “wide moving jam” phase propagates through the locations of bottlenecks $B_2$ and $B_3$, whereas the downstream front of the “synchronized flow” phase is fixed at freeway bottleneck $B_2$.

The three-phase traffic theory explains the complexity of traffic phenomena based on phase transitions among these three traffic phases, and on their complex nonlinear spatiotemporal features. For example, in Fig. 1.2, the “wide moving jam” phase propagating through the location of bottleneck $B_2$ induces the “synchronized flow” phase at this bottleneck (this induced phase transition is labeled “induced F→S transition” in Fig. 1.2a). Synchronized flow is self-sustaining for a very long time (more than an hour) upstream of bottleneck $B_2$.

Rather than reviewing various approaches to traffic flow models and theories, the main focus of this book is a comprehensive review of reproducible empirical spatiotemporal congested pattern features and their possible engineering applications.¹ The methodology of the empirical study of congested traffic patterns in this book is as follows. Firstly, spatiotemporal congested pattern features and phase transitions, which are the reasons for pattern occurrence, are studied. Only after an empirical spatiotemporal structure of a congested pattern has been found, investigations of local dynamics of traffic flow variables at a freeway location are performed. Examples of this local dynamics are dependencies of the flow rate, vehicle density, and vehicle speed on time measured at the freeway location. It is important in an empirical study that the relationship of the local dynamics to the type of the pattern and to the spatiotemporal pattern dynamics always should be known. Based on the three-phase traffic theory, a theoretical analysis of spatiotemporal congested traffic patterns is considered, so this theory might explain empirical spatiotemporal structures of congested patterns and empirical features of the phase transitions causing pattern emergence.

The focus of the book was chosen for the following reasons. On the one hand, there have been a huge number of empirical studies of local dynamics of congested traffic. On the other hand, the complete spatiotemporal structure of congested patterns is known for only a few of these empirical results. In other words, in most of these empirical investigations the relationship of the local dynamics of congested traffic to the spatiotemporal features of the congested pattern have not been adequately studied. However, an understanding of empirical spatiotemporal dynamics of congested patterns and of empirical features of the phase transitions in traffic forms the basis for any

¹ Note that reproducible empirical congested pattern features are those qualitative features of spatiotemporal congested patterns that are observed on different freeways on many different days. Only such reproducible qualitative features of spatiotemporal congested patterns will be considered in this book.
traffic flow theory. As for previous earlier traffic flow theories and mathematical traffic flow models, it should be noted that recent reviews of traffic flow models cover the subject of mathematical traffic flow modeling adequately [20, 23, 25, 26, 31, 33, 35, 36, 38]. It turns out that only a few of these model results show a correspondence to empirical spatiotemporal pattern features [221] (see the related criticism of earlier traffic flow theories in Sect. 3.3). This also explains why the author questions results of these theories and has therefore introduced the concept “synchronized flow” and the related three-phase traffic theory [205–215, 218].

Nevertheless, due to the great importance of earlier traffic flow theories, Chap. 3 presents historical results for the reader’s consideration. In this chapter, both achievements and a critical analysis of earlier traffic flow theories are discussed (Chap. 3).

However, we first discuss in Chap. 2 the deep connection between traffic flow phenomena and many nonlinear phenomena in physical, chemical, and biological systems that are studied in the nonlinear science called by Haken “synergetics” [223–225]. In this chapter, the main empirical features of spatiotemporal congested patterns and of phase transitions will be briefly introduced. In addition, some terms like metastable state, hysteresis effect, nucleation effect, spontaneous local phase transition, and standard other terms of the nonlinear physics will be defined and applied to traffic flow phenomena. These definitions are necessary to understand empirical spatiotemporal traffic features described in this book.

One of the most important aims of traffic science is to provide an understanding of freeway traffic that can be used for effective traffic management, traffic control, organization, and other engineering applications, which should increase freeway capacity, improve traffic safety, and result in high-quality mobility (e.g., [20, 21, 27]). The physics of empirical congested pattern features provides the basis for all engineering applications. In particular, the detection, tracking, and prediction of spatiotemporal congested traffic patterns, which is possible if the physics of these patterns is understood, is an important research field in traffic technology for efficient freeway management in traffic control centers. An understanding of the spatiotemporal congested pattern features enables us to forecast pattern development. Traffic pattern features give necessary information for efficient collective management strategy. This strategy can include such well-known methods as ramp metering and traffic assignment. In turn, these methods can be used either to dissolve existing congested patterns or to avoid the emergence of congested patterns.

In particular, based on the physics of traffic the FOTO and ASDA models for spatiotemporal pattern recognition, tracking, and prediction have recently been developed [226–235, 237–239] and applied at a traffic control center [234, 235, 237]. Furthermore, a study of spatiotemporal congested patterns shows that these patterns possess some characteristic, i.e., predictable and reproducible features that remain the same over days and even years. This has
The “free flow” (white), “synchronized flow” (gray) and “wide moving jam” (black) traffic phases are shown on a workday over an approximately 30-km long stretch of the A5-North freeway near Frankfurt, Germany. Printed from the online operation of the FOTO and ASDA models based on three-phase traffic theory. Taken from [237]

been used for the development of pattern-based methods for reliable pattern prediction [231] (Figs. 1.3 and 1.4).

An understanding of spatiotemporal traffic pattern features can also be applied to a variety of applications in which drivers play an active role. For example, analogous to a weather forecast that enables people to adjust their behavior based on predicted information, actual and predicted traffic patterns influence individual driver’s reactions and decisions in a similar manner.

Another example of an application of three-phase traffic flow theory is the recent concept of “cooperative driving” [240]. This is an approach to improving safety and efficiency in congested traffic, based on the physics of traffic and on technical improvements in communication and sensor technology. In cooperative driving, an individual vehicle can serve both as sensor and actuator. Traffic-adaptive behavior and harmonious driving can influence the
Fig. 1.4. Application of three-phase traffic theory. The “free flow” (white), “synchronized flow” (gray) and “wide moving jam” (black) traffic phases reconstructed from empirical data with the FOTO and ASDA models. Data from January 27, 2003 measured on the Interstate freeway I405-South, Orange County near Los Angeles, California, USA (12-km section between freeway exits “Bolsa Chica” and “Euclid Street”). Taken from [236]

congested traffic patterns in a positive way, i.e., eliminate or reduce actual congested patterns and/or hinder the emergence of new congested patterns. The various applications of cooperative driving or any kind of driver information and assistance systems are strongly dependent on actual and predicted traffic patterns. To increase individual driver security, information and data communication between neighboring vehicles, including visualization of this information in the car, should be examined [241]. This information can also include actual and future spatiotemporal traffic features that are important for driving safety and comfort. Another interesting aspect is the influence of individual drivers’ decisions on spatiotemporal pattern features. To make all these applications possible and efficient, sufficient information about actual and predicted spatiotemporal congested pattern features is needed. This is only possible if the physics of congested spatiotemporal patterns is understood.

In this book, we limit attention to dynamic traffic phenomena due to spatiotemporal effects determined by intrinsic traffic features, i.e., by drivers’
interactions in traffic. In particular, these intrinsic traffic features play the most important role on freeways. In contrast, in city traffic, light signals and other traffic regulations at road intersections can often almost fully determine traffic dynamics, rather than driver interaction effects (e.g., [19,28,31,244–291,316,324–328]). Traffic dynamics in city traffic that is fully determined by light signals and other traffic regulations at road intersections is beyond the scope of this book.\footnote{The only exception is examined in Sect. 22.4, where some well-known results of traffic dynamics at road intersections are applied for a discussion of traffic prediction in urban areas.} We will also not consider traffic phenomena due to traffic accidents or due to extremely bad weather and road conditions (like icy road) when vehicles cannot move on the road at all.

The book consists of four parts. Each chapter begins with an “Introduction” section where the aims of the chapter are given. Each of the chapters ends with a “Conclusions” section where the main results of the chapter are listed.

In Part I, the following two main topics will be considered:

1. The main empirical features of spatiotemporal congested patterns and phase transitions in traffic flow (Chap. 2). These main empirical features are the basis of all further considerations in Part I. In particular, these features are important for a discussion of previous traffic flow theories and mathematical models (Chap. 3).

2. Three-phase traffic theory, which is covered in Chaps. 4–8. This theory is a qualitative theory. The most important aim of the three-phase traffic theory is to explain in a simple, qualitative way the main empirical features of spatiotemporal congested traffic patterns discussed in Chap. 2. This theory does not use complex mathematical models. However, this theory should both explain empirical freeway traffic features and introduce to a more complex microscopic three-phase traffic theory.

A comprehensive consideration of empirical spatiotemporal freeway traffic patterns is undertaken in Part II. Congested patterns at isolated bottlenecks are considered in Chap. 9. The breakdown phenomenon responsible for congested pattern emergence in free flow is considered in Chap. 10. In Chap. 11, the characteristic parameters of wide moving jams are studied. Diverse effects of wide moving jam emergence is discussed in Chap. 12. Evolution of the patterns and diverse transformations among various patterns that arise when traffic demand and other traffic characteristics vary form the subject of Chap. 13. In Chap. 14, we consider sometimes very complex patterns that result from peculiarities of freeway infrastructure when several neighboring bottlenecks exist on a freeway.

In Part III, a microscopic three-phase traffic theory that can explain congested pattern features will be presented. This theory is based on spatial continuum and discrete-time microscopic models of Kerner and Klenov [329,
330], and on cellular automata models of Kerner, Klenov, and Wolf [331], which fall within the scope of the author’s three-phase traffic theory.

The empirical features and the theory of the spatiotemporal patterns will be the basis for various engineering applications in Part IV, where we consider methods for spatiotemporal traffic pattern reconstruction, tracking, and prediction, the results of online application of these methods at a traffic control center, and a control theory for spatiotemporal congested patterns in freeway traffic.
Part I

Historical Overview
and Three-Phase Traffic Theory
2 Spatiotemporal Pattern Formation in Freeway Traffic

2.1 Introduction

In this chapter, we consider the following topics:

(i) The deep connection between spatiotemporal pattern formation in physics and freeway traffic.
(ii) A brief consideration of measurement techniques based on induction loop detectors installed on freeways and examples of detector arrangement.
(iii) Definitions and some features of free and congested traffic.
(iv) The main empirical features of spatiotemporal congested traffic patterns and of the phase transitions that lead to pattern emergence.
(v) Some terms and definitions that are important for an understanding of traffic flow phenomena.

It must be stressed that both in the overview of traffic pattern features and elsewhere in the book we will pay attention only to those empirical freeway traffic features that have been found to be reproducible on different freeways, and over many days and years of observations. The reproducible character of empirical results will be the main criterion for all further consideration of empirical spatiotemporal pattern features discussed in this book. This is also because only such reproducible empirical results can be considered a “reliable” empirical basis for successful engineering applications and the development of accurate traffic flow theories.

We will see that many reproducible empirical traffic flow phenomena and spatiotemporal traffic pattern features resemble certain nonlinear effects and phenomena in a variety of different nonlinear non-equilibrium (dissipative) physical, chemical, and biological spatially distributed systems. In particular, features of phase transitions and spatiotemporal pattern formation in these systems are very similar to those observed in traffic flow. Therefore, it seems reasonable to use the same terminology to describe phase transitions and pattern formation in traffic flow that has been accepted as state-of-the-art in the very old science of nonlinear phenomena in nonlinear physics.

The brief consideration of the main empirical spatiotemporal traffic features has two aims:
(1) We would like to show at the beginning of the book what kind of empirical nonlinear phenomena that govern the dynamic process known as "traffic" will be discussed in this book.

(2) We would like to explain the empirical basis of three-phase traffic flow theory.

In more detail, such subjects as empirical (objective) criteria for traffic phases and the line $J$ will be considered in Chap. 4. Deeper consideration of the empirical features of freeway traffic will be found in Part II.

It turns out that to understand the empirical features of spatiotemporal congested patterns in freeway traffic, such physics terminology as spontaneous phase transition and induced phase transition is extremely important. It is also important when one applies an understanding of the empirical features of spatiotemporal congested patterns in traffic engineering methods to correct and optimal traffic control. The need to introduce this terminology also stems from the author's attempt to make three-phase traffic flow theory (which will be considered in Chaps. 4–8) understandable to readers who have never been involved in investigations of nonlinear physics and mathematics, and in particular to those readers who are not comfortable with the terminology of nonlinear physics.

In this chapter, based on a discussion of the main empirical features of spatiotemporal congested traffic patterns we apply only some of the terms and definitions of nonlinear physics to traffic flow phenomena. A more detailed account of these and other terms, and of definitions used in this book, can be found in Appendix A.

### 2.2 Traffic and Synergetics

It has already been stressed that the complex dynamic behavior of traffic can be understood only if spatiotemporal features of this process are studied. This is related to the nature of traffic, which occurs both in space and time. Therefore, traffic phenomena measured at one location are often strongly dependent on what happens at another location. Thus, a study of spatiotemporal traffic patterns is the key to understanding of the dynamic "traffic" process.

We will see that spatiotemporal traffic patterns in real traffic flow exhibit such phenomena as phase transitions, hysteresis effects, nucleation effects, and many other nonlinear effects. These effects are also the subject of the nonlinear physics of a huge number of distributed dynamic systems (e.g., [223–225, 333, 334, 336–355]) studied in the nonlinear science called "synergetics" by Haken [223–225].

Because phase transitions, hysteresis effects, and other nonlinear effects of synergetics determine spatiotemporal traffic pattern features, spatiotemporal freeway traffic phenomena may be considered an aspect of synergetics.
2.3 Free and Congested Traffic

2.3.1 Local Measurements of Traffic Variables

There are several measurement techniques of traffic variables (see, e.g., [21,35] and references therein). In this book, we introduce only material necessary for an understanding of the empirical data to be discussed. In particular, results of measurements of traffic variables will be considered, which have been performed at some freeway locations through induction double loop detectors installed at those locations.

Each detector consists of two induction loops spatially separated by a given small distance $\ell_d$. The induction loop registers a vehicle $i$ moving on the freeway by producing a pulse electric current that begins at some time $t_{i,b}$ when the vehicle reaches the induction loop and it ends some time later $t_{i,f}$ when the vehicle leaves the induction loop. The duration of this current pulse

\[ \Delta t_i = t_{i,f} - t_{i,b} \]  

(2.1)

is therefore related to the time taken by the vehicle to traverse the induction loop.

Every vehicle that passes the induction loop produces a related current pulse. This enables us to calculate the gross time gap between two vehicles $i$ and $i + 1$ that have passed the induction loop one after the other:

\[ \tau_{i,i+1}^{(\text{gross})} = t_{i+1,b} - t_{i,b} \]  

(2.2)

We can further calculate the flow rate $q$ as the measured number of vehicles $\Delta N$ passing the induction loop during a given time interval $\Delta T$ (e.g., one
Because there are two different induction loops in each detector, separated by a known distance $\ell_d$ from one another, the detector is able to measure the individual vehicle speed $v_i$. Indeed, due to the distance $\ell_d$ between two loops of the detector, the first (upstream) loop registers the vehicle earlier than the second (downstream) one. Therefore, if the vehicle speed $v_i$ is not zero, there will be a time lag $\delta t_i$ between the current pulses produced by the two detector induction loops when the vehicle passes both. It is assumed that by virtue of the small value of $\ell_d$, the vehicle speed does not change between the induction loops. This enables us to calculate the individual vehicle speed $v_i$:

$$v_i = \frac{\ell_d}{\delta t_i} . \tag{2.4}$$

and the vehicle length $d_i$

$$d_i = v_i \Delta t_i . \tag{2.5}$$

From (2.5) and (2.2) it is possible to calculate the net time gap:

$$\tau_{i,i+1}^{(net)} = \tau_{i,i+1}^{(gross)} - \frac{d_i}{v_i} = \tau_{i,i+1}^{(gross)} - \Delta t_i \tag{2.6}$$

and the net distance (space gap) $g_{i,i+1}$ between two vehicles $i$ and $i+1$:

$$g_{i,i+1} = v_{i+1} \tau_{i,i+1}^{(net)} . \tag{2.7}$$

The individual vehicle speed also enables us to calculate the average (arithmetic) vehicle speed $v$ of $\Delta N$ vehicles passing the detector in time interval $\Delta T$,

$$v = \frac{1}{\Delta N} \sum_{i=1}^{\Delta N} v_i \tag{2.8}$$

and the speed variance

$$\sigma^2 = \frac{1}{\Delta N - 1} \sum_{i=1}^{\Delta N} (v_i - v)^2 . \tag{2.9}$$

The vehicle density (the number of vehicles per unit length of a freeway, e.g., vehicles per km) can be estimated from the relation

$$\rho = \frac{q}{v} . \tag{2.10}$$

However, it should be noted that the vehicle density $\rho$ is related to vehicles on a freeway section of a given length whereas the vehicle speed is measured at the location of the detector only and is averaged over the time interval
In addition, low vehicle speeds can usually be measured to a lower accuracy than higher vehicle speeds. As a result, at higher vehicle densities (lower average vehicle speed), the vehicle density estimated via (2.10) can lead to a considerable discrepancy in comparison with the real vehicle density. For this reason, when discussing results of measurements we will not usually consider higher vehicle densities (more than 70 vehicles/km) estimated by (2.10).

There are also other reasons why the estimation of the density via (2.10) can lead to a considerable mistake at higher vehicle densities. In particular, this can occur when the vehicle speed and flow rate are strongly spatially inhomogeneous. Thus, the averaging of the vehicle speed through the formula (2.8) gives a temporal averaging of the speed at the detector location made during some time interval. If traffic flow is spatially inhomogeneous, this temporal averaging of the speed can give a very different average speed in comparison with a spatial averaging of the speed of vehicles made at a given instant on a freeway section of a given length.

### 2.3.2 Examples of Freeway Infrastructures and Detector Arrangements

Examples of freeway sections with double induction loop detectors are shown in Figs. 2.1 and 2.2. Measurements of traffic variables on these sections will often be considered in this book. For this reason, we now discuss these freeway sections in more detail.

The section of the freeway A5-South (Fig. 2.1) has three intersections with other freeways (I1, “Friedberg,” I2, “Bad Homburger Kreuz,” and I3, “Nordwestkreuz Frankfurt”) where on- and off-ramps are located, which can be considered potential bottlenecks. This section is equipped with 24 sets of
double induction loop detectors (D1, ..., D24) (Fig. 2.1). Each of the sets D4–D6, D12–D15, and D23, D24 consists of three detectors for a left (passing), middle, and right lane, plus detectors for the lanes associated with on-ramps and off-ramps (the detectors on on- and off-ramps will be designated D4-off, D5-on, ..., D24-off-2; see Fig. 2.1). The other sets of detectors are situated on the one-way three-lane road without on- and off-ramps, where each consists of only three detectors.

The section of the freeway A5-North has four intersections with other freeways (I1, “Westkreuz Frankfurt,” I2 “Nordwestkreuz Frankfurt,” I3 “Bad Homburger Kreuz,” and I4 “Friedberg”). This section is equipped with 30 sets of double induction loop detectors (D1, ..., D30) (Fig. 2.2) whose labeling is the same as those for the section of A5-South.

2.3.3 Free Traffic Flow

Empirical Limit (Maximum) Point of Free Flow

Free flow is the most empirically and theoretically investigated traffic phase (e.g., [20, 21]). In particular, it has been found that if empirical data corresponding to free flow is measured at a freeway location, then in the flow–density plane these empirical points for each freeway lane are well described by a certain curve with a positive slope; the average speed is a decreasing density function (Fig. 2.3a–c). This curve is cut off at some empirical limit (maximum) point of free flow [21, 88]

\[
\left( \rho_{\text{max}}^{(\text{free, emp})}, q_{\text{max}}^{(\text{free, emp})} \right) .
\] (2.11)

At this limit (maximum) point, the average vehicle speed has a minimum possible value for free flow \(v_{\text{min}}^{(\text{free, emp})}\):

\[
v_{\text{min}}^{(\text{free, emp})} = \frac{q_{\text{max}}^{(\text{free, emp})}}{\rho_{\text{max}}^{(\text{free, emp})}} .
\] (2.12)
In some cases, there is a trivial reason for the limit point of free flow: traffic demand is low enough. In this case, congestion does not occur at all. Then the limit point of free flow is related to the maximum flow rate in free flow, which is related to the maximum traffic demand. However, when congested traffic occurs on a freeway, there can be several reasons for the existence of the limit point of free flow. We will consider this point in Sect. 10.3.2.

Let us give some explanation to the designation of the empirical limit point of free flow in (2.12) and below in this book. When traffic data is measured at a freeway location, we can find from the analysis of the data in the flow–density plane that there is a limit point of free flow. However, firstly, we do not usually know the reason for this empirical limit point. We can only see that the empirical points are fit by a curve, which is cut off at the limit (maximum) point of free flow. For this reason, independent of the reason for the empirical limit point of free flow, we will use the same designation for the traffic variables related to the empirical limit point of free flow with the upper index “(free, emp).” To find the reason for the empirical limit point in each particular case, the spatiotemporal behavior of traffic measured at many different freeway locations should be studied. Usually only after the spatiotemporal analysis of traffic can the reason for the empirical limit point of free flow be found (see discussion in Sect. 10.3.2). When these empirical limit points for each of the freeway lanes are considered, then they are denoted by \((\rho_{\text{max}}^\text{free, emp}, q_{\text{max}}^\text{free, emp})\), \((\rho_{\text{max}}^\text{free, emp}, q_{\text{max}}^\text{free, emp})\), \((\rho_{\text{max}}^\text{free, emp}, q_{\text{max}}^\text{free, emp})\), and \((\rho_{\text{max}}^\text{free, emp}, q_{\text{max}}^\text{free, emp})\) for the left lane, the middle lane, and the right freeway lane, respectively (Fig. 2.3a–c).

### About Traffic Regulations on German Freeways

We will mostly use empirical data measured over sections of the freeways A5-South and A5-North in Germany (see explanations in Sect. 2.4.8). Note that traffic in Germany has a strong asymmetry between different lanes. In particular, long vehicles do not usually move in the left (passing) lane of the one-way three-lane freeway sections shown in Figs. 2.1 and 2.2. Therefore,

---

1 The empirical limit point of free flow can be related to the onset of congestion in traffic flow. However, there can be several varieties related to the onset of congestion in traffic flow (see Sect. 2.4). Consequently, when the empirical limit point of free flow (2.12) is related to the onset of congestion in traffic flow, there can be several different theoretical explanations for this empirical limit point. To distinguish between these different theoretical cases, we will use different labels for different theoretical limit points of free flow (Chaps. 5 and 17). This should not lead to confusion because in none of these labels for the theoretical limit points of free flow we will use the upper index “(free, emp).” For example, the maximum flow rate and maximum density at the theoretical limit point of free flow for the hypothetical case of a homogeneous road (road without bottlenecks) are denoted by \(q_{\text{max}}^\text{free}\) and \(\rho_{\text{max}}^\text{free}\), respectively.
quantitative parameters of the traffic dynamics in the left lane do not usually
depend on the percentage of long vehicles on the freeway. For this reason, to
show some dynamic effects clearly, we will often use empirical data related
to the left (passing) lane.

Another important regulation is that long vehicles have a special speed
limit, which is considerably lower than for other vehicles. For these reasons,
the limit point in free flow (2.11) is different for the left (passing) (Fig. 2.3a),
middle (Fig. 2.3b), and right lanes (Fig. 2.3c). The latter can be explained by
the different percentage of long vehicles in these lanes and by the mentioned
speed limit for long vehicles, as well as other differences in vehicle parameters
and driver behavior.

In the example in Fig. 2.3 in the left (passing) lane almost no long vehicles
move, in the middle lane the percentage of long vehicles for the data shown
in Fig. 2.3 was about 10%. In the right lane it was about 40%. Thus, the
minimum average vehicle speed (2.12) at this limit point for free flow is
different for the different freeway lanes. In Fig. 2.3, for the left lane the
minimum speed (2.12) was about 100 km/h, for the middle lane it was about
95 km/h and for the right lane it was about 85 km/h. The asymmetry between
different lanes also causes different average vehicle speeds in the freeway lanes
(Fig. 2.3d).

Another well-known effect in free flow related to this asymmetry between
lanes is the following. In free flow, the percentage of vehicles that move in the
left lane increases with increasing in the flow rate. Beginning at some density
of free flow, the percentage of vehicles in the left lane becomes higher than
the percentage of vehicles in the right lane (e.g., [357]) (Fig. 2.3f). This effect
in free flow is probably associated with the following driver behavior. The
average speed in the left (passing) lane is higher than in the right freeway
lane: there are many long vehicles (trucks) in the right lane that move slowly.
When the density is very low, a driver can easily overtake such a slow moving
vehicle and later continues moving in the right lane. When the density in
free flow increases, the probability of passing decreases. For this reason, all
vehicles change to the left lane that want to move faster than these slow
moving vehicles in the right lane. Although the density in the left lane is
considerably higher than in the right lane, the speed in the left lane can be
higher than in the right lane. However, this can occur only up to some limit
density in the left lane. When this density is exceeded, the onset of congestion
is realized.

Note that in other countries there are different traffic regulations in com­
parison with those mentioned for German freeways. In particular, the highest
flow rate can be observed in the middle lane rather than in the left one (see,
e.g., the review by Banks [37]). It is important to note that these and other
traffic regulation rules in different countries do not detract from common
qualitative results about spatiotemporal congested patterns presented in this
book. We will see from the empirical study of congested patterns (Part II)
2.3 Free and Congested Traffic

A5-North, October 9, 1992

![Diagram showing traffic flow and density](image)

**Fig. 2.3.** Explanation of features of free traffic flow [20, 21, 357]. (a–c) Free flow in the flow–density plane for different lanes. (d, e) Average speed (d) and flow rate (e) as functions of time. (f) Fractions of vehicles that move in the left and right lanes at detectors D6 as a function of total flow rate across the freeway. In (a–c, f) data from 7:00 to 12:00 are shown. 1-min data from October 09, 1992 on the A5-North freeway

and from a microscopic theory that explains and predicts these empirical congested pattern features (Part III) that these qualitative results are associated with intrinsic features of the dynamic “traffic” process rather than different traffic regulations and/or possible different vehicle and driver characteristics.

### 2.3.4 Congested Traffic

When the vehicle density is high enough, traffic is usually in a congested regime (e.g., [21, 23–28, 33–35, 39–45]). In this book, traffic phenomena will be defined in terms of their features rather than their genesis. Often congested
traffic is defined as a traffic state merely results from a freeway bottleneck as the inevitable consequence of an upstream flow that exceeds the downstream capacity of the bottleneck. In this definition it seems obvious to suggest that congested flow is not self-organized, but rigidly controlled by external constraints. However, it has been shown that a variety of self-organized processes are responsible for features of real traffic congestion [166, 168, 169, 203, 205, 208–211, 218]. For these reasons, the following definition of congested traffic will be used in this book.

Definition of Congested Traffic

Congested traffic states will be defined as complementary to states of free flow (e.g., [88]). It has already been noted that empirical points related to free flow can be approximately presented by a curve with a positive slope in the flow–density plane (Fig. 2.3a–c). Congested traffic will be defined, therefore, as a state of traffic where the average vehicle speed is lower than the minimum possible speed in free flow $v^{(\text{free, emp})}$ (2.12), which is related to the limit point $(\rho_{\text{max}}^{(\text{free, emp})}, q_{\text{max}}^{(\text{free, emp})})$ (2.11) in free flow (Fig. 2.4a). Thus, empirical points of congested traffic lie to the right of the dashed line $FC$ in the flow–density plane whose slope equals the minimum possible speed in free flow (2.12) (dashed line $FC$ in Fig. 2.4).

In this definition of congested traffic, nothing is said about the origin of the congestion. It is related to empirical facts that a congested state can occur spontaneously also away from bottlenecks and that many diverse self-organizing effects are responsible for traffic congestion [166, 168, 169, 203, 205, 208–211, 218].

2.3.5 Empirical Fundamental Diagram

Although the complexity of traffic is related to the occurrence of spatiotemporal patterns, some traffic features can be understood if average traffic characteristics are considered. Thus, one of the important empirical methods in traffic science is the study of empirical flow–density and speed–density relationships, which are related to measurements of some average traffic variables at some freeway location (e.g., [21]). In particular, the empirical relationship of the average vehicle speed to the vehicle density must be related to an obvious result observed in real traffic flow: the higher the vehicle density, the lower the average vehicle speed.

The product of the vehicle density $\rho$ and the average speed $v$ is the flow rate:

$$q = v \rho .$$  \hspace{1cm} (2.13)

When the flow rate $q$ (2.13) is plotted as a function of the vehicle density $\rho$, we have what is known as the “fundamental diagram of traffic flow.” It is obvious that this curve must pass through the origin (when the density is
2.3 Free and Congested Traffic

Fig. 2.4. Explanations of the definition and complex local dynamics of congested traffic. (a) Free flow (points left of the dashed line FC) and congested traffic (points right of the dashed line FC) in the flow–density plane. (b) Vehicle speed in congested traffic, related to points right of the line FC in (a). The slope of the dashed line FC is equal to the minimum vehicle speed (2.12) at the limit point of free flow (2.11). In (a), data from 9:00 until 15:30 is shown. Within the interval 9:00–12:06 free flow and within the interval 12:07–15:30 congested traffic is observed at D20. 1-min average data measured on March 23, 2001 at the detector D20 on a section of the freeway A5-North (Fig. 2.2). Data that is averaged across all freeway lanes is shown per a freeway lane. The data is qualitatively similar to earlier observed data in other countries (e.g., [64,88]).

zero, so is the flow rate). There are a huge number of papers where different forms of the fundamental diagram have been suggested. In particular, in 1935 Greenshields [358] proposed a linear relationship between the average speed and the density for congested flow, which is used in many traffic flow theories. According to many traffic scientists the empirical fundamental diagram consists of two distinct curves, one with a positive slope for free traffic flow, and the other with a negative slope for congested traffic (e.g., [21,30,55,88]). In particular, Koshi et al. [88] have proposed a reverse-λ-shaped empirical fundamental diagram, which is widely used by many researchers.
The empirical fundamental diagram is a very useful and powerful method for some engineering applications, such as estimation of the freeway capacity and “level of service” in traffic. The latter and many other important applications are summarized in the Highway Capacity Manual, a special report of the Transportation Research Board (Washington, DC) [285,286].

We will see in Chap. 15 that the qualitative form of the empirical fundamental diagram depends strongly both on the type of congested pattern at a freeway bottleneck and on the freeway location where traffic variables are measured inside the pattern.

2.3.6 Complex Local Dynamics of Congested Traffic

Often empirical points measured at a freeway location are considered in the flow–density plane (Fig. 2.4a). It can be seen from this figure that in contrast to points related to free flow, the points for congested traffic show a large spread in the flow–density plane, rather than conforming to a curve (e.g., [88]). These well-known behaviors of congested traffic are usually explained as fluctuations, instabilities, and moving jam formation (e.g., [20,21]).

However, from an analysis of these empirical points in the flow–density plane it is usually not possible to come to a detailed understanding even of the related local traffic dynamics. To understand this, we consider the example of local traffic dynamics of the vehicle speed in congested traffic shown in Fig. 2.4b. This local traffic dynamics is related to empirical points for congested traffic in the flow–density plane in Fig. 2.4a.

It can be seen in Fig. 2.4b that traffic dynamics of the vehicle speed measured at a freeway location can be extremely complex. However, the dynamics cannot be adequately studied with methods of nonlinear dynamic systems whose variables depend only on time. The complex variations in vehicle speed measured at D20 \((x = 18.1 \text{ km})\) are most closely related to previous events in traffic flow at other freeway locations downstream, rather than at D20 only (Fig. 2.4b). To see this, we consider the traffic dynamics at two downstream locations (at detectors D22 and D21) and compare these traffic dynamics with the dynamics at location D20 (Fig. 2.5). For simplicity we restrict this comparison by a consideration of empirical data for the left freeway lane only.

Firstly, it can be seen that at the downstream locations D22 \((x = 20.5 \text{ km})\) and D21 \((x = 19.4 \text{ km})\) very complex local traffic dynamics also occur. Secondly, if the traffic dynamics at D21 and at D20 are compared, one can see that many speed drops measured at D21 are correlated with speed drops at D20. In particular, three different speed drops at D21 are marked by arrows in Fig. 2.5. Two of these speed drops (arrows 1 and 2) occur at 12:31 and 12:35, respectively. The third (arrow 3) occurs at 15:19, i.e., about three hours later. Studying the speed dynamics at D20, we find that the three speed drops (marked by arrows 1, 2, and 3) are correlated with the related speed drops at D21. Indeed, the speed drop marked by arrow 1 at D20 occurs at 12:36,
i.e., 5 min later than the speed drop 1 at the downstream detectors D21. The speed drop 2 at D20 occurs at 12:40, i.e., also 5 min later than the drop 2 at D21. Finally, the speed drop 3 at D20 occurs at 15:24, i.e., also 5 min later than the speed drop 3 at D21. This conclusion is also valid for some other speed drops at D20 and D21.

This confirms the assumption that the traffic dynamics at the upstream location D20 \((x = 18.1 \text{ km})\) has a strong spatiotemporal correlation with the traffic dynamics at the downstream location at D21 \((x = 19.4 \text{ km})\). Furthermore, it is important to note that the time delay between all correlated drops in vehicle speed at D21 and at D20 is almost independent of time. It is about 5 min during the three hours long speed distributions shown in Fig. 2.5.
The distance between detectors D21 and D20 is 1.3 km. The time delay between the speed drops is about 5 min. Thus, the velocity of speed drop propagation is nearly the same for different drops in speed during all three hours of traffic dynamics in Fig. 2.5: \(v_{\text{prop}} \approx -15.6 \text{ km/h} \).

Some of the speed drops at D20 do not correlate with the speed drops at D21. This can suggest that complex spatiotemporal dynamics also occurs between these detectors, leading to the emergence of new speed drops at D20 and to the disappearance of some other speed drops that can still exist at D21. We will see below that the speed drops under consideration are related to moving traffic jam dynamics, in particular moving jam propagation (Sect. 12.5).

It is difficult to understand the dynamics of this moving jam propagation and the moving jam dynamics if empirical points measured at different freeway locations are studied only in the flow–density plane. Indeed, if the traffic dynamics at D22 and at D21 (Fig. 2.6) were compared with the traffic dynamics at D20 (Fig. 2.4a) only in the flow–density plane (i.e., without consideration of the spatiotemporal speed distributions), one can see that moving jam propagation could not be found in the flow–density plane. It is related to the obvious fact that there is no time coordinate in the flow–density plane. For this reason, aforementioned time delay between different moving jams and other spatiotemporal correlation between different nonlinear effects in traffic are almost impossible to find and follow in the flow–density plane.

**Fig. 2.6.** Traffic dynamics at D22 and D21 in the flow–density plane (left lane) for a comparison with the traffic dynamics at D20 shown in Fig. 2.4a. Points for congested traffic (right of the dashed line FC) are related to the time interval 12:20–15:30 shown in Fig. 2.5
2.4 Main Empirical Features of Spatiotemporal Congested Patterns

2.4.1 Three Traffic Phases

A traffic phase is a traffic state considered in space and time that possesses specific empirical spatiotemporal features. These features are specific (unique) only to this traffic phase. Note that a traffic state is characterized by a certain set of statistical properties of traffic variables. Examples of traffic variables are the flow rate $q$ (vehicles/h), vehicle speed $v$ (km/h), vehicle space gap $g$ (m), and vehicle density $\rho$ (vehicles/km).

Based on investigations of congested spatiotemporal patterns measured on different freeways over many days and years [203, 207, 208, 210, 218], the three-phase traffic theory [205, 207–211] suggests that besides the “free flow” traffic phase there are two other traffic phases in congested traffic: “synchronized flow” and “wide moving jam.”

Traffic flow consists of free flow and congested traffic. Congested traffic consists of the “synchronized flow” phase and the “wide moving jam” phase. Thus, there are three traffic phases in three-phase traffic theory:

- Free flow
- Synchronized flow
- Wide moving jam

To distinguish between the “synchronized flow” phase and the “wide moving jam” phase, the following objective criteria for the traffic phases in congested traffic are applied [205, 208, 212, 213, 215]. These empirical criteria are related to some characteristic spatiotemporal empirical features of the traffic phases:

[J] The “wide moving jam” traffic phase is a moving jam that maintains the mean velocity of the downstream front of the jam, $v_g$, as the jam propagates. Vehicles accelerate within the downstream jam front from low speed states (sometimes as low as zero) inside the jam to higher speeds downstream of the jam. On average a wide moving jam maintains the mean velocity of the downstream jam front, even as it propagates through other (possible very complex) traffic states or freeway bottlenecks [166, 205, 208, 212, 213, 215, 218]. This is a characteristic feature of the “wide moving jam” traffic phase.

[S] The “synchronized flow” traffic phase is defined as follows. In contrast to the “wide moving jam” traffic phase, the downstream front of the “synchronized flow” traffic phase does not maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often fixed at a freeway bottleneck. In other words, the “synchronized flow” traffic phase does not show aforementioned characteristic feature of the “wide moving jam” traffic phase.
The downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front.

It must be noted that there are at least two well-known empirical effects in congested traffic: 1. Synchronization of the average vehicle speed between different freeway lanes (e.g., [88,114]). 2. A wide spread of empirical data in the flow–density plane (e.g., [88]). These effects can occur in both “synchronized flow” and “wide moving jam” traffic phases. To distinguish between these traffic phases, the objective criteria \[J\] and \[S\] should be applied rather than the speed synchronization effect or the wide spread of empirical data in the flow–density plane.

One example of application of these objective criteria, to distinguish between the “synchronized flow” phase and the “wide moving jam” phase, is shown in Fig. 1.2. The downstream front of synchronized flow is fixed at the bottleneck. In contrast, a wide moving jam propagates through this bottleneck with the mean velocity of the downstream jam front remaining unchanged.

Another example is shown in Fig. 2.7. It can be seen that a sequence of two moving jams propagates through different states of traffic flow and through a bottleneck while maintaining the downstream jam front velocity. Therefore, these moving jams belong to the “wide moving jam” phase. In contrast, there is a congested traffic where speed is much lower than in free flow (compare vehicle speeds in Fig. 2.7c,d). The downstream front of this congested traffic flow, where vehicles accelerate to free flow, is fixed at the bottleneck (dashed line in Fig. 2.7a). Therefore, this congested traffic belongs to the “synchronized flow” phase.

We can see that whereas in wide moving jams both the speed and flow rate are very low (sometimes as low as zero), in synchronized flow the flow rate is high (compare the flow rates within the wide moving jams and within synchronized flow in Fig. 2.7d). The vehicle speed in synchronized flow is considerably lower than in free flow. However, the flow rates in the “free flow” and “synchronized flow” phases can be close to one another (Fig. 2.7c,d).

### 2.4.2 Characteristic Parameters of Wide Moving Jams

#### The Line \( J \)

The constancy of the mean velocity \( v_g \) of the downstream front of a wide moving jam during jam propagation is the characteristic feature of the “wide moving jam” phase. This characteristic feature can be shown by a line in the flow–density plane. This line is called the line \( J \) [166,203,367] (Fig. 2.8).
The line $J$ represents the steady propagation of the downstream front of a wide moving jam in the flow-density plane. The slope of the line $J$ is equal to the characteristic velocity $v_g$. The right coordinates of the line $J$ are related to the traffic variables within the jam, the density $\rho_{\text{max}}$ and the average vehicle speed $v_{\text{min}}$. For simplicity we assume here that $v_{\text{min}} = 0$. The left coordinates of the line $J$ are related to the traffic variables in the wide moving jam outflow.

We restrict our attention in this section to the case in which free flow occurs in the jam outflow. Then the flow rate $q_{\text{out}}$, the density $\rho_{\text{min}}$, and the

---

2 A discussion of a relationship between the line $J$ and empirical fundamental diagram appears in Sects. 15.1.3 and 15.4.
average speed $v_{\text{max}} = \frac{q_{\text{out}}}{\rho_{\text{min}}}$ in this jam outflow are also characteristic parameters. These parameters do not depend on initial conditions (density, speed, and flow rate). The mean values of these characteristic parameters as well as the mean value of the velocity $v_g$ do not depend on time during wide moving jam propagation. These characteristic parameters are the same for different wide moving jams. This is true when “control” parameters of traffic like weather, percentage of long vehicles, and other road conditions do not change.

Characteristics of Wide Moving Jam Outflow and Free Flow Metastability

The important feature of a wide moving jam is that the flow rate $q_{\text{out}}$ in the wide moving jam outflow can be considerably lower than the maximum flow rate in free flow $q_{\text{max}}^{(\text{free, emp})}$ (Fig. 2.8):

$$q_{\text{max}}^{(\text{free, emp})} > q_{\text{out}}.$$  \hspace{1cm} (2.14)

To understand the importance of this feature, let us consider free flow where the flow rate $q^{(\text{free})}$ satisfies the condition

$$q^{(\text{free})} > q_{\text{out}}.$$  \hspace{1cm} (2.15)

In this case, wide moving jams and free flow can exist. To see this, we assume that in free flow, where the condition (2.15) is satisfied, a wide moving jam
propagates. Then the flow rate in this free flow $q^{\text{free}}$ is the flow rate $q_{\text{in}}$ in the inflow into the wide moving jam

$$q_{\text{in}} = q^{\text{free}}.$$  

(2.16)

The flow rate in the jam outflow is equal to $q_{\text{out}}$. Corresponding to (2.16) and (2.15), the flow rate in the jam inflow is higher than the flow rate in the jam outflow (Fig. 2.9):

$$q_{\text{in}} > q_{\text{out}}.$$  

(2.17)

Thus, the jam width $L_{J}$ (in the longitudinal direction) increases over time. This means that a wide moving jam can be excited in free flow under the condition (2.17). In other words, this free flow is in a metastable state with respect to wide moving jam emergence [166, 203, 367, 403].

![Fig. 2.9. Empirical example of the flow rate in a wide moving jam that propagates through free flow. $q_{\text{in}}$ is the flow rate in the jam inflow, $q_{\text{out}}$ is the flow rate in the jam outflow. $q_{\text{in}} > q_{\text{out}}$ (2.17). The propagation of this wide moving jam is shown in Fig. 1.2. The flow rate is the total flow rate across the three-lane freeway, which is measured at freeway location $x = 7.9 \text{ km}$ (see the distance axis in Fig. 1.2)](image)

In contrast, if

$$q_{\text{in}} < q_{\text{out}},$$  

then the jam width $L_{J}$ decreases over time: the wide moving jam dissolves. This means that under the condition (2.18), no wide moving jam can be excited or exist for a long time in the free flow. In other words, this free flow is stable against wide moving jam emergence.

However, empirical investigations have shown that wide moving jams do not emerge spontaneously in free flow [208]. Wide moving jams can emerge spontaneously only in the “synchronized flow” phase. This contradiction in empirical observations of the metastability of free flow (2.17) with respect to wide moving jam formation, on the one hand, and the fact that wide moving jams do not emerge spontaneously in free flow, on the other, is one of the very important and complex features of traffic flow.
2.4.3 Spontaneous Breakdown Phenomenon (Spontaneous F→S Transition)

The onset of congestion in an initial free flow is accompanied by a sharp decrease in average vehicle speed in the free flow to a considerably lower speed in congested traffic (Figs. 2.10 and 2.11). This speed breakdown occurs mostly at freeway bottlenecks and is called the breakdown phenomenon (see, e.g., [30, 61, 64, 71, 74, 163]).

The breakdown phenomenon (F→S transition) usually occurs at the same freeway bottlenecks of a freeway section. These bottlenecks are called effectual bottlenecks. An effectual bottleneck is a bottleneck where an F→S transition most frequently occurs on many different days. Examples of effectual bottlenecks are adjacent bottlenecks B₁, B₂, and B₃ in Fig. 1.2b.

It has been found that the breakdown phenomenon has a probabilistic nature [61, 163]: the mean probability of the breakdown phenomenon at a bottleneck is an increasing function of flow rate (Fig. 2.12).

![Image](image_url)

**Fig. 2.10.** Average speed on the main road in space and time during and after the breakdown phenomenon at an on-ramp bottleneck. Empirical 1-min data from March 26, 1996 is averaged across all lanes of the three-lane (one-way) freeway A5-South in Germany in the vicinity of the on-ramp bottleneck.
2.4 Main Empirical Features of Congested Patterns

(a) free flow
congested traffic
(synchronized flow)

(b) 1000
0 06:38 06:58 07:18
flow rate [veh/h]
time

Fig. 2.11. Breakdown phenomenon at an on-ramp bottleneck. Vehicle speed (a) and flow rate (b) as functions of time related to Fig. 2.10. Empirical 1-min average data measured at detectors on the main road in the vicinity of the on-ramp bottleneck during the breakdown phenomenon. Data is averaged across all three freeway lanes. This empirical example of the breakdown phenomenon is qualitatively the same as many other examples of the well-known breakdown phenomenon, which has been observed in various countries (e.g., [30, 61, 64, 71, 74, 163])

Fig. 2.12. Probability for the breakdown phenomenon (F→S transition) at an on-ramp bottleneck for various flow rate averaging time constants \( T_{av} \) (\( T_{av} = 3, 5, 10, 15 \) min). Data is measured on a freeway in Toronto, Canada. Taken from Persaud et al. [61]

Congested traffic in Fig. 2.11 that has occurred due to the breakdown phenomenon shows a further development in space and time (Fig. 2.10), with the following effects:

(i) The upstream front of congested traffic propagates upstream. The upstream front separates free flow upstream from congested traffic downstream. Thus, the region of congestion broadens in the upstream direction.
(ii) The downstream front of congested traffic is fixed at the bottleneck (dashed line in Fig. 2.10). Corresponding to the objective criteria for traffic phases in congested traffic considered above, this means that the congested traffic resulting from the breakdown phenomenon belongs to the “synchronized flow” phase.

In other words, the well-known breakdown phenomenon is associated with the phase transition from the “free flow” phase to the “synchronized flow” phase (F→S transition) [167, 203, 208]. For this reason, the breakdown phenomenon is labeled “F→S transition” in Figs. 2.10 and 2.11. Thus, the terms “F→S transition,” “breakdown phenomenon,” and “speed breakdown” are synonyms related to the same effect: the onset of congestion in free flow. In this book, the term “F→S transition” is often used because it reflects the physical reason for the breakdown phenomenon.

The breakdown phenomenon (F→S transition) in Figs. 2.10 and 2.11 is caused by an *internal* local disturbance in traffic flow. Indeed, there are no *external* disturbances in traffic flow responsible for this phase transition. For this reason, it is a spontaneous breakdown phenomenon (spontaneous F→S transition).

The most common reason for an internal local disturbance in traffic flow is a freeway bottleneck: on the main road in the vicinity of the bottleneck, vehicle speed decreases locally while vehicle density increases. This is the reason for a deterministic (permanent) local perturbation in traffic flow at the bottleneck (Sect. 5.3). There are also many reasons for a random internal local disturbance in traffic flow, such as an unexpected decrease in the speed or lane-changing of one of the vehicles. The related random local perturbation can occur at the bottleneck or away from bottlenecks. In the latter case, the spontaneous breakdown phenomenon (spontaneous F→S transition) can occur away from bottlenecks. This phenomenon is observed in real traffic flow (Sect. 10.6). However, speed breakdown away from bottlenecks occurs very seldom in comparison with speed breakdown at freeway bottlenecks.

It can be seen in Fig. 2.11 that whereas there is a sharp decrease in average vehicle speed, the flow rate does not necessarily abruptly decrease during an F→S transition. This is an important feature of the breakdown phenomenon. In contrast, if a moving jam is produced, then both the speed and flow rate decrease sharply within the moving jam (Fig. 2.9). We have already mentioned this important feature of a moving jam in discussing Fig. 2.7 above.

### 2.4.4 Induced Breakdown Phenomenon

In Figs. 2.10 and 2.11, the spontaneous breakdown phenomenon (spontaneous F→S transition) at a freeway bottleneck is shown. The spontaneous F→S transition occurs without *additional external disturbance* of an initial free flow at the bottleneck. However, there can also be an induced breakdown phenomenon (induced F→S transition) [205, 214]. The induced F→S
2.4 Main Empirical Features of Congested Patterns

transition is caused by a short-time *external* disturbance in traffic flow. This external disturbance can be related to the propagation of a moving spatiotemporal congested pattern that initially occurs at a *different* freeway location than that of the induced F→S transition.

In particular, the breakdown phenomenon (F→S transition) can be induced by a wide moving jam propagating upstream through a bottleneck [214]. This case is shown in Fig. 1.2a: propagating through the bottleneck where free flow has already been, the wide moving jam causes synchronized flow emergence at the bottleneck. Synchronized flow remains at the bottleneck even after the wide moving jam is far upstream of the bottleneck. The induced phase transition is labeled “induced F→S transition” in this figure. In this case, after the wide moving jam has passed the bottleneck a *synchronized flow pattern* (SP for short) is formed. The SP remains at the bottleneck for a long time (this SP is labeled “synchronized flow” in Fig. 1.2a).

Induced speed breakdown (induced F→S transition) at a bottleneck can also occur when a region of synchronized flow first occurs downstream of this bottleneck, and the region later reaches the bottleneck due to the upstream propagation of synchronized flow. However, in contrast to the above case of induced speed breakdown caused by wide moving jam propagation, the initial synchronized flow is caught at the bottleneck [218] (see this induced speed breakdown and the catch effect in Sect. 2.4.6 below).

2.4.5 Synchronized Flow Patterns

The F→S transition shown in Figs. 2.10 and 2.11 can lead to the emergence of a synchronized flow pattern (SP) at the bottleneck. The SP is a congested pattern that consists of synchronized flow only, i.e., no wide moving jams emerge in an SP.

There are three main types of SPs [218]:

(1) Localized SP (LSP for short): The downstream front of an LSP is fixed at the bottleneck. The upstream front of the LSP does not continuously propagate upstream over time: this front is localized at some distance upstream of the bottleneck. In other words, the region of synchronized flow in the LSP is localized in space. However, the location of the upstream front of synchronized flow in the LSP and therefore the LSP width (in the longitudinal direction) can exhibit complex oscillations over time. Note that the upstream front of synchronized flow separates free flow upstream from synchronized flow downstream of this front. Vehicles must slow down within the upstream front of synchronized flow.

(2) Widening SP (WSP for short): As in the LSP, the downstream front of an WSP is fixed at the bottleneck. In contrast to the LSP, the upstream front of the WSP propagates upstream continuously over time. In other words, the width of synchronized flow (in the longitudinal direction) in the WSP widens in the upstream direction. It must be noted that due to this
widening, the upstream front of synchronized flow must eventually reach an upstream adjacent effectual bottleneck. At this upstream bottleneck another congested pattern either already exists or can be induced. In both cases, the WSP is caught at this upstream bottleneck (catch effect; see below). Thus, in real traffic the WSP can exist only for a finite time, before the WSP reaches the upstream effectual bottleneck.

(3) Moving SP (MSP for short): In contrast to the LSP and WSP, an MSP is a localized pattern that propagates as a whole pattern on the freeway over time: both downstream and upstream MSP fronts propagate on the freeway. The MSP propagates between freeway bottlenecks: if the MSP propagates upstream and the MSP reaches an upstream adjacent effectual bottleneck, the MSP is caught at the bottleneck (catch effect). Thus, the MSP that propagates upstream can exist only for a finite time, before the MSP reaches the upstream bottleneck.

Empirical examples of these SPs and the catch effect will be considered below.

An example of an LSP is shown in Fig. 2.13. This LSP is the result of the subsequent development of synchronized flow shown in Fig. 2.10. It can be seen that there are no wide moving jams in the LSP. Moreover, we can see that whereas the speed is considerably lower within the LSP (Fig. 2.13a), the flow rate is close to the flow rate in free flow (Fig. 2.13b, where the increase in flow rate downstream of the bottleneck is related to the on-ramp inflow).

The upstream front of synchronized flow in the LSP (Figs. 2.13a and 2.14) does not continuously propagate upstream of the bottleneck over time: this upstream front is localized at some finite distance upstream of the bottleneck. This distance is a function of time. In other words, this SP is indeed an empirical example of an LSP.

The LSP (Figs. 2.13a and 2.14) exists for about one hour at the bottleneck: at 7:40 free flow occurs at the bottleneck. This restoration of free flow is related to a reverse phase transition from the “synchronized flow” phase to the “free flow” phase (S→F transition for short) at the bottleneck. The F→S transition (speed breakdown) and the S→F transition are accompanied by a hysteresis effect and hysteresis loop in the flow–density plane (Fig. 2.15).

For a quick overview of congested patterns and an estimation of their spatiotemporal structure we often use such a simple representation of traffic phases shown in the time–space plane (see also Figs. 2.16–2.19, 2.21, and 2.23–2.27 below). In these figures, free flow (white), synchronized flow (gray), and moving jams (black) are shown. This simple illustration of the spatiotemporal behavior of congested patterns is made based on a spatiotemporal analysis of the average vehicle speed and flow rate in space and time using the objective criteria for traffic phases [J] and [S] discussed above. It must be noted that these figures are only approximate spatiotemporal representations of the traffic phases. In the figures, synchronized flow is related to the average speed lower than 75 km/h; moving jams are related to the average speed lower than 25 km/h and the flow rate lower than 700 vehicles/h per lane.
2.4 Main Empirical Features of Congested Patterns

breakdown phenomenon (F → S transition)

synchronized flow pattern

Fig. 2.13. Empirical example of a localized synchronized flow pattern (LSP) at on-ramp bottleneck $B_3$. (a) Averaged vehicle speed in the LSP in space and time. (b) Total flow rate across the freeway in space and time. The bottleneck is the same as the bottleneck labeled $B_3$ in Fig. 1.2. The dashed line in (a) shows the location of the bottleneck. Traffic data from March 26, 1996 measured on a section of the freeway A5-South in Germany. Taken from [218]

Fig. 2.14. Overview of an LSP in space and time for the LSP shown in Fig. 2.13a. Free flow (white), synchronized flow (gray). Taken from [218]

2.4.6 Catch Effect

Let us consider a region of synchronized flow downstream of a freeway bottleneck. If this region of synchronized flow propagates further upstream and reaches the bottleneck, then it is caught at the bottleneck [218]. This catch effect is inconsistent with a characteristic feature of wide moving jam propagation through a bottleneck, namely maintaining the velocity of the downstream jam front. Thus, the catch effect can also be used to distinguish the “synchronized flow” phase from the “wide moving jam” phase in congested traffic.
Fig. 2.15. Empirical example of the hysteresis effect in the flow–density plane due to an F→S transition (“F→S transition” arrow) and a reverse S→F transition (“S→F transition” arrow) on the freeway A5-South near an on-ramp bottleneck. Data from March 26, 1996. Free flow (black squares), synchronized flow (circles). The data averaged across all freeway lanes is shown per freeway lane. This hysteresis effect is qualitatively similar to the numerous hysteresis effects observed first at onset and then dissipation of congestion, as previously observed near on-ramp bottlenecks and other freeway bottlenecks in various countries (see references in the reviews [30,37]).

Propagation of MSP and Induced Speed Breakdown

An empirical example of the catch effect causing an induced F→S transition is shown in Fig. 2.16. We consider two adjacent bottlenecks $B_1$ and $B_2$. The bottleneck downstream will be called “downstream bottleneck” and the bottleneck upstream will be called “upstream bottleneck.” Synchronized flow first occurs spontaneously at the downstream bottleneck $B_1$ (this speed breakdown is labeled “spontaneous F→S transition”). The synchronized flow subsequently propagates upstream as a distinct localized structure. This is an example of an MSP (MSP is labeled “MSP” in Fig. 2.16).

When the MSP reaches the upstream bottleneck $B_2$, it does not propagate through the bottleneck as a localized structure: the MSP is caught at the bottleneck (catch effect). This is in contrast to wide moving jam propagation, as shown in Figs. 1.2 and 2.7a.

When the MSP has not yet reached bottleneck $B_2$, free flow is there (Fig. 2.16). After the MSP is caught at the bottleneck, synchronized flow occurs at bottleneck $B_2$. This synchronized flow continues for a long time (more than 2 hours) at bottleneck $B_2$. The upstream front of synchronized flow is now localized at some finite distance upstream of bottleneck $B_2$. This means that after the MSP has induced another SP at the bottleneck, instead of MSP propagation through the bottleneck, the MSP is caught at the
2.4 Main Empirical Features of Congested Patterns

A5-South, April 20, 1998

Fig. 2.16. Empirical example of the catch effect with induced speed breakdown at bottleneck $B_2$. Free flow (white), synchronized flow (gray), and moving jams (black). Bottleneck $B_2$, the downstream bottleneck $B_1$, and the upstream bottleneck $B_3$ are the same as those in Fig. 1.2. See an explanation of these bottlenecks in Sect. 9.2.1. Data from the A5-South freeway near Frankfurt, Germany. Taken from [218]

bottleneck. Thus, the catch of the MSP at the bottleneck causes an induced speed breakdown at the bottleneck (these two effects are labeled “catch effect and induced $F \rightarrow S$ transition” in Fig. 2.16).

Propagation of WSP and Induced Speed Breakdown

A different example of the catch effect is shown in Fig. 2.17. In this case, first synchronized flow due to a spontaneous speed breakdown is formed at the downstream bottleneck $B_{North_1}$. The downstream front of this synchronized flow is fixed at this bottleneck. The upstream front of synchronized flow propagates continuously upstream over time. Thus, this is an empirical example of an WSP (Fig. 2.17a).

When the upstream front of the WSP reaches the upstream bottleneck $B_{North_2}$, this synchronized flow causes an induced speed breakdown at this upstream bottleneck. Synchronized flow is localized at some distance from the bottleneck $B_{North_2}$, rather than the upstream front of this synchronized
flow propagating further upstream continuously over time. This means that the initial WSP is caught at the upstream bottleneck. This catch effect causes the induced speed breakdown at the bottleneck (Fig. 2.17b).

![Fig. 2.17. Empirical examples of an WSP (a) and the catch effect (b). Synchronized flow in the initial WSP (a) is caught at the upstream bottleneck (b). Free flow (white), synchronized flow (gray). Data from March 23, 2001 on the freeway A5-North near Frankfurt, Germany. Taken from [218]]

Thus, the WSP exists only for a relatively short time, which is determined by the propagation of the upstream front of the WSP to the upstream bottleneck. In this case, the lifetime of the WSP is about 20 min. After synchronized flow of the WSP is caught at the upstream bottleneck $B_{\text{North} 2}$, a new congested pattern is formed. Synchronized flow in this pattern affects both bottlenecks $B_{\text{North} 1}$ and $B_{\text{North} 2}$. This congested pattern is called an expanded congested pattern, or EP for short (Fig. 2.17b).

It must be noted that over time the average speed in synchronized flow of an WSP can decrease, and wide moving jams can emerge spontaneously. In other words, the WSP can transform into another type of congested pattern (see Sect. 2.4.7 below).

**MSP Emerging between Bottlenecks and Catch Effect**

Another empirical example of the catch effect is shown in Fig. 2.18. A local region of synchronized flow occurs between the bottleneck $B_{\text{North} 2}$ and the upstream bottleneck $B_{\text{North} 3}$ (this speed breakdown is labeled “F→S” in Fig. 2.18). A localized region of synchronized flow departs from where the initial synchronized flow has emerged: an MSP is formed. As in the case considered above (Fig. 2.16), this MSP is a local moving SP, where both the downstream and upstream fronts of synchronized flow propagate upstream. Due to the upstream propagation of the MSP, it eventually reaches the upstream bottleneck $B_{\text{North} 3}$. As a result, the MSP is caught at the bottleneck:
we can see that rather than the MSP propagating upstream of the bottleneck $B_{\text{North } 3}$, the upstream front of synchronized flow at this bottleneck is now localized at some finite distance upstream of $B_{\text{North } 3}$.

![Diagram](image)

**Fig. 2.18.** Empirical example of the catch effect. Emergence and propagation of synchronized flow in space and time. Free flow *(white)*, synchronized flow *(gray)*. Data from March 23, 2001 on the freeway A5-North near Frankfurt, Germany. Taken from [218]

From these two empirical examples it can be seen that there is a very important difference from wide moving jam propagation shown in Fig. 2.7a: after the jams have reached the bottleneck due to their upstream propagation (the bottleneck location is shown by dashed line in Fig. 2.7a), the jams propagate through the bottleneck and through synchronized flow upstream of the bottleneck while maintaining the velocity of the downstream jam fronts.

### 2.4.7 Moving Jam Emergence in Synchronized Flow: General Pattern

We have already mentioned that wide moving jams do not emerge spontaneously in free flow: no spontaneous phase transition from the “free flow” phase to the “wide moving jam” phase (F$\rightarrow$J transition for short) has been observed [208,218].
Wide moving jams emerge \textit{spontaneously only} in the “synchronized flow” phase [208,218]. In synchronized flow at higher speeds, wide moving jams should not necessarily emerge spontaneously. The lower the speed and the higher the density in synchronized flow, the more likely is spontaneous moving jam emergence in that synchronized flow.

Let us assume that a moving jam emerges spontaneously in synchronized flow. This moving jam can later transform into a wide moving jam. As a result, a \textit{general pattern} (GP) appears. This means that an initial SP at the bottleneck transforms into an GP. The GP is a congested pattern, which consists of synchronized flow upstream of an effectual bottleneck and wide moving jams that emerge spontaneously in that synchronized flow [208,218]. Thus, in the GP there are two traffic phases of congested traffic, the “synchronized flow” phase and the “wide moving jam” phase.

**Transformation of SP into GP**

One scenario of GP emergence involves the transformation of an initial SP into an GP [218,221]. Figure 2.19 presents an empirical example of the transformation of an initial WSP into an GP. Firstly, the WSP occurs at bottleneck $B_{\text{North 1}}$, as explained in the discussion of Fig. 2.17a. The average speed in synchronized flow slowly decreases and the density increases over time.

![Empirical example of a general pattern (GP). Transformation of the initial WSP in Fig. 2.17 into an GP. Free flow (white), synchronized flow (gray), and moving jams (black) in space and time. Data from March 23, 2001 on the freeway A5-North near Frankfurt, Germany. Taken from [218]](image-url)
This self-compression of synchronized flow is called the pinch effect. Likewise, the region of synchronized flow where the pinch effect occurs is called the pinch region of synchronized flow. In this self-compressed region, narrow moving jams emerge spontaneously. The narrow moving jams propagate upstream. Some of these narrow moving jams grow over time. Finally, the growing narrow moving jams can transform into wide moving jams: the initial WSP turns into an GP. The wide moving jams propagate on the freeway while maintaining the velocity of the downstream jam fronts (see the discussion of Fig. 2.23 below).

During the dynamics of transformation of a narrow moving jam into a wide moving jam the velocity of the downstream jam front tends to the characteristic velocity $v_g$. Respectively, parameters of the jam outflow tend to values that are related to points on the line $J$ in the flow-density plane (Sect. 2.4.2) [203]. In other words, characteristic parameters of wide moving jams are self-formed during the dynamics of wide moving jam emergence. This self-formation of characteristic parameters of wide moving jams can be considered an example of “self-organization” in traffic flow.

Note that a narrow moving jam consists of the upstream and downstream jam fronts only. The vehicle speed within the narrow moving jam should not be necessarily be equal to zero. The narrow moving jam does not possess a characteristic feature of a wide moving jam, maintaining the downstream jam front velocity through any state of traffic and any bottlenecks. In particular, a narrow moving jam can be caught at an upstream bottleneck. Narrow moving jams occur in congested traffic. Corresponding to the objective criteria of traffic phases in congested traffic, any state of congested traffic that does not exhibit this characteristic feature of a wide moving jam is related to the “synchronized flow” phase. Thus, narrow moving jams are states of the “synchronized flow” phase [218].

**GP Formation during Synchronized Flow Propagation**

Another scenario for GP emergence is the following (Figs. 2.20 and 2.21) [208]. Firstly, an F→S transition (breakdown phenomenon) occurs at a bottleneck where free flow has been disrupted. This phase transition leads to synchronized flow at the bottleneck. The upstream front of synchronized flow propagates upstream. However, a pinch effect ensues, i.e., a pinch region occurs spontaneously very quickly upstream of the bottleneck in this propagating synchronized flow, rather than an SP. In the pinch region of synchronized flow, narrow moving jams emerge spontaneously.

The narrow moving jams propagate upstream and grow. Finally, some transform into wide moving jams. The freeway locations where this transformation takes place are the locations of the phase transition from the
Fig. 2.20. Empirical example of a general pattern (GP) at bottleneck $B_3$ due to an on-ramp. (a) Average vehicle speed in the GP in space and time. (b) Total flow rate across the freeway in space and time. Firstly, synchronized flow is formed upstream of the bottleneck. Later, and at other freeway locations wide moving jams emerge in that synchronized flow. These wide moving jams propagate further upstream while maintaining the velocity of their downstream fronts. Dashed line in (a) shows the location of the bottleneck. The bottleneck is the same as bottleneck $B_3$ in Fig. 1.2. Traffic data from January 13, 1997 measured on a section of the freeway A5-South in Germany. Taken from [218]
“synchronized flow” phase to the “wide moving jam” phase (S→J transition for short). The result of wide moving jam formation is GP emergence. It can be seen in Fig. 2.21 that wide moving jam formation is GP emergence. It can be seen in Fig. 2.21 that wide moving jams appear later and at freeway locations other than the bottleneck location where the F→S transition occurred. Thus, the GP emerges spontaneously from a pair of phase transitions: firstly, an F→S transition occurs, followed at the other freeway locations by an S→J transition (F→S→J transitions for short). This self-generation of wide moving jams and consequent an GP can exist for several hours after the GP has emerged.

Comparing these two examples of GPs (Figs. 2.19 and 2.21), an important difference emerges: the frequency of spontaneous moving jam emergence in the GP in Fig. 2.19 is appreciably lower than the frequency of moving jam emergence in the GP in Fig. 2.21. This is because the average vehicle speed in the pinch region of the GP in Fig. 2.19 is higher than that of the GP in Fig. 2.21. This is a common empirical result of spontaneous moving jam emergence in synchronized flow: the lower the vehicle speed and the higher the density in synchronized flow, the higher the frequency of spontaneous moving jam emergence in that synchronized flow [208,218].

Wide moving jams of the GP can propagate many kilometers from the pinch region of the GP where moving jams emerge, with the width of the GP (in the longitudinal direction) continuously widening upstream. In contrast, the pinch region of the GP is localized at the bottleneck.

These two empirical examples of GPs (Figs. 2.19 and 2.21) are considered in more detail in Sect. 9.4.
Stop-and-Go Traffic

It must be noted that the well-known and very old term “stop-and-go” traffic, which is related to a sequence of moving traffic jams, will not be used in this book. For a traffic observer, both a sequence of narrow moving jams and a sequence of wide moving jams constitutes “stop-and-go” traffic. However, as already mentioned, narrow moving jams belong to the “synchronized flow” traffic phase, whereas wide moving jams belong to the qualitative different “wide moving jam” traffic phase.

Theoretical Congested Pattern Diagram

There is a diverse multitude of SPs and GPs that can be formed at bottlenecks, and that will be considered in this book. If the bottleneck is associated with an on-ramp, then different SPs and GPs occur spontaneously, depending on both the flow rate $q_{in}$ in an initial free flow on the main road upstream of the on-ramp and on the flow rate to the on-ramp $q_{on}$ (Fig. 2.22a). This leads to a theoretical diagram of congested patterns in the flow-flow plane where pattern emergence upstream of the bottleneck is presented as a function of the flow rates $q_{in}$ and $q_{on}$ (Fig. 2.22b) [218].

This diagram has the following physical meaning. At a very low flow rate to the on-ramp $q_{on}$, and if the flow rate upstream of the bottleneck $q_{in}$ is not very high, free flow occurs at the bottleneck. If the flow rate $q_{on}$ is increased, speed breakdown ($F \rightarrow S$ transition) occurs at the bottleneck: vehicles merging from the on-ramp onto the main road force the vehicles on the main road to reduce speed. However, if the flow rate to the on-ramp $q_{on}$ is still not very high, the influence of the on-ramp inflow on vehicle speed on the main road is not very strong. In this case, vehicle speed in the emergent synchronized flow should not be reduced to very low values. Thus, wide moving jams should not necessarily emerge in synchronized flow. As a result, SPs occur at lower flow rates $q_{on}$. If the flow rate $q_{on}$ then increases, the influence of the on-ramp inflow on vehicle speed on the main road upstream of the bottleneck increases too; the speed decreases and the density increases in the SP. In this compressed synchronized flow, wide moving jams emerge spontaneously, i.e., an GP should occur.

2.4.8 Expanded Congested Patterns

The GP and the SP that have briefly been discussed above appear at an “isolated” effectual bottleneck; the influence of other possible adjacent freeway bottlenecks on pattern formation at the isolated bottleneck should be negligible.

On real freeways there are many bottlenecks where various congested patterns can emerge almost simultaneously. If two or more adjacent bottlenecks exist close to one another, then an expanded congested pattern (EP) can
be formed [218]. In the EP, synchronized flow affects at least two adjacent effectual bottlenecks.

For example, consider two adjacent effectual bottlenecks close to one another. Let us assume that an F→S transition occurs at the downstream bottleneck, i.e., synchronized flow emerges there. Due to the upstream propagation of synchronized flow, this flow can reach the upstream bottleneck. In this case, synchronized flow can propagate upstream of the upstream bottleneck, affecting both adjacent bottlenecks. In other words, an EP appears.

There can be many different types of EPs. As shown in [218], EPs can be explained within the scope of three-phase traffic theory. This conclusion is related to the empirical fact [218] that congested traffic of all known EP types consists of either the “synchronized flow” phase only or the “synchronized flow” phase and the “wide moving jam” phase.

This result of an empirical study is illustrated in Fig. 2.23, where an example of an EP is shown. There are three adjacent effectual bottlenecks at this freeway section, the downstream bottleneck $B_{\text{North 1}}$, and two upstream adjacent bottlenecks labeled $B_{\text{North 2}}$ and $B_{\text{North 3}}$ (these bottleneck locations will be discussed in Sect. 9.2). Firstly, synchronized flow occurs
Fig. 2.23. Empirical example of an expanded congested pattern (EP). Free flow (white), synchronized flow (gray), moving jams (black). Data from March 23, 2001 on the freeway A5-North near Frankfurt, Germany. Taken from [218]

at the downstream bottleneck $B_{\text{North}1}$. This synchronized flow propagates upstream (this is the example of the WSP considered in Fig. 2.17a).

The upstream front of synchronized flow propagates upstream and reaches the upstream bottleneck $B_{\text{North}2}$. This synchronized flow propagates further upstream: the initial WSP transforms into an EP. In the EP, synchronized flow affects the adjacent effectual bottlenecks $B_{\text{North}1}$ and $B_{\text{North}2}$ (Fig. 2.17b).

Let us explain the effect of transformation of an WSP into an EP in more detail. The upstream front of synchronized flow in the WSP propagates continuously upstream. The upstream front of synchronized flow separates free flow upstream from synchronized flow downstream of the front. The downstream front of the WSP where vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow is fixed at a bottleneck. For these reasons, the width of synchronized flow of the WSP increases over time continuously. This means that synchronized flow of the WSP reaches sometimes an upstream bottleneck where upstream synchronized flow (i.e., upstream congestion) can be induced by the propagation of the upstream front of the WSP. This induced F→S transition leads to EP formation. The same conclusion can be valid for GPs. Thus, strictly speaking, WSPs and GPs can be considered as congested patterns at the isolated bottleneck only while
the upstream front of these patterns has not reached an upstream bottleneck where another synchronized flow can be induced.

We have mentioned above that over time, wide moving jams emerge in synchronized flow of the initial WSP, i.e., the WSP is transformed into an GP (Fig. 2.19). These wide moving jams propagate through synchronized flow of the EP. Thus, if we consider the EP in Fig. 2.17b for a longer time interval, we find the “complete” EP shown in Fig. 2.23. Congested traffic in this EP consists of both synchronized flow and wide moving jams.

Examples of wide moving jams in this EP are the moving jams labeled 1, 2, ..., 6 in Fig. 2.23. These moving jams propagate through complex states of synchronized flow and through bottlenecks while maintaining the mean velocity of the downstream jam fronts. We see that the EP (Fig. 2.23) qualitatively resembles an GP at an isolated bottleneck (Fig. 2.21). However, there is an important difference between the GP and the EP.

To understand this difference, note that upstream of each of the upstream adjacent effectual bottlenecks $B_{\text{North}2}$ and $B_{\text{North}3}$, new moving jams emerge spontaneously in synchronized flow of the EP (Fig. 2.23). Thus, we can expect a complex interactive process among various moving jams. This dynamic jam interaction occurs when the wide moving jams that previously emerged downstream of the upstream adjacent bottlenecks $B_{\text{North}2}$ and $B_{\text{North}3}$ propagate through synchronized flow upstream of those adjacent bottlenecks (see Sect. 11.3).

Note that an EP can consist of the “synchronized flow” phase only. The example is shown in Fig. 2.17b. However, this EP consists of synchronized flow for only about 30 min. Subsequently, wide moving jams begin to emerge in synchronized flow. This is a common feature of EPs; EPs, consisting solely of synchronized flow for the whole time of EP existence, occur very seldom.

If several adjacent effectual bottlenecks are close to one another, a qualitatively different case usually ensues (Fig. 2.24): EPs appear where synchronized flow covers many adjacent bottlenecks, and moving jams emerge in that synchronized flow. This and other empirical examples of congested patterns observed in traffic data of the PeMS (Freeway Performance Measurement System) database in the USA (California) show that congested traffic consists of two traffic phases, synchronized flow and wide moving jams. These traffic phases satisfy the objective criteria for traffic phases in congested traffic (Sect. 2.4.1), they possess other empirical features of congested patterns considered in this book, and those congested patterns can also be explained by three-phase traffic theory. Other empirical examples of EPs on this freeway section, which are reconstructed based on the FOTO and ASDA models, will be shown in Figs. 22.9 and 22.10 of Sect. 22.2.2.

However, it must be noted that in these cases (Figs. 2.24, 22.9, and 22.10) it is very difficult (often impossible) to distinguish moving jams that have emerged in synchronized flow between one pair of adjacent effectual bottlenecks from moving jams that have emerged in synchronized flow between the
Fig. 2.24. Empirical example of predictable and reproducible patterns on American freeways that confirms three-phase traffic theory. Free flow (white), synchronized flow (gray), and moving jams (black) in space and time. Data from the Interstate freeway I405-South, Orange County near Los Angeles, California. A section between the freeway exits “Bolsa Chica” and “Euclid Street” (about 12-km section). There are 12 on- and off-ramps on this section. Taken from [236]

neighboring pair of adjacent effectual bottlenecks. This result is related to the empirical fact that moving jam emergence (S→J transition) takes much longer (about 10 min) than an F→S transition (about 1 min). In 10 min a moving jam propagates about 2–3 km along the freeway. For this reason, the empirical result that moving jams do not usually emerge directly at bottlenecks can only clearly be seen if the distances between adjacent bottlenecks are high enough. In contrast, if the distances between adjacent effectual bottlenecks are about 2–3 km or less, we cannot usually find the real reason for wide moving jam emergence. In particular, it is almost impossible to see that moving jams do not emerge at the bottlenecks. This is the case for the infrastructure of the freeway section in Fig. 2.24. In Figs. 2.16 and 2.23, the distances between adjacent effectual bottlenecks are appreciably greater than 2–3 km. For this reason, we can much more easily understand traffic phenomena observed on such freeway sections.4

4 On the freeway section shown in Fig. 2.24, which is about only 12 km long, there are 12 on- and off-ramps, which can be potential effectual bottlenecks where speed breakdown (F→S transition) can occur. This case is common to many other freeway sections in the USA and Europe where local traffic measurements are available. This is one of the reasons why in examining spatiotemporal congested pattern features in this book, traffic data has usually been acquired over sections
2.4.9 Foreign Wide Moving Jams

We mentioned above that wide moving jams in Fig. 2.23, which initially emerged upstream of the downstream bottleneck $B_{North 1}$, propagate through the upstream bottleneck $B_{North 2}$ while maintaining the velocity of the downstream jam front. These wide moving jams subsequently propagate through the next upstream bottleneck $B_{North 3}$ as well.

The wide moving jams that initially emerged downstream of the bottleneck $B_{North 2}$ are called “foreign” wide moving jams when they propagate through synchronized flow upstream of the bottleneck $B_{North 2}$ [213,215,218]. Examples of the foreign wide moving jams are labeled 1, 2, . . . , 6 in Fig. 2.23 when these moving jams are upstream of the bottleneck $B_{North 2}$ or $B_{North 3}$.

“United” Sequence of Wide Moving Jams

The identification of foreign wide moving jams is associated with the following feature of wide moving jams [213,218]. When a foreign wide moving jam propagates through synchronized flow upstream of the upstream bottleneck (e.g., the bottleneck $B_{North 3}$), jam interaction can occur. Specifically, the wide moving jam can exert a considerable influence on other moving jams that are just emerging in synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam. In contrast, if a narrow moving jam is far enough away from the downstream front of the foreign wide moving jam, this narrow jam can turn into a wide moving jam.

Thus, in addition to foreign wide moving jams, new wide moving jams can emerge in the pinch region of synchronized flow upstream of the upstream bottleneck. As a result, an “united” sequence of wide moving jams is formed from the joining of the foreign wide moving jams and the wide moving jams that have emerged in synchronized flow upstream of the upstream bottleneck. This occurs in the example shown in Fig. 2.23.

Spatially Separated GP Forming Complex Congested Patterns

It should be noted that when the distances between adjacent effectual bottlenecks are great enough, spatially separated regions of synchronized flow can occur between adjacent effectual bottlenecks where the onset of congestion is realized. Often pinch regions are formed in these synchronized flows, leading to GP emergence. These GPs at different adjacent bottlenecks have spatially separated synchronized flow regions. We call these GPs “spatially separated” GPs [218].

the freeway A5-North and A5-South (e.g., Figs. 2.16 and 2.23). On the latter freeway sections, the distances between adjacent effectual bottlenecks, where congested patterns usually occur spontaneously, are more than 4 km.
An empirical example is shown in Fig. 2.25, where two spatially separated GPs occur at adjacent effectual bottlenecks \( B_2 \) and \( B_3 \). Synchronized flow upstream of the GP at bottleneck \( B_2 \) does not cover bottleneck \( B_3 \), i.e., no EP occurs between them. In other words, a congested pattern can consist of several spatially separated GPs, if there is enough distance between adjacent effectual bottlenecks (distances should usually be more than 4–5 km).

There can also be an intermediate case. An example can be seen between downstream bottleneck \( B_1 \) and upstream bottleneck \( B_2 \) in Fig. 2.25. We find that during some time intervals synchronized flow from the downstream GP at bottleneck \( B_1 \) covers bottleneck \( B_2 \), i.e., EPs appear. During other time intervals two spatially separated GPs at adjacent bottlenecks \( B_1 \) and \( B_2 \) are realized.

However, wide moving jams have emerged upstream of the downstream bottleneck \( B_1 \). The wide moving jams propagate through adjacent bottlenecks \( B_2 \) and \( B_3 \) while maintaining the velocity of the downstream jam fronts. These wide moving jams are foreign wide moving jams when they are upstream of adjacent bottlenecks \( B_2 \) and \( B_3 \), where other moving jams are emerging. Thus, these GPs have only spatially separated regions of

---

Fig. 2.25. Empirical example of foreign wide moving jam propagation. Free flow (white), synchronized flow (gray), and moving jams (black) in space and time. Data from the freeway A5-South
synchronized flow. In contrast to the synchronized flow regions, wide moving jams of the GP at the downstream bottleneck propagate through the GP at the upstream bottlenecks. This foreign jam propagation can considerably influence moving jam emergence at the upstream bottleneck. This usually occurs when spatially separated GPs appear at the freeway. Thus, due to foreign wide moving jam propagation, spatially separated GPs can form a complex congested pattern on the freeway (Fig. 2.25).

2.4.10 Reproducible and Predictable Congested Patterns

It has been found that for each freeway bottleneck or each set of several adjacent effectual bottlenecks, where congested patterns occur, the spatiotemporal structure of congested patterns exhibits predictable, i.e., characteristic, unique, and reproducible features [218]. These predictable and reproducible pattern features can be almost the same for different days and years. They can also persist over a large range of flow rates (traffic demand) at which the patterns exist. The predictable and reproducible pattern features remain in traffic flows with very different driver behavioral characteristics and different vehicle parameters. The predictable and reproducible pattern features can be used to forecast of congested patterns at freeway bottlenecks.

To find predictable pattern features, a chosen freeway section should be studied over a large number of days when congested patterns occur. Firstly, effectual bottlenecks, i.e., the bottlenecks where an F → S transition (speed breakdown) usually occurs and the downstream front of synchronized flow is usually fixed should be identified. Secondly, the duration \( T\text{type} \) for each of the types of congested patterns should be determined. Then the probability of a certain pattern type is found as \( P\text{type} = T\text{type}/T\text{cong} \), where \( T\text{cong} \) is the total duration of congestion on a freeway section on all these days (in \( T\text{type} \) and \( P\text{type} \) the low index “type” is for example “WSP” for an WSP, or “GP” for an GP or else “EP” for an EP). Finally, dependencies of the pattern type and pattern features on bottleneck characteristics and on traffic demand should be studied.

Furthermore, free flow should be observed both downstream of the downstream front of synchronized flow and upstream of the upstream front of synchronized flow. Thus, synchronized flow in congested patterns should spatially be localized on a freeway section where measurements are made. In this case correct pattern recognition and classification can be made. The latter requirement of spatial pattern localization does not concern wide moving jams. These jams can propagate through the entire freeway section. This is because wide moving jams possess predictable and reproducible characteristics that do not depend on traffic demand and on bottleneck characteristics. In accordance with this scheme, a congested pattern study has been made in [218] and in Part II of this book based on data measured during 1995–2003 on about 30-km long sections of the freeways A5-South and A5-North (Figs. 2.1 and 2.2).
Some of the reproducible congested pattern features can be seen in Figs. 2.26 and 2.27. Here, empirical examples of congested patterns, which were measured on different days and years, on two different freeway sections on German freeways, are presented.

There are pattern features that have a probabilistic nature. This means that in different realizations (days) congested pattern features can be different for the same traffic control parameters (weather, other road conditions), traffic demand, and initial conditions. However, we can identify different “degrees of predictability” for different pattern features. “Deterministic” pattern features and pattern features with a high “degree of predictability” are as follows:

(1) Speed breakdown (F→S transition) occurs mostly at the same effectual freeway bottlenecks on different days and years. Examples are shown in Figs. 2.23 and 2.26, where speed breakdown occurs mostly at adjacent effectual bottlenecks $B_{\text{North } 1}$, $B_{\text{North } 2}$, and $B_{\text{North } 3}$. Other examples can be seen in Figs. 1.2b and 2.27, where speed breakdown occurs mostly at adjacent effectual bottlenecks $B_1$, $B_2$, and $B_3$. It can be seen from these examples that speed breakdown occurs either spontaneously or it is induced (by the upstream propagation of a wide moving jam or a region of synchronized flow) at these bottlenecks.

(2) In contrast to synchronized flow, which occurs mostly at freeway bottlenecks, moving jam emergence is usually observed in synchronized flow a finite distance upstream of bottlenecks. In some cases, this distance can be 3 km or more. For example, this can be seen in Fig. 2.23, where moving jam emergence is observed in synchronized flow upstream of bottleneck $B_{\text{North } 1}$. For this reason, the empirical result that moving jams do not emerge right at bottlenecks can only be clearly seen if the distances between bottlenecks are great enough (see footnote 4 of Sect. 2.4.8).

(3) Wide moving jams propagate through any complex state of traffic and through any bottleneck at the mean velocity $v_g$ of the downstream jam front.

(4) If free flow emerges in the wide moving jam outflow, then the flow rate in this jam outflow is considerably lower than the maximum possible flow rate in free flow.

(5) The catch effect: a local synchronized flow region reaching a bottleneck is caught at the bottleneck.

(6) If a narrow moving jam is very close to the downstream front of a foreign wide moving jam, the foreign jam suppresses the growth of this narrow jam.

Pattern features with a middle “degree of predictability” are as follows:

(i) The instant of spontaneous speed breakdown at an effectual bottleneck at the same dependence of traffic demand on time. This reveals the probabilistic nature of an F→S transition.
2.4 Main Empirical Features of Congested Patterns

Fig. 2.26. Empirical examples of predictable and reproducible patterns. Overviews of free flow (white), synchronized flow (gray), and moving jams (black) in space and time. Data from the freeway A5-North near Frankfurt, Germany. The effectual bottlenecks are the same as those in Fig. 2.23.
Fig. 2.27. Empirical examples of predictable and reproducible patterns. Overviews of free flow (white), synchronized flow (gray), and moving jams (black) in space and time. Data from the freeway A5-South near Frankfurt, Germany. The effectual bottlenecks are the same as those in Fig. 2.16.
(ii) For a bottleneck or for a set of adjacent effectual bottlenecks that are close to one another there is a certain type of spatiotemporal congested pattern that appears with the highest probability, given the same traffic demand. At the same traffic demand various types of congested patterns can occur spontaneously or can be induced (one of the SPs, GPs, EPs). Furthermore, specific bottleneck characteristics can restrict a choice of types of congested patterns, which can occur in comparison with all possible alternatives. As a result, some of the pattern types do not appear at a set of the adjacent effectual bottlenecks. Often one of the congested patterns strongly predominates, i.e., the probability of a certain pattern type is much higher than the sum of the probabilities of the occurrence of all other possible patterns. For example, if a congested pattern occurs at bottleneck $B_3$ in Fig. 2.27, this pattern is an GP with the probability more than 90%.

Pattern features with a low “degree of predictability” are as follows:

(a) Whether a spontaneous or an induced speed breakdown occurs, leading to synchronized flow at an effectual bottleneck.
(b) The instant of moving jam emergence in synchronized flow.

There are also other predictable and reproducible features of congested patterns that will be considered in Part II of this book.

2.4.11 Methodology for Empirical Congested Pattern Study

A correct understanding of the dynamics of congested traffic is possible only if a spatiotemporal analysis of traffic flow is performed on a sufficiently large section of freeway. The spatiotemporal analysis of traffic flow means the following:

(i) Traffic flow variables must be measured at many freeway locations simultaneously.
(ii) Freeway locations where congested traffic occurs should be determined.
(iii) Freeway locations where the onset of congestion first occurs should be identified.
(iv) The study associated with item (i)–(iii) should be repeated over the course of many days.
(v) The freeway locations where the onset of congestion occurs on most days should be identified.
(vi) Based on results of item (v) and on freeway infrastructure plans, effectual bottlenecks should be found.
(vii) To study some specific congested pattern, the farthest downstream freeway bottleneck where the downstream front of synchronized flow within the pattern is localized should be found. Recall that the downstream front of synchronized flow at an effectual bottleneck separates synchronized flow upstream of the bottleneck from free flow downstream of the
bottleneck. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front.

(viii) It should be found whether the pattern in item (vii) due to the upstream propagation of congestion reaches the next upstream effectual bottleneck or not.

(ix) Based on the objective criteria for traffic phases in congested traffic (Sect. 2.4.1; for more detail see Sect. 4.2.1), the “synchronized flow” and “wide moving jam” phases should be identified.

(x) Free flow should be observed both downstream of the downstream front of synchronized flow and upstream of the upstream front of synchronized flow. Thus, synchronized flow in congested patterns should spatially be localized within a freeway section where measurements are made.

(xi) The latter requirement of the spatial localization does not concern wide moving jams. These jams can propagate through the freeway section. This is because wide moving jams possess predictable and reproducible characteristics that do not depend on traffic demand and on bottleneck characteristics. In particular, the velocity $v_g$ of the downstream jam front possesses this feature.

(xii) A sufficient time interval for traffic variable measurements and the congested pattern analysis should be chosen: free flow should be observed before congested pattern formation at the effectual bottlenecks of the freeway section and also after congested patterns have disappeared.

(xiii) Based on measurements of traffic variables at various freeway locations upstream of the bottleneck in item (vii)–(xii), the type of congested pattern (whether the pattern is one of the SPs, GPs, EPs), the spatiotemporal structure of the pattern, the pattern evolution over time, and the pattern characteristics depending on traffic demand should be found.

Such spatiotemporal analysis is the main methodology of the study of empirical congested patterns in this book. We already used this methodology to discuss the main empirical spatiotemporal congested pattern features (Sects. 2.4.1–2.4.10). It should be noted that after the main spatiotemporal pattern features are understood, additional study of some nonlinear pattern features can be performed in the flow–density plane (or in the speed–density plane).

### 2.5 Conclusions. Fundamental Empirical Features of Spatiotemporal Congested Patterns

There are empirical features of phase transitions and spatiotemporal congested patterns at freeway bottlenecks that are reproducible in traffic obser-
2.5 Fundamental Empirical Features of Traffic Patterns

Observations on numerous days and years on different freeways in various countries. Thus, these qualitative empirical features remain in traffic flows with very different driver behavioral characteristics and vehicle parameters. Such fundamental empirical features of phase transitions and congested patterns at freeway bottlenecks are as follows:

(i) Traffic can be either “free” or “congested.” The onset of congestion in free traffic flow at an effectual freeway bottleneck (i.e., speed breakdown in free flow) (see Figs. 2.10 and 2.11) and freeway capacity possess a probabilistic nature (Fig. 2.12).

(ii) There are two types of speed breakdown in free flow at an effectual bottleneck, spontaneous (Fig. 2.10) and induced (Figs. 1.2a and 2.16).

(iii) An induced speed breakdown in free flow at an effectual bottleneck can be caused either by moving jam propagation through the bottleneck (Fig. 1.2a) or the upstream propagation of a synchronized flow pattern (SP) that has initially occurred downstream of the bottleneck (Fig. 2.16).

(iv) In congested traffic, two different traffic phases can be distinguished: the “synchronized flow” phase and the “wide moving jam” phase. To distinguish between these traffic phases, the empirical (objective) criteria \[ J \] and \[ S \] should be applied (Sect. 2.4.1; see Figs. 1.2 and 2.7). Thus, there are three traffic phases in freeway traffic: (1) free flow; (2) synchronized flow; (3) wide moving jam.

(v) Speed breakdown in free flow at an effectual bottleneck causes the emergence of the “synchronized flow” phase (Fig. 2.10) at the bottleneck, rather than the emergence of the “wide moving jam” phase. Speed breakdown at the freeway bottleneck is associated with a phase transition from the “free flow” phase to the “synchronized flow” phase (F→S transition).

(vi) Wide moving jams do not emerge spontaneously in the “free flow” phase.

(vii) Wide moving jams can emerge spontaneously only in the “synchronized flow” phase (Fig. 2.19). Wide moving jams emerge in free flow due to the sequence of F→S→J transitions. Firstly, a region of synchronized flow occurs in free flow. Then a self-compression of that synchronized flow (the pinch effect) is realized. In the pinch region of synchronized flow, narrow moving jams emerge and grow. Wide moving jams will result due to the growth of these narrow moving jams (Figs. 2.19 and 2.21).

(viii) The higher the density in synchronized flow, the higher the frequency of spontaneous moving jam emergence in that synchronized flow (Figs. 2.19 and 2.21). In synchronized flow of higher speeds and lower densities, wide moving jams should not necessarily emerge spontaneously (Fig. 2.17).
During the dynamics of transformation of a narrow moving jam into a wide moving jam, the mean velocity of the downstream jam front tends to the characteristic velocity $v_g$ and parameters of the jam outflow tend to values that are related to points on the line $J$ in the flow–density plane (Fig. 2.8). Wide moving jams possess characteristic parameters and features that do not depend on initial conditions and perturbations in traffic. The characteristic parameters are the same for different wide moving jams. The characteristic parameters can depend on control parameters of traffic (weather, road conditions, etc.). These characteristic parameters and features are as follows:

(a) The mean velocity of the downstream front of a wide moving jam: the wide moving jam propagates upstream through any traffic state and through any bottleneck while maintaining the velocity of the downstream jam front.

(b) When free flow occurs in the wide moving jam outflow, the flow rate and density in the jam outflow are also characteristic parameters.

(c) The flow rate in the jam outflow is lower than the maximum possible flow rate in free flow.

There are two main types of congested patterns at an isolated effectual freeway bottleneck: synchronized flow (SP) (Figs. 2.13, 2.16, and 2.17) and general patterns (GP) (Figs. 2.19 and 2.21).

In contrast to wide moving jam propagation (item ix (a)), an SP that propagates upstream is caught at an effectual bottleneck (catch effect) when the SP reaches the bottleneck. Due to the catch effect, an induced $F\rightarrow S$ transition can occur (Fig. 2.16).

There are three types of SPs: a moving SP (Fig. 2.16), a widening SP (Fig. 2.17), and a localized SP (Fig. 2.13).

There can be complex spontaneous transformations between various congested patterns at an effectual bottleneck over time (Fig. 2.19).

At the same traffic demand various congested patterns can be formed at an effectual bottleneck (the probabilistic nature of spatiotemporal congested patterns at the bottleneck).

If two or more adjacent effectual bottlenecks are close to one another, expanded congested patterns (EP) often emerge. In an EP, synchronized flow affects at least two adjacent effectual bottlenecks (Figs. 2.23 and 2.24).

When a wide moving jam that has initially emerged downstream of an effectual bottleneck propagates through synchronized flow upstream of the bottleneck, this foreign wide moving jam can influence considerably on other moving jams, which are just emerging in that synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam (Fig. 2.23).
(xvii) For each effectual freeway bottleneck, or each set of several adjacent effectual bottlenecks where congested patterns occur, the spatiotemporal structure of a congested pattern possesses some predictability, i.e., characteristic, unique, and reproducible features (Figs. 2.26 and 2.27).
3 Overview of Freeway Traffic Theories and Models: Fundamental Diagram Approach

3.1 Introduction: Hypothesis About Theoretical Fundamental Diagram

A huge number of models have been developed to explain empirical spatiotemporal traffic dynamics (e.g., [1–11,13–18,20,21,42–46,362–367,370–395, 397–401,420–422,436–465]; see also references in the reviews by Gartner et al. (eds.) [31], Wolf [32], Chowdhury et al. [33], Helbing [34,35], Nagatani [36], Nagel et al. [38]).

In 1955 Lighthill and Whitham [2] wrote in their classic work (see p. 319 in [2]): “...The fundamental hypothesis of the theory is that at any point of the road the flow (vehicles per hour) is a function of the concentration (vehicles per mile)…” (Fig. 3.1).

To explain this fundamental hypothesis, let us consider hypothetical model solutions where all vehicles move at the same distances with respect to one another and with the same time-independent vehicle speed. These solutions are called “homogeneous,” “equilibrium,” “stationary” states, or

![Fig. 3.1. Qualitative example of the theoretical fundamental diagram [2]. A steady model state where the vehicle speed $v$ is related to only one vehicle density, which is given by the intersection of the dotted line “slope $v$” with the flow–density relation, i.e., with the fundamental diagram](image-url)
“steady-state” model solutions. In this book, we call these hypothetical traffic states “steady states.” The fundamental hypothesis about the fundamental diagram means that all steady-state model solutions lie on a theoretical fundamental diagram, i.e., on a curve(s) in the flow–density plane (Fig. 3.1) \cite{31,33,35,36,38}. This approach to traffic flow modeling and theory whose fundamental hypothesis is associated with the fundamental diagram for steady-state model solutions can be called the fundamental diagram approach. Thus, in this approach \cite{31,33,35,36,38} the whole multitude of the steady-state model solutions cover a one-dimensional region in the flow–density plane.

Concerning the hypothesis about the theoretical fundamental diagram, it must be noted that complex spatiotemporal traffic patterns are observed in congested traffic (e.g., \cite{80,82,85–88,218}). Thus, at higher density the empirical fundamental diagram is related to averaged characteristics of spatiotemporal congested patterns measured at a freeway location rather than to features of the hypothetical steady states of congested traffic. This means that the existence of the theoretical fundamental diagram is only a hypothesis. The hypothesis about the theoretical fundamental diagram underlies almost all traffic flow modeling approaches up to now \cite{31,33,35,36} in the sense that the models are constructed such that in the unperturbed, noiseless limit they have a fundamental diagram of steady states, i.e., the steady states form a curve in the flow–density plane (e.g., \cite{2–4,6–11,13–18,20,21,42–46,362–367,370–395,397–401,420–422,436–464}; see also references in the reviews \cite{31–36,38}).

In this chapter, the following main points are considered:

(i) achievements of the fundamental diagram approach, which is the theoretical basis for most of the earlier mathematical traffic flow models and theories;

(ii) a critical analysis of the fundamental diagram approach for a mathematical description of spatiotemporal features of congested traffic.

### 3.2 Achievements of Fundamental Diagram Approach to Traffic Flow Modeling and Theory

The first models of traffic flow in the fundamental diagram approach were based on collective properties of traffic such as conservation of the number of vehicles, the balance of average vehicle speed, and other more complex macroscopic properties of the flow. The first macroscopic models were proposed in 1955 by Lighthill and Whitham \cite{2}, in 1956 by Richards \cite{3}, by Prigogine in 1959 \cite{14}, and by Payne in 1971 \cite{15,16}. The Payne model was further developed by many authors (e.g., \cite{27,363,366,367,395}).

There has also been a huge number of microscopic models in which the individual behavior of each vehicle is taken into account. Examples include the
“car-following” approach of Reuschel [359–361], Pipes [1], Herman, Montroll, Potts, Gazis, Rothery [4,5], and Kometani and Sasaki [6].

In the approach of Herman, Montroll, Potts, Gazis, and Rothery [4,5] a formula connects the acceleration of a car with the distance between it and the car ahead of it (the preceding vehicle), and the relative velocity of two cars taken with some time delay. Independent of distances between vehicles, the solutions of the car-following model in which all the vehicles move at the same constant speed are steady-state model solutions. In some region of model parameters the car-following model shows instability of initial steady states [4]. However, the model of Herman et al. [4] is not capable of describing traffic beyond the onset of instabilities, because it lacks a mechanism that limits oscillations to realistic values, i.e., to values limited by the acceleration and braking capabilities of vehicles [38]. In [5], to have model solutions that can be compared with empirical data, the theoretical fundamental diagram was introduced in the car-following model by integrating the vehicle equation of motion with an appropriate choice of integration constant.

In this case, the car-following model in [5] reduces to the model proposed by Newell in 1961 [10] (for more detail see the review by Nagel et al. [38]), in which the vehicle speed is some function of the distance between vehicles. In the “optimal velocity” model developed by Whitham in 1990 [362] and by Bando, Sugiyama, and colleagues in Japan [369,370], each driver tries to achieve an optimal velocity that depends on the distance to the preceding vehicle and the speed difference between vehicles. In the “intelligent driver” model by Treiber and Helbing the vehicle acceleration assumed is a continuous function of vehicle speed, space gap (distance to the preceding vehicle), and speed difference between the vehicle and the preceding vehicle [396,397]. The model of Gazis, Herman, Rothery with the fundamental diagram [5], the optimal velocity models [10,362,370], and other traffic flow models that utilize the fundamental diagram approach [33,35,36,38] can all describe some traffic dynamics beyond the instability.

Wiedemann has developed a psychophysical model in which a complex set of rules governs the reaction of the driver to the motion of other cars [25].

Besides “deterministic” models, there are many stochastic traffic flow models. In 1981 Gipps introduced a discrete in time and spatial continuum stochastic model based on safe speed that prevents vehicle collisions [17]. There are stochastic traffic flow models in which random model fluctuations play an important role in the nonlinear model dynamics. Examples are “cellular automata” models [364, 373–375]. This approach was pioneered by Nagel, Shreckenberg, and coworkers [364, 373, 374]. Mahnke et al. [376,380,402] developed a master equation approach that enables one to calculate the probability of moving jam formation in an initially homogeneous free flow.
3.2.1 Conservation of Vehicle Number on Road and Front Velocity

The familiar particle conservation law of classical physics was applied to the earliest traffic flow models to address vehicle number conservation on a road \([2, 20, 21, 362]\). Indeed, all traffic flow models and theories must satisfy the law of conservation of the number of vehicles on the road. This means that the balance equation

\[
\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial q(x, t)}{\partial x} = 0
\]  

must always be satisfied either explicitly or implicitly in any traffic flow model and theory. Here \(x\) is a spatial coordinate in the direction of traffic flow, and \(t\) is time.

Let us assume that at \(t = t_0\) a spatial change in vehicle density \(\rho(x)\) and flow rate \(q(x)\) exists and is localized at \(x = x_0\); at \(x < x_0\), \(\rho = \rho_1\) and \(q = q_1\). At \(x > x_0\), \(\rho = \rho_2\) and \(q = q_2\). Recall that the region in traffic where one or more traffic variables spatially change abruptly from one state of traffic to another state of traffic is called a front between two different traffic states (Sect. A.2). The velocity of the front (shock-wave velocity) \(v_p\) can be found from the usual balance equation (3.1). In coordinate system moving at the shock-wave velocity \(v_p\), this front is motionless. In this coordinate system the flow rate \(\rho_1 u_1\) left of the front is equal to the flow rate \(\rho_2 u_2\) right of the front, i.e.,

\[
\rho_1 u_1 = \rho_2 u_2 ,
\]

where

\[
u_1 = v_1 - v_p , \quad u_2 = v_2 - v_p .
\]

\(v_1\) is the speed of vehicles related to the state \((\rho_1, q_1)\); \(v_2\) is the speed of vehicles related to the state \((\rho_2, q_2)\);

\[
q_1 = v_1 \rho_1 , \quad q_2 = v_2 \rho_2 .
\]

According to a historical overview of the theory of shock waves made in the book by Courant and Friedrichs [466] (published in 1948), (3.2) was first derived by Stokes [467] in 1848. In the motionless coordinate system, taking into account (3.3) and (3.4) the Stokes shock-wave formula (3.2) can be rewritten for the shock-wave velocity \(v_p\) as follows

\[
v_p(t_0) = \frac{q_2 - q_1}{\rho_2 - \rho_1} .
\]

Therefore, we refer to this formula as the Stokes shock-wave formula. The Stokes shock-wave formula will often be used in this book.
3.2.2 The Lighthill–Whitham–Richards Model and Shock Wave Theory

There are two independent variables, \( \rho \) and \( q \), in the balance equation (3.1) and in the Stokes shock-wave formula (3.5).

To explain traffic flow phenomena based on shock wave theory, Lighthill and Whitham [2], and Richards [3] used the hypothesis about the fundamental diagram. They assumed that the flow rate \( q \) is a function of the vehicle density \( \rho \):

\[
q = q(\rho) .
\]

This function is related to the fundamental diagram for traffic flow (Fig. 3.1). Using (3.6), the balance equation (3.1) takes the form

\[
\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial q(\rho(x, t))}{\partial x} = 0 .
\]

There­fore, there is now only one independent variable in the balance equation, the vehicle density \( \rho \) (Fig. 3.1). This makes it possible to solve this equation if initial and boundary conditions are given. Thus, the application of the hypothesis about the fundamental diagram leads to a solvable traffic flow model.

Solutions of the Lighthill–Whitham–Richards (LWR) model (3.6), (3.7) are kinematic waves (KW) moving with the velocity [2,3]

\[
c = \frac{dq(\rho)}{d\rho} .
\]

This KW velocity is positive on the part of the fundamental diagram where the flow rate increases with density, and it is negative on the part of the fundamental diagram where the flow rate decreases with density.

The KW velocity (3.8) is a function of the density. In some cases, this function leads to turnover of the KW front [2]. Then a shock wave is formed (Fig. 3.2a). This shock wave should propagate at the velocity [2]

\[
v_s = \frac{q(\rho_2) - q(\rho_1)}{\rho_2 - \rho_1} ,
\]

where \( q(\rho_2) \), \( q(\rho_1) \) are related to the solutions of (3.6), i.e., corresponding points in the flow–density plane lie on the fundamental diagram for traffic flow; \( \rho_1 \) and \( \rho_2 \) are the densities upstream and downstream of the shock wave, respectively (Fig. 3.2). The shock wave can be represented in the flow–density plane by the line \( S \) whose slope is equal to the shock velocity \( v_s \) (Fig. 3.2b).

In contrast to the Stokes shock-wave formula (3.5), in the Lighthill–Whitham (LW) theory of shock waves (3.6)–(3.9) there is only one independent variable, the density. The flow rate here is a function of the density (3.6). Thus, in the Lighthill–Whitham–Richards theory, it is possible to make
Fig. 3.2. Representation of a shock wave in space (a) and in the fundamental diagram with the line $S$ whose slope is equal to the velocity of the shock wave (b) [2]

a traffic prediction for both the density and flow rate based on the LWR model (3.6), (3.7).

The kinematic wave (KW) theory was used by Lighthill and Whitham to explain many traffic phenomena, in particular congested traffic upstream of a freeway bottleneck. The LWR theory is also the basic theory used by many traffic researchers to describe congested traffic (e.g., [23,28,39–41]).

### 3.2.3 Collective Flow Concept and Probability of Passing

In 1971 Prigogine and Herman introduced the concept of collective flow in the context of the fundamental diagram approach [24]. Collective flow should occur at higher vehicle density than free flow. In this two-phase traffic theory (the “free flow” phase and the “collective flow” phase), vehicles move almost at their desired speeds in free flow, but they cannot do so in collective flow, where a collective motion of vehicles should occur. In this theory, the collective flow should be related to the part of the fundamental diagram at higher density.

This behavior should be related to a dependence of the probability of passing in traffic flow, which in this traffic flow theory is a monotonic decreasing function of vehicle density [24]. This is also responsible for synchronization of vehicle speed between different lanes of a multilane road in congested traffic, as observed in empirical investigations (e.g., [88,114]).

1 The fact that the probability of passing in congested traffic is considerably lower than in free flow [24] and that there is often synchronization of vehicle speed between different lanes of a multilane road in congested traffic [88,114] is also used in three-phase traffic theory. In this theory, as in the classic theory of Prigogine and Herman, it is also assumed that the probability of passing in synchronized flow is considerably lower than that in free flow.

However, in the theory of Prigogine and Herman, which is related to the fundamental diagram approach, steady states of collective flow should be associated
### 3.2.4 Scenarios for Moving Jam Emergence

Two qualitatively different scenarios for the formation of moving traffic jams have emerged from different mathematical models in the fundamental diagram approach. As long ago as 1958–1959, Herman, Montroll, Potts, and Rothery [4], and Komentani and Sasaki [6] applied ideas from statistical physics – such as instabilities, critical points, and so forth – in an attempt to explain why moving traffic jams form. The traffic flow instability is related to a finite reaction time of drivers. This reaction time is responsible for the vehicle over-deceleration effect: if the preceding vehicle begins to decelerate unexpectedly, a driver decelerates stronger than it is needed to avoid collisions. In [4, 6] it has been proposed that due to this effect there should be vehicle densities where traffic cannot flow freely because of instabilities of traffic flow with respect to small-amplitude fluctuations.

A different scenario for moving jam formation was proposed in [367, 368]. Below the critical density for the onset of moving traffic jams due to growth of small-amplitude fluctuations, in [367, 368] was proposed that there should be a broad range of lower vehicle densities where seemingly homogeneous and stable states of free traffic flow are in fact metastable with respect to moving jam formation. A metastable state is a state that is stable against small-amplitude fluctuations. However, if the amplitude of a local perturbation exceeds some critical amplitude, the perturbation will grow and lead to a moving jam.

### 3.2.5 Wide Moving Jam Characteristics

If the width of a moving jam (in the longitudinal direction) considerably exceeds the width of the jam fronts, then it is called a wide moving jam (Fig. 3.3).\(^2\)

with higher density in the theoretical fundamental diagram. In contrast, the fundamental hypothesis of the three-phase traffic theory is that steady states of the “synchronized flow” traffic phase cover a two-dimensional region in the flow–density plane, i.e., there is no fundamental diagram for the steady states of synchronized flow in this theory [207–209] (see Sect. 4.3). Furthermore, in the Prigogine and Herman theory the probability of passing is a monotonic function of density. In contrast, in the three-phase traffic theory, in accordance with empirical results [221, 222] (Sect. 10.7.3), the probability of passing is a Z-shaped function of vehicle density (see Sect. 5.2.5).

It must also be recalled (see Sect. 2.4.1) that the speed synchronization effect in synchronized flow is not a criterion to distinguish between the “synchronized flow” phase and the “wide moving jam” phase in congested traffic. These traffic phases are defined based on the objective criteria [S] and [J] in Sect. 2.4.1 (for more detail see Sect. 4.2) that are associated with qualitative different spatiotemporal behavior of these traffic phases.

\(^2\) It must be noted this is a theoretical definition of wide moving jams. The exact definition of wide moving jams, however, follows from the empirical (objective)
Fig. 3.3. Schematic representation of a wide moving jam at a fixed time [367]. Spatial distribution of vehicle speed $v$, flow rate $q$, and vehicle density $\rho$ in the wide moving jam, which propagates through a homogeneous state of free flow with speed $v_h$, flow rate $q_h$, and density $\rho_h$.

Corresponding to theory of wide moving jams, first derived in 1994 by Kerner and Konhäuser, wide moving jams and free flow should possess the following characteristic features [367]:

(i) The downstream front of a wide moving jam possesses a feature of steady propagation. Within the downstream front of the wide moving jam vehicles accelerate from the standstill inside the jam to free flow downstream of the jam. The feature of steady propagation of the downstream jam front means that the mean velocity of the downstream jam front $v_g$ is a constant. This is true if control parameters of traffic (road conditions, weather, etc.) do not change. Thus, the mean velocity $v_g$ is a characteristic, i.e., unique, coherent, and reproducible parameter.

criteria for two traffic phases in congested traffic (Sects. 2.4.1 and 4.2). It turns out (Chap. 16) that if the width of moving jams considerably exceeds the width of the jam fronts, then these wide moving jams satisfy the empirical definition of the “wide moving jam” traffic phase – they have the same mean velocity of the downstream jam front, and also they maintain this velocity while propagating through any other complex state of traffic flow or bottlenecks.
(ii) If in the wide moving jam outflow free flow is formed, then the flow rate \( q_{\text{out}} \) in this jam outflow, the related vehicle density, \( \rho_{\text{min}} \), and the mean vehicle speed \( v_{\text{max}} \) are characteristic parameters. At a given control parameter of traffic characteristic parameters do not depend on initial conditions, and they are the same for different wide moving jams.

Note that in an asymptotic mathematical theory of wide moving jams [403], some simple algebraic equations for these characteristic parameters of wide moving jams were found.

### 3.2.6 Flow Rate in Wide Moving Jam Outflow.

#### The Line \( J \)

The steady propagation of the downstream front of a wide moving jam in free flow with the mean velocity \( v_g \) can be represented in the flow–density plane by a line. This line is called “the line \( J \).” The slope of the line \( J \) is equal to the velocity \( v_g \) of this front (Fig. 3.4) [367]. The left coordinates of the line \( J \) are related to the parameters of free flow \( (\rho_{\text{min}}, q_{\text{out}}) \) exhibited by vehicles that have accelerated from the standstill inside the jam. The right coordinates of the line \( J \), \( (\rho_{\text{max}}, 0) \), are related to the vehicle density inside the jam \( \rho_{\text{max}} \) where the vehicle speed \( v_{\text{min}} \) is zero (Fig. 3.3). The line \( J \), the metastability of free flow with respect to moving jam formation (Sect. 3.2.7), and the characteristic parameters of wide moving jams were first predicted and revealed in the theory of wide moving jams [367]. These features have further been found in empirical studies of wide moving jam propagation by Kerner and Rehborn [166,203].

![Fig. 3.4. Qualitative representation of the characteristic line for the downstream front of a wide moving jam (line \( J \)) whose slope is equal to the downstream jam front velocity \( v_g \), and explanation of the metastability of states of free flow (curve \( F \), which is a part of the fundamental diagram associated with low density in Fig. 3.1) with respect to moving jam formation [367]](image-url)
Concerning the flow rate in the wide moving jam outflow \( q_{\text{out}} \), some general assumptions can be made [168]. It can be expected that the mean flow rate \( q_{\text{out}} \) does not change over time during wide moving jam propagation. This assumption can be explained as follows. Each driver standing within a wide moving jam can start to accelerate from the standstill inside the jam to free flow downstream after two conditions have been satisfied:

(i) The preceding vehicle has already begun to move away from the jam.
(ii) Due to the preceding vehicle motion, after some time the distance between the two drivers has exceeded some “safety distance” \( g_{\text{del}} \).

In other words, there is some finite mean time delay \( \tau_{\text{del}}^{(a)} \) in vehicle acceleration. This time delay corresponds to the safety distance \( g_{\text{del}} \) between two successive drivers that accelerate from the standstill within the wide moving jam to free flow downstream. This delay in vehicle acceleration is the time delay on some “expected” event. This expected event occurs when a driver is at the downstream front of the wide moving jam. The downstream jam front spatially separates the standstill inside the jam from free flow downstream of the jam.

The motion of the downstream front of a wide moving jam results from acceleration of drivers from the standstill inside the jam to flow downstream of the jam. Because the average distance between vehicles inside the jam, including average length of each vehicle, equals \( 1/\rho_{\text{max}} \), the velocity of the downstream front of the wide moving jam is

\[
v_{\text{g}} = - \frac{1}{\rho_{\text{max}} \tau_{\text{del}}^{(a)}}.
\]

It has been assumed that the flow rate inside the wide moving jam is zero. Therefore, the velocity of the downstream front of the jam is obviously related to the flow rate in the wide moving jam outflow corresponding to the Stokes shock-wave formula (3.5):

\[
v_{\text{g}} = - \frac{q_{\text{out}}}{\rho_{\text{max}} - \rho_{\text{min}}}.
\]

Equations (3.10) and (3.11) make it possible to write the flow rate in the wide moving jam outflow in the form

\[
q_{\text{out}} = \frac{1}{\tau_{\text{del}}^{(a)}} \left( 1 - \frac{\rho_{\text{min}}}{\rho_{\text{max}}} \right).
\]  

It has been assumed that free flow ensues downstream of the wide moving jam. In this case, the density of vehicles \( \rho_{\text{min}} \) in the jam outflow can be considerably lower than the density of vehicles within the jam \( \rho_{\text{max}} \), thereby yielding the following approximate value for \( q_{\text{out}} \) from (3.12):

\[
q_{\text{out}} \approx \frac{1}{\tau_{\text{del}}^{(a)}}.
\]
The mean time delay $\tau_{del}^{(a)}$ between two successive vehicles that accelerate from the standstill inside a wide moving jam to free flow downstream is constant when the control parameters of traffic (weather, etc.) do not change. This means that $\tau_{del}^{(a)}$ is a characteristic parameter of traffic flow. Therefore, the flow rate in the wide moving jam outflow $q_{out}$ corresponding to (3.13) is also a characteristic parameter of traffic flow. Note that the approximation (3.13) can only be applied if free traffic flow is formed downstream of the jam. A different case occurs if synchronized traffic flow is formed in the outflow from the jam (Sect. 11.2.3).

Corresponding to this definition of the line $J$ [367, 403], it satisfies the obvious equation

$$q(p) = \frac{1}{\tau_{del}^{(a)}} \left( 1 - \frac{\rho}{\rho_{max}} \right).$$

Equation (3.12) corresponds to (3.14) at $\rho = \rho_{min}$ and $q = q_{out}$. The maximum flow rate $q(\rho_{min}) = q_{out}$ on the line (3.12) is related to the flow rate in the free flow achieved in the wide moving jam outflow. This traffic flow is formed by vehicles that have accelerated from the standstill inside the jam to free flow downstream of the jam.

Because $q = v\rho$ we can rewrite (3.14) in the form

$$v(p) = \frac{1}{\tau_{del}^{(a)}} \left( \frac{1}{\rho} - \frac{1}{\rho_{max}} \right),$$

where the speed $v$ and density $\rho$ are related to points on the line $J$.

This shows that the downstream front of wide moving jams is a steady moving structure whose parameters do not depend on initial conditions in traffic flow. This explains the existence of characteristic parameters of traffic flow.

3.2.7 Metastable States of Free Flow with Respect to Moving Jam Emergence

Another result of the theory of wide moving jams [367] is that the density $\rho_{min}$ is the threshold density in free flow for the excitation and persistence of a wide moving jam. Respectively, the flow rate $q_{out}$ is the related threshold flow rate. Thus, there is the threshold point $(\rho_{min}, q_{out})$ in free flow for the excitation and persistence of wide moving jams. This is illustrated in Fig. 3.4:

(1) If the density in free flow $\rho$ (curve $F$) is lower than $\rho_{min}$, no wide moving jams can be excited or persist in this flow. In other words, states of free flow (curve $F$) below the threshold density $\rho_{min}$, i.e., states in which the
vehicle density is within the density range\(^3\)

\[ 0 < \rho < \rho_{\text{min}} \quad (0 < q < q_{\text{out}}) , \]  

are stable states.

(2) If the density in free flow \( \rho \) equals or exceeds \( \rho_{\text{min}} \), a wide moving jam occurs in this flow when a local perturbation of traffic variables appears whose amplitude exceeds the critical amplitude. In other words, free flow is in a metastable state with respect to moving jam formation if the density in this flow \( \rho \) equals or exceeds \( \rho_{\text{min}} \). Thus, states of free flow above this density, i.e., states in which the vehicle density is in the density range

\[ \rho_{\text{min}} \leq \rho < \rho_{\text{max}}^{(\text{free})} \quad \left( q_{\text{out}} \leq q < q_{\text{max}}^{(\text{free})} \right) , \]  

are metastable with respect to moving jam formation (Fig. 3.4). In (3.17)

\[ \rho = \rho_{\text{max}}^{(\text{free})} \]  

and

\[ q = q_{\text{max}}^{(\text{free})} \]  

are the critical (limit) density and flow rate in free flow, respectively.

To explain the metastability of free flow with respect to wide moving jam emergence \([367]\), note that wide moving jams cannot be formed in any state of free flow situated below the line \( J \), i.e., when the density in free flow \( \rho < \rho_{\text{min}} \) (the flow rate \( q < q_{\text{out}} \)). Let the state of free flow directly upstream of a wide moving jam be related to a point \( k \) in the flow–density plane below the line \( J \) (Fig. 3.5a,b). Because the velocity of the upstream front of the wide moving jam \( v_{\text{g}}^{(\text{up})} \) equals the slope of a line \( K \) (from a point \( k \) in free flow to the point \( (\rho_{\text{max}}, 0) \)), the related absolute value \( |v_{\text{g}}^{(\text{up})}| \) is always lower than that of the downstream jam front \( |v_{\text{g}}| \), which is determined by the slope of the line \( J \):

\[ |v_{\text{g}}^{(\text{up})}| < |v_{\text{g}}| . \]  

Therefore, the width of the wide moving jam gradually decreases.

Now consider another wide moving jam, in which a free flow upstream of the jam is related to a state in the flow–density plane above the line \( J \) (Fig. 3.5c,d). In this case, the velocity of the upstream front of the wide moving jam \( v_{\text{g}}^{(\text{up})} \) equals the slope of a line \( N \) (from a point \( n \) in free flow to the point \( (\rho_{\text{max}}, 0) \)), i.e., the related absolute value \( |v_{\text{g}}^{(\text{up})}| \) is always greater

\(^3\) There is the well-known relationship between the density and the flow rate in free flow (curve \( F \) in Fig. 3.4). For this reason, the density range (3.16) is associated with the flow rate range given in parentheses in (3.16).
Fig. 3.5. Metastability of free flow with respect to moving jam emergence [367].

(a, c) Qualitative forms of wide moving jams at two different densities in free flow upstream of the jams, \( \rho = \rho_k \) (a), and \( \rho = \rho_n \) (c). (b, d) Representation of the propagation of the downstream front of a wide moving jam (line \( J \)) and the upstream front of the wide moving jam (lines \( K \) (b) and \( N \) (d)) in the flow–density plane for wide moving jams in (a) and (c), respectively.

than that of the downstream front \( |v_g^{(\text{up})}| \):

\[
|v_g^{(\text{up})}| > |v_g|.
\]  

(3.21)

Therefore, the width of the wide moving jam in Fig. 3.5c should be gradually increasing. For these reasons, wide moving jams can be formed in states of free flow that lie on or above the line \( J \), i.e., under the conditions (3.17).

The threshold flow rate \( q_{\text{out}} \) and the threshold density \( \rho_{\text{min}} \) for wide moving jam emergence are related to the condition that the velocity \( v_g \) of the downstream front of a wide moving jam is equal to the velocity of the upstream front of the jam \( v_g^{(\text{up})} \):

\[
v_g = v_g^{(\text{up})}.
\]  

(3.22)

In this case, the line that represents the propagation of the upstream front of the wide moving jam coincides with the line \( J \) in the flow–density plane.
The condition (3.22) separates metastable states of free flow in which (3.21) is valid and stable states of free flow in which (3.20) is valid. The velocity $v_{g}^{(\text{up})}$ of the upstream front of the wide moving jam corresponding to the Stokes shock-wave formula (3.5) is equal to

$$v_{g}^{(\text{up})} = -\frac{q_{h}}{\rho_{\text{max}} - \rho_{h}},$$

(3.23)

where $q_{h}$ and $\rho_{h}$ are the flow rate and density in free flow upstream of the wide moving jam, respectively. Corresponding to (3.11) and (3.23), the condition (3.22) can be satisfied only if in (3.23) the flow rate $q_{h} = q_{\text{out}}$ and the density $\rho_{h} = \rho_{\text{min}}$. This explains the above statement that the condition (3.22) determines the threshold flow rate $q_{\text{out}}$ and the threshold density $\rho_{\text{min}}$ with respect to wide moving jam emergence. Thus, we find that corresponding to (3.12) the time delay $\tau_{\text{del}}^{(a)}$ in vehicle acceleration (Sect. 3.2.6) is responsible for the threshold point for wide moving jam emergence in free flow.

Metastability of free flow with respect to wide moving jam emergence can also be explained from driver behavior. First note that the mean flow rate $q_{\text{max}}^{(\text{free})}$ at the limit (maximum) point of free flow is equal to

$$q_{\text{max}}^{(\text{free})} = \frac{1}{\tau_{\text{min}}^{(\text{free})}},$$

(3.24)

where $\tau_{\text{min}}^{(\text{free})}$ is the mean gross time gap between vehicles at the limit (maximum) point of free flow ($\rho_{\text{max}}^{(\text{free})}$, $q_{\text{max}}^{(\text{free})}$). The mean time delay $\tau_{\text{del}}^{(a)}$ in (3.13) is considerably higher than $\tau_{\text{min}}^{(\text{free})}$:

$$\tau_{\text{del}}^{(a)} > \tau_{\text{min}}^{(\text{free})}.$$  

(3.25)

The condition (3.25) is related to the following feature of driver behavior. When the average vehicle speed is nearly homogeneous in space and time in a freeway lane, drivers accept considerably lower time gaps in free flow than in the case when they accelerate at the downstream front of a wide moving jam from the standstill within the jam to free flow. The condition (3.25) explains empirical formula (2.14). This condition (3.25) is also responsible for (3.17) that determines the range of metastable states of free flow with respect to wide moving jam formation.

In the theory of moving jams [367], there are also solutions for narrow moving jams. As a wide moving jam, a narrow moving jam (Fig. 3.6) propagates upstream. However, the density $\rho_{\text{max}}^{(\text{narrow})}$ within this narrow moving jam is lower than the density $\rho_{\text{max}}$ within the wide moving jam. When a growing local perturbation occurs in free flow, first a narrow moving jam emerges. The narrow moving jam can later transform into a wide moving jam. This is the usual scenario of spontaneous wide moving jam emergence in free flow (F-$\rightarrow$J transition) within the scope of the fundamental diagram approach [367].
The predictions of this theory of wide moving jams about characteristic jam parameters [367] have indeed been verified in empirical investigations [166]. It has also been found that wide moving jam propagates through bottlenecks and complex states of traffic flow while maintaining the velocity of the downstream jam front [166, 207, 208]. In other words, independent of the state of traffic flow and existence of freeway bottlenecks, the velocity of the downstream jam front remains the characteristic parameter.

However, in models within the scope of the fundamental diagram approach under consideration, at the limit point of free flow (3.18), (3.19) (Fig. 3.4), the amplitude of the critical local perturbation for an $F \rightarrow J$ transition is zero. This means that the critical point of free flow for the $F \rightarrow J$ transition is given by

$$\rho_{cr} = \rho_{\text{max}}^{(\text{free})}, \quad q_{cr} = q_{\text{max}}^{(\text{free})}$$

in these traffic flow theories and models. If the density

$$\rho > \rho_{cr} = \rho_{\text{max}}^{(\text{free})}$$

and consequently the flow rate

$$q > q_{cr} = q_{\text{max}}^{(\text{free})}$$

then free flow is unstable against moving jam emergence, i.e., the $F \rightarrow J$ transition must occur spontaneously in these theories.

Thus, the existence of critical (limit) density in free flow (3.18) in the fundamental diagram approach is associated with moving jam emergence. This means that if the density in free flow gradually increased and the density reaches the critical density (3.18), small local perturbations (in fact, infinitesimal perturbations) in an initial steady model state of free flow grow and lead to one or more moving jams in that free flow.

The theory by Kerner and Konhäuser on the metastability of free flow with respect to moving jam formation and on the characteristic parameters of wide moving jam propagation [367] has subsequently been incorporated into a large number of theories and models of traffic flow. These include theories based on the optimal velocity model of Bando, Sugiyama, and coworkers [371], on cellular automata (CA) of Barlovic, Schreckenberg, and coworkers [378], Helbing and Schreckenberg [379], and other models, in particular the model...
of Krauß et al. [377], Mahnke, Kühne et al. [380, 402], Helbing, Treiber, and coworkers [191, 393, 397], Nishinari and Takahashi [401], Fukui et al. [404] (see also references in the reviews [33, 35, 36, 38]). The theory of wide moving jams and their evolution along a homogeneous road [367] and the theory of a deterministic emergence of wide moving jams at a freeway bottleneck due to an on-ramp [381] are also the basis for the diagram of congested patterns at the on-ramp in the context of the fundamental diagram approach first developed by Helbing et al. [35, 393].

It must be noted that in recent years many traffic simulation tools based on macroscopic and microscopic traffic flow models within the scope of the fundamental diagram approach have been developed. These tools are widely used in a variety of engineering applications. Examples of these application fields are simulations of traffic congestion, congestion control, and dynamic traffic assignment in traffic networks, studies of the effects of new mechanical and electronic systems such as assistance systems on the whole system of traffic, computer-assisted testing of vehicle motor and transmission, etc. (e.g., [27, 279, 288, 291–323, 420, 485–492, 520]).

3.3 Drawbacks of Fundamental Diagram Approach in Describing of Spatiotemporal Congested Freeway Patterns

3.3.1 Shock Wave Theory

Note that if in a traffic flow model within the scope of the fundamental diagram approach, at any vehicle density in an initial homogeneous traffic flow, fluctuations do not grow, these models cannot divulge the spontaneous occurrence of moving jams. These models also cannot usually reveal the characteristic parameters of wide moving jams as observed in empirical investigations [166].

As shown in [403], this for example applies to the classic Lighthill–Whitham–Richards theory [2, 3] (3.6), (3.7). It must be noted that the Lighthill–Whitham–Richards model (3.6), (3.7) and the related kinematic wave (KW) theory are pioneering works that were the source for many other traffic flow models and theories. However, it is now clear that this theory cannot explain the majority of important empirical spatiotemporal congested pattern features that have recently been found [203, 205, 207–215, 218] (Sect. 2.4). In particular, the LWR theory cannot show the following fundamental empirical phenomena in traffic:

(i) the probabilistic nature of the F→S transition (breakdown phenomenon)
(ii) the spontaneous formation of GPs at bottlenecks.

Let us compare some features of shock waves in the LWR theory and of wide moving jams in the Kerner–Konhäuser theory discussed above. In both
3.3 Drawbacks of Fundamental Diagram Approach

theories, a solution nearly consists of two neighboring steady states separated by a discontinuity.\(^4\) On the one hand, the velocity of the downstream front of a wide moving jam can be formally written as a shock wave velocity (3.9) of the Lighthill–Whitham–Richards KW theory,

\[
v_g = -\frac{q_{\text{out}}}{\rho_{\text{max}} - \rho_{\text{min}}},\tag{3.29}
\]

between two densities \(\rho_1 = \rho_{\text{max}}\) and \(\rho_2 = \rho_{\text{min}}\), where it has been taken into account that \(q(\rho_2) = q(\rho_{\text{min}}) = q_{\text{out}}\) and \(q(\rho_1) = q(\rho_{\text{max}}) = 0\). On the other hand, in the Lighthill–Whitham–Richards KW theory, the densities \(\rho_1\) and \(\rho_2\) upstream and downstream of the shock can be arbitrary points on the fundamental diagram, i.e., an infinite multitude of different shocks with different velocities can be formed. In contrast to the Lighthill–Whitham–Richards theory, in the theory of wide moving jams [367] the densities \(\rho_1 = \rho_{\text{max}}\) and \(\rho_2 = \rho_{\text{min}}\) are the two distinct characteristic densities that determine the downstream front of a wide moving jam. There is also only one distinct characteristic velocity \(v_g\) of this jam front in the latter theory [367].

Furthermore, for a describing of the process of vehicle acceleration from a wide moving jam one should take into account a time delay between any two vehicles following one another accelerating from the standstill inside the jam. However, there is no such time delay in the Lighthill–Whitham–Richards model (3.6), (3.7). Apparently for this reason, it has been shown [403] that the Lighthill–Whitham–Richards model and the KW theory cannot describe the characteristic parameters of wide moving jams observed in empirical investigations [166]. The empirical results represented by (2.14) and in Fig. 2.8 also cannot be predicted by the Lighthill–Whitham–Richards theory. As in empirical observations (Sect. 2.4.7), in theory of moving jams [367], characteristic parameters of wide moving jams (Sects. 3.2.5 and 3.2.6) occur as a result of “self-organization” during the dynamics of wide moving jam emergence. There is no self-formation of wide moving jams in the Lighthill–Whitham–Richards model and the KW theory. For these reasons, the terms “kinematic waves” and “shock waves,” which are automatically associated with the Lighthill–Whitham–Richards theory, will not be used in this book.

\(^4\) Note that such kinds of spatiotemporal solutions are well-known in the physics of nonlinear phenomena (e.g., [336, 337, 339]). An example is the theory of waves in connected excitable elements, which was developed in 1946 by Wiener and Rosenblueth [468]. This theory is the basis for further theories of autowaves in chemical, biological, and physical systems (e.g., [336, 337]). However, it must be noted that solutions to reasonable problems in traffic flow theory should not necessarily consist of steady states separated by some discontinuities. In particular, there are important spatiotemporal features of synchronized flow (like transformations between different synchronized flow states that are far from steady states of synchronized flow or behavior of growing small amplitude perturbations in synchronized flow) that cannot be described by solutions consisting of steady states separated by discontinuities (Chaps. 4–7).
again. The only exception is the Stokes shock-wave formula (3.5), which will often be used in this book.

3.3.2 Models and Theories of Moving Jam Emergence in Free Flow

How well do traffic flow theories within the scope of the fundamental diagram approach predict spontaneous moving jam emergence? How well do they really explain the phenomena observed in traffic flow? The relationship between theory and experiment holds true for the characteristic parameters of a wide moving jam, such as the flow rate in the jam outflow, the metastability of free flow with respect to jam formation, and wide moving jam propagation first predicted in [367] and then incorporated into a large number of models and theories (see references in the reviews [33,35,36,38]). It seems that these are the only theoretical results in models that follow the fundamental diagram approach, which are in accordance with empirical congested pattern features. We will see that these theories within the fundamental diagram approach did not predict, and cannot explain, important features of various phase transitions and spatiotemporal traffic patterns that have been observed in real traffic flow [208,218].

However, it should first be noted that in recent years, considerable progress has been made in understanding spatiotemporal pattern formation and pattern features that can support traffic flow models in the fundamental diagram approach (e.g., [42–46,362,366,367,370–386,391–395,397–399,401, 422,436,442–448]; for reviews see [33,35,36,38]). In particular, in this approach two main classes of traffic flow models can be identified that claim to exhibit moving jam emergence and other congested patterns upstream of an on-ramp:

(i) models in which moving jams occur spontaneously in free flow at an on-ramp if the flow rate upstream of the on-ramp $q_{\text{in}}$ is sufficiently high and the flow rate to the on-ramp $q_{\text{on}}$ gradually increases from zero. However, the range of the flow rate to the on-ramp where moving jams occur spontaneously is limited. Beginning at a high enough flow rate to the on-ramp $q_{\text{on}}$, spatial homogeneous states of traffic flow (e.g., [191,367,370, 371,381,395]), which have been called “homogeneous congested traffic” (HCT) [35,393], occur upstream of the on-ramp. In these homogeneous states no moving jams emerge spontaneously;

(ii) models in which moving jams occur spontaneously in free flow at the on-ramp beginning at some flow rate to the on-ramp $q_{\text{on}}$. However, no HCT occurs in these models.

How do the “synchronized flow” and “wide moving jam” traffic phases emerge in an initially free flow at an isolated bottleneck (i.e., a bottleneck far from other effectual bottlenecks), e.g., at the bottleneck due to an on-ramp? Empirical observations suggest that the following scenarios are responsible
for the phase transitions and pattern evolution in traffic flow at the on-ramp [208,218]:

(1) Moving jams do not emerge spontaneously in an initial free flow at the on-ramp when the flow rate on the main road upstream of the on-ramp is high enough and the flow rate to the on-ramp gradually increases from zero. Instead of moving jams, an $F \rightarrow S$ transition occurs at the on-ramp. As a result of the $F \rightarrow S$ transition, synchronized flow occurs at the bottleneck. This synchronized flow possesses the following empirical features.

(2) At a low enough flow rate to the on-ramp, the vehicle speed in synchronized flow upstream of the on-ramp is relatively high. Moving jams do not necessarily emerge in that synchronized flow. If the flow rate to the on-ramp is high, then the vehicle speed in the synchronized flow is low and moving jams, in particular wide moving jams, emerge in that synchronized flow.

(3) The latter means that moving jams emerge in an initial free flow due to the following sequence of phase transitions: first an $F \rightarrow S$ transition occurs. Later, and usually at another freeway location, an $S \rightarrow J$ transition occurs, i.e., wide moving jams can occur spontaneously only due to $F \rightarrow S \rightarrow J$ transitions.

(4) The lower the average vehicle speed in synchronized flow upstream of the on-ramp, the higher the frequency of moving jam emergence in that synchronized flow. This means that moving jam emergence persists up to the highest possible values of the flow rate to the on-ramp. Traffic states of high density and low vehicle speed, where moving jams do not emerge, are not observed in synchronized flow upstream of the on-ramp.

The empirical results in item (1)–(3) are in qualitative contradiction with both model classes (i) and (ii) in the fundamental diagram approach. In these models, at high enough initial flow rates upstream of the on-ramp moving jams must emerge in an initial free flow if the flow rate to the on-ramp gradually increases from zero [35]. The last empirical result in item (4) means that on average the higher the vehicle density, the lower the stability of traffic flow with respect to moving jam emergence in that flow. This result of observations, which seems to be intuitively obvious to any driver, is in qualitative contradiction with the models of class (i) in the fundamental diagram approach where HCT, i.e., homogeneous congested traffic with high density and low vehicle speed must occur where moving jams do not emerge [35].

To clarify these conclusions, we note that at a high enough given initial flow rate on a freeway upstream of the on-ramp, the main model features of congested patterns upstream of the on-ramp [35, 36, 393, 395, 397] may be illustrated with the following simplified theoretical schemes:

- high flow rate on the main road upstream of the on-ramp and low flow rate to the on-ramp $\rightarrow$ different kind of moving jams emerge spontaneously;
high enough flow rate to the on-ramp → HCT occurs in which the density is high, the speed is very low and no moving jams emerge spontaneously.

In empirical observations [218, 221, 222], however, the following *empirical schemes* are observed:

- low flow rate to the on-ramp → synchronized flow in which the density is relatively low and the speed relatively high occurs; moving jams should not necessarily emerge;
- high enough flow rate to the on-ramp → moving jams emerge spontaneously in synchronized flow upstream of the on-ramp.

It can be seen that the theoretical schemes [35, 36, 393, 395, 397] and the empirical schemes [218, 221, 222] are inconsistent.

However, it must be noted that this critical consideration does not affect many other mathematical ideas introduced and developed in models based on the fundamental diagram, for example those concerning with modeling vehicle safety conditions, fluctuations, vehicle acceleration and deceleration, various vehicle time delays, and other important effects.

In particular, pertinent pioneering mathematical ideas were introduced in earlier models and theories by Lighthill and Whitham [2], Richards [3], Herman, Montroll, Potts, Rothery, Gazis [4, 5], Komentani and Sasaki [6], Prigogine [14], Newell [10], Whitham [362], Payne [15], Gipps [17], Wiedemann [25], Nagel, Schreckenberg, Schadschneider, Santen, and coworkers [364, 373, 374, 378, 426, 429], Bando, Sugiyama, and colleagues [370], Takayasu and Takayasu [428], Krauß et al. [377], Mahnke, Kühne, Lubashevsky, and coworkers [376, 402, 424, 435], Helbing, Treiber, and coworkers [191, 422, 431], Wolf [32], Nishinari and Takahashi [401], Fukui, Ishibashi, and coworkers [375, 404], Havlin, Tomer, and coworkers [405, 430], Wagner, and coworkers [432–434], and by many other groups (see references in the reviews by Gartner et al. (eds.) [31], Wolf [32], Chowdhury et al. [33], Helbing [35], Nagatani [36], Nagel et al. [38]). These mathematical ideas are also very important elements of the three-phase traffic theory [207–210, 221] (Chaps. 4–8) and of a microscopic three-phase traffic theory [329–331] considered in Part III of this book. The main feature of the three-phase traffic theory is that this theory *rejects* the basic hypothesis concerning the fundamental diagram adopted by previous traffic flow theories and models. The three-phase traffic theory introduces a new phase of traffic flow, synchronized flow, whose steady states cover a 2D region in the flow–density plane (Sect. 4.3). This enables us to overcome the cited problems of the fundamental diagram approach, and to explain empirical spatiotemporal congested pattern features found in [208, 218, 221, 222].
3.3.3 Models and Theories with Variety of Vehicle and Driver Characteristics

There are many behavioral traffic flow theories where spatiotemporal traffic patterns should be explained by different vehicle and/or driver characteristics. Daganzo [242, 243] proposed a behavioral theory of traffic dynamics in which traffic flow phenomena are to be explained in terms of different driver characteristics. There are at least two qualitatively different driver classes in this theory: “aggressive” drivers and “timid” drivers. Treiber and Helbing [394] proposed a model in which there are two fundamental diagrams, one for standard vehicles and the other for long vehicles (trucks).

However, these and other models and theories based on the fundamental diagram approach have the same qualitative problems in describing spatiotemporal congested patterns discussed above. In particular, the model with different vehicle characteristics [394] exhibits qualitatively the same features of moving jam emergence in free flow as those discussed in Sect. 3.3.2 for other models based on the fundamental diagram approach where only one type of vehicle is considered. These model features are qualitatively inconsistent with empirical results of congested pattern emergence [218].

The behavioral traffic flow theory [242, 243] cannot show main empirical phenomena in traffic flow discussed in Sect. 2.4. In particular, neither the probabilistic feature of the breakdown phenomenon nor spontaneous moving jam emergence can mathematically be shown in [242, 243].

In contrast to the behavioral theories [242, 243, 394], we will see in Part III of this book that the main phenomena of spatiotemporal traffic pattern formation (Sect. 2.4) can be shown and predicted in a microscopic three-phase traffic theory where all drivers and all vehicles have the same characteristics and parameters [330]. Thus, to explain these main qualitative features of traffic phenomena, a consideration of various vehicle and driver characteristics should not necessarily be made in a behavioral traffic flow theory. Examples of these fundamental congested traffic pattern features include (see Sect. 2.4 for more detail)

(i) empirical breakdown phenomenon in an initial free flow associated with an $F \rightarrow S$ transition rather than an $F \rightarrow J$ transition. The $F \rightarrow S$ transition has a probabilistic nature;

(ii) wide moving jams that emerge in an initial free flow due to the sequence of $F \rightarrow S \rightarrow J$ transitions. Firstly, a region of synchronized flow occurs in the free flow, and only later (and usually at other than the location of the $F \rightarrow S$ transition) can wide moving jams emerge in that synchronized flow;

(iii) two main types of congested patterns at an isolated freeway bottleneck: the SP and the GP. In particular, if a bottleneck is due to an on-ramp, at higher flow rates to the on-ramp an GP occurs, and at lower flow rates to the on-ramp an SP can appear.
This set of fundamental congested pattern features could not be found in any theory and model based on the fundamental diagram approach, regardless of whether different driver characteristics and/or various vehicle parameters are used in the theory.

Thus, different driver characteristics and/or different vehicle parameters cannot resolve the cited fundamental problems of a theoretical description of empirical spatiotemporal congested patterns if these models and theories conform to the fundamental diagram approach.

Obviously, in real traffic there are drivers with very different characteristics, and vehicles that have very different parameters (e.g., differing desired and safe driving speeds, aggressive and timid driver behavior, standard vehicles and long vehicles, and so on). These differing characteristics and parameters can change certain quantitative spatiotemporal congested pattern parameters and conditions for pattern emergence. There can also be some secondary and specific effects in congested traffic that can be caused by these different driver behavioral characteristics and different vehicle parameters. This is confirmed by recent numerical simulations of a microscopic model based on three-phase traffic theory where different driver and vehicle characteristics have been used [332] (see Chap. 20). These simulations show qualitatively the same fundamental features of congested patterns as those in the case when all drivers and all vehicles have the same characteristics and parameters [330].

Behavioral assumptions of three-phase traffic flow theory that enable us to explain and to predict main qualitative features of empirical traffic flow dynamics will be discussed in Sects. 4.3.1 and 8.6.

### 3.3.4 Application of Classical Queuing Theories to Freeway Congested Traffic Patterns

It has been mentioned that there are two traffic phases in the congested regime, wide moving jams and synchronized flow, with totally different properties. Which of the dynamics of the two phases can be correctly described by the various classical queuing theories (e.g., [23, 28])? Is the “queue” on a freeway a sequence of vehicles that are remaining and waiting within a wide moving jam, or is it a sequence of vehicles in synchronized flow?

In a model of freeway operation as a queuing at multiple servers (e.g., [23, 28]), the freeway is considered as consisting of a number of bottlenecks regarded as “servers,” each with its own “service rate,” which cannot only be different at different locations, but can change suddenly if an “incident” occurs. The location of the bottleneck is assumed known. Sometimes every meter of the freeway is regarded as a server at which a queue might form. Therefore, in the models of freeway operation as queuing at multiple servers, the values of the service rates are the given boundary conditions for the description of the spatiotemporal behavior of the queue of vehicles formed upstream of the related server. However, neither do conditions upstream of a
bottleneck result solely from the flow rate at the downstream boundary of congestion, nor does the flow rate at the downstream boundary result solely from conditions upstream: spatiotemporal competition both between feedback from upstream to the downstream boundary of congestion and feedback from the downstream boundary of congestion to upstream determines features of traffic patterns at a bottleneck. Indeed, empirical data related to acceleration of vehicles from lower speeds in synchronized flow at a bottleneck to higher speed in flow downstream of the bottleneck (see Sect. 13.4) show that the discharge flow rate depends on conditions in congested regime upstream if free flow is realized downstream. In contrast to these empirical results, in any queuing theory conditions upstream result solely from the flow rate at the downstream boundary of congestion. In other words, in contrast to queuing theories, on unsignalized freeways “service rate” at the downstream boundary must not be either given or predefined, even as a function of time. The flow rate downstream of congestion must be found, along with other spatiotemporal distributions of traffic variables upstream, in particular inside the congestion region.

Therefore, queuing theories cannot be relied upon for a correct description of congestion on unsignalized freeways, and the term queue will henceforth not be used to describe traffic on freeways.

However, it must be stressed that this critical conclusion does not apply to many other mathematical ideas and queuing theories (e.g., [19, 28, 31, 244–291, 316]) introduced and developed in models for city traffic. In this case, the traffic dynamics is determined by light signals and other traffic regulations at road intersections, rather than driver interactions alone. We also use the important achievements of these traffic flow theories [324–328] (see Sect. 22.4).

### 3.4 Conclusions

(i) Ideas of traffic flow theories within the fundamental diagram approach about different driver time delays, mathematical description of driver acceleration and deceleration, safety conditions, and other effects are very important elements of any traffic theory. However, probably due to the fundamental hypothesis of these theories and models about the existence of the theoretical fundamental diagram for steady-state model solutions (hypothetical model solutions in the unperturbed and noiseless limit in which all moving vehicles remain the same distance apart and move at the same time-independent vehicle speed), these theories cannot predict fundamental empirical features of phase transitions and spatiotemporal congested pattern features of real traffic.

(a) The classic Lighthill–Whitham–Richards model of shock (kinematic) waves cannot predict empirical results relating to spontaneous moving jam emergence and empirical features of wide moving jams.
(b) Models and theories based on the fundamental diagram approach that claim to predict spontaneous moving jam emergence cannot predict
(1) the empirical features of the F→S transition (breakdown phenomenon);
(2) empirical spatiotemporal features of the “synchronized flow” traffic phase, in particular features related to spontaneous wide moving jam emergence in synchronized flow (S→J transition);
(3) the empirical observed sequence of F→S→J transitions responsible for spontaneous wide moving jam emergence in free flow.

(ii) Queuing theories cannot accurately describe empirical spatiotemporal congested pattern features on freeways and other unsignalized highways.

(iii) In contrast, queuing theories are highly applicable to city traffic: they can be used for a correct mathematical description of empirical features of traffic dynamics when the dynamics are completely governed by light signals and other traffic regulations at road intersections.

(iv) In the theory of wide moving jams within the fundamental diagram approach, there is a range of the vehicle density where free flow is metastable with respect to wide moving jam formation. There are characteristic parameters and features of wide moving jams. One of these characteristic jam features is that the downstream front of wide moving jams propagates steadily along a road. The line $J$ represents this steady propagation of the downstream front of a wide moving jam in the flow–density plane. The line $J$ is defined by the point in the flow–density plane that is related to the jam outflow and by the slope of that line, given by the velocity of the downstream jam front.
4 Basis of Three-Phase Traffic Theory

4.1 Introduction and Remarks on Three-Phase Traffic Theory

To explain empirical spatiotemporal features of congested patterns, in 1996–1999 the author introduced three-phase traffic theory, in which the empirical complexity of freeway traffic was explained in terms of diverse spatiotemporal features of three traffic phases: free flow, synchronized flow, and wide moving jam [205,207–211].

In this and the next chapters of Part I of this book, we will try to show from the fundamental hypothesis of three-phase traffic theory (Sect. 4.3) and from empirical spatiotemporal traffic pattern features [166,167,203,205,207–211] (Sect. 2.4) how the main conclusions and results of the three-phase traffic theory [205,207–211,218] can be derived without complex mathematical formulae or mathematical traffic flow models.

The three-phase traffic theory is a qualitative theory. This should help readers understand the main ideas and results of this theory. Detailed considerations of the empirical basis [166, 167, 203, 205, 207–211, 218] and of mathematical results of three-phase traffic theory [329–331] are considered in Parts II and III of this book, respectively.

To make the three-phase traffic theory as easy as possible to understand, we neglect and simplify certain real traffic features. However, we try to retain salient freeway traffic features. For example, in this simplification we do not consider the very real difference between traffic in different lanes of multilane traffic. All vehicles and all drivers are considered to have the same mean parameters and characteristics, e.g., all vehicles have the same length, all drivers have the same desired (maximum) speed in free flow, they exhibit the same reaction time, and so on. We also neglect or simplify some other (sometimes seemingly important) effects. These simplifications enable us to highlight in the three-phase traffic theory some of the most important features of traffic.

We will see that fundamental qualitative features of congested pattern emergence observed in empirical investigations [208,218] (Sect. 2.4) can be explained in the three-phase traffic theory even when all drivers are suggested to have the same characteristics and all vehicles are suggested to have the
same parameters [218,329,331]. Obviously, in real traffic there can be very large differences in driver behavioral characteristics and in vehicle parameters. However, these differences are responsible for some specific effects and they change usually quantitative characteristics of fundamental spatiotemporal congested pattern features only (see Chap. 20). Therefore, differences in driver characteristics and in vehicle parameters can first be neglected when fundamental spatiotemporal congested pattern features are studied.

Three-phase traffic theory discussed in this chapter and Chaps. 5–8 below is an attempt to explain the author’s three-phase traffic theory [205, 207–211, 218] in the simplest way, beginning with some basic empirical background [166,167,203,205,207–211].

In 1992–1995, when the author began to work in traffic science, he had strongly believed that the fundamental diagram approach would be the correct mathematical approximation for a description of empirical traffic flow phenomena (see early papers by the author et al. [366,367,381,382]). The empirical findings [205,207–211] were the reasons why the author was forced in 1996–1999 to change his mind fundamentally about the correctness of the well-known and generally accepted fundamental hypothesis of the fundamental diagram approach to traffic flow theory and modeling (see Chap. 3).1 That is why the author introduced and developed the three-phase traffic theory [205,207–211], which rejects almost all earlier theoretical results about features of spatiotemporal congested patterns and phase transitions in freeway traffic.

4.2 Definition of Traffic Phases in Congested Traffic Based on Empirical Data

4.2.1 Objective Criteria for Traffic Phases in Congested Traffic

In Chap. 1 when we discussed Fig. 1.2 and also in Sect. 2.4, we mentioned the empirical (objective) criteria for two phases in congested traffic, “synchronized flow” and “wide moving jam.” These objective criteria are related to some spatiotemporal congested patterns features. Here we consider these objective criteria in more detail.

---

1 The author was fortunate to work in 1994–1995 for the traffic engineering consulting firm Heusch/Boesefeldt GmbH in Aachen (Germany), where he met many good traffic engineers. Here he had his first opportunity to begin a study of empirical data on congested freeway traffic. The author thanks all colleagues at this firm, in particular Heinz Heusch, Jochen Boesefeldt, Heribert Kirschfink, Hubert Rehborn, Ulrich Uerlings, and Thomas Scheiderer, for their help, critical discussion of earlier traffic flow theories, and their fruitful cooperation.
Numerous empirical studies of congested spatiotemporal patterns at bottlenecks enable us to distinguish between the different traffic phases in congested traffic based on the following objective criteria [205,208,212,213,215]:

[J] a local spatiotemporal congested traffic pattern moving upstream, i.e., a pattern that is spatially restricted by two moving (downstream and upstream) fronts, is identified with the “wide moving jam” phase if for the given “traffic control parameters” (e.g., weather and other environmental and road conditions), the pattern possesses the following characteristic (i.e., unique, coherent, predictable, and reproducible feature): the pattern as a whole local structure propagates through any state of free and synchronized flow and through any bottlenecks (e.g., at on-ramps and off-ramps) while maintaining the mean velocity of the downstream front of the pattern. This velocity is the same for different wide moving jams.

[S] The downstream front of the “synchronized flow” phase is usually fixed at the bottleneck. Upstream of a freeway bottleneck the “synchronized flow” phase characteristically forms diverse spatiotemporal patterns. Even if a moving synchronized flow pattern (MSP) occurs, the velocity of the downstream front of this pattern is not a characteristic parameter: it can vary over a wide range during pattern propagation, and it can be different for different MSPs.

There are many various empirical dynamic effects in congested traffic. In particular, the average vehicle speed between different freeway lanes can be synchronized in congested traffic (e.g., [88,114]). Furthermore, in congested traffic there can be a wide spread of empirical data in the flow–density plane (e.g., [30,88]). These dynamic effects can occur in both traffic phases in congested traffic, synchronized flow and wide moving jam. There are also different hysteresis effects associated with congested traffic (e.g., [21,86]). However, to distinguish between the “synchronized flow” and “wide moving jam” traffic phases in congested traffic the objective criteria [S] and [J] should be used rather than: the speed synchronization effect, or the effect of the wide spread of empirical data in the flow–density plane, or the hysteresis effects.

An example of the application of the objective criteria to distinguish between the “wide moving jam” phase and the “synchronized flow” phase is shown in Fig. 4.1. In this figure, empirical congested traffic states in the time–space plane over a freeway section 24 km long (Fig. 2.1) are shown. This section has three effectual bottlenecks marked $B_1$, $B_2$, and $B_3$ in Fig. 4.1. It can be seen that a moving jam propagates through these three bottlenecks and through different states of synchronized flow (as labeled in Fig. 4.1) while maintaining the velocity of the downstream jam front $v_g$ [166]. Therefore, this is a wide moving jam. The distances between detectors are known. For this reason, the velocity of the downstream jam front can be easily calculated. In this case, it is found that $v_g = -16$ km/h.

In contrast to the wide moving jam, after a congested pattern has occurred upstream of the bottleneck $B_2$ the downstream front of the pattern is fixed
Fig. 4.1. Mean vehicle trajectory related to vehicles moving through all three traffic phases. Traffic data measured on A5-South (Fig. 2.1) are taken from Fig. 1.2 at the bottleneck \( B_2 \). Thus, this pattern is related to the “synchronized flow” phase.

4.2.2 Explanation of Terms “Synchronized Flow” and “Wide Moving Jam”

The term “synchronized flow” is meant to reflect the following features of this traffic phase:

(i) It is a continuous traffic flow with no significant stoppage, as often occurs inside a wide moving jam. The word “flow” reflects this feature.

(ii) There is a tendency towards synchronization of vehicle speeds across different lanes on a multilane road in this flow. In addition, there is a tendency towards synchronization of vehicle speeds in each of the road lanes (bunching of vehicles) in synchronized flow. This is due to a relatively low probability of passing. The word “synchronized” reflects these speed synchronization effects.

The term “wide moving jam” is meant to reflect the characteristic feature of the jam to propagate through any other state of traffic flow and through any bottleneck while maintaining the velocity of the downstream jam front. The phrase “moving jam” reflects the jam propagation as a whole localized structure on a road. To distinguish wide moving jams from other moving jams, which do not characteristically maintain the mean velocity of the downstream jam front, we use the term “wide moving jam.” This relates to the fact that if
a moving jam has a width (in the longitudinal direction) considerably greater than the widths of the jam fronts, and if the vehicle speed inside the jam is zero, the jam always exhibits the characteristic feature of maintaining the velocity of the downstream jam front. Thus, the word “wide” reflects this characteristic feature.

However, one should recall that to distinguish between the “wide moving jam” and “synchronized flow” traffic phases in congested traffic the objective criteria \([J]\) and \([S]\) for traffic phases should be applied rather than the speed synchronization effect (Sect. 4.2.1).

4.2.3 Mean Vehicle Trajectories

Let us consider a mean vehicle trajectory for the last example of traffic data (Fig. 4.1) for a vehicle that begins to move on this freeway section at 7:30. It can be seen that first the vehicle moves in free flow at a high speed (about 120 km/h). About 5 min later the vehicle must slow down because it reaches the wide moving jam. After the vehicle has accelerated from low speed states inside the jam, the vehicle can move again in free flow.

However about 2 min later the vehicle must slow down once more, as it reaches the “synchronized flow” phase. In this traffic phase the vehicle moves very slowly (with an average speed about 30 km/h), and continues moving at this slow rate for about 10 min. Only after the vehicle has accelerated from these low speed states in synchronized flow at the downstream front of synchronized flow can it move at a high speed in free flow.

We can see that drivers cannot usually recognize whether they are in the “wide moving jam” phase or the “synchronized flow” phase. To distinguish between these two traffic phases of congested traffic, some knowledge of the spatiotemporal characteristics of these traffic phases is necessary.

4.2.4 Flow Rate in Synchronized Flow

We must distinguish between the “wide moving jam” phase and the “synchronized flow” phase because it is very important to an understanding of the nature of traffic. Let us consider the example of traffic data in Fig. 4.1 in more detail (Fig. 4.2).

We find that the “synchronized flow” phase is difficult to see if the flow rate is plotted in space and time. Whereas, the “wide moving jam” phase can be seen clearly (Fig. 4.2a). Indeed, although the vehicle speed in synchronized flow can be very low the flow rate in this synchronized flow can be as high as in free flow (Fig. 4.2b). This can also be seen in Fig. 4.2b, if the data at the detectors D14 for free flow and synchronized flow in the flow–density plane is considered.

The empirical result that the flow rate in this synchronized flow can be as high as in free flow is also confirmed by investigations of the \(F \rightarrow S\) transition. It turns out that whereas speed breakdown occurs during this phase
Fig. 4.2. Explanation of the difference between the “wide moving jam” and “synchronized flow” traffic phases. (a) Vehicle speed averaged across all freeway lanes (left) and total flow rate across the freeway (right) as functions of time and space measured at the detectors D1–D21 on June 23, 1998 on a section of the freeway A5-South (Fig. 2.1). (b) Top and middle lines: average vehicle speed \( v \) and flow rate \( q \) as functions of time for free flow (left), synchronized flow (middle), and the wide moving jam (right) for each of the freeway lanes. Bottom line: representation of the corresponding traffic phases in the flow–density plane (\( F \) – free flow, \( S \) – synchronized flow, the line \( J \) – the characteristic line for the downstream front of the wide moving jam). The slope of the line \( J \) equals the velocity of the downstream front of the wide moving jam \( v_g \). In (b) (bottom line) the data per one lane is shown, and is averaged across all three freeway lanes. Taken from [212–214]
transition, flow rate during the F→S transition should not necessarily change (compare Fig. 2.11a and Fig. 2.11b).

### 4.2.5 Empirical Line J

In contrast to this feature of synchronized flow, the average flow rate inside a wide moving jam can be almost zero (Figs. 2.9 and 4.2b). Thus, the density and flow rate inside the jam are approximately related to the traffic state \((\rho_{\text{max}}, 0)\) in the flow–density plane. The steady propagation of the downstream jam front through all other traffic states and through bottlenecks at velocity \(v_g\) can be represented by the empirical line \(J\) (Fig. 4.2b, bottom line, right column). The slope of the line \(J\) is equal to the velocity \(v_g\). The right coordinates of the line \(J\) are related to the traffic state \((\rho_{\text{max}}, 0)\) inside the jam. The left coordinates of the line \(J\) are related to a traffic state in the jam outflow. When free flow prevails in the jam outflow, the flow rate in this jam outflow is equal to \(q_{\text{out}}\) (Fig. 4.2b). Therefore, the left coordinate of the line \(J\) is given by \((\rho_{\text{min}}, q_{\text{out}})\), where \(\rho_{\text{min}}\) is the density in the jam outflow relative to the flow rate \(q_{\text{out}}\). These empirical results confirm theoretical conclusions about wide moving jam propagation made in Sect. 3.2.6.

We can also see empirically that the flow rate in this free flow \(q_{\text{out}}\) is considerably lower than the maximum flow rate in free flow \(q_{\text{max}}^{(\text{free, emp})}\):

\[
q_{\text{max}}^{(\text{free, emp})} > q_{\text{out}} \tag{4.1}
\]

(a more detailed consideration of the empirical line \(J\) can be found in Chap. 11).

If this empirical line \(J\) is drawn together with empirical states of synchronized flow taken from Fig. 4.2b (states \(S\), bottom line, middle column), we find that some of the states of synchronized flow lie above the line \(J\) (Fig. 4.3). This is related to the above conclusion that the flow rate in synchronized flow

![Fig. 4.3. Synchronized flow states in the flow–density plane (circles S) and the line J. Synchronized flow states and the empirical line J are taken from Figs. 4.2b](image-url)
can be as high as in free flow. Thus, in some cases the line $J$ divides empirical states of synchronized flow in two: synchronized flow states above the line $J$, and states below the line $J$.

### 4.2.6 Propagation of Two Wide Moving Jams

Let us consider in more detail another example of wide moving jam propagation shown in Fig. 2.7, which was briefly discussed in Sect. 2.4. The sequence of two moving jams propagates through at least three bottlenecks (the intersections I1, I2, and I3, Fig. 2.2) and through various states of synchronized flow (Fig. 2.7c, bottom). The velocity of the downstream front of each of these moving jams can be found by measuring the time at which the downstream jam front appears at different detectors. Two such measurements of vehicle speed and flow rate measured at detectors D12 and the detectors D6 are shown in Figs. 2.7c,d, respectively. Because distances between detectors are known (Fig. 2.2), it is easy to calculate the velocity of the downstream jam fronts. To within the accuracy of the measurements (about 10%), this velocity is the same for the two moving jams, $v_g = -15 \text{ km/h}$, and the velocity is maintaining during jam propagation through bottlenecks and synchronized flow. Therefore, each of these moving jams belongs to the “wide moving jam” phase.

From these two empirical examples we see that the velocity $v_g$ of the downstream front of a wide moving jam, is indeed a characteristic parameter, i.e., this velocity does not depend on traffic demand or initial conditions. Under the same traffic control parameters this velocity is also the same for two different wide moving jams in Fig. 2.7. These conclusions are confirmed by an empirical study of many other wide moving jams [169,218]. Furthermore, if free flow prevails in the wide moving jam outflow, the flow rate $q_{out}$ in that outflow (as well as the related density $\rho_{min}$ and average speed $v_{max}$) is also a characteristic parameter. However, it must be noted that the characteristic velocity $v_g$ and other characteristic jam parameters ($q_{out}$, $\rho_{min}$, and $v_{max}$) can depend considerably on control parameters of traffic including weather and other road (travel) conditions (e.g., whether it is a workday or weekend, etc. (Chap. 11)).

In contrast to the wide moving jams, after a congested pattern has occurred upstream of the on-ramp at the detectors D7, the downstream front of the pattern is fixed at the on-ramp rather than propagating upstream. The downstream front of the pattern is shown by the dashed line Fig. 2.7a. This pattern therefore belongs to the “synchronized flow” phase. Note that inside the downstream front of synchronized flow (that is fixed at the detectors D7), vehicles accelerate from synchronized flow to free flow. In other words, the downstream front of synchronized flow separates synchronized flow upstream of D7 from free flow downstream.

In this example, we can also see that synchronized flow cannot almost be distinguished from free flow in the spatiotemporal distribution of the flow
rate (Fig. 2.7b). In contrast, the flow rate inside wide moving jams can be zero (Fig. 2.7c,d). For this reason, jam propagation can also be clearly seen in the flow rate distribution (Fig. 2.7b). Thus, this feature of the relatively high flow rate in synchronized flow can be used as an additional criterion for the “synchronized flow” phase.

However, it must be stressed that neither the mentioned feature of the relatively high flow rate in synchronized flow nor some other isolated local measurements of traffic variables (e.g., measurements at only one freeway location) can be used alone to reliably distinguish a wide moving jam from synchronized flow. To make such a distinction, empirical spatiotemporal features of these phases, as formulated in the objective criteria \([J]\) and \([S]\), should be used.

Thus, only simultaneous measurements of traffic both in space and time together with a knowledge of effectual freeway bottlenecks, can yield the information necessary to precisely and accurately distinguish between a wide moving jam and synchronized flow in congested traffic.

### 4.3 Fundamental Hypothesis of Three-Phase Traffic Theory

Three-phase traffic theory is a qualitative theory. As in many other qualitative traffic flow theories, the starting point of the three-phase traffic theory is an assumption about features of hypothetical steady states of traffic flow. In the three-phase traffic theory, an assumption about features of hypothetical steady states of synchronized flow is important. This is because this assumption is the basis for other hypotheses of this qualitative theory, which should explain phase transitions in traffic flow and spatiotemporal congested patterns (Chaps. 5–8). For this reason, this assumption is called the fundamental hypothesis of the three-phase traffic theory. The fundamental hypothesis of the three-phase traffic theory \([205, 208]\) determines features of hypothetical steady states of the “synchronized flow” phase. Steady states of the “synchronized flow” phase are hypothetical synchronized flow states in which vehicles move at the same distance from one another with the same time-independent vehicle speed.

The fundamental hypothesis of three-phase traffic theory reads as follows:

- Hypothetical steady speed states of synchronized flow, i.e., synchronized flow states in which all vehicles move at the same distances from one another and with the same time-independent speed, cover a two-dimensional region in the flow–density plane (Fig. 4.4). This means that in these hypothetical steady speed states of synchronized flow a given vehicle speed \(v\) is relative to an infinite multitude of different vehicle densities (a finite density range, \([\rho_1, \rho_2]\) in Fig. 4.5), and a given vehicle density \(\rho\) is relative to an infinite multitude of different vehicle speeds (a finite speed range,
Fig. 4.4. Qualitative representation of the two-dimensional region of steady states of synchronized flow in the flow density plane (dashed region). (a) Multilane (one-way) road. (b) Single-lane road. The curve \( F \) represents states of free flow. Taken from [205]

Fig. 4.5. Explanation of the fundamental hypothesis of the three-phase traffic theory. A hypothetical model steady state in which vehicle speed \( v \) is related to an infinite multitude of vehicle densities within the density range \([\rho_1, \rho_2]\), which is given by the intersection of the dotted line “slope \( v \)” with the two-dimensional region of steady states of synchronized flow. A hypothetical steady state with vehicle density \( \rho \) is related to an infinite multitude of vehicle speeds within the speed range \([v_1, v_2]\), which is given by the intersection of the dotted line \( \rho \) with boundaries of the two-dimensional region of steady states of synchronized flow. The states of synchronized and free flows are the same as those in Fig. 4.4a

\([v_1, v_2]\) in Fig. 4.5). This hypothesis also means that in the three-phase traffic theory there is no fundamental diagram for hypothetical steady states of synchronized flow.
When the hypothesis regarding the theoretical fundamental diagram was discussed (Sect. 3.1), it was mentioned that in congested traffic, characteristics of spatially inhomogeneous and time-dependent patterns can usually be measured, unlike hypothetical steady states of traffic flow. Thus, empirical observations in which a broad spread of vehicle gaps at the same speed is observed (e.g., [88]; see empirical points of congested traffic in Fig. 2.4a) can only be a hint (but obviously not a proof) of the fundamental hypothesis regarding a 2D region of steady states of synchronized flow in three-phase traffic theory. As a proof of the three-phase traffic theory, its mathematical results can be considered [329–331] (Part III). These mathematical results enable us to overcome the indicated problems of the fundamental diagram approach for a correct description of empirical spatiotemporal features of congested traffic [208, 218].

As recently postulated on general grounds [218] and demonstrated for a microscopic traffic model [329], whether the steady states of a mathematical description of traffic belong to a curve or a 2D-region in the flow–density plane can qualitatively change the basic nonlinear spatiotemporal features of the congested patterns that the model allows. The same conclusion follows from a recent cellular automata approach to the three-phase traffic theory [331].

Of course, in models with a fundamental diagram of steady states, fluctuations and external perturbations let the system evolve in time through a 2D region in the flow–density plane as well. This 2D region is related to certain “dynamic” model solutions, which can also be steady solutions in the form of different steady moving jams. Examples of the 2D region in the flow–density plane for the “dynamic” model states are Nagel-Schreckenberg CA models (e.g., [426]), the models of Tomer, Havlin et al. [430], Helbing, Treiber et al. [422, 431], Nishinari and Takahashi [442, 443], Nishinari [445], Fukui et al. [404] (see other references in the reviews [33, 35, 36, 38]). However, in these traffic flow models, steady model states are related to a one-dimensional region in the flow–density plane, i.e., to the fundamental diagram.

The dynamics of a traffic flow model is often governed locally by steady-state properties. If the steady states form a 2D region in the flow–density plane and have properties postulated in the three-phase traffic theory [205, 208–210], the dynamics is fundamentally different [329–331] from the models within the scope of the fundamental diagram approach. This also leads to qualitative differences between congested patterns obtained in the three-phase traffic theory and the fundamental diagram approach as shown in [329, 331]. Specifically, at a high flow rate to the on-ramp, instead of homogeneous congested traffic (HCT) without moving jams [35, 393], we find that moving jams always emerge spontaneously in synchronized flow at low vehicle speed and high density. Furthermore, at low flow rate to the on-ramp and high flow rate on the freeway instead of the moving jams found in [35, 36, 381, 393, 395], synchronized flow at higher vehicle speed can exist for a long time without the occurrence of moving jams [218, 329]. This agrees with the empirical observations [218, 221, 222].
For further analysis it is useful to consider the fundamental hypotheses of three-phase traffic theory in the speed–density plane (Fig. 4.6).

In the speed–density plane (Fig. 4.6b), as follows from the fundamental hypotheses of the three-phase traffic theory, steady states of synchronized flow also cover a two-dimensional region. This has a different form in comparison to steady states of synchronized flow in the flow–density plane (Fig. 4.6a).

Fig. 4.6. Symbolic illustration of the states of free flow (curve \( F \)) and 2D steady states of synchronized flow (dashed region) in the speed–density plane (b) that correspond to the related states in the flow–density plane (a). In (a) states of free flow and synchronized flow are taken from Fig. 4.4a

### 4.3.1 Three-Phase Traffic Theory as Driver Behavioral Theory

Like other traffic theories, three-phase traffic theory is a *driver behavioral theory*. This means that hypotheses of this theory are based on certain observed...
common behavioral characteristics of drivers in traffic flow. In particular, the
fundamental hypothesis of three-phase traffic theory is related to a driver’s
ability to recognize whether the space gap to the preceding vehicle is increas­
ing or decreasing [205]. If space gaps between vehicles are not too large, the
driver’s ability is unchanged even if the difference between the vehicle speed
and the speed of the preceding vehicle is negligible.

In synchronized flow, the mean space gap between vehicles is relatively
small (i.e., the density is relatively high) in comparison with free flow at
the same flow rate (Fig. 4.4). In other words, drivers are able to maintain a
time-independent space gap (without taking fluctuations into account) to the
preceding vehicle in an initially steady state of synchronized flow. The ability
of drivers to maintain a time-independent space gap should be valid for a finite
range of space gaps. Therefore, a given steady speed in synchronized flow can
be related to an infinite multitude of steady states with different densities in
a limited range (Fig. 4.5). For this reason, the multitude of steady states of
synchronized flow covers a two-dimensional region in the flow–density plane
(hatched region in Fig. 4.4).

The upper boundary $S_{upper}$ in the two-dimensional region of steady states
of synchronized flow in the flow–density plane (Fig. 4.6) can be determined in
terms of the minimal possible time gap between vehicles, which is accepted by
drivers in synchronized flow. This time gap is dictated by security conditions
to avoid a collision between vehicles.

States of free flow (curve $F$ in Fig. 4.4a) and steady states of synchronized
flow (hatched region) overlap in density. However, in free flow on a multilane
road (in one direction), due to the possibility of passing, the average vehi­
cle speed can be higher than the maximum speed in synchronized flow at the
same density. Therefore, there is a gap in the flow rate between states of free
flow and synchronized flow at a given density (Fig. 4.4a).

Note that on a single-lane road vehicles cannot pass regardless of the vehi­
cle density. Therefore, steady states of flow on the single-lane road at higher
densities are identical to steady states of synchronized flow on a multilane
road (hatched region in Fig. 4.4b).

In some traffic flow theories, both driver behavioral characteristics and
empirical results are considered as the basis of these theories. However, there
are many driver behavioral characteristics, which are in reality only theore­
tical hypotheses rather than empirical evidence. This is because it is difficult
to determine reproducible empirical features of many driver behavioral char­
acteristics from traffic measurements. In contrast, spatiotemporal congested
pattern features discussed in Sect. 2.4 are reproducible empirical features of
traffic, which are observed in data measured on many different days and years
on different freeways. For this reason, rather than driver behavioral char­
acteristics these reproducible empirical spatiotemporal pattern features are the
empirical basis of the three-phase traffic theory. However, possible driver be­
behavioral characteristics, which can explain the hypotheses of the three-phase
traffic theory, will be discussed when the three-phase traffic theory is considered. A summary of main driver behavioral assumptions in the three-phase traffic theory will be made at the end of the consideration of the three-phase traffic theory (Sect. 8.6).

4.3.2 Synchronization Distance and Speed Adaptation Effect in Synchronized Flow

The lower boundary \( S_{\text{low}} \) in the two-dimensional region of steady states of synchronized flow in the flow–density plane (Fig. 4.6) is determined by a “synchronization distance.”

The synchronization distance is caused by the speed adaptation effect in synchronized flow: if the distance of a vehicle from the preceding vehicle is greater than the synchronization distance, the vehicle simply accelerates if vehicle speed is lower than a desired vehicle speed. However, if the vehicle cannot pass the preceding vehicle, then within the synchronization distance the vehicle tends to adjust its speed to the preceding vehicle, i.e., the vehicle decelerates if it is faster, and accelerates if the vehicle becomes slower than the preceding vehicle.\(^3\)

Within the synchronization distance a driver tends to adapt the vehicle speed to the speed of the preceding vehicle without caring, what the precise distance to the preceding vehicle is, as long as this distance is greater than a safe one. This driver behavior is responsible for an infinite number of possible various distances between vehicles at the same vehicle speed in synchronized flow. We can see that the speed adaptation effect is related to the fundamental hypothesis of the three-phase traffic theory.

4.3.3 Random Transformations (“Wandering”) Within Synchronized Flow States

Small enough space gap perturbations related to some initial fluctuation in the braking of a vehicle should not grow in steady states of synchronized flow. Indeed, small enough changes in space gap are allowed, so drivers should not immediately react to. For this reason, even after a time delay due to the finite reaction time of drivers, drivers upstream should not brake harder than drivers in front of them to avoid an accident. As a result, a local perturbation of traffic variables (e.g., flow rate, density, or vehicle speed) with small amplitude does not grow.

\(^3\) Instead of the synchronization distance, a “synchronization time gap” can be used for the definition of the lower boundary \( S_{\text{low}} \). In this case, if the time gap between a vehicle and the preceding vehicle is higher than the synchronization time gap, the vehicle accelerates. However, if the vehicle cannot pass the preceding vehicle, then within the synchronization time gap the vehicle tends to adjust its speed to the preceding vehicle, i.e., it decelerates if it is faster, and accelerates if it becomes slower than the preceding vehicle.
An occurrence of a small amplitude perturbation associated with vehicle deceleration (or vehicle acceleration) can cause a spatiotemporal transition to another state of synchronized flow. In other words, local perturbations can cause continuous spatiotemporal transitions between different states of synchronized flow [205]. These random transformations ("wandering") within synchronized flow states can lead to complex dynamic synchronized flow states, which cover a 2D region in the flow–density plane. This 2D region can be very close to the 2D region of steady states of synchronized flow discussed above (Fig. 4.4).

Thus, steady states of synchronized flow are related to a hypothetical unperturbed and noiseless vehicle motion that does not occur in reality. Already small amplitude random perturbations in synchronized flow destroy steady states of synchronized flow. This means that after an F→S transition has occurred in an initial free flow rather than steady states some dynamic spatiotemporal synchronized flow states appear.

4.3.4 Dynamic Synchronized Flow States

Such dynamic spatiotemporal states of synchronized flow that can be very complex can also occur due to different dynamic effects. However, these complex dynamic synchronized flow states can be close to steady states of synchronized flow. This is because there is a competition of dynamic effects destroying steady states with the speed adaptation effect in synchronized flow (Sect. 4.3.2). This speed adaptation effect attracts vehicles to a region of small speed differences in synchronized flow (see Sect. 18.2.2). The related spatiotemporal dynamic synchronized flow states should exhibit features of steady states of synchronized flow discussed above, and which will be considered in other hypotheses of the three-phase traffic theory below in Part I of this book. In other words, for such dynamic synchronized flow states we can expect to find qualitatively the same features of phase transitions in traffic flow and spatiotemporal congested patterns as those postulated in the three-phase traffic theory (Chaps. 5–8).

This is confirmed by numerical simulations of the microscopic models of [329–331]. In these models, after an F→S transition in an initial free flow has occurred, rather than steady states of synchronized flow some dynamic synchronized flow states appear. The latter states exhibit the main features of steady states with respect to phase transitions and congested pattern formation. As a result, these models reproduce features of phase transitions in traffic flow and congested patterns in accordance with empirical results (Part III).

4 Recall that the three-phase traffic theory is a qualitative theory. For this reason, the fundamental hypothesis of this theory is associated with features of hypothetical steady states of traffic flow. Rather than some hypothesis about model
4.4 Empirical Basis of Three-Phase Traffic Theory

Based on a consideration of the main empirical features of spatiotemporal congested patterns in Sect. 2.4 and on empirical results considered above, we can formulate the empirical basis for the three-phase traffic theory as follows [207,208]:

1. There are three traffic phases: the well-known “free flow” phase and two phases in congested traffic – synchronized flow and wide moving jam. These phases are distinguished by qualitatively different spatiotemporal features corresponding to the objective criteria \([S]\) and \([J]\) of Sect. 4.2.1.

2. A wide moving jam propagates upstream through any traffic state and through any bottleneck while maintaining the mean velocity of the downstream jam front \(v_g\).

3. In the flow–density plane, empirical data measured for free flow is cut off at a limit (maximum) point for free flow, where the flow rate and density have maximum values and the vehicle speed has a minimum value for free flow. When traffic demand is high enough, the maximum empirical flow rate \(q_{\text{max, emp}}\) at the limit (maximum) point of free flow is higher than the flow rate \(q_{\text{out}}\) in the jam outflow:

steady states, the fundamental hypothesis of a mathematical traffic flow model can be a hypothesis about some dynamic model features.

In other words, in contrast to the three-phase traffic theory, which is a qualitative theory, the fundamental hypothesis of a mathematical traffic flow model based on this three-phase traffic theory should not necessarily be a hypothesis about hypothetical steady states of synchronized flow. Indeed, in the mathematical model, the fundamental hypothesis can be, for example, a hypothesis about features of dynamic synchronized flow model states. When these dynamic synchronized flow model states qualitatively exhibit features of steady states in the three-phase traffic theory with respect to phase transitions and pattern formation (Chaps. 5–8), we can expect that the model can explain some empirical congested traffic patterns. The latter can also be true even when steady-state solutions of this model cover a one-dimensional region(s) in the flow–density plane.

It can also be assumed that in mathematical traffic flow models, rather than the fundamental hypothesis of the three-phase traffic theory, some other important consequences of this fundamental hypothesis can be used. One of these consequences can be the hypothesis of the three-phase traffic theory about a double Z-characteristic for \(F\rightarrow S\rightarrow J\) transitions in traffic flow (Chap. 6), which is used in a mathematical model. Also in this case, we can expect that the mathematical model can explain some empirical features of phase transitions in traffic flow.

In general, we can expect that the more precisely hypotheses of the three-phase traffic theory discussed in this chapter and Chaps. 5–8 are used in a mathematical traffic flow model, the more precisely the model explains and predicts empirical features of phase transitions and spatiotemporal congested patterns.
\[ q_{\text{max}}^{(\text{free, emp})} > q_{\text{out}} \]  

(4.2)

(4) Empirical points related to synchronized flow can lie above the line \( J \) (Fig. 4.2b).

(5) The onset of congestion (breakdown phenomenon) is associated with a local first-order \( F \rightarrow S \) transition.

(6) Independent of the initial density of free flow, no spontaneous \( F \rightarrow J \) transition is observed, i.e., wide moving jams do not emerge spontaneously in free flow.

(7) Wide moving jams can occur spontaneously in synchronized flow (\( S \rightarrow J \) transition) rather than due to an \( F \rightarrow J \) transition.

(8) The lower the vehicle speed in a dense synchronized flow, the higher the frequency of wide moving jam emergence in the synchronized flow.

(9) In a state of synchronized flow at a higher vehicle speed, a wide moving jam will not necessarily emerge.

(10) Wide moving jams emerge in free flow due to \( F \rightarrow S \rightarrow J \) transitions:

(a) first an \( F \rightarrow S \) transition is observed, i.e., synchronized flow appears;

(b) only then and as a rule at another freeway location is an \( S \rightarrow J \) transition observed in that synchronized flow.

A more detailed discussion of this empirical basis for three-phase traffic theory can be found in Part II of this book.

4.5 Conclusions

(i) Simultaneous measurements of traffic in both space and time, together with an understanding of effectual freeway bottlenecks are necessary to properly distinguish the “wide moving jam” phase from the “synchronized flow” phase in congested traffic. The objective (empirical) criteria for the traffic phases are based on the propagation of the downstream front of these traffic phases (within the downstream front vehicles accelerate from lower speeds in the associated traffic phase to higher speeds in the other traffic phase downstream). The downstream front of a wide moving jam propagates through any other traffic states and any bottlenecks while maintaining the velocity of the downstream jam front. In contrast, the downstream front of synchronized flow does not possess this characteristic feature; in particular, this synchronized flow front is usually fixed at a bottleneck.

(ii) Three-phase traffic theory is based on empirical features of phase transitions among the three traffic phases, as well as on empirical features of spatiotemporal congested patterns.

(iii) In three-phase traffic theory, it is suggested that hypothetical steady states of synchronized flow cover a two-dimensional region in the flow–density plane.
5 Breakdown Phenomenon (F→S Transition) in Three-Phase Traffic Theory

5.1 Introduction

In this chapter, based on the fundamental hypothesis and the empirical basis of three-phase traffic theory (Chap. 4), we start to consider a qualitative theory of phase transitions in traffic flow and congested traffic patterns [205, 208–212, 218].

We begin with a discussion of an F→S transition, i.e., the phase transition from the “free flow” phase to the “synchronized flow” phase. In three-phase traffic theory, the F→S transition explains the well-known breakdown phenomenon (speed breakdown) in an initial free flow. Thus, in this book, the term “F→S transition” is a synonym for the terms “breakdown phenomenon” and “speed breakdown” in free flow.

The F→S transition is observed mostly at freeway bottlenecks. However, we will first analyze the F→S transition on a homogeneous road, i.e., a road without any bottlenecks. The aim of the analysis of the hypothetical case of the homogeneous road is as follows.

We would like to show that an F→S transition can occur spontaneously on a road without any bottlenecks. In other words, the F→S transition is fundamentally associated with intrinsic features of the dynamic process “traffic.” It turns out that the intrinsic traffic features responsible for an F→S transition on a homogeneous road are also responsible for the occurrence of an F→S transition at a freeway bottleneck.

However, there is an important difference between these two cases. When the flow rate is high enough, a bottleneck introduces some permanent (deterministic) local disturbance into free flow in the vicinity of the bottleneck. This disturbance does not move and it is localized at the bottleneck. The disturbance leads to a permanent local decrease in the average vehicle speed and to the related local increase in the vehicle density at the bottleneck. This speed decrease and the density increase in the vicinity of the bottleneck will be called a deterministic local perturbation at the bottleneck.

The deterministic local perturbation leads to an F→S transition at the bottleneck at considerably lower vehicle density in free flow downstream of the bottleneck than the density at which an F→S transition occurs on a homogeneous road. In other words, on real freeways with bottlenecks at the same flow rate in free flow, due to a deterministic local perturbation, the probability
of the breakdown phenomenon (F→S transition) at the bottleneck, $P_{FS}^{(B)}$, is much higher than the probability of the breakdown phenomenon away from bottlenecks, $P_{FS}$:

$$P_{FS}^{(B)} \gg P_{FS} .$$

This can explain why the breakdown phenomenon (F→S transition) occurs in empirical observations mostly at freeway bottlenecks.

5.2 Breakdown Phenomenon on Homogeneous Road

5.2.1 Speed Breakdown at Limit Point of Free Flow

Let us discuss the breakdown phenomenon on a homogeneous road in free flow. We denote the flow rate, density, and speed in free flow by $q^{(\text{free})}$, $\rho^{(\text{free})}$, and $v^{(\text{free})}$, respectively. We have already mentioned that the branch for free flow in the flow–density plane is cut off at some limit (maximum) point $(\rho^{(\text{free})}_{\text{max}}, q^{(\text{free})}_{\text{max}})$. At vehicle densities higher than the limit density $\rho^{(\text{free})}_{\text{max}}$, free flow does not exist. This means that congested traffic should occur. In three-phase traffic theory, there are two traffic phases in congested traffic, synchronized flow and wide moving jam.

However, corresponding to the point (6) of the empirical basis for three-phase traffic theory (Sect. 4.4), wide moving jams do not emerge spontaneously in free flow. Thus, at a vehicle density in free flow higher than the limit density $\rho^{(\text{free})}_{\text{max}}$, synchronized flow must occur, i.e., an F→S transition is realized. In other words, the limit point of free flow is also the critical point of free flow $(\rho^{(\text{free})}_{\text{cr}}, q^{(\text{free})}_{\text{cr}})$, at which the F→S transition must occur:

$$\rho^{(\text{free})}_{\text{cr}} = \rho^{(\text{free})}_{\text{max}} \quad (q^{(\text{free})}_{\text{cr}} = q^{(\text{free})}_{\text{max}} (\rho^{(\text{free})}_{\text{max}})).$$

The critical speed associated with this critical point of free flow is

$$v^{(\text{free})}_{\text{cr}} = v^{(\text{free})}_{\text{min}} = q^{(\text{free})}_{\text{max}} / \rho^{(\text{free})}_{\text{max}} .$$

---

1 In empirical observations of free flow, the breakdown phenomenon occurs most frequently at bottlenecks. As a result, when characteristics of free flow away from bottlenecks are measured and traffic demand is high enough, the empirical limit (maximum) point of free flow $(\rho^{(\text{free, emp})}_{\text{max}}, q^{(\text{free, emp})}_{\text{max}})$ is usually associated with the flow rate limitation caused by congestion at a freeway bottleneck upstream of where the free flow is studied. Thus, the empirical limit point of free flow away from bottlenecks is usually not related to the theoretical critical point of free flow $(\rho^{(\text{free})}_{\text{max}}, q^{(\text{free})}_{\text{max}})$. For the theoretical analysis under consideration this is not important. However, we should not forget that the limit (maximum) point of free flow $(\rho^{(\text{free})}_{\text{max}}, q^{(\text{free})}_{\text{max}})$ for a homogeneous road is only a hypothetical limit point of free flow, which cannot usually be determined in empirical observations.
Let us consider states of free flow and steady states of synchronized flow in Fig. 4.6 that follow from the fundamental hypothesis of three-phase traffic theory. We can see from Fig. 4.6b that after the F→S transition has occurred at the limit point of free flow, the vehicle speed should decrease in the emergent synchronized flow in comparison with the speed in free flow. This speed breakdown is symbolically shown in Fig. 5.1 by the dotted arrow from the critical point \( K \) given by (5.2) to a point \( M \) inside the 2D-region of states of synchronized flow.

**Fig. 5.1.** Qualitative illustration of the breakdown phenomenon (F→S transition) at the limit point of free flow (dotted arrow \( K \rightarrow M \)) and a possible reverse transition (S→F transition, dotted arrow \( N \rightarrow P \)) on a homogeneous road (away from freeway bottlenecks). (a) States of traffic in the flow–density plane. (b) States of traffic in the speed–density plane. States of free flow and synchronized flow are taken from Fig. 4.6

**Hysteresis Effect. First-Order Phase Transition**

After an F→S transition occurs, we first have a local region of synchronized flow associated with the point \( M \), surrounded by the initial free flow. Later
there can be a “wandering” of the speed and density inside the 2D states of synchronized flow, e.g., from the initial point \( M \) to another point \( N \) with a lower density (Sect. 4.3.3).

This can lead to a reverse \( S \rightarrow F \) transition from the point \( N \) to a point \( P \) in free flow states (dotted up-arrow \( N \rightarrow P \) in Fig. 5.1). This is one possible hysteresis effect, and the related hysteresis loop \( K \rightarrow M \rightarrow N \rightarrow P \) reflects the empirical point (5) of Sect. 4.4 that the \( F \rightarrow S \) transition is a \textit{first-order} phase transition (see definition of “first-order phase transition” in Appendix A).

\section*{5.2.2 Critical Local Perturbation for Speed Breakdown}

In the foregoing discussion, we assumed that after the critical point of free flow (5.2) has been reached, an \( F \rightarrow S \) transition occurs. This is correct if there are no random fluctuations (perturbations) in free flow. In real free flow, there are always random local perturbations caused by random flow disturbances, such as an unexpected breaking of a vehicle. These random perturbations can lead to an \( F \rightarrow S \) transition at the free-flow density \( \rho^{(\text{free})} \), which is lower than the critical density \( \rho^{(\text{free})}_{\text{max}} \) (5.2).

However, if the density in free flow

\[
\rho^{(\text{free})} < \rho^{(\text{free})}_{\text{max}} \quad \left( q^{(\text{free})} < q^{(\text{free})}_{\text{max}} \right),
\]

then a random local perturbation must satisfy certain requirements to cause an \( F \rightarrow S \) transition. These requirements are associated with the amplitude and spatial form of the perturbation. Note that the amplitude of a local perturbation is the difference between the value of a traffic variable (density or speed) within the perturbation and in homogeneous free flow.

The requirement on the amplitude of the local perturbation is as follows. Only if the amplitude of the local perturbation exceeds some \textit{critical amplitude} of the local perturbation will the \( F \rightarrow S \) transition occur under the condition (5.4). A local perturbation with the critical amplitude is called a \textit{critical local perturbation} for the \( F \rightarrow S \) transition (breakdown phenomenon) in an initial free flow.

Thus, if the amplitude of an initial local perturbation that occurs randomly in the free flow exceeds the critical amplitude of the critical local perturbation, the initial local perturbation will grow and lead to an \( F \rightarrow S \) transition. Otherwise, the initial local perturbation will decay.\footnote{Strictly speaking, the \textit{critical} local perturbation is the local perturbation that has both the \textit{critical spatial form} and the \textit{critical amplitude}. However, for simplicity of the further qualitative consideration we will neglect the feature of the critical form of the critical local perturbation in the three-phase traffic theory.}

Corresponding to the above definition, the critical amplitude of the local perturbation for the \( F \rightarrow S \) transition is
\[
\Delta v_{\text{cr}}^{(FS)} = \nu^{(\text{free})} - v_{\text{cr}}^{(FS)},
\]
where \(\nu^{(\text{free})}\) is the speed in an initial free flow and \(v_{\text{cr}}^{(FS)}\) is the vehicle speed within the critical local perturbation.

Let us assume that in an initial free flow a local perturbation randomly occurs whose amplitude exceeds the critical amplitude (5.5). This means that the speed within the initial perturbation, \(\nu_{\text{pert}}^{(\text{initial})}\), is lower than the speed \(v_{\text{cr}}^{(FS)}\) within the critical local perturbation (Fig. 5.2a). Then this random local perturbation must begin to grow, leading to a local region of the “synchronized flow” phase within the initial “free flow” phase. The random perturbations that grow into an F→S transition initially result from internal effects in traffic flow (deceleration of one of the vehicles, vehicle lane changing, and so on). For this reason, this F→S transition is a spontaneous F→S transition.

\[
\nu^{(\text{free})} = \nu^{(\text{free})} \quad < \quad v_{\text{cr}}^{(FS)} < \quad \nu_{\text{pert}}^{(\text{pert})} \quad < \quad \nu_{\text{initial}}^{(pert)}
\]

Fig. 5.2. Qualitative illustration of local perturbations of vehicle speed in an initial free flow where the speed \(v = \nu^{(\text{free})}\). (a) Growing local perturbation whose amplitude exceeds the critical amplitude \(\Delta v_{\text{cr}}^{(FS)}\) (5.5). (b) Local perturbation that decays because the perturbation amplitude is less than the critical amplitude. The arrows symbolically illustrate the growth (a) or decay (b) of the initial local perturbation.

In this case, we have spontaneous local speed breakdown from a state of free flow with a lower flow rate and a lower density than the flow rate and density at the critical point (5.2). This F→S transition is symbolically shown by dotted arrow \(K' \rightarrow M'\) in Fig. 5.1a. The point \(K'\) lies left of the limit point \(K\) (5.2) in the states for free flow \(F\).

In contrast, when the amplitude of a random local perturbation in the initial free flow is less than the critical amplitude \(\Delta v_{\text{cr}}^{(FS)}\) (5.5), the local perturbation decays over time (Fig. 5.2b). Thus, in this case, no local region is formed of synchronized flow in the initial free flow, i.e., speed breakdown does not occur.

There should be a critical point \((\rho_{\text{max}}^{(\text{free})}, q_{\text{max}}^{(\text{free})})\) (5.2) at which the amplitude of the critical local perturbation \(\Delta v_{\text{cr}}^{(FS)}\) (5.5) vanishes:

\[
\Delta v_{\text{cr}}^{(FS)} \bigg|_{\rho_{\text{max}}^{(\text{free})}} = \nu^{(\text{free})} - v_{\text{cr}}^{(FS)} = 0, \quad (5.6)
\]
where $v^{(\text{free})} = v^{(\text{free})}_\text{min}$. In this case, an $F \rightarrow S$ transition must occur at the critical point for free flow, as there are always random perturbations in free flow whose amplitude exceeds zero. These perturbations grow in the limiting free flow, leading to an $F \rightarrow S$ transition. This means that at this critical point of free flow the probability $P_{FS}$ for the $F \rightarrow S$ transition is

$$P_{FS} \mid \rho = \rho^{(\text{free})}_\text{max} = 1 .$$  \hspace{1cm} (5.7)

The critical density $\rho^{(\text{free})}_\text{max}$ (critical flow rate $q^{(\text{free})}_\text{max}$) is found from the condition (5.7). Because the critical point of free flow is determined by calculating the probability $P_{FS}$ for the $F \rightarrow S$ transition, we consider the definition of this probability below.

### 5.2.3 Probability for Breakdown Phenomenon

The probability $P_{FS}$ for an $F \rightarrow S$ transition is defined as follows. We consider a large number of realizations, $N_{FS}$, where the $F \rightarrow S$ transition in an initial free flow is studied. Each realization should refer to the same flow rate of free flow, other initial conditions, and the same time interval $T_{ob}$ for observing the spontaneous $F \rightarrow S$ transition on a chosen road section of length $L_{ob}$. Let us assume that the $F \rightarrow S$ transition occurs in $n_{FS}$ of these $N_{FS}$ realizations. Then the probability for this transition in time interval $T_{ob}$ and length of road $L_{ob}$ is

$$P_{FS} = \frac{n_{FS}}{N_{FS}} .$$  \hspace{1cm} (5.8)

When the density in free flow is lower than the critical density, i.e., when the condition (5.4) is satisfied, then the critical amplitude of the local perturbation is higher than zero:

$$\Delta v^{(\text{FS})}_\text{cr} \bigg|_{\rho^{(\text{free})} < \rho^{(\text{free})}_\text{max}} = v^{(\text{free})} - v^{(\text{FS})}_\text{cr} > 0 .$$  \hspace{1cm} (5.9)

The lower the density in free flow in comparison with the critical density (5.2), the higher should be the critical amplitude of the local perturbation $\Delta v^{(\text{FS})}_\text{cr}$. However, the higher the amplitude of a local perturbation should occur, the lower the probability of this perturbation occurrence. This means that the probability for the $F \rightarrow S$ transition $P_{FS}$ (5.8) should decrease with the decrease in the vehicle density in free flow.

---

3 Both the probability $P_{FS}$ for a spontaneous $F \rightarrow S$ transition and the critical density (critical flow rate) (5.2) depend on the time interval $T_{ob}$ for observing free flow. In empirical observations of free flow away from bottlenecks on a given road section of the length $L_{ob}$, the flow rate and density in free flow are not usually constant in time. In this case, the probability $P_{FS}$ and the critical traffic variables at the critical point of free flow (5.2) depend on the averaging time $T_{av}$ for the traffic variables.

4 Strictly speaking, the exact probability for a spontaneous $F \rightarrow S$ transition is the limit as $N_{FS} \rightarrow \infty$. 

5.2.4 Threshold Flow Rate and Density, Metastability, and Nucleation Effects

When the density in free flow further decreases, there will be a characteristic density at which the critical amplitude of the local perturbation $\Delta v_{\text{cr}}^{(FS)}$ reaches its maximum. This characteristic density in free flow is called the threshold density for an $F \rightarrow S$ transition, $\rho_{\text{th}}$. The flow rate associated with this density is called the threshold flow rate for the $F \rightarrow S$ transition, $q_{\text{th}}$.

The threshold point for an $F \rightarrow S$ transition in free flow

$$
\rho^{(\text{free})} = \rho_{\text{th}} \quad \left( q^{(\text{free})} = q_{\text{th}} \right)
$$

is defined as follows. Consider a free flow in which the density $\rho^{(\text{free})}$ is lower than the critical density $\rho_{\text{max}}^{(\text{free})}$; an $F \rightarrow S$ transition can still occur in this free flow. If the density then further decreases the threshold density $\rho_{\text{th}}$ is the minimum density in free flow at which an $F \rightarrow S$ transition can still occur. If the density $\rho^{(\text{free})}$ is lower than the threshold density $\rho_{\text{th}}$, no $F \rightarrow S$ transition can occur in this free flow. The latter is true regardless of the amplitude of a time-limited local perturbation in the free flow. This means that when

$$
\rho^{(\text{free})} < \rho_{\text{th}} \quad \left( q^{(\text{free})} < q_{\text{th}} \right),
$$

the probability $P_{FS}$ for the $F \rightarrow S$ transition satisfies the condition

$$
P_{FS} \mid \rho^{(\text{free})} < \rho_{\text{th}} = 0.
$$

Corresponding to the above definitions of the critical point (5.2) and threshold point (5.10) for an $F \rightarrow S$ transition, within the free-flow density range

$$
\rho_{\text{th}} \leq \rho^{(\text{free})} < \rho_{\text{max}}^{(\text{free})} \quad \left( q_{\text{th}} \leq q^{(\text{free})} < q_{\text{max}}^{(\text{free})} \right)
$$

the mean probability of an $F \rightarrow S$ transition is

$$
P_{FS} < 1.
$$

Assume now that the density in free flow satisfies (5.13). If the density then decreases and approaches the threshold point, the critical amplitude of the local perturbation for an $F \rightarrow S$ transition reaches a maximum:

$$
\Delta v_{\text{cr}}^{(FS)} \mid \rho^{(\text{free})} = \rho_{\text{th}} = \Delta v_{\text{cr}, \text{max}}^{(FS)}.
$$

The related dependencies of the critical amplitude of the local perturbation and of the probability $P_{FS}$ for an $F \rightarrow S$ transition on the density are shown by the curves $F_S$ and $P_{FS}$ in Fig. 5.3b,c, respectively.

Within the free-flow density range (5.13) (Fig. 5.4), spontaneous nucleation can take place, and the free flow can be metastable against an $F \rightarrow S$ transition.
Fig. 5.3. Qualitative illustration of the nucleation and metastability effects in free flow due to an F→S transition. (a) States of free and synchronized flow taken from Fig. 4.4a. (b) Critical amplitude $\Delta v_{cr}^{(FS)}$ of local perturbation for the F→S transition (5.5) as a function of density (curve $F_S$). (c) Probability $P_{FS}$ for the F→S transition as a function of density (curve $P_{FS}$)
Fig. 5.4. Qualitative illustration of the metastability of free flow with respect to an F→S transition (breakdown phenomenon) in the flow–density plane (a) and in the speed–density plane (b, c). In (b, c) the critical branch $v_{cr}^{(FS)}(\rho^{(free)})$ (dashed curve) gives the speed within the critical local perturbation. In (a) and (b) states of free and synchronized flows are taken from Fig. 4.6. In (c) a part of figure (b) for lower density range is shown in a larger scale. The black point in (a)–(c) on the curve $F$ for free flow shows the threshold point (5.10). In (a) the critical branch $q_{cr}^{(FS)}(\rho^{(free)}) = \rho^{(free)}v_{cr}^{(FS)}(\rho^{(free)})$
transformation. Nucleation in F→S transitions means that a local perturbation in free flow whose amplitude exceeds the critical amplitude of the local perturbation can play the role of a nucleus for an F→S transition. The necessity of a finite amplitude local perturbation in free flow whose amplitude must exceed the critical amplitude (5.5) for the occurrence of an F→S transition is called the *metastability effect* with respect to F→S transitions. A state of free flow where the metastability effect can occur is called a *metastable state* of free flow with respect to the F→S transition.

### 5.2.5 Z-Shaped Speed–Density and Passing Probability Characteristics

The curve $F$ for states of free flow, the “critical branch” $v_{cr}^{(FS)}(\rho^{\text{free}})$ for the F→S transition, and the 2D region for steady states of synchronized flow form together a Z-shaped function of speed in terms of density (Fig. 5.4b,c) [221]. To explain this Z-characteristic, first let us discuss the definition of the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ for the F→S transition and its dependence on density. Each point on the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ is the speed within the critical local perturbation for an F→S transition. The dependence of this speed on density is given by the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$. At a given density $\rho^{\text{free}}$ the difference between the vehicle speed $v_{cr}^{(FS)}(\rho^{\text{free}})$ within the critical local perturbation and the initial speed in free flow $v_{cr}^{(FS)}$ is the critical amplitude of the local perturbation $\Delta v_{cr}^{(FS)}$ (5.5) for an F→S transition.

To understand the dependence $v_{cr}^{(FS)}(\rho^{\text{free}})$, let us first consider the critical density (5.2), $\rho^{\text{free}} = \rho_{\text{max}}$. At this density the speed is

$$v^{\text{free}}(\rho_{\text{max}}) = v_{\text{min}}^{\text{free}}.$$  

In accordance with (5.6), at the critical density the speed $v_{cr}^{(FS)}$ on the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ is

$$v_{\text{cr}}^{(FS)}(\rho_{\text{max}}) = v_{\text{min}}^{\text{free}}.$$  

From (5.16) and (5.17) it follows that the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ in Fig. 5.4b merges with the states of free flow $F$ at the critical point. The critical amplitude of the local perturbation $\Delta v_{cr}^{(FS)}$ (5.5) increases when the density in free flow decreases (Sect. 5.2.3). As follows from (5.15), at the threshold density (5.10) the critical amplitude $\Delta v_{cr}^{(FS)} = v^{\text{free}} - v_{cr}^{(FS)}$ has a maximum (5.15). This is reflected in the dependence of the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ in Fig. 5.4b,c.

The Z-characteristic in Fig. 5.4b,c explains an F→S transition as follows. Let us consider a behavior of a local random decrease in an initial homogeneous speed $v^{(\text{free})}$. In the free-flow density range (5.13), there can be two different cases of a development of this random speed perturbation.
(1) The speed within this local perturbation $v_{\text{initial}}^{(\text{pert})}$ satisfies the condition

$$v_{\text{initial}}^{(\text{pert})} < v_{\text{cr}}^{(\text{FS})}(\rho^{(\text{free})}).$$  

This means that in the speed–density plane the speed within the perturbation is related to a point below the critical branch $v_{\text{cr}}^{(\text{FS})}(\rho^{(\text{free})})$ in Fig. 5.4c. In this case, the amplitude of this perturbation

$$\Delta v_{\text{initial}}^{(\text{pert})} = v^{(\text{free})} - v_{\text{initial}}^{(\text{pert})}$$  

is higher than the critical amplitude (5.5) for the F─S transition:

$$v^{(\text{free})} - v_{\text{initial}}^{(\text{pert})} > \Delta v_{\text{cr}}^{(\text{FS})}.$$  

Thus, the initial perturbation should grow and lead to an F─S transition (down-arrow in Fig. 5.2a). As a result, a local region of synchronized flow appears (down-arrow in Fig. 5.4c).

(2) The speed within this local perturbation $v_{\text{initial}}^{(\text{pert})}$ satisfies the condition

$$v_{\text{initial}}^{(\text{pert})} > v_{\text{cr}}^{(\text{FS})}(\rho^{(\text{free})}).$$  

This means that in the speed–density plane the speed within the perturbation is related to a point above the critical branch $v_{\text{cr}}^{(\text{FS})}(\rho^{(\text{free})})$ in Fig. 5.4c. In this case, the amplitude of this perturbation (5.19) is lower than the critical amplitude (5.5) for an F─S transition:

$$v^{(\text{free})} - v_{\text{initial}}^{(\text{pert})} < \Delta v_{\text{cr}}^{(\text{FS})}.$$  

Thus, the initial perturbation decays and no F─S transition occurs (up-arrow in Fig. 5.2a). As a result, free flow remains (up-arrow in Fig. 5.4c).

To see the Z-shaped function of speed in terms of density clearly, we average all different steady states of synchronized flow for a given density inside the 2D region (dashed region in Fig. 5.4b) to only one averaged synchronized flow speed (curve $v_{\text{av}}^{(\text{syn})}(\rho)$ in Fig. 5.5).

This simplified Z-shaped speed–density relationship (Fig. 5.5) is qualitatively correlated with the hypothesis about the probability $P$ of passing on a multilane road (Fig. 5.6). In the three-phase traffic theory, the following hypothesis is valid [211]:

The probability $P$ of passing on a multilane (one-way) road is a Z-shaped function of density (Fig. 5.6b).

At very low vehicle density in free flow, vehicles can freely pass. Thus, the probability $P$ of passing in free flow should reach 1 as $\rho^{(\text{free})} \to 0$ (curve $P_P$ in Fig. 5.6b). To find the probability of passing, a large number $N_P$ of different realizations (runs) must be examined for the same initial conditions.
Fig. 5.5. Simplified Z-shaped speed–density relationship. The curve $v_{\text{av}}(\rho)$ results from averaging all different synchronized flow speeds at a given density in Fig. 5.4b (dashed region) to the average synchronized flow speed. The black point on the curve $F$ for free flow shows the threshold point (5.10). For simplicity, certain of the average synchronized flow speeds at density $\rho < \rho_{\text{th}}$ are not shown.

In each of these realizations, there should be a driver who moves faster than the preceding vehicle. When approaching the preceding vehicle, the driver should try to pass using a passing freeway lane. It can turn out that the driver is able to pass in some realizations, but not in others. The latter is because the passing lane was occupied by other drivers. If the number of realizations where the driver is able to pass is $n_{\text{P}}$, the probability of passing is\footnote{Strictly speaking, the exact probability of passing is defined as $N_{\text{P}} \to \infty$.}

$$P = \frac{n_{\text{P}}}{N_{\text{P}}} \cdot \tag{5.23}$$

To explain the hypothesis about the Z-shaped form of the probability of passing, note that in accordance with the fundamental hypothesis of the three-phase traffic theory steady states of synchronized flow overlap states of free flow in the vehicle density in a density range (Fig. 5.6a)

$$\left[ \rho_{\text{min}}^{(\text{syn})}, \rho_{\text{max}}^{(\text{free})} \right]. \tag{5.24}$$

However, it is well-known that the probability of passing in congested traffic (in our case, in one of the two traffic phases in congested traffic, synchronized flow) $P$ is considerably lower than in free flow. Thus, in free flow the probability $P$ of passing (the curve $P_F$ in Fig. 5.6b) should be higher than that in synchronized flow (the curve $P_S$). In the range of the density (5.24), at the same given density there can be either a state of free flow where $P = P_F$ is high, or a steady-state of synchronized flow where $P = P_S$.
Fig. 5.6. Explanation of the hypothesis about the Z-shaped dependence of the probability $P$ of passing in the three-phase traffic theory [209, 210]. (a) States of free (curve $F$) and synchronized flow (hatched region), which are the same as those in Fig. 4.4a. (b) Qualitative dependence of the probability $P$ of passing (averaged over all different steady states of synchronized flow at a given density) as a function of density.

is low. This leads to the Z-shaped function of the probability of passing in the three-phase traffic theory.

The curve $P_S$ in Fig. 5.6b, as well as the curve $v_{\text{aver}}^{(\text{syn})}$ in Fig. 5.5, are associated with some simplifications: these curves result from averaging of different synchronized flow speeds at a given density to one average speed. If we consider all these different steady-state synchronized flow speeds (as
Fig. 5.7. Influence of 2D region of steady states of synchronized flow on the density dependence of the passing probability $P$. (a) States of free (curve $F$) and synchronized flow (hatched region), which are the same as those in Fig. 4.4a. (b) Qualitative density dependence of the passing probability $P$ without averaging of the probability for all different steady states of synchronized flow at a given density. The black point in (a, b) is related to the threshold point (5.10)

in Fig. 5.4b,c) we naturally come to Fig. 5.7 for the probability of passing, which reflects the actual 2D region of steady states of synchronized flow.

The Z form of the dependence of the probability $P$ of passing as a function of density, as well as the related Z-shape of the speed–density relationship, enables us to explain the $F\rightarrow S$ transition on a multilane homogeneous road. As the density in free flow gradually increases, the probability of passing $P$ must drop when the density reaches the limit density for free flow $\rho_{\text{max}}^{(\text{free})}$ (Fig. 5.6b). As a result of this drop, the probability $P$ of passing decreases abruptly to the low values of $P$ associated with synchronized flow.
Thus, a spontaneous $F \rightarrow S$ transition must occur at the limit density for free flow $\rho_{\text{max}}^{(\text{free})}$.

### 5.2.6 Physics of Breakdown Phenomenon: Competition Between Over-Acceleration and Speed Adaptation

The physics of a first-order $F \rightarrow S$ transition and the associated $Z$-characteristic (Sect. 5.2.5) can be explained by a competition between two opposing tendencies within a random local perturbation (Fig. 5.2) in an initial free flow:

(i) a tendency towards the initial free flow due to over-acceleration;
(ii) a tendency towards synchronized flow due to vehicle adaptation to the speed of the preceding vehicle.

To explain vehicle over-acceleration, let us assume that a driver is within the synchronization distance to the preceding vehicle. If the driver believes that there is a possibility to pass, the driver accelerates. This can occur even if the preceding vehicle does not accelerate. However, it can turn out that although the driver accelerates, nevertheless the driver cannot pass. In this case, the driver must decelerate to the speed of the preceding vehicle and wait for another possibility to pass. This vehicle over-acceleration with a subsequent vehicle deceleration to the speed of the preceding vehicle can be repeated several times before the driver can pass.

The tendency towards synchronized flow can occur when the driver is within the synchronization distance to the preceding vehicle (Sect. 4.3.2). This tendency results from the need to adapt to the speed of the preceding vehicle when passing is not possible or is difficult.

Thus, the need to decrease the speed to the speed of the slower moving preceding vehicle, i.e., vehicle speed adaptation, which occurs when the vehicle cannot pass this slow moving preceding vehicle, explains the physics of the onset of congestion ($F \rightarrow S$ transition) in free flow. In contrast, vehicle over-acceleration, which enables the vehicle to pass the slow moving preceding vehicle, explains the physics of the reverse transition from synchronized flow (congested traffic) to free flow ($S \rightarrow F$ transition).

Over-acceleration is stronger at higher vehicle speed, or more precisely, at lower density (the flow rate is nearly the same in the initial free flow and in the local perturbation). In this case, the initial local perturbation in free flow whose amplitude is less than the critical amplitude decays (up-arrow in Fig. 5.2b). Over-acceleration is also responsible for the phase transition from synchronized flow to free flow ($S \rightarrow F$ transition).

In contrast, the tendency towards speed adaptation is stronger at lower speed, i.e., at higher density. This causes a decrease in the average vehicle speed within the perturbation whose amplitude exceeds the critical amplitude (down-arrow in Fig. 5.2a) and maintenance of the emergent synchronized flow.
5.2.7 Physics of Threshold Point in Free Flow

To understand the meaning of the threshold point (5.10), we consider two different states of free flow. In the first state, the initial density \( \rho^{(\text{free})} \) and the initial flow rate \( q^{(\text{free})} \) in free flow satisfy the condition

\[
\rho^{(\text{free})} > \rho_{\text{th}} \quad \left( q^{(\text{free})} > q_{\text{th}} \right). \tag{5.25}
\]

In the second state, we have the opposite condition

\[
\rho^{(\text{free})} < \rho_{\text{th}} \quad \left( q^{(\text{free})} < q_{\text{th}} \right). \tag{5.26}
\]

Further we consider the behavior of local perturbations in each of these two states (5.25) and (5.26) (Fig. 5.8). We assume that the speed \( u^{(\text{pert})}_\text{initial} \) and the density \( \rho^{(\text{pert})}_\text{initial} \) within each of these perturbations is the same for both states (5.25) and (5.26). Furthermore, we assume that the amplitude of the perturbation in the first case (5.25) is greater than the critical amplitude \( \Delta v^{(\text{FS})}_{\text{cr}} \) (5.5) for the F\( \rightarrow \)S transition (Fig. 5.8a).

Note that the velocity \( v^{\text{down}}_\text{initial} \) of the downstream front of the initial perturbations in free flow shown in Fig. 5.8 at \( t = t_0 \) (the front that separates free flow downstream from incipient synchronized flow within the perturbation) can be found from the relation

\[
v^{\text{down}} = u^{(\text{pert})}_\text{initial} \left( 1 - \frac{1}{\tau_{\text{del, syn}}^{(a)} \rho^{(\text{pert})}_\text{initial}} \right). \tag{5.27}
\]

In this formula, \( \tau_{\text{del, syn}}^{(a)} \) is the average time delay in vehicle acceleration. This time delay is usually a reaction to an “expected event.” This event is associated with safety conditions for vehicles accelerating from the lower speed \( u^{(\text{pert})}_\text{initial} \) within the perturbation to free flow downstream of the perturbation,

\(6\) To explain (5.27), note that the downstream front of the perturbation, i.e., the front between synchronized flow within the perturbation \( u^{(\text{pert})}_\text{initial} \) and free flow downstream of the perturbation, is defined by vehicles that accelerate from synchronized flow at the vehicle speed \( u^{(\text{pert})}_\text{initial} \) to free flow at a higher speed \( u^{(\text{free})} \) (Fig. 5.8 at \( t = t_0 \)). This vehicle acceleration process occurs with vehicle acceleration time delay \( \tau_{\text{del, syn}}^{(a)} \) – after the preceding vehicle has begun to accelerate, a vehicle must wait some time before acceleration. Let us consider a spatial coordinate system moving at the velocity \( u^{(\text{pert})}_\text{initial} \), i.e., together with vehicles in synchronized flow within the initial perturbation. On average, each vehicle that accelerates from synchronized flow causes a spatial shift of the downstream front of the perturbation opposite the direction of flow, in accordance with the mean distance between vehicles \( 1/\rho^{(\text{pert})}_\text{initial} \), i.e., this shift is equal to \( -1/\rho^{(\text{pert})}_\text{initial} \). Thus, in the chosen moving coordinate system the velocity of the downstream perturbation front is \( -1/\tau_{\text{del, syn}}^{(a)} \rho^{(\text{pert})}_\text{initial} \). The velocity of this coordinate system is \( u^{(\text{pert})}_\text{initial} \). The sum of the latter two velocities is the velocity of the front (5.27).
5.2 Breakdown Phenomenon on Homogeneous Road

Fig. 5.8. Qualitative illustration of local perturbations of vehicle speed in an initial free flow with speed \( v^{(\text{free})} \). (a) Growing local perturbation whose amplitude exceeds the critical amplitude \( \Delta v^{(\text{FS})} = v^{(\text{free})} - v^{(\text{FS})} \) at the density in the initial free flow \( \rho^{(\text{free})} > \rho_{\text{th}} \). The width \( L^{(\text{pert})} \) of the perturbation increases over time. (b) The same initial amplitude local perturbation as those in (a), whose width, however, decreases over time, while the perturbation decays because in (b) \( \rho^{(\text{free})} < \rho_{\text{th}} \).

where \( v^{(\text{free})} > v^{(\text{pert})} \). After the preceding vehicle has begun to accelerate, a vehicle should wait until the distance to the preceding vehicle has increased before accelerating.

Corresponding to the Stokes shock-wave formula (3.5), the mean velocity \( v_{\text{up}} \) of the upstream front of the perturbations in Fig. 5.8 at \( t = t_0 \) (the front that separates free flow upstream of the perturbation from incipient synchronized flow within the perturbation) is

\[
v_{\text{up}} = \frac{q^{(\text{pert})} - q^{(\text{free})}}{\rho^{(\text{initial})} - \rho^{(\text{free})}}.
\]  

(5.28)
Let us consider the case in which $v_{\text{up}} < 0$ and $v_{\text{down}} < 0$ in (5.27) and (5.28), i.e., both initial perturbations in Fig. 5.8a,b propagate upstream.

The velocity $v_{\text{down}}$ (5.27) does not change when the flow rate and density $q^{\text{(free)}}$ and $\rho^{\text{(free)}}$ in free flow outside the perturbations are changing, i.e., this velocity is the same for both initial perturbations in Fig. 5.8a,b.

In contrast, the absolute value of the velocity $|v_{\text{up}}|$ decreases when the flow rate $q^{\text{(free)}}$ and density $\rho^{\text{(free)}}$ decrease. Thus, the threshold flow rate (threshold density) in free flow should yield

$$v_{\text{down}} = v_{\text{up}}. \quad (5.29)$$

At a flow rate higher (density higher) than the threshold flow rate (threshold density) associated with (5.29), the absolute value of the velocity $|v_{\text{up}}|$ (5.28) increases, i.e.,

$$|v_{\text{down}}| < |v_{\text{up}}|. \quad (5.30)$$

At a flow rate lower (density lower) than the threshold flow rate (threshold density) associated with (5.29), the absolute value of the velocity $|v_{\text{up}}|$ (5.28) decreases, i.e.,

$$|v_{\text{down}}| > |v_{\text{up}}|. \quad (5.31)$$

Under the condition (5.30), the width of the region of synchronized flow in the local perturbation $L^{\text{(pert)}}$ increases over time (from $t = t_0$ to $t = t_1 > t_0$, Fig. 5.8a).

In contrast, under the condition (5.31), the width of the local perturbation $L^{\text{(pert)}}$ decreases over time (from $t = t_0$ to $t = t_1 > t_0$ in Fig. 5.8b). In this case, the local perturbation must dissolve over time.

The condition (5.29) is satisfied at the threshold point (5.10) at which $\rho^{\text{(free)}} = \rho_{\text{th}} (q^{\text{(free)}} = q_{\text{th}})$. In states of free flow above the threshold point, i.e., under the condition (5.25) the condition (5.30) is valid, whereas in states of free flow below the threshold point, i.e., under the condition (5.26), the opposite condition (5.31) is true. Thus, the physics of the threshold point (5.10) is related to the condition (5.29) at which the velocities of upstream and downstream fronts of local perturbations in free flow become equal to one another.

5.2.8 Moving Synchronized Flow Pattern

The decay of an initial local perturbation due to the condition (5.31), i.e., in the states of free flow (5.26) that are below the threshold point (5.10), is confirmed by the numerical simulation shown in Fig. 5.9.

In contrast to this decay effect, when the condition (5.30) is satisfied, i.e., in the states of free flow (5.25) that are above the threshold point (5.10) spontaneous nucleation should occur if the perturbation amplitude exceeds the critical amplitude. It can be expected that due to the F$\rightarrow$S transition
5.3 Breakdown Phenomenon at Freeway Bottlenecks

5.3.1 Deterministic Local Perturbation

In real traffic, the breakdown phenomenon (F→S transition) occurs much more frequently at a freeway bottleneck. This is related to the occurrence of some permanent (“deterministic”) local perturbation on the main road in the vicinity of the bottleneck. To explain this, let us note that the bottleneck plays the role of a permanent (deterministic) inhomogeneity localized on a moving synchronized flow pattern (MSP) appears in the initial free flow. This is indeed confirmed by numerical simulations (Fig. 5.10).

The MSP is a localized spatiotemporal pattern surrounded by an initial metastable state of free flow. Within the MSP, synchronized flow is realized. The speed in this synchronized flow is appreciably lower than the speed in the initial free flow (Fig. 5.10b). However, the flow rate within the MSP is only slightly less than the flow rate in the initial free flow (Fig. 5.10c). The width of MSP $L_{MSP}$ is an increasing function over time if the condition (5.30) is satisfied.

In contrast, if the amplitude of the initial perturbation in the initial free flow is less than the critical amplitude, the perturbation decays over time even if (5.30) is satisfied (up-arrow in Fig. 5.2b). A numerical simulation of this decay effect is shown in Fig. 5.11.

Fig. 5.9. Numerical simulation of the decay of a high-amplitude initial local perturbation under the condition (5.31). The perturbation has been applied at time $t = t_0$. Taken from [330]
Fig. 5.10. Numerical simulation of moving synchronized pattern (MSP) emergence in a microscopic three-phase traffic theory. (a) Time and space dependence of vehicle speed during MSP emergence. MSP emergence occurs under the condition (5.30) after a local perturbation has been applied to an initial free flow at time $t = t_0$. The amplitude of the perturbation is greater than the critical amplitude. (b, c) Temporal behavior of the speed (b) and flow rate (c) within the MSP at the constant spatial coordinate $x = 9$ km. Taken from [330]

road in the vicinity of the bottleneck. At a high enough flow rate, this road inhomogeneity causes a local decrease in average vehicle speed and an increase in vehicle density in the vicinity of the bottleneck. This local decrease in speed and increase in density is permanent in free flow, and is localized at the bottleneck. This disturbance in free flow can be considered a deterministic local perturbation in free flow at the bottleneck (Fig. 5.12b–d).

Let us consider free flow at a time-independent flow rate downstream of a bottleneck. The disturbance caused by the bottleneck in free flow is permanent and on average motionless. In other words, the spatial distribution of vehicle density $\rho(x)$ within the deterministic perturbation does not depend
on time. This means that
\[ \frac{\partial \rho}{\partial t} = 0. \] (5.32)
Therefore, from the vehicle balance equation (3.1) we find
\[ \frac{\partial q}{\partial x} = 0. \] (5.33)
From this equation it follows that the flow rate \( q \)
\[ q = v(x) \rho(x) \] (5.34)

does not depend on the spatial coordinate \( x \) within the deterministic perturbation.\(^8\) Here \( v(x) \) and \( \rho(x) \) are the freeway location dependencies of the average speed and density within the deterministic perturbation on the main road in the vicinity of the bottleneck.

\(^8\) This is qualitatively the same effect as independence of the flow rate on spatial coordinate in the vicinity of a congested bottleneck, which is well-known from investigations of speed breakdown at freeway bottlenecks (see references in the papers by Persaud and Hurdle [59] and by Hall, Hurdle, and Banks [30]).

\(^9\) Note that in the vicinity of the bottleneck, the total number of freeway lanes (also taking into account on- and off-ramps) usually depends on the freeway location. For this reason, here and below in this Part I when the flow rate and density are considered to be functions of the spatial coordinate, they are taken per freeway lane for the same chosen freeway location (e.g., a freeway location downstream of the bottleneck). This means that to find the flow rate \( q \) (5.34), the total flow rate is divided by the same chosen number of freeway lanes regardless of a freeway location.
5 Breakdown Phenomenon (F→S Transition)

Fig. 5.12. Qualitative explanation of a deterministic local perturbation in free flow in the vicinity of an on-ramp bottleneck. (a) Sketch of the on-ramp bottleneck. (b–d) Amplitude growth of the deterministic local perturbation in speed \( v(x) \) on the main road at a given \( q_{in} \) when the flow rate \( q_{on} \) increases from \( q_{on} = q_{on}^{(1)} \) (b) through \( q_{on} = q_{on}^{(2)} > q_{on}^{(1)} \) (c) to the critical flow rate \( q_{on} = q_{on}^{(\text{determin}, \text{FS})} > q_{on}^{(2)} \) (d). (e) Probability density \( p_{FS} \) for an F→S transition as a function of road location \( x \). At some location in the vicinity of the bottleneck, the probability density has a maximum \( p_{FS, \text{max}} \).
For simplicity, we restrict further discussion of traffic phenomena at a freeway bottleneck to on-ramps (Fig. 5.12a). At a freeway bottleneck, vehicles merge from the on-ramp, where the initial flow rate is \( q_{\text{on}} \), onto the main road, where the initial flow rate upstream of the on-ramp is \( q_{\text{in}} \) (Fig. 5.12a). The flow rate downstream of the on-ramp (that will be denoted by \( q_{\text{sum}} \)) is assumed to be related to a spatially homogeneous free flow. When free flow conditions are realized both upstream and at the bottleneck, the flow rate \( q_{\text{sum}} \) is given by the obvious condition

\[
q_{\text{sum}} = q_{\text{on}} + q_{\text{in}}. 
\]  

(5.35)

**Speed within Deterministic Perturbation**

Vehicles merge from the on-ramp onto the main road within some vehicle merging region near the on-ramp. In the vicinity of this merging region, vehicles on the main road have to decelerate, if the flow rate \( q_{\text{in}} \) is high enough. This deceleration leads to a local decrease in vehicle speed in the vicinity of the bottleneck (Fig. 5.12a, b).

At some given time-independent flow rates \( q_{\text{on}} \) and \( q_{\text{in}} \), the decrease in speed in free flow at the bottleneck leads to an increase in density, corresponding to (5.34), which in the present case reads

\[
q_{\text{sum}} = q_{\text{on}} + q_{\text{in}} = v(x)p(x) = \text{const},
\]

(5.36)

where \( v(x) \) and \( p(x) \) are the speed and density on the main road as functions of the freeway location in the vicinity of the bottleneck. Note that the remark in footnote 9 also applies to the flow rates \( q_{\text{on}} \) and \( q_{\text{in}} \), as well as to the density. The decrease in vehicle speed (Fig. 5.12b–d) and the related increase in density at the bottleneck can be considered as a deterministic effect. This explains the local deterministic perturbation on the main road in the vicinity of the on-ramp.

We denote the vehicle speed in free flow on the main road within the deterministic perturbation at the bottleneck by \( v_{\text{free}}^{(B)} \). The speed \( v_{\text{free}}^{(B)} \) is lower than the speed \( v^{(\text{free})} \) in free flow downstream of the bottleneck (Fig. 5.12b,c):

\[
v_{\text{free}}^{(B)} < v^{(\text{free})}. 
\]

(5.37)

As follows from (5.36),

\[
q_{\text{sum}} = v^{(\text{free})}\rho^{(\text{free})} = v_{\text{free}}^{(B)}\rho_{\text{free}}^{(B)},
\]

(5.38)

where \( \rho_{\text{free}}^{(B)} \) is the density within the deterministic perturbation at the bottleneck associated with the speed \( v_{\text{free}}^{(B)} \), \( \rho^{(\text{free})} \) is the density in free flow downstream of the bottleneck associated with the speed \( v^{(\text{free})} \). In accordance with (5.37) and (5.38), the density \( \rho_{\text{free}}^{(B)} \) is higher than the density \( \rho^{(\text{free})} \).
This speed $v_{\text{free}}^{(B)}$ is a function of the flow rate $q_{\text{sum}}$, and consequently of the density downstream of the bottleneck $\rho^{(\text{free})}$:

$$v_{\text{free}}^{(B)} = v_{\text{free}}^{(B)} \left( \rho^{(\text{free})} \right).$$

The higher the vehicle density in free flow downstream of the bottleneck, the lower the speed $v_{\text{free}}^{(B)}(\rho^{(\text{free})})$ (Fig. 5.13). To understand this, we consider the qualitative behavior of the speed in free flow at the bottleneck when the flow rate $q_{\text{in}}$ has a given high value and the flow rate $q_{\text{on}}$ increases.

It is clear that the higher the flow rate $q_{\text{on}}$, the lower the speed $v_{\text{free}}^{(B)}$ – the more vehicles merge from the on-ramp lane onto the main road, the greater the disturbance in free flow on the main road caused by these merging vehicles. Thus, the speed $v_{\text{free}}^{(B)}$ within the deterministic perturbation continuously decreases and the density $\rho_{\text{free}}^{(B)}$ increases at the bottleneck when the flow rate $q_{\text{on}}$ further increases. If the flow rate $q_{\text{in}}$ is given and the flow rate $q_{\text{on}}$ increases, then the flow rate $q_{\text{sum}}$, and consequently the density $\rho^{(\text{free})}$ downstream of the bottleneck, both increase. For this reason, we can plot the effect of a decrease in speed at the bottleneck to the speed within the deterministic perturbation $v_{\text{free}}^{(B)}$ as a function of density $\rho^{(\text{free})}$ (Fig. 5.13b).

**Amplitude of Deterministic Perturbation**

We explained above the existence of deterministic perturbations in free flow at the bottleneck: at a high enough flow rate $q_{\text{on}}$ to the on-ramp, the speed $v_{\text{free}}^{(B)}$ in free flow on the main road in the vicinity of the on-ramp (Fig. 5.13), i.e., the speed within the deterministic perturbation (Fig. 5.12b,c), is less than the speed $v^{(\text{free})}$ in free flow, which is related to the flow rate $q_{\text{sum}}$ in free flow downstream of the bottleneck (5.37).

The difference between the speed $v^{(\text{free})}$ downstream of the bottleneck and the speed $v_{\text{free}}^{(B)}$ within the perturbation at the bottleneck can be considered as the amplitude of the deterministic perturbation in the speed:

$$\Delta v_{\text{determ}}^{(\text{pert})} = v^{(\text{free})} - v_{\text{free}}^{(B)}.$$

**5.3.2 Deterministic F→S Transition**

**Critical Point for Deterministic Speed Breakdown in Free Flow at Bottleneck**

Let us consider the behavior of the speed $v_{\text{free}}^{(B)}$ at the bottleneck, when the flow rate $q_{\text{in}}$ is given and the flow rate $q_{\text{on}}$ gradually increases. In this case, as discussed above, the speed $v_{\text{free}}^{(B)}$ should gradually decrease.

However, there must be some constraint on this gradual decrease in the speed $v_{\text{free}}^{(B)}$ at the bottleneck. This constraint is related to the empirical
Fig. 5.13. Qualitative illustration of the deterministic breakdown phenomenon (deterministic $F \rightarrow S$ transition) in an initial free flow at a freeway bottleneck. (a) States of free flow and steady states of synchronized flow (dashed region) in the flow–density plane. (b) Speed–density characteristics of (a). Solid curves $F(q_{\text{sum}}(\rho_{\text{free}}))$ in (a) and $v_{\text{free}}^{(B)}(\rho_{\text{free}})$ in (b) related to a road with the bottleneck; dashed branches show states of free flow on a homogeneous road. The steady states of synchronized flow are taken from Fig. 5.1.
result (3) of Sect. 4.4. Corresponding to this result, there is a limit (maximum) point of free flow at a bottleneck. At this limit point, the vehicle speed reaches the minimum possible in free flow at the bottleneck. Free flow at the bottleneck cannot exist at a lower vehicle speed than this minimum speed, and must transform into synchronized flow.

Accordingly, we can assume that there must be a minimum value of the speed \( v_{\text{free}}^{(B)} \) within the deterministic local perturbation at a bottleneck. We denote this minimum (critical) speed at the bottleneck by \( v_{\text{determ, FS}}^{(B)} \). Thus, if \( q_{\text{on}} \) increases continuously, and consequently the speed in free flow at the bottleneck \( v_{\text{free}}^{(B)} \) decreases, the speed \( v_{\text{free}}^{(B)} \) should reach the critical speed \( v_{\text{determ, FS}}^{(B)} \). However, from the above discussion we see that free flow cannot exist when the vehicle speed \( v_{\text{free}}^{(B)} \) at the bottleneck becomes lower than the critical value \( v_{\text{determ, FS}}^{(B)} \).

This means that there is a critical point of free flow

\[
v_{\text{free}}^{(B)} = v_{\text{determ, FS}}^{(B)} \quad \left( \rho_{\text{free}}^{(B)} = \rho_{\text{determ, FS}}^{(B)} \right), \tag{5.41}
\]

where \( \rho_{\text{determ, FS}}^{(B)} \) is the critical density at the bottleneck associated with the critical speed at the bottleneck \( v_{\text{determ, FS}}^{(B)} \). At the critical point (5.41) synchronized flow must occur at the bottleneck. This \( F \rightarrow S \) transition is related to the existence of a deterministic local perturbation at the bottleneck. This phase transition occurs even if there were no random perturbations in free flow on the main road in the vicinity of the bottleneck. For these reasons, this \( F \rightarrow S \) transition is deterministic at the bottleneck. This breakdown is a precipitous decrease in speed, and a related increase in density at the location of the deterministic perturbation. Symbolically this deterministic local speed breakdown is shown by the down-arrow in Fig. 5.12d.

The flow rate \( q_{\text{sum}} \) and the density downstream of the bottleneck increase when the flow rate \( q_{\text{on}} \) increases at a given flow rate \( q_{\text{in}} \). There is a critical flow rate downstream of the bottleneck. We denote this critical flow rate by \( q_{\text{determ, FS}}^{(B)} \). When the flow rate downstream of the bottleneck \( q_{\text{sum}} \) reaches the critical value, i.e.,

\[
q_{\text{sum}} = q_{\text{determ, FS}}^{(B)}, \tag{5.42}
\]

the speed at the bottleneck \( v_{\text{free}}^{(B)} \) is equal to the critical speed \( v_{\text{determ, FS}}^{(B)} \) (5.41). Thus, (5.42) determines the flow rate \( q_{\text{sum}} \) downstream of the bottleneck associated with the critical point of free flow

\[
\left( \rho_{\text{determ, FS}}^{(B)}, v_{\text{determ, FS}}^{(B)} \right), \tag{5.43}
\]

for a deterministic \( F \rightarrow S \) transition at the bottleneck (Fig. 5.13).
5.3 Breakdown Phenomenon at Freeway Bottlenecks

Critical Amplitude of Deterministic Perturbation

At the critical point (5.41), the amplitude of the deterministic perturbation in free flow at the bottleneck reaches the critical amplitude. Corresponding to (5.40) and (5.41), this critical amplitude (Fig. 5.12d) is

\[ \Delta v_{\text{determ, cr}}^{(\text{pert})} = v_{\text{determ, FS}}^{(\text{free})} - v_{\text{determ, FS}}^{(B)}. \] (5.44)

The speed at the bottleneck \( v_{\text{free}}^{(B)} \) is a function of density (Fig. 5.13b). However, the vehicle density depends on both the flow rate \( q_{\text{on}} \) and the flow rate \( q_{\text{in}} \). For this reason, it is also convenient to consider the speed \( v_{\text{free}}^{(B)} \) as a function of the flow rate \( q_{\text{on}} \) at a given flow rate \( q_{\text{in}} \) (Fig. 5.14). At a given flow rate upstream of the bottleneck \( q_{\text{in}} \) the critical flow rate \( q_{\text{sum}} = q_{\text{determ, FS}}^{(B)} \) corresponds to some critical flow rate to the on-ramp. We denote this critical flow rate to the on-ramp by \( q_{\text{on}}^{(\text{determ, FS})} \) (Fig. 5.14). Thus, at the given flow rate \( q_{\text{in}} \) a deterministic F→S transition occurs at the critical flow rate to the on-ramp \( q_{\text{on}}^{(\text{determ, FS})} \).

It should be noted that the flow rate downstream of the bottleneck \( q_{\text{sum}} \) (5.35) is a function of the flow rate upstream of the bottleneck \( q_{\text{in}} \). Thus, the critical flow rate \( q_{\text{sum}} = q_{\text{determ, FS}}^{(B)} \) downstream of the bottleneck and the critical flow rate to the on-ramp \( q_{\text{on}}^{(\text{determ, FS})} \) as well as the critical speed \( v_{\text{determ, FS}}^{(B)} \) and the critical density \( \rho_{\text{determ, FS}}^{(B)} \) are functions of the flow rate \( q_{\text{in}} \). This is because at the same \( q_{\text{sum}} \) depending on the flow rate \( q_{\text{in}} \) different local deterministic perturbations in the speed can occur in the vicinity of the bottleneck.

Features of Deterministic Speed Breakdown

The breakdown phenomenon described above is a deterministic F→S transition. During this transition a precipitous decrease in speed at the bottleneck at the critical point of free flow (5.41) occurs even if there are no random perturbations (fluctuations) in free flow.

It must be noted that even if random perturbations in free flow are negligible for the deterministic F→S transition, this phase transition is a spontaneous effect, i.e., it is a spontaneous F→S transition. Indeed, this phase transition must occur at the critical point (5.41) without any additional external perturbation of traffic flow for an intrinsic reason, which has been explained above: after the critical speed (5.41) has been reached at the bottleneck, free flow can no longer exist at the bottleneck. The spontaneous local breakdown in free flow within the deterministic perturbation is denoted by dotted arrows \( K_{\text{determ}} \rightarrow M_{\text{determ}} \) in Fig. 5.13.

Within the deterministic perturbation, the speed \( v_{\text{free}}^{(B)} \) is lower and the density \( \rho_{\text{free}}^{(B)} \) is higher than these traffic variables are downstream of the bottleneck, respectively. For this reason, the deterministic F→S transition occurs
Fig. 5.14. Qualitative illustration of a Z-shaped speed–flow characteristic for the spontaneous breakdown phenomenon (spontaneous \( F \rightarrow S \) transition) of initial free flow at a freeway bottleneck due to an on-ramp. (a) Free flow states (curve \( v^{(B)}_{\text{free}} \)) and steady states of synchronized flow (dashed region) in the flow–density plane, together with the critical branch \( v^{(B)}_{\text{cr, FS}} \) (dashed curve), which gives the speed within the critical perturbation. (b) Simplified Z-shaped dependence of the speed on the flow rate \( q_{\text{on}} \), which is related to (a)

at a lower maximum flow rate \( q_{\text{sum}} = q^{(B)}_{\text{determ, FS}} \) downstream of the bottleneck than the limiting flow rate (5.2) on a homogeneous road (Fig. 5.13a):

\[
q^{(B)}_{\text{determ, FS}} < q^{(\text{free})}_{\text{max}}.
\]

This formula may explain the well-known empirical result that the breakdown phenomenon (\( F \rightarrow S \) transition) is most frequently observed at freeway bottlenecks: the critical point of free flow (5.2) for a homogeneous road is usually
very difficult to reach, because at the lower flow rate $q_{\text{sum}} = q_{\text{det}, \text{FS}}^{(B)}$ (5.42) the deterministic $F \to S$ transition occurs at the bottleneck.

Due to the deterministic perturbation at the bottleneck, the probability density for a spontaneous $F \to S$ transition in the given time interval $T_{\text{ob}}$, $p_{FS}(x)$ (probability for an $F \to S$ transition per km as a function of freeway location) has a sharp maximum, $p_{FS, \text{max}}$, at the bottleneck (Fig. 5.12e). In other words, the decrease in speed and increase in density within the deterministic perturbation cause a spontaneous $F \to S$ transition at the bottleneck with much higher probability than at freeway locations away from the bottleneck. This conclusion correlates with the condition (5.45).

5.3.3 Physics of Deterministic Speed Breakdown at Bottleneck

As in the hypothetical case of a homogeneous road (Sect. 5.2.6), the physics of the deterministic $F \to S$ transition (deterministic onset of congestion) at the bottleneck is associated with the competition between a tendency towards the initial free flow due to over-acceleration and a tendency towards synchronized flow due to vehicle adaptation to the speed of the preceding vehicle. However, in the case under consideration this competition occurs within the deterministic local perturbation in the vicinity of the bottleneck.

The lower the speed $v_{\text{free}}^{(B)}$ within the deterministic perturbation, the stronger the tendency towards speed adaptation. As a result, when the critical speed $v_{\text{det}, \text{FS}}^{(B)}$ within the deterministic perturbation at the bottleneck is reached, the deterministic onset of congestion occurs at the bottleneck (down-arrow in Fig. 5.12d). The speed adaptation effect causes further maintenance of the emergent synchronized flow at the bottleneck.

5.3.4 Influence of Random Perturbations

The $F \to S$ transition at the on-ramp is a first-order phase transition, which is accompanied by nucleation, metastability, and hysteresis (Figs. 5.13 and 5.14). Nucleation, metastability, and hysteresis are common attributes of first-order phase transitions (see Appendix A).

In particular, real random perturbations in free flow near a bottleneck can lead to a spontaneous $F \to S$ transition even if the critical point of free flow for a deterministic $F \to S$ transition (5.41), (5.42) has not yet been achieved. In other words, at

$$q_{\text{sum}} < q_{\text{det}, \text{FS}}^{(B)}$$

(5.46)

there are metastable states in which the spontaneous nucleation of synchronized flow at an on-ramp can occur. This spontaneous $F \to S$ transition is symbolically indicated by dotted arrows $K' \to M'$ in Fig. 5.13a. At a given flow rate upstream of the bottleneck $q_{\text{in}}$ the condition (5.46) can be rewritten as follows
Breakdown Phenomenon (F→S Transition)

\[ q_{on} < q_{on}^{(\text{deterministic, FS})}. \]  

(5.47)

The spontaneous F→S transition that occurs due to a random local perturbation under the condition (5.47) is symbolically indicated by down-arrows labeled “F→S” in Fig. 5.14.

To explain the influence of random perturbations on an F→S transition at the bottleneck, note that in real traffic flow a local perturbation due to the vehicle merging from the on-ramp onto the main road consists of two components:

(i) The first is the deterministic local perturbation considered above. The speed within the deterministic local perturbation at the bottleneck is \( v^{(B)}_{\text{free}} \) (Fig. 5.12b,c).

(ii) The second is a random component of the local perturbation (Fig. 5.15).

We will be interested in those random perturbations that lead to a decrease in the speed in free flow at the bottleneck in comparison with the speed \( v^{(B)}_{\text{free}} \).

Let us denote the speed within the random local perturbation due to these deterministic and random components by \( v^{(\text{pert})}_{\text{random}} \). Then the amplitude of this random perturbation is (Fig. 5.15b,c)

\[ \Delta v^{(\text{pert})}_{\text{random}} = v^{(B)}_{\text{free}} - v^{(\text{pert})}_{\text{random}}. \]  

(5.48)

Critical Perturbation Amplitude

Under the condition (5.47), there is a critical amplitude of the random local perturbation

\[ \Delta v^{(B)}_{\text{cr, FS}} = v^{(B)}_{\text{free}} - v^{(B)}_{\text{cr, FS}}, \]  

(5.49)

where \( v^{(B)}_{\text{cr, FS}} \) is the critical speed at the bottleneck, which is a function of the flow rate \( q_{on} \) at a given flow rate \( q_{in} \):

\[ v^{(B)}_{\text{cr, FS}} = v^{(B)}_{\text{cr, FS}}(q_{on}). \]  

(5.50)

The critical amplitude \( \Delta v^{(B)}_{\text{cr, FS}}(q_{on}) \) (5.49) and the speed at the bottleneck \( v^{(B)}_{\text{free}}(q_{on}) \) are also functions of the flow rate \( q_{on} \) at the given flow rate \( q_{in} \).

The physical sense of the critical amplitude of the local perturbation (5.49) is as follows. Let us assume that a random local perturbation appears on the main road near an on-ramp (Fig. 5.15). If the amplitude of the random perturbation \( \Delta v^{(\text{pert})}_{\text{random}} \) exceeds the critical amplitude \( \Delta v^{(B)}_{\text{cr, FS}}(q_{on}) \), i.e., if

\[ \Delta v^{(\text{pert})}_{\text{random}} > \Delta v^{(B)}_{\text{cr, FS}}(q_{on}), \]  

(5.51)
5.3 Breakdown Phenomenon at Freeway Bottlenecks

Fig. 5.15. Qualitative explanation of a random local perturbation in free flow on the main road near an on-ramp. (a) Sketch of the on-ramp. (b) The random local perturbation in the speed at the bottleneck grows (down-arrow) if the amplitude of the perturbation \( \Delta v^{(\text{pert})}_{\text{random}} = v^{(B)}_{\text{free}} - v^{(pert)}_{\text{random}} \) exceeds the critical amplitude \( \Delta v^{(B)}_{\text{cr}, \text{FS}} = v^{(B)}_{\text{free}} - v^{(B)}_{\text{cr}, \text{FS}} \). (c) The random local perturbation in the speed at the bottleneck decays (up-arrow) if the amplitude of the perturbation \( \Delta v^{(pert)}_{\text{random}} \) is less than the critical amplitude \( \Delta v^{(B)}_{\text{cr}, \text{FS}} \).

The perturbation grows and leads to a spontaneous F→S transition at the bottleneck (Fig. 5.15b and down-arrows in Fig. 5.14). In this case, synchronized flow appears at the bottleneck.

Otherwise, if

\[
\Delta v^{(pert)}_{\text{random}} < \Delta v^{(B)}_{\text{cr}, \text{FS}}(q_{\text{on}}),
\]

the perturbation decays (Fig. 5.15c and up-arrows in Fig. 5.14). In this case, free flow remains at the bottleneck.
5.3.5 Z-Characteristic for Speed Breakdown at Bottleneck

Dependence of Critical Perturbation Amplitude on Flow Rate

To explain the critical speed \( v_{cr, FS}(q_{on}) \) (5.50), let us note that at the limit case

\[ q_{on} = q_{on}^{(\text{determin, FS})} \]  

(5.53)
a deterministic F→S transition occurs at the bottleneck. In this case, the critical amplitude of a random local perturbation (5.49) is zero:

\[ \Delta v_{cr, FS}(q_{on}^{(\text{determin, FS})}) = 0 \]  

(5.54)

This leads to the condition

\[ v_{cr, FS}(q_{on}^{(\text{determin, FS})}) = v_{cr, FS}(q_{on}^{(\text{determin, FS})}) = v_{determin, FS} \]  

(5.55)

We can see from (5.55) that under the condition (5.53) the critical speed at the bottleneck (5.50) coincides with the critical speed at the critical point (5.41) for the deterministic F→S transition. This is because at the critical point (5.41) an F→S transition occurs at the bottleneck even if there are no random perturbations at the bottleneck.

Under the condition (5.47), the critical point (5.41) is not achieved and a deterministic F→S transition cannot occur. In this case, an F→S transition can only occur if random perturbations appear at the bottleneck. However, when at the same flow rate upstream of the bottleneck \( q_{in} \) the flow rate \( q_{on} \) is lower than the critical flow rate (5.53), we can expect that the critical speed at the bottleneck for an F→S transition should also decrease, i.e.,

\[ v_{cr, FS}(q_{on}) < v_{cr, FS}(q_{on}^{(\text{determin, FS})}) \]  

(5.56)

This condition corresponds to an obvious assumption that at the same flow rate \( q_{in} \) the lower the flow rate \( q_{on} \), the higher the decrease in the speed at the bottleneck due to a local perturbation should be to cause an F→S transition. In other words, when the condition (5.47) is satisfied, the critical speed (5.50) for an F→S transition due to a random perturbation should be lower than the critical speed is at the critical point (5.41) for a deterministic F→S transition. Furthermore, the lower the flow rate \( q_{on} \), the lower the critical speed (5.50).

This dependence of the speed \( v_{cr, FS}(5.50) \) on the flow rate \( q_{on} \) determines “the critical branch” \( v_{cr, FS}(q_{on}) \) for a spontaneous F→S transition at the bottleneck (Fig. 5.14). To explain the term “the critical branch,” note that if a random local perturbation makes the speed at the bottleneck lower than the speed on the “the critical branch” \( v_{cr, FS}(q_{on}) \), the condition (5.51) will be satisfied. Corresponding to this condition, the perturbation grows and leads
to a spontaneous F→S transition at the bottleneck (Fig. 5.15b). Otherwise, let us assume that the speed within a random perturbation at the bottleneck is higher than the speed on the critical branch $v^{(B)}_{cr, FS}(q_{on})$. Then the condition (5.52) is satisfied. In this case, the perturbation decays.

Let us consider the qualitative behavior of the speed within the deterministic perturbation at the bottleneck $v^{(B)}_{free}(q_{on})$ as a function of the flow rate $q_{on}$ at a given flow rate $q_{in}$. When the flow rate of vehicles merging onto the main road from the on-ramp $q_{on}$ increases, the speed within the deterministic perturbation $v^{(B)}_{free}(q_{on})$ decreases (Fig. 5.14a). If the flow rate $q_{on}$ increases further, $v^{(B)}_{free}(q_{on})$ decreases up to the critical point (5.41) where the condition (5.55) is satisfied. This means that the difference

$$v^{(B)}_{free}(q_{on}) - v^{(B)}_{cr, FS}(q_{on})$$

should be a decreasing function of the flow rate $q_{on}$ at the given flow rate $q_{in}$. This qualitative result is illustrated in Fig. 5.14, where the critical branch $v^{(B)}_{cr, FS}(q_{on})$ is shown. At each value $q_{on}$, corresponding to (5.49), the critical amplitude of the critical perturbation at the bottleneck $\Delta v^{(B)}_{cr, FS}$ is equal to the difference between the speed at the bottleneck $v^{(B)}_{free}(q_{on})$ and the critical speed on the critical branch $v^{(B)}_{cr, FS}(q_{on})$ (Fig. 5.14). Therefore, the critical amplitude $\Delta v^{(B)}_{cr, FS}$ required for an F→S transition decreases when $q_{on}$ increases. The critical amplitude $\Delta v^{(B)}_{cr, FS}$ reaches zero at the limit point (5.53) where a deterministic F→S transition occurs.

### Z-Characteristic as Feature of Speed Breakdown

The dependence on the flow rate to the on-ramp $q_{on}$ of the speed in free flow at the bottleneck $v^{(B)}_{free}$ and of the critical speed $v^{(B)}_{cr, FS}$, along with a 2D region of states of synchronized flow (dashed region), together form a Z-shaped characteristic in the speed–flow plane (Fig. 5.14a).

To explain the Z-characteristic, note that at the critical point of free flow the condition (5.55) is satisfied. This means that at the critical point of free flow the speed branch for free flow at the bottleneck $v^{(B)}_{free}(q_{on})$ merges with the critical branch $v^{(B)}_{cr, FS}(q_{on})$ (Fig. 5.14).

This critical speed $v^{(B)}_{determ, FS}$ should be reached when the flow rate $q_{on}$ reaches some critical flow rate to the on-ramp $q^{(determ, FS)}_{on}$. The latter critical flow rate depends on the flow rate $q_{in}$:

$$q^{(determ, FS)}_{on} = q^{(determ, FS)}_{on}(q_{in}) .$$

The flow rate downstream of the bottleneck, $q_{sum} = q_{in} + q_{on}$ (5.35), which is related to the critical point (5.41) where the deterministic F→S transition must occur, is
\[ q_{\text{sum}} = q_{\text{determin, FS}} = q_{\text{in}} + q_{\text{on, FS}}(q_{\text{in}}). \] (5.59)

From Fig. 5.14a we can see that at a given flow rate \( q_{\text{in}} \), the lower the flow rate \( q_{\text{on}} \), the greater the amplitude of the random critical perturbation (5.49) required for an F→S transition at the bottleneck. This means that the lower the flow rate \( q_{\text{on}} \), the greater the difference between the speed branch \( v_{\text{free}}^{(B)}(q_{\text{on}}) \) for free flow and the critical branch \( v_{\text{cr, FS}}^{(B)}(q_{\text{on}}) \).

As in the case of a homogeneous road, there should be a threshold flow rate \( q_{\text{on}}^{(th)} \) for an F→S transition: at

\[ q_{\text{on}} < q_{\text{on}}^{(th)} \quad (q_{\text{in}} = \text{const}) \] (5.60)

the F→S transition at the bottleneck cannot occur (for more detail about this threshold point see Sect. 8.3.3). This means that the critical branch \( v_{\text{cr, FS}}^{(B)} \) should merge with states of synchronized flow at the threshold flow rate \( q_{\text{on}} = q_{\text{on}}^{(th)} \) (Fig. 5.14a). Thus, the Z-characteristic for an F→S transition at the bottleneck in Fig. 5.14a consists of

(a) the speed branch \( v_{\text{free}}^{(B)}(q_{\text{on}}) \) that yields the speed within the deterministic perturbation in free flow at the bottleneck as a function of the flow rate \( q_{\text{on}} \);

(b) the critical branch \( v_{\text{cr, FS}}^{(B)}(q_{\text{on}}) \) that yields the speed within the critical local perturbation for an F→S transition at the bottleneck as a function of the flow rate \( q_{\text{on}} \);

(c) a 2D region of infinite speeds in synchronized flow (dashed region).

To show the Z-shaped character of the speed-flow dependence more clearly, in the simplified Fig. 5.14b we have averaged the infinity of different synchronized flow speeds for each given flow rate \( q_{\text{on}} \) (dashed region in Fig. 5.14a) to one speed, \( v_{\text{syn, aver}}^{(B)}(q_{\text{on}}) \).

Thus, an F→S transition at a bottleneck is associated with the Z-shaped traffic flow characteristic shown in Fig. 5.14. As discussed above, the Z-characteristic explains the F→S transition in free flow at the bottleneck as follows.

(i) Let us assume that a random local perturbation occurs at the bottleneck that leads to a decrease in the speed on the main road in the vicinity of the bottleneck. This vehicle speed at the bottleneck is equal to the speed within the random perturbation \( v_{\text{random}}^{(pert)} \) (Fig. 5.15b). There are two different cases:

(1) The amplitude of this perturbation \( \Delta v_{\text{random}}^{(pert)} \) corresponds to the condition (5.51). Then the speed within the perturbation at the bottleneck \( v_{\text{random}}^{(pert)} \) is lower than that on the critical branch \( v_{\text{cr, FS}}^{(B)}(q_{\text{on}}) \) of the Z-characteristic (Fig. 5.14). As a result, the perturbation grows and leads to a spontaneous F→S transition at the bottleneck (down-arrows “F→S” in Fig. 5.14). In this case, some state of synchronized flow occurs at the bottleneck.
(2) In contrast, we assume that the amplitude of a local perturbation \( \Delta v_{\text{random}}^{(\text{pert})} \) at the bottleneck corresponds to the condition (5.52) (Fig. 5.15c), the speed in free flow at the bottleneck \( v_{\text{random}}^{(\text{pert})} \) will be higher than the speed on the critical branch \( v_{\text{cr, FS}}^{(\text{B})}(q_{on}) \) (Fig. 5.14). In this case, the perturbation decays and no spontaneous F→S transition can occur, i.e., free flow remains at the bottleneck (up-arrows in Fig. 5.14).

(ii) If at a given \( q_{in} \) the flow rate \( q_{on} \) is lower than the threshold value \( q_{on}^{(\text{th})} \) (5.60), then regardless of how large the amplitude of a random local perturbation is, no spontaneous F→S transition can occur, i.e., free flow remains at the bottleneck.

Note that the threshold flow rate \( q_{on}^{(\text{th})} \) is a function of the flow rate \( q_{in} \) – the lower the flow rate \( q_{in} \), the higher the threshold flow rate \( q_{on}^{(\text{th})} \).

5.3.6 Physics of Speed Breakdown at Bottleneck

As in the hypothetical case of a homogeneous road (Sect. 5.2.6), the physics of the spontaneous F→S transition at the bottleneck (onset of congestion at the bottleneck) is associated with the competition between a tendency towards the initial free flow due to over-acceleration and a tendency towards synchronized flow due to vehicle adaptation to the speed of the preceding vehicle. However, in the case under consideration this competition occurs within a random local perturbation, which is localized in the vicinity of the bottleneck (Fig. 5.15b,c).

The lower the speed \( v_{\text{random}}^{(\text{pert})} \) within the local perturbation at the bottleneck, the stronger the tendency towards speed adaptation. As a result, when the amplitude of this random perturbation \( \Delta v_{\text{random}}^{(\text{pert})} \) is greater than the critical amplitude \( \Delta v_{\text{cr, FS}}^{(\text{B})}(q_{on}) \), the perturbation grows and leads to a spontaneous F→S transition at the bottleneck (down-arrows “F→S” in Fig. 5.14 and down-arrow in Fig. 5.15b).

In contrast, the tendency towards the initial free flow due to over-acceleration is stronger at a higher speed within the perturbation. As a result, when the amplitude of this random perturbation \( \Delta v_{\text{random}}^{(\text{pert})} \) is less than the critical amplitude \( \Delta v_{\text{cr, FS}}^{(\text{B})}(q_{on}) \), the perturbation decays and no spontaneous F→S transition can occur, i.e., free flow remains at the bottleneck (up-arrows in Fig. 5.14 and up-arrow in Fig. 5.15c).

After an F→S transition occurs at the bottleneck, the speed adaptation effect causes subsequent maintenance of the emergent synchronized flow. Let us assume that there is a continuous vehicle inflow onto the main road from the on-ramp. These merging vehicles prevent acceleration of vehicles in this synchronized flow to free flow. In other words, the merging vehicles cause a permanent disturbance for vehicle acceleration within synchronized flow on
the main road. This permanent disturbance is localized on the main road in a small neighborhood of the bottleneck. As a result, the downstream front of the synchronized flow is fixed in the vicinity of the bottleneck (see also Sect. 7.4.3).

5.3.7 Time Delay of Speed Breakdown

It has already been mentioned that only if the amplitude of a random local perturbation exceeds the critical amplitude will that random perturbation grow and lead to a spontaneous F→S transition at a bottleneck. During a finite time interval random perturbations with different perturbation amplitudes can occur at the bottleneck. Because for the F→S transition the perturbation amplitude must exceed some distinct value, it can turn out that there is a time delay of this perturbation random occurrence. Thus, under the condition (5.46), one must usually wait a finite time $T^{(B)}_{FS}$ for the occurrence of a spontaneous F→S transition at the bottleneck. In other words, there is a time delay $T^{(B)}_{FS}$ for the spontaneous F→S transition at an on-ramp (Fig. 5.16).

![Fig. 5.16. Numerical simulation of a random time delay of the breakdown phenomenon (spontaneous F→S transition) in an initial free flow at an on-ramp. Uparrow shows the time of speed breakdown below a given speed level (dashed horizontal line). Taken from [331]](image)

The time delay $T^{(B)}_{FS}$ consists of two components:

(i) the random time of occurrence of a random component of the local perturbation whose amplitude is greater than the critical amplitude, precisely the time of occurrence of a local perturbation that begins to grow at the bottleneck, $\tau^{(B)}_{pert}$;

(ii) the time needed for this perturbation to grow to a synchronized flow state, $\tau^{(grow \ B)}_{pert}$:

$$T^{(B)}_{FS} = \tau^{(B)}_{pert} + \tau^{(grow \ B)}_{pert}.$$  

(5.61)

We denote the difference between the current amplitude of a random local perturbation at the bottleneck $\Delta v^{(pert)}_{random}$ (5.48) (Fig. 5.15b,c) and the critical perturbation amplitude $\Delta v^{(B)}_{cr,FS}$ (5.49) by
5.3 Breakdown Phenomenon at Freeway Bottlenecks

\[ \Delta v_{cr, FS}^{(pert)} = v_{cr, FS} - v_{random}^{(pert)} \quad (5.62) \]

We assume that

\[ \Delta v_{cr, FS}^{(pert)} > 0 \quad (5.63) \]

i.e., the initial perturbation grows (Fig. 5.15b).

On the one hand, the greater the difference \( \Delta v_{cr, FS}^{(pert)} \) (5.62), the smaller the time delay component \( \tau_{pert}^{(grow \, B)} \) should be. On the other hand, the greater the amplitude of the critical perturbation, the greater the time delay component \( \tau_{pert}^{(B)} \). Because the time delay \( T_{FS}^{(B)} \) depends on the both components (5.61), this is a very complex stochastic (random) characteristic of an F→S transition: in different realizations with the same flow rates \( q_{on} \) and \( q_{in} \) (and the same other initial conditions), the time delay can take on very different values.

One should recall that at the critical flow rate to the on-ramp \( q_{on}^{(determ, \, FS)} \) (Fig. 5.14a) a deterministic F→S transition occurs (Sect. 5.3.2): any random local perturbation within the deterministic perturbation at the bottleneck leads to synchronized flow emergence at the bottleneck. For this reason, in contrast to \( T_{FS}^{(B)} \) (5.61), the delay time of the deterministic F→S transition consists of one component:

\[ T_{determ, \, FS}^{(B)} = \tau_{determ}^{(grow \, B)} \quad (5.64) \]

where \( \tau_{determ}^{(grow \, B)} \) is the time required for average vehicle speed to decrease from the speed within the deterministic perturbation in free flow at the bottleneck to a synchronized flow speed at the bottleneck after the critical point (5.41) is reached.

**Mean Time Delay**

The mean time delay of a spontaneous F→S transition at the on-ramp, \( T_{FS}^{(B, \, mean)} \) must be determined by averaging random time delays \( T_{FS}^{(B)} \) (5.61) obtained in different realizations (e.g., different days where the spontaneous F→S transition at given flow rates \( q_{in} \) and \( q_{on} \) is realized at the on-ramp).

If we consider a flow rate \( q_{sum} \) close to the critical flow rate \( q_{sum} = q_{determ, \, FS}^{(B)} \) (5.59), i.e., the difference

\[ \Delta q = q_{determ, \, FS}^{(B)} - q_{sum} \quad (5.65) \]

is small, then relatively small random local perturbations can lead to an F→S transition at the bottleneck. In other words, the closer a given flow rate \( q_{sum} \) is to the critical value \( q_{sum} = q_{determ, \, FS}^{(B)} \), the smaller the amplitude of the random local perturbation at the bottleneck can be to lead to a spontaneous F→S transition. Thus, the mean time delay of the phase transition \( T_{FS}^{(B, \, mean)} \) should be a decreasing function of the flow rate \( q_{sum} \) (Fig. 5.17).
5.4 Conclusions

(i) The well-known breakdown phenomenon in traffic is associated with a local first-order phase transition from free flow to synchronized flow (F→S transition).

(ii) The existence of a critical (limit) flow rate in free flow is associated with an F→S transition rather than with moving jam emergence in free flow.

(iii) The critical flow rate in free flow downstream of a freeway bottleneck is less than the limit flow rate on a homogeneous freeway. This should explain why the breakdown phenomenon is mostly observed at a bottleneck.

(iv) The conclusion of item (iii) is related to the occurrence of a deterministic local perturbation in the vicinity of the bottleneck: within this deterministic perturbation the vehicle speed in free flow is less than the speed in free flow downstream of the bottleneck.

(v) The speed decrease within the deterministic local perturbation at the bottleneck causes a deterministic breakdown phenomenon (deterministic F→S transition). This speed decrease is also responsible for the fact that the deterministic breakdown phenomenon occurs at a lower flow rate and lower density than the flow rate and density required for an F→S transition away from the bottleneck. The deterministic breakdown phenomenon occurs in a deterministic way, i.e., even when random fluctuations (perturbations) are negligible.
(vi) Real random local perturbations in the vicinity of the bottleneck lead to the spontaneous breakdown phenomenon (spontaneous F→S transition) when the flow rate does not reach the flow rate at the critical point for free flow related to the deterministic F→S transition. This spontaneous breakdown phenomenon is realized if some random component of the perturbation at the bottleneck has amplitude greater than the critical amplitude.

(vii) If the flow rate is lower than the flow rate at the critical point, there is a time delay of a spontaneous F→S transition. The time delay is a random variable. The greater the flow rate $q_{\text{sum}}$, the shorter the mean time delay for a spontaneous F→S transition at the bottleneck.

(viii) There is a threshold point for an F→S transition. When the flow rate in free flow is lower than the threshold flow rate, the F→S transition cannot occur. This is regardless of the amplitude of a local time-limited random perturbation in free flow.
6 Moving Jam Emergence
in Three-Phase Traffic Theory

6.1 Introduction

Explaining the fascinating phenomenon of moving jam emergence in free traffic flow has always been one of the most important aims of traffic researchers.

In three-phase traffic theory, moving jam emergence in an initial free flow is explained by the sequence of two phase transitions. Firstly, the phase transition from the “free flow” phase to the “synchronized flow” phase should occur. Later, and usually at a different freeway location the phase transition from the “synchronized flow” phase to the “wide moving jam” phase is realized (F→S→J transitions).

This conclusion of three-phase traffic theory explains the empirical result (Sect. 4.4) that a moving jam cannot emerge spontaneously in the “free flow” phase. This seems to contradict another empirical result: free flow should be in a metastable state with respect to moving jam emergence in free flow (i.e., with respect to an F→J transition) when the flow rate in this flow \( q^{(\text{free})} \) satisfies the condition [166,203]

\[
q^{(\text{free})} \geq q_{\text{out}},
\]

(6.1)

where \( q_{\text{out}} \) is the flow rate in free flow formed in the wide moving jam outflow.

The condition (6.1) also appeared in the nonlinear theory of moving jams first developed in 1994 by Kerner and Konhäuser [367] within the scope of the fundamental diagram approach. As discussed in Sect. 3.2.7 (Fig. 3.5), under the condition (6.1) this theory predicts that free flow should be metastable with respect to moving jam emergence in free flow, i.e., in free flow the F→J transition should be possible.

The metastability of free flow with respect to an F→J transition [166, 203,367] means that if under the condition (6.1) a random local perturbation whose amplitude exceeds some critical amplitude for the F→J transition appears in an initial free flow, a wide moving jam must occur spontaneously in this free flow. Otherwise, if the amplitude of the local perturbation in free flow is less than the critical amplitude, the perturbation decays.

Both the nonlinear theory of moving jams within the scope of the fundamental diagram approach [367] and empirical studies [166,203] have shown that the threshold point for an F→J transition is determined by the flow rate
$q_{\text{out}}$ of the wide moving jam outflow. In other words, at the threshold point for an F$\rightarrow$J transition

$$q^{(\text{free})} = q_{\text{out}}, \quad \rho^{(\text{free})} = \rho_{\text{min}}, \quad v^{(\text{free})} = v_{\text{max}}, \quad (6.2)$$

where $\rho_{\text{min}}$ and $v_{\text{max}} = q_{\text{out}}/\rho_{\text{min}}$ are the density and vehicle speed in free flow, which are related to the flow rate $q_{\text{out}}$ of the wide moving jam outflow.

We know that at the threshold of a local phase transition, the critical amplitude of the local perturbation for this phase transition should have reached its maximum (Sect. A.7). If the flow rate increases more and more above this threshold $q_{\text{out}}$, i.e., if

$$q^{(\text{free})} > q_{\text{out}} \quad (6.3)$$

and the difference

$$q^{(\text{free})} - q_{\text{out}} \quad (6.4)$$

gradually increases, then the amplitude of the critical perturbation should decrease.

Thus, one would expect that at some high enough flow rate $q^{(\text{free})}$ (6.3) above the threshold $q_{\text{out}}$, the critical amplitude of the local perturbation for a spontaneous F$\rightarrow$J transition should tend to zero. This would mean that the wide moving jam should occur spontaneously in free flow at some high enough critical flow rate (3.28) (the critical density (3.27)) in this free flow. This theoretical conclusion is one of the main conclusions of the fundamental diagram approach for any traffic flow theory and model that claims to explain spontaneous moving jam emergence in free flow (see Sect. 3.2). It is, however, in contradiction with empirical results: the spontaneous F$\rightarrow$J transition is not observed in real free flow (see the related critical discussion in Sect. 3.3.2).

In models and theories based on the fundamental diagram approach, the critical flow rate (3.28) (the critical density (3.27)) for a spontaneous F$\rightarrow$J transition is related to the critical flow rate in free flow (5.2). This means the following:

(i) The critical amplitude of the critical perturbation required for moving jam emergence in free flow is zero at the critical point of free flow (5.2).

(ii) The spontaneous F$\rightarrow$J transition determines the existence of the critical point of free flow (5.2). This theoretical conclusion of all theories and models in the fundamental diagram approach that claim to predict moving jam emergence is inconsistent with the empirical results (5) and (6) listed in Sect. 4.4.

In this chapter, we address the question: Under what conditions do moving jams emerge spontaneously in traffic flow? One of the main aims is to show how three-phase traffic theory [205, 208–212, 218] resolves the contradiction between the empirical metastability of free flow with respect to an F$\rightarrow$J transition under the condition (6.1) and another empirical fact – that no spontaneous F$\rightarrow$J transition is observed.
6.2 Wide Moving Jam Emergence in Free Flow

In three-phase traffic theory, free flow is also in a metastable state within the density range (3.17).

However, to resolve the contradiction between theories and models based on the fundamental diagram approach and empirical observations, in three-phase traffic theory the following hypothesis is suggested. At any given flow rate (density) in free flow, the critical amplitude of a local perturbation required for moving jam emergence (i.e., for an $F \rightarrow J$ transition) is considerably greater than the critical perturbation required for synchronized flow emergence (i.e., for an $F \rightarrow S$ transition) in the same free flow [205, 209–211]. In particular, as discussed in Sect. 5.2.4, the critical point of free flow (5.2) is related to the critical point of an $F \rightarrow S$ transition, rather than to an $F \rightarrow J$ transition.

To explain this hypothesis, we assume that the $F \rightarrow S$ transition occurs if a local perturbation in speed causes the probability of passing to fall below certain critical value (curve $P_{cr}$ in Fig. 5.6b). The critical amplitude of perturbation required to significantly reduce the probability of passing can be considerably less than that needed to cause the $F \rightarrow J$ transition. Indeed, the initial perturbation must only cause the probability of passing to fall. In contrast, such a reduction of the probability of passing is not enough for the $F \rightarrow J$ transition to occur. In this case, the perturbation must grow up to the speed within the perturbation as low as zero to lead to wide moving jam emergence in free flow.

This hypothesis of three-phase traffic theory means that even at the critical point of free flow (5.2) the critical amplitude of the critical perturbation for moving jam emergence in free flow on a homogeneous road is a relatively large finite value, rather than zero [207, 210]. The same conclusion is valid for a freeway bottleneck: even at the critical point of free flow at the bottleneck (5.59), the critical amplitude of the critical perturbation for moving jam emergence in free flow at the bottleneck is a relatively large finite value, rather than zero (see Sect. 6.5.2).

To explain this hypothesis of three-phase traffic theory, we first reconsider Fig. 5.4a, which used to explain the breakdown phenomenon ($F \rightarrow S$ transition) in Sect. 5.2.4. We present the line $J$ in the flow–density plane, together with states of free flow and steady states of synchronized flow, as it shown in Fig. 6.1a. Here we have used empirical points (3) and (4) of Sect. 4.4. The latter point (4) enables us to draw the line $J$ through the multitude of steady states of synchronized flow. Thus, the line $J$ divides these steady states on two classes: those of synchronized flow, which lie above the line $J$, and those that lie below the line $J$. Note that the line $J$ in the flow–density plane (Fig. 6.1a) is related to the hyperbolic function in the speed–density plane that is given by the equation (3.15) (curve $J$ in Fig. 6.1b).

The hypothesis of three-phase traffic theory under consideration can be explained if it is assumed that besides the critical point of an $F \rightarrow S$ transition
Why moving jams do not emerge in free flow on a homogeneous road in three-phase traffic theory [208]. (a) Qualitative concatenation of states of free flow (curve $F$) and steady states of synchronized flow (hatched region) with the line $J$ in the flow-density plane. (b) States in the speed-density plane related to (a), where the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ is shown. States of free flow (curve $F$), steady states of synchronized flow (hatched region), and the critical branch $v_{cr}^{(FS)}(\rho^{\text{free}})$ are taken from Fig. 5.4 (5.2), (5.3), which determines the critical point of free flow on a homogeneous road, there is another critical point in free flow. At the latter critical point (the point $\rho^{\text{free}} = \rho_{\text{max, FJ}}^{\text{free}}$ among the states of free flow in Fig. 6.1b), the critical amplitude of the critical perturbation for moving jam emergence in free flow is zero. However, the density in free flow related to this critical density $\rho^{\text{free}} = \rho_{\text{max, FJ}}^{\text{free}}$ is much greater than the limiting density $\rho^{\text{free}} = \rho_{\text{max}}^{\text{free}}$ at which synchronized flow must occur in the initial free flow. The latter means that the critical point $\rho^{\text{free}} = \rho_{\text{max, FJ}}^{\text{free}}$ of free flow in Fig. 6.1b lies in some hypothetical subset of the states of free flow above the limiting density $\rho_{\max}^{\text{free}}$. These hypothetical free flow states and the critical
6.2 Wide Moving Jam Emergence in Free Flow

hypothetical point \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max, FJ}} \) cannot be reached in reality, because synchronized flow must occur at a much lower vehicle density in free flow \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max}} \). For this reason, no states of free flow between \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max}} \) and \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max, FJ}} \) are realizable. These hypothetical inaccessible free flow states are shown by the dashed curve \( F' \) in Fig. 6.1b.

At the critical point of these inaccessible free flow states \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max, FJ}} \), the critical amplitude of the critical perturbation for moving jam emergence in free flow is zero. This means that there should be a critical branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) in the speed–density plane (Fig. 6.1b). The critical branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) gives the speed within the critical perturbation as a function of density. For a given density, the corresponding point on the branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) determines the critical amplitude of the critical local perturbation \( \Delta v^{(\text{FJ})}_{\text{cr}} \) for an \( F \rightarrow \text{J} \) transition:

\[
\Delta v^{(\text{FJ})}_{\text{cr}} = v^{(\text{free})} - v^{(\text{FJ})}_{\text{cr}}. \tag{6.5}
\]

The qualitative shape of the branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) in Fig. 6.1b can be explained as follows. If the density \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max, FJ}} \), the critical amplitude of the critical perturbation (6.5) for moving jam emergence in free flow is zero:

\[
\Delta v^{(\text{FJ})}_{\text{cr}} = 0. \tag{6.6}
\]

From this condition and (6.5) we obtain

\[
v^{(\text{free})} = v^{(\text{FJ})}_{\text{cr}} \at \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max, FJ}}. \tag{6.7}
\]

Therefore, at the density \( \rho^{(\text{free})} = \rho^{(\text{free})}_{\text{max, FJ}} \) the branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) will merge with the branch of the inaccessible states of free flow \( F' \).

At lower density

\[
\rho^{(\text{free})} < \rho^{(\text{free})}_{\text{max, FJ}} \tag{6.8}
\]

the critical amplitude of the critical perturbation for moving jam emergence in free flow should be greater than zero: the lower the density of free flow in comparison with \( \rho^{(\text{free})}_{\text{max, FJ}} \), the greater the critical amplitude \( \Delta v^{(\text{FJ})}_{\text{cr}} \) (6.5). Thus, the critical branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) will be associated with lower speeds as the density in free flow gradually decreases.

This decrease in speed on the branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) has the limit \( v^{(\text{FJ})}_{\text{cr}} = 0 \). This limit is related to the speed within a wide moving jam \( v_{\text{min}} = 0 \). At the threshold density of free flow for wide moving jam emergence (6.2), a local perturbation is required, in which the speed is equal to that within a wide moving jam:

\[
v_{\text{min}} = 0. \tag{6.9}
\]

Thus, at this threshold density (6.2) the critical branch \( v^{(\text{FJ})}_{\text{cr}}(\rho^{(\text{free})}) \) will merge with the horizontal line associated with the speed \( v_{\text{min}} = 0 \) within a wide moving jam (Fig. 6.1b).
Thus, we see from Fig. 6.1b that the critical amplitude of the local perturbation for moving jam emergence in an initial free flow $\Delta v_{\text{cr}}^{(FJ)}$ (6.5) is greater at any density in free flow than the critical amplitude of the local perturbation for synchronized flow emergence in free flow $\Delta v_{\text{cr}}^{(FS)}$ (5.5). This explains why a spontaneous emergence of moving jams is not observed in free flow: the probability of the random occurrence in free flow of the much larger perturbation required for an $F \rightarrow J$ transition in comparison to that required for an $F \rightarrow S$ transition is extremely low.

This is illustrated in Fig. 6.2. Independent of the density in free flow above the threshold point for an $F \rightarrow S$ transition (5.10), the critical amplitude of the local perturbation for an $F \rightarrow S$ transition (curve $F_S$ in Fig. 6.2b) is less than that for an $F \rightarrow J$ transition (curve $F_J$). This can also be seen directly if the critical branches $v_{\text{cr}}^{(FS)}(\rho^{\text{free}})$ and $v_{\text{cr}}^{(FJ)}(\rho^{\text{free}})$ in Fig. 6.1b are compared; these branches give the speeds required within a critical perturbation for an $F \rightarrow S$ transition or an $F \rightarrow J$ transition, respectively.

Consequently, under the same conditions, the probability $P_{FS}$ for an $F \rightarrow S$ transition (curve $P_{FS}$ in Fig. 6.2c) is much greater than the probability $P_{FJ}$ for an $F \rightarrow J$ transition (curve $P_{FJ}$).\footnote{The probability $P_{FJ}$ is defined as follows. We consider a large number of realizations, $N_{FJ}$, where a spontaneous $F \rightarrow J$ transition in an initial free flow is studied. Each realization should refer to the same flow rate of free flow, other initial conditions, the same time interval $T_{ob}$ for observing the spontaneous $F \rightarrow J$ transition on a chosen freeway section of length $L_{ob}$. Let us assume that the $F \rightarrow J$ transition occurs in $n_{FJ}$ of these $N_{FJ}$ realizations. Then the probability for this transition is

$$P_{FJ} = \frac{n_{FJ}}{N_{FJ}}.$$}

6.3 Wide Moving Jam Emergence in Synchronized Flow

The limit point of free flow is thus associated with an $F \rightarrow S$ transition rather than to an $F \rightarrow J$ transition. This $F \rightarrow S$ transition induces a region of synchronized flow in the initial free flow. The $F \rightarrow S$ transition occurs mostly at freeway bottlenecks (Sect. 5.3). In this usual case, synchronized flow appears at a bottleneck. How do moving jams emerge in synchronized flow?

6.3.1 Hypothesis for Moving Jam Emergence in Synchronized Flow

In three-phase traffic theory, the following hypothesis holds for moving jam emergence in synchronized flow [205,208]:

$$P_{FJ} = \frac{n_{FJ}}{N_{FJ}}. \quad (6.10)$$

Strictly speaking, the exact probability for the spontaneous $F \rightarrow J$ transition is the limit as $N_{FJ} \rightarrow \infty$. 
Fig. 6.2. Hypotheses of three-phase traffic theory [209, 210]. (a) States of free (curve $F$) and synchronized flow (hatched region), which are the same as those in Fig. 4.4. (b) Qualitative dependencies of the amplitude of the critical speed local perturbation $\Delta v_{ct}^{(FS)}$ on density. (c) Qualitative dependencies of the probability for phase transitions. In (b, c) curves $F_S$ and $P_{FS}$ are related to an $F\rightarrow S$ transition and curves $F_J$ and $P_{FJ}$ are related to an $F\rightarrow J$ transition. The curves $F_S$ and $P_{FS}$ in (b, c) are taken from Figs. 5.3b,c, respectively.
(i) The line $J$ (as in Fig. 6.3a) determines the threshold for wide moving jam existence and excitation in synchronized flow. In other words, all (an infinite number!) steady states of synchronized flow associated with the line $J$ in the flow–density plane are threshold states with respect to wide moving jam emergence, i.e., with respect to an $S \rightarrow J$ transition.

(ii) This means that the line $J$ separates all steady states of synchronized flow into two qualitatively different classes:

(a) In states associated with points in the flow–density plane below the line $J$ (see axes in Fig. 6.3a), no wide moving jams can either continue to exist or be excited.

(b) States associated with points in the flow–density plane on or above the line $J$ are metastable states with respect to wide moving jam emergence. Only local perturbations of traffic variables whose amplitude exceeds some critical value grow and can lead to wide moving jams (up-arrows, curves $S_J^{(1)}$ and $S_J^{(2)}$ in Fig. 6.3b); otherwise wide moving jams do not occur (down-arrows, curves $S_J^{(1)}$ and $S_J^{(2)}$ in Fig. 6.3b). These critical local perturbations act as nucleation centers for wide moving jam emergence in synchronized flow.

To understand this hypothesis, note that wide moving jams cannot be formed in any states of synchronized flow situated below the line $J$. Let us assume that a state of synchronized flow directly upstream of a wide moving jam is associated with a point $k$ in the flow–density plane. This point is below the line $J$ (Fig. 6.4a,b). Because the velocity of the upstream front of the wide moving jam $v_g^{(\text{up})}$ equals the slope of the line $K$ (from a point $k$ in free flow to the point $(\rho_{\text{max}}, 0)$), the absolute value $|v_g^{(\text{up})}|$ is always less than that of the downstream front $v_g$ determined by the slope of the line $J$, i.e., the formula $|v_g^{(\text{up})}| < |v_g|$ (3.20) is valid. Therefore, the width of the wide moving jam gradually decreases.

On the other hand, assume that a state of synchronized flow upstream of another wide moving jam is associated with a point $n$ in the flow–density plane. This state is above the line $J$ (Fig. 6.4c,d). In this case, the velocity of the upstream front of the wide moving jam $v_g^{(\text{up})}$ equals the slope of the line $N$ (from a point $n$ in synchronized flow to the point $(\rho_{\text{max}}, 0)$), i.e., the absolute value $|v_g^{(\text{up})}|$ is always higher than that of the downstream front $|v_g|$, i.e., the formula $|v_g^{(\text{up})}| > |v_g|$ (3.21) is valid. Therefore, the width of the wide moving jam in Fig. 6.4c should gradually increase. For these reasons, wide moving jams can be formed in states of synchronized flow that lie on or above the line $J$, i.e., these states are metastable with respect to $S \rightarrow J$ transitions.

As in the case of wide moving jam emergence in free flow (see (3.22) in Sect. 3.2.7), the condition

$$v_g = v_g^{(\text{up})}$$

(6.11)
Fig. 6.3. Nucleation effect for jam emergence in synchronized flow in three-phase traffic theory [208]. (a) Two different constant synchronized flow speeds, \( v_1^{(\text{syn})} \) and \( v_2^{(\text{syn})} \), in the flow–density plane. (b) Critical amplitude of critical local perturbations for an S→J transition associated with these two different synchronized flow speeds \( v_1^{(\text{syn})} \) and \( v_2^{(\text{syn})} \) in (a), curves \( S_{J}^{(1)} \) and \( S_{J}^{(2)} \), respectively. (c) Probability for an S→J transition associated with the speed \( v_1^{(\text{syn})} \) (curve \( P_{SJ}^{(1)} \)) and with the speed \( v_2^{(\text{syn})} \) (curve \( P_{SJ}^{(2)} \)). In (a), states of free (curve \( F \)), steady states of synchronized flow (hatched region), and the line \( J \) are taken from Fig. 6.1a.
Fig. 6.4. Metastability of synchronized flow with respect to moving jam emergence [208]. (a, c) Qualitative forms of wide moving jams at two different densities in synchronized flow upstream of the jams, $\rho^{(\text{syn})} = \rho_k^{(\text{syn})}$ (a) and $\rho^{(\text{syn})} = \rho_n^{(\text{syn})}$ (c). (b, d) Representation of the propagation of downstream wide moving jam fronts (line J) and upstream wide moving jam fronts (lines K (b) and N (d)) in the flow–density plane for the wide moving jams in (a) and (c), respectively. In (b, d) states for free flow (curve F), steady states of synchronized flow (dashed region) and the line J are taken from Fig. 6.1a.

determines threshold points with respect to wide moving jam emergence in synchronized flow. However, in the case of synchronized flow the velocity $v_g$ of the downstream front of a wide moving jam and the velocity $v_g^{(\text{up})}$ of the upstream front of the wide moving jam, corresponding to the Stokes shock-wave formula (3.5), are found from

$$v_g = -\frac{q_{\text{out}}^{(\text{syn})}}{\rho_{\text{max}}^{(\text{syn})} - \rho_{\text{min}}^{(\text{syn})}}$$  \hspace{1cm} (6.12)

and

$$v_g^{(\text{up})} = -\frac{q^{(\text{syn})}}{\rho_{\text{max}} - \rho^{(\text{syn})}}$$  \hspace{1cm} (6.13)
6.3 Wide Moving Jam Emergence in Synchronized Flow

respectively. In (6.12) and (6.13), \( q_{\text{out}}^{\text{(syn)}} \) and \( \rho_{\text{min} J}^{\text{(syn)}} \) are the flow rate and density in synchronized flow in the jam outflow; \( q^{\text{(syn)}} \) and \( \rho^{\text{(syn)}} \) are the flow rate and density in synchronized flow in the jam inflow.

As explained in the hypothesis above moving jam emergence in synchronized flow, \( q_{\text{out}}^{\text{(syn)}} \) and \( \rho_{\text{min} J}^{\text{(syn)}} \) are related to a point on the line \( J \) in the flow-density plane, i.e., they are associated with only one of the infinite number of different threshold points for moving jam emergence in synchronized flow. If (3.10) is substituted into (6.12), we obtain

\[
q_{\text{out}}^{\text{(syn)}} = \frac{1}{\tau_{\text{del}}^{(a)}} \left( 1 - \frac{\rho_{\text{min} J}^{\text{(syn)}}}{\rho_{\text{max}}} \right).
\]  

Equation (6.14) yields the line \( J \) in the flow-density plane at the flow rate \( q^{\text{(syn)}} = q_{\text{out}}^{\text{(syn)}} \) and density \( \rho^{\text{(syn)}} = \rho_{\text{min} J}^{\text{(syn)}} \) in the jam outflow.

From (6.14) it can be seen that the time delay \( \tau_{\text{del}}^{(a)} \) in vehicle acceleration is responsible for the threshold point of wide moving jam emergence in synchronized flow. The same result has been discussed in Sect. 3.2.6 when the relation (3.12) has been considered that is related to wide moving jam emergence in free flow. A difference between (6.14) and (3.12) is as follows. There are an infinite number of solutions for threshold flow rates \( q_{\text{out}}^{\text{(syn)}} \) (and for threshold densities \( \rho_{\text{min} J}^{\text{(syn)}} \)). These solutions are related to (6.14), i.e., to the case of wide moving jam emergence in synchronized flow. All these infinite number of solutions correspond to the line \( J \) in the flow density plane. This is related to the fact that the condition (6.11) separates metastable states of synchronized flow (in which the condition (3.21) for moving jam emergence in synchronized flow is valid) and stable states of synchronized flow (in which the opposite condition (3.20) is valid).

6.3.2 Features of Metastable Synchronized Flow States

To explain the empirical results (8) and (9) of Sect. 4.4, we consider two dependencies of the critical amplitude of a local perturbation for an S→J transition (Fig. 6.3a,b). These dependencies are related to two different vehicle speeds in synchronized flow, \( v_1^{\text{(syn)}} \) and \( v_2^{\text{(syn)}} \) where \( v_1^{\text{(syn)}} > v_2^{\text{(syn)}} \) (curves \( S_1^{(1)} \) and \( S_1^{(2)} \) for \( v_1^{\text{(syn)}} \) and \( v_2^{\text{(syn)}} \), respectively). For each of these two speeds there is a threshold density for an S→J transition. The threshold density is related to the intersection of the line of a constant speed with the line \( J \) in the flow-density plane. Thus, for the speed \( v_1^{\text{(syn)}} \) the threshold density is \( \rho_{\text{min} J, 1}^{\text{(syn)}} \) and for the speed \( v_2^{\text{(syn)}} \) the threshold density is \( \rho_{\text{min} J, 2}^{\text{(syn)}} \) (Fig. 6.3).

Let us also consider two synchronized flow states above the line \( J \) (Fig. 6.3a). These states are metastable with respect to moving jam emergence. We assume that the first state is related to the speed \( v = v_1^{(\text{syn})} \). The density in this state is \( \rho = \rho_1^{(\text{syn})} \). The second state is related to the speed
$v = v_2^{(\text{syn})}$. The density in this state is $\rho = \rho_2^{(\text{syn})}$. We choose the same difference between the densities $\rho_1^{(\text{syn})}$, $\rho_2^{(\text{syn})}$ and the threshold densities $\rho_{\text{min} \ j, 1}$, $\rho_{\text{min} \ j, 2}$ associated with the speeds $v_1^{(\text{syn})}$ and $v_2^{(\text{syn})}$, respectively:

$$\rho_1^{(\text{syn})} - \rho_{\text{min} \ j, 1} = \rho_2^{(\text{syn})} - \rho_{\text{min} \ j, 2}. \quad (6.15)$$

Under the condition (6.15), for synchronized flow states that are metastable with respect to moving jam emergence the following hypothesis is suggested in three-phase traffic theory [209]:

- the lower the speed in synchronized flow, the lower the critical amplitude of a local density perturbation $\Delta \rho_{\text{cr}}^{(\text{SJ})}$ (Fig. 6.3b) (critical amplitude of a local speed perturbation) required for an $\text{S} \rightarrow \text{J}$ transition. Therefore, the lower the initial speed in synchronized flow, the greater the probability for $\text{S} \rightarrow \text{J}$ transition (wide moving jam emergence) in synchronized flow $P_{\text{SJ}}$ (Fig. 6.3c).\(^2\)

This is reflected in the density dependencies of the probability for an $\text{S} \rightarrow \text{J}$ transition in synchronized flows for these two different vehicle speeds, $v_1^{(\text{syn})}$ and $v_2^{(\text{syn})}$ (curves $P_{\text{SJ}}^{(1)}$ and $P_{\text{SJ}}^{(2)}$ in Fig. 6.3c).

### 6.3.3 Stable High Density Synchronized Flow States

As follows from the hypothesis for moving jam emergence in synchronized flow (Sect. 6.3.1), synchronized flow states associated with points in the flow–density plane below the line $J$ are stable against moving jam emergence. Among these synchronized flow states, there are also states in which the density is very high and the speed is low (e.g., a state labeled “stable” in Fig. 6.5). We can see that at a given low synchronized flow speed $v_3^{(\text{syn})}$ a synchronized flow state that is stable against moving jam emergence is related to a lower density $\rho_3^{(\text{syn})}$ than the density $\rho_4^{(\text{syn})}$ in a metastable synchronized flow state (this metastable state with respect to moving jam emergence is

\[^2\] The probability $P_{\text{SJ}}$ is defined as follows. We consider a large number of realizations, $N_{\text{SJ}}$, where a spontaneous $\text{S} \rightarrow \text{J}$ transition in a synchronized flow region is studied. Each realization should refer to the same flow rate of synchronized flow, other initial conditions, the same time interval $T_{\text{ob}}$ for observing the spontaneous $\text{S} \rightarrow \text{J}$ transition on a chosen freeway section of length $L_{\text{syn}}$ associated with the synchronized flow region. Let us assume that the $\text{S} \rightarrow \text{J}$ transition occurs in $n_{\text{SJ}}$ of these $N_{\text{SJ}}$ realizations. Then the probability for this transition is

$$P_{\text{SJ}} = \frac{n_{\text{SJ}}}{N_{\text{SJ}}}. \quad (6.16)$$

Strictly speaking, the exact probability for the spontaneous $\text{S} \rightarrow \text{J}$ transition is the limit as $N_{\text{SJ}} \rightarrow \infty$. 
6.3 Wide Moving Jam Emergence in Synchronized Flow

Fig. 6.5. Qualitative illustration of high density and low speed synchronized flow states. States for free flow (curve F), steady states of synchronized flow (dashed region), and the line $J$ are taken from Fig. 6.1a.

labeled “metastable” in Fig. 6.5a). At a given high density $\rho_3^{(\text{syn})}$ a synchronized flow state that is stable against moving jam emergence is related to lower speed $v_3^{(\text{syn})}$ than the speed $v_4^{(\text{syn})}$ in a metastable synchronized flow state (Fig. 6.5b).

In accordance with the discussion in Sect. 4.3.4, the existence of high density and low speed synchronized flow states that are stable against moving jam emergence means that there can also be dynamic synchronized flow states of very high density and low speed where moving jams cannot emerge. These dynamic synchronized flow states that are stable against moving jam emergence are related to points in the flow–density plane that are below the line $J$.

However, such high density dynamic synchronized flow states in which no moving jams emerge have not been found in empirical observations up to now. The latter is probably because usually high density dynamic synchronized flow states result from the pinch effect in synchronized flow upstream of bottlenecks. In this case, states of synchronized flow are associated with points in the flow–density plane that are above the line $J$ (see Sect. 7.6.2). In other words, these high density dynamic synchronized flow states are metastable with respect to moving jam emergence. Nevertheless, in three-phase traffic theory there can also be dynamic synchronized flow states of very high density and low speed that are stable against moving jam emergence, i.e., in these synchronized flow states moving jams do not appear. We can expect that such stable dynamic synchronized flow states of very high density and low speed will be found in the future in empirical data.\(^3\)

\(^3\) We should note that stable high density and low speed synchronized flow states, which are stable against moving jam emergence, can be metastable with respect to an S→F transition. This remark is related to both hypothetical steady states and dynamic states of synchronized flow.
Critical Branches of Speed for $S \rightarrow J$ Transition

The above assumptions of three-phase traffic theory can also be illustrated in the speed–density plane (Fig. 6.6) [221]. In the speed–density plane, for any speed in synchronized flow (for example, $v_1^{(\text{syn})}$ and $v_2^{(\text{syn})}$ in Fig. 6.6, which are the same as the corresponding speeds in Fig. 6.3a) there is the critical branch of the speed for an $S \rightarrow J$ transition ($v_{cr, 1}^{(SJ)}$ and $v_{cr, 2}^{(SJ)}$ for the speeds $v_1^{(\text{syn})}$ and $v_2^{(\text{syn})}$, respectively).

![Diagram of speed-density plane for $S \rightarrow J$ transition](image)

**Fig. 6.6.** Critical branches $v_{cr, 1}^{(SJ)}$ and $v_{cr, 2}^{(SJ)}$ for an $S \rightarrow J$ transition at two different steady synchronized flow speeds $v_1^{(\text{syn})}$ and $v_2^{(\text{syn})}$, respectively [221]. States for free flow (curve $F$), steady states of synchronized flow (dashed region), and the curve $J$ are taken from Fig. 6.1b.

The threshold densities $\rho_{\text{min J, 1}}^{(\text{syn})}$ and $\rho_{\text{min J, 2}}^{(\text{syn})}$ for an $S \rightarrow J$ transition, which correspond to the speeds $v_1^{(\text{syn})}$ and $v_2^{(\text{syn})}$ (Fig. 6.3), can also be found in the speed–density plane (Fig 6.6). In this case, every threshold density corresponds to an intersection point of a horizontal line of constant speed and the curve $J$. 
In general, the critical amplitude of the critical speed perturbation for an S→J transition is

\[ \Delta v_{cr}^{(SJ)} = v^{(syn)} - v_{cr}^{(SJ)} \].

(6.17)

In (6.17), \(v^{(syn)}\) is one of the infinity of possible synchronized flow speeds; \(v_{cr}^{(SJ)}\) is the speed within the critical local perturbation for an S→J transition in the initial state of synchronized flow with speed \(v^{(syn)}\).

The critical speed \(v_{cr}^{(SJ)}\) in (6.17) is a function of density. This function \(v_{cr}^{(SJ)}(\rho)\) is called the “critical branch” of the speed for an S→J transition. To explain this term, let us consider a state of synchronized flow with a speed \(v^{(syn)}\) where an initial local perturbation occurs. The amplitude of this perturbation is

\[ \Delta v_{initial}^{(pert)} = v^{(syn)} - v_{initial}^{(pert)} \],

(6.18)

where \(v_{initial}^{(pert)}\) is the initial speed within the perturbation. If this speed is lower than \(v_{cr}^{(SJ)}\), in accordance with (6.17), we obtain

\[ \Delta v_{initial}^{(pert)} > \Delta v_{cr}^{(SJ)} \].

(6.19)

In this case, the amplitude of the perturbation is greater than the critical amplitude. As a result, the perturbation grows and leads to an S→J transition. In contrast, when the speed \(v_{initial}^{(pert)}\) is higher than \(v_{cr}^{(SJ)}\), we obtain

\[ \Delta v_{initial}^{(pert)} < \Delta v_{cr}^{(SJ)} \].

(6.20)

In this case, the perturbation decays, i.e., an S→J transition does not occur.

Because there are infinitely many different speeds in synchronized flow (Sect. 6.3), there are also infinitely many threshold densities for these different synchronized flow speeds, and a corresponding infinity of critical branches \(v_{cr}^{(SJ)}\) that are functions of density (only two of them, \(v_{cr,1}^{(SJ)}\) and \(v_{cr,2}^{(SJ)}\), are shown in Fig. 6.6 for the speeds \(v_{1}^{(syn)}\) and \(v_{2}^{(syn)}\), respectively).

**Threshold Density for S→J Transition**

In discussing an F→S transition (breakdown phenomenon), we have mentioned that for a first-order phase transition, the maximum critical amplitude of the local perturbation for a phase transition is related to the threshold point of the phase transition (Sect. 5.2.4). In the case of the S→J transition under consideration, this maximum perturbation amplitude corresponds to the threshold density that determines the threshold point for the S→J transition at the given speed in synchronized flow.

Thus, at the threshold density the critical branch \(v_{cr}^{(SJ)}\) should merge with the horizontal line \(v_{min} = 0\) related to the speed within a wide moving jam (Fig. 6.6).
Critical Point for S→J Transition

The minimum critical amplitude of the critical perturbation for an S→J transition is

$$\Delta v_{cr}^{(SJ)} = 0 .$$  \hspace{1cm} (6.21)

This minimum critical amplitude of a local perturbation is related to the critical point for an S→J transition. At this critical point an S→J transition must occur. This is because in real synchronized flow there are always local random perturbations of finite amplitudes that exceed the critical amplitude (6.21). At the critical point, corresponding to the value of $\Delta v_{cr}^{(SJ)}$ (6.17) the condition (6.21) leads to

$$v^{(syn)} = v_{cr}^{(SJ)} .$$  \hspace{1cm} (6.22)

However, there are infinitely many speeds in synchronized flow that can satisfy this condition. Thus, the condition (6.22) determines the infinite number of the critical points for an S→J transition for different speeds in synchronized flow.

To find the multitude of the infinite synchronized flow speeds that satisfy the condition (6.22), let us consider a given synchronized flow speed $v = v^{(syn)} . \hspace{1cm} (6.23)$

We assume that this speed is related to a horizontal line in the speed-density plane that intersects the boundary $S_{upper}$ above the curve $J$. It can be seen from Fig. 6.6 that in this case the condition (6.22) can be satisfied. The condition (6.22) also means that at the critical point for an S→J transition the critical branch $v_{cr}^{(SJ)}$ merges with this horizontal line for the speed (6.23).

In contrast, if the speed (6.23) is related to a horizontal line in the speed-density plane, which intersects the boundary $S_{upper}$ below the curve $J$, the condition (6.22) cannot be satisfied.

Thus, a synchronized flow speed satisfies the condition (6.22) only if the speed is related to a horizontal line in the speed-density plane that intersects the boundary $S_{upper}$ above the curve $J$. There are an infinite number of these synchronized flow speeds and the related critical points for an S→J transition.

For each of these infinite critical points there is the critical branch $v_{cr}^{(SJ)}$. This branch merges with a horizontal line for the related synchronized flow speed. Examples are the horizontal lines that are related to the speeds $v = v_{1}^{(syn)}$ and $v = v_{2}^{(syn)}$ in Fig. 6.6. The horizontal lines for the speeds $v = v_{1}^{(syn)}$ and $v = v_{2}^{(syn)}$ merge at the related critical points with the critical branches $v_{cr, 1}(\rho)$ and $v_{cr, 2}(\rho)$, respectively.

The critical points of synchronized flow with respect to an S→J transition under consideration are related to the well-known over-deceleration effect (Sect. 3.2.4). If the preceding vehicle begins to decelerate unexpectedly, due to the driver time delay the driver decelerates stronger than it is required to avoid collisions.
At the upper boundary of steady states of synchronized flow (the boundary $S_{\text{upper}}$ in Fig. 6.6; see Sect. 4.3.1), the critical amplitude of the local perturbation for an $S\rightarrow J$ transition can vanish, but does not necessarily do so. For this reason, in Fig. 6.3b at the boundary $S_{\text{upper}}$, the critical amplitudes are not shown to vanish.\(^4\) The latter corresponds to the assumed stability of steady states of synchronized flow with respect to infinitesimal perturbations (Sect. 4.3.3) [205]. For this reason, it can turn out that the merge point between the critical branch $v_{\text{cr}}^{(SJ)}$ with the line for the related synchronized flow speed deviates to a higher density from the density on the boundary $S_{\text{upper}}$. Examples are shown in Fig. 6.6, where lines corresponding to the constant speeds $v = v_{\text{syn}}^{(1)}$ and $v = v_{\text{syn}}^{(2)}$ merge with the associated critical branches $v_{\text{cr}}^{(SJ), 1}$ and $v_{\text{cr}}^{(SJ), 2}$, respectively.

The synchronized flow states, the critical branches $v_{\text{cr}}^{(SJ)}$, and the line $v_{\text{min}} = 0$ for the speed within a wide moving jam together yield a Z-shaped characteristic for an $S\rightarrow J$ transition (Fig. 6.6). Thus, this Z-characteristic for the $S\rightarrow J$ transition consists of

(i) a 2D region of steady states of synchronized flow (dashed region);
(ii) an infinite number of critical branches $v_{\text{cr}}^{(SJ)}(\rho)$, each of which gives the speed within the critical local perturbation for a spontaneous $S\rightarrow J$ transition as a function of density for a given synchronized flow speed (two of these critical branches $v_{\text{cr}}^{(SJ), i(\rho)}$, $i = 1, 2$ for the related two speeds of synchronized flow $v_{\text{syn}, i}$, $i = 1, 2$ are shown in Fig. 6.6);
(iii) the speed $v_{\text{min}} = 0$ within wide moving jams.

### 6.4.2 Cascade of Two Phase Transitions ($F\rightarrow S\rightarrow J$ Transitions)

Now we add to Fig. 6.6 the critical branch for an $F\rightarrow S$ transition $v_{\text{cr}}^{(FS)}$, as considered in Chap. 5 (Fig. 5.4b). Then we obtain a double Z-shaped density dependence of the speed (Fig. 6.7).

The double Z-characteristic at the bottleneck (Fig. 6.7) consists of

(a) states of free flow (curve $F$) as a function of density;
(b) the critical branch $v_{\text{cr}}^{(FS)}(\rho(\text{free}))$, which gives the speed within the critical local perturbation for a spontaneous $F\rightarrow S$ transition as a function of density;
(c) a 2D region of steady states of synchronized flow (dashed region);

\(^4\) Note that in models based on three-phase traffic theory discussed in Chap. 16, another case is considered, where the condition (6.21) is satisfied at the upper boundary of steady states of synchronized flow $S_{\text{upper}}$. The critical amplitude of the local perturbation for an $S\rightarrow J$ transition vanishes at the boundary $S_{\text{upper}}$. In this case, a part of the boundary $S_{\text{upper}}$ right of the curve $J$ in the speed–density plane determines an infinite number of the critical points for an $S\rightarrow J$ transition related to different synchronized flow speeds.
Fig. 6.7. Double Z-shaped speed–density characteristic in three-phase traffic theory [221]. The critical branches for an F→S transition and for an S→J transitions in the speed–density plane, together with states of free flow (curve F), steady states of synchronized flow (dashed region), and the state within a wide moving jam (the point W associated with \( \rho_{\text{max}}, 0 \)) form a double Z-shaped speed–density relationship. States of free flow (curve F), steady states of synchronized flow (dashed region), and the line J are taken from Fig. 6.6.

(d) an infinite number of critical branches \( v_{\text{cr}}^{(SJ)}(\rho) \), each of which yields the speed within the critical local perturbation for a spontaneous S→J transition as a function of density for a given synchronized flow speed (two such critical branches \( v_{\text{cr}, i}^{(SJ)}(\rho) \), \( i = 1, 2 \) are shown in Fig. 6.7) for the related two speeds of synchronized flow \( v_{i}^{(syn)} \), \( i = 1, 2 \); 
(e) the speed \( v_{\text{min}} = 0 \) within wide moving jams.

The double Z-shaped density dependence of the speed in Fig. 6.7 explains the empirical result (10) of Sect. 4.4: moving jams emerge spontaneously due to a sequence (cascade) of two phase transitions, first an F→S transition, and subsequently an S→J transition (F→S→J-transitions).

To see this double Z-shaped speed–density characteristic more clearly, we drastically simplify Fig. 6.7. In this simplification, which is shown in Fig. 6.8a, the two-dimensional region of states of synchronized flow in Fig. 6.7 is replaced by an average branch \( v_{\text{aver}}^{(syn)} \), and some of these states have been neglected. Furthermore, we average all the various critical branches \( v_{\text{cr}}^{(SJ)}(\rho) \)
Fig. 6.8. Simplified double Z-shaped relationships in three-phase traffic theory. (a) Simplified illustration of Fig. 6.7 for the double Z-shaped speed–density relationship. An infinite number of steady states of synchronized flow for each density in Fig. 6.7 (dashed region) are averaged to the one synchronized flow speed for each given density \((\text{curve } v_{\text{aver}}^{(\text{syn})}(\rho))\), while possible steady states of synchronized flow (left of the threshold density \(\rho_{\text{th}}\) in Fig. 6.7) are neglected. (b) Simplified qualitative illustration of the double Z-shaped flow–density characteristic corresponding to (a) into a single critical branch \(v_{\text{cr, aver}}^{(\text{SJ})}(\rho)\), which corresponds to the average synchronized flow speed on the branch \(v_{\text{aver}}^{(\text{syn})}(\rho)\).

As a result, we have the simplified double Z-shaped speed–density characteristic. This relationship can be considered to consist of the following two distinct Z-characteristics:

(i) The first Z-shaped speed–density characteristic between free flow (curve \(F\)) and averaged states of synchronized flow (branch \(v_{\text{aver}}^{(\text{syn})}(\rho)\)). This Z-characteristic describes an \(F \rightarrow S\) transition.
(ii) The second Z-shaped speed-density characteristic between the averaged states of synchronized flow and the state within a wide moving jam (6.9) (the point \( W \) corresponding to \((\rho_{\text{max}}, 0)\), Fig. 6.8a). This Z-characteristic describes an \( S \rightarrow J \) transition.

The corresponding simplified double Z-shaped flow-density characteristic is shown in Fig. 6.8b.

The simplified double Z-characteristic qualitatively accounts for the \( F \rightarrow S \rightarrow J \) transitions on a homogeneous road as follows:

(i) Let us first assume that a random local perturbation occurs in free flow. If the speed within this local perturbation is lower than the critical speed \( v_{\text{cr}}^{(FS)} \), the perturbation grows, i.e., the speed precipitously decreases. This leads to synchronized flow emergence (down-arrow “\( F \rightarrow S \)” in Fig. 6.8a).

(ii) Let us now assume that the average speed \( v_{\text{aver}}^{(\text{syn})} \) in synchronized flow resulting from an \( F \rightarrow S \) transition decreases. We consider a random local perturbation in this synchronized flow. If the speed within this local perturbation is lower than the critical speed \( v_{\text{cr, aver}}^{(SJ)} \), the perturbation grows, i.e., the speed decreases precipitously. This leads to wide moving jam emergence in the synchronized flow (down-arrow “\( S \rightarrow J \)” in Fig. 6.8a). Otherwise, the perturbation decays and synchronized flow is self-sustaining (up-arrow to the synchronized flow states \( v_{\text{aver}}^{(\text{syn})} \)).

An \( F \rightarrow S \) transition and an \( S \rightarrow J \) transition, the corresponding cascade of \( F \rightarrow S \rightarrow J \) transitions, as well as reverse phase transitions, are qualitatively illustrated by dashed arrows in Fig. 6.9:

(i) The arrow \( A \rightarrow B \) corresponds to an \( F \rightarrow S \) transition.
(ii) The arrow \( C \rightarrow D \) corresponds to a reverse \( S \rightarrow F \) transition.
(iii) The arrow \( E \rightarrow W \) corresponds to an \( S \rightarrow J \) transition.
(iv) The arrow \( W \rightarrow G \) corresponds to a reverse \( J \rightarrow S \) transition.
(v) The sequence of two arrows \( A \rightarrow B \) and \( E \rightarrow W \) corresponds to the cascade of \( F \rightarrow S \rightarrow J \) transitions.

Furthermore, the dotted arrow \( W \rightarrow H \) in Fig. 6.9 illustrates a possible direct reverse \( J \rightarrow F \) transition. The latter transition from a wide moving jam to free flow can result from the dissolution of the wide moving jam. For example, this can occur if free flow is formed downstream of the jam and the flow rate in free flow upstream of the jam becomes considerably less than the flow rate \( q_{\text{out}} \) of the jam outflow for a long enough time.

**Complex Character of Nucleation Effects and Phase Transitions in Freeway Traffic**

To conclude this discussion, we present some figures that summarize the hypotheses of three-phase traffic theory about two qualitatively different
Fig. 6.9. Qualitative illustration of double hysteresis effects on the simplified double Z-shaped relationships in three-phase traffic theory [209, 221]. (a) Double hysteresis effects in the speed–density plane. (b) The same double hysteresis effects in the flow–density plane. The simplified double Z-shaped relationships in (a) and (b) are taken from Fig. 6.8 (a) and (b), respectively.
nucleation and hysteresis effects in free flow, which are responsible for an F→S transition and for an F→J transition, and about moving jam emergence in synchronized flow, i.e., an S→J transition.

In particular, in Fig. 6.10 all critical branches are shown that give the speeds within the critical local perturbations for different possible phase transitions and the related nucleation effects. These critical branches are:

(i) The critical branch $v_{cr}^{(FS)}$ for an F→S transition.

(ii) The critical branch $v_{cr}^{(FJ)}$ for a hypothetical F→J transition.

(iii) An infinite number of critical branches $v_{cr}^{(SJ)}$ for an S→J transition. Two of these branches, $v_{cr,1}^{(SJ)}$ and $v_{cr,2}^{(SJ)}$ are shown in Fig. 6.10. The branch $v_{cr,1}^{(SJ)}$ is related to the synchronized flow speed $v_{1}^{(syn)}$ and the branch $v_{cr,2}^{(SJ)}$ is related to the speed $v_{2}^{(syn)}$.

![Fig. 6.10. Qualitative explanation of various nucleation effects in free and synchronized flows in the speed–density plane [221]. All critical branches determined due to nucleation effects from both Fig. 6.7a and Fig. 6.1b are presented.](image)

In particular, we can see in Fig. 6.10 that in free flow there are two different critical points:

$$\rho^{(\text{free})} = \rho_{\text{max}}$$  \hspace{1cm} (6.24)

$$\rho^{(\text{free})} = \rho_{\text{max}, \text{FJ}}$$  \hspace{1cm} (6.25)
and two different threshold points:

$$\rho^{(\text{free})} = \rho_{th},$$  \hspace{1cm} (6.26)

$$\rho^{(\text{free})} = \rho_{\text{min}}.$$  \hspace{1cm} (6.27)

The points (6.24) and (6.26) are related to an F\(\rightarrow\)S transition. In contrast, the points (6.25) and (6.27) are related to an F\(\rightarrow\)J transition. The existence of two qualitatively different phase transitions can explain the complexity of the theory of phase transitions in free flow. However, as we have already stressed,

$$\rho^{(\text{free})}_{\text{max}} < \rho^{(\text{free})}_{\max, FJ}$$  \hspace{1cm} (6.28)

(Fig. 6.10). For this reason, in reality only an F\(\rightarrow\)S transition can occur spontaneously in free flow.

Note that above we have considered a possible case where the threshold density in free flow for the F\(\rightarrow\)S transition, \(\rho_{th}\), is lower than the threshold density in free flow for an F\(\rightarrow\)J transition \(\rho_{\text{min}}\): \(\rho_{th} < \rho_{\text{min}}\) (Fig. 6.10). Correspondingly, for the threshold flow rate \(q_{th}\) for an F\(\rightarrow\)S transition and the threshold flow rate \(q_{\text{out}}\) for the F\(\rightarrow\)J transition we have \(q_{th} < q_{\text{out}}\). All qualitative results above are also valid for the opposite case \(\rho_{th} \geq \rho_{\text{min}}\) \((q_{th} \geq q_{\text{out}})\). To see this, one can look at Fig. 6.11 where summarized nucleation effects in free and synchronized flows are illustrated in the speed–flow plane for both these cases.

### 6.4.3 Wide Moving Jam Emergence

**Within Initial Moving Synchronized Flow Pattern**

After a moving synchronized flow pattern (MSP) due to an F\(\rightarrow\)S transition has occurred, the MSP propagates away from a road location where the MSP has emerged.

During MSP propagation the average speed within the MSP can gradually decrease. In this case, corresponding to the consideration of moving jam emergence in synchronized flow made above one can expect that when the speed within the MSP becomes low enough one or sometimes several moving jams should begin to emerge within the initial MSP. This S\(\rightarrow\)J transition has indeed been found both in numerical simulations (Fig. 6.12) and in empirical observations (see Sect. 12.4).

Firstly, due to an external local perturbation, an MSP occurs the same way as discussed in Sect. 5.2.8 (the arrow “F\(\rightarrow\)S” in Fig. 6.12). Then during MSP propagation and its widening the average speed of synchronized flow within the MSP decreases and an S\(\rightarrow\)J transition occurs in synchronized flow leading to wide moving jam formation within the initial MSP (the arrow “S\(\rightarrow\)J” in Fig. 6.12). This S\(\rightarrow\)J transition occurs later and at another road location in comparison with the location where the initial F\(\rightarrow\)S transition has occurred leading to MSP formation.
Fig. 6.11. Qualitative explanation of various nucleation effects in free and synchronized flows in the speed–flow plane. The summarized figure where all critical branches, which are shown in Fig. 6.10, are presented. (a) For the case $q_{th} > q_{out}$. (b) For the case $q_{th} < q_{out}$. Curve $J$ is related to the line $J$ in the flow–density plane.
6.5 Moving Jam Emergence in Synchronized Flow at Bottlenecks

In the above consideration outlined in this chapter, we have considered an F→J transition, an S→J transition, and the double Z-shaped traffic flow characteristic that explains F→S→J transitions on a homogeneous road. Below in this chapter we use these results for the explanation of wide moving jam emergence in synchronized flow that has initially occurred at a bottleneck. However, firstly, we explain why moving jams do not emerge spontaneously in an initial free flow at a freeway bottleneck.

6.5.1 Why Moving Jams Do not Emerge in Free Flow at Bottlenecks

To understand why moving jams do not emerge spontaneously in an initial free flow at a freeway bottleneck, we compare the critical amplitude of the critical perturbation for an F→J transition, and the critical amplitude of the critical perturbation for an F→S transition. It is convenient to consider these critical amplitudes of local perturbations as functions of the flow rate to the on-ramp \( q_{on} \) at a given flow rate on the main road upstream of the on-ramp, \( q_{in} \) (Fig. 6.13).

The critical amplitude of the critical perturbation for an F→J transition at a bottleneck is

\[
\Delta v_{cr, FJ}^{(B)} = v_{free}^{(B)} - v_{cr, FJ}^{(B)} .
\]  

(6.29)

The critical amplitude of the local perturbation for an F→S transition at the

---

**Fig. 6.12.** Emergence of a wide moving jam within an initial MSP: F→S→J transitions. Results of numerical simulations of a microscopic model based on three-phase traffic theory [329]
Fig. 6.13. Qualitative comparison of the critical perturbation amplitude for an F→S transition (6.30) and the critical perturbation amplitude for an F→J transition (6.29) in free flow at an on-ramp bottleneck. States of free flow $v_{\text{free}}^{(B)}$, the critical branch $v_{\text{cr, FS}}^{(B)}$ that gives the speed within the critical local perturbation for a spontaneous F→S transition, the critical branch $v_{\text{cr, FJ}}^{(B)}$ that gives the speed within the critical local perturbation for a hypothetical F→J transition, and a two-dimensional region of steady states of synchronized flow (dashed region). The states of free flow and synchronized flow are taken from Fig. 5.14.

bottleneck corresponding to (5.49) is

$$\Delta v_{\text{cr, FS}}^{(B)} = v_{\text{free}}^{(B)} - v_{\text{cr, FS}}^{(B)}. \quad (6.30)$$

In (6.29), $v_{\text{cr, FJ}}^{(B)}$ is the speed within the critical local perturbation for an F→J transition at the bottleneck. This speed determines the critical branch $v_{\text{cr, FJ}}^{(B)}(q_{on})$ for the F→J transition in Fig. 6.13. The physics of this critical branch is the same as those of the critical branch $v_{\text{cr}}^{(FJ)}$ in Fig. 6.1b.

Let us assume that a local perturbation occurs in free flow at a bottleneck that leads to a decrease in the speed at the bottleneck. If this speed is lower than $v_{\text{cr, FJ}}^{(B)}$, the perturbation grows and leads to wide moving jam emergence. If this speed is higher than $v_{\text{cr, FJ}}^{(B)}$, the perturbation decays.

The speed $v_{\text{free}}^{(B)}$ is the speed within the deterministic perturbation at the bottleneck. The physical sense of the speed $v_{\text{free}}^{(B)}$ as well as of the critical branch $v_{\text{cr, FS}}^{(B)}(q_{on})$ for an F→S transition at the bottleneck have already been discussed in Sect. 5.3 (see the explanation of Fig. 5.14a).
From Fig. 6.13 we can see that at the same flow rate to the on-ramp $q_{on}$ the critical amplitude of the critical perturbation for an $F\rightarrow J$ transition (6.29) is considerably greater than the critical amplitude of the critical perturbation $\Delta v_{cr, FS}^{(B)}$ (6.30) for an $F\rightarrow S$ transition at a bottleneck:

$$\Delta v_{cr, FS}^{(B)} \ll \Delta v_{cr, FJ}^{(B)} .$$

This conclusion that is similar to that made above for a homogeneous road is illustrated in Fig. 6.13. For this reason, the probability for an $F\rightarrow S$ transition at the bottleneck is much higher than the one for an $F\rightarrow J$ transition. Thus, in three-phase traffic theory, in free flow at the bottleneck an $F\rightarrow S$ transition (breakdown phenomenon, Sect. 5.3) occurs rather than an $F\rightarrow J$ transition.

### 6.5.2 Z-Characteristic for $S\rightarrow J$ Transition at Bottlenecks

We will see in Chap. 18 that synchronized flow that has occurred due to an $F\rightarrow S$ transition at an effectual bottleneck is usually widening upstream of the bottleneck. In this synchronized flow upstream of the bottleneck, wide moving jams can emerge due to a spontaneous $S\rightarrow J$ transition. In other words, this $S\rightarrow J$ transition occurs upstream of the bottleneck rather than at the effectual bottleneck. The bottleneck has only an indirect influence on the $S\rightarrow J$ transition: bottleneck characteristics can change the speed and density and other parameters of synchronized flow upstream of the bottleneck. This can influence the condition for the $S\rightarrow J$ transition.

Because an $S\rightarrow J$ transition occurs within synchronized flow at road locations, which can be far upstream of a bottleneck, we can expect that results of the theory of $S\rightarrow J$ transitions considered above for a homogeneous road are also valid for the case of an isolated bottleneck due to an on-ramp under consideration, i.e., the bottleneck that is far from other freeway bottlenecks. The only difference with a homogeneous road is that synchronized flow on the main road upstream of the on-ramp depends on the flow rates $q_{on}$ and $q_{in}$. For this reason, it is convenient to consider dependencies of characteristics of an $S\rightarrow J$ transition at the bottleneck as functions of $q_{on}$ at a given $q_{in}$ (Fig. 6.14).

In particular, as for the case of an $S\rightarrow J$ transition on a homogeneous road (Fig. 6.6), there is a Z-shaped characteristic that explains an $S\rightarrow J$ transition at bottlenecks (Fig. 6.14a). The Z-characteristic consists of

(i) a 2D region of steady states of synchronized flow (dashed region);

(ii) an infinite number of critical branches $v_{cr, SJ}^{(B)}(q_{on})$ each of which yields the speed within the critical local perturbation for a spontaneous $S\rightarrow J$ transition as a function of the flow rate $q_{on}$ for a given synchronized flow speed (two such critical branches $v_{cr, SJ, i}(q_{on})$, $i = 1, 2$ are shown in Fig. 6.14a for the corresponding two speeds of synchronized flow $v_{i}^{(syn)}$, $i = 1, 2$);

(iii) the speed $v_{min} = 0$ within wide moving jams.
6.5.3 Physics of Moving Jam Emergence in Synchronized Flow

To explain an S→J transition based on the Z-characteristic for this transition, we drastically simplify the Z-characteristic shown in Fig. 6.14a. We average different states of synchronized flow into a single speed $v^{(B)}_{\text{syn, aver}}$ at a given flow rate $q_{\text{on}}$. When $q_{\text{in}} = \text{const}$, this average synchronized flow speed $v^{(B)}_{\text{syn, aver}}(q_{\text{on}})$ is a decreasing function of the flow rate $q_{\text{on}}$ (Fig. 6.14b). Moreover, we average all the various critical branches $v^{(B)}_{\text{cr, SJ}}$ into a single critical...
branch \( v_{cr, SJ, aver}^{(B)}(q_{on}) \) associated with the average synchronized flow speed \( v_{syn, aver}^{(B)}(q_{on}) \) (Fig. 6.14b).

The Z-characteristic qualitatively explains an S→J transition at a bottleneck as follows:

(i) Let us assume that the speed in synchronized flow upstream of the bottleneck is low and that a local perturbation occurs in that synchronized flow. This local perturbation leads to a local decrease in synchronized flow speed. If this speed is lower than \( v_{cr, SJ, aver}^{(B)} \), the perturbation grows and leads to wide moving jam emergence in synchronized flow (down-arrow “S→J” in Fig. 6.14b).

(ii) Alternatively, let us assume that a local perturbation occurs in this synchronized flow and the speed within the perturbation is higher than \( v_{cr, SJ, aver}^{(B)} \). Then the perturbation decays and synchronized flow remains (up-arrow to the synchronized flow states \( v_{syn, aver}^{(B)} \)).

(iii) The Z-characteristic can qualitatively explain the empirical results (8) and (9) of Sect. 4.4. To show this, let us note that the lower the speed in synchronized flow, the lower the difference

\[
\Delta v_{cr, SJ, aver}^{(B)} = v_{syn, aver}^{(B)} - v_{cr, SJ, aver}^{(B)}. \tag{6.32}
\]

This difference \( \Delta v_{cr, SJ, aver}^{(B)} \) is the average critical amplitude of the local perturbation required for an S→J transition: if the amplitude of an initial local perturbation \( 6.18 \) in synchronized flow

\[
\Delta v_{pert}^{(B)} > \Delta v_{cr, SJ, aver}^{(B)}, \tag{6.33}
\]

we have the case of item (i), i.e., an S→J transition occurs (down-arrow “S→J” in Fig. 6.14b). If, in contrast,

\[
\Delta v_{initial}^{(pert)} < \Delta v_{cr, SJ, aver}^{(B)}, \tag{6.34}
\]

we have the case of item (ii), i.e., an S→J transition does not occur (up-arrow to the synchronized flow states \( v_{syn, aver}^{(B)} \)). We can distinguish two different cases:

(a) The flow rate \( q_{on} \) increases and, consequently, the average synchronized flow speed \( v_{syn, aver}^{(B)}(q_{on}) \) decreases (Fig. 6.14b). In this case, the average critical amplitude of the local perturbation \( \Delta v_{cr, SJ, aver}^{(B)}(q_{on}) \) as a function of flow rate \( q_{on} \) decreases, too. Thus, the lower the synchronized flow speed, the lower the amplitude of a random perturbation required for wide moving jam emergence in this synchronized flow. This explains why the lower the vehicle speed in dense synchronized flow, the higher the frequency of wide moving jam emergence in the synchronized flow (the empirical result (8) of Sect. 4.4). Therefore, the lower vehicle speed in the synchronized flow, the higher the probability of wide moving jam emergence in this flow.
(b) In contrast, if the flow rate $q_{on}$ decreases, the average synchronized flow speed $v_{syn, \text{aver}}^{(B)}(q_{on})$ increases (Fig. 6.14b). In this case, the average critical amplitude of the local perturbation $\Delta v_{cr, \text{SJ}, \text{aver}}^{(B)}(q_{on})$ as a function of the flow rate $q_{on}$ increases, too. Thus, the higher the synchronized flow speed, the higher the amplitude of a random perturbation required for wide moving jam emergence in this synchronized flow. For this reason, in a state of synchronized flow with a higher vehicle speed a wide moving jam should not necessarily emerge (the empirical result (9) of Sect. 4.4).

(iv) If the flow rate to the on-ramp further decreases, then the average speed in synchronized flow $v_{syn, \text{aver}}^{(B)}(q_{on})$ continues to increase (Fig. 6.14b). Finally, the flow rate $q_{on}$ becomes lower than the threshold flow rate (Fig. 6.14b), i.e., it satisfies the condition

$$q_{on} < q_{on}^{(\text{th, SJ})}. \tag{6.35}$$

In this case, synchronized flow at higher speeds is stable with respect to an $S \rightarrow J$ transition: independent of how high the amplitudes of random local perturbations in this synchronized flow are, wide moving jams cannot emerge in this synchronized flow. This also correlates with the empirical result (9) of Sect. 4.4.

Moving Jam Nucleation and Over-Deceleration in Synchronized Flow

The physics of moving jam emergence in metastable states of synchronized flow (Sect. 6.3.1) and the associated Z-characteristic (Sect. 6.5.2) can be explained by the well-known vehicle over-deceleration effect (Sect. 3.2.4). This effect occurs within a random local perturbation in high density synchronized flow: if in this synchronized flow the preceding vehicle begins to decelerate unexpectedly, then due to the driver time delay the driver decelerates stronger than is required to avoid collisions.

Let us consider a metastable steady state of synchronized flow that is associated with a point above the line $J$ in the flow–density plane (Sect. 6.3.2). We assume that in this synchronized flow state an initial local perturbation occurs. If the speed $v_{\text{initial}}^{(\text{pert})}$ within the perturbation is lower than the critical speed $v_{cr, \text{SJ}}^{(B)}$, the amplitude of the perturbation is greater than the critical perturbation amplitude. In this case, the density within the perturbation is high. Each driver that comes into the high density region within the perturbation decelerates stronger than the deceleration of the preceding vehicle within the perturbation. This occurs due to the over-deceleration effect. As a result, the speed within the perturbation decreases. In other words, the perturbation amplitude grows continuously over time. This case is associated with the condition (6.33). Due to this nucleation effect a wide moving jam emerges, i.e., an $S \rightarrow J$ transition occurs.
Otherwise, let us consider an opposite case at which in the same initial state of synchronized flow a local perturbation occurs whose amplitude $\Delta v_{\text{initial}}^{(\text{pert})}$ is less than the critical perturbation amplitude. In this case, drivers that come into the region within the perturbation will not necessarily decelerate stronger than the deceleration of the preceding vehicle inside the perturbation, i.e., the vehicle over-deceleration effect is not realized. This is because within the perturbation the density is not very high, i.e., space gaps between vehicles are not very small. Therefore, a driver has enough time to adjust the speed to that of the preceding vehicle. As a result, the speed within the initial perturbation does not decrease. This case is associated with the condition (6.34). The perturbation either decays or a transformation to another state of synchronized flow occurs (Sect. 4.3.3).

### Nucleation-Interruption Effects in Dynamic Synchronized Flow States

In real traffic flow, states of synchronized flow are dynamic states (Sect. 4.3.4). In these synchronized flow states, distances between vehicles can vary within the range between the safe distance and the synchronization distance (Sect. 4.3.2). A dynamic state of synchronized flow where the mean space gap is associated with a point above the line $J$ in the flow-density plane is a metastable synchronized flow state with respect to wide moving jam formation.

In these metastable dynamic synchronized flow states, moving jam nucleation, i.e., the growth of an initial local perturbation, can be interrupted. This interruption of moving jam nucleation can occur if a space gap of one of the drivers that comes into the growing local perturbation is considerably greater than the mean space gap within this perturbation. In this case, this driver will not necessarily decelerate stronger than the preceding vehicle: this driver has enough time to adjust the speed to that of the preceding vehicle. As a result, the growth of the perturbation, i.e., the nucleation effect, can be interrupted.

Due to very different space gaps between vehicles there can be many such nucleation-interruption effects in real metastable synchronized flow before a wide moving jam emerges. In particular, due to the nucleation-interruption effect some narrow moving jams that have formed in the synchronized flow can begin to dissolve. Later, a subset of these moving jams can begin to grow once more, and so on. Thus, the nucleation-interruption effect can be one of the reasons for very complex spatiotemporal behavior observed in dense synchronized flow states (Chap. 12).
6.5.4 Double Z-Characteristic and $F \rightarrow S \rightarrow J$ Transitions at Bottlenecks

Just as for a homogeneous road (Fig. 6.7), there is also a double Z-characteristic that explains $F \rightarrow S \rightarrow J$ transitions at bottlenecks (Fig. 6.15a).

The double Z-characteristic at the bottleneck consists of

(a) states of free flow $v_{\text{free}}^{(B)}(q_{on})$ as a function of the flow rate $q_{on}$;

(b) the critical branch $v_{\text{cr, FS}}^{(B)}(q_{on})$ that gives the speed within the critical local perturbation for a spontaneous $F \rightarrow S$ transition as a function of the flow rate $q_{on}$;

(c) a 2D region of steady states of synchronized flow (dashed region);

(d) an infinite number of critical branches $v_{\text{cr, SJ}}^{(B)}(q_{on})$ each of which gives the speed within the critical local perturbation for a spontaneous $S \rightarrow J$ transition as a function of the flow rate $q_{on}$ for a given synchronized flow speed (two of these critical branches $v_{\text{cr, SJ}}^{(B)}(q_{on}), i = 1, 2$ for the corresponding two speeds of synchronized flow $v_{i}^{(\text{syn})}$, $i = 1, 2$ are shown in Fig. 6.15a);

(e) the speed $v_{\text{min}} = 0$ within wide moving jams.

Now we simplify Fig. 6.15a by averaging of all different states of synchronized flow to the one speed at a given flow rate $q_{on}$ as this has been done in Fig. 6.14b. The averaged synchronized flow speed is a function of the flow rate to the on-ramp: $v_{\text{syn, aver}}^{(B)}(q_{on})$. Further we average all different critical branches $v_{\text{cr, SJ}}^{(B)}(q_{on})$ to one branch $v_{\text{cr, SJ, aver}}^{(B)}(q_{on})$ that is related to the average speed $v_{\text{syn, aver}}^{(B)}(q_{on})$. Thus, we get much simpler double Z-shaped traffic flow characteristic for the explanation of $F \rightarrow S \rightarrow J$ transitions at the bottleneck (Fig. 6.15b). The latter resembles the simplified double Z-shaped traffic flow characteristics for a homogeneous road (Fig. 6.8a).

The double Z-characteristic qualitatively explains $F \rightarrow S \rightarrow J$ transitions as follows:

(i) Let us first assume free flow at a bottleneck. In this free flow, a random local perturbation occurs at the bottleneck. If the speed within this local perturbation is lower than the critical speed $v_{\text{cr, FS}}^{(B)}$, the perturbation grows, i.e., the speed decreases precipitously. This leads to synchronized flow emergence at the bottleneck (down-arrow “$F \rightarrow S$” in Fig. 6.15b). Otherwise, the perturbation decays and free flow remains at the bottleneck (up-arrow to the free flow states $v_{\text{free}}^{(B)}$).

(ii) The synchronized flow that has occurred at the bottleneck propagates later upstream. Let us now assume that the average speed in this flow decreases and that a local perturbation in this synchronized flow occurs at some location upstream of the bottleneck. If the speed within this local perturbation is lower than the critical speed $v_{\text{cr, SJ, aver}}^{(B)}$, the perturbation grows, i.e., the speed decreases precipitously. This leads to wide moving
6.5 Moving Jam Emergence in Synchronized Flow at Bottlenecks 177

Fig. 6.15. Qualitative illustration of $F \rightarrow S \rightarrow J$ transitions and the related double Z-shaped characteristic of traffic flow at an on-ramp bottleneck. (a) States of free flow $v_{\text{free}}^{(B)}$, the critical branch $v_{\text{cr, FS}}^{(B)}$ that gives the speed within the critical local perturbation required for a spontaneous $F \rightarrow S$ transition, a two-dimensional region of steady states of synchronized flow (dashed region), two critical branches $v_{\text{cr, SJ, } i}^{(B)}$, $i = 1, 2$ that gives the speed within the critical local perturbation required for spontaneous $S \rightarrow J$ transitions for the corresponding two speeds of synchronized flow $v_{\text{syn, } i}^{(syn)}$, $i = 1, 2$, and the speed $v_{\text{min}} = 0$ within wide moving jams. (b) Simplified double Z-shaped speed–flow characteristic related to (a) where an infinite number of synchronized flow speeds for each given flow rate $q_{\text{on}}$ are averaged to the speed $v_{\text{syn, aver}}^{(B)}(q_{\text{on}})$, and an infinite number of the critical branches $v_{\text{cr, SJ, aver}}^{(B)}$ are averaged to the critical branch $v_{\text{cr, SJ, aver}}^{(B)}(q_{\text{on}})$. In (a) states of free flow (the curve $v_{\text{free}}^{(B)}$), states of synchronized flow (dashed region), and the critical branch $v_{\text{cr, FS}}^{(B)}$ are taken from Fig. 5.14; the critical branches $v_{\text{cr, SJ, } i}^{(B)}$, $i = 1, 2$ are taken from Fig. 6.14a.
jam emergence in synchronized flow (down-arrow “S→J” in Fig. 6.15b). Otherwise, the perturbation decays and synchronized flow remains (up-arrow to the synchronized flow states $v_{\text{syn, aver}}^{(B)}$).

6.6 Conclusions

(i) In three-phase traffic theory, there are two qualitatively different nucleation effects in free flow:

1. The nucleation effect that is responsible for jam emergence in free flow, i.e., for an $F\rightarrow J$ transition.
2. The nucleation effect that is responsible for an $F\rightarrow S$ transition. At any flow rate (density) in free flow where synchronized flow could emerge the critical amplitude of a local perturbation required for moving jam emergence is considerably higher than the critical amplitude of a local perturbation that is required for synchronized flow emergence ($F\rightarrow S$ transition). For this reason, a local region of synchronized flow can occur spontaneously in free flow rather than a moving jam. This also explains why the existence of the critical (limit) density for free flow is related to an $F\rightarrow S$ transition rather than to spontaneous moving jam emergence.

(ii) At each given density in free flow the probability for an $F\rightarrow S$ transition during a given time interval is considerably higher than the probability for an $F\rightarrow J$ transition.

(iii) Moving jams can emerge in synchronized flow, i.e., due to $F\rightarrow S\rightarrow J$ transitions.

(iv) The line $J$ separates all steady states of synchronized flow in the flow–density plane into two qualitatively different classes:

1. In states that are related to points in the flow–density plane lying below the line $J$ no wide moving jams can continue to exist nor be excited.
2. States that are related to points in the flow–density plane lying on and above the line $J$ are metastable steady states of synchronized flow with respect to wide moving jam emergence ($S\rightarrow J$ transition). All (infinite number) states that are on the line $J$ are the threshold states for an $S\rightarrow J$ transition.

(v) Like an $F\rightarrow S$ transition, an $S\rightarrow J$ transition leads to a Z-shaped speed–density characteristic of traffic flow as well.

(vi) The $F\rightarrow S$ transition and the $S\rightarrow J$ transition lead to a double Z-shaped speed–density (flow rate) characteristic of traffic flow.

(vii) All of the above conclusions are also qualitatively valid for phase transitions at a freeway bottleneck.
7 Congested Patterns at Freeway Bottlenecks in Three-Phase Traffic Theory

7.1 Introduction

The three-phase traffic theory considered in Chaps. 4–6 gives a theoretical basis for an explanation of empirical phase transitions in real traffic (Chaps. 10 and 12).

In this chapter, a further development of three-phase traffic theory will be made. Here, based on results presented in [208,218,221] we will try to answer the question about spatiotemporal congested patterns that should appear due to the onset of congestion at effectual freeway bottlenecks. We will see that these qualitative theoretical results can explain the main empirical congested pattern features (Sect. 2.4 and Chaps. 9, 12). This consideration is also the basis for a microscopic theory of spatiotemporal congested traffic patterns at effectual bottlenecks in Part III.

7.2 Two Main Types of Spatiotemporal Congested Patterns

Recall that if the breakdown phenomenon (F→S transition) occurs at a freeway bottleneck, the bottleneck is called an effectual bottleneck. After a congested pattern due to an F→S transition has occurred upstream of the effectual bottleneck then the downstream front of this pattern is spatially fixed at a freeway location in the vicinity of the effectual bottleneck. This downstream front of the congested pattern separates free flow downstream from synchronized flow upstream of the effectual bottleneck. Within this downstream front vehicles accelerate from synchronized flow (congested traffic) upstream of the front to free flow downstream of the front. The freeway location where the downstream front of the congested pattern is spatially fixed will be called an effective location of the effectual bottleneck.

For simplicity, we discuss only features of congested patterns that occur upstream of an isolated effectual freeway bottleneck (or an isolated bottleneck for short). An isolated effectual freeway bottleneck is a freeway bottleneck

1 The exception is the case of MSP formation: the downstream front of an MSP propagates on the freeway rather than this front is spatially fixed.
that is far away from all other possible effectual freeway bottlenecks. By this
definition, it is assumed that the influence of all other effectual bottlenecks
on pattern formation at the isolated bottleneck can be neglected.

There are two main types of congested patterns that can occur sponta­
neously upstream of an isolated bottleneck:

(i) *The general pattern* or *GP* for short. An GP is the congested pattern at
an isolated bottleneck where synchronized flow occurs upstream of the
bottleneck and wide moving jams emerge spontaneously in that synchro­
nized flow (see empirical example of an GP in Fig. 2.21). Thus, the GP
consists of both traffic phases in congested traffic: “synchronized flow”
and “wide moving jam.”

(ii) *The synchronized flow pattern* or *SP* for short. An SP consists of synchro­
nized flow upstream of the isolated bottleneck only, i.e., no wide moving
jams emerge in that synchronized flow (see empirical example of an SP
in Fig. 2.14).

However, depending on the bottleneck features and on traffic demand, GPs
and SPs show a diverse variety of special cases whose consideration is one of
the main aims of this book.

### 7.3 Simplified Diagram of Congested Patterns
**at Isolated Bottlenecks**

As discussed above, a congested pattern upstream of an isolated bottleneck
occurs due to an F→S transition at this bottleneck. The probability of the
F→S transition at the bottleneck is much higher than away from the bottle­
neck. This is because the bottleneck causes a deterministic perturbation in
free flow that is localized at the bottleneck (Sect. 5.3.1).

As in Chaps. 5 and 6, we continue a consideration of an on-ramp bot­
leneck. The higher the flow rate to the on-ramp \(q_{on}\), the greater the amplitude
of this deterministic local perturbation caused by the bottleneck on the main
road. Therefore, F→S transition occurrence at the bottleneck and hence con­
gested pattern emergence upstream of the bottleneck should depend on both
the flow rate in an initial free flow upstream of the bottleneck \(q_{in}\) and on the
flow rate to the on-ramp \(q_{on}\).

This leads to a diagram of congested patterns in the flow–flow plane,
i.e., the plane that has the coordinates \((q_{on}, q_{in})\). In this diagram, pattern
emergence upstream of the bottleneck is presented as a function of the flow
rates \(q_{in}\) and \(q_{on}\). In three-phase traffic theory, the diagram of congested
patterns at freeway bottlenecks has been postulated in [218]. A simplified
version of this diagram is shown in Fig. 7.1.

There are two main boundaries at this diagram, \(F_{S}^{(B)}\) and \(S_{j}^{(B)}\). Below
and left of the boundary \(F_{S}^{(B)}\) free flow occurs upstream of the on-ramp.
Between the boundaries $F_S^{(B)}$ and $S_J^{(B)}$ different synchronized flow patterns (SP) emerge on the main road upstream of the on-ramp. Right of the boundary $S_J^{(B)}$ wide moving jams emerge in synchronized flow of these SPs, i.e., different general patterns (GP) occur upstream of the on-ramp.

At the boundary $F_S^{(B)}$ (Fig. 7.1) an F→S transition (speed breakdown) occurs at a bottleneck during a given time interval $T_{ob}$ with the probability $P_{FS}^{(B)} = 1$. In other words, at this boundary the “synchronized flow” phase occurs spontaneously at the bottleneck, i.e., an SP emerges. Thus, we define the boundary $F_S^{(B)}$ in the diagram of congested patterns at the bottleneck as follows.

- The boundary $F_S^{(B)}$ in the diagram of congested patterns (Fig. 7.1) is related to the minimum traffic demand (i.e., the minimum flow rates $q_{sum}(q_{on}, q_{in})$ (5.35)) at which an F→S transition occurs at a bottleneck during a given time interval $T_{ob}$ with the probability $P_{FS}^{(B)} = 1$.

The form of the boundary $F_S^{(B)}$ can be found from the following qualitative consideration. The critical amplitude for an F→S transition tends to zero if the flow rate $q_{in}$ approaches the limit (maximum) point of free flow $q_{lim} = q_{max}^{(free)}$ (Sect. 5.2; the curve $F_S$ in Fig. 6.2b). Thus, the F→S transition at the bottleneck must occur already at $q_{on} \to 0$ at some flow rate $q_{in} = q_{max, lim}^{(free)}$ that should be close to the limit flow rate $q_{max}^{(free)}$ (about a possible difference between the flow rates $q_{max, lim}^{(free)}$ and $q_{max}^{(free)}$ see the next Sect. 7.4.1). The higher

---

2 The definition of the probability $P_{FS}^{(B)}$ for an F→S transition at a bottleneck appears in Sect. 8.3.1.
the flow rate $q_{on}$, the greater the amplitude of a permanent local perturbation caused by the bottleneck. Therefore, the higher the flow rate $q_{on}$, the lower the flow rate should be in free flow upstream of the bottleneck $q_{in}$ at which the F→S transition occurs at the bottleneck. This explains the form of the boundary $F_{S(B)}^{(B)}$ for spontaneous SP emergence at the bottleneck in Fig. 7.1.

The boundary $S_{J(B)}^{(B)}$ is determined by spontaneous wide moving jam emergence in synchronized flow (S→J transition) on the main road upstream of a bottleneck during a given time interval $T_{ob}$ with the probability $P_{SJ} = 1$.\(^3\) In other words, at this boundary the “wide moving jam” phase occurs spontaneously in synchronized flow upstream of the bottleneck, i.e., an GP emerges. Thus, we define the boundary $S_{J(B)}^{(B)}$ in the diagram of congested patterns at the bottleneck as follows.

- The boundary $S_{J(B)}^{(B)}$ in the diagram of congested patterns (Fig. 7.1) is related to the minimum traffic demand (i.e., the minimum flow rates $q_{sum}(q_{on}, q_{in})$ (5.35)) at which an S→J transition occurs in synchronized flow upstream of a bottleneck during a given time interval $T_{ob}$ with the probability $P_{SJ} = 1$.

The form of the boundary $S_{J(B)}^{(B)}$ can be found from the following qualitative consideration. On the one hand, between the boundaries $F_{S(B)}^{(B)}$ and $S_{J(B)}^{(B)}$ the vehicle speed in an SP should decrease when the flow rate $q_{on}$ increases. On the other hand, the lower the vehicle speed in synchronized flow, the higher the probability of wide moving jam emergence in that synchronized flow (Sect. 6.5.3). Thus, in comparison with the boundary $F_{S(B)}^{(B)}$, the boundary $S_{J(B)}^{(B)}$ should be shifted to the right in the flow-flow plane in Fig. 7.1.\(^4\)

\(^3\) Recall that $P_{SJ}$ is the probability for an S→J transition (Sect. 6.3).

\(^4\) As discussed in Sect. 6.3.3, there are high density and low speed synchronized flow states that are stable against moving jam emergence. These states are associated with points in the flow–density plane that are below the line $J$. Thus, it can turn out that when the flow rate $q_{on}$ increases and the vehicle speed in an SP decreases rather than metastable synchronized flow states high density and low speed synchronized flow states appear within the SP that are stable against moving jam emergence. In this case, moving jams do not occur in these synchronized flow states of the SP, i.e., GPs do not appear at a bottleneck in some range of the flow rate $q_{in}$. However, this scenario of congested pattern evolution in which moving jams do not emerge in synchronized flow upstream of the bottleneck when the flow rate $q_{on}$ increases and takes on large values has not been observed in empirical data up to now. Nevertheless, we can expect that stable dynamic synchronized flow states of very high density and low speed that are predicted by three-phase traffic theory (Sect. 6.3.3) will be found in the future in empirical data.
7.4 Synchronized Flow Patterns

7.4.1 Influence of Fluctuations on Limit Point for Free Flow at Bottlenecks

It is clear that bottlenecks can introduce additional fluctuations in comparison with fluctuations in an initial free flow on a homogeneous road. Indeed, if an effectual bottleneck is an on-ramp bottleneck, then vehicles, which merge onto the main road from the on-ramp, force vehicles on the main road upstream of the merging region of the on-ramp to decelerate or to change the lane. Both effects can cause additional random fluctuations localized on the main road in the vicinity of the merging region of the on-ramp. These localized fluctuations (random perturbations) are responsible for a random component of the local perturbation on the main road in the vicinity of the on-ramp bottleneck that can cause a spontaneous F→S transition on the main road at the on-ramp bottleneck (Sect. 5.3.4).

To show an influence of these localized random perturbations on the limit (critical) point of free flow clearly, let us consider a hypothetical case when the flow rate \( q_{\text{in}} \) is high but the flow rate \( q_{\text{on}} \) is extremely small, i.e., when

\[
q_{\text{on}} \to 0 \quad \text{but} \quad q_{\text{on}} \neq 0.
\]  

(7.1)

The flow rate \( q_{\text{on}} \) in (7.1) is considered to be so small that its influence on the flow rate in free flow downstream of the on-ramp \( q_{\text{sum}} \) (5.35) can be neglected: \( q_{\text{sum}} = q_{\text{in}} \).

We know (Sect. 5.2) that if we have \( q_{\text{on}} = 0 \), then the critical point of free flow where a spontaneous F→S transition must occur would be the same as on a homogeneous road, i.e., this would be the limit point of free flow \((\rho_{\text{max}}^{(\text{free})}, q_{\text{max}}^{(\text{free})})\). When, however, the conditions (7.1) are satisfied we find a qualitatively different case compared with the homogeneous road.

In this case, there are only single vehicles merging onto the main road from the on-ramp: at \( q_{\text{on}} \to 0 \) (7.1) there can be large enough time intervals between the merging vehicles. During these time intervals traffic flow on the main road can be considered undisturbed by the on-ramp vehicles. Thus, there cannot be a deterministic local perturbation on the main road in the vicinity of the on-ramp. However, if the flow rate upstream of the on-ramp \( q_{\text{in}} \) is high enough, then these single vehicles during their merging cause time-limited random disturbances in free flow on the main road in the vicinity of the on-ramp. These random disturbances can be considered to be some local fluctuations, which appear in addition to the fluctuations that occur in free flow at \( q_{\text{on}} = 0 \), i.e., when no vehicles merge from the on-ramp onto the main road.

These additional fluctuations (additional random perturbations) are localized in the vicinity of the on-ramp. These perturbations can obviously cause an F→S transition at the bottleneck at a lower flow rate \( q_{\text{in}} \) than it
would be in the case at $q_{on} = 0$, i.e., on a homogeneous road.\(^5\) Thus, we can expect that due to these additional random local perturbations in the vicinity of the on-ramp at $q_{on} \to 0$ (7.1), the critical flow rate and the critical density for the F$\to$S transition become lower than these critical traffic variables on the homogeneous road (5.2), i.e., at $q_{on} = 0$.

\[ q_{\text{max, lim}}^{(\text{free B})} < q_{\text{max}}^{(\text{free B})} \quad (\rho_{\text{max, lim}}^{(\text{free B})} < \rho_{\text{max}}^{(\text{free B})}). \quad (7.2) \]

\(^5\) This is true when fluctuations in traffic flow caused by vehicles merging onto the main road from the on-ramp are higher than other fluctuations on the main road.

---

**Fig. 7.2.** Diagram of congested patterns at an on-ramp bottleneck in three-phase traffic theory with regions of different possible synchronized flow patterns (SP). Taken from [218]
Consequently, for the new critical speed in free flow on the main road in the vicinity of the on-ramp at $q_{on} \rightarrow 0$ (7.1) we have

$$v_{\min, \lim}^{(\text{free B})} = \frac{q_{\max, \lim}^{(\text{free B})}}{\rho_{\max, \lim}^{(\text{free B})}}.$$  (7.3)

### 7.4.2 Moving Synchronized Flow Pattern Emergence at Bottlenecks

Additional random local perturbations on the main road in the vicinity of the on-ramp at $q_{on} \rightarrow 0$ (7.1) discussed above can also lead to another effect. This effect is moving synchronized flow pattern (MSP) emergence at a high enough flow rate $q_{in}$ and a small enough flow rate $q_{on}$. This MSP emergence can be realized in a small vicinity of the critical point $q_{a_{\max, \lim}^{(\text{free B})}}$ in the region of SP emergence in the diagram of the congested patterns (this region is labeled “MSP” right of the boundary $F_{S}^{(B)}$ in Fig. 7.2).

Let us assume that in free flow on the main road in the vicinity of the on-ramp due to the merging of a single vehicle from the on-ramp onto the main road a random local perturbation appears whose amplitude exceeds the critical amplitude. Then the perturbation grows and leads to synchronized flow emergence in a local region on the main road near the on-ramp.

### Velocity of Upstream Front of Moving Synchronized Flow Patterns

At a high enough flow rate $q_{in}$ under consideration (the region labeled “MSP” in Fig. 7.2) the upstream front of this local region of synchronized flow propagates continuously upstream; the velocity of this upstream front of synchronized flow is $v_{up} < 0$. Recall that at the upstream front of synchronized flow vehicles moving in free flow upstream of the on-ramp must begin to slow down approaching synchronized flow.

To explain the condition $v_{up} < 0$ at a high enough flow rate $q_{in}$, let us consider the upstream front of a local region of synchronized flow that moves with a constant time-independent velocity $v_{p} = v_{up}$. In this case, for a qualitative analysis the Stokes shock-wave formula (3.5) can be used, which we rewrite in the form

$$v_{up} = \frac{q_{up}^{(\text{syn})} - q_{in}}{\rho_{up}^{(\text{syn})} - \rho_{in}}.$$  (7.4)

In this case, $q_{up}^{(\text{syn})}$ and $\rho_{up}^{(\text{syn})}$ are the flow rate and vehicle density in synchronized flow, just downstream of the upstream front of the synchronized flow region; $q_{in}$, $\rho_{in}$ are the flow rate and vehicle density in free flow upstream of this front. Within the synchronized flow region the density should be higher.
than in the initial free flow: \( \rho_{up}^{(syn)} > \rho_{in} \). Then from (7.4) we find that \( v_{up} < 0 \) if the flow rate within synchronized flow is lower than the flow rate \( q_{in} \):
\[
q_{in} > \rho_{up}^{(syn)}. \tag{7.5}
\]
Because we consider the region in the diagram of congested patterns of the highest possible flow rate \( q_{in} \) (the region labeled “MSP” in Fig. 7.2) the condition (7.5) is easily satisfied.

**Physics of Moving Synchronized Flow Pattern Emergence**

To explain the physics of MSP emergence, let us consider the above hypothetical case \( q_{on} \rightarrow 0 \) (7.1). In this case, time gaps between vehicles merging from the on-ramp lane onto the main road can be very large: vehicles following one another, which merge onto the main road from the on-ramp, can be considered as single vehicles. Each of these single vehicles causes independent additional local perturbations in flow on the main road in the vicinity of the on-ramp.

Let us now assume that one of these independent random perturbations causes an F→S transition at the on-ramp. At \( q_{on} \rightarrow 0 \) (7.1), after the downstream front of the synchronized flow region due to the F→S transition has been formed in the vicinity of the on-ramp, there can be a long enough time interval when there are no vehicles that merge onto the main road from the on-ramp. This means that during this time interval \( q_{on} = 0 \). Then the downstream front of the synchronized flow region begins to propagate upstream. Indeed, during this time interval there are no vehicles merging from the on-ramp. Thus, during this time interval the main road can be considered as a homogeneous road (without a bottleneck due to the on-ramp).\(^6\)

In the case of a homogeneous road, we know that if a local region of synchronized flow emerges in an initial free flow, an MSP is formed (Sect. 5.2.8). Thus, in the case \( q_{on} \rightarrow 0 \) (7.1) and a high enough flow rate \( q_{in} \) we can expect MSP emergence at the on-ramp. In this case, the on-ramp plays only the role of a source of an additional random disturbance in the free flow of a high flow rate \( q_{in} \) on the main road: if an MSP due to its upstream propagation is already far upstream of the on-ramp, a new vehicle that merges onto the main road from the on-ramp cannot influence the MSP any more.

To explain the upstream propagation of the downstream front of an MSP, we use the formula
\[
v_{down} = v^{(syn)} - \frac{1}{\tau_{del, syn} \rho^{(syn)}}, \tag{7.6}
\]

\(^6\) It must be noted that this case of the upstream propagation of the downstream front of a congested pattern that has emerged at an isolated freeway bottleneck is an exceptional case. In most other situations (at higher flow rates \( q_{on} \)), as mentioned in Sect. 7.2, the downstream front of congested patterns is spatially fixed at the bottleneck.
which has the same physical meaning as those of (5.27) (see footnote 6 of Sect. 5.2.7). In (7.6), \( v^{(\text{syn})} \) and \( \rho^{(\text{syn})} \) are the average vehicle speed and vehicle density in synchronized flow just upstream of the downstream front of synchronized flow; \( \tau_{\text{del, syn}}^{(a)} \) is the average time delay in vehicle acceleration as in (5.27). However, this time delay is related to the vehicle speed \( v^{(\text{syn})} \).

If the time delay in vehicle acceleration \( \tau_{\text{del, syn}}^{(a)} \) is low enough, the velocity of the downstream front of synchronized flow (7.6) is negative. It should be noted that values \( v^{(\text{syn})} \) and \( \tau_{\text{del, syn}}^{(a)} \) in (7.6) can change within an MSP over time and they can also be very different for different MSPs. Thus, in contrast to the velocity \( v_d \) of the downstream front of a wide moving jam that is a characteristic parameter (Sects. 2.4.1, 3.2.6, and 4.2.1), the velocity \( v_{\text{down}} \) (7.6) is not a characteristic parameter.

Thus, both downstream and upstream fronts of synchronized flow move upstream, i.e., an MSP occurs (Fig. 7.3). The reason for emergence of this MSP at the on-ramp is different from MSP emergence on a homogeneous road. However, after the MSP that has emerged at the on-ramp is far upstream of the on-ramp, the subsequent development and propagation of the MSP is not different from the case of an MSP on a homogeneous road discussed in Sect. 5.2.8.

As in the case of a wide moving jam where the equality in the velocities of the downstream and upstream jam fronts (3.22) determines the threshold of wide moving jam emergence, the condition

\[
v_{\text{down}} = v_{\text{up}}
\]  

(7.7)

determines the threshold of MSP emergence. To explain this, note that if \( |v_{\text{up}}| > |v_{\text{down}}| \) the width of an MSP (in the longitudinal direction) should gradually increase. Otherwise, if \( |v_{\text{up}}| < |v_{\text{down}}| \) the width of the MSP should decrease. The condition (7.7) together with (7.4) and (7.6) enables us to
estimate the threshold flow rate in free flow for an $F \rightarrow S$ transition. To do this, we take into account that at the threshold point $q_{in} = q_{th}$ and $\rho_{in} = \rho_{th}$. Then we obtain

$$q_{th} = \frac{1}{T_{del, syn}^{(a)}} \frac{1 - \rho_{th}/\rho^{(syn)}}{1 - v^{(syn)}/v_{th}},$$  

(7.8)

where $v^{(syn)} = q^{(syn)}/\rho^{(syn)}$, $v_{th} = q_{th}/\rho_{th}$, it is assumed that the speed and density do not depend on the spatial coordinate within the MSP. However, there can be an infinite number of sets of the values $(\rho^{(syn)}, q^{(syn)})$ for different MSPs that can satisfy the condition (7.7). Thus, to find $q_{th}$, a set of these values should be chosen that is related to the lowest solution for $q_{th}$ in (7.8). For a rough estimation of the flow rate $q_{th}$ from (7.8) one can take into account that $v^{(syn)} < v_{th}$ and $\rho^{(syn)} > \rho_{th}$:

$$q_{th} \approx \frac{1}{T_{del, syn}^{(a)}}.$$

(7.9)

At the maximum point of free flow $(\rho^{(free)}_{\text{max}}, q^{(free)}_{\text{max}})$ on a homogeneous road a driver accepts a lower mean gross time gap $T_{\text{min}}^{(free)}$ (3.24) than the mean time delay $T_{del, syn}^{(a)}$ in vehicle acceleration at the downstream front of synchronized flow:

$$T_{\text{min}}^{(free)} < T_{del, syn}^{(a)}.$$

(7.10)

This driver behavior can explain two conclusions of the theory of $F \rightarrow S$ transitions of Sect. 5.2.4:

(i) The flow rate at the maximum point of free flow $q^{(free)}_{\text{max}} = 1/T_{\text{min}}^{(free)}$ is higher than $q_{th}$ (7.9):

$$q_{th} < q^{(free)}_{\text{max}}.$$

(7.11)

(ii) There is a range of the flow rate (density) (5.13) where free flow is metastable with respect to an $F \rightarrow S$ transition.

**Numerical Example of MSP**

The latter conclusions are confirmed by numerical simulations of an MSP at the on-ramp bottleneck\(^7\) shown in Fig. 7.4. Corresponding to the above discussion, this MSP emerges only in the case when the flow rate $q_{in}$ is high enough and the flow rate $q_{on}$ is very low.

It can be seen that a region of synchronized flow that has first emerged at the on-ramp departs from the bottleneck, i.e., an MSP appears. The MSP propagates further upstream as a whole local structure. Features of this MSP do not differ from an MSP on a homogeneous road (Sect. 5.2.8).

\(^7\) An empirical example of an MSP is shown in Fig. 2.16 of Sect. 2.4.6 and in Fig. 9.10 of Sect. 9.3.3.
7.4 Synchronized Flow Patterns

Fig. 7.4. Results of numerical simulations of a moving synchronized flow pattern (MSP) at an on-ramp [329]. (a) Vehicle speed as a function of time and distance. (b) Speed (left) and flow rate (right) as functions of time at some locations upstream of the on-ramp that is at the location $x = 16 \text{ km}$. At $t = t_0$ the on-ramp inflow has been switched. Taken from [329]

7.4.3 Pinning of Downstream Front of Synchronized Flow at Bottlenecks

Let the flow rate in free flow $q_{in}$ be as high as necessary for MSP emergence on the main road at the on-ramp bottleneck discussed above. If now the flow rate $q_{on}$ increases, some new phenomena occur. In this case, time intervals between vehicles merging onto the main road from the on-ramp decrease due to higher flow rate to the on-ramp $q_{on}$.

Therefore, at a high enough flow rate $q_{on}$ we cannot consider merging vehicles as single independent vehicles that make independent disturbances in flow on the main road. These vehicles cause a permanent disturbance in free flow on the main road at the bottleneck: a deterministic local perturbation appears in the vicinity of the bottleneck (Sect. 5.3).

This deterministic perturbation exists permanently on the main road at the on-ramp bottleneck. Thus, it can be expected that the downstream front of synchronized flow, which occurs due to the $F \rightarrow S$ transition at the on-ramp, should be pinned at the on-ramp (Fig. 7.5a) rather than this front moves upstream as in the case of MSP emergence above.
The physics of this catch effect is related to the frequent merging of vehicles from the on-ramp onto the main road. These vehicles, causing a permanent disturbance in free flow on the main road, maintain the synchronized flow state on the main road in the vicinity of the on-ramp. As a result, the downstream front of synchronized flow is fixed in the vicinity of the bottleneck. This means that the velocity of the downstream front of synchronized flow $v_{\text{down}} = 0$.

**Physics of Widening SP**

Corresponding to the Stokes shock-wave formula (3.5), the velocity $v_{\text{down}}$ can be written as

$$v_{\text{down}} = \frac{q^{(\text{syn})} - q^{(\text{bottle})}}{\rho^{(\text{syn})} - \rho^{(\text{bottle})}}.$$  \hspace{1cm} (7.12)

In this formula, $q^{(\text{syn})}$ and $\rho^{(\text{syn})}$ are the flow rate and average density in synchronized flow just upstream of the downstream front of synchronized flow. Recall that within the downstream front of synchronized flow vehicles accelerate from synchronized flow upstream to free flow downstream. The flow rate in this free flow is called the discharge flow rate, $q_{\text{out}}^{(\text{bottle})}$. Thus, the discharge flow rate is the flow rate in free flow on the main road downstream of the downstream front of synchronized flow. The density $\rho_{\text{out}}^{(\text{bottle})}$ is the vehicle density in free flow in which the flow rate $q_{\text{out}}^{(\text{bottle})}$. Here and below we should take into account the remark of footnote 9 of Sect. 5.3.1.
The vehicle density within synchronized flow on the main road in the vicinity of the on-ramp is different from the density in free flow on the main road downstream of synchronized flow. Thus, the condition \( v_{\text{down}} = 0 \) can be satisfied only if

\[
q_{\text{out}}^{(\text{bottle})} = q^{(\text{syn})}.
\]  

(7.13)

Because the downstream front of synchronized flow on the main road is fixed at the bottleneck, the flow rate within this front of synchronized flow cannot be a function of spatial coordinate (Fig. 7.5b): this flow rate is the same in synchronized flow just upstream of the front and in free flow just downstream of the front (7.13).

Note that the above conclusions about spatial independence of the total flow rate on a freeway location under congestion conditions at bottlenecks, the behavior of the discharge flow rate, and a spatial dependence of vehicle speed downstream of congested bottlenecks are well-known effects, which have intensively been studied and discussed in the literature (see references in the papers by Persaud and Hurdle [59] and by Hall, Hurdle, and Banks [30]). The velocity of the upstream front of synchronized flow is given by (7.4).

Within the upstream front of synchronized flow vehicles slow down from free flow upstream of the front to synchronized flow downstream of the front. The flow rate \( q_{\text{in}} \) is assumed as high as in the case of MSPs discussed in Sect. 7.4.2. In other words, we assume that the condition (7.5) is also satisfied. For this reason, the velocity of the upstream front of synchronized flow \( v_{\text{up}} < 0 \). Because \( v_{\text{down}} = 0 \) and \( v_{\text{up}} < 0 \), the region of synchronized flow is continuously widening upstream over time. This means that a widening SP (WSP) occurs. Thus, at a higher flow rate \( q_{\text{on}} \), rather than an MSP, an WSP should be formed. For these reasons, there should be a boundary (labeled \( M \) in Fig. 7.2) at a higher flow rate \( q_{\text{on}} \) and a high enough flow rate \( q_{\text{in}} \) that separates the region of MSPs and the region of WSPs in the diagram of congested patterns.

**Numerical Simulation of WSP**

This qualitative consideration is confirmed by numerical simulations of an WSP at an isolated bottleneck\(^8\) within a microscopic three-phase traffic theory (Fig. 7.6).

It can be seen that at the downstream front vehicles accelerate from synchronized flow on the main road upstream of the on-ramp to free flow downstream of the on-ramp. This downstream front of an WSP is fixed at the on-ramp bottleneck. At the upstream front of the WSP vehicles must slow down: within this front free flow upstream transforms into synchronized flow within the WSP downstream of the front. The upstream front moves continuously upstream.

\(^8\) Empirical examples of WSPs are shown in Fig. 2.17 of Sect. 2.4.6 and Fig. 9.5 of Sect. 9.3.1.
7.4.4 Transformation Between Widening and Localized Synchronized Flow Patterns

If the flow rate $q_{in}$ is lower than for the above case of an WSP but the flow rate to the on-ramp $q_{on}$ remains high, the downstream front of synchronized flow is also fixed at the on-ramp bottleneck. However, the lower the flow rate $q_{in}$, the lower the absolute value $|v_{up}|$ should be of the negative velocity of the upstream front of the WSP, $v_{up}$. This follows from the formula for this velocity (7.4). Thus, there should be a relatively low flow rate $q_{in}$ when the velocity of the upstream front of synchronized flow is zero: $v_{up} = 0$, i.e., rather than the WSP, a localized synchronized flow pattern (LSP) occurs. This means that when the flow rate $q_{in}$ decreases, the WSP should transform into the LSP. Therefore, at a lower flow rate $q_{in}$ and at a high enough flow rate $q_{on}$ there should be a boundary (labeled $W$ in Fig. 7.2), which separates the region of WSPs and the region of LSPs in the diagram of congested patterns.

In the case of an LSP, corresponding to the formula (7.4) rather than the condition (7.5) on average we have at the upstream front of the LSP

$$q_{in} = q_{up}^{(syn)}.$$  

(7.14)
7.4 Synchronized Flow Patterns

Fig. 7.7. Results of numerical simulations of a localized synchronized flow pattern (LSP) at an on-ramp bottleneck. (a) Vehicle speed on the main road as a function of time and distance. (b) Speed (left) and flow rate (right) as functions of time at some locations upstream of the on-ramp, which is at the location $x = 16$ km. At $t = t_0$ the on-ramp inflow has been switched. (c) Density as a function of time at some location (the dashed line is related to the density at the limit point of free flow, $\rho_{(free)}^{\text{(free)}}$, see Fig. 4.4a). (d) LSP width (in the longitudinal direction) $L_{\text{LSP}}$ as a function of time. Taken from [329]

Numerical Simulation of LSP

These conclusions are confirmed by numerical simulations of LSPs at an isolated bottleneck\(^9\) due to the on-ramp that are made within a microscopic three-phase traffic theory (Fig. 7.7).

It can be seen that at the downstream front vehicles accelerate from synchronized flow on the main road upstream of the on-ramp to free flow on the main road downstream of the on-ramp. This downstream front of an LSP is fixed at the on-ramp. At the upstream front of the LSP vehicles must slow down when they reach synchronized flow in the LSP (Figs. 7.7a,b). The flow

---

\(^9\) Empirical examples of LSPs are shown in Fig. 2.14 of Sect. 2.4.5 and Fig. 9.9 of Sect. 9.3.2.
rate in the initial free flow $q_{in}$ and the flow rate in synchronized flow in the LSP $q_{up}^{(syn)}$ are on average equal to one another in accordance with (7.14) (Fig. 7.7b). However, the flow rate $q_{up}^{(syn)}$ is a function of time, i.e., the condition (7.14) is only satisfied on average. For this reason, the upstream front of the LSP possesses oscillations over time. Thus, the LSP width $L_{LSP}$ is a complicated function of time (Fig. 7.7d).

7.5 General Patterns

Right of the boundary $S_j^{(B)}$ in the diagram of congested patterns at the on-ramp wide moving jams emerge in synchronized flow of an initial SP, i.e., an GP occurs spontaneously (Fig. 7.1). However, there are a diverse multitude of various GPs that exhibit qualitatively different features. Some of these various types of GPs that occur in three-phase traffic theory will be considered below.

7.5.1 Spatiotemporal Structure of General Patterns

The general pattern (GP) is the most often observed congested pattern at an isolated effectual bottleneck [218]. An GP consists of the following parts:

(i) A region of synchronized flow: the downstream front of this synchronized flow is fixed at the bottleneck.

(ii) The pinch region within synchronized flow: in the pinch region of synchronized flow, a self-compression of synchronized flow occurs. This self-compression of synchronized flow is called the pinch effect in synchronized flow. In the pinch region of synchronized flow, the vehicle density is high and the average vehicle speed is low. However, the flow rate can be relatively high in the pinch region. The pinch region in synchronized flow is the region where narrow moving jams emerge spontaneously. The narrow moving jam is a moving jam that consists of the downstream and upstream jam fronts only. Firstly, vehicles must slow down when they reach a narrow moving jam. Then with only a small delay vehicles can accelerate from low speed states within the jam. A narrow moving jam does not possess the feature of a wide moving jam to propagate through any other states of traffic flow and through any bottleneck while maintaining the mean velocity of the downstream jam front.

(iii) The narrow moving jams that have emerged spontaneously in the pinch region of synchronized flow propagate upstream and grow in amplitude: the vehicle speed decreases and the density increases within a narrow moving jam during jam propagation in the upstream direction. Finally, some of the narrow moving jams transform into wide moving jams. This is related to S→J transitions in synchronized flow.

(iv) The region of wide moving jams, i.e., a sequence of wide moving jams. Because wide moving jams propagate continuously upstream the region of wide moving jams is widening over time.
In particular, there can be the following two types of GPs:

1. an GP in which the synchronized flow region is upstream bordered by the region of wide moving jams. The upstream front of the GP is related to the upstream front of the farthest upstream wide moving jam in the region of wide moving jams. This case is symbolically shown in Fig. 7.8a;
2. an GP in which the upstream front is related to the upstream front of synchronized flow. In this case, the region of wide moving jams is within the synchronized flow region. This case is symbolically shown in Fig. 7.8b.

Fig. 7.8. Symbolic spatial structures of an GP at an isolated effectual bottleneck at a given time. (a) GP of type (1). (b) GP of type (2)

**GP of Type (1)**

In this GP (Fig. 7.8a), the upstream front of synchronized flow coincides with the upstream front of the pinch region. This upstream front of synchronized flow is determined by the location where a narrow moving jam has just transformed into a wide moving jam, i.e., where an S→J transition occurs. The upstream front separates synchronized flow downstream from the region of wide moving jams upstream.

Because the transformation of different narrow moving jams into wide moving jams can occur at different freeway locations, the upstream front of synchronized flow can perform complex spatial oscillations over time.
The mean width $L_{\text{syn}}$ of synchronized flow in an GP is spatially limited. The width $L_{\text{syn}}$ does not depend on traffic demand on the main road upstream of the GP.

A successive process of narrow moving jam transformation into wide moving jams at the upstream front of synchronized flow causes a region of wide moving jams in the GP. Due to upstream wide jam propagation, the region of wide moving jams is continuously widening upstream. Consequently, the quantity of wide moving jams within the region of these jams can increase over time. Between wide moving jams both synchronized flow and free flow can be formed. These flows will be considered a part of the region of wide moving jams.

**GP of Type (2)**

In this GP (Fig. 7.8b), the upstream front of synchronized flow is continuously widening upstream. Therefore, the width of synchronized flow $L_{\text{syn}}$ in the GP increases over time rather than being spatially limited. In other words, in contrast to an GP of type (1), the upstream front of synchronized flow in an GP of type (2) is not determined by the location of an S→J transition (wide moving jam emergence). Firstly, synchronized flow propagates continuously upstream of the bottleneck and only later wide moving jams emerge in that synchronized flow.

Note that an GP of type (2) can persist only when the velocity of the upstream front of synchronized flow is more negative than the velocity of the upstream front of the farthest upstream wide moving jam in the GP. GPs are also possible when the spatial structure changes over time between various GP types.

**Numerical Examples of GP**

These results are confirmed by numerical simulations of GPs. In an GP of type (1), the upstream front of synchronized flow separates the pinch region where growing narrow moving jams emerge spontaneously from the region of wide moving jams (Fig. 7.9). The pinch region is related to the freeway locations $x = 15.8\, \text{km}$ and $x = 14.5\, \text{km}$ in Fig. 7.9b and the region of wide moving jams is related to $x = 9\, \text{km}$.

In an GP of type (2), the upstream front of synchronized flow is upstream of the farthest upstream wide moving jam in the GP (Figs. 7.10). The propagation of this upstream front of synchronized flow is shown on vehicle trajectories within the GP in Fig. 7.11 by the dashed line.

Empirical examples of an GP of type (1) is shown in Figs. 2.21 and 9.12 and of an GP of type (2) in Figs. 2.19 and 9.14 (see explanations in Sect. 9.4).
Fig. 7.9. Numerical simulations of an GP of type (1) at an on-ramp bottleneck. 
(a) Speed on the main road in space and time. (b) Speed (left) and flow rate (right) 
within the GP as a function of time at different freeway locations. Taken from [329]

Fig. 7.10. Numerical simulations of an GP of type (2) at an on-ramp bottleneck. 
Speed on the main road in space and time. Taken from [330]
7.5.2 Dissolving General Pattern and Pattern Transformation

Besides GPs where wide moving jams continuously emerge over time, there can be a dissolving GP (DGP). In an DGP, the process of continuous wide moving jam emergence is interrupted at some stage. As a result, a single wide moving jam or a finite number of wide moving jams emerge over time only. After the wide moving jam due to upstream jam propagation is already far upstream of a bottleneck, free flow or one of the SPs remains at the bottleneck.

To understand DGP formation, let us consider a case when a wide moving jam has already emerged within synchronized flow in an incipient GP. We assume that free flow prevails in the jam outflow and the flow rate $q_{\text{in}}$ in free flow upstream of the GP is considerably higher than the flow rate of the wide moving jam outflow:

$$q_{\text{in}} > q_{\text{out}}. \tag{7.15}$$

Now the jam outflow $q_{\text{out}}$ determines the flow rate upstream of the pinch region of the incipient GP: this jam outflow is the inflow into the pinch region of the GP. Recall that before the wide moving jam has been formed, the inflow into the pinch region has been equal to the flow rate $q_{\text{in}}$. We have assumed above that $q_{\text{in}} > q_{\text{out}}$. Thus, after the wide moving jam has been formed, the new flow rate upstream of synchronized flow in the pinch region is $q_{\text{out}}$. The latter flow rate can be considerably lower than the initial flow rate $q_{\text{in}}$. This means, that at this new lower flow rate upstream of the on-ramp $q_{\text{out}}$ synchronized flow at the on-ramp can dissolve or an LSP occurs at the on-ramp. In the first case, free flow returns at the on-ramp. In the second case, the LSP remains at the on-ramp after the wide moving jam is far upstream.
of the bottleneck. In other words, under the condition (7.15) an DGP can emerge (Fig. 7.12), rather than the process of moving jam emergence in the pinch region continuously occurring.

![Diagram of speed, time, and distance](image)

**Fig. 7.12.** Numerical simulations of an DGP at an on-ramp bottleneck. Speed on the main road in space and time. Formation of the DGP after the first wide moving jam has emerged within an initial synchronized flow upstream of the on-ramp. When the wide moving jam in the DGP is already far upstream of the bottleneck, an LSP remains at the bottleneck. Taken from [329]

There should be a region of DGPs in the diagram of congested patterns right of the boundary $S_{j}^{(B)}$ at a higher flow rate $q_{in}$ that satisfies the condition (7.15) (Fig. 7.13). This region (labeled "DGP" in Fig. 7.13) is indeed found in numerical simulations (see Fig. 7.14 below). The above conclusion about DGP formation is, however, only true if the flow rate to the on-ramp $q_{on}$ is not very high.

At a higher flow rate to the on-ramp $q_{on}$ it can be expected that the speed in the pinch region of synchronized flow upstream of the on-ramp can decrease considerably. Therefore, wide moving jams are much more likely to emerge in synchronized flow (Sect. 6.3). Thus, under the condition (7.15) there should be a boundary (labeled $G$ in Fig. 7.13) at the higher flow rate to the on-ramp $q_{on}$ that should separate DGPs (between the boundaries $S_{j}^{(B)}$ and $G$) from GPs (right of the boundary $G$).

The regions of MSPs, WSPs, and DGPs in the diagram of congested patterns (Fig. 7.13) are very close to one another; a relatively small change in the flow rate to the on-ramp $q_{on}$ can lead to a change in the pattern type in the diagram. Thus, it is expected that with real empirical conditions, where the flow rates $q_{in}$ and $q_{on}$ are not some constants, diverse pattern transformation over time between different MSPs, WSPs, and DGPs should be observed. This conclusion of three-phase traffic theory is confirmed by empirical investigations of these congested patterns at isolated bottlenecks (Chap. 14).
Fig. 7.13. Diagram of congested patterns at an on-ramp bottleneck in three-phase traffic theory with all possible patterns at the bottleneck. Taken from [218,221,222]

Numerical Diagram of Congested Patterns

These qualitative results of three-phase traffic theory about the diagram of congested patterns (Fig. 7.13) are confirmed by numerical simulations of the diagram at the on-ramp shown in Fig. 7.14.

All boundaries between different SPs and also between DGPs and GPs considered above can be found in this diagram.

7.6 Physics of General Patterns

7.6.1 Region of Wide Moving Jams

Characteristics of Wide Moving Jam Outflow

In the region of wide moving jams of an GP, wide moving jams propagate upstream from the upstream boundary of the pinch region in synchronized flow where these jams have initially emerged. For this reason, the GP expands continuously in the upstream direction over time. During wide moving jam
propagation both free flow and synchronized flow can occur between the jams in the region of wide moving jams of the GP.

Let us denote the flow rate of the outflow from a wide moving jam by \( q_{\text{out}}^{(J)} \) and the related vehicle density by \( \rho_{\text{min}}^{(J)} \). The flow rate \( q_{\text{out}}^{(J)} \) and density \( \rho_{\text{min}}^{(J)} \) of the outflow from a wide moving jam satisfy the conditions

\[
q_{\text{out}}^{(J)} \leq q_{\text{out}}, \quad \rho_{\text{min}}^{(J)} \geq \rho_{\text{min}}, 
\]

where \( q_{\text{out}} \) is the flow rate and \( \rho_{\text{min}} \) is the density of the jam outflow when free flow is formed in this jam outflow. The point \((\rho_{\text{min}}, q_{\text{out}})\) in free flow is the threshold point for an \( F \rightarrow J \) transition (Sect. 6.2). This point is related to the leftmost point on the line \( J \) in the flow–density plane (Fig. 6.1).

To explain the conditions (7.16) and (7.17), note that when synchronized flow is formed in the jam outflow, then the flow rate of this outflow \( q_{\text{out}}^{(\text{syn})} \) is lower than \( q_{\text{out}} \):

\[
q_{\text{out}}^{(\text{syn})} < q_{\text{out}} \tag{7.18}
\]

and the vehicle density \( \rho_{\text{min}}^{(\text{syn})} \) is higher than \( \rho_{\text{min}} \):

\[
\rho_{\text{min}}^{(\text{syn})} > \rho_{\text{min}} \tag{7.19}
\]

A point \((\rho_{\text{min}}^{(\text{syn})}, q_{\text{out}}^{(\text{syn})})\) in synchronized flow is one of the threshold points for an \( S \rightarrow J \) transition: there, an infinite number of different threshold densities
\( \rho_{\text{min}}^{(\text{syn})} \) and threshold flow rates \( q_{\text{out}}^{(\text{syn})} \) exist (Sect. 6.3). Each of the threshold points \((\rho_{\text{min}}^{(\text{syn})}, q_{\text{out}}^{(\text{syn})})\) lies on the line \( J \) in the flow–density plane. These points are to the right of the point \((\rho_{\text{min}}, q_{\text{out}})\) in this plane. Two of these threshold densities \( \rho_{\text{min}}^{(\text{syn})}, i, i = 1, 2 \) are shown in Fig. 6.3 for the related vehicle speeds in synchronized flow \( v_{i}^{(\text{syn})}, i = 1, 2 \).

Thus, when synchronized flow is formed in the jam outflow, we obtain

\[
q_{\text{out}}^{(J)} = q_{\text{out}}^{(\text{syn})} < q_{\text{out}} \tag{7.20}
\]

and

\[
\rho_{\text{min}}^{(J)} = \rho_{\text{min}}^{(\text{syn})} > \rho_{\text{min}} \cdot \tag{7.21}
\]

When free flow is formed in the jam outflow, we get

\[
q_{\text{out}}^{(J)} = q_{\text{out}} \tag{7.22}
\]

and

\[
\rho_{\text{min}}^{(J)} = \rho_{\text{min}} \cdot \tag{7.23}
\]

**Width of Wide Moving Jams**

The mean velocity of the downstream front of a wide moving jam \( v_{g} \) does not depend on time. This characteristic velocity is the same for all different wide moving jams in the region of wide moving jams in an GP. However, the velocity of the upstream front of a wide moving jam is not a characteristic parameter. This is because the upstream jam velocity is determined by the flow rate and density of the jam inflow. In particular, for the farthest upstream wide moving jam of the GP this jam inflow is determined by traffic demand upstream of the GP.

Let us consider two adjacent wide moving jams in the region of wide moving jams of an GP. One of these jams can be called “the upstream jam” and the other one can be called “the downstream jam.” If there are no on- and off-ramps between these jams the outflow from the upstream wide moving jam is also the inflow into the downstream jam. In this case, the velocity of the upstream front of the downstream jam depends on the characteristics of the outflow from the upstream jam. The difference between the velocities of the downstream and upstream fronts of the jam determines the variation of the jam width (in the longitudinal direction) over time. Thus, the width of a wide moving jam during jam propagation depends on the position of the jam in comparison with other wide moving jams of the GP.

The dependence of the width of a wide moving jam on time is determined from the obvious equation:

\[
\frac{dL_{j}}{dt} = v_{g} - v_{g}^{(\text{up})} \cdot \tag{7.24}
\]
where \( v_g \) is the velocity of the downstream jam front, \( v_{g(\text{up})} \) is the velocity of the upstream jam front. Both velocities are negative.

However, the velocity \( v_g \) is the characteristic parameter whose mean value is constant during wide moving jam propagation. This mean velocity is given by (3.10). In contrast, the velocity \( v_{g(\text{up})} \) is not a characteristic parameter. This velocity is determined through the Stokes shock-wave formula (3.5). If we take into account the flow rate \( q_{\text{in},J} \) and vehicle density \( \rho_{\text{in},J} \) in traffic flow just upstream of the wide moving jam, we obtain:

\[
v_{g(\text{up})} = -\frac{q_{\text{in},J}}{\rho_{\text{max}} - \rho_{\text{in},J}},
\]

(7.25)

where \( \rho_{\text{max}} \) is the density within the jam (the jam density) and it is assumed that the flow rate within the jam is equal to zero. Thus, from (7.24), (3.10) and (7.25) we find

\[
L_J(t) = L_J(t_0) + \int_{t_0}^{t} \left( \frac{q_{\text{in},J}}{\rho_{\text{max}} - \rho_{\text{in},J}} - \frac{1}{\rho_{\text{max}} x_{\text{del}}^{(a)}} \right) \, dt.
\]

(7.26)

Let us consider an GP at a time when only one wide moving jam has emerged in the GP. We call this jam the “first” wide moving jam of the GP. In an GP of type (1), the first wide moving jam is at the upstream boundary of synchronized flow. This first wide moving jam can be considered as a region where “superflous” vehicles, which cannot immediately pass through the pinch region of synchronized flow in the GP, are virtually stored: the average flow rate upstream of the wide moving jam \( q_{\text{in},J} = q_{\text{in}} \) (\( q_{\text{in}} \) is the flow rate in free flow upstream of the GP) is assumed to be higher than the mean flow rate in the pinch region \( q^{(\text{pinch})} \). The wide moving jam interrupts flow. Due to the difference between \( q_{\text{in}} \) and \( q^{(\text{pinch})} \) the width of the jam \( L_J \) increases. In other words, there is no influence of the inflow rate \( q_{\text{in}} \) on the width of synchronized flow in the GP and on other parameters of synchronized flow downstream of the jam even if the flow rate \( q_{\text{in}} \) is very high for a long time.

Note that this interruption of traffic flow is a general effect of each wide moving jam. This traffic flow interruption means that traffic flow upstream of a wide moving jam has no influence on flow downstream of the jam. The width of the wide moving jam \( L_J \) only changes when the characteristics of flow upstream change over time. In contrast, narrow moving jams and other states of synchronized flow do not interrupt traffic flow upstream and downstream. For this reason, the interruption flow effect is an additional criterion to distinguish between wide moving jams and synchronized flow in congested traffic.

In an GP of type (2), the first wide moving jam can also be considered as a region where “superflous” vehicles, which cannot immediately pass through the pinch region of synchronized flow in the GP, are virtually stored. However, in this case, the average flow rate upstream of the wide moving
jam $q_{in,j}$ is lower than $q_{in}$. This is because upstream of the wide moving jam rather than free flow, synchronized flow occurs. Thus, at the same flow rate $q_{in}$ the increase in the jam width $L_J$ (7.26) can occur over time appreciably slower than for the first wide moving jam in the GP of type (1).

Later, a new wide moving jam is formed in the GP. This “second” wide moving jam is downstream of the first wide moving jam. Thus, the first wide moving jam is the upstream jam and the second jam is the downstream wide moving jam in the GP. The outflow from the upstream wide moving jam is the inflow into the downstream wide moving jam. For this reason, it can turn out that the width of the downstream wide moving jam does not increase over time. This case is approximately realized in the empirical example of two wide moving jams shown in Fig. 2.7a, where the width of the downstream jam does not change appreciably during jam propagation. This conclusion can also be valid for other downstream wide moving jams in the region of wide moving jams of the GP.

**Width of Wide Moving Jam Propagating Through Bottlenecks**

In real traffic flow, wide moving jams can propagate many kilometers away from the pinch region of an GP where the jams have initially emerged. Thus, the wide moving jams can reach an adjacent upstream effectual bottleneck on the freeway. In this case, they are called foreign wide moving jams (Sect. 2.4.9). Wide moving jams propagate through any bottleneck while maintaining their downstream front velocity, $v_g$ (Sect. 4.2.1). Let us assume that the upstream bottleneck is due to an on-ramp and free flow is realized at this bottleneck. After the farthest upstream wide moving jam of the GP is already upstream of the on-ramp, then the flow rate of the inflow into the downstream wide moving jam of the GP increases. This increase is due to vehicles that merge from the on-ramp onto the main road. In other words, this jam inflow is determined by the sum of the flow rate of the outflow from the upstream wide moving jam plus the flow rate of the vehicles merging onto the main road from the on-ramp. For this reason, the width of the downstream moving jams can increase when the jams propagate on the main road through the bottleneck due to the on-ramp (see wide moving jams in Fig. 2.23).

A qualitatively different case can occur when wide moving jams propagate on the main road through a bottleneck due to an off-ramp. In this case, the width of wide moving jams can decrease and the jams can even dissolve (see empirical examples of this case in Sect. 14.3.2).

**7.6.2 Narrow Moving Jam Emergence in Pinch Region**

In the pinch region of synchronized flow, there is a compression of synchronized flow. Due to this pinch effect in synchronized flow, the vehicle density is high and the speed is low. This compression of synchronized flow in the
pinch region causes another effect: the speed and density within this synchronized flow are related to traffic flow states that lie above the line $J$ in the flow–density plane (Fig. 6.3a).

As has been shown in Sect. 6.3, these states of synchronized flow are metastable states with respect to moving jam emergence. In this case, local fluctuations of finite amplitude in real synchronized flow can begin to grow precipitously when their amplitude exceeds some critical value. Because of the low speed in the pinch region the probability of this precipitous growth is high.

As a result, growing narrow moving jams emerge in the pinch region of synchronized flow. The higher the density in the pinch region, the lower the speed in the pinch region and therefore the lower the critical amplitude of the critical perturbations (Fig. 6.3b). Thus, the higher the density, the higher the frequency of narrow moving jam emergence in the pinch region.

To explain the physics of the pinch effect and moving jam emergence in synchronized flow, let us consider a hypothetical case when only one narrow moving jam emerges within the pinch region during the time interval of jam propagation through the pinch region (Fig. 7.15).

At $t = t_0$ there is a wide moving jam at the upstream boundary of the pinch region $x = x_{\text{up}}^{\text{(pinch)}}$ (down-arrow 1 in Fig. 7.15b at $t = t_0$). It is assumed that at $t = t_0$ there are no narrow moving jams within the pinch region whose downstream boundary is at $x = x_{\text{down}}^{\text{(pinch)}}$ upstream of the effective location of the effectual bottleneck at $x = x_{\text{eff}}^{\text{(bottle)}}$. A narrow moving jam can emerge at a later time $t = t_1 > t_0$ (down-arrow 2) only if the flow rate of the inflow into the pinch region, $q_{\text{in}}^{\text{(pinch)}}$ at $x = x_{\text{up}}^{\text{(pinch)}}$, is higher than the flow rate of the outflow from the pinch region, $q_{\text{out}}^{\text{(pinch)}}$ at $x = x_{\text{down}}^{\text{(pinch)}}$:

$$q_{\text{in}}^{\text{(pinch)}} > q_{\text{out}}^{\text{(pinch)}} , \quad t_1 \leq t \leq t_2 .$$

This follows from the vehicle balance equation (3.1). Integrating this equation over the pinch region under the condition (7.27), we get

$$\frac{\partial N}{\partial t} = q_{\text{in}}^{\text{(pinch)}} - q_{\text{out}}^{\text{(pinch)}} > 0 , \quad t_1 \leq t \leq t_2 ,$$

where

$$N = \int_{x_{\text{up}}^{\text{(pinch)}}}^{x_{\text{down}}^{\text{(pinch)}}} \rho dx ;$$

$t_1 \leq t \leq t_2$ is the time interval when the narrow moving jam is within the pinch region (Fig. 7.15b).

Corresponding to Fig. 7.15b, during the time interval $t_1 \leq t \leq t_2$ the flow rate into the pinch region $q_{\text{in}}^{\text{(pinch)}}$ is equal to the flow rate of the wide moving jam outflow, $q_{\text{out}}^{(J)}$:

$$q_{\text{in}}^{\text{(pinch)}} = q_{\text{out}}^{(J)} , \quad t_1 \leq t \leq t_2 .$$
Fig. 7.15. Explanation of the physics of moving jam emergence in the pinch region of synchronized flow. (a) Symbolic spatial structures of an GP at a given time taken from Fig. 7.8a. (b) Qualitative spatial distributions of the vehicle speed in the pinch region at different times. (c) Qualitative illustration of the shift of states of synchronized flow in the flow–density plane due to the self-compression of synchronized flow in the pinch region. (d) Qualitative illustration of the downstream front (line $J^{(\text{down})}_{\text{narrow}}$) and of the upstream front (line $J^{(\text{up})}_{\text{narrow}}$) of the narrow moving jam in (b) (up-arrow 2) at $t = t_2$ in the flow–density plane. F: free flow, J: the line $J$.
When synchronized flow prevails in the jam outflow (Fig. 7.15), the flow rate of the jam outflow \( q_{\text{out}}^{(J)} \) (7.20) should be related to a point on the line \( J \) in the flow–density plane (point labeled “Jam outflow” in Fig. 7.15c). On the other hand, in Sect. 6.3 it has been shown that only points of synchronized flow, which lie above the line \( J \) in the flow–density plane, are metastable states where moving jams can emerge spontaneously. Points on the line \( J \) are the threshold points for moving jam emergence where moving jams can occur only if a local perturbation with a very high amplitude appears in synchronized flow. For this reason, synchronized flow within the pinch region should be related to points labeled “Pinch” in Fig. 7.15c,d, which are shifted to the right from the point labeled “Jam outflow” in the flow–density plane (this shift is symbolically shown by dashed arrow in Fig. 7.15c). This shift is related to an increase in density and a decrease in speed at the same flow rate, i.e., to a compression of synchronized flow. This should explain why moving jams emerge spontaneously after the pinch effect, i.e., the self-compression of synchronized flow has occurred.

Corresponding to (7.28), vehicles are stored within a narrow moving jam whose amplitude (i.e., the difference between the speed within the jam and the speed away from the jam) grows over time \( t = t_1, t_2 \) in Fig. 7.15b). The absolute value of the velocity of the upstream front of the narrow moving jam, \( |v_{\text{narrow}}^{(\text{up})}| \), should be slightly higher than the absolute value of the velocity of the downstream front of the narrow moving jam, \( |v_{\text{narrow}}^{(\text{down})}| \): \( |v_{\text{narrow}}^{(\text{up})}| > |v_{\text{narrow}}^{(\text{down})}| \). The upstream and downstream fronts of the narrow moving jam can approximately be presented by lines in the flow–density plane (line \( J_{\text{narrow}}^{(\text{up})} \) and line \( J_{\text{narrow}}^{(\text{down})} \) for the related fronts in Fig. 7.15d). The slopes of these lines are equal to the velocities of the upstream and downstream narrow moving jam fronts, respectively. Values of these negative velocities change over time. Therefore, positions of the lines \( J_{\text{narrow}}^{(\text{up})} \) and \( J_{\text{narrow}}^{(\text{down})} \) in Fig. 7.15d depend on time. However, because states of synchronized flow upstream of the narrow moving jam are related to points, which lie above the line \( J \) (points labeled “Pinch” in Fig. 7.15c,d), the narrow moving jam propagates upstream with the negative velocity whose absolute value can be higher than the absolute value of the characteristic velocity of the downstream front of a wide moving jam, \( |v_{\text{g}}| \).

Finally, the narrow moving jam transforms into a wide one at some \( t = t_4 \), later a new narrow moving jam emerges in the pinch region, and so on. During the whole time interval \( t_1 \leq t \leq t_4 \) of narrow moving jam emergence and transformation into a wide moving jam, the condition

\[
\int_{t_1}^{t_4} \left( q_{\text{in}}^{(\text{pinch})} - q_{\text{out}}^{(\text{pinch})} \right) \, dt = 0
\]  

(7.31)

is satisfied. In other words, during the time interval \( t_1 \leq t \leq t_4 \) the whole number of vehicles within the pinch region does not change. This is because
within this time interval there is a finite time interval when the narrow moving jam propagates through the location of the upstream boundary of the pinch region (see $t = t_3$ in Fig. 7.15b). During this jam propagation through the upstream boundary of the pinch region, rather than formula (7.30) the following formula is valid:

$$q_{\text{in}}^{(\text{pinch})} \ll q_{\text{out}}^{(J)}.$$  

(7.32)

As a result, the flow rate of the inflow into the pinch region is considerably lower than the flow rate of the pinch region outflow:

$$q_{\text{in}}^{(\text{pinch})} \ll q_{\text{out}}^{(\text{pinch})}.$$  

(7.33)

This condition is the opposite one to the condition (7.27). This explains formula (7.31). As follows from (7.30), (7.32), and (7.16) the average flow rate in the pinch region

$$q^{(\text{pinch})} = \frac{1}{t_4 - t_1} \int_{t_1}^{t_4} q_{\text{in}}^{(\text{pinch})}(t) dt$$  

(7.34)

satisfies the condition

$$q^{(\text{pinch})} < q_{\text{out}}^{(J)} \leq q_{\text{out}}.$$  

(7.35)

Thus, the average flow rate of the outflow from the farthest downstream wide moving jam in an GP $q_{\text{out}}^{(J)}$ is higher than the average flow rate in the pinch region $q^{(\text{pinch})}$. The difference between these flow rates determines narrow moving jam emergence within the pinch region.

The condition (7.35) can also be explained in another way. Let us average the flow rate in the region of wide moving jams of an GP over a large time interval

$$T_{\text{av}} > T_{\text{J}}^{(\text{wide})}.$$  

(7.36)

In (7.36), $T_{\text{J}}^{(\text{wide})}$ is the mean time between the downstream fronts of wide moving jams in the GP. This averaged flow rate in the region of wide moving jams, $q_{\text{J}}^{(\text{wide})}$, is approximately

$$q_{\text{J}}^{(\text{wide})} = q_{\text{out}} \frac{T_{\text{int}}^{(\text{wide})}}{T_{\text{J}}^{(\text{wide})}},$$  

(7.37)

where

$$T_{\text{int}}^{(\text{wide})} = T_{\text{J}}^{(\text{wide})} - \tau_{\text{J}},$$  

(7.38)

$\tau_{\text{J}}$ is the mean duration of the wide moving jams.\(^{11}\) We see that

$$q_{\text{J}}^{(\text{wide})} < q_{\text{out}}.$$  

(7.39)

\(^{11}\)To explain the terms “the mean time between the downstream fronts of wide moving jams” $T_{\text{J}}^{(\text{wide})}$, “the mean duration of wide moving jams” $\tau_{\text{J}}$, and “the
For the large time averaging interval $T_{av}$ (7.36) we can approximate
\[ q^{(\text{pinch})} = q^{(\text{wide})}. \] (7.40)

The relations (7.39) and (7.40) can explain (7.35) under the condition (7.36).

It should be noted that the above qualitative explanation is a very simplified one. In reality, locations of the upstream and downstream boundaries of the pinch region can be complicated functions of time. This concerns essentially the upstream boundary of the pinch region because narrow moving jams can transform into wide ones at different locations. Moreover, during the time interval $t_1 \leq t \leq t_4$ rather than only one narrow moving jam, there can be many different narrow moving jams, which have randomly emerged within the pinch region. However, if distances between these narrow moving jams are low enough, then only some of these narrow jams can transform into wide ones. This is related to the wide moving jam suppression effect.

7.6.3 Moving Jam Suppression Effect

When a wide moving jam occurs at the upstream boundary of the pinch region, the wide moving jam suppresses further growth of narrow moving jams that are very close to the downstream front of this wide moving jam. As a result of this jam suppression effect, some of the narrow moving jams can disappear without transformation into a wide moving one [208].

The physics of the jam suppression effect is related to the metastability of those synchronized flow states in the pinch region that are above the line $J$ in the flow–density plane (Fig. 6.3a). A narrow moving jam can grow in the pinch region only if the synchronized flow state just upstream of the jam is related to one of these metastable states. Corresponding to the explanation of Fig. 6.4 made in Sect. 6.3, in this case, the velocity of the upstream front $v^{(\text{up})}_{\text{narrow}}$ and of the downstream front of the narrow moving jam $v^{(\text{down})}_{\text{narrow}}$ can satisfy the condition $|v^{(\text{up})}_{\text{narrow}}| > |v^{(\text{down})}_{\text{narrow}}|$ that is necessary for the jam growth.

It has been mentioned that the outflow from the farthest downstream wide moving jam is also the inflow into a narrow moving jam in the pinch...
region downstream of the wide moving jam \((t = t_1\) in Fig. 7.15b). There can be two different cases:

(i) If this narrow moving jam is close enough to the downstream front of the wide moving jam, then the synchronized flow state in the narrow moving jam inflow cannot usually be related to a metastable point above the line \(J\) in the flow–density plane. Directly downstream of the wide moving jam the state of traffic flow is related to a point on the line \(J\) in the flow–density plane (point labeled “Jam outflow” in Fig. 7.15c) rather than a point above the line \(J\). Thus, this narrow moving jam cannot grow.

(ii) In contrast, if a narrow moving jam is far enough from the downstream front of the wide moving jam, then the synchronized flow state in the narrow moving jam inflow can be related to a metastable point above the line \(J\) in the flow–density plane. This occurs at the same flow rate of the outflow from the farthest downstream wide moving jam due to the compression of synchronized flow in the pinch region (this compression effect is symbolically shown by the dashed arrow in Fig. 7.15c). Therefore, the narrow moving jam can grow. Thus, if the frequency of narrow moving jam emergence in the pinch region of GP is low enough, then each narrow moving jam can transform into a wide moving jam. In this case, the jam suppression effect does not occur.

7.6.4 Width of Pinch Region

Due to the growth of the amplitude of a narrow moving jam, there is a time interval required for the transformation of the growing narrow moving jam into a wide moving jam. We denote by \(T_{\text{narrow}}\) the time interval required for the transformation of a narrow moving jam into a wide moving jam. For simplicity we neglect a difference between the velocities of the downstream and upstream fronts of the narrow moving jam. We assume that the velocity of the narrow moving jam is equal to \(v_{\text{narrow}}\). Then the width of the pinch region (in the longitudinal direction) (Fig. 7.15)

\[
L_{\text{syn}}^{(\text{pinch})} = x_{\text{down}}^{(\text{pinch})} - x_{\text{up}}^{(\text{pinch})} \tag{7.41}
\]

over which the narrow moving jam moves on the freeway before the jam transforms into a wide moving jam is

\[
L_{\text{syn}}^{(\text{pinch})} = |v_{\text{narrow}}| T_{\text{narrow}}. \tag{7.42}
\]

Because narrow moving jams can transform into wide moving jams at different freeway locations, the width of the pinch region \(L_{\text{syn}}^{(\text{pinch})}\) (7.42) is a random value. Thus, the mean value of this width should determine the mean width of the pinch region in an GP (Fig. 7.8). The mean width (in the longitudinal direction) of synchronized flow in the pinch region, which we denote by \(L_{\text{syn}}^{(\text{mean})}\), can be estimated by the formula

\[
L_{\text{syn}}^{(\text{mean})} = |v_{\text{narrow}}| T_{\text{narrow}}^{(\text{mean})}. \tag{7.43}
\]
7.6 Physics of General Patterns

where \( v^{(\text{mean})}_{\text{narrow}} \) and \( T^{(\text{mean})}_{\text{narrow}} \) are the mean values of \( v_{\text{narrow}} \) and \( T_{\text{narrow}} \), respectively.

At a given bottleneck, the type of the GP depends on the flow rates \( q_{\text{on}} \) and \( q_{\text{in}} \). Let us explain GPs of type (1) and (2) when the flow rate \( q_{\text{on}} \) is a given value. We consider the time interval of GP formation when wide moving jams have not occurred in synchronized flow upstream of the bottleneck. Then the velocity of the upstream front of synchronized flow \( v_{\text{up}} \) is given by the formula (7.4). If the flow rate \( q_{\text{in}} \) is high enough the absolute value of this negative velocity \( |v_{\text{up}}| \) can be high: there is a quick widening of synchronized flow upstream of the bottleneck. In this case, the first wide moving jam in an GP can occur downstream of the upstream front of synchronized flow. As a result, the GP of type (2) (Fig. 7.8b) should be expected where the upstream front of synchronized flow is upstream of the farthest upstream wide moving jam.

If in contrast the flow rate \( q_{\text{in}} \) is relatively low, then \( |v_{\text{up}}| \) is not high. In this case, the effect of the propagation of synchronized flow upstream of the farthest upstream wide moving jam in an GP should not occur. As a result, the GP of type (1) (Fig. 7.8a) should be expected. The upstream front of synchronized flow in the GP of type (1) coincides with the upstream boundary of the pinch region of the GP. This boundary is determined by the location of an S→J transition, i.e., by the location where a narrow moving jam has just transformed into a wide one.

7.6.5 Wide Moving Jam Propagation Through Bottlenecks

After a wide moving jam has emerged in an GP, the jam propagates upstream. When the wide moving jam is far from the pinch region of the GP, this upstream propagation of the wide moving jam is regardless of subsequent GP development. On real freeways there can be many adjacent bottlenecks. Thus, due to upstream jam propagation the wide moving jam can approach a freeway bottleneck upstream of the bottleneck where the GP occurred. In accordance with the objective criteria for wide moving jams (Sect. 4.2.1), the wide moving jam should propagate through this upstream bottleneck while maintaining the downstream jam front velocity, \( v_{\text{g}} \). This velocity can be found from (3.10):

\[
v_{\text{g}} = -\frac{1}{\rho_{\text{max}} r^{(a)}_{\text{del}}}.
\]  

(7.44)

When a wide moving jam is still downstream of the bottleneck (Fig. 7.16a), then the downstream jam front propagates upstream with the velocity \( v_{\text{g}} \). Let us first assume that free flow is formed in the jam outflow, i.e., the condition (7.22) is satisfied. In this case, the flow rate in free flow downstream of the jam on the main road \( q_{\text{down}} \) is (Fig. 7.16a):

\[
q_{\text{down}} = q_{\text{out}}.
\]  

(7.45)
Due to further upstream jam propagation on the main road, the wide moving jam is later in the vicinity of the on-ramp merging region of the bottleneck (Fig. 7.16b,c).

It can be seen from (7.44) that the velocity $v_g$ is always negative: the downstream jam front cannot be pinned at the bottleneck as it was the case with an WSP and an LSP (Sect. 7.4.3). To explain this, note that if the downstream jam front is within the merging region of the on-ramp there can be two cases:

(i) The upstream jam propagation causes the occurrence of two wide moving jams in the vicinity of the on-ramp: the wide moving jam on the main road and the wide moving jam in the on-ramp lane (Fig. 7.16b).

(ii) The wide moving jam is only on the main road, i.e., there is no wide moving jam in the on-ramp lane (Fig. 7.16c).

In the case of item (i), the speed of vehicles in the on-ramp lane, which are within the wide moving jam, is zero. Within the downstream jam front vehicles accelerate from a stationary position within the jam in the on-ramp lane. Vehicles within the wide moving jam on the main road accelerate within the downstream jam front from a standstill within the jam, too. As a result, the flow rate in free flow downstream of the bottleneck is
where \( q_{\text{out, sum}}^{(J)} \) is the flow rate of both jam outflows (Fig. 7.16b). The vehicle acceleration from these two wide moving jams cause upstream propagation of the downstream fronts of both jams. The velocity of this upstream propagation corresponding to (7.44) is the same on the main road and in the on-ramp lane.

In the case of item (ii) (Fig. 7.16c), the flow rate of vehicles \( q_{\text{on}}^{(\text{on})} \) that merge onto the main road can be lower than \( q_{\text{on}} \). This is because the wide moving jam on the main road in the vicinity of the on-ramp can hinder these vehicles from merging freely onto the main road (Fig. 7.16c). The latter can lead to a decrease in the speed on the main road in the vicinity of the on-ramp downstream of the downstream front of the wide moving jam.

In contrast to the case of an WSP and an LSP, where a speed reduction on the main road in the vicinity of the on-ramp leads to the pinning of the downstream front of synchronized flow, there is no such catch effect in the case of the wide moving jam. This is because traffic flow is discontinuous within the wide moving jam. Vehicles that are in a standstill within the wide moving jam can accelerate from the wide moving jam at the downstream jam front only. Due to this vehicle acceleration from the wide moving jam, the downstream jam front propagates further upstream. The velocity of this downstream jam front propagation is independent of speed reduction in the vicinity of the on-ramp. This follows from (7.44): the velocity of the downstream jam front \( v_g \) does not depend on the vehicle speed downstream of the wide moving jam as long as this speed is higher than zero. The speed reduction in the vicinity of the on-ramp can lead to synchronized flow occurrence on the main road downstream of the wide moving jam. As a result, the flow rate \( q_{\text{out}}^{(J)} \) of the wide moving jam outflow on the main road decreases in comparison with \( q_{\text{out}} \): \( q_{\text{out}}^{(J)} < q_{\text{out}} \) (7.20). In this case, the flow rate downstream of the bottleneck is

\[
q_{\text{down}} = q_{\text{out, sum}}^{(J)} + q_{\text{on}}^{(\text{on})}. \tag{7.47}
\]

Thus, independent of whether free flow or synchronized flow is formed in the jam outflow the downstream front of the wide moving jam propagates through the bottleneck upstream with a constant mean velocity \( v_g \) rather than being pinned at the bottleneck.

7.7 Conclusions

(i) In three-phase traffic theory, there are two main types of congested patterns: (1) the synchronized flow pattern (SP) that consists of the “synchronized flow” phase only, and (2) the general pattern (GP) that consists of both traffic phases in congested traffic, synchronized flow and wide moving jam.
(ii) SPs exist between the boundaries $F^{(B)}_S$ and $S^{(B)}_J$ in the diagram of congested patterns on the main road at an on-ramp bottleneck. This diagram gives regions of various congested pattern emergence at the bottleneck depending on traffic demand. At the boundary $F^{(B)}_S$ an $F \rightarrow S$ transition occurs that leads to synchronized flow emergence at the on-ramp. Right of the boundary $S^{(B)}_J$ wide moving jams emerge in the synchronized flow, i.e., different GPs exist.

(iii) There are three types of SPs:
1. Moving synchronized flow pattern (MSP).
2. Widening synchronized flow pattern (WSP).
3. Localized synchronized flow pattern (LSP).

(iv) Depending on the behavior of the upstream front of synchronized flow in an GP, there are two types of GPs. In the GP of type (1), this front separates the pinch region of synchronized flow (where growing narrow moving jams emerge) from the region of wide moving jams. In the GP of type (2), the upstream front of synchronized flow is upstream of the farthest upstream wide moving jam in the GP.

(v) Besides GPs where wide moving jams emerge continuously over time, there is a dissolving GP (DGP). In the DGP, the process of the continuous wide moving jam emerging is interrupted at some stage. As a result, a single wide moving jam or a finite number of wide moving jams emerge over time only. After the wide moving jam(s) due to upstream jam propagation are already far upstream of the bottleneck, free flow or one of the SPs remain at the bottleneck.

(vi) Diverse pattern transformations over time between different MSPs, WSPs, and DGPs can be expected at isolated bottlenecks.

(vii) In the case of a bottleneck due to an on-ramp, these different congested patterns occur depending on the flow rate on the main road upstream of the on-ramp $q_{in}$ and the flow rate to the on-ramp $q_{on}$. In particular,
1. an MSP exists at a high flow rate $q_{in}$ and a very low flow rate $q_{on}$.
2. At a higher flow rate $q_{on}$ an WSP emerges.
3. If the flow rate $q_{in}$ decreases, the upstream front of synchronized flow at the on-ramp cannot move far enough upstream of the on-ramp beginning from some low enough flow rate $q_{in}$: an LSP occurs at the bottleneck.
4. If right of the boundary $S^{(B)}_J$ and left of the boundary $G$ the flow rate $q_{in} > q_{out}$, an DGP occurs rather than an GP.

(viii) The occurrence of a wide moving jam leads to interruption of traffic flow: traffic flow upstream of the wide moving jam has no influence on flow downstream of the jam. The width of the wide moving jam only changes when the characteristics of flow upstream change over time.

(ix) The self-compression of synchronized flow in the pinch region of an GP leads to a decrease in the flow rate within the pinch region. As a result, the flow rate of the pinch region outflow, i.e., the flow rate at
the downstream boundary of the pinch region is lower than the flow rate of the outflow from the farthest downstream wide moving jam in the GP. This jam outflow is the inflow into the pinch region of the GP. The difference between the flow rates of the pinch region inflow and the pinch region outflow governs narrow moving jam emergence in the pinch region of the GP.
8 Freeway Capacity
in Three-Phase Traffic Theory

8.1 Introduction
The determination of freeway capacity is one of the important applications of any traffic theory. For this reason, there is a huge number of empirical and theoretical scientific works devoted to this topic (e.g., references in [21, 285, 286, 470]).

Here, based on the conclusions about the breakdown phenomenon (F→S transition) considered in Chap. 5 and congested pattern features (Chap. 7) we give a qualitative theory of freeway capacity and of the capacity drop based on three-phase traffic theory [208, 212, 221, 222, 330, 331].

Freeway capacity depends on whether a homogeneous road (without bottlenecks) or a freeway bottleneck is considered. A general theory of freeway capacity within the scope of three-phase traffic theory presented below will be confirmed by numerical results.

8.2 Homogeneous Road
On a homogeneous multilane road (i.e., road without bottlenecks), freeway capacity depends on which traffic phase the traffic is in (Fig. 6.1a):

(i) The maximum freeway capacity in the “free flow” phase is equal to the maximum possible flow rate in free flow, \( q_{\text{max}}^{(\text{free})} \).

(ii) The maximum freeway capacity in the “synchronized flow” phase is equal to the maximum possible flow rate in synchronized flow, \( q_{\text{max}}^{(\text{syn})} \).

(iii) The maximum freeway capacity downstream of the “wide moving jam” phase is equal to the flow rate in the wide moving jam outflow, \( q_{\text{out}} \).

Because of first-order phase transitions between the traffic phases, each of these maximum freeway capacities has a probabilistic nature.

In particular, the probabilistic nature of the freeway capacity in the “free flow” phase means the following:

(1) The maximum freeway capacity in the “free flow” phase \( q_{\text{max}}^{(\text{free})} \) is determined from the condition that the probability \( P_{FS} \) for an F→S transition during a chosen time interval \( T_{ob} \) and for a road length \( L_{ob} \) is 1. This
means that the maximum freeway capacity in free flow is found from (5.7) that can be rewritten as

\[ P_{FS} \Big| q = q_{\text{max}}^{(\text{free})} = 1. \]  

Thus, the maximum freeway capacity in free flow \( q = q_{\text{max}}^{(\text{free})} \) is defined as the flow rate in free flow at which an F→S transition occurs during the time interval \( T_{ob} \) for a given road length \( L_{ob} \) with probability 1. The maximum capacity \( q = q_{\text{max}}^{(\text{free})} \) depends on the time interval \( T_{ob} \) (at least in some range of \( T_{ob} \)) and on the road length \( L_{ob} \) where the F→S transition is studied.

(2) There is a threshold flow rate \( q_{\text{th}} \) (Sect. 5.2.4) that is lower than \( q_{\text{max}}^{(\text{free})} \) (Fig. 6.2a). At the threshold point of free flow \( (\rho_{\text{th}}, q_{\text{th}}) \) the critical perturbation, which can cause the F→S transition, reaches the maximum value. Below this threshold point, i.e., at \( q < q_{\text{th}} \) \( (\rho < \rho_{\text{th}}) \) the probability for the F→S transition \( P_{FS} = 0 \).

(3) If the flow rate in free flow \( q \) is in the range \( [q_{\text{th}}, q_{\text{max}}^{(\text{free})}] \), then the higher the flow rate in free flow \( q \), the higher the probability \( P_{FS} \) for the F→S transition. Thus, the attribute of this \textit{probabilistic freeway capacity} is the probability

\[ P_{C} = 1 - P_{FS} \]  

that free flow remains on the road of length \( L_{ob} \) during the time interval \( T_{ob} \) for observing this capacity in free flow.

However, if a bottleneck exists on the road, then more complex nonlinear phenomena determine freeway capacity in free flow. This is because due to the bottleneck the road is spatially inhomogeneous: freeway capacity of free flow can depend on freeway location (e.g., [21,286]).

### 8.3 Freeway Capacity in Free Flow at Bottlenecks

#### 8.3.1 Definition of Freeway Capacity

In three-phase traffic theory, the breakdown phenomenon at a freeway bottleneck is explained by an F→S transition at the bottleneck (Chap. 5). Consequently, freeway capacity of free flow at a bottleneck is determined by the condition that the F→S transition occurs at the bottleneck during a given time interval \( T_{ob} \) with a given probability \( P_{FS}^{(B)} \) [221]. In other words, an attribute of this \textit{probabilistic freeway capacity at a bottleneck} is the probability

\[ P_{C}^{(B)} = 1 - P_{FS}^{(B)} \]  

that free flow remains at the bottleneck during the time interval \( T_{ob} \). Freeway
8.3 Freeway Capacity in Free Flow at Bottlenecks

Freeway Capacity in Free Flow at Bottlenecks

Thus, we define freeway capacity in free flow at a bottleneck as follows.\textsuperscript{1}

- Freeway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability \( P_C^{(B)} < 1 \) (8.4) during a given time interval \( T_{ob} \) for observing traffic flow. This means that at this flow rate with the probability \( P_{FS}^{(B)} > 0 \) an F→S transition (breakdown phenomenon) occurs at the bottleneck during the time interval \( T_{ob} \).

Because there can be an infinite number of flow rates downstream of the bottleneck for which the condition (8.4) is satisfied, there can also be an infinite number of freeway capacities in free flow at the bottleneck. Each of the freeway capacities has two attributes:

1. The probability \( P_C^{(B)} \) (8.4) that free flow remains at the bottleneck during a given time interval \( T_{ob} \) for observing traffic flow.
2. The time interval \( T_{ob} \).

When for free flow at the bottleneck rather than the condition (8.4) the condition

\[
P_C^{(B)} = 1
\]  

is satisfied, then the flow rate downstream of the bottleneck in this free flow is lower than any of the freeway capacities. In this case, the probability that an F→S transition (breakdown phenomenon) occurs at the bottleneck during the time interval \( T_{ob} \) is \( P_{FS}^{(B)} = 0 \). This means that freeway capacity is not reached in this free flow at the bottleneck.

**Maximum Freeway Capacities**

As discussed in Sect. 7.3, an F→S transition occurs at a bottleneck during a given time interval \( T_{ob} \) with the probability \( P_{FS}^{(B)} = 1 \) if traffic demand is related to the boundary \( F_{S}^{(B)} \) in the diagram of congested patterns (Fig. 7.1). Thus, there are \textit{an infinite} number of maximum freeway capacities of free flow at the bottleneck that are given by the points on the boundary \( F_{S}^{(B)} \). We denote these capacities by \( q_{\text{max}}^{(\text{free } B)} \):

\[
q_{\text{max}}^{(\text{free } B)} = q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}) \big|_{F_{S}^{(B)}}, \quad (8.6)
\]

where \( q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}) \) (5.35) is the flow rate downstream of the bottleneck. Thus, the maximum freeway capacities of free flow at the bottleneck are

\textsuperscript{1} The definition of freeway capacity in empirical observations of free flow where the flow rate downstream of the bottleneck can be a complicated function of time appears in Sect. 10.3.1.
found from the condition

\[ P^{(B)}_{FS} \big|_{q_{sum}=q_{\text{free B}}} = 1 . \] (8.7)

From the condition (8.7) we see that the maximum freeway capacities of free flow at the bottleneck depend on the flow rates \( q_{on} \), \( q_{in} \), and on the time interval \( T_{ob} \). Accordingly, the boundary \( F_{S}^{(B)} \) in the diagram of congested patterns (Fig. 8.1a), where the F→S transition occurs at the bottleneck with the probability \( P^{(B)}_{FS} = 1 \), also depends on \( T_{ob} \).

Inserting (8.7) into (8.3), we obtain

\[ P^{(B)}_{C} \big|_{q_{sum}=q_{\text{max B}}} = 0 . \] (8.8)

Thus, the probability, that free flow remains at a bottleneck during a given time interval \( T_{ob} \) while observing each of the maximum freeway capacities in free flow, is zero. This means that on average each of the maximum freeway capacities at the bottleneck \( q^{(\text{free B})}_{\text{max}} \) cannot be observed during a longer time interval than the time interval \( T_{ob} \). This is related to an F→S transition that occurs in this free flow during the time interval \( T_{ob} \) with the probability \( P^{(B)}_{FS} = 1 \) (8.7).

We see that to study freeway capacity at a bottleneck, the probability \( P^{(B)}_{FS} \) for an F→S transition at the bottleneck should be found.

### 8.3.2 Probability for Speed Breakdown at Bottlenecks

The maximum freeway capacities of free flow at the bottleneck are determined from the condition (8.7), i.e., through the probability for an F→S transition at the bottleneck. We see that the probability \( P^{(B)}_{FS} \) is a very important characteristic of the freeway capacity in free flow at the bottleneck. For this reason, we consider the dependence of the probability for the F→S transition at the bottleneck \( P^{(B)}_{FS} \) on the flow rate. However, firstly, we give the definition of this probability.

#### Definition of Probability for Breakdown Phenomenon

The probability for an F→S transition at the bottleneck, \( P^{(B)}_{FS} \), is defined as follows. A large number of different realizations, \( N_{FS} \), are performed where a spontaneous F→S transition in an initial free flow at the bottleneck is studied. Each of the realizations should be made at the same flow rates \( q_{on} \) and \( q_{in} \), other initial conditions, and during the same time interval \( T_{ob} \) for observing the spontaneous F→S transition at the bottleneck.
Fig. 8.1. Explanation of freeway capacity. (a) Threshold boundary $F_{th}^{(B)}$ in the diagram of congested patterns. (b) Maximum freeway capacities $q_{max, lim}^{(free B)}$ in free flow at an on-ramp bottleneck as functions of the flow rate to the on-ramp $q_{on}$. The diagram in (a) (without the threshold boundary) is taken from Fig. 7.13. Taken from [218,221,222]
If in \( n_{FS} \) of these \( N_{FS} \) realizations the \( F \rightarrow S \) transition occurs, then

\[
P_{FS}^{(B)} = \frac{n_{FS}}{N_{FS}}. \tag{8.9}
\]

Here the remark in footnote 4 of Sect. 5.2.3 is also valid.\(^2\)

**Dependence of Maximum Capacities of Free Flow on Flow Rate to On-Ramp**

There can be a number of different dependencies of the maximum freeway capacity in free flow at the bottleneck \( q_{\text{max}}^{(\text{free } B)} \) on \( q_{\text{on}} \) (e.g., curves 1, 2, 3 in Fig. 8.1b). In particular, it can be expected that the higher the flow rate to the on-ramp \( q_{\text{on}} \), the lower the maximum freeway capacity in free flow at the bottleneck \( q_{\text{max}}^{(\text{free } B)} \). In this case, the maximum freeway capacity \( q_{\text{max}}^{(\text{free } B)} \) is a decreasing function of \( q_{\text{on}} \) (curve 2 in Fig. 8.1b).

The highest is the maximum freeway capacity of free flow at \( q_{\text{on}} = 0 \). This case is obviously related to the maximum capacity \( q_{\text{max}}^{(\text{free})} \) on a homogeneous (without bottlenecks) road (Fig. 6.2a). However, as explained in Sect. 7.4.1, at a very small flow rate \( q_{\text{on}} \), specifically at a limit case \( q_{\text{on}} \rightarrow 0 \) (but \( q_{\text{on}} \neq 0 \))

\[
q_{\text{max}}^{(\text{free } B)} \bigg|_{q_{\text{on}} \rightarrow 0} = q_{\text{max}}^{(\text{free})}, \quad \lim_{q_{\text{on}} \rightarrow 0} q_{\text{max}}^{(\text{free})} < q_{\text{max}}^{(\text{free })} \quad \text{at } q_{\text{on}} \neq 0. \tag{8.11}
\]

The condition (8.11) is related to the effect of random perturbations that occur on the main road in the vicinity of the on-ramp bottleneck due to single vehicles merging onto the main road from the on-ramp (Sect. 7.4.1). For this reason, in Fig. 8.1b rather than the flow rate \( q_{\text{max}}^{(\text{free })} \) the flow rate \( q_{\text{max}}, \lim \) is shown as the maximum freeway capacity at the bottleneck at the limit case \( q_{\text{on}} \rightarrow 0 \).

A decrease of the maximum freeway capacity \( q_{\text{max}}^{(\text{free } B)} \) at the on-ramp bottleneck when the flow rate \( q_{\text{on}} \) increases can have saturation at a sufficient flow rate to the on-ramp \( q_{\text{on}} \) (Fig. 8.1b, dotted curve 3). In this case, the maximum freeway capacity \( q_{\text{max}}^{(\text{free } B)} \) does not reduce below some saturation value \( q_{\text{max}}, \text{ sat} \) even at a very high \( q_{\text{on}} \).\(^2\)

---

\(^2\) Strictly speaking, the definitions of the boundary \( F_S^{(B)} \) in the diagram of congested patterns (Sect. 7.3) and the definition of the maximum freeway capacities (8.7) are only correct if \( N_{FS} \) in (8.9) is a given finite value. When \( N_{FS} \rightarrow \infty \), the boundary \( F_S^{(B)} \) and the related maximum freeway capacities can be found from the condition

\[
P_{FS}^{(B)} \bigg|_{q_{\text{sum}} = q_{\text{max}}^{(\text{free } B)}} = 1 - \epsilon, \tag{8.10}
\]

where \( \epsilon \) is a given positive value, \( \epsilon \ll 1 \).
8.3.3 Threshold Boundary for Speed Breakdown

For some values \( q_{on} \) and \( q_{in} \), which are related to points \((q_{on}, q_{in})\) below and left of the boundary \( F^{(B)}_S \), an F→S transition nevertheless occurs at the bottleneck during the time interval \( T_{ob} \) with a probability

\[
P^{(B)}_{FS} < 1.
\]  

(8.12)

In this case, the more distant a point \((q_{on}, q_{in})\) from the boundary \( F^{(B)}_S \), the lower the probability \( P^{(B)}_{FS} \).

The region in the flow–flow plane with the coordinates \((q_{on}, q_{in})\), where this probabilistic effect occurs, is restricted by a threshold boundary \( F^{(B)}_{th} \) (Figs. 8.1a and 8.2). The threshold boundary \( F^{(B)}_{th} \) is below and left of the boundary \( F^{(B)}_S \) in the flow–flow plane with the coordinates \((q_{on}, q_{in})\). This threshold boundary is related to an infinite multitude of threshold flow rates

\[
q^{(B)}_{th} = q_{sum}(q_{on}, q_{in}) \big|_{F^{(B)}_{th}},
\]  

(8.13)

where \( q_{sum}(q_{on}, q_{in}) \) (5.35). At the threshold boundary \( F^{(B)}_{th} \) the critical amplitudes of the local perturbation for an F→S transition reach maximum values.

![Fig. 8.2.](image)

**Fig. 8.2.** Numerical simulations of the diagram of congested patterns at an on-ramp bottleneck with the threshold boundary \( F^{(B)}_{th} \). Taken from [331]
The threshold points $q_{th}^{(B)}(q_{on}, q_{in})$ (8.13) on the threshold boundary $F_{th}^{(B)}$ are the minimum flow rates $q_{sum}$ at which an $F \rightarrow S$ transition can still occur during a given time interval $T_{ob}$ at a bottleneck: if at a given flow rate $q_{in}$ the condition

$$q_{sum}(q_{on}, q_{in}) < q_{th}^{(B)}(q_{on} | q_{th}^{(B)}, q_{in})$$  \hspace{1cm} (8.14)

is satisfied, then the $F \rightarrow S$ transition cannot occur in free flow at the bottleneck. This means that below and left of the threshold boundary $F_{th}^{(B)}$, i.e., under the condition (8.14), the probability $P_{FS}^{(B)}$ for the $F \rightarrow S$ transition is given by:

$$P_{FS}^{(B)} | q_{sum}(q_{on}, q_{in}) < q_{th}^{(B)}(q_{on} | q_{th}^{(B)}, q_{in}) = 0.$$  \hspace{1cm} (8.15)

Thus, regardless of the amplitude of a time-limited local perturbation at the bottleneck, an $F \rightarrow S$ transition (speed breakdown) does not occur when the condition (8.14) is satisfied.

The limit point on the threshold boundary $F_{th}^{(B)}$ at $q_{on} = 0$, i.e., when $q_{sum} = q_{in}$, is the threshold point for an $F \rightarrow S$ transition on a homogeneous road (Fig. 8.1a):

$$q_{th}^{(B)}(0, q_{in}) = q_{th}.$$  \hspace{1cm} (8.16)

The threshold flow rate $q_{th}$ for the $F \rightarrow S$ transition on the homogeneous road has been defined in Sect. 5.2.4.

Other threshold points on the threshold boundary $F_{th}^{(B)}$ at $q_{on} > 0$ are related to the flow rate $q_{in} < q_{th}$. To understand these threshold points, we consider a point $(q_{on}, q_{in})$ in the diagram of congested patterns (Fig. 8.1a) that is below and left of the boundary $F_{S}^{(B)}$. However, an $F \rightarrow S$ transition occurs in this free flow with the probability

$$0 < P_{FS}^{(B)} < 1.$$  \hspace{1cm} (8.17)

If now the flow rate $q_{sum}$ decreases (due to a gradual decrease in the flow rate $q_{on}$ and/or the flow rate $q_{in}$) the distance of the point $(q_{on}, q_{in})$ from the critical boundary $F_{S}^{(B)}$ increases and the probability $P_{FS}^{(B)}$ for the $F \rightarrow S$ transition decreases.

**Minimum Freeway Capacities**

There are an infinite number of minimum freeway capacities of free flow at a bottleneck that are given by the points on the boundary $F_{th}^{(B)}$. These minimum capacities are equal to the threshold flow rates $q_{th}^{(B)}$ (8.13). Thus, the minimum freeway capacities in free flow at the bottleneck are found from the condition that free flow remains at the bottleneck with the probability $P^{(B)}_{C}$, which is still lower than 1 during the time interval $T_{ob}$. In contrast, if the flow rate in free flow downstream of the bottleneck $q_{sum}$ is related to a
point \((q_{on}, q_{in})\) below and left of the threshold boundary \(F^{(B)}_{th}\) in the diagram of congested patterns (Fig. 8.1a), then this flow rate is lower than freeway capacity. The latter means that this free flow remains with the probability \(P^{(B)}_{C} = 1\) at the bottleneck during the time interval \(T_{ob}\).

Thus, if the flow rate \(q_{sum}\) downstream of an on-ramp bottleneck is less than freeway capacity, specifically the condition (8.14) is satisfied, then the probability \(P^{(B)}_{FS}\) for speed breakdown at the bottleneck during a given time interval \(T_{ob}\) for observing traffic flow is zero (8.15). If the flow rate \(q_{sum}\) increases and exceeds one of the minimum freeway capacities \(q^{(B)}_{th}\) (8.13), specifically the condition \(P^{(B)}_{FS} > 0\) is satisfied, then the probability \(P^{(B)}_{FS}\) increases continuously with the flow rate \(q_{sum}\) (Fig. 8.3), in accordance with empirical results (Fig. 2.12). The probability \(P^{(B)}_{FS}\) is equal to 1 (8.7) when the flow rate \(q_{sum}\) reaches one of the maximum freeway capacities \(q^{(free\ B)}_{max}\) (Fig. 8.3).

![Fig. 8.3.](image)

**Infinite Number of Freeway Capacities**

Maximum freeway capacities are related to the boundary \(F^{(B)}_{S}\) in the diagram of congested patterns (Fig. 8.1a). Minimum freeway capacities are related to the boundary \(F^{(B)}_{th}\) in this diagram. Different points \((q_{on}, q_{in})\) between
226 8 Freeway Capacity

the boundaries $F_S^{(B)}$ and $F_{th}^{(B)}$ in the diagram of congested patterns are related to the infinity of the flow rates in free flow downstream of a bottleneck $q_{\text{sum}}(q_{on}, q_{in})$ (5.35). For each of these flow rates the probability $P_C^{(B)}$ (8.3) satisfies the condition

$$0 < P_C^{(B)} < 1.$$  \hspace{1cm} (8.18)

This corresponds to (8.4).

Taking into account the definition of freeway capacity of Sect. 8.3.1, we can conclude that there are the infinity of freeway capacities in free flow at a freeway bottleneck. We denote the freeway capacities in free flow at the freeway bottleneck by $q_C^{(B)}$. These freeway capacities are equal to the flow rates $q_{\text{sum}}(q_{on}, q_{in})$ (5.35) downstream of the bottleneck associated with the infinite number of points $(q_{on}, q_{in})$ on and between the boundaries $F_S^{(B)}$ and $F_{th}^{(B)}$ in the diagram of congested patterns at the bottleneck (Fig. 8.1a). This means that at a given value of the flow rate $q_{in}$ the freeway capacity $q_C^{(B)}(q_{on}, q_{in})$ in free flow at the bottleneck satisfies the condition

$$q \left( q_{on} \left| p_{th}^{(B)} \right., q_{in} \right) \leq q_C^{(B)}(q_{on}, q_{in}) \leq q_{\text{max}}^{(free \ B)} \left( q_{on} \left| F_S^{(B)} \right., q_{in} \right).$$  \hspace{1cm} (8.19)

8.3.4 Features of Freeway Capacity at Bottlenecks

From the discussion in the previous Sects. 8.3.1–8.3.3, we can draw the following conclusions about the probabilistic nature of the freeway capacity of free flow at bottlenecks.

(1) Freeway capacity in free flow at a bottleneck is limited by the occurrence of an $F \rightarrow S$ transition at the bottleneck. This transition is realized with the probability $P_{FS}^{(B)}$ during a given time interval $T_{ob}$. For this reason, freeway capacity is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability $P_C^{(B)} = 1 - P_{FS}^{(B)} < 1$ (8.4) during the time interval $T_{ob}$. The probability $P_C^{(B)}$ is the necessary attribute of freeway capacity at the bottleneck.

(2) There are an infinite number of such freeway capacities at a bottleneck. Considering one realization where an $F \rightarrow S$ transition (speed breakdown) has occurred at the bottleneck for a given time interval $T_{ob}$ it is not possible to determine all these freeway capacities. To find the freeway capacities, an analysis of the probability $P_{FS}^{(B)}$ for the $F \rightarrow S$ transition in free flow at the bottleneck as a function of traffic demand should be made. To find this probability, many realizations are necessary.

\footnote{In empirical observations, where the flow rate in free flow cannot usually be time-independent, the role of the time interval $T_{ob}$ for observing traffic flow is played by the averaging time interval $T_{av}$ for the flow rate. For more detail see Sect. 10.3.1.}
(3) For a bottleneck due to the on-ramp freeway capacities in free flow are equal to the flow rates downstream of the bottleneck \( q_{\text{sum}} = q_{\text{on}} + q_{\text{in}} \) that are related to a two-dimensional region in the diagram of congested patterns at the bottleneck (Fig. 8.1a). The boundaries of this two-dimensional region, \( F_S^{(B)} \) and \( F_{\text{th}}^{(B)} \), correspond to the occurrence of an \( F \rightarrow S \) transition during a given time interval \( T_{\text{ob}} \) and to the threshold points for the \( F \rightarrow S \) transition, respectively. In this region of the diagram of congested patterns at the bottleneck (Fig. 8.1a), the probability \( P_{FS}^{(B)}(q_{\text{on}}, q_{\text{in}}) \) for the \( F \rightarrow S \) transition during the time interval \( T_{\text{ob}} \) satisfies the condition

\[
0 < P_{FS}^{(B)}(q_{\text{on}}, q_{\text{in}}) \leq 1 .
\]

(8.20)

In this case, for the probability that free flow remains at the bottleneck during the time interval \( T_{\text{ob}} \) we obtain the following conditions:

\[
0 \leq P_{C}^{(B)}(q_{\text{on}}, q_{\text{in}}) < 1 .
\]

(8.21)

These conditions are related to (8.4) that determines freeway capacities.

(4) There are an infinite number of maximum freeway capacities. The maximum freeway capacities correspond to the condition that the probability for an \( F \rightarrow S \) transition during a given time interval \( T_{\text{ob}} \) is equal to 1:

\[
P_{FS}^{(B)}(q_{\text{on}}, q_{\text{in}}) = 1 .
\]

The maximum freeway capacities are equal to the flow rates \( q_{\text{max}}^{(\text{free B})} \) (8.6). These flow rates are related to points in the diagram of congested patterns that lie on the boundary \( F_S^{(B)} \).

(5) In some range of the time interval \( T_{\text{ob}} \), the higher the time interval \( T_{\text{ob}} \), the lower the flow rate \( q_{\text{on}} \) at which the condition \( P_{FS}^{(B)} = 1 \) is satisfied at a given \( q_{\text{in}} \). When the flow rate \( q_{\text{on}} > 0 \) is a given value, the higher the time interval \( T_{\text{ob}} \), the lower the flow rate \( q_{\text{in}} \) at which the condition \( P_{FS}^{(B)} = 1 \) is satisfied. The maximum freeway capacities \( q_{\text{max}}^{(\text{free B})}(q_{\text{on}}, q_{\text{in}}) \) at a given value \( T_{\text{ob}} \) can be complex functions of the flow rates \( q_{\text{in}} \) and \( q_{\text{on}} \).

(6) There are an infinite number of minimum freeway capacities. The minimum freeway capacities are equal to the flow rates \( q_{\text{th}}^{(B)} \) (8.13). These threshold flow rates are related to points in the diagram of congested patterns that lie on the threshold boundary \( F_{\text{th}}^{(B)} \). The minimum freeway capacities \( q_{\text{th}}^{(B)}(q_{\text{on}}, q_{\text{in}}) \) at a given value \( T_{\text{ob}} \) can be complex functions of the flow rates \( q_{\text{in}} \) and \( q_{\text{on}} \).

(7) The probability for the random occurrence of small amplitude perturbations in free flow at the bottleneck is high. However, this probability sharply decreases when the amplitude of the local perturbations required for an \( F \rightarrow S \) transition should increase. This occurs when the flow rate downstream of the bottleneck \( q_{\text{sum}} \) is related to a point \( (q_{\text{on}}, q_{\text{in}}) \) in the diagram of congested patterns that is below and left of the boundary \( F_S^{(B)} \). For this reason, in the vicinity of the critical points \( q_{\text{max}}^{(\text{free B})} \) at
which \( P_{FS}^{(B)} = 1 \) the probability \( P_{FS}^{(B)} \) should be a sharp function of the flow rate \( q_{sum} \) (Fig. 8.3). This can explain the empirical Fig. 2.12.

**Numerical Simulation of Capacity at Bottlenecks**

The above general conclusions are confirmed by numerical simulations in a microscopic three-phase traffic theory. In all investigated cases at \( q_{on} \to 0 \) the condition (8.11) is valid. Moreover, in the examples shown in Fig. 8.4 the limit value \( q_{max, lim}^{(free \ B)} \) is related to the highest capacity \( q_{max}^{(free \ B)} \) at \( q_{on} \to 0 \) (but \( q_{on} \neq 0 \)).

For various models and model parameters, different dependencies of the maximum capacity at the on-ramp \( q_{max}^{(free \ B)} \) on \( q_{on} \) were found. In particular, both a weak function \( q_{max}^{(free \ B)}(q_{on}) \) with a minimum (Fig. 8.4a) and an initially rapidly decreasing function with saturation at a higher flow rate to the on-ramp (Fig. 8.4b) have been found.

![Fig. 8.4. Numerical simulations of freeway capacity at an on-ramp bottleneck. (a) A spatial continuum model based on the three-phase traffic theory. (b) KKW cellular automata model. Taken from [330,331]](image)

**8.4 Z-Characteristic and Probability for Speed Breakdown**

In this section, we consider a connection between the Z-characteristic for an F→S transition (speed breakdown) (Sect. 5.3.5) and the probability \( P_{FS}^{(B)} \) for the F→S transition. In particular, we find that the Z-characteristic for an F→S transition (Fig. 5.14) can explain the dependence of the probability \( P_{FS}^{(B)} \) on the flow rates \( q_{on} \) and \( q_{in} \) discussed above.

To show this, let us first consider the boundaries \( F_{S}^{(B)} \) and \( F_{th}^{(B)} \) in the diagram of congested patterns (Fig. 8.1a) together with simplified Z-characteristics for an F→S transition (Fig. 5.14b). This is possible in a 3D space with coordinates \( q_{on} \), \( q_{in} \), and the vehicle speed (Fig. 8.5).
In Fig. 8.5, we show two simplified Z-characteristics for the F→S transition taken from Fig. 5.14b for two different flow rates $q_{in}^{(1)} = \text{const}$ (the Z-characteristic labeled $Z_1$) and $q_{in}^{(2)} = \text{const}$ (the Z-characteristic labeled $Z_2$), where $q_{in}^{(2)} > q_{in}^{(1)}$. For clarity, we have also marked the branch for free flow at the bottleneck $v_{\text{free}}^{(B)}(q_{on})$, the critical branch $v_{\text{cr, FS}}^{(B)}(q_{on})$, and the averaged branch for synchronized flow $v_{\text{syn, aver}}^{(B)}(q_{on})$ (Fig. 5.14b) on one of these simplified Z-characteristics in Fig. 8.5 (at the Z-characteristic labeled $Z_2$).

For the explanation of the relation between the Z-characteristics and the boundaries $F_{\text{th}}^{(B)}$ and $F_{\text{FS}}^{(B)}$ in Fig. 8.5 we should recall that the difference

$$v_{\text{free}}^{(B)} - v_{\text{cr, FS}}^{(B)}$$

(8.22)

gives the critical amplitude (5.49)
of the critical local perturbation for speed breakdown in free flow at the bottleneck. In Sect. 5.3.4, we explained that only if the current amplitude of a random local perturbation in free flow at the bottleneck exceeds the critical amplitude (8.22) does speed breakdown occur.

The difference (8.22) is a decreasing function of the flow rate $q_{on}$ at a given flow rate $q_{in}$ (Sect. 5.3.5). At the critical point $q_{on} = q_{on}^{(deterministic, FS)}(q_{in})$ (5.58) the difference (8.22) is zero, i.e., the critical amplitude of the critical perturbation for an $F \rightarrow S$ transition $\Delta v_{cr, FS}^{(B)} = 0$. In this limiting case, even if there were no random perturbations in free flow, a deterministic $F \rightarrow S$ transition would take place at the bottleneck (Sect. 5.3.2). When $\Delta v_{cr, FS}^{(B)} = 0$, then during the relatively small time delay $T_{deterministic, FS}^{(B)}$ (5.64) (Sect. 5.3.7), the deterministic $F \rightarrow S$ transition occurs at the bottleneck with the probability $P_{FS}^{(B)} = 1$. The time delay $T_{deterministic, FS}^{(B)}$ of the deterministic $F \rightarrow S$ transition is comparable with the duration of the $F \rightarrow S$ transition. Accordingly, the maximum freeway capacities of free flow at the bottleneck $q_{\text{max}}^{\text{free}, B}$ given by the condition (8.7) are related to the deterministic $F \rightarrow S$ transition at the bottleneck. Thus, in this limiting case:

$$q_{\text{max}}^{(\text{free}, B)} \approx q_{\text{deterministic, FS}}^{(B)},$$

where $q_{\text{deterministic, FS}}^{(B)}$ (5.59) is the critical flow rate downstream of the bottleneck for the deterministic $F \rightarrow S$ transition.

For simplicity we will consider cases at which values of the mean time delay $T_{FS}^{(B, \text{mean})}$ for an $F \rightarrow S$ transition are considerably greater than the duration of the $F \rightarrow S$ transition. Then at the boundary $F_{S}^{(B)}$ in Figs. 8.1a and 8.5, where the $F \rightarrow S$ transition occurs with the probability $P_{FS}^{(B)} = 1$ (8.7) during a given time interval $T_{ob}$ for observing the $F \rightarrow S$ transition, the difference (8.22) and therefore the critical amplitude of the critical perturbation $\Delta v_{cr, FS}^{(B)}$ (8.23) are appreciably greater than zero. In other words, at the boundary $F_{S}^{(B)}$ in the diagram of congested patterns (Fig. 8.1a) the deterministic $F \rightarrow S$ transition cannot occur. For an $F \rightarrow S$ transition in free flow

\[\Delta v_{cr, FS}^{(B)} = v_{\text{free}}^{(B)} - v_{cr, FS}^{(B)} \quad (8.23)\]

\[q_{\text{max}}^{(\text{free}, B)} = q_{\text{deterministic, FS}}^{(B)} \quad (8.24)\]

\[q_{\text{max}}^{(\text{free}, B)} \approx q_{\text{deterministic, FS}}^{(B)} \quad (8.25)\]

\[q_{\text{max}}^{(\text{free}, B)} \approx q_{\text{deterministic, FS}}^{(B)} \quad (8.25)\]

\[q_{\text{max}}^{(\text{free}, B)} \approx q_{\text{deterministic, FS}}^{(B)} \quad (8.25)\]
at the bottleneck, a random local perturbation must occur with amplitude greater than the critical amplitude (8.23). For this reason, in Fig. 8.5 the boundary $F_{S}^{(B)}$ intersects the $Z$-characteristics at the flow rates to the onramp $q_{on}$ where the difference (8.22) is greater than zero. Therefore, these flow rates $q_{on}$ are lower than the critical values $q_{on} = q_{on,1}^{(determ, FS)}$ (for the $Z$-characteristic $Z_1$) and $q_{on} = q_{on,2}^{(determ, FS)}$ (for the $Z$-characteristic $Z_2$).

Below and left of the boundary $F_{S}^{(B)}$, when at a given flow rate $q_{in}$ the flow rate $q_{on}$ decreases, the shortest distance between the point $(q_{on}, q_{in})$ and the boundary $F_{S}^{(B)}$ in the diagram of congested patterns increases (Fig. 8.1a). This increase is accompanied by the increase in the difference (8.22) in Fig. 8.5.

On the one hand, the amplitude of a current random local perturbation in free flow at the bottleneck should exceed the difference (8.22) (see the condition (5.51)) to cause an $F \rightarrow S$ transition at the bottleneck. On the other hand, the greater the amplitude of a random perturbation, the less probable its occurrence. This is a well-known result of statistical physics.

Thus, the probability for a spontaneous $F \rightarrow S$ transition at the bottleneck, $P_{FS}^{(B)}$ should decrease when at the given flow rate $q_{in}$ the flow rate $q_{on}$ decreases.

If we further decrease the flow rate $q_{on}$, we reach the threshold boundary $F_{th}^{(B)}$ in Fig. 8.5. From this figure we can also see that at the threshold boundary the difference (8.22) and the critical amplitude of the critical local perturbation $\Delta v_{cr, FS}^{(B)}$ (8.23) reach maximum values at the $Z$-characteristics. The flow rate $q_{on}$ is equal to the threshold flow rate: $q_{on} = q_{on,1}^{(th)}$ (for the $Z$-characteristic $Z_1$) and $q_{on} = q_{on,2}^{(th)}$ (for the $Z$-characteristic $Z_2$). The probability $P_{FS}^{(B)}$ is very low (close to zero) at the threshold boundary $F_{th}^{(B)}$.

When the flow rate $q_{on}$ decreases below the threshold flow rate (below and left of the boundary $F_{th}^{(B)}$ in Fig. 8.5), then there are no states for synchronized flow at the $Z$-characteristics. For this reason, in this case there cannot be speed breakdown in free flow at the bottleneck, i.e., the probability for an $F \rightarrow S$ transition $P_{FS}^{(B)} = 0$. This is independent of how high the current amplitude of a random time-limited perturbation is in free flow at the bottleneck.

From Fig. 8.5 and the above explanations we see that for each given flow rate $q_{in}$ there is a correspondence between the $Z$-characteristic for an $F \rightarrow S$ transition, the boundaries $F_{S}^{(B)}$ and $F_{th}^{(B)}$ in the diagram of congested patterns (Fig. 8.1a), and the probability $P_{FS}^{(B)}$ for the $F \rightarrow S$ transition in free flow at the bottleneck. Because there are an infinite number of different flow rates $q_{in}$ associated with the boundaries $F_{S}^{(B)}$ and $F_{th}^{(B)}$, there are an infinite number of different $Z$-characteristics for the $F \rightarrow S$ transition for these different values $q_{in}$ (only two of them, $Z_1$ and $Z_2$, are shown in Fig. 8.5).
8.5 Congested Pattern Capacity at Bottlenecks

In this section, we consider a theory of freeway capacity at a bottleneck where a congested pattern has already occurred. In order to study freeway capacity downstream of the congested bottleneck, one has to consider the outflow from the congested pattern at the bottleneck, \( q_{\text{out}}^{(\text{bottle})} \) (the discharge flow rate), which is measured downstream of the bottleneck, where free flow conditions are reached.

In three-phase traffic theory, the discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \) is not just a characteristic property of the type of bottleneck under consideration. It also depends on the type of congested pattern, which is actually formed upstream of the bottleneck. Thus, in three-phase traffic theory, freeway capacity in free flow downstream of the congested bottleneck depends on the type of congested pattern upstream of the bottleneck, the pattern characteristics, and on the parameters of the bottleneck (see results of numerical simulations in Sect. 18.8.3). We call this freeway capacity congested pattern capacity.

In the case of a bottleneck due to the on-ramp, \( q_{\text{out}}^{(\text{bottle})} \) is expected to vary with \((q_{\text{on}}, q_{\text{in}})\). Obviously, \( q_{\text{out}}^{(\text{bottle})} \) only limits the freeway capacity, if it is smaller than the traffic demand upstream of the on-ramp, \( q_{\text{sum}} = q_{\text{in}} + q_{\text{on}} \), i.e., if
\[
q_{\text{out}}^{(\text{bottle})}(q_{\text{on}}, q_{\text{in}}) < q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}) .
\] (8.26)

Note that in (8.26) in contrast to the discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \), the flow rate \( q_{\text{sum}} \) is related to free flow conditions at the bottleneck, i.e., when no congested pattern exists upstream of the bottleneck. Under the condition (8.26), the congested pattern on the main road upstream of the on-ramp simply expands, while the throughput remains limited by \( q_{\text{out}}^{(\text{bottle})} \). For example, if an GP is formed at the bottleneck, an increase of \( q_{\text{in}} \) does not influence the discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \). Instead, the width of the wide moving jam, which is farthest upstream in the GP, simply grows.

We denote the congested pattern capacity by \( q_{\text{cong}}^{(B)} \). In the case (8.26), \( q_{\text{cong}}^{(B)} \) is equal to \( q_{\text{out}}^{(\text{bottle})} \):
\[
q_{\text{cong}}^{(B)} = q_{\text{out}}^{(\text{bottle})} .
\] (8.27)

It must be noted that \( q_{\text{out}}^{(\text{bottle})} \) can strongly depend on the congested pattern type and congested pattern parameters. These in turn depend on initial conditions and the flow rates \( q_{\text{on}} \) and \( q_{\text{in}} \) (Fig. 7.13). Thus, the congested pattern capacity \( q_{\text{cong}}^{(B)} \) implicitly depends on the flow rates \( q_{\text{on}} \) and \( q_{\text{in}} \).

Capacity Drop

The capacity drop is the difference between freeway capacity in free flow at the bottleneck and in a situation where there is a congested pattern upstream and free flow downstream of the bottleneck.
Assuming that (8.26) is satisfied, we can define the capacity drop as follows:

\[ \delta q = q_{\text{free}}^{(B)} - q_{\text{cong}}^{(B)}, \]  

where the congested pattern capacity \( q_{\text{cong}}^{(B)} \) is given by (8.27); \( q_{\text{max}}^{(B)} \) is the maximum freeway capacity in free flow at \( q_{\text{on}} = 0 \).

Let us consider all kinds of congested patterns that can occur upstream of the bottleneck. Then the minimum value, which \( q_{\text{out}}^{(bottle)} \) can take under the condition (8.26), should be a characteristic quantity for the type of bottleneck under consideration. We denote this quantity by \( q_{\text{min}}^{(bottle)} \). The maximum of \( q_{\text{out}}^{(bottle)} \) (denoted by \( q_{\text{max}}^{(bottle)} \)) is predicted to be the maximum flow rate that can be realized in synchronized flow, \( q_{\text{max}}^{(bottle)} = q_{\text{max}}^{(syn)} \). To explain the latter condition, recall that the downstream front of a congested pattern at the bottleneck due to the on-ramp separates free flow downstream of the front and synchronized flow upstream of the front. This downstream front is fixed at the bottleneck. Thus, within the front the total flow rate across the road (together with the on-ramp lane) does not depend on the coordinates along the road (Sect. 7.4.3). Just downstream of the front, i.e., in free flow, this flow rate is equal to \( q_{\text{out}}^{(bottle)} \). Just upstream of the front it is assumed that synchronized flow occurs both on the main road and in the on-ramp lane. The maximum possible flow rate in synchronized flow is equal to \( q_{\text{max}}^{(syn)} \). Hence, the capacity drop at a bottleneck that is defined by formula (8.28) cannot be smaller than

\[ \delta q_{\text{min}} = q_{\text{max}}^{(free)} - q_{\text{max}}^{(syn)}. \]  

Note that there can also be another definition of the capacity drop:

\[ \delta q = q_{\text{free}}^{(B)} - q_{\text{cong}}^{(B)}, \]  

where the congested pattern capacity \( q_{\text{cong}}^{(B)} \) is given by (8.27) and \( q_{\text{max}}^{(free \ B)} \) is given by (8.6). However, there can be a difficulty in the application of the definition (8.30): there are an infinite number of different maximum freeway capacities in free flow at a bottleneck, \( q_{\text{max}}^{(free \ B)} \) (see the formula (8.6) and its explanation in Sect. 8.3).

### Congested Pattern Capacity by LSP Formation

There can be one exception to the condition (8.26): if an LSP (see Sect. 7.4) occurs both on the main road and in the on-ramp lane upstream of the merging region of the on-ramp.

In this case, the discharge flow rate is equal to traffic demand:

\[ q_{\text{out}}^{(bottle)}(q_{\text{on}}, q_{\text{in}}) = q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}). \]  

The congested pattern capacity related to this LSP should be determined by the maximum discharge flow rate at which the LSP still exists upstream of the on-ramp.
8.6 Main Behavioral Assumptions of Three-Phase Traffic Theory

In this section, a summary of main driver behavioral assumptions in three-phase traffic theory is made. These driver behavioral assumptions are associated with human expectation of local driving conditions.

There are two reasons for this summary. (1) We would like to show a correlation between hypotheses about traffic flow characteristics and driver’s behavior. (2) This consideration should give a link between hypotheses of the three-phase traffic theory and a microscopic three-phase traffic theory where these driver behavioral assumptions are realized (Part III).

The following main behavioral assumptions have been made in the three-phase traffic theory [205, 207–211]:

(i) In synchronized flow, a driver accepts a range of different hypothetical steady state speeds at the same space gap to the preceding vehicle: there is no optimal speed in synchronized flow.

In synchronized flow, the mean space gap between vehicles is low in comparison with the mean space gap in free flow at the same flow rate. In synchronized flow, a driver is able to recognize whether the net distance (space gap) to the preceding vehicle becomes bigger or smaller over time. This is true even if the difference between the vehicle speed and the speed of the preceding vehicle is negligible. The ability of drivers to maintain a time-independent space gap (without taking fluctuations into account) should be valid for a finite range of space gaps. This leads to the fundamental hypothesis of the theory (Sect. 4.3): steady states of synchronized flow (steady states are hypothetical states in which all vehicles move with the same speed and at the same distance to one another) cover a two-dimensional region in the flow–density plane. A given steady speed in synchronized flow can be related to an infinite multitude of steady states with different densities in a limited range. A given density in synchronized flow can be related to an infinite multitude of steady states with different speeds in a limited range. In some density range, steady states of synchronized flow overlap states of free flow in density in the flow–density plane.

(ii) The higher the density and the lower the speed difference between mean speeds in different freeway lanes, the lower the probability that a driver can pass.

At the same density the speed difference between mean speeds in different lanes in free flow is higher than in synchronized flow. Thus, the probability of passing is lower in synchronized flow than in free flow. As a result of the overlap of states of free flow and steady states of synchronized flow in density, the probability of passing and the speed as functions of density have $Z$-shaped forms. The $Z$-characteristics consist of the upper branch for free flow and the lower two-dimensional region
for synchronized flow. The Z-characteristics explain $F \rightarrow S$ transitions and reverse $S \rightarrow F$ transitions (Sect. 5.2.5).

(iii) **To avoid collisions, drivers do not accept lower space gap than a safe space gap.**

The upper boundary of the two-dimensional region of steady states of synchronized flow in the flow–density plane is related to safety conditions (Sect. 4.3.1).

(iv) **A driver tends to adjust the speed to the preceding vehicle within a “synchronization distance.”**

At a sufficiently high distance to the preceding vehicle a driver accelerates. However, if the driver cannot pass the preceding vehicle, then within the synchronization distance the driver tends to adjust its speed to the preceding vehicle, i.e., the driver decelerates if it is faster, and accelerates if it is slower than the preceding vehicle.

The synchronization distance determines the lower boundary of the two-dimensional region of steady states of synchronized flow in the flow–density plane (Sect. 4.3.2). The lower the speed, the lower the synchronization distance.

The synchronization distance reflects a low driver motivation in synchronized flow in comparison with a high driver motivation in free flow. In other words, the synchronization distance and the existence of a 2D-region of steady states of synchronized flow in the flow–density plane are associated with the behavioral assumption that driver psychology in synchronized flow (specifically when a driver almost cannot pass) is qualitatively different from driver psychology in free flow where the driver can pass and move with a higher average speed than other drivers.

(v) **In synchronized flow of a lower density, a driver searches for the opportunity to accelerate and to pass.**

The dynamics of $F \rightarrow S$ and $S \rightarrow F$ transitions can be explained by a competition between a tendency towards free flow due to an over-acceleration and a tendency towards synchronized flow due to the adaptation of the vehicle speed to the speed of the preceding vehicle (Sect. 5.2.6). This speed adaptation occurs within the synchronization distance. For example, let us consider a case when a driver reaches a region of flow with a lower speed downstream and the driver does not see a possibility to pass. Then the driver decelerates within the synchronization distance, i.e., much earlier than required from safety conditions. The speed adaptation effect predominates at higher density and can lead to an $F \rightarrow S$ transition. This speed adaptation also causes the self-maintenance of synchronized flow. The over-acceleration is the driver acceleration in the case when the driver is in synchronized flow. To explain the over-acceleration, let us assume that a driver is within the synchronization distance to the preceding vehicle. If the driver believes
that there is a possibility to pass, the driver accelerates. This can occur even if the preceding vehicle does not accelerate. However, it can turn out that although the driver accelerates, nevertheless the driver cannot pass. In this case, the driver must decelerate to the speed of the preceding vehicle and wait for another possibility to pass. This vehicle over-acceleration with a subsequent vehicle deceleration to the speed of the preceding vehicle can be repeated several times before the driver can pass.

The over-acceleration predominates at lower density of synchronized flow. The over-acceleration can lead to an S→F transition.

(vi) *In free flow of high enough density at a bottleneck, a driver slows down.* At high enough density in free flow drivers slow down in the vicinity of the bottleneck. This leads to a “deterministic perturbation” in free flow, i.e., to a permanent local decrease in the speed on the main road at the bottleneck (Sect. 5.3.1). This is responsible for the empirical result that an F→S transition occurs mostly at bottlenecks. After the F→S transition has occurred drivers continue to decelerate in the vicinity of the bottleneck. This driver behavior can explain the result that the downstream front of synchronized flow in WSPs, LSPs, and GPs is fixed at the bottleneck (Sects. 2.4 and 7.4.3).

(vii) *In high density flow, a driver decelerates stronger than it is required to avoid collisions if the preceding vehicle begins to decelerate unexpectedly (vehicle over-deceleration).* Moving jam emergence in synchronized flow is associated with the well-known vehicle over-deceleration effect: if in high density synchronized flow the preceding vehicle begins to decelerate unexpectedly, then due to the driver time delay the driver decelerates stronger than it is required to avoid collisions (Sect. 6.5.3).

(viii) *At the downstream front of synchronized flow, which separates this synchronized flow and free flow downstream, a driver in synchronized flow does not accelerate before the preceding vehicle has begun to accelerate.* Let us assume that a driver is in synchronized flow at the downstream front of synchronized flow. To start acceleration to free flow, the driver has to wait some time before the preceding vehicle that has just begun to accelerate is at some safe distance from the driver. This time delay \( \tau_{del, syn}^{(a)} \) is responsible for the threshold density in free flow where an F→S transition still can occur: at the density in free flow that is lower than the threshold density the F→S transition cannot occur (Sect. 5.2.7).

(ix) *At the maximum point of free flow a driver accepts a lower mean time gap than the driver time delay \( \tau_{del, syn}^{(a)} \).* This driver behavior can explain conclusions of the theory of F→S transition (Sect. 5.2.4) and of the probabilistic theory of freeway capacity (Sect. 8.3). In particular, there is a range of density (5.13) and there
is a region of flow rates $q_{on}$ and $q_{in}$ between the boundaries $F_{th}^{(B)}$ and $F_{S}^{(B)}$ in the diagram of congested patterns at bottlenecks (Fig. 8.1a) where free flow is metastable with respect to an $F\rightarrow S$ transition.

(x) At the downstream front of a wide moving jam, which separates this jam from either free flow or synchronized flow downstream, a driver within the jam does not accelerate before the preceding vehicle has begun to accelerate.

Let us assume that a driver is within a wide moving jam at the downstream front of the jam that separates the jam and either free flow or synchronized flow downstream. There is a time delay in driver acceleration. This time delay $\tau_{del}^{(a)}$ has the same nature as $\tau_{del, syn}^{(a)}$. However, the time delay $\tau_{del}^{(a)}$ can be higher than $\tau_{del, syn}^{(a)}$. The time delay in acceleration for the driver within the jam is responsible for threshold densities for wide moving jam excitation either in free flow or in synchronized flow: at densities lower than the related threshold density the wide moving jam cannot occur (Sect. 6.3).

(xi) At the maximum point of free flow a driver accepts a lower mean time gap than the driver time delay $\tau_{del}^{(a)}$.

This driver behavior explains the metastability of free flow with respect to moving jam emergence (Sect. 3.2.7).

(xii) Moving in synchronized flow, a driver comes closer to the preceding vehicle than the synchronization distance.

After synchronized flow has already occurred, a driver can come closer to the preceding vehicle over time. There can be several reasons for this effect. For example, the driver would like to prevent the merging of a vehicle from an adjacent freeway lane. The reduction of the space gap in synchronized flow can explain the pinch effect in synchronized flow, i.e., a self-compression of synchronized flow (Sect. 7.6).

### 8.7 Conclusions

(i) For a homogeneous road (without bottlenecks) there are three kinds of freeway capacity depending on which one of the three traffic phases prevails.

(ii) Freeway capacity of free flow at an effectual freeway bottleneck is limited by an $F\rightarrow S$ transition at the bottleneck. The $F\rightarrow S$ transition can occur at the bottleneck due to a random growing local perturbation in the vicinity of the bottleneck. This random feature explains the probabilistic nature of freeway capacity. The probability $P_{FS}^{(B)}$ for the $F\rightarrow S$ transition at the bottleneck can be a complex function of traffic demand and of the time interval $T_{ob}$ for observing traffic flow. There are infinitely many freeway capacities in free flow at the bottleneck. These freeway capacities are defined through the condition that free flow remains at
the bottleneck during a given time interval $T_{ob}$ with the probability $P_{C}^{(B)} = 1 - P_{FS}^{(B)} < 1$.

(iii) The congested pattern capacity as well as a capacity drop, which occur due to congested pattern formation, depend both on the bottleneck characteristics and on the pattern type formed at the bottleneck. The congested pattern capacity is a function of the spatiotemporal pattern parameters, which can depend heavily on traffic demand.
Part II

Empirical Spatiotemporal Congested Traffic Patterns
9 Empirical Congested Patterns at Isolated Bottlenecks

9.1 Introduction

In this and other chapters of this Part II, we will discuss empirical spatiotemporal congested pattern features [208,217,218,221]. Here we consider congested patterns at effectual isolated bottlenecks. These empirical patterns have been briefly discussed in Sect. 2.4. We will find that the pattern classification made in three-phase traffic theory (Chap. 7) is associated with empirical spatiotemporal pattern features.

We will also see that if “effectual” bottlenecks and “isolated effectual” bottlenecks are not identified on a freeway section where traffic measurements are made, then some incorrect conclusions about congested pattern features can very easily be made. One mistake would be to conclude that an observed congested pattern occurs at an isolated bottleneck whereas in reality the observed spatiotemporal pattern features are related to a great degree to the existence of several different adjacent freeway bottlenecks on this freeway section.

In the latter case, on the one hand, we can expect very complex spatiotemporal patterns (Sects. 2.4.8–2.4.10 and Chap. 14). These patterns cannot be classified by the same scheme as those in Chap. 7, where spatiotemporal patterns were classified at effectual isolated bottlenecks only. On the other hand, if complex spatiotemporal patterns occur due to the existence of two or more different adjacent freeway bottlenecks that are close to one another or other freeway peculiarities, then such patterns can exhibit new features, which are qualitatively different from features of congested patterns at isolated bottlenecks. It can also turn out that there are no bottlenecks, which can be considered as isolated ones at a freeway section, where traffic measurements are made. Rather than isolated bottlenecks, only different adjacent freeway bottlenecks that are close enough to one another exist. In this case, which occurs very often, it can be expected that spatiotemporal patterns possess special features associated with peculiarities of the freeway infrastructure under consideration (Chap. 14; see also the discussion of expanded patterns and foreign wide moving jams in Sect. 2.4).
9.2 Effectual Bottlenecks and Effective Locations of Bottlenecks

A consideration of many different days where congested patterns appear shows that the congested patterns occur at approximately the same freeway locations. These locations are in the vicinity of “effectual” freeway bottlenecks. The definition of an effectual bottleneck is as follows. The effectual bottleneck is a freeway bottleneck that exhibits the following two empirical features [212]:

1. A phase transition from free flow to synchronized flow (F→S transition) occurs considerably more frequently in the vicinity of the effectual bottleneck in comparison to other locations on the freeway.

2. After the F→S transition, synchronized flow emerges at the effectual bottleneck. The downstream front of the “synchronized flow” phase is fixed at a freeway location in the vicinity of the effectual bottleneck (an exception is MSP formation, Sects. 2.4.6 and 7.4.2). Within the downstream front of synchronized flow vehicles accelerate from synchronized flow upstream of the front to free flow downstream of the front.

To find effectual bottlenecks on a freeway section, congested patterns on many different days should be analyzed on the freeway section. The methodology of this empirical pattern analysis has been discussed in Sect. 2.4.11.

An effectual bottleneck is usually caused by some inhomogeneities on the freeway. The reason of an inhomogeneity can be on- and off-ramps, a decrease in the number of freeway lanes in the direction of traffic flow, roadworks, and so on. The location of this inhomogeneity can be considered as the location of the effectual bottleneck. In particular, the beginning of the off-ramp lane can be considered as the location of a bottleneck due to the off-ramp. The beginning of the merging region of the on-ramp lane with the main road can be considered as the location of a bottleneck due to the on-ramp.

The freeway location where an F→S transition has occurred on a particular day at an effectual bottleneck is not necessarily the location of the bottleneck: empirical investigations show that the location of the F→S transition can be either upstream or downstream of the location of the effectual bottleneck. It can be assumed that there is a sharp maximum in the spatial dependence of the probability density for an F→S transition in the vicinity of the bottleneck (Fig. 5.12e). An F→S transition occurs with the highest probability at the location of this maximum. As a result of this F→S transition, synchronized flow occurs in the vicinity of the effectual bottleneck.

It must be noted that the location where the downstream front of synchronized flow is fixed can also be different from the location of the effectual bottleneck: the location of the downstream front of synchronized flow can be either upstream or downstream of the location of the effectual bottleneck.
We make the following definition:

- A location on a freeway where the downstream front of synchronized flow is spatially fixed is called “the effective location” of an effectual bottleneck (the effective location of a bottleneck, for short).

The effective location of the bottleneck is in the vicinity of the effectual bottleneck. However, as mentioned above, the effective location of the bottleneck can be downstream and upstream of the location of the bottleneck. Furthermore, the effective location of the bottleneck does not necessarily coincide with the location of an F→S transition at the bottleneck.

We will also see that after a congested pattern has already been formed at a bottleneck, the effective location of the bottleneck can randomly change over time. It can be assumed that the effective location of the bottleneck has a probabilistic nature. From a study of effective locations of a bottleneck, made on a number of different days, the probability density (the probability per km) of locations for the downstream synchronized flow front (i.e., for effective locations of the bottleneck) as a function of road location can be found. Thus, we can find the probability for an event that the downstream front of synchronized flow in a congested pattern is spatially fixed in the vicinity of the effectual bottleneck in a given range of road locations.

In particular, for effectual bottlenecks due to on-ramps effective bottleneck locations have empirically been studied. We can conclude that the probability density for the effective location of a bottleneck due to the on-ramp has a sharp maximum. This maximum is at some distinct location in the vicinity of the bottleneck. In other words, we can expect that with a high probability the effective location of the bottleneck due to the on-ramp does not change over time. Moreover, the location of an F→S transition usually coincides with the effective location of the bottleneck (to within the accuracy of measurements, which is determined by distances between detector locations).

In contrast, effective locations of an effectual bottleneck due to an off-ramp can often change on different days. These effective locations can also change over time on a particular day. These effective locations of the bottleneck can also be different from the location of an F→S transition, which causes pattern formation at the bottleneck. For the same bottleneck the effective location of the bottleneck can depend on the type of congested pattern that is formed at the bottleneck. However, we have already mentioned that at the same traffic demand there can be several types of congested patterns that can occur at a bottleneck with different probabilities (Sect. 2.4.10). This is one of the reasons for the probabilistic nature of the effective location of the bottleneck. Even for the same type of congested pattern the effective location of the bottleneck can randomly change over time. Empirical examples of this complex behavior will be considered in Sects. 9.3.1 and 9.4.3.
9.2.1 Effectual Bottlenecks on Freeway A5-South

To explain effectual freeway bottlenecks, let us consider two typical congested patterns on a section of the freeway A5 on two different days (Figs. 9.1, 9.2, and 9.3). In these examples, there are moving jams that propagate through freeway bottlenecks and through all other traffic states of traffic while maintaining the velocity of the downstream jam fronts. These moving jams therefore belong to the “wide moving jam” phase (these moving jams are labeled “moving jam A” and “moving jam B” in Figs. 9.1 and 9.2 and “moving jam” in Fig. 9.3).

![Diagram of effectual bottlenecks on A5-South](image)

**Fig. 9.1.** Explanation of effectual bottlenecks. An overview of congested traffic patterns on a section of the freeway A5-South (Fig. 2.1) on June 18, 1997. (a, b) Dependencies of average vehicle speed (across all lanes) (a) and total flow rate across the freeway in time and space (b). Taken from [218]
9.2 Effectual Bottlenecks and Effective Locations of Bottlenecks

A5-South, June 18, 1997

![Graph of free flow (white), synchronized flow (gray), and moving jams (black) in time and space for the pattern shown in Fig. 9.1](image)

**Fig. 9.2.** Graph of free flow (white), synchronized flow (gray), and moving jams (black) in time and space for the pattern shown in Fig. 9.1

Besides wide moving jams, in the vehicle speed distributions (Figs. 9.1a and 9.3a) it is also possible to distinguish synchronized flow. The downstream front of synchronized flow, i.e., the boundary, which separates synchronized flow upstream and free flow downstream, is fixed at some freeway location. Corresponding to the above definition, this location is the effective location of a bottleneck. In Figs. 9.1 and 9.3, there are three effective locations of bottlenecks where three different downstream synchronized flow fronts are localized. These effective locations of bottlenecks are in the vicinity of the same bottlenecks for both days (Figs. 9.1, 9.2, and 9.3).

The locations of these bottlenecks are labeled \( B_1, B_2, \) and \( B_3 \) in Figs. 9.1, 9.2, and 9.3. A study of congested pattern formation on this section of the freeway A5-South made during 1995–2003 showed that an \( F \rightarrow S \) transition occurred most in the vicinity of these bottlenecks. Also in the most cases, after a congested pattern has occurred, the downstream front of this congested pattern is localized in the vicinity of one of these bottlenecks. Corresponding to the definition of the effectual bottleneck, these bottlenecks are effectual bottlenecks on this section of the freeway A5-South (Fig. 2.1).

The effectual bottleneck \( B_1 \) is related to the off-ramp D23-off \((x \approx 23.4 \text{ km})\). The effectual bottleneck \( B_2 \) is related to the on-ramp D15-on about 100 m upstream of D16 \((x \approx 17.1 \text{ km})\). The effectual bottleneck \( B_3 \) is related to the on-ramps D6-on and D5-on about 100 m upstream of D6 \((x \approx 6.4 \text{ km})\).
Fig. 9.3. Explanation of effectual bottlenecks. An overview of congested traffic patterns on a section of the freeway A5-South on March 17, 1997. (a, b) Dependencies of average vehicle speed (across all lanes) (a) and total flow rate across the freeway in time and space (b). (c) Graph of free flow (white), synchronized flow (gray), and moving jams (black) in time and space. Taken from [218]
9.2.2 Effectual Bottlenecks on Freeway A5-North

Traffic observations on a section of the freeway A5-North have shown that there are at least three effectual bottlenecks there (Fig. 9.4). The first one, labeled \( B_{\text{North1}} \), is related to the off-ramp D25-off \((x \approx 22.4 \text{ km})\) (Fig. 2.2). The second effectual bottleneck, labeled \( B_{\text{North2}} \), is related to the on-ramp D15-on about 100 m upstream of D16 \((x \approx 13 \text{ km})\). The third effectual bottleneck, labeled \( B_{\text{North3}} \), is related to the on-ramp upstream of D6 \((x \approx 4.4 \text{ km})\). The features of \( F \rightarrow S \) transitions, which initially occur at the off-ramp D25-off and the on-ramps (D16 and D6), are qualitatively similar to those observed at the effectual bottlenecks at the off-ramp D23-off \((B_1)\) and the on-ramps at D16 \((B_2)\) and D6 \((B_3)\) in the section of the freeway A5-South (Fig. 2.1), respectively.

An example of a typical congested pattern on a section of the freeway A5-North is shown in Fig. 9.4. This is the same pattern that has been shown in Fig. 2.23 and briefly explained in Sect. 2.4.8. In contrast to congested patterns on a section of the freeway A5-South shown in Figs. 9.1, 9.2, and 9.3, from Fig. 9.4a one might have a first impression that the latter congested pattern would be related to different congested states rather than to a coexistence and to an interaction of two traffic phases in congested traffic, synchronized flow and wide moving jams. It seems that there is no possibility to distinguish wide moving jams in Fig. 9.4a. However, this first impression turns out to be incorrect if synchronized flow regions and moving jams in Fig. 9.4c are examined and a more detailed analysis is made (Sects. 9.4.2 and 12.5).

It should be noted that there are also freeway inhomogeneities that do not act as an effectual bottleneck, i.e., where an \( F \rightarrow S \) transition does not occur. For example, while the off-ramp at D23-off is often an effectual bottleneck on the section of the freeway A5-South (Fig. 2.1), the off-ramps at D5-off and D13-off on this section do not act as an effectual bottleneck. The latter follows from a study of congested patterns on this freeway section during 1996–2003: congested patterns have never occurred at the mentioned potential bottlenecks. A potential bottleneck can act as an effectual bottleneck if some high enough flow rates (high enough traffic demand) are realized in the vicinity of the bottleneck.

9.2.3 Isolated Effectual Bottleneck

In this chapter, we restrict further consideration to features of congested patterns occurring at such effectual bottlenecks, which are located far enough away from other adjacent bottlenecks. Effects of other adjacent bottlenecks and/or any other inhomogeneities on the freeway away from an effectual bottleneck should not have a qualitative influence on features of congested patterns at the effectual bottleneck. Such an effectual bottleneck is called an *isolated* effectual bottleneck (an isolated bottleneck).
Fig. 9.4. Explanation of effectual bottlenecks. An overview of congested traffic patterns on a section of the freeway A5-North in Fig. 2.2 on March 23, 2001. (a, b) Dependencies of average vehicle speed (across all lanes) (a) and total flow rate across the freeway in time and space (b). (c) Graph of free flow (white), synchronized flow (gray), and moving jams (black) in time and space.
Let us consider an example of an isolated bottleneck. Inside the intersection I1 on the freeway A5-South there are several on- and/or off-ramps. However, an F→S transition is observed on all days at the same location in the vicinity of the effectual bottleneck at the detectors D6 (Fig. 2.1). This effectual bottleneck is labeled \( B_3 \) in Figs. 9.1 and 9.3. Other possible effectual bottlenecks are located on the road far enough from D6.

However, in this case, there are two on-ramps, D5-on and D6-on, which are very close to one another (Fig. 2.1). Because the distance between the on-ramps D5-on and D6-on is appreciably shorter than the length of each of the on-ramps [218], they can also be considered as one effectual on-ramp on this section with the effective flow rate \( q_{\text{eff-on}} \):

\[
q_{\text{eff-on}} = q_{D6-on} + q_{D5-on} - q_{D6-off} .
\]  

The possibility to consider these two on-ramps at D6-on and D5-on as the effectual on-ramp at D6 is related to the empirical fact that if a congested pattern has occurred in the intersection I3 on the section of the freeway A5-South (Fig. 2.1) this pattern has always occurred only at the same location – at the detectors D6. This means that there is only one effectual bottleneck in this case, in the vicinity of the detectors D6, rather than two different effectual bottlenecks due to the two related on-ramps at D6-on and D5-on.

It must be noted that a consideration of an effectual bottleneck as an isolated one is always a simplification of the reality. Indeed, on real freeways there are many adjacent effectual bottlenecks. Therefore, a congested pattern that has occurred at one of the effectual bottlenecks can reach the next upstream effectual bottleneck if the upstream boundary of this pattern widens upstream over time. In this case, this upstream effectual bottleneck can have a great influence on the future development of the pattern. Empirical investigations have shown that this influence concerns significantly a region in the congested pattern in the vicinity of this upstream bottleneck and upstream of the bottleneck (Chap. 14). Thus, only the region in the congested pattern that is far from the upstream effectual bottleneck can be considered as the pattern at the isolated bottleneck.\(^1\) Nevertheless, the concept of the isolated bottleneck seems to be an important one to study elementary congested pattern features and to make the pattern classification. Congested patterns whose features are essentially related to the existence of two or more adjacent freeway bottlenecks have briefly been considered in Sect. 2.4.8. A more detailed discussion appears in Chap. 14.

\(^{1}\) Strictly speaking, an effectual bottleneck can only be considered as an isolated one during a finite time interval while the upstream boundary of a congested pattern does not reach the upstream bottleneck. An exception can be an LSP. In this case, the upstream front of synchronized flow of the LSP is at a finite distance upstream of the bottleneck.
9.3 Empirical Synchronized Flow Patterns

In Sect. 5.3.1, we have already mentioned that an effectual freeway bottleneck introduces a permanent disturbance in an initial free flow. This occurs if an initial vehicle density on the main road upstream of the bottleneck is high enough. For example, if a bottleneck due to an on-ramp exists, then vehicles merging from the on-ramp onto the main road increase the vehicle density in the vicinity of the on-ramp merging region. This local increase in vehicle density should lead to the related decrease in vehicle speed. When this local permanent decrease in speed (a permanent local disturbance) is high enough, then it can be expected that free flow conditions cannot be maintained at this location any more: an $F \rightarrow S$ transition occurs at the on-ramp.

The other example is a bottleneck due to an off-ramp that exists on a one-way multilane road. Let us assume that only vehicles that move in the right freeway lane can leave to the off-ramp. Thus, some of the vehicles must change to the right lane upstream of the merging region of the off-ramp to have a possibility to leave to the off-ramp. This lane changing occurs only in a local region of the freeway upstream of the off-ramp. Naturally, lane changing causes a permanent local increase (a permanent local disturbance) in vehicle density in the right lane upstream of the off-ramp. This density increase leads to the related decrease in vehicle speed in the right lane. Nevertheless, if the share of vehicles, which must leave to the off-ramp, is small enough a free flow condition can be maintained on the road. If, however, the share of vehicles leaving to the off-ramp increases, the permanent local disturbance in speed in the right lane increases in amplitude. Beginning at some flow rate to the off-ramp the initial free flow condition cannot exist any more: an $F \rightarrow S$ transition occurs upstream of the off-ramp. These qualitative explanations are confirmed by numerous empirical investigations, some of which will be considered in Chap. 10.

Empirical investigations of congested patterns [208, 218] show that in agreement with conclusions of three-phase traffic theory (Sect. 7.4) an $F \rightarrow S$ transition can lead to a synchronized flow pattern (SP) upstream of a bottleneck. An SP consists of synchronized flow upstream of the bottleneck only, i.e., no wide moving jams emerge in the SP. Depending on the flow rates in the vicinity of the bottleneck and bottleneck characteristics, three different types of SPs are possible [218]: a widening SP (WSP), a localized SP (LSP), and a moving SP (MSP). In this section, empirical examples of these SPs will be considered.

9.3.1 Widening Synchronized Flow Pattern

Spatiotemporal Structure of WSP

An empirical example of an WSP is shown in Figs. 9.5 and 9.6. The WSP occurs at the bottleneck due to an off-ramp D25-off on a section of the freeway
A5-North (the bottleneck $B_{\text{North}}$, Sect. 9.2.2). It can be seen that downstream of the off-ramp (D25, Fig. 9.5b) free flow occurs. An $F \rightarrow S$ transition occurs at the detectors D20 upstream of the bottleneck. The $F \rightarrow S$ transition leads to WSP formation: the upstream front of the WSP propagates continuously upstream. This can be seen in Fig. 9.6: synchronized flow occurs at D17 at 16:42, i.e., 3 min later than at the downstream detector D18 (the times of $F \rightarrow S$ transitions at D18–D16 are marked by up-arrows in Fig. 9.6a). Correspondingly, synchronized flow occurs at D16 at 16:45, i.e., also 3 min later than at the downstream detectors D17.

![Fig. 9.5. Widening synchronized flow pattern (WSP). (a) Average speed and flow rate in the WSP as functions of time and location. (b) Time dependencies of average speed (left) and flow rate (right) downstream of the bottleneck and in the off-ramp lane D25-off. Layout of the freeway infrastructure and local measurements on a section of the freeway A5-North is shown in Fig. 2.2. Taken from [221]]
Fig. 9.6. Widening synchronized flow pattern (WSP). Time dependencies of average speed (a) and flow rate (b) at different detectors upstream of the bottleneck due to the off-ramp D25-off. Layout of the freeway infrastructure and local measurements on a section of the freeway A5-North is shown in Fig. 2.2. Taken from [221]
It can be seen in Figs. 9.5b and 9.6 that vehicle speeds slowly decrease within the WSP in the upstream direction whereas the flow rate does not change considerably within the WSP. This is a peculiarity of synchronized flow, which we will consider further. It can be seen that in this case synchronized flow can exist upstream of the bottleneck on a long stretch of the freeway (about 4 km between D21 and D18) during a long time (more than 60 min) without wide moving jam emergence in that synchronized flow.

We can also see that the flow rates in different freeway lanes are appreciably different in free flow (D25, Fig. 9.5). In synchronized flow, the flow rates in different freeway lanes can become close to one another (D17, Fig. 9.6). This flow rate synchronization effect is well-known from other observations (e.g., references in the review by Banks [37]). However, it should be noted that in many other empirical examples of synchronized flow the flow rate synchronization effect is very weak (see next Sect. 9.3.2).

While the vehicle speed within the WSP decreases further in the upstream direction, some moving jams emerge in this synchronized flow of lower vehicle speed (D17 and D16, Fig. 9.6). However, D16 is already related to another upstream bottleneck due to the on-ramp D15-on (Sect. 9.2.2).

This means that the WSP in Fig. 9.5 can be considered as a congested pattern at the isolated bottleneck due to the off-ramp D25-off only up to the detectors D17. Synchronized flow in the WSP, which has emerged upstream of the bottleneck at the off-ramp D25-off, propagates upstream on the main road and reaches the on-ramp location. Due to upstream propagation of synchronized flow a congested pattern is formed on the main road where synchronized flow affects two effectual bottlenecks. Such a congested pattern is called an expanded congested pattern (EP) (Sects. 2.4.8 and 14.2). However, in the part of the EP downstream of the detectors D16, i.e., in the pattern related to the detectors D17–D20, no wide moving jams occur. Thus, the WSP upstream of the off-ramp at D25-off and downstream of the on-ramp at D16 is only a part of this complex EP.

This conclusion concerns each WSP that is forming on real freeways. Indeed, the width of WSP $L_{\text{syn}}$ (Fig. 7.5) increases without limit in the upstream direction. Therefore, on real freeways the upstream front of an WSP over time always reaches another upstream freeway bottleneck. As a result, this bottleneck can change features of the WSP considerably, in particular rather than WSP remaining, another congested pattern can be formed.

Effective Location of Bottleneck

The downstream front of synchronized flow in the WSP is localized at a freeway location between the detectors D20 and D21. The location of the downstream front of the WSP is the effective location of the bottleneck. This means that in the case under consideration the effective location of the bottleneck $B_{\text{North}}$ due to the off-ramp D25-off is at the location $x \approx 18.5$ km. This effective location of the bottleneck is about 3.5 km upstream of
the location of the bottleneck (we assume that the location of the bottleneck is at the beginning of the off-ramp lane). In other words, in this case, free flow conditions are at the location of the bottleneck due to the off-ramp. This behavior will be discussed in Sect. 10.5 where the physics of $F \rightarrow S$ transitions at effectual bottlenecks due to off-ramps will be considered.

The effective location of the bottleneck between the detectors D20 and D21 remains during the most time of WSP existence. However, during a relatively short time interval the effective location of the bottleneck moves downstream to the location of the detectors D23 (Fig. 9.7). The beginning and the end of this time interval are marked by the solid and dashed up-arrows in Fig. 9.7, respectively. In other words, during this short time interval the effective location of the bottleneck is about only 1 km upstream of the location of the bottleneck. Thus, in this case, the distance between the beginning of the off-ramp lane (the location of the bottleneck) and the effective location of the bottleneck decreases. This effect is related to an $F \rightarrow S$ transition that occurs at the detectors D23 at about 17:34. As a result, synchronized flow occurs
at D23 and D22 and later at D21. This synchronized flow exists, however, only several minutes. Respectively, in this particular case the change in the effective location of the bottleneck occurs only during a short time interval. Another empirical example, where this time interval can be a relatively long one, will be considered in Sect. 9.4.3.

9.3.2 Localized Synchronized Flow Pattern

An example of an LSP with a strong oscillating upstream front is shown in Figs. 9.8 and 9.9. The LSP occurs upstream of the bottleneck due to the off-ramp \((x = 23.4\, \text{km}, \text{D23-off in Fig. 2.1})\) It can be seen that downstream of the bottleneck free flow occurs (Fig. 9.8, \(x = 24\, \text{km}, \text{D24}\)). Free flow occurs also upstream of the bottleneck at the detectors D17 (Fig. 9.9, \(x = 17.9\, \text{km}\)). However, upstream of the bottleneck at the detectors D21 synchronized flow is realized. As usual for synchronized flows, in synchronized flow the vehicle speed is lower than the speed at the maximum (limit) point of free flow and the flow rate is relatively high (Fig. 9.9, D21). The flow rate is relatively high during the whole time interval 9:00–10:00 (Fig. 9.8b). No wide moving jams are formed in synchronized flow between the detectors D24 and D17 (Figs. 9.8 and 9.9). Thus, the congested pattern is indeed an LSP.

From the dependence of speed in space and time (Fig. 9.8a) it can be seen that the downstream front of synchronized flow in the LSP where vehicles accelerate from synchronized flow at D21 to free flow at D23 and D24 is fixed at one location (between the detectors D22 and D21).

In contrast to this behavior of the downstream front of synchronized flow, the upstream front of synchronized flow in the LSP exhibits large amplitude oscillations over time. In other words, the width of the LSP \(L_{\text{syn}}\) is a complicated function of time.

To see this conclusion clearly, first note that synchronized flow occurs at the detectors D21 during the whole time interval from 9:00 through 10:00. The maximum width of the LSP \(L_{\text{syn}} \approx 4\, \text{km}\) is realized only within the short time intervals 9:25–9:35 and 9:47–9:54 when synchronized flow is measured at the detectors D18. The minimum width of the LSP \(L_{\text{syn}}\) is about 1.5 km when free flow conditions are measured at the detectors D20. Thus, the amplitude of the oscillation of the LSP width is approximately 2.5 km, i.e., this amplitude is more than 60% of the maximum LSP width.

It can also be seen that the oscillation of the LSP width has a very complicated irregular character. Furthermore, in some time intervals regions of free flow appear within synchronized flow in the vicinity of the initial upstream front of the LSP (Figs. 9.8a and 9.9). This phenomenon of the appearance of alternations of free and synchronized flow in congested patterns will be considered in Sect. 13.2.2.

As mentioned above, flow rates in different freeway lanes can be synchronized in synchronized flow (D17, Fig. 9.6). However, in the example of the
Fig. 9.8. Localized synchronized flow pattern (LSP) upstream of the bottleneck $B_1$ due to the off-ramp D23-off on the freeway A5-South (Fig. 2.1). (a, b) Speed averaged across the freeway (a) and total flow rate (b) in time and space (the detectors D24–D17). (c, d) Time dependencies of average speed (c) and flow rate (d) in different freeway lanes at the detectors D24–D22. Taken from [218]

LSP this flow rate synchronization effect is very weak. This can be seen at D20–D18 in Fig. 9.9b.

9.3.3 Moving Synchronized Flow Pattern

An empirical example of an MSP that occurs at a bottleneck due to an off-ramp is shown in Fig. 9.10. An overview that covers a congested pattern on the entire section of the freeway A5-South is shown in Fig. 2.16.
Fig. 9.9. Localized synchronized flow pattern (LSP) upstream of bottleneck $B_1$ due to the off-ramp D23-off on the freeway A5-South (Fig. 2.1). Time dependencies of average speed (a) and flow rate (b) in different freeway lanes at the detectors D21–D17. Taken from [218]
Fig. 9.10. Moving synchronized flow pattern (MSP) upstream of bottleneck $B_1$ due to the off-ramp D23-off on the freeway A5-South (Fig. 2.1). (a) Average speed (left) and total flow rate across the freeway (right) as functions of space and time. (b) Time dependencies of average speed (left) and flow rate (right) in different freeway lanes at different detectors. Taken from [218]
The condition of MSP occurrence is more easily satisfied at a bottleneck due to the off-ramp. In this case, a disturbance caused by the bottleneck is related to vehicles leaving the main road to the off-ramp. For this reason, if a percentage of vehicles leaving to the off-ramp is high enough an F→S transition occurs upstream of the off-ramp (for more detail see Sect. 10.5). It can be seen in Fig. 9.10b that downstream of the off-ramp D23-off (at D24, Fig. 9.10b) free flow occurs. Upstream of the off-ramp a local region of synchronized flow appears (D21, marked by the down-arrow “MSP” in Fig. 9.10b). This region departs from the off-ramp, and free flow returns upstream of the off-ramp (D21, D20). Thus, an MSP appears. The MSP propagates upstream as a localized structure (D21, D20, and D17).

The flow rate does not change considerably in synchronized flow in comparison with the flow rate in free flow that appears before MSP occurrence and after MSP propagation (D21, D20, and D17, Fig. 9.10b). This conclusion can also be drawn from empirical points in the flow–density plane for free flow (black squares in Fig. 9.11) and for synchronized flow within the MSP (circles).

![Graph](image)

Fig. 9.11. Moving synchronized flow pattern (MSP) in the flow–density plane at the location of D20. Black squares: free flow; circles: synchronized flow. Taken from [218]

9.4 Empirical General Patterns

In three-phase traffic theory, there are general patterns (GP) of type (1) and type (2) (Sect. 7.5.1).

9.4.1 Empirical General Pattern of Type (1)

In Fig. 9.12 an overview of the speed and flow rate in an GP of type (1) is shown. The downstream front of the GP is fixed at a bottleneck (the bottleneck labeled $B_3$). The region of synchronized flow is widening upstream (this
widening of synchronized flow is marked by up-arrows S in Fig. 9.13). The widening of synchronized flow can be considered as a wave of induced F→S transitions propagating upstream over time. At the location of D5 (Fig. 9.13) a pinch region can be seen. Note that this GP has already briefly been considered in Sect. 2.4.7 (Figs. 2.20 and 2.21).

![Fig. 9.12. Overview of an GP at the effectual on-ramp at D6 (bottleneck B3, see Sect. 9.2) on the freeway A5-South (Fig. 2.1). Average speed (left) and total flow rate across the freeway (right) as functions of time and distance. Taken from [218]](image)

At the time marked by the up-arrow P in Fig. 9.13a at D5 the vehicle speed sharply decreases. There is a decrease in flow rate. However, the latter is much lower than the decrease in average speed. This means that the density increases. This is an example of the self-compression of synchronized flow. This self-compression is called the pinch effect in synchronized flow. As a result of the pinch effect, a pinch region is formed in synchronized flow: the speed is low and the density is high in the pinch region. In the pinch region of synchronized flow, narrow moving jams emerge spontaneously (D4, left in Fig. 9.13). These narrow moving jams grow in their amplitude when they propagate upstream (D3). Some of these narrow moving jams transform into wide moving jams (D2). The other narrow moving jams disappear.

The upstream front of synchronized flow is related to a location where a narrow moving jam has just transformed into a wide moving jam. This location, however, oscillates over time (in the vicinity of D3, D2). Because some of the initial narrow moving jams disappear, the mean time between wide moving jams (D1) is considerably higher than an initial mean time between narrow moving jams (D4).

The downstream front of synchronized flow in the GP of type (1) at the on-ramp bottleneck is localized at the detectors D6 during the whole time interval after the GP has been formed (Fig. 9.13). Thus, in this case, the effective location of the bottleneck is at D6. This effective location does not change over time (to within the accuracy of the distance between detectors). An F→S transition also occurs at D6. This means that in this case the
Fig. 9.13. GP at the effectual on-ramp at D6 (the bottleneck \(B_3\), see Sect. 9.2) on the freeway A5-South (Fig. 2.1). Average speed (a) and flow rate (b) shown in each freeway lane at different detectors. Taken from [218]
effective location of the bottleneck coincides with the location of the F→S transition.

9.4.2 Empirical General Pattern of Type (2)

In Fig. 9.14, the formation of an GP of type (2) is shown. Note that this GP has briefly been discussed in Sect. 2.4.7 (Fig. 2.19). Also in this case first an F→S transition occurs. However, in contrast to the case shown in Fig. 9.13, this speed breakdown leads first to WSP formation (this WSP is labeled “WSP” at the detectors D22–D17, Fig. 9.14). The downstream front of the WSP is further fixed at the bottleneck (an overview of this WSP is shown in Fig. 2.17a). Synchronized flow is widening upstream. This upstream widening is marked by up-arrow S in Fig. 9.14: synchronized flow occurs at D22 at about 12:00, i.e., about 4 min later than at the downstream detector D23; synchronized flow occurs at D21 at about 12:03, i.e., about 3 min later than at the downstream detectors D22. Correspondingly, synchronized flow occurs at D20 at about 12:07, i.e., about 4 min later than at the downstream detectors D21, and so on (up-arrows S at D20–D17). Synchronized flow reaches the upstream bottleneck $B_{\text{North2}}$ due to the on-ramp (D16) at $t \approx 12:24$, i.e., about 30 minutes later than the initial F→S transition has occurred at the location of the detectors D23.

Over time the average speed in synchronized flow at the detectors D21–D20 decreases and the density increases: the pinch region is formed in synchronized flow. Narrow moving jams emerge in the pinch region of synchronized flow of the initial WSP. Some of these narrow moving jams grow and transform into wide moving jams (wide moving jams are marked by the down-arrow 1–4 in Fig. 9.14 at the detectors D19–D16): the initial WSP transforms into an GP at the bottleneck due to the off-ramp (this GP is labeled “GP” at the detectors D22–D17).

In contrast to the GP shown in Fig. 9.13, in the GP at the off-ramp bottleneck the upstream front of synchronized flow in the GP is determined by the widening of synchronized flow of the former WSP (up-arrows S in Fig. 9.14) rather than wide moving jam emergence. To see this difference between GPs of type (1) and type (2) clearly, one should compare the propagation of the upstream front of synchronized flow in the GP of type (2) marked by up-arrow $S$ and the propagation of the farthest upstream wide moving jam marked by the down-arrow 1 in Fig. 9.14.

As we have mentioned, in the GP of type (2) the pinch region also occurs where narrow moving jams emerge (D22–D20, Fig. 9.14). These narrow moving jams propagate upstream and grow in their amplitude (the speed decreases and the density increases within these jams). However, the
frequency of narrow moving jam emergence is appreciably lower than in the GP of type (1) shown Fig. 9.13. As a result, almost every growing narrow moving jam transforms into a wide moving jam (the jams marked by down-arrows in Fig. 9.14). These wide moving jams propagate within the initial synchronized flow that has been formed in the process of GP formation.
9.4.3 Dependence of Effective Location of Bottleneck on Time

The downstream front of synchronized flow in the initial WSP and the further GP of type (2) discussed in Sect. 9.4.2 is localized at a freeway location in the vicinity of the detectors D23. The location of the downstream front of synchronized flow in these patterns is the effective location of the bottleneck. This means that in the case under consideration the effective location of the bottleneck $B_{\text{North} \ 1}$ due to the off-ramp D25-off is at the location $x \approx 21$ km. This effective location of the bottleneck is about 1 km upstream of the location of the bottleneck (the beginning of the off-ramp lane). Note that in the case of the WSP considered in Sect. 9.3.1 this effective location of the bottleneck can be observed only during a very short time (between up-arrows in Fig. 9.7). In contrast, in the case under consideration this effective location of the bottleneck is realized after the $F \rightarrow S$ transition has occurred at $t \approx 11:56$ up to $t \approx 15:40$: this effective location of the bottleneck remains during about 3.7 hours.

This result can be seen in Figs. 9.14 and 9.15. However, after $t = 15:40$ the downstream front of synchronized flow is at the location of the detectors D20. This means that the effective location of the bottleneck moves upstream to the location of the detectors D20 (dashed up-arrow at $t = 15:40$ in Fig. 9.15a; D23). This new effective location of the bottleneck is at the location $x \approx 18$ km. This effective location of the bottleneck is about 4 km upstream of the location of the bottleneck (the beginning of the off-ramp lane). The change of the effective location of the bottleneck is accompanied by pattern transformation that will be considered in Sect. 13.3.

The new effectual location of the bottleneck at D20 remains up to $t \approx 16:36$, i.e., during about one hour. Later, the downstream front of synchronized flow is at the location of the detectors D23, i.e., downstream of D20. Thus, the effective location of the bottleneck returns to the location of the detectors D23. The time of this effect is marked in Fig. 9.15a by the right dashed up-arrow at $t = 16:36$. This new effective location of the bottleneck is approximately the same effective location of the bottleneck, which was before $t \approx 15:40$ (Fig. 9.15a).

Thus, we can conclude the following. Firstly, during about 3.7 hours the effective location of the bottleneck is about 1 km upstream of the location of the bottleneck. Later, during about one hour the effective location of the bottleneck is about 4 km upstream of the location of the bottleneck. Then the effective location of the bottleneck is about 1 km upstream of the location of the bottleneck. An overview of this effect is shown in Fig. 9.16.

The dependence of the effective location of the bottleneck on time is also accompanied by another effect. This effect is the appearance of the region where synchronized flow alternates with free flow. This alternation of free and synchronized flows occurs in space and time.
Fig. 9.15. Dependence of an effective location of a bottleneck on time. (a) Average speed at different detectors upstream of the bottleneck due to the off-ramp D25-off as functions of time. (b) Percentage of vehicles η that want to leave the main road to the off-ramp as a function of time. (c) Total discharge flow rate across the freeway as a function of time. In (b, c) 10-min average data is shown. Layout of the freeway infrastructure and local measurements on a section of the freeway A5-North is shown in Fig. 2.2
This effect is observed when the effective location of the bottleneck is at the detectors D20, i.e., during the time interval 15:40–16:36 (the time interval between dashed up-arrows in Fig. 9.15a). The freeway region where the alternating effect occurs is between the detectors D20 and D23. It can be seen that during the time interval 15:40–16:36 many $F \rightarrow S$ transitions occur in free flow at the detectors D23–D21 (these transitions are marked by solid up-arrows in Fig. 9.15a). Short-time living regions (1–2 minutes long time intervals) of synchronized flow should appear on the freeway between D23–D21 due to these local $F \rightarrow S$ transitions. Thus, although the effective location of the bottleneck is at the detectors D20 during the time interval 15:40–16:36 there are small and very short-time living regions of synchronized flow within free flow between the detectors D23–D21. The occurrence of these synchronized flow regions within free flow can correlate to one another at different detectors, but they do not necessarily correlate.

An important characteristic of a bottleneck due to an off-ramp is the percentage of vehicles $\eta$ that want to leave the main road to the off-ramp. In empirical data, this percentage is a function of time. We determine empirical function $\eta(t)$ at D25, i.e., at the location just downstream of the end of the merging region of the off-ramp:

\[
\eta = \frac{q_{\text{off}}}{q_{\text{off}} + q_d},
\]

where $q_{\text{off}}$ is the total flow rate to the off-ramp and $q_d$ is the total flow rate across the freeway just downstream of the end of the merging region of the
off-ramp. In the case under consideration, \( q_d = q_{D25} \); \( q_{D25} \) is the total flow rate across the freeway measured at the detectors D25 on the main road.

The function \( \eta(t) \) is shown in Fig. 9.15b. However, we cannot see some peculiarities in the dependence of \( \eta \) on time, which should explain the sudden change in the effective location of the bottleneck at \( t \approx 15:40 \). The same conclusion can be made if the dependence of total discharge flow rate on time is considered (Fig. 9.15c). It can be assumed that the change of the effective location of the bottleneck is related to the random effect of lane changing from the left and middle lanes to the right lane of vehicles that want to leave the main road via the off-ramp (see also an explanation of an \( F \rightarrow S \) transition at the off-ramp in Sect. 10.5).

**Effective Location of Bottleneck in Off-Ramp Lane**

Above we have considered two examples of pattern formation at an effectual bottleneck due to an off-ramp. We found that the effective location of the bottleneck, where the downstream front of the pattern is fixed, is upstream of the location of the bottleneck (the beginning of the off-ramp lane). However, it can turn out that the effective location of the bottleneck can also be downstream of the location of the bottleneck. This happens that an \( F \rightarrow S \) transition occurs in the off-ramp lane(s) rather than on the main road upstream of the location of the bottleneck. In this case, the downstream front of synchronized flow can be fixed somewhere in the off-ramp lane and a congested pattern expands upstream in the off-ramp lane. After this pattern has reached the main road, the pattern propagates upstream on the main road. The effect of synchronization of speeds between different freeway lanes on the main road upstream of the off-ramp should not necessarily occur [214]. There are at least the following scenarios of pattern formation in this case:

1. A moving jam emerges in synchronized flow already in the off-ramp lane(s). This moving jam propagates upstream and reaches the main road. If there are still free flow conditions on the main road upstream of the off-ramp, then there are two possibilities. If the flow rate on the main road is lower than the characteristic flow rate \( q_{out} \) in the wide moving jam outflow the moving jam dissolves. If, in contrast, the flow rate is higher than \( q_{out} \) the moving jam transforms into a wide moving jam. This case will be considered in Sect. 12.6.

2. Synchronized flow propagates upstream and causes traffic congestion first only in the right lane of the main road. In the other freeway lanes (the middle and left lanes) free flow first remains. Congestion in the right lane leads to an \( F \rightarrow S \) transition upstream of the off-ramp on the main road.

3. In comparison with the scenario (2), the effect of synchronization of speeds between different freeway lanes on the main road upstream of the off-ramp does not occur, i.e., congestion in the right lane does not lead to synchronized flow upstream of the off-ramp across all lanes of the
main road: in the left lane (sometimes also in the middle lane) free flow remains. However, a more detailed consideration of such two-dimensional spatial effects is beyond the scope of this book.

In conclusion to this discussion of the GP of type (2), let us note that besides the GPs whose empirical examples have been considered above, a dissolving GP (DGP) has also often been observed at isolated bottlenecks. The peculiarity of DGPs is that up to now this type of pattern has been found during pattern evolution or pattern transformation at isolated bottlenecks only. Concerning pattern transformation (Chap. 13), it should be noted that diverse pattern transformations between different SPs and DGPs have very often been observed over time at isolated bottlenecks due to an off-ramp.

9.5 Conclusions

(i) An identification of effectual isolated bottlenecks is important for an empirical traffic congested pattern study and classification.

(ii) In accordance with three-phase traffic theory, there are two main types of empirical congested patterns at effectual isolated bottlenecks: synchronized flow patterns (SP) and general patterns (GP).

(iii) There are three types of empirical SPs at isolated bottlenecks: an WSP, an LSP, and an MSP.

(iv) In synchronized flow within an SP, the flow rate does not usually change significantly in comparison with the flow rate in free flow before and after SP emergence.

(v) There are empirical GPs of type (1) and (2). In an GP of type (1), the upstream front of synchronized flow in the pinch region of the GP is bounded by the region of wide moving jams. In an GP of type (2), the upstream front of synchronized flow is upstream of the farthest upstream wide moving jam in the GP.

This effect can be explained as follows [214]. The synchronization of speeds between different freeway lanes can be hindered at an off-ramp bottleneck (or at all diverge bottlenecks where traffic is splitting up into at least two routes). If a fraction of vehicles, which have to choose the off-ramp, is high enough, these vehicles in the vicinity of the off-ramp cannot change to those lanes, which are related to the straight route, even if the vehicle speed is higher there. Therefore, an F—S transition should be hindered in this case. This effect has also been observed on US freeways [185,186].
10 Empirical Breakdown Phenomenon: Phase Transition from Free Flow to Synchronized Flow

10.1 Introduction

We have already noted in Chaps. 1 and 2 that the breakdown phenomenon has empirically been studied by many scientific groups in various countries (e.g., [30, 37, 58, 59, 61–79, 182]). In three-phase traffic theory, the breakdown phenomenon is explained by a first-order local F \rightarrow S transition [205, 208]. A qualitative theory of this phenomenon has been discussed in Chap. 5.

In this chapter, we present results of empirical studies of the breakdown phenomenon. These results should help us understand the probabilistic nature of this phenomenon and the consequences of this phenomenon for subsequent congested pattern formation. In particular, it will be found that

(i) after the breakdown phenomenon in an initial free flow at an effectual isolated bottleneck, the emerging congested traffic at the bottleneck first belongs to the “synchronized flow” phase rather than to the “wide moving jam” phase [208].

(ii) There can be either spontaneous or induced breakdown phenomena (spontaneous or induced F \rightarrow S transitions) at the bottleneck [213, 218].

(iii) After the onset of congestion due to one of these breakdown phenomena, qualitative features of a congested pattern that is formed upstream of the bottleneck do not necessarily depend on whether a spontaneous F \rightarrow S transition or an induced F \rightarrow S transition has been the reason for congested pattern formation [213, 218].

(iv) The formation and the dissolution of congested patterns are usually accompanied by a hysteresis effect (see references in e.g., [21, 30, 37]).

These features alone, which are typical of first-order phase transitions in non-equilibrium spatially distributed physical systems (e.g., [339]), can confirm that the empirical breakdown phenomenon is a first-order local F \rightarrow S transition (see also Appendix A).

To explain the possibility of an induced F \rightarrow S transition at a bottleneck, note that from theory and experimental studies of local first-order phase transitions in non-equilibrium spatially distributed physical systems (e.g., [339]) it is well-known that in many cases an induced phase transition occurs rather than a spontaneous phase transition. In particular, a phase transition in a
physical distributed system can be induced by the propagation of a spatiotemporal pattern through the system. We have already discussed in Sect. 2.4.4 that an induced F→S transition (induced breakdown phenomenon) can also occur in traffic flow at an effectual bottleneck. This is possible if traffic demand is high enough for the occurrence of an F→S transition in an initial free flow at the bottleneck. The F→S transition can be induced when a wide moving jam propagates through the bottleneck or a local region of synchronized flow that has initially occurred downstream of the bottleneck reaches the bottleneck.

After the “synchronized flow” phase has emerged (due to either a spontaneous F→S transition or an induced F→S transition), synchronized flow can exhibit very complex behavior. In particular, growing narrow moving jams can emerge spontaneously in synchronized flow, leading to an S→J transition. Besides this feature of synchronized flow, which will be considered in Chap. 12, there are some other interesting peculiarities of synchronized flow, which will be discussed in Sect 10.7.

10.2 Spontaneous Breakdown Phenomenon (Spontaneous F→S Transition) at On-Ramp Bottlenecks

Both spontaneous and induced speed breakdown phenomena often occur at a bottleneck due to an on-ramp. We discuss empirical features of a spontaneous F→S transition at the bottleneck.

The usual scenario of the spontaneous F→S transition at the bottleneck due to the on-ramp is the following. Firstly, free flow exists on the main road both at the bottleneck (D6) and upstream (D5) and also downstream (D7) of the bottleneck (t < 06:35 in Fig. 10.1a, b). Then an F→S transition at the bottleneck occurs (up-arrow in Fig. 10.1c (D6) and arrow in Fig. 10.1e). The speed at D6 becomes considerably lower than the minimum vehicle speed at the limit point for free flow \( \rho_{\text{max}}^{(\text{free emp})} \), \( q_{\text{max}}^{(\text{free emp})} \).

During the F→S transition at D6 free flow is observed both upstream (D5) and downstream (D7). The downstream front of a pattern that is developing after this transition has occurred is fixed in the vicinity of the location of

Note that in the theoretical consideration made in Chap. 5 we have distinguished between the limit point for free flow away from freeway bottlenecks (a hypothetical case of a homogeneous road) and the limit point for free flow at an effectual freeway bottleneck. Here and below in this Part II where empirical features of traffic are considered we cannot make a distinction between these two limit points. This is because in empirical data presented in the book (and all other empirical data that the author knows) the flow rate in free flow away from bottlenecks is essentially determined by the maximum flow rate downstream of the related effectual bottleneck. This is true even when free flow occurs at the bottlenecks (e.g., freeway sections in Figs. 2.1 and 2.2).
Fig. 10.1. F→S transition at the on-ramp (D6). (a, b) Overview of vehicle speed averaged across all freeway lanes (a) and total flow rate across the freeway (b). (c, d) Vehicle speed (c) and flow rate (d) at different detectors. (e, f) F→S transition in the flow-density plane. In (e, f) data averaged across all three freeway lanes (per lane) (e) and for the left lane (f) is shown. Taken from [218]
Empirical F→S Transition

The detectors D6 (Fig. 10.1a). There is no external reason for this speed breakdown. Thus, this speed breakdown is a spontaneous F→S transition at the on-ramp.

The F→S transition leads to an abrupt speed decrease in traffic flow. The duration of the F→S transition, i.e., the time interval when the speed during the F→S transition decreases from an initial speed in free flow to a synchronized flow speed is usually not appreciably higher than about 1 min (Fig. 10.1c; D6).

It must be noted that even though there is an abrupt speed decrease, there should not necessarily be a decrease in flow rate during an F→S transition. This can be seen in Fig. 10.1d and in more detail in Fig. 10.1e,f at the detectors D6. It can be seen that the flow rate in the left lane almost does not change during the F→S transition (Fig. 10.1f). However, if we consider the flow rate averaged across all three freeway lanes (Fig. 10.1e), we find that the flow rate is considerably higher in synchronized flow than this flow rate has been in free flow at the time of the F→S transition (t = 06:37). The average flow rate in synchronized flow just after the F→S transition is even a slightly higher than the average maximum flow rate in free flow (Fig. 10.1e,f). This is a very important feature of an F→S transition, which qualitatively distinguishes this phase transition from moving jam formation. In the latter case, both the speed and flow rate decrease abruptly during moving jam emergence (see Chap. 12).

After the F→S transition at D6, the average vehicle speed in synchronized flow exhibits only relatively small changes of about 10% over time near 65 km/h at D6 (Fig. 10.1). The same behavior of the speed in synchronized flow is observed on all other days at D6.

It must be noted that the F→S transition leads to further self-sustaining of synchronized flow at the on-ramp bottleneck (during about 2.5 hours for the case in Fig. 10.1). However, there are many cases when an F→S transition at the on-ramp does not lead to the effect of the self-sustaining of synchronized flow: synchronized flow exists on the main road at the on-ramp bottleneck only during a short time interval. Such cases are shown in Fig. 10.2a (up-arrows 1, 2, and 3; D6) where for comparison the F→S transition at t ≈ 06:37 (marked by the up-arrow 4) considered above (Fig. 10.1) is shown.

To understand this different behavior, recall that the vehicle speed in synchronized flow is always lower than the minimum vehicle speed in free flow. The density in synchronized flow after an F→S transition is usually higher than an initial density in free flow just before the transition. Therefore, corresponding to the Stokes shock-wave formula (3.5), after a local region of synchronized flow has occurred at the on-ramp, the upstream front of this synchronized flow can propagate upstream only if the average flow rate in free flow upstream is higher than the average flow rate in synchronized flow. A comparison of the total flow rate qD6 across the freeway at D6 (solid
10.2 Spontaneous Breakdown Phenomenon at On-Ramps

(a) A5-South, March 17, 1997
- left lane -- middle lane ··· right lane

Fig. 10.2. Comparison of different F→S transitions at the on-ramp bottleneck at D6. (a) Vehicle speed at D6 and D5. (b) Flow rates at D6 and upstream of D6 ($q_{up}$). (c) F→S transitions in the flow–density plane, which do not lead to pattern formation (up-arrows 1–3 in (a, b), respectively). Taken from [218]

curve “D6” in Fig. 10.2b) with the sum of the flow rates $q_{up}$ (dashed curve) upstream of D6 for these different F→S transitions is made in Table 10.1. Here

$$q_{up} = q_{D5} + q_{D6-on} + q_{D5-on} - q_{D6-off}, \quad (10.1)$$

where $q_{D5}$, $q_{D6-on}$, $q_{D5-on}$, and $q_{D6-off}$ are the total flow rate across the freeway at D5, the flow rates at the on-ramp D6-on, at the on-ramp D5-on, and at the off-ramp D6-off, respectively.

In the cases 1, 2, and 3, $q_{up} < q_{D6}$ (table 10.1). This explains why the upstream front of synchronized flow at the on-ramp (D6) does not propagate upstream. In contrast, for the F→S transition, which is marked by up-arrows 4 in Fig. 10.2a,b, $q_{up} > q_{D6}$ (Table 10.1). As a result, the upstream front of synchronized flow propagates upstream and reaches the location of D5 (up-arrow 4 on Fig. 10.2a; D5). In this case, the self-sustaining of synchronized flow on the main road at the on-ramp bottleneck occurs.
Table 10.1. Average flow rates during F→S transitions. Time intervals are given over which the flow rate after the related F→S transition has occurred has been averaged. These intervals are chosen to be greater than the trip time of vehicles between D5 and D6 (the latter is less than 3 min)

<table>
<thead>
<tr>
<th>F→S transition</th>
<th>Interval</th>
<th>$q_{D6}$ [vehicles/h]</th>
<th>$q_{up}$ [vehicles/h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrow 1</td>
<td>(t = 6:18)</td>
<td>6:18-6:20</td>
<td>5080</td>
</tr>
<tr>
<td>arrow 2</td>
<td>(t = 6:27)</td>
<td>6:27-6:31</td>
<td>6230</td>
</tr>
<tr>
<td>arrow 3</td>
<td>(t = 6:34)</td>
<td>6:34-6:36</td>
<td>6180</td>
</tr>
<tr>
<td>arrow 4</td>
<td>(t = 6:37)</td>
<td>6:37-6:39</td>
<td>6020</td>
</tr>
</tbody>
</table>

10.3 Probability for F→S Transition

Studying the probability for breakdown phenomenon (F→S transition) at a freeway bottleneck, Persaud et al. [61] have analyzed the flow rate $q$ in free flow on different days at which an F→S transition occurs at the bottleneck. All observed flow rates have been divided into groups in multiples of 100 vehicles/h. The frequency of flow rates, $N_i$, in each group $i$ in free flow has been studied before the F→S transition occurs. For each flow rate group $i$ over a large number of days, the number of instances $n_i$ is found in which the flow rate $q_i$ of the F→S transition at the bottleneck falls within the range of flow rate group $i$. Thus, for the calculation of the probability $P_{FS}^{(B)}$ for an F→S transition for the flow rate group $i$ the following formula is used [61]:

$$P_{FS}^{(B)} = \frac{n_i}{N_i}.$$  \hspace{1cm} (10.2)

It can be seen from Fig. 10.3 that the probability $P_{FS}^{(B)}$ is a sharp function of the flow rate in free flow in the vicinity of the critical point, where the probability $P_{FS}^{(B)} = 1$. This feature confirms that the breakdown phenomenon is a first-order F→S transition (compare with the theoretical discussion of the probability $P_{FS}^{(B)}$ in Sect. 5.3.4). This behavior of the probability $P_{FS}^{(B)}$ is qualitatively independent of the averaging time interval $T_{av}$ for the flow rate. However, the increase in this averaging time interval leads to a decrease in the critical flow rate (curves 1 min and 10 min in Fig. 10.3). One of the reasons is that in empirical data the flow rate in free flow usually changes considerably over time. As a result, in 10 minutes of average data (Fig. 10.3) there can be points of 1 minute of average data where the flow rate is considerably higher than the average flow rate in 10 minutes of data [61].
10.3 Probability for F→S Transition

Fig. 10.3. Probability for the breakdown phenomenon (F→S transition) at an on-ramp bottleneck for two different averaging time intervals $T_{av}$. Taken from Persaud et al. [61]

The probability for an F→S transition at freeway bottlenecks as a function of flow rate has also been found by Lorenz and Elefteriadou [60]. The results are qualitatively similar to the findings of Persaud et al. [61].

10.3.1 Empirical and Theoretical Definitions of Freeway Capacities at Bottlenecks

A definition of freeway capacity made in Sect. 8.3.1 is related to a time-independent flow rate in free flow. In contrast, in empirical observations of free flow, the flow rate is usually a complicated function of time and empirical data is averaged for a time interval $T_{av}$. In this case, the following definition of freeway capacity in free flow at a freeway bottleneck can be made:

- Freeway capacity in free flow is equal to the average flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability

$$P_C^{(B)} = 1 - P_{FS}^{(B)} < 1$$  \hspace{1cm} \text{(10.3)}

during a given averaging time interval $T_{av}$ for traffic variables. This means that at this average flow rate with the probability $P_{FS}^{(B)} > 0$ an F→S transition (breakdown phenomenon) occurs at the bottleneck during the time interval $T_{av}$.

There can be an infinite number of the flow rates downstream of the bottleneck for which the condition (10.3) is satisfied. As a result, there can also be an infinite number of freeway capacities in free flow at the bottleneck. Every freeway capacity has two attributes:
(1) The averaging time interval $T_{av}$ for empirical data.

(2) The probability $P^{(B)}_C$, which is lower than 1, that free flow remains at the bottleneck during the time interval $T_{av}$.

When for free flow at a bottleneck rather than the condition \((10.3)\) the opposite condition $P^{(B)}_C = 1$ is satisfied, the flow rate in this free flow is lower than freeway capacity. This means that freeway capacity is not reached in this free flow at the bottleneck.

From \((10.3)\) we can see that to study empirical freeway capacity at a bottleneck, the empirical probability $P^{(B)}_{FS}$ for an $F\rightarrow S$ transition at the bottleneck should be found. However, the $F\rightarrow S$ transition at the bottleneck does not necessarily lead to congested pattern formation at the bottleneck. An example is shown in Fig. 10.2a. There are three $F\rightarrow S$ transitions at the bottleneck due to the on-ramp (up-arrows 1, 2, and 3; D6 in Fig. 10.2a), which do not cause the formation of congested patterns upstream of the bottleneck: synchronized flows occurring at the bottleneck due to these transitions exist during short time intervals only. These time intervals are comparable with the averaging time interval $T_{av}$ for empirical data ($T_{av} = 1$ min in Fig. 10.2a). In contrast to these transitions, the $F\rightarrow S$ transition, which is marked by up-arrow 4 in Fig. 10.2a, leads to congested pattern formation. This congested pattern exists at the bottleneck during a long time interval that is considerably longer than $T_{av}$. This congested pattern is an GP at bottleneck $B_3$ (the bottleneck at D6) shown in Fig. 9.3a.

We can assume that when empirical freeway capacity is measured, one can neglect $F\rightarrow S$ transitions that do not lead to congested pattern formation at a bottleneck (e.g., up-arrows 1, 2, and 3; D6 in Fig. 10.2a). Thus, under the above assumption when the probability $P^{(B)}_C = 1 - P^{(B)}_{FS}$ is studied, only those of the $F\rightarrow S$ transitions at the bottleneck, which lead to congested pattern formation at the bottleneck, should be taken into account. The congested pattern should further exist for a time interval that is considerably higher than $T_{av}$.

We see that to find the probability $P^{(B)}_{FS}$ \((10.2)\) as a function of the flow rate $q_i$, many different realizations (days) are required where $F\rightarrow S$ transitions occur and lead to congested pattern emergence at a bottleneck. However, conditions for these $F\rightarrow S$ transitions can considerably depend on traffic control parameters (weather, etc.). Therefore, the function $P^{(B)}_{FS}(q_i)$ is related to the probability for the onset of congestion at the bottleneck \textit{averaged over different traffic control parameters}, which can be realized on different days. In other words, when empirical freeway capacity in free flow at the bottleneck associated with given values of probability $P^{(B)}_C$ \((10.3)\) and the time interval $T_{av}$ should be found, this empirical freeway capacity is always \textit{average freeway capacity} in free flow at the bottleneck. This empirical freeway capacity is averaged over different traffic control parameters on the days where the onset
of congestion has been observed and used by a calculation of the probability $P_{FS}^{(B)}(q_i)$ (10.2).

In the three-phase traffic theory, we have noted that there are an infinite number of the maximum freeway capacities in free flow at a bottleneck due to the on-ramp (Sect. 8.3). In empirical observations, the maximum freeway capacities $q_{\text{max}}^{(\text{free B})}$ in free flow at the bottleneck are also determined through the condition: the maximum freeway capacities are the flow rates $q_i$ at which the probability $P_{FS}^{(B)}$ (10.2) for an $F \rightarrow S$ transition reaches 1:

$$P_{FS}^{(B)} = \frac{n_i}{N_i} \bigg|_{q_i = q_{\text{max}}^{(\text{free B})}} = 1.$$  \hspace{1cm} (10.4)

In this case, the probability $P_{FS}^{(B)}$ (10.3) satisfies the condition (8.8): $P_{C}^{(B)} = 0$. This means that on average each maximum freeway capacity at a bottleneck cannot be observed during a longer time interval than the time interval $T_{av}$. This is related to an $F \rightarrow S$ transition leading to congested pattern formation. This $F \rightarrow S$ transition occurs in this free flow during the time interval $T_{av}$ with the probability $P_{FS}^{(B)} = 1$ (10.4).

However, if the definition (10.2) is used, the probability $P_{FS}^{(B)}$ is the probability for an $F \rightarrow S$ transition averaged over different sets of flow rates $q_{on}$ and $q_{in}$ that give the same flow rate $q_{\text{sum}}(q_{on}, q_{in}) = q_i$. In other words, in the empirical formulae (10.2) and (10.4), different sets of flow rates $q_{on}$, $q_{in}$ that give the same flow rate $q_i$ are not taken into account. In this case, the related average maximum freeway capacity is given by the formula (10.4) where the existence of different sets of flow rates $q_{on}$, $q_{in}$ that give the same flow rate $q_i(q_{on}, q_{in}) = q_{on} + q_{in}$ is neglected.

A more precise definition of the empirical probability of an $F \rightarrow S$ transition at a bottleneck due to the on-ramp is

$$P_{FS}^{(B)} = \frac{n_i}{N_i} \big|_{q_{on}, q_{in}}.$$  \hspace{1cm} (10.5)

Thus, the empirical maximum freeway capacities in free flow at the bottleneck should be derived using

$$P_{FS}^{(B)} = \frac{n_i}{N_i} \big|_{q_{on} + q_{in} = q_{\text{max}}^{(\text{free B})}} = 1.$$  \hspace{1cm} (10.6)

The maximum freeway capacities $q_{\text{max}}^{(\text{free B})}$ are functions of both the flow rates $q_{on}$, $q_{in}$ and the averaging time interval $T_{av}$.

In the definition (10.5), the probability $P_{FS}^{(B)}$ is found separately for every different set of flow rates $q_{on}$, $q_{in}$ that gives the same flow rate $q_i$. This definition is compatible with the definition of the maximum freeway capacities (10.6). However, in comparison with formula (10.4) from the paper by Persaud et al. [61], to measure the maximum freeway capacities (10.6) one needs more observations of different $F \rightarrow S$ transitions at a bottleneck. The
observations of an F→S transitions should be related to different sets of flow rates \( q_{on} \) and \( q_{in} \) that give the same flow rate \( q_{sum} = q_i \).

Let us compare the theoretical definition of the probability \( p_{FS}^{(B)} \) for an F→S transition at a bottleneck (8.9) with the empirical definition (10.5). For simplicity we assume that the flow rate within each of the averaging time intervals \( T_{av} \) is time-independent. Then the frequency of flow rates \( N_i \) in each group \( i \) in free flow in the empirical definition (10.5) is related to the number of the realizations \( N_{FS} \) in the theoretical definition (8.9), i.e.,

\[
N_i = N_{FS} , \quad (10.7)
\]

the flow rate \( q_i \) in (10.5) is related to the flow rate \( q_{sum} \) in (8.9), i.e.,

\[
q_i = q_{sum} , \quad (10.8)
\]

and the time interval \( T_{av} \) in the empirical definition (10.5) is related to the time interval \( T_{ob} \) in the theoretical definition (8.9), i.e.,

\[
T_{av} = T_{ob} . \quad (10.9)
\]

Accordingly, the number of instances \( n_i \) in (10.5) is related to \( n_{FS} \) in the theoretical definition (8.9) under the conditions (10.7)–(10.9). In other words, the empirical (10.5) and the theoretical definition (8.9) are equivalent under the conditions (10.7)–(10.9).

It must be noted that in empirical observations, the flow rate is not usually constant during the averaging time interval \( T_{av} \) for the flow rate. In addition, in empirical observations of a large number of flow rates \( N_i \), in each group \( i \), many different days should be studied. Traffic control parameters (weather, etc.) on these days can be very different. In other words, the empirical probabilities (10.2) and (10.5) are the probability for an F→S transition at the bottleneck averaged over different traffic control parameters observed on different days.

### 10.3.2 Pre-Discharge Flow Rate

**Definition of Pre-Discharge Flow Rate**

The flow rate \( q_i \) in free flow at which an F→S transition (speed breakdown) leads to congested pattern emergence at the bottleneck is often called the pre-discharge flow rate or the breakdown flow (e.g., [37, 61]). We denote the pre-discharge flow rate by \( q_{FS}^{(B)} \).

The term “pre-discharge” flow rate can be explained as follows. After the onset of congestion at a bottleneck, a congested pattern is formed on the main road upstream of the bottleneck. At this “congested bottleneck” the flow rate in free flow downstream of the bottleneck is determined by vehicles that accelerate from lower speeds in synchronized flow (congestion) at the bottleneck.
to higher speeds in free flow downstream of the bottleneck. This process of
the vehicle acceleration is called the discharge of congestion. Likewise, the
flow rate in free flow downstream of the congested bottleneck is called the
discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ (e.g., [37, 61]). We already used this term when we
considered a theory of congested patterns at isolated bottlenecks (Chap. 7)
and a theory of congested pattern capacity, i.e., capacity of congested bot-
tlenecks in Sect. 8.5.

**Pre-Discharge Flow Rate and Maximum Freeway Capacities**

In Sect. 8.3.2, we have seen that there are an infinite number of flow rates
in free flow downstream of a bottleneck due to an on-ramp associated with
points $(q_{\text{on}}, q_{\text{in}})$ on and between the boundaries $F_{\text{FS}}^{(B)}$ and $F_{\text{th}}^{(B)}$ in the diagram
of congested patterns (Fig. 8.1a). The pre-discharge flow rate $q_{\text{FS}}^{(B)}$ can be one
of these flow rates. This is because an $F \rightarrow S$ transition can occur and lead
to congested pattern emergence at the bottleneck with the probability $P_{\text{FS}}^{(B)} > 0$
during the time interval $T_{\text{av}}$ for any of these flow rates.

Recall that flow rates in free flow downstream of a bottleneck associ­
ated with the boundary $F_{\text{FS}}^{(B)}$ are the maximum freeway capacities in free
flow at the bottleneck, $q_{\text{max}}^{(B)}(q_{\text{on}}, q_{\text{in}})$ (8.6). For the maximum freeway
 capacities the probability for an $F \rightarrow S$ transition at the bottleneck $P_{\text{FS}}^{(B)} = 1$
(Sect. 8.3.2). Flow rates in free flow downstream of the bottleneck associ­
ated with the boundary $F_{\text{th}}^{(B)}$ are the minimum freeway capacities in free flow
at the bottleneck, $q_{\text{th}}^{(B)}(q_{\text{on}}, q_{\text{in}})$ (8.13). Thus, at a given flow rate $q_{\text{in}}$ the
pre-discharge flow rate $q_{\text{FS}}^{(B)}(q_{\text{on}}, q_{\text{in}})$ satisfies the condition

$$q_{\text{th}}^{(B)}(q_{\text{on}}, q_{\text{in}}) \leq q_{\text{FS}}^{(B)}(q_{\text{on}}, q_{\text{in}}) \leq q_{\text{max}}^{(B)}(q_{\text{on}}, q_{\text{in}}).$$

Consequently, the pre-discharge flow rate $q_{\text{FS}}^{(B)}(q_{\text{on}}, q_{\text{in}})$ is related to the
probability $P_{\text{FS}}^{(B)}$ for the $F \rightarrow S$ transition at the bottleneck, which satisfies the
condition

$$0 < P_{\text{FS}}^{(B)} \leq 1.$$  

From a comparison (8.19) with (10.10) we can conclude that the pre-discharge
flow rate $q_{\text{FS}}^{(B)}(q_{\text{on}}, q_{\text{in}})$ is equal to one of the freeway capacities $q_{C}^{(B)}(q_{\text{on}}, q_{\text{in}})$
in free flow on the main road at the bottleneck.

**Pre-Discharge Flow Rate and Limit Point of Free Flow**

Let us compare the pre-discharge flow rate $q_{\text{FS}}^{(B)}$ and the empirical flow rate
at the limit point of free flow $q_{\text{max}}^{(\text{free, emp})}$ measured downstream of a bottleneck (Sect. 2.3.3). The empirical flow rate at the limit point of free flow,
$

\rho_{\text{max}}^{(\text{free, emp})}, q_{\text{max}}^{(\text{free, emp})}$, can be restricted
(1) either by traffic demand upstream of the bottleneck,
(2) or by an F→S transition (onset of congestion) at the bottleneck,
(3) or else by the discharge flow rate from a congested pattern at another
effectual adjacent bottleneck that is upstream of the bottleneck.

It must be noted that the flow rate $q^{(\text{free, emp})}$ does not necessarily coincide with the pre-discharge flow rate $q^{(B)}$.

An empirical example where the pre-discharge flow rate $q^{(B)}_{FS}$ is lower than the flow rate at the empirical limit point of free flow $q^{(\text{free, emp})}$ is shown in Figs. 10.1f and 10.4. The flow rate at the empirical limit point of free flow $q^{(\text{free, emp})}_{FS}$ is related to the time $t = 06:29$ that is earlier than the time of an F→S transition ($t = 06:37$, up-arrows in Fig. 10.4) leading to congested pattern emergence at the bottleneck. This empirical result can be explained in at least two ways.

(1) Recall that the maximum freeway capacity is related to the probability for an F→S transition $P^{(B)}_{FS} = 1$. At the flow rate $q^{(\text{free, emp})}_{FS}$ no F→S transition has occurred at the bottleneck (Fig. 10.4). This means that the flow rate $q^{(\text{free, emp})}_{FS}$ corresponds to $P^{(B)}_{FS} < 1$. In other words, in the case under consideration the flow rate $q^{(\text{free, emp})}_{FS}$ is lower than the maximum freeway capacity. The pre-discharge flow rate $q^{(B)}_{FS}$ is lower than $q^{(\text{free, emp})}_{FS}$ (Fig. 10.4). It can be assumed that the flow rate $q^{(B)}_{FS}$ is also related to $P^{(B)}_{FS} < 1$, i.e., the pre-discharge flow rate is lower than the related maximum freeway capacity. If the latter is correct, then it was a random event on this day that the F→S transition occurred at the flow rate $q^{(B)}_{FS}$, which is lower than $q^{(\text{free, emp})}_{FS}$.

Fig. 10.4. Explanation of the pre-discharge flow rate $q^{(B)}_{FS}$ and the flow rate at the empirical limit point of free flow $q^{(\text{free, emp})}_{FS}$. Flow rate (a) and speed (b), which are averaged across three lanes of the freeway A5-South. 1-min data is measured at the detectors D6 on March 17, 1997. In the flow–density plane, this data is shown in Fig. 10.1f. Up-arrows labeled “F→S” show an F→S transition leading to congested pattern emergence at a bottleneck.
(2) In empirical data presented in Fig. 10.4 during the time interval 06:00–06:37 the flow rate to the on-ramp $q_{on}$ was on average an increasing function of time. Therefore, the flow rate $q^{(\text{free, emp})}$ is related to lower flow rate $q_{on}$ than the pre-discharge flow rate $q_{FS}^{(B)}$. Thus, we can assume that at a given flow rate in free flow downstream of the bottleneck $q_{sum}$, the higher the flow rate $q_{on}$, the more likely the onset of congestion at the bottleneck.

If the pre-discharge flow rate $q_{FS}^{(B)}$ is equal to one of the maximum freeway capacities $q^{(\text{free B})}$, then the pre-discharge flow rate is associated with the probability for an $F \rightarrow S$ transition $P_{FS}^{(B)} = 1$ (8.7). However, even in this limit case, the pre-discharge flow rate is not necessarily equal to $q^{(\text{free, emp})}$. To explain this, recall that the maximum freeway capacity $q^{(\text{free B})}$ can appreciably depend on the flow rates $q_{on}$ and $q_{in}$. As mentioned above, there can be other flow rates $q_{on}$ and $q_{in}$ for which the maximum freeway capacity $q^{(\text{free B})}$ is higher than the pre-discharge flow rate $q_{FS}^{(B)}$. Thus, there can be the flow rate $q_{sum} = q_{on} + q_{in}$ that is lower than the latter maximum freeway capacity, however that is higher than the pre-discharge flow rate. For this flow rate $q_{sum}$ an $F \rightarrow S$ transition does not necessarily occur at the bottleneck. In the latter case, the flow rate $q_{sum}$ that is higher than the pre-discharge flow rate $q_{FS}^{(B)}$ can be observed at the bottleneck before the $F \rightarrow S$ transition associated with the pre-discharge flow rate $q_{FS}^{(B)}$ occurs at this bottleneck.

It can turn out that the empirical limit point of free flow is associated with the onset of congestion at the bottleneck. In this special case, $q^{(\text{free, emp})} = q_{FS}^{(B)}$. Nevertheless, the flow rate $q^{(\text{free, emp})}$ can be lower than the maximum freeway capacity. This follows from the condition (10.11): the pre-discharge flow rate $q_{FS}^{(B)}$ can be related to $P_{FS}^{(B)} < 1$.

It should be noted that as the maximum freeway capacities $q^{(\text{free B})}$, the empirical maximum flow rate $q^{(\text{free, emp})}$, and the pre-discharge flow rate $q_{FS}^{(B)}$ usually decrease when the time interval $T_{av}$ increases (at least over some range of $T_{av}$).

From the above discussion we can conclude that the maximum freeway capacities $q^{(\text{free B})}$, the flow rate at the empirical limit point of free flow $q^{(\text{free, emp})}$, and the pre-discharge flow rate $q_{FS}^{(B)}$ can in general take on very different values.

### 10.4 Induced Speed Breakdown at On-Ramp Bottlenecks

Thus, there is a wide range of the flow rate downstream of a bottleneck where the probability for an $F \rightarrow S$ transition at the bottleneck $P_{FS}^{(B)}(q_{on}, q_{in}) > 0$. This means that a time-limited high enough decrease in the speed at the
bottleneck, which is caused by an external disturbance in free flow, can be the reason for the onset of congestion on the main road at the bottleneck. Congestion persists on the main road at the bottleneck even if the external reason for this decrease does not exist any more. In this case, one speaks of an induced F→S transition (induced speed breakdown) [218]. This time-limited external disturbance in the speed at the bottleneck occurs when either a wide moving jam or a region of synchronized flow propagates on the main road through the bottleneck.

10.4.1 F→S Transition Induced by Wide Moving Jam Propagation Through Effectual Bottleneck

An example of an F→S transition on the main road at an on-ramp bottleneck, which is induced by wide moving jam propagation through the bottleneck (bottleneck B2 due to the on-ramp at D16) on the freeway A5-South, is shown in Figs. 4.2 and 10.5. Firstly, it should be noted that during the whole time before the wide moving jam reaches the bottleneck free flow conditions are realized on the main road at the bottleneck at D16, upstream (D15), and also downstream (D17) of the bottleneck (Fig. 10.5). However, after the wide moving jam has passed the bottleneck synchronized flow is formed on the main road at the bottleneck. This synchronized flow exists for a long time at the bottleneck, and the downstream front of synchronized flow is fixed at the bottleneck. Thus, the F→S transition is induced at the on-ramp bottleneck during jam propagation.

It can be assumed that after the wide moving jam has just passed the bottleneck, the still slowly moving vehicles on the main road that accelerate from low speed states within the moving jam force vehicles in the on-ramp lane to move slowly, too. In turn, the latter slow vehicles prevent the acceleration of vehicles on the main road to free flow speed in the vicinity of the on-ramp bottleneck, even when the moving jam due to jam upstream propagation is far upstream of the bottleneck.

10.4.2 Induced Speed Breakdown at Bottlenecks Caused by Synchronized Flow Propagation

A synchronized flow pattern (SP) has occurred at the bottleneck B1 (the off-ramp, D23-off) in Fig. 10.6a. Both the upstream and downstream fronts of this SP are moving upstream, i.e., an MSP occurs as discussed in Sect. 9.3.3. When the upstream front of the MSP reaches the bottleneck B2 (the on-ramp, D16), another SP that is further localized at the bottleneck B2 is formed. The downstream front of this LSP is fixed at the on-ramp and this LSP is self-sustaining from 06:49 to 09:25 (Figs. 2.16 and 10.6a). Thus, the upstream propagation of the initial MSP indeed induces an F→S transition at the on-ramp.
10.4 Induced Speed Breakdown at On-Ramps

(a) A5-South, June 23, 1998

Fig. 10.5. $F \rightarrow S$ transition at the on-ramp bottleneck at D16. (a) Vehicle speed (left) and flow rate (right) at different detectors. (b) $F \rightarrow S$ transition in the flow-density plane. Taken from [218]
Fig. 10.6. F→S transition at the on-ramp bottleneck at D16. (a, b) Overview of speed (a) and flow rate (b) during the time interval 06:00–08:00. (c, d) Vehicle speed (c) and flow rate (d) at different detectors. (e) F→S transition in the flow–density plane. MSP: the moving SP, LSP: the localized SP. Overview of the congested pattern during the whole time of pattern existence is shown in Fig. 2.16. Taken from [218]
This induced F→S transition is qualitatively different from the induced F→S transition at the bottleneck due to wide moving jam propagation discussed above. In contrast to the case of wide moving jam propagation where the jam propagates through the bottleneck B₂ while maintaining the downstream jam front velocity (Figs. 4.2 and 10.5), the MSP is caught at the on-ramp (the catch effect).² Indeed, after the MSP has reached the bottleneck a qualitatively different pattern (LSP) occurs there. This LSP is determined mostly by bottleneck characteristics, traffic demand, and freeway peculiarities upstream (Sect. 14.3.2). There is also another difference between a wide moving jam and an MSP: in contrast to the wide moving jam, within the MSP the vehicle speed is relatively high (about 40–70 km/h, Fig. 10.5) and the average flow rate within the MSP is only a slightly lower than in the initial free flow.

In another example, there are first at least four spontaneous F→S transitions at the on-ramp bottleneck that do not lead to the effect of self-sustaining of synchronized flow at the on-ramp (up-arrows 1–5 in Fig. 10.7). The latter cases are similar to the ones considered above (see Fig. 10.2, up-arrows 1, 2, and 3).

After a local region of synchronized flow (D7) that has initially occurred between the bottlenecks B₂ and B₃ has reached the bottleneck B₃ (D6), synchronized flow emerges upstream of the bottleneck B₃ (D5). This synchronized flow is further self-sustaining for 2:20 hours whereas the initial local region of synchronized flow disappears after about 15 min. Thus, the upstream propagation of synchronized flow also induces an F→S transition on the main road at the on-ramp bottleneck at D6.

10.5 Breakdown Phenomenon at Off-Ramp Bottlenecks

It is well-known that traffic congestion on the main road upstream of an off-ramp bottleneck can occur if the fraction of vehicles that leave the main road to the off-ramp is high enough (e.g., [21]). An example is shown in Fig. 10.8 where the ratio $\delta = \frac{q_{\text{right}}}{q_{\text{total}}}$ of the flow rate $q_{\text{right}}$ of vehicles moving in the right lane of the main road to the total flow rate on the road $q_{\text{total}}$ is a continuously increasing function of the distance from intersection I₂ to intersection I₃ (Fig. 10.8b). As a result, the flow rate in the right lane increases drastically from D₂₁ (21.8 km) to D₂₂ (22.9 km).

The latter effect causes an F→S transition in the right lane at D₂₂ (up-arrow at D₂₂, Fig. 10.8a). Most of the vehicles in the middle and left lanes at D₂₂ have a different route in comparison to those vehicles that want to

² A brief discussion of the catch effect can be found in Sect. 2.4.6.
Fig. 10.7. Induced F→S transition at the on-ramp bottleneck at D6 caused by the propagation of a local region of synchronized flow. (a) Vehicle speed (left) and flow rate (right) at different detectors. (b) F→S transition in the flow-density plane (black points: free flow, circles: synchronized flow). Taken from [218]

leave via the off-ramp. This can be the reason why the synchronization of the speeds in different freeway lanes does not occur at D22.

Synchronized flow occurs only at D21, i.e., about 1.5 km upstream of the off-ramp. The reason for this synchronized flow and for the related F→S transition can be the fall in vehicle speed at the bottleneck $B_1$ due to the off-ramp: between D21 and D22 many vehicles change to the right lane (Fig. 10.8b),
10.5 Breakdown Phenomenon at Off-Ramp Bottlenecks

(a) A5-South, April 20, 1998

![Graphs showing vehicle speed and flow rate at different detectors.](image)

(b) The ratio $\delta = q_{right}/q_{total}$ as a function of the distance.

(c) F→S transition in the flow–density plane (black points: free flow, circles: synchronized flow). Taken from [218]
where the speed is lower. These vehicles can force the vehicles in the middle
and left lanes, which want to continue on the freeway, to slow down.³

Due to this F→S transition at the bottleneck due to the off-ramp, an
MSP (Fig. 10.6a) occurs whose upstream propagation later causes an in­
duced F→S transition at the upstream adjacent bottleneck $B_2$ discussed
above (Sect. 10.4.2).

At both on- and off-ramps an F→S transition is accompanied by a drop
in vehicle speed (arrows in Figs. 10.1c, 10.5b, 10.6c, 10.7b, and 10.8a,c) last­
ing about one minute. Moreover, although the speed decreases appreciably
during the F→S transition, the flow rate in the emergent synchronized flow
can remain of the same order of magnitude as those in the initial free flow.
However, in comparison with F→S transitions at on-ramp bottlenecks, in
the case of an off-ramp bottleneck synchronized flow occurs at some distance
upstream of the off-ramp.

This effect is related to lane changing of vehicles that want to leave the
main road via the off-ramp: this lane changing (from the left and middle lanes
to the right lane) occurs upstream of the off-ramp lane. Apparently the same
reason is responsible for another result. The effective location of the bottle­
neck due to the off-ramp is located at some distance upstream of the off-ramp.
This result we already discussed in Sects. 9.3.1 and 9.4.3 when congested pat­
terns at the bottleneck $B_{\text{North}}$ were considered.

In the latter cases, the F→S transition (see up-arrows in Fig. 9.6, D18 and
in Fig. 9.14, D23) is also related to lane changing from the left and middle
lanes to the right lane. This lane changing occurs upstream of the off-ramp lane. Indeed, in both cases Figs. 9.6 and 9.14 the ratio $\delta = \frac{q_{\text{right}}}{q_{\text{total}}}$ is
a continuously increasing function of the spatial coordinate upstream of the
off-ramp at D25-off, as for example, on June 26, 1996 (Fig. 9.6) $\delta \approx 0.2$ at
D20 and $\delta \approx 0.3$ at D24.

It can be seen in Fig. 10.9b,d,f that in these three different cases (Figs. 10.8,
9.6 and 9.14) F→S transitions occur when the related total flow rates increase
over time. The percentage of vehicles $\eta$ that want to leave the main road to
the off-ramp also slightly increases on April 20, 1998. On March 23, 2001
and on June 26, 1996 in the related time intervals when the F→S transitions
occur the percentage $\eta$ does not change considerably.

³ On some freeways there are more than three lanes in the vicinity of the off-ramp.
Then vehicles in the leftmost lane are almost undisturbed by vehicles that leave
the freeway via the off-ramp. Thus, synchronization of vehicle speeds across the
freeway does not necessarily occur in some of the freeway lanes far from the right
lane. Nevertheless congested patterns in freeway lanes close to the right lane can
be expected. This possible two-dimensional spatial effect, however, requires a
separate study.
10.6 Breakdown Phenomenon Away from Bottlenecks

Let us first consider a dependence of average vehicle speed and flow rate within the time interval 06:20–06:40 (Fig. 10.10) in the vicinity of two freeway bottlenecks. The first one is due to the off-ramp D23-off (Fig. 2.1). The second bottleneck is due to the on-ramp at D16 (Fig. 2.1). It can be seen from the dependence of average vehicle speed (Fig. 10.10a) and flow rate (Fig. 10.10b) that in the whole time interval 06:20–06:40 in the vicinity of both freeway bottlenecks free flow is realized. In contrast, away from both of the bottlenecks an F→S transition occurs at D18 (up-arrow in Fig. 10.10a, D18). The speed appreciably decreases due to the F→S transition. However, the flow rate remains approximately the same (Fig. 10.10b, D18).

Let us show that the transition from free flow to synchronized flow at \( t = 06:36 \) in the vicinity of the detectors D18 (Fig. 10.10a, D18, up-arrow) is an F→S transition away from freeway bottlenecks. To do this, the spatiotemporal distributions of vehicle speed and flow rate both upstream and
downstream from the detectors D18 at later times should be studied. The results of such an investigation in the time interval 06:35–06:50 are shown in Fig. 10.11. It can be seen from Fig. 10.11 that transitions from free to synchronized flow both upstream (D17, D16 in Fig. 10.11a, up-arrows) and downstream (D19, D20 in Fig. 10.11a, up-arrows) occur later than at the detectors D18 (up-arrow at \( t = 06:36 \) in Fig. 10.11a). Moreover, the greater the distance from the detectors D18, the later the transition from free to synchronized flow. This is true both upstream (D17, D16 in Fig. 10.11a, up-arrows) and downstream (D19, D20 in Fig. 10.11a, up-arrows) of the detectors D18. This confirms that the transition from free flow to synchronized flow first
occurs only in the vicinity of the detectors D18, i.e., it is really the F→S transition away from freeway bottlenecks.

Note that the transitions from free flow to synchronized flow, which occur upstream (D17, D16 in Fig. 10.11a, up-arrows) of the detectors D18, are related to the appearance of a wave of induced F→S transitions. Correspondingly, the F→S transition downstream of the detectors D18 (D19, D20 in Fig. 10.11a, up-arrows) is related to a wave of the propagating synchronized flow. The propagation of synchronized flow supplants free flow downstream.
Indeed, as already mentioned, the transitions from free flow to synchronized flow, which occur upstream (D17, D16 in Fig. 10.11a, up-arrows) and downstream (D19, D20 in Fig. 10.11a, up-arrows) of the detectors D18, occur later than at the detectors D18. Furthermore, the greater the distance from the detectors D18, the later the F→S transitions. Therefore, the F→S transitions that occur upstream (D17, D16 in Fig. 10.11a, up-arrows) and downstream (D19, D20 in Fig. 10.11a, up-arrows) are not local F→S transitions, but induced F→S transitions. The induced F→S transitions upstream and propagating synchronized flow downstream cause a widening of synchronized flow, which first occurs spontaneously in the vicinity of the detectors D18: the region of localization of synchronized flow is widening both upstream and downstream over time. It must be noted that the detectors D20–D17 are located away from bottlenecks.

Comparison of F→S Transitions
Away from and at Freeway Bottlenecks

Let us compare the F→S transition that occurs away from freeway bottlenecks (Fig. 10.10) with an F→S transition at freeway bottlenecks.

First note that in the case of the F→S transition, which occurs away from freeway bottlenecks, the wave of induced F→S transition upstream of the location of the initial local phase transition \( t = 06:36, \) D18, Fig. 10.10a) at time \( t = 06:43 \) reaches the detectors D17 (up-arrow, Fig. 10.11a) and at time \( t = 06:44 \) reaches the detectors D16 (up-arrows in Fig. 10.11a). The latter detector is already located at a bottleneck due to the on-ramp. It might be expected that if the wave of induced F→S transitions propagates further upstream, then synchronized flow will be caught in the vicinity of this freeway bottleneck, i.e., it will be self-sustaining in the vicinity of the freeway bottleneck for a long time, as explained in Sect. 10.4.2. However, in the case under consideration it does not occur. In contrast, a reverse S→F transition is realized at the detectors D16 (dotted up-arrow in Fig. 10.12a).

Free flow that has occurred due to the mentioned reverse S→F transition has existed at the detectors D16 only during about 5 min. After this time, an F→S transition occurs in the vicinity of the freeway bottleneck due to the on-ramp at D16 (solid up-arrow in Fig. 10.12a).\(^4\)

It should be noted that at a distance of about 12.6 km from D16 at time \( t = 06:37 \) another F→S transition occurs at the detectors D6 at the freeway bottleneck due to the on-ramp (up-arrow in Fig. 10.1c). Both F→S transitions (solid up-arrow in Fig. 10.12a and up-arrow in Fig. 10.1c) show qualitatively the same peculiarities of first-order phase transitions considered above.

\(^4\) A discussion of moving jam emergence (down-arrows in Fig. 10.12) in synchronized flow that has occurred away from bottlenecks appears in Sect. 12.4.
Fig. 10.12. Results of empirical observations of an S→F transition (*dotted up-arrow* at D16), an F→S transition (*solid up-arrow* at D16) at the on-ramp bottleneck, and an S→J transition (*down-arrows*, see explanations in Sect. 12.4) on the main road away from freeway bottlenecks. Dependence of vehicle speed (a) and flow rate (b) is plotted at the different detectors upstream of the location of the S→J transition (D20). Down-arrows symbolically show the location of the moving jam at different detectors. Taken from [210]
Thus, there are at least two types of $F \rightarrow S$ transitions in empirically observed traffic flow on the same day:

(i) the $F \rightarrow S$ transition away from bottlenecks (up-arrow, D18, Fig. 10.10a);
(ii) the $F \rightarrow S$ transitions at the bottleneck (the bottleneck $B_2$ at D16, solid up-arrow in Fig. 10.12a and the bottleneck $B_3$ at D6, up-arrow in Fig. 10.1c).

The difference between these $F \rightarrow S$ transitions is that in the $F \rightarrow S$ transitions at the bottlenecks, synchronized flow can be self-sustaining in the vicinity of the bottleneck for several hours after the transition. In contrast, after the $F \rightarrow S$ transition away from bottlenecks has occurred synchronized flow usually exists for a relatively short time interval and it can propagate both upstream and downstream from the location where the $F \rightarrow S$ transition initially occurred. The second difference is that the $F \rightarrow S$ transition away from bottlenecks is observed very seldom.

However, these two types of $F \rightarrow S$ transitions show qualitatively the same breakdown effect in free flow (Fig. 10.13) that is well-known from numerous observations of freeway bottlenecks. One can see from the discussion above that the breakdown phenomenon in free traffic is not exclusively a property of freeway bottlenecks: it can occur spontaneously away from bottlenecks. The latter is related to the fact that these breakdown phenomena have the same nature: they are local first-order $F \rightarrow S$ transitions.

### 10.7 Some Empirical Features of Synchronized Flow

#### 10.7.1 Complex Behavior in Flow–Density Plane

It has been found that empirical synchronized flow has different properties in contrast to free traffic flow. Let us first note that in free flow in accordance
with a huge number of empirical results (e.g., [20,21]) an increase in the flow rate is accompanied, as a rule, by an increase in density and by corresponding decrease in average speed (Fig. 10.14a).

It is well-known that data for congested traffic shows a wide spread of points in the flow–density plane (see e.g., Fig. 2.4a [21,30]). If all wide moving jams are omitted in empirical data for congested traffic we find that remaining empirical data for the “synchronized flow” phase also exhibits a wide spread of points in the flow–density plane (Fig. 10.14b).

In contrast to free flow, in synchronized flow an increase in flow rate can be accompanied by either an increase or a decrease in density (Fig. 10.14b). Correspondingly, the average speed of vehicles can either decrease or increase
when the flow rate increases. The measurement points can perform random transference in all directions with either positive or negative slope in the flow–density plane. This transference of the measurement points for synchronized flow covers a two-dimensional region in the flow–density plane (Fig. 10.14b). The latter circumstance can distinctly be seen in Fig. 10.14c, where both free flow (black points in Fig. 10.14c) and synchronized flow (circles in Fig. 10.14c) over longer time intervals are presented. In Fig. 10.14c all points related to wide moving jams are omitted.

10.7.2 Three Types of Synchronized Flow

There can be three different types of states of synchronized flow [203]:

(1) Approximately both time-independent and spatially homogeneous states, where both the average speed and flow rate do not change appreciably during a time interval of about 2–5 min (e.g., points 1–3 in Fig. 10.14b). These states are called “homogeneous states of synchronized flow.”

(2) States in which the average speed is approximately constant during a relatively long time interval, but the flow rate and density change appreciably during this time interval. It can be assumed that waves of flow rate and density can propagate with positive velocity in such synchronized flow. The flow rate does not necessarily change synchronously in different lanes of a freeway. It can be assumed additionally that different density waves propagate in different lanes. These states of synchronized flow are called “homogeneous-in-speed” states. An example of synchronized flow that is approximately related to homogeneous-in-speed states is shown in Fig. 10.15. Homogeneous-in-speed states of synchronized flow in the flow–density plane are related to a line whose slope equals the vehicle speed in synchronized flow (dotted line in Fig. 10.15c).

![Fig. 10.15. Example of homogeneous-in-speed states of synchronized flow. Flow rate (a) and vehicle speed (b). (c) Free flow (black points) and synchronized flow (circles) in the flow–density plane for the left lane at the detectors D6 on the freeway A5-North (Fig. 2.2)). Taken from [213]
Fig. 10.16. Example of nonstationary and inhomogeneous states of synchronized flow. (a, b) Flow rate. (c, d) Vehicle speed. (e, f) Free flow (black points) and synchronized flow (circles) in the flow–density plane for the left lane at the detectors D6 (e) and D7 (f) on the freeway A5-North. Taken from [213]

(3) Essentially nonstationary and inhomogeneous states of synchronized flow, in which both the average vehicle speed and flow rate abruptly change from one empirical point to the next (Fig. 10.16).

Note that in the flow–density plane each of these three types of states of synchronized traffic flow covers a two-dimensional region.

Even though in free flow only small-amplitude waves with positive velocity can propagate, the dynamics of synchronized flow can be very complex: small-amplitude waves with positive and negative velocity are both possible. Waves with positive velocity are more probable in synchronized flow states at higher speeds. In synchronized flow states at lower speeds, small-amplitude waves with negative velocity are more likely to propagate.

In homogeneous-in-speed states of synchronized flow, changes in flow rate and density are correlated in approximately the same way as those in free flow. In Fig. 10.15 the related correlation coefficient is $K \approx 0.9$. In contrast, in inhomogeneous and nonstationary states of synchronized flow there is almost no correlation between variations in flow rate and density: an increase in flow
rate can be accompanied either by an increase or a decrease in density [203]. In Fig. 10.16e the correlation coefficient between flow rate and density is $K \approx -0.11$.

Fig. 10.17. Fluctuations in free and synchronized flow. (a, b) Example of homogeneous-in-speed-states of synchronized flow (between D19 and D16 on the freeway A5-South, Fig. 2.1). (c) Speed variations at D19 for synchronized flow (left) and free flow (right). Taken from [213]

The speed variance associated with fluctuations in synchronized flow can be considerably lower than in free flow (Fig. 10.17c). Moreover, in contrast to free flow, the speed variance in different lanes in synchronized flow can be approximately the same. Both results can be related to a bunching of vehicles in synchronized flow. Indeed, the bunching of vehicles, i.e., a building of vehicle platoons on the road, restricts the speed variance between neighboring vehicles. On the other hand, such a bunching of vehicles should lead to long-range spatial correlation of vehicle speeds in synchronized flow. The latter has indeed been found in a study of single vehicle data by Neubert et al. [195,196] (Fig. 10.18).
10.7 Synchronized Flow

The average vehicle speed in the “synchronized flow” phase can be synchronized across different freeway lanes. However, this is often the case only away from bottlenecks. In particular, empirical observations show that in non-stationary and inhomogeneous states of synchronized flow the vehicle speed is not necessarily synchronized over different freeway lanes at all locations and over time. This is often the case inside the pinch region of an GP when the average vehicle speed in synchronized flow of the pinch region is very low. However, there is a tendency towards speed synchronization in all states of synchronized flow. This tendency is the reason for the term “synchronized” in the “synchronized flow” phase (Sect. 4.2.2).

10.7.3 Overlapping States of Free Flow and Synchronized Flow in Density

To show the effect of the overlapping states of free flow and synchronized flow in density, let us consider the example of the WSP shown in Fig. 9.5. When free flow (black squares in Fig. 10.19) and synchronized flow within the WSP (circles in Fig. 10.19) are shown in the flow–density-plane, it can be seen that at least at the detectors D20–D18, states of synchronized flow partially overlap free flow in density. The same conclusion can be drawn if the vehicle speed as a function of density is plotted (Fig. 10.20a,b). This means
Fig. 10.19. Measurement points in the flow–density plane for the WSP shown in Fig. 9.5 at the detectors D20–D17 (left line). Free traffic: black squares; synchronized flow: circles. Overlapping states of free flow and synchronized flow at the detectors D19 is in the density range from about 18 to 36 vehicles/km. 1-min average data. Taken from [221]

that at the same density either a state of synchronized flow or a state of free flow is possible.

If now the average absolute values of the vehicle speed difference between the left lane and the middle lane $\Delta v$ for free flow (curve $F$ in Fig. 10.20c,d) and for synchronized flow (curve $S$ in Fig. 10.20c,d) are shown, then we see that there is overlap of the speed difference in density. This overlapping can lead to a hysteresis between the free flow states and the synchronized flow states. It can be assumed that this overlap is related to Z-shaped form of the dependence of $\Delta v$ on density.\textsuperscript{5}

\textsuperscript{5} The assumption that the overlap of the speed difference $\Delta v$ in density should be related to a Z-shaped characteristic $\Delta v(\rho)$ is made by analogy to a huge number of physical, chemical, and biological systems in which states of two different phases overlap in a system parameter. Naturally, there is usually a hysteresis effect due to this overlap. In the theory of these spatially distributed systems, depending on the form of this overlap it is usually assumed that there is one of the N-, S- or Z-shaped characteristics of the system [335–337, 339]. The middle branch of these characteristics cannot usually be observed in experiments. This is because this branch should be related to unstable states of the system [335–337, 339]. The assumption under consideration can be considered correct if the theory predicts features of spatiotemporal patterns, which are
Fig. 10.20. Empirical features of synchronized flow at the detectors D19 (a, c, e, g) and D18 (b, d, f, h) in the WSP shown in Figs. 9.5b and 9.6. (a, b) Measurement points in the speed–density plane (free flow: black squares, synchronized flow: circles; left lane; 1-min average data), (c, d) Average difference in speeds between the left and middle freeway lanes as a function of density; curve $F$: free flow, curve $S$: synchronized flow. (e–h) Distribution of the number of vehicles as a function of different speed classes associated with individual single vehicle data for synchronized flow (e, f) and for free flow (g, h). Curve 1: vehicles in the left lane, 2: vehicles in the middle lane, 3: vehicles in the right lane, 4: long vehicles in the right lane, 5: long vehicles in the middle lane. In (e–h) measured single vehicle data is shown whereby the number of vehicles in each of 15 different classes with regard to the vehicle speed is used separately for vehicles and for long vehicles. Taken from [221]
This result is in agreement with the conclusion about the Z-shape of the probability for passing (see Sect. 5.2.5). The physical meaning of the result in Fig. 10.20c,d is as follows. In free flow the difference in average speed on German freeways between the left (passing) freeway lane and the middle lane due to high probability of passing is considerably higher than that in synchronized flow. However, at the same density in a limited range (e.g., at D19 from 18 vehicles/km to 26 vehicles/km) either states of free flow or synchronized flow can exist. This leads to a nearly Z-form of the dependency of $\Delta v$ on density. The lower the vehicle speed in synchronized flow, the less the density range of the overlap of the curves $F$ and $S$ in Fig. 10.20d. This overlap disappears fully if the vehicle speed in synchronized flow decreases further.

10.7.4 Analysis of Individual Vehicle Speeds

To clarify the difference between free flow and synchronized flow and features of synchronized flow, distributions of the number of vehicles as a function of the individual vehicle speed for synchronized flow (Fig. 10.20e,f) and for free flow (Fig. 10.20g,h) are shown.

Firstly, it can be seen that in synchronized flow the mean vehicle speed of vehicles and long vehicles are almost the same for different freeway lanes whereas for free flow these mean values are strongly shifted to one another.

Secondly, we see that at the detectors D19 individual vehicle speeds in synchronized flow were not lower than 40 km/h (Fig. 10.20e); there are no narrow moving jams in synchronized flow between D19 and D18. Nevertheless, these states of synchronized flow cover 2D regions in the flow–density plane (Fig. 10.19, D19-left, D18-left).

10.8 Conclusions

(i) The breakdown phenomenon (speed breakdown) is related to an F→S transition.

(ii) This speed breakdown effect in free traffic is not exclusively a property of freeway bottlenecks. There are two types of F→S transitions in empirically observed traffic flow:

1. a local F→S transition away from bottlenecks;
2. a local F→S transition in the vicinity of a freeway bottleneck.

observed in experiments [335–337,339]. The three-phase traffic theory (where the Z-shaped traffic flow characteristic is used [221,330]) predicts the basic features of empirical spatiotemporal congested traffic patterns [329–331]. For this reason, we will use the assumption about the Z-shaped traffic flow characteristic made above in the further consideration.
After the $F \rightarrow S$ transition (2) has occurred, synchronized flow can be self-sustaining in the vicinity of the bottleneck for several hours. In contrast, after the $F \rightarrow S$ transition (1) has occurred, synchronized flow usually exists for a relatively short time interval and the entire region of this synchronized flow can propagate both upstream and downstream from the location of the phase transition. In the flow-density plane, both $F \rightarrow S$ transitions show qualitatively the same speed breakdown effect in free flow, which is well-known from numerous observations of freeway bottlenecks.

(iii) The duration of an $F \rightarrow S$ transition is not appreciably higher than about 1 min.

(iv) Whereas during an $F \rightarrow S$ transition there is an abrupt speed breakdown, there should not necessarily be a decrease in flow rate.

(v) There are spontaneous and induced local first-order $F \rightarrow S$ transitions at isolated bottlenecks.

(vi) Induced $F \rightarrow S$ transitions at an isolated bottleneck can be caused by either wide moving jam propagation through the bottleneck or the upstream propagation of a region of synchronized flow that initially occurred downstream of the bottleneck.

(vii) In the latter case, this region of synchronized flow is caught at the bottleneck (catch effect).

(viii) An important attribute of empirical freeway capacity of free flow at a bottleneck is the probability $P_{FS}^{(B)}$ for a spontaneous $F \rightarrow S$ transition at the bottleneck. The maximum freeway capacity of free flow at the bottleneck is found from the condition that $P_{FS}^{(B)} = 1$. The maximum freeway capacity is a function of the averaging time interval $T_{av}$ for the flow rate. If a bottleneck is due to the on-ramp, then the maximum freeway capacity is also a function of the flow rates $q_{on}$ and $q_{in}$.

(ix) The maximum freeway capacity, the flow rate at the empirical limit (maximum) point of free flow $q_{\text{emp}}^{(\text{free, emp})}$, and the pre-discharge flow rate $q_{FS}^{(B)}$ can in general take on very different values. In particular, if the empirical limit (maximum) point of free flow is related to a spontaneous $F \rightarrow S$ transition at a bottleneck, the measured flow rate $q_{\text{max, emp}}^{(\text{free, emp})}$ at this limit point will not necessarily be related to the maximum freeway capacity. This is because this point can be related to the flow rate where the $F \rightarrow S$ transition occurs with the probability $P_{FS}^{(B)} < 1$. This means that in the latter case this flow rate is lower than the maximum freeway capacity.

(x) The conclusion of item (ix) can also be drawn for the pre-discharge flow rate $q_{FS}^{(B)}$, the pre-discharge flow rate can be lower than the maximum freeway capacity.
(xi) If the pre-discharge flow rate $q_{FS}^{(B)}$ is lower than the maximum freeway capacity, this flow rate can also be considerably lower than the flow rate at the empirical limit (maximum) point of free flow $q_{\text{max, emp}}^{(\text{free, emp})}$.

(xii) Statistical features of free flow and synchronized flow can be very different: speed variance in synchronized flow can be lower than in free flow. In contrast to free flow, there is long-range spatial speed correlation in synchronized flow.

(xiii) There are three types of synchronized flow: (1) homogeneous states of synchronized flow, (2) homogeneous-in-speed states of synchronized flow, and (3) nonstationary and inhomogeneous states of synchronized flow. The whole multitude of each of these states covers a two-dimensional region in the flow–density plane. Only a part of this multitude is usually observed in a single measurement.

(xiv) At higher speeds in synchronized flow, states of free flow and synchronized flow overlap in density in the flow–density plane.
11 Empirical Features of Wide Moving Jam Propagation

11.1 Introduction

The characteristic feature of wide moving jams to propagate through other states of traffic flow and through any bottlenecks while maintaining the mean velocity of the downstream jam front is the main criterion for the “wide moving jam” phase in congested traffic (Sect. 4.2).

In this chapter, we will consider this criterion and its consequences for other features of spatiotemporal congested patterns in more detail [166, 203, 205, 218]. In particular, an empirical determination of the line $J$ in different conditions will be studied. The other aim is a discussion of other possible characteristic parameters of wide moving jams. A dependence of characteristic jam parameters on weather and other road conditions (control parameters of traffic) is the next aim of this empirical analysis.

When a wide moving jam, which has initially appeared downstream of a freeway bottleneck, propagates through the bottleneck an interesting effect can occur. This effect is realized if a different congested pattern has already emerged or is emerging at this bottleneck. In this case, when the wide moving jam propagates through this congested pattern the wide moving jam is a foreign wide moving jam for the pattern where other wide moving jams can also emerge. This foreign wide moving jam propagation and the effect of foreign wide moving jams on pattern features will be another aim of the empirical analysis.

11.2 Characteristic Parameters of Wide Moving Jams

When control parameters of traffic (weather, road conditions, number of freeway lanes, etc.) are given, there are some characteristic parameters of wide moving jams that do not depend on the initial conditions in traffic. In particular, if in the outflow from a wide moving jam free flow is formed we find the following [166, 169]:

(i) Downstream fronts of different wide moving jams are approximately the same steady moving structure. This means that the mean values of
parameters of the downstream front of wide moving jams are approximately the same for different wide moving jams. These parameters are the velocity of the downstream wide moving jam front $v_g$, the flow rate $q_{\text{out}}$, average speed $v_{\text{max}}$, vehicle density $\rho_{\text{min}}$ in the outflow from the jam, and the vehicle density within the jam $\rho_{\text{max}}$. The mean values of these characteristic parameters remain constant over time.

(ii) The maximum possible flow rate in free flow $q_{\text{free, emp}}^{(\text{free, emp})}$ can be considerably higher than the flow rate $q_{\text{out}}$.

(iii) The development of moving jams whose widths monotonically increase over time shows a tendency towards the self-organization of the downstream front of the jams to a steady moving structure related to the downstream front of a wide moving jam.

(iv) A stable localized complex structure consisting of a few wide moving jams can propagate on a freeway (e.g., Fig. 2.7a).

### 11.2.1 Empirical Determination of Line $J$

Thus, a wide moving jam exhibits characteristic (i.e., unique, reproducible, and predictable) parameters. These parameters depend only on the control parameters of traffic such as freeway infrastructure, weather, and other environmental conditions. One of these parameters is the mean velocity of the downstream front of the wide moving jam $v_g$. It is important that this velocity remains the characteristic parameter independent of states of flow in the outflow of the wide moving jam. The related steady movement of the downstream front of a wide moving jam can be represented in the flow–density plane by the line $J$ whose slope equals $v_g$.

To determine the line $J$ with empirical data, the following procedure has been used [166,203]. When a wide moving jam passes a detector, the time series of vehicle speed and traffic flow rate are for example like in Fig. 2.7c,d for two wide moving jams. Similar measurements of the time series of vehicle speed can be performed at different detectors due to jam propagation through the freeway (Fig. 11.1). Using the distances between the different detectors (Fig. 2.2) it is easy to calculate the mean velocity of the downstream front of the wide moving jam $v_g$, to within the accuracy of measurements (one-minute intervals).

For the example shown in Fig. 2.7, the mean velocity $v_g \approx -15$ km/h for each of two wide moving jams. For some other examples of wide moving jams results of such measurements of the downstream jam velocity $v_g$ for wide moving jams observed on different days are shown in Table 11.1.

The line $J$ is defined through the coordinates $(\rho_{\text{min}}, q_{\text{out}})$ with a line slope that is equal to the mean value of the velocity of the downstream front $v_g$. Thus, besides the mean velocity $v_g$, the flow rate and vehicle speed downstream of a wide moving jam must be measured. If in the outflow from the jam free flow is formed, one average point of this free flow rate $q_{\text{out}}$ and
11.2 Characteristic Parameters of Wide Moving Jams

Fig. 11.1. Flow rate (per lane) and vehicle speed averaged across all freeway lanes at different detectors related to Fig. 2.7a,b. Taken from [169]

Table 11.1. Parameters of wide moving jams on three different days on a section of the freeway A5-South

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_g$ [km/h]</td>
<td>-16</td>
<td>-15</td>
<td>-14</td>
</tr>
<tr>
<td>$q_{out, 1}$ [vehicles/h, left lane]</td>
<td>2000</td>
<td>1800</td>
<td>1650</td>
</tr>
<tr>
<td>$q_{out}$ [vehicles/h, total flow rate]</td>
<td>5000</td>
<td>4500</td>
<td>4200</td>
</tr>
</tbody>
</table>
speed $v_{\text{max}}$ can be represented in the flow–density plane in the region of free flow with the coordinates $(q_{\text{out}}, \rho_{\text{min}})$, where $\rho_{\text{min}} = q_{\text{out}}/v_{\text{max}}$. Thus, we have these coordinates and the slope of the line $J$ from the measurements. This enables us to draw the line $J$ and to estimate the average vehicle density within the jam $\rho_{\text{max}}$ as the intersection point of the line $J$ and the density axis (x-axis) (Fig. 2.8).

The dependence of the characteristic parameters $q_{\text{out}}$, $\rho_{\text{min}}$ associated with free flow in the wide moving jam outflow and the jam density $\rho_{\text{max}}$ on different vehicle parameters can be seen from Fig. 11.2 where the line $J$ for each of the freeway lanes is shown. In this case shown in Fig. 11.2, there are no long vehicles (trucks) in the left (passing) lane, about 10% long vehicles in the middle lane, and about 40% long vehicles in the right lane. This explains the quantitative differences in the mentioned characteristic parameters of wide moving jams. It must be noted that the mean velocity of the downstream jam front $v_{g}$ is independent of the percentage of trucks. For this reason, the slope of the line $J$ in Fig. 11.2 is the same for different freeway lanes. However, traffic variables in the jam outflow are different for these lanes. The same conclusion is valid for the jam densities, which are different for the different freeway lanes. A study of other empirical examples of wide moving jam propagation shows that different vehicle characteristics usually have only a quantitative influence on the characteristic wide moving jam features rather than a qualitative one.

It should be noted that to avoid an influence of fluctuations, for the determination of the flow rate in the wide moving jam outflow the flow rate during a few minutes interval after the jam has passed the detectors is usually used. However, during this averaging time interval one should take care that on the one hand vehicles accelerating from a wide moving jam are taken into account only, and on the other hand all of these vehicles are taken into account. In other words, the detectors where the jam outflow rate is measured should be far enough away from any on- and off-ramps. It must also be noted that the line $J$ is related to the downstream jam front only and that only the flow rate in the outflow of the downstream front and the velocity of this front are used for the definition and for the determination of the line $J$, rather than that traffic states within the wide moving jam necessarily fall on the line $J$.\(^1\)

We have already mentioned that the empirical maximum (limit) flow rate in free flow $q_{\text{max}}^{\text{(free, emp)}}$ can be considerably higher than the flow rate $q_{\text{out}}$:

$$q_{\text{max}}^{\text{(free, emp)}} > q_{\text{out}} \cdot (11.1)$$

It has been found that for the flow rate $q_{\text{max}}^{\text{(free, emp)}}$ averaging over 1 min intervals ($T_{\text{av}} = 1 \text{ min}$) \([166]\)

\(^1\) Note that a discussion of a relationship between the line $J$ and empirical fundamental diagrams appears in Sects. 15.1.3 and 15.4.
The relation (11.2) for $T_{av} = 1\ \text{min}$ holds on average for different control parameters (weather, different infrastructure of freeways, etc.) [169]. The empirical maximum (limit) flow rate in free flow $q_{\text{max}}^{(\text{free, emp})}$ usually decreases when the averaging time interval $T_{av}$ for the flow rate increases (Sect. 10.3.2). For this reason, although the condition (11.1) usually holds even for $T_{av} = 10\ \text{min}$, the relation $q_{\text{max}}^{(\text{free, emp})}/q_{\text{out}}$ can be appreciably lower than 1.5 if the value $T_{av}$ increases.
It should be noted that in the example shown in Fig. 11.2 the condition (11.1) is satisfied for each of the freeway lanes separately:
\[
q_{\text{max}, \, l}^{(\text{free, emp})} > q_{\text{out}, \, l}, \quad q_{\text{max}, \, m}^{(\text{free, emp})} > q_{\text{out}, \, m}, \quad q_{\text{max}, \, r}^{(\text{free, emp})} > q_{\text{out}, \, r}.
\] (11.3)

However, it has also been observed that on some other days in the right freeway lane the maximum flow rate in free flow is not appreciably higher than the flow rate in the wide moving jam outflow \(q_{\text{out}, \, r}\).

Let us estimate the mean gross time gap between vehicles at the maximum point of free flow for each of the freeway lanes \(\tau_{\text{min}, \, i}^{(\text{free})}\), \(i = l, \, m, \, r\) and the mean time delay in vehicle acceleration at the downstream front of a wide moving jam \(\tau_{\text{del}, \, i}^{(a)}\), \(i = l, \, m, \, r\) (see Sect. 3.2.6) where index \(i = l\) is for the left lane, index \(i = m\) is for the middle lane, and index \(i = r\) is for the right lane. For the example shown in Fig. 11.2 from (3.24) we obtain
\[
\tau_{\text{min}, \, l}^{(\text{free})} \approx 1.3 \, [s], \quad \tau_{\text{min}, \, m}^{(\text{free})} \approx 1.4 \, [s], \quad \tau_{\text{min}, \, r}^{(\text{free})} \approx 2.2 \, [s].
\] (11.4)

For the estimation of the mean time delay we use (3.10) where the densities within the wide moving jam \(\rho_{\text{max}, \, i}\), \(i = l, \, m, \, r\) are calculated through the use of the lines \(J\) shown in Figs. 11.2a–c:
\[
\tau_{\text{del}, \, l}^{(a)} \approx 1.7 \, [s], \quad \tau_{\text{del}, \, m}^{(a)} \approx 1.8 \, [s], \quad \tau_{\text{del}, \, r}^{(a)} \approx 2.8 \, [s].
\] (11.5)

We can see from (11.4) and (11.5) that the condition (3.25) is satisfied in this example for each of the freeway lanes.

### 11.2.2 Dependence of Characteristic Jam Parameters on Traffic Conditions

Some other examples of the line \(J\) are shown in Figs. 11.3–11.5. In Fig. 11.3, the mean velocity of the downstream jam front \(v_{g}\) is about \(-15\) km/h as it was the case for two wide moving jams in Fig. 2.7. However, on other days one can find different values of the characteristic velocity \(v_{g}\) and the flow rate in the wide moving jam outflow \(q_{\text{out}}\) (Table 11.1). The relatively low flow rate \(q_{\text{out}}\) and the absolute value of the characteristic velocity \(|v_{g}|\) on March 23, 1998 is probably related to weather conditions; there was an intense snow fall on that day.

Two other examples in which the absolute value of the characteristic velocity \(|v_{g}|\) was appreciably lower than \(15\) km/h is shown in Figs. 11.6 and 11.7. In the first case, it was raining on the day of wide moving jam propagation. In the second case, it was a weekend: it can be assumed that drivers were in no hurry on that day. Therefore, they accelerate from wide moving jams with higher mean delay time \(\tau_{\text{del}}^{(a)}\). Corresponding to (3.10), this can explain the lower value \(|v_{g}|\).
11.2 Characteristic Parameters of Wide Moving Jams

**Fig. 11.3.** Wide moving jam propagation on January 13, 1997. (a, b) Vehicle speed (a) and flow rate (b) in a wide moving jam at D11 on a section of the freeway A5-South (Fig. 2.1). (c, d) Free flow (black points) and the line \( J \) in the flow–density plane for data related to the left lane (c) and for data averaged across all three freeway lanes (d). Taken from [218].

### 11.2.3 Propagation of Wide Moving Jams Through Synchronized Flow

The flow rate in the wide moving jam outflow \( q_{out} \), density \( \rho_{min} \), and speed \( v_{\text{max}} \) are characteristic parameters of traffic flow as long as no hindrances to traffic flow form, specifically if free flow is formed in the jam outflow [166]. The latter condition is not satisfied when a wide moving jam either reaches the location of synchronized flow or synchronized flow is formed in the wide moving jam outflow. In this case shown in Figs. 11.4c,d, 11.5c,d, and 11.8, instead of free flow formation, synchronized flow occurs in the jam outflow. We find that the flow rate, density, and average vehicle speed in the jam outflow do not maintain the property to be characteristic parameters [169].

If synchronized flow is formed downstream of a wide moving jam the average flow rate in this flow \( q_{\text{out}}^{(\text{syn})} \) is lower than \( q_{out} \):

\[
q_{\text{out}}^{(\text{syn})} < q_{out}.
\]  
(11.6)
Fig. 11.4. Wide moving jam propagation on April 15, 1996. (a, b) Vehicle speed (a) and flow rate (b) in a wide moving jam at D1 on a section of the freeway A5-South (Fig. 2.1). (c, d) Free flow (back points) and the line $J$ in the flow–density plane for data related to the left lane (c) and for data averaged across all three freeway lanes (d). Taken from [218]

The density in the jam outflow $\rho_{\text{min}, J}^{(\text{syn})}$ can be considerably higher than the characteristic density $\rho_{\text{min}}$:

$$\rho_{\text{min}, J}^{(\text{syn})} > \rho_{\text{min}}.$$

However, we find that the mean velocity of the downstream wide moving jam front $v_g$ maintains the property to be a characteristic parameter. This is related to the definition of the “wide moving jam” phase (Sect. 4.2).

It is found that a point in the flow–density plane, which is related to the flow rate $q_{\text{out}}^{(\text{syn})}$, lies (to within the accuracy of measurements) on the line $J$ (Figs. 11.8 and 11.4c). Therefore, the line $J$ can be approximately found if the flow rate $q_{\text{out}}^{(\text{syn})}$ and speed in synchronized flow $v_{\text{max}, J}^{(\text{syn})}$ averaged during 5–10 min are taken for calculation of the density $\rho_{\text{min}, J}^{(\text{syn})}$:

$$\rho_{\text{min}, J}^{(\text{syn})} = \frac{q_{\text{out}}^{(\text{syn})}}{v_{\text{max}, J}^{(\text{syn})}}.$$

This gives the left coordinates of the line $J$, $(\rho_{\text{min}, J}^{(\text{syn})}, q_{\text{out}}^{(\text{syn})})$. 
These features of wide moving jam propagation through synchronized flow enable us to estimate to within a good accuracy characteristic parameters of wide moving jams. In particular, the flow rate $q_{\text{out}}$ can be estimated at the point in which the line $J$ crosses the region of free flow in the flow-density plane (Fig. 11.4c,d). Analogous results for the other days are shown in Figs. 11.3 and 11.5.

Thus, when a wide moving jam propagates through synchronized flow, the jam nearly maintains the same mean velocity $v_g$ of the downstream front of the jam as when the jam propagates through free flow [205]. This result confirms the formula (3.10), because the parameters $P_{\text{max}}$ and $\tau_{\text{del}}^{(a)}$ obviously do not depend on a state of flow far downstream of the jam.

The flow rate $q_{\text{out}}^{(\text{syn})}$ and the density $\rho_{\text{min}, J}^{(\text{syn})}$ can appreciably change over time. Therefore, the left coordinates of the line $J$ associated with the point $(\rho_{\text{min}, J}^{(\text{syn})}, q_{\text{out}}^{(\text{syn})})$ (Fig. 11.8) can perform complicated slides along the line $J$. 

---

**Fig. 11.5.** Wide moving jam propagation on March 23, 1998. (a, b) Vehicle speed (a) and flow rate (b) in a sequence of wide moving jams at D1 on a section of the freeway A5-South (Fig. 2.1). (c, d) Free flow (back points) and the line $J$ in the flow-density plane for the jam marked by arrow 3 in (a, b) for data related to the left lane (c) and for data averaged across all three freeway lanes (d). Taken from [218]
11.2.4 Moving Blanks Within Wide Moving Jams

Considering characteristics of wide moving jams, we did not take into account that within wide moving jams there can be a complicated spatiotemporal structure of flow rate \[166\]. For a fixed observer, the flow rate can change from zero (a standstill) to some values, then become zero once more, and so on during wide moving jam propagation. Sometimes the flow rate within a wide
Fig. 11.7. Characteristic parameters of wide moving jams on a weekend. (a) Layout of a section of the freeway A44-West in Germany. (b) Flow rate (left) and speed (right) at three detectors on September 29, 1996. (c) Free flow (black squares) and the line J in the flow–density plane. Data is averaged across all lanes of the freeway. Taken from [169]
moving jam increases up to 400–800 vehicles/h per lane (see also Fig. 11.3b). For example, the flow rate within the first of two wide moving jams shown in Fig. 2.7 increases up to 600 vehicles/h per lane. A form of this changing in the flow rate within the wide moving jam, corresponding to empirical data from different sets of detectors, does not seem to be periodical with regard to time and space.

Apparently this effect can be explained as follows. When drivers suddenly met a wide moving jam, they have to slow down very sharply up to a standstill. As a result, the “blanks” between vehicles within the jam can be very different from one pair of vehicles to the others. After some time, some of the drivers within the wide moving jam can begin to reduce the largest “blanks.” As a result, new blanks appear upstream covering the latter blanks, and so on. In this case, the moving upstream blanks between vehicles within the wide moving jam can be created. This can explain the complicated character of flow rate within wide moving jams.

11.3 Features of Foreign Wide Moving Jams

A real freeway has many effectual adjacent bottlenecks. Thus, two or more GPs where the related different sequences of wide moving jams emerge can appear almost simultaneously.

Wide moving jams from downstream sequences of wide moving jams are called foreign wide moving jams (the jams marked by down-arrows “A” and “B” in Figs. 2.1 and 11.9) when they propagate through an upstream GP (an GP at D6) where other narrow moving jams are just emerging (the
11.3 Features of Foreign Wide Moving Jams

A5-South, June 18, 1997

--- left lane -- middle lane --- right lane

Fig. 11.9. Foreign wide moving jams. Propagation of the foreign wide moving jams “A” and “B” through an GP that has been formed due to an induced F→S transition at 07:15 (up-arrow, D6) induced by synchronized flow (up-arrow, D7) at the on-ramp D6 on a section of the freeway A5-South (Fig. 2.1). Vehicle speed at different detectors. Taken from [218]

jams marked by down-arrows 1–9 in Fig. 11.9). We already briefly considered foreign wide moving jams in Sect. 2.4.9.

Upstream of the foreign wide moving jam “A,” narrow jam emergence is not influenced by this foreign wide moving jam. Downstream of the foreign wide moving jam “A,” the narrow moving jams marked by arrows 6 and 7 disappear. This jam suppression effect is apparently the same as the suppression of narrow moving jams by a wide moving jam at the upstream boundary of the pinch region in an GP [208]. The latter effect will be considered in Sect. 12.2. The physics of the jam suppression effect has been discussed in Sect. 7.6.
However, far downstream of the jam “A” the features of jam emergence are not influenced by foreign wide moving jam propagation. Thus, the narrow moving jam marked by arrow 8, which is far downstream of the foreign wide moving jam “A,” grows and leads to the formation of the wide moving jam 8 at D2 (Fig. 11.9). All mentioned features of the propagation of the foreign wide moving jam “A” are characteristic for all other foreign wide moving jams, in particular for the foreign wide moving jam “B.”

As a result of these effects, instead of initial isolated sequences of wide moving jams (a sequence of the foreign wide moving jams “A” and “B” and a sequence of the moving jams 1–8) an “united” sequence of wide moving jams is finally formed (the jams marked by down-arrows 1, 3, “A,” 8, “B” in Fig. 11.9; D2, D1).

11.4 Conclusions

(i) The downstream wide moving jam front propagates steadily through different states of free flow and synchronized flow, and through congested freeway bottlenecks while maintaining the mean velocity of the downstream jam front $v_g$. This characteristic velocity depends on control traffic parameters, like weather and other road conditions.

(ii) The line $J$ represents the steady propagation of wide moving jams with the characteristic velocity $v_g$ in the flow–density plane. For the empirical determination of the line $J$ only this characteristic velocity and measurement of traffic variables in the wide moving jam outflow are necessary. These measurements give the left coordinates of the line $J$ in the flow–density plane. The right coordinates of the line $J$ are determined by the intersection of the line $J$ with the density axis: it is suggested that the average flow rate within the jam is zero.

(iii) When free flow is formed in the wide moving jam outflow, the flow rate, density, and average vehicle speed in this jam outflow are also characteristic parameters that do not depend on initial conditions. These characteristic parameters depend on traffic control parameters.

(iv) When synchronized flow is formed in the wide moving jam outflow or a wide moving jam propagates through complex states of synchronized flow, the flow rate, density, and speed in the jam outflow are not characteristic parameters. These traffic variables can change over time considerably. However, if values of traffic variables in the jam outflow are averaged over some minutes, a point in the flow–density plane related to this jam outflow lies on the line $J$.

(v) The left coordinates of line $J$ related to the wide moving jam outflow can perform complicated slides along the line $J$ over time. This occurs when in the wide moving jam outflow complicated states of synchronized flow are formed or a wide moving jam propagates through synchronized flow where traffic variables change in space and time.
(vi) A wide moving jam that has emerged in an GP downstream of an effectual freeway bottleneck propagates through this bottleneck while maintaining the downstream jam velocity $v_g$. This is true even, if another GP is forming upstream of the bottleneck. In this case, the wide moving jam is called foreign wide moving jam.

(vii) A foreign wide moving jam propagating through the pinch region of an GP suppresses the growth of narrow moving jams in the pinch region of the GP, which are very close to the downstream front of the foreign wide moving jam.

(viii) As a result of foreign wide moving jam propagation through an GP, an united sequence of wide moving jams is finally formed. The united sequence consists of wide moving jams that have emerged in the GP and foreign wide moving jams.
12 Empirical Features of Moving Jam Emergence

12.1 Introduction

In this chapter, we will try to answer the following question: how do moving jams emerge in real traffic flow? The importance of this question can already be seen from the huge number of publications devoted to theoretical studies of moving jam emergence in freeway traffic (see references in the reviews [33, 35, 36, 38]). This is also the point where the fundamental diagram approach to traffic flow theories and modeling fails to explain and predict results of empirical observations (Sect. 3.3.2).

In three-phase traffic theory, the breakdown phenomenon in an initial free flow is related to an F→S transition rather than moving jam emergence (F→J transition) (Chap. 10). In this theory, moving jams emerge spontaneously only in synchronized flow (S→J transition). The latter is confirmed by all presently known reproducible empirical results; discussing those is the main goal of this chapter.

Firstly, we consider some empirical features of narrow moving jam emergence in the pinch region of synchronized flow upstream of an on-ramp bottleneck. A discussion of the transformation of narrow moving jams into wide moving jams (S→J transition) and peculiarities of wide moving jam emergence at bottlenecks due to off-ramps will be the additional goals [208, 218].

12.2 Pinch Effect in Synchronized Flow

Let us consider an GP of type (1) at the effectual bottleneck $B_3$ on a section of the freeway A5-South in more detail (Fig. 12.1). This GP was briefly discussed in Sect. 9.4 (Fig. 9.12).

In Sect. 10.2, it was noted that there can be many F→S transitions at an on-ramp bottleneck that do not lead to a congested pattern upstream of the bottleneck (Fig. 10.2a,b; arrows 1–3 at D6). The congested pattern can occur only if the upstream front of synchronized flow starts to propagate upstream (up-arrows S at D5 and D4 in Fig. 12.1).

An GP appears upstream of the on-ramp bottleneck if the pinch effect is realized within synchronized flow upstream of the bottleneck [208]. The
Fig. 12.1. GP upstream of the bottleneck due to the effectual on-ramp D6 (bottleneck $B_3$) on a section of the freeway A5-South (Fig. 2.1) on January 13, 1997. Vehicle speed is displayed in different freeway lanes within the GP at different detectors. Taken from [218]

pinch effect is the self-compression of synchronized flow, i.e., an increase in density and a decrease in average vehicle speed in synchronized flow.

The occurrence of this pinch effect at D5 is marked by up-arrow $P$ in Fig. 12.1 (D5). As a result of the pinch effect, the pinch region is formed in synchronized flow. Within the pinch region, the average density is relatively high and the average speed is low.

The pinch effect can occur with a time delay after an F $\rightarrow$ S transition has reached D5 (compare up-arrows $S$ and $P$ in Fig. 12.1, D5). On other days,
no time delay has been observed: synchronized flow was already compressed when synchronized flow was measured at D5 (Fig. 12.2).

12.2.1 Narrow Moving Jam Emergence

In synchronized flow in the GP in Fig. 12.1, a pinch region is formed where narrow moving jams emerge, and grow propagating upstream (D5, D4). The downstream front (boundary) of synchronized flow is located at the effective location of the bottleneck. The upstream front (boundary) of synchronized flow is determined by the location where a narrow moving jam has just transformed into a wide moving jam, i.e., where a phase transition from synchronized flow to a wide moving jam (S→J transition) has occurred. The upstream front separates synchronized flow downstream and the region of wide moving jams upstream. When a narrow moving jam occurs, the jam suppresses the further growth of a narrow moving jam, which is very close to the downstream front of this wide moving jam. Due to this effect, either the narrow moving jam dissolves or the narrow moving jam merges with the wide moving jam. As a result of the jam suppression effect, some narrow moving jams can disappear without their transformation into wide moving jams [208].

Because the transformation of different narrow moving jams into wide moving jams can occur at different locations, the upstream front of synchronized flow performs complicated spatial oscillations over time. The mean width (in the longitudinal direction) of the pinch region of synchronized flow in the GP, $L_{\text{syn}}^{(\text{pinch})}$, is limited: $L_{\text{syn}}^{(\text{pinch})} \approx 3$–4 km. The mean width of the pinch region of synchronized flow $L_{\text{syn}}^{(\text{pinch})}$ does not significantly depend on traffic demand. These empirical results confirm theoretical features of the GP of type (1) discussed in Sect. 7.5.

In the pinch region, the vehicle speed and density vary over a wide range (Fig. 12.1, D5 and Fig. 12.3a). This spread of vehicle speed and density increases in the upstream direction within the pinch region (Fig. 12.1, D4 and Fig. 12.3c). This behavior is explained by growing narrow moving jam emergence in the pinch region.

Empirical points associated with synchronized flow between narrow moving jams usually lie above the line $J$ in the flow–density plane (Fig. 12.3b,d). Narrow moving jams usually have a slightly more negative velocity $v_{\text{narrow}}$ than the characteristic velocity $v_g$ of the downstream front of wide moving jams: $|v_{\text{narrow}}| > |v_g|$. These two empirical results confirm the theoretical treatment of moving jam emergence in synchronized flow (Sect. 7.6). Recall that in this treatment metastable synchronized flow states in which moving jams can emerge should lie above the line $J$ in the flow–density plane (see explanation of Fig. 6.4 of Sect. 6.3).
Fig. 12.2. GP upstream of the bottleneck at the effectual on-ramp D6 (bottleneck $B_3$) on a section of the freeway A5-South (Fig. 2.1) on March 23, 1998. (a) Vehicle speed averaged across all lanes (left) and total flow rate across the freeway (right) in space and time. (b) Vehicle speed in different freeway lanes within the GP at different detectors. Taken from [218]
A5-South, January 13, 1997

12.2 Pinch Effect in Synchronized Flow

Limit Flow Rate in Pinch Region

After the pinch region is formed, the average flow rate $q^{(\text{pinch})}$ in the pinch region often falls abruptly (D5, Fig. 12.4, January 13, 1997). It should be noted that this decrease in $q^{(\text{pinch})}$ occurs before the first wide moving jam is formed upstream of the pinch region (D2, D1 in Fig. 12.1). Therefore, the flow rate $q^{(\text{pinch})}$ falls within the pinch region, rather than being due to a possible decrease in flow rate upstream of the pinch region. The averaging time interval $T_{av}$ for the flow rate in the pinch region $q^{(\text{pinch})}$ in Fig. 12.4 (D5) is chosen to be greater than the mean time between narrow moving jams that emerge in the pinch region (this time interval between narrow moving jams is about 5–6 min at D5).

The decrease in the average flow rate $q^{(\text{pinch})}$ has a limit $q^{(\text{pinch})}_{\text{lim}}$: after this limit has been reached, the flow rate $q^{(\text{pinch})}$ displays only minor changes of approximately 10% in the vicinity of $q^{(\text{pinch})}_{\text{lim}}$. Thus, we can write the following approximate condition:

$$q^{(\text{pinch})} = q^{(\text{pinch})}_{\text{lim}}.$$  \hspace{1cm} (12.1)

This saturation feature of the pinch effect remains even if traffic demand changes during a long time interval. However, the limit flow rate $q^{(\text{pinch})}_{\text{lim}}$ can exhibit considerable variations on different days (Table 12.1).
Fig. 12.4. Average flow rate within the pinch region $q_{\text{lim}}^{(\text{pinch})}$ (solid curves) (a), the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ (b), and flow rates to the on-ramps D5-on, D6-on, and $q_{\text{eff-on}}$ (labeled “eff-on”; see (9.1)) (c) on three different days. 10-min average data for total flow rates across the freeway. Dashed lines in (a, b) are related to the flow rate in the wide moving jam outflow $q_{\text{out}}$ when free flow is formed downstream of the jam. Taken from [218].

Table 12.1. Limit flow rate in the pinch region of GPs on three different days on a section of the freeway A5-South

<table>
<thead>
<tr>
<th>Day</th>
<th>Parameter/Characteristic</th>
<th>$q_{\text{lim}}^{(\text{pinch})}$ [vehicles/h, total flow rate]</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 15, 1996</td>
<td></td>
<td>3700</td>
</tr>
<tr>
<td>January 13, 1997</td>
<td></td>
<td>3400</td>
</tr>
<tr>
<td>March 23, 1998</td>
<td></td>
<td>2800</td>
</tr>
</tbody>
</table>

Pinch Effect in On-Ramp Lane

It has been mentioned that the effectual on-ramp of the bottleneck at D6 consists of two on-ramps: D6-on and D5-on (Fig. 2.1). The vehicle speed at the downstream on-ramp D6-on is maintained almost without change after an F→S transition has occurred at D6 (D6-on in Figs. 12.1, 12.2, and 12.5). However, in the on-ramp lane of the upstream on-ramp (D5-on), a pinch effect that is very similar to the one on the freeway (D5) is realized on January 13, 1997 and on March 23, 1998 when the flow rate to this on-ramp increases...
Fig. 12.5. GP upstream of the bottleneck at the effectual on-ramp D6 on a section of the freeway A5-South (Fig. 2.1) on April 15, 1996. (a) Vehicle speed averaged across all lanes (left) and total flow rate across the freeway (right) in space and time. (b) Vehicle speed in different freeway lanes within the GP at different detectors. Taken from [218]
(D5-on in Figs. 12.1, 12.2, and 12.4c). This pinch effect often occurs with a time delay after the F→S transition reaches D5. Probably due to this pinch effect in the on-ramp lane D5-on, as on the freeway (D5), the flow rate in the on-ramp lane D5-on varies by less than 12% in the vicinity of the limit (minimum) flow rate \(q_{\text{lim pinch}}^{(\text{on})}\) after the pinch effect occurs there (D5-on, Fig. 12.4c). Note that the limit flow rate in the on-ramp lane \(q_{\text{lim pinch}}^{(\text{pinch})}\) and the limit flow rate on the main road \(q_{\text{lim pinch}}^{(\text{pinch})}\) can differ from one another.\(^1\)

### 12.2.2 Wide Moving Jam Emergence (S→J Transition)

The successive transformation of narrow moving jams into wide moving jams at the upstream boundary of the pinch region in synchronized flow leads to the formation of a region of wide moving jams. Due to upstream wide moving jam propagation, the region of wide moving jams widens continuously upstream. Consequently, the number of wide moving jams within the region of these moving jams can increase over time. Between wide moving jams synchronized flow and free flow can be formed. When the distances between wide moving jams are small (about 1.5 km or less), as a rule, synchronized flow is realized between these wide moving jams. Flow between wide moving jams is considered a part of the region of wide moving jams [208].\(^2\)

Empirical studies confirm the obvious result that a narrow moving jam transforms into a wide moving jam when the flow rate in the narrow moving jam inflow is higher than the flow rate in the jam outflow. To explain this, note that the flow rate in the outflow from the farthest downstream wide moving jam of an GP is greater than the average flow rate in the pinch region \(q_{\text{pinch}}^{(\text{pinch})}\). However, the outflow from this wide moving jam is also the inflow into the pinch region of the GP (see the discussion of this effect in Sect. 7.6.2).

The development of every narrow moving jam, whose width monotonically increases over time, displays a tendency towards self-organization of the downstream jam front into the line \(J\) [203]. This means that average empirical data corresponding to the jam outflow approaches the line \(J\). When the jam transforms into a wide moving jam, the average empirical data associated with the wide moving jam outflow lies on the line \(J\).

---

1 When these flow rates are considered per freeway lane.

2 Probably this spontaneous wide moving jam emergence in dense synchronized flow explains also driver experiments [200] presented in TGF'01 [45] and in TGF'03. In these experiments, moving jams emerged spontaneously in very dense traffic flow on a circular road (parameters of the first experiment: 23 drivers, the road length 0.23 km; parameters of the second experiment: 25 drivers, the road length 0.25 km).
Empirical observations show that the mean velocity $v^{(up)}_g$ can be a complicated function of time during jam propagation.\(^3\) In contrast, the mean value of the characteristic velocity $v_g$ does not vary during wide moving jam propagation. For this reason, it is often observed that during some time intervals we find that $|v^{(up)}_g| > |v_g|$, i.e., the jam width increases, but during other time intervals $|v^{(up)}_g| < |v_g|$, i.e., the jam width decreases over time (see Sect. 7.6). Thus, in the flow–density plane, the slope of the line that represents the propagation of the upstream jam front can vary considerably. This slope depends on the time interval when the mean velocity $v^{(up)}_g$ is calculated.

We have already noted that after a narrow moving jam has transformed into a wide moving jam, the wide moving jam suppresses the growth of those narrow moving jams in the pinch region that are very close to the downstream front of the wide moving jam. This suppression effect occurs only if distances between narrow moving jams that emerge in the pinch region are less than some minimum characteristic distance between wide moving jams $L_{\text{min},J}$ in the region of wide moving jams of an GP. This minimum distance $L_{\text{min},J}$ is about 2.5 km in the examples of the GPs under consideration. The wide moving jam suppression effect is responsible for many other saturation and dynamic features of wide moving jam emergence, as considered below. The physics of the wide moving jam suppression effect is discussed in Sect. 7.6.

**GP on Different Days**

Let us compare features of GPs that emerged at the effectual bottleneck at D6 on the freeway A5-South (Fig. 2.1) in three different years. This enables us to study common features of GPs. The first GP from January 13, 1997 is shown in Fig. 12.1. The second and the third GP from April 15, 1996 (Fig. 12.5) and March 23, 1998 (Fig. 12.2) are very similar to the first. In these three cases, free flow conditions are realized downstream of the GPs (D7) (Figs. 12.1, 12.2, and 12.5). Therefore, every of these congested patterns can indeed be considered to be GP at an isolated bottleneck due to the effectual on-ramp (D6). Up-arrow at D6 in Figs. 12.1, 12.2b, and 12.5b marks the time $t_{FS}$ (Table 12.2) at which an F→S-transition at the bottleneck occurs, leading to GP emergence upstream (D6–D1).

To compare the initial conditions of GP formation, the time dependencies of average flow rates to on- and off-ramps, as well as flow rate to the effectual on-ramp $q_{\text{eff-on}}$ (“eff-on”) are shown in Fig. 12.6 on these three days. Up-arrows 1, 2, and 3 in Fig. 12.6 (D6-on and “eff-on”; see (9.1)) are associated with the corresponding times $t_{FS}$ (Table 12.2). The time intervals where congested patterns exist upstream of the bottleneck due to the effectual on-ramp (D6) are also marked in Fig. 12.6 (D6 and D1).

\(^3\) In the study under consideration, the mean velocity $v^{(up)}_g$ is calculated from the observation of upstream jam front propagation between a pair of adjacent detectors at different freeway locations.
Table 12.2. Parameters for GP emergence and evolution on three different days on a section of the freeway A5-South

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{FS}$</td>
<td>06:40</td>
<td>06:27</td>
<td>06:23</td>
</tr>
<tr>
<td>$q_{\text{eff-on}}$</td>
<td>1680</td>
<td>1440</td>
<td>1500</td>
</tr>
<tr>
<td>$t_{\text{eff-on}}$</td>
<td>07:09</td>
<td>07:43</td>
<td>07:10</td>
</tr>
<tr>
<td>$q_{\text{eff-on, max}}$</td>
<td>1980</td>
<td>1830</td>
<td>1800</td>
</tr>
<tr>
<td>$q_{\text{eff-on}}^{\text{(trans)}}$</td>
<td>900–600</td>
<td>570–480</td>
<td>780–600</td>
</tr>
</tbody>
</table>

Flow Rates in GP

The downstream front of synchronized flow where vehicles accelerate from synchronized flow upstream to free flow downstream is fixed at the bottleneck due to the effectual on-ramp (D6). The flow rate within this front – the discharge flow rate $q_{\text{out}}^{\text{(bottle)}}$ – does not depend on the spatial coordinate (see explanation in Sect. 7.4.3). Recall that the discharge flow rate is the flow rate in free flow downstream of the congested pattern (in this case, a GP). Thus, the discharge flow rate is measured in free flow, which is related to the GP outflow at the bottleneck. However, the discharge flow rate can appreciably depend on time (Fig. 12.6, D7). It must be noted that whereas flow rates to on-ramps are very similar on the three days, the discharge flow rates on these three days are quite different.

This is because the average flow rate $q_{\text{pinch}}^{\text{(pinch)}}$ within the pinch region (D5) is also quite different on these days, and on the other hand, the sum of $q_{\text{pinch}}^{\text{(pinch)}}$ and $q_{\text{eff-on}}$ gives approximately the average discharge flow rate at D6, $q_{\text{out}}^{\text{(bottle)}}$:

$$q_{\text{out}}^{\text{(bottle)}} \approx q_{\text{pinch}}^{\text{(pinch)}} + q_{\text{eff-on}}.$$ (12.2)

Note that the discharge flow rate must be measured at the downstream detectors D7 (Fig. 2.1) where free flow occurs. However, the travel time between D6 and D7 is less than 2 min even when synchronized flow is measured at D6. There are no on- and off-ramps between these detectors. Because there are also no moving jams at the locations of the detectors D6 and D7 during the time intervals shown in Fig. 12.6, to within the accuracy of measurements the total flow rate across the freeway (10-min average data) is the same at the locations of the detectors D6 and D7. This explains the label “D6, D7” in Fig. 12.6.
Fig. 12.6. Comparison of flow rates and percentage of long vehicles $A_{\text{long}}$ at various locations related to congested patterns observed on three different days. 10-min average data. The averaging time interval $T_{\text{av}}$ is appreciably greater than the travel time between D6 and D5 that is less than 3 min. Taken from [218]
12.2.3 Correlation of Characteristics for Pinch Region and Wide Moving Jams

The cited differences in discharge flow rate $q^{(bottle)}_{\text{out}}$ and flow rate within the pinch region $q^{(pinch)}_{\text{pinch}}$ are correlated with the differences in the flow rate in the outflow of a wide moving jam, $q_{\text{out}}$, when free flow is formed in this jam outflow.

To show this, parameters of the pinch region and of wide moving jams must be compared. First recall that within the pinch region, traffic variables of synchronized flow, which are measured during time intervals between narrow moving jams, are associated with the points in the flow–density plane that usually lie above the line $J$ (circles in Fig. 12.3b,d). Nevertheless these points are associated with flow rates that are lower than $q_{\text{out}}$. For all observed cases (Table 12.1)

$$q^{(pinch)}_{\lim} < q_{\text{out}}. \quad (12.3)$$

It turns out that $q^{(pinch)}_{\lim}$ is correlated with $q_{\text{out}}$: the higher the flow rate in the wide moving jam outflow $q_{\text{out}}$, the higher the limit flow rate in the pinch region $q^{(pinch)}_{\lim}$ (Fig. 12.4). A study of data measured on different days suggests that

$$1.2 \lesssim \frac{q_{\text{out}}}{q^{(pinch)}_{\lim}} \lesssim 1.5. \quad (12.4)$$

Recall that corresponding to (11.2), $q^{(free, emp)}_{\text{max}}/q_{\text{out}} \approx 1.5$. Thus, for GPs at on-ramps, one derives from (12.4) and (11.2) the empirical relation

$$1.8 \lesssim q^{(free, emp)}_{\text{max}}/q^{(pinch)}_{\lim} \lesssim 2.25. \quad (12.5)$$

It should be noted that the condition (11.2) is true when the maximum (limit) flow rate in free flow $q^{(free, emp)}_{\text{max}}$ is associated with the averaging time interval $T_{\text{av}} = 1 \text{ min}$ (Sect. 11.2.1). For greater values $T_{\text{av}}$, the flow rate $q^{(free, emp)}_{\text{max}}$ usually decreases. Thus, although

$$q^{(free, emp)}_{\text{max}} > q^{(pinch)}_{\lim} \quad (12.6)$$

for greater values $T_{\text{av}}$, the ratio $q^{(free, emp)}_{\text{max}}/q^{(pinch)}_{\lim}$ decreases in comparison with (12.5) when $T_{\text{av}}$ increases.

12.2.4 Frequency of Narrow Moving Jam Emergence

Let us denote the mean time between narrow moving jams in an GP by $T_{\text{J}}$. In GPs under consideration, the average flow rate in the pinch region $q^{(pinch)}_{\text{pinch}}$ is limited by the limit flow rate $q^{(pinch)}_{\lim}$, i.e., the condition (12.1) is satisfied. In this case, the mean time between narrow moving jams when the jams are just emerging in the pinch region also reaches a minimum, $T_{\text{J, lim}}$, i.e., $T_{\text{J}}$
is limited by the limit (minimum) mean time between moving jams, $T_{J, \text{lim}}$. Thus, we can write the following approximate condition for this limit case:

$$T_J = T_{J, \text{lim}}.$$  \hfill (12.7)

For the examples of GPs under consideration this minimum mean time between narrow moving jams is about 5–7 min on different days (Table 12.3, D5).

**Table 12.3.** Mean time between moving jams $T_J$ and period of speed correlation functions $T_c$ at different locations

<table>
<thead>
<tr>
<th>Day</th>
<th>Time/Period</th>
<th>Detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_J$ [min]</td>
<td>D1</td>
</tr>
<tr>
<td>April 15, 1996</td>
<td>11.2</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td>10.5</td>
<td>11</td>
</tr>
<tr>
<td>January 13, 1997</td>
<td>15.7</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>9.7</td>
<td>6.4</td>
</tr>
<tr>
<td>March 17, 1997</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>16.7</td>
</tr>
<tr>
<td>March 23, 1998</td>
<td>9.9</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>8.3</td>
<td>7.8</td>
</tr>
<tr>
<td>April 20, 1998</td>
<td>12.8</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>17.7</td>
<td>19.2</td>
</tr>
</tbody>
</table>

The minimum mean time between moving jams $T_{J, \text{lim}}$ is correlated with the minimum of the mean distance between narrow jams $\bar{R}_{\text{narrow}}$ in Fig. 12.7. This mean minimum distance is about 1.5 km. This corresponds to the mean time between narrow moving jams $T_{J, \text{lim}} = 6$ min in the pinch region at a narrow jam velocity about $-15$ km/h.

In other words, the values $T_{J, \text{lim}}$ and $\bar{R}_{\text{narrow}}$ reach a minimum when the average speed of synchronized flow in the pinch region decreases. Correspondingly, the maximum frequency of narrow moving jam emergence in the pinch region of an GP $f_{\text{narrow}}$ reaches a maximum

$$f_{\text{narrow}} = f_{\text{max}} = 1/T_{J, \text{lim}}.$$  \hfill (12.8)

The existence of the maximum frequency of narrow moving jam emergence in the pinch region of an GP as well as of the minimum of $T_{J, \text{lim}}$ and $\bar{R}_{\text{narrow}}$ is a saturation effect. When the flow rate to the on-ramp continuously increases the frequency of narrow moving jam emergence $f_{\text{narrow}}$ is limited by the maximum frequency $f_{\text{max}}$ (12.8).
This saturation effect is realized when the speed is low and the density is high in the pinch region of synchronized flow (Fig. 12.7). If due to an increase in flow rate to the on-ramp $q_{on}$ the speed in the pinch region further decreases, narrow moving jams emerge continuously on average at the minimum distances $R_{narrow}$ from one another, i.e., with the maximum frequency $f_{narrow}$.

### 12.2.5 Saturation and Dynamic Features of Pinch Effect

The pinch effect exhibits some saturation and some dynamic features. The limitation of the average flow rate $q^{(\text{pinch})}$ by the limit flow rate (12.1) is one of the saturation features of the pinch effect (Sect. 12.2.1). If 1-min data is considered, then a dynamic feature of the pinch effect is found. This feature is narrow moving jam emergence in the pinch region. These growing narrow moving jams propagate upstream.

In particular, it has been mentioned that if the mean distance between narrow moving jams in the pinch region $R_{narrow}$ is less than some minimum distance between wide moving jams (about 2.5 km in the examples under consideration), some of these narrow jams disappear during their transformation into wide moving jams. This narrow moving jam dissolution is probably associated with the jam suppression effect considered above: a wide moving jam suppresses the growth of narrow moving jams that are very close to the wide moving jam downstream front. As a result, the mean distance between wide moving jams is markedly greater than the initial distance between narrow jams in the pinch region. This result is illustrated for data on the three days under consideration in Figs. 12.1, 12.2, and 12.5, and Table 12.3. Nevertheless, even this dynamic process in the pinch region exhibits a saturation
12.2 Pinch Effect in Synchronized Flow

This feature is the saturation effect in narrow moving jam emergence discussed above.

The other saturation feature of the pinch effect is as follows. The mean time interval \( T_{\text{narrow}}^{(\text{mean})} \) between narrow moving jam emergence and an S→J transition, i.e., the transformation of a narrow moving jam into a wide moving jam, is also almost constant for different moving jams. This holds when traffic conditions (traffic control parameters) do not change. For example, on January 13, 1997 we have \( T_{\text{narrow}}^{(\text{mean})} \approx 11 \text{ min.} \)

However, these saturation features are valid only during the time interval when the average flow rate \( q^{(\text{pinch})} \) is also almost constant in the vicinity of \( q_{\text{lim}}^{(\text{pinch})} \) (12.1) (Fig. 12.4a).

The time interval \( T_{\text{narrow}}^{(\text{mean})} \) determines the mean duration of an S→J transition. Recall that the duration of an F→S transition is about 1 min (Sect. 10.2). This means that the duration of the S→J transition is about 10 times greater than the duration of the F→S transition.

The high value of \( T_{\text{narrow}}^{(\text{mean})} \) and complex dynamics of wide moving jam emergence can be responsible for the spatial dependence of the speed correlation function discussed below.

12.2.6 Spatial Dependence of Speed Correlation Function

The speed correlation function that is calculated during moving jam emergence at different freeway locations (D5–D1) is a strong function of the spatial coordinate (Fig. 12.8 and Table 12.3): the period of this function can vary in space from about 5 min to about 20 min.

The speed correlation function is calculated for time series \( v_n = v(t_n) \) where \( v \) is the vehicle speed, \( t_n = n \Delta t + t_0 \) is the time within a given time interval \( t_0 < t_n \leq t_N, \) \( n = 1, 2, \ldots, N, \) \( \Delta t = 1 \text{ min,} \) \( t_0 \) is the initial time, \( N \) is the number of points in the time interval. The correlation function \( R_{VV}(k \Delta t), k = 0, 1, 2, \ldots \) is defined to be [471]

\[
R_{VV}(k \Delta t) = \frac{1}{\sigma^2(N - k)} \sum_{n=1}^{N-k} (v_n - \langle v \rangle)(v_{n+k} - \langle v \rangle), \tag{12.9}
\]

where \( \langle v \rangle \) is the average speed over the time interval \( t_0 < t_n \leq t_N, \)

\[
\langle v \rangle = \frac{1}{N} \sum_{n=1}^{N} v_n, \quad \sigma^2 = \frac{1}{N-1} \sum_{n=1}^{N} (v_n - \langle v \rangle)^2. \tag{12.10}
\]

The maximum \( k \) in (12.9) is \( N/2. \)

The period of the speed correlation function \( T_c \) (Table 12.3) has a minimum (between 5 and 7 min on different days) in the pinch region (D5). This is correlated with the saturation effect in narrow moving jam emergence (Sect. 12.2.4). When narrow moving jams propagate upstream, their
amplitude and period can increase. In this case, the period $T_c$ increases to a maximum value (9–20 min on different days). This maximum is reached after narrow moving jams have transformed into wide moving jams.

While $T_c$ is nearly the same in the pinch region of the GPs (D5) on the days under consideration, this period is very different for wide moving jams formed on the different days (D1). Note that from the results in Table 12.3 it can be concluded that in some cases, e.g., on April 20, 1998, the mean time between moving jams $T_J$ is lower than the period of the correlation function $T_c$. This is because moving jams of lower amplitude decrease the mean time between moving jams $T_J$, but they only slightly influence the period of the speed correlation function $T_c$. 

Fig. 12.8. Spatial dependence of the speed correlation function (12.9) during moving jam emergence upstream of the bottleneck at D6 on a section of the freeway A5-South at different detectors (Fig. 2.1). Taken from [218]
12.3 Strong and Weak Congestion

12.2.7 Effect of Wide Moving Jam Emergence in Pinch Region of General Pattern

As discussed above, wide moving jams are usually formed at the distance \( L_{\text{syn}}^{(\text{pinch})} \approx 3-4 \text{ km} \) upstream of the bottleneck at D6. Now a somewhat exceptional case is discussed, when a wide moving jam appears at a very small distance upstream of the bottleneck at D6. The wide moving jam is marked by down-arrow “B” in Fig. 12.9a.

The wide moving jam “B” occurs between D6 and D5. Indeed, at D6 and at D6-on the vehicle speed does not decrease either before or after the wide moving jam is measured at D5 and D5-on (Fig. 12.9a). After the wide moving jam has occurred, the jam propagates through the GP (D5–D1, down-arrow “B”). The occurrence and propagation of the wide moving jam causes two effects.

The first effect is a return S→F transition at the bottleneck (D6, up-arrow F_1 in Fig. 12.9a. We have noted that the wide moving jam “B” occurs at the small distance upstream of the bottleneck at D6. Therefore, the flow rate at D6 during jam formation decreases (Fig. 12.9b, D6). This can explain the return S→F transition. When the jam propagates upstream of D5, the flow rate at D6 increases (it is the sum of the outflow rate from the jam “B” and \( q_{\text{eff-on}} \)) and an F→S transition at D6 occurs once more (D6, up-arrow S_2 in Fig. 12.9a,b).

The second effect is the suppression of the growth of narrow moving jams in the pinch region that are very close to the downstream front of the wide moving jam “B.” Apparently due to this effect the moving jams 1 and 2 disappear (D5–D3 in Fig. 12.9a). This suppression effect is the same as those in the vicinity of the upstream front of the pinch region in the GP (Sect. 12.2.2).

12.3 Strong and Weak Congestion

The flow rate \( q_{\text{eff-on}} \) as a function of time has a maximum point \( q_{\text{eff-on}, \text{ max}} \) at \( t = t_{\text{eff-on}} \). The time \( t = t_{\text{eff-on}} > t_{\text{FS}} \), where \( t_{\text{FS}} \) is the time of an F→S transition at the bottleneck leading to GP emergence (Table 12.2 and Fig. 12.6, eff-on). Whereas when the flow rate \( q_{\text{eff-on}}(t) \) is high, the average flow rate in the pinch region \( q_{\text{pinch}}^{(\text{pinch})} \) is self-sustaining in the vicinity of the limit (minimum) flow rate \( q_{\text{lim}}^{(\text{pinch})} \) (12.1) and the average vehicle speed \( v_{\text{av}} \) is low (the region labeled “strong” in Fig. 12.10a,b). This case is called “strong” congestion.

In the strong congestion condition, the mean time between narrow moving jams \( T_J \) reaches the lowest possible value \( T_{J, \text{ lim}} \) (12.7) and the mean width \( L_{\text{syn}}^{(\text{pinch})} \) of synchronized flow in the GP is limited. This width is independent of traffic demand. Thus, the pinch effect and the related GPs considered in Sect. 12.2 are related to the strong congestion condition. Such an GP is
Fig. 12.9. Wide moving jam emergence in the pinch region of an GP. Vehicle speed (a) and flow rate (b) at different detectors on a section of the freeway A5-South (Fig. 2.1). Taken from [218]
339

A5-South, March 23, 1998

Fig. 12.10. Transition from strong congestion to weak congestion. (a) Vehicle speed in different lanes (1-min data) where narrow moving jams are marked by down-arrows and average speed $v_{av}$ (fat solid curve, 10-min average data). (b–d) Flow rates in the pinch region (D5) (b), to the on-ramp D6-on (c), to the effectual on-ramp (labeled “eff-on”) and a part of the flow rate to the effectual on-ramp $q_{on-up} = q_{eff-on} - q_{D6-on}$ (labeled “on-up”) (d) with 10-min average data. Taken from [218]

called “a GP under the strong congestion condition.” Thus, the GP of type (1) that is shown in Figs. 2.21 and 9.12 is the GP under the strong congestion condition.

In contrast, when the flow rate $q_{eff-on}$ decreases below some value, the average speed in the pinch region $v_{av}$ begins to gradually increase. In addition, the flow rate $q^{(pinch)}$ ceases to be a self-sustaining value that is close to the limit flow rate $q^{(lim)}$ (Fig. 12.10). This case is called “weak” congestion. Accordingly, the pinch effect in the GP is related to the weak congestion condition. Such an GP is called “a GP under the weak congestion condition.”

Under the weak congestion condition, the mean time between narrow moving jams that emerge in the pinch region of an GP increases with average speed in the pinch region (Fig. 12.10a). This is correlated with Fig. 12.7 where it can be seen that the higher the speed in the pinch region away from narrow moving jams, the greater the mean initial distance between narrow moving jams $R_{narrow}$.

The distance $R_{narrow}$ can sometimes equal or exceed the minimum distance between wide moving jams $L_{min}$, (Sect. 12.2.2). In this case, every
growing narrow moving jam can lead to the occurrence of a wide moving jam. If the speed in synchronized flow is high enough then no moving jams emerge in this flow.

It will be shown below that under the weak congestion condition diverse transformations between different congested patterns can occur at freeway bottlenecks (Chap. 13).

12.4 Moving Jam Emergence in Synchronized Flow Away from Bottlenecks

In the above examples of GPs, an GP emerges due to F→S→J transitions. The first F→S transition occurs at the bottleneck (D6). The second S→J transition is realized in synchronized flow that has emerged at this bottleneck. However, the S→J transition occurs some kilometers upstream from the bottleneck (D3–D2). Thus, this S→J transition is a phase transition away from bottlenecks.

It can be expected that if synchronized flow has occurred away from a freeway bottleneck, then later the pinch region and narrow moving jam emergence in this synchronized flow can also occur. The narrow moving jam grows and leads to S→J transition. This supposition is confirmed by empirical observations.

To show this effect, we consider synchronized flow that emerged away from bottlenecks (Sect. 10.6). A wide moving jam appears spontaneously in this synchronized flow due to the sequence of F→S→J transitions:

(i) Firstly, an F→S transition occurs away from bottlenecks (D18 in Fig. 10.10). The vehicle density is relatively low in this synchronized flow (D18, Fig. 12.11a).

![image](image-url)

**Fig. 12.11.** Pinch effect in synchronized flow. Concatenation of states of free flow (*black squares*), synchronized flow (*circles*), and the line *J* at D18 (a) and D20 (b). Synchronized flow at D20 is related to higher densities than at D18. Data from March 17, 1997, the freeway A5-South (Fig. 2.1). Taken from [218]
12.4 Moving Jam Emergence Away from Bottlenecks

(ii) Later, the pinch effect in synchronized flow (a self-compression of synchronized flow) is observed. States of this dense synchronized flow lie above the line $J$ in the flow–density plane (D20, Fig. 12.11b). This pinch region is away from bottleneck. In the pinch region, a narrow moving jam emerges spontaneously. To see that the narrow moving jam has emerged at D20, we show the speed and flow rate at downstream locations in Fig. 12.12. It must be noted that the pinch region of synchronized flow where the narrow moving jam emerges (D20) is located downstream of the freeway location where synchronized flow has earlier occurred spontaneously away from bottlenecks (up-arrow at D18 in Fig. 10.10).

(iii) The narrow moving jam grows propagating upstream (down-arrows in Fig. 10.12, D20–D17). The subsequent growth of this narrow moving jam leads to wide moving jam formation (Figs. 10.12 and 12.13). We call this spatiotemporal congested pattern “a GP away from bottlenecks.” Thus, the GP away from bottlenecks consists of a synchronized flow region that has occurred away from bottlenecks and wide moving jam(s) that emerge within this synchronized flow region. This GP can be considered...
Fig. 12.13. Wide moving jam propagation (see Fig. 10.12) through the freeway bottleneck at D16 and upstream of the bottleneck. Dependence of vehicle speed \textit{(left)} and flow rate \textit{(right)} at different detectors. Down-arrows show the location of the moving jam at different detectors. Up-arrow at D16 shows an F→S transition in the vicinity of the freeway bottleneck (Fig. 2.1). Taken from [210]
a GP away from bottlenecks before the synchronized flow region of the GP reaches an upstream adjacent effectual bottleneck.

In contrast to a local F→S transition, where the flow rate does not necessarily decrease when the F→S transition occurs (D18 in Figs. 10.10 and 10.11), both the vehicle speed and flow rate have decreased simultaneously in the narrow moving jam (down-arrows left and right in Fig. 10.12, D20). Narrow moving jam propagation exhibits the process of spontaneous self-formation of a wide moving jam (down-arrows in Fig. 10.12, D19–D15): the narrow moving jam gradually increases in amplitude during jam propagation upstream (from the detectors D20 to D15, Fig. 10.12). Both the vehicle speed and flow rate in the moving jam decrease over time up to zero (D16). It should be noted that the detectors D20–D17, where the wide moving jam emerges, are located away from bottlenecks. After the wide moving jam has been formed, the jam propagates upstream while maintaining the mean velocity of the downstream jam front through complex states of traffic flow and through bottlenecks (Fig. 12.13, down-arrows).

12.5 Pattern Formation at Off-Ramp Bottlenecks

All types of patterns that occur at an isolated effectual on-ramp bottleneck can also emerge at an isolated effectual off-ramp bottleneck. However, congested patterns at the off-ramp bottleneck and moving jam emergence in an GP show some peculiarities.

The peculiarities of GPs at off-ramp bottlenecks are as follows.

(i) An F→S-transition usually occurs at some distance upstream of the off-ramp bottleneck (up-arrows S in Fig. 9.14). The downstream front of the GP is also located at some distance upstream of the bottleneck.

(ii) In all measurements up to now, GPs have been observed at off-ramp bottlenecks only under weak congestion conditions.

An example of an GP at an off-ramp bottleneck that exhibits these peculiarities can be seen in Fig. 9.14. The upstream propagation of synchronized flow (up-arrows at D22, D21) that has initially occurred upstream of the off-ramp leads to GP formation. However, in the pinch region of this GP the weak congestion condition is realized. In contrast to the GPs at the bottleneck due to the effectual on-ramp D6 discussed in Sect. 12.2 (Figs. 12.1, 12.2, and 12.5), in the GP at the off-ramp bottleneck there is almost no difference between the flow rate at D22 in free flow (the time interval before up-arrow in Fig. 12.14) and the flow rate in the pinch region of synchronized flow at D22 (the time interval after up-arrow). In the pinch region, the vehicle speed away from moving jams (about 40–60 km/h, Fig. 9.14; D22, D21) is appreciably higher than in the pinch region of the GPs under the strong congestion
condition that appear at the on-ramp bottleneck at the detectors D6 (about 20–30 km/h, Figs. 12.1, 12.2, and 12.5; D5).

The upstream front of synchronized flow in the GP at the off-ramp bottleneck widens continuously upstream (up-arrows in Fig. 9.14). Therefore, the width $L_{syn}$ of synchronized flow in the GP at the off-ramp bottleneck increases over time rather than this width is spatially limited. In contrast to the GPs at the on-ramp bottleneck discussed in Sect. 12.2 (Figs. 12.1, 12.2, and 12.5), the upstream boundary of synchronized flow in the GP at the off-ramp bottleneck in Fig. 9.14 is not determined by the location of an $S\rightarrow J$ transition (wide moving jam emergence): firstly, synchronized flow propagates continuously upstream of the bottleneck (up-arrows in Fig. 9.14) and only later wide moving jams emerge in that synchronized flow.

Thus, the pinch effect and the GP of type (2) in Figs. 2.19 and 9.14 are related to the weak congestion condition.

### 12.6 Induced F→J Transition

In this section, we show that an induced $F\rightarrow J$ transition, i.e., induced wide moving jam emergence in free flow is possible [214]. The induced $F\rightarrow J$ transition can occur when the flow rate in free flow is greater than the flow rate $q_{out}$ in the wide moving jam outflow. This confirms empirical and theoretical conclusions that free flow is in a metastable state with respect to moving jam formation at higher flow rates in free flow than $q_{out}$ [166,367] (Sect. 2.4.2). This induced effect can be observed when a high-amplitude local perturbation associated with a decrease in vehicle speed occurs in the off-ramp lane of an off-ramp bottleneck.

An example of an induced $F\rightarrow J$ transition in the vicinity of bottleneck $B_1$ (bottleneck due to the off-ramp D23-off in Fig. 2.1) is shown in Figs. 12.15...
A wide moving jam that results from this induced transition is labeled “wide moving jam” in Fig. 12.15. To understand features of this effect, we should note that a few hundred meters downstream of the detector D23-off the off-ramp lane is splitting up into two roads running into two directions (west and east) on the freeway A66 that crosses the freeway A5 at the intersection I3. Therefore, there are two detectors: D24-off-1 for the east direction, and D24-off-2 for the west direction (Fig. 2.1).

Firstly, an abrupt decrease in average vehicle speed is observed in the off-ramp lane at the detector D23-off (down-arrow 1 at \( t = 06:53 \) in Fig. 12.16a). However, the speed downstream of this detector (at the detectors D24-off-1 and D24-off-2) does not change appreciably over time (D24-off-1 and D24-off-2, Fig. 12.16a). This means that the decrease in speed in the off-ramp lane is a local effect that occurs in the vicinity of the detector D23-off. This local speed decrease can be considered a local speed perturbation in the off-ramp lane. In addition, we can see that the flow rate remains approximately the same before and after local perturbation occurrence (D23-off, Fig. 12.16b). Thus, there is an increase in vehicle density within the local perturbation.

This local perturbation propagates upstream in the off-ramp lane. As a result, the perturbation reaches the main road. This is realized at \( t = 06:56 \) (down-arrow 1 in Fig. 12.16c, D22), i.e., 3 minutes later than the time of perturbation emergence in the off-ramp lane at D23-off. This perturbation that has initially emerged in the off-ramp lane can be considered an “external” local perturbation for traffic flow on the main road. If free flow on the main road is in a metastable state with respect to wide moving jam formation and the speed within this external perturbation is low enough, i.e., the perturbation amplitude is high enough, then the perturbation grows and leads to an induced F\( \rightarrow \)J transition on the main road (Fig. 12.16c, down-arrows 1).
Fig. 12.16. Induced $F \rightarrow J$ transition. Time dependencies of vehicle speed (a, c) and flow rate (b, d) in freeway lanes at different detectors in the vicinity of the off-ramp bottleneck $B_1$ at D23-off (Sect. 9.2.1). The horizontal arrow in (d) at D19 shows the flow rate $q_{\text{out},1}$ in the wide moving jam outflow in the left lane on the main road. Taken from [214]
Features of this induced F→J transition on the main road are as follows.

(i) Upstream of the off-ramp lane the flow rate in the right lane on the main road becomes approximately as high as in the left lane (Fig. 12.16d, D22). The external local perturbation appears only in the right lane on the main road (down-arrow 1 at $t = 06:56$ in Fig. 12.16c, D22). This perturbation does not cause an F→S transition on the main road, i.e., no synchronization of speeds between different lanes occurs [214] (see explanation in footnote 2 of Sect. 9.4.3). This is correlated with the empirical result that the F→S transition at the bottleneck due to the off-ramp can occur only some distance upstream of the off-ramp (Sect. 10.5).

(ii) The growth of the initial external local perturbation (down-arrow 1, D23-off, Fig. 12.16a) leads to narrow moving jam emergence in free flow on the main road upstream (down-arrow 1, D19, Fig. 12.16c). When the speed within the narrow moving jam becomes low enough, the jam exhibits characteristic features of wide moving jams. In particular, the flow rate in the jam outflow ($q_{out, 1} \approx 1900$ vehicles/h in the left lane on the main road) is considerably lower than the flow rate in free flow. In other words, free flow is in a metastable state with respect to wide moving jam emergence (Sect. 6.1).

(iii) It turns out that other local perturbations of vehicle speed (dotted down-arrows 2–4, D21, Fig. 12.16c) in this metastable free flow do not lead to jam emergence. This confirms the theoretical supposition that the critical amplitude of a local perturbation required to cause an F→J transition should be much greater than that required to cause an F→S transition (Sect. 6.3).

(iv) The flow rate in the right lane on the main road spatially increases in the downstream direction upstream of the off-ramp. Accordingly, the flow rates in the other (especially left) lanes on the main road decrease from the detectors D19 to D22, i.e., when vehicles approach the off-ramp. Almost all vehicles in the right lane on the main road at D22 leave the freeway via the off-ramp: traffic flow in the right lane on the main road downstream of D22 and in the off-ramp lane could be considered a traffic flow on a single-lane road.

This can explain why in contrast to the abrupt speed decrease in the right lane on the main road (dotted up arrow, D22, Fig. 12.16c), the other lanes display only a very small decrease in speed. On the other hand, if such a decrease in speed in the right lane on the main road in the vicinity of on-ramps or away from bottlenecks is realized, an F→S transition usually occurs. However, in the vicinity of the off-ramp bottleneck, only those vehicles that want to continue on the main road change from the right lane to other lanes on the main road. In the example under consideration, this fraction is negligible. As a result, an F→S transition on the main road in the vicinity of the off-ramp is hindered. As a result, the induced F→J transition occurs.
12.7 Conclusions

(i) In empirical observations, wide moving jams do not emerge spontaneously in free flow. Firstly, an F→S transition should occur. Later and at other freeway locations, wide moving jams can emerge spontaneously in that synchronized flow (S→J transition).

(ii) The lower the speed in synchronized flow, the higher the frequency of narrow moving jam emergence in an GP.

(iii) In synchronized flow, cases of strong congestion and weak congestion should be distinguished. In strong congestion, the average flow rate in the pinch region is a self-sustaining value that is close to some limit (minimum) flow rate, the mean time between narrow moving jams reaches the least possible value and the mean width of synchronized flow in an GP is limited, and this width is independent of traffic demand. In contrast, under weak congestion the average speed in the pinch region begins to increase and the flow rate in the pinch region ceases to be a self-sustaining value that is close to the limit flow rate.

(iv) A local speed perturbation that initially emerges in an off-ramp lane can propagate upstream in this lane and reaches free flow on the main road. For free flow on the main road this local speed perturbation is an external local perturbation. If this free flow is in a metastable state with respect to an F→J transition and the amplitude of the perturbation is high enough, the external perturbation can cause induced wide moving jam formation. This is because no F→S transition can occur at road locations that are very close to the off-ramp. Thus, upstream speed perturbation propagation in the off-ramp lane can lead to an induced F→J transition on the main road.

(v) The duration of an S→J transition (about 10 min) is about 10 times the duration of an F→S transition.

(vi) Whereas both the vehicle speed and flow rate decrease abruptly during moving jam emergence (S→J transition), the flow rate does not necessarily decrease during an F→S transition.
13 Empirical Pattern Evolution
and Transformation at Isolated Bottlenecks

13.1 Introduction

From the theoretical diagram of congested patterns (Fig. 7.13) at isolated on-ramp bottlenecks we can expect that if the flow rate to the on-ramp $q_{on}$ and the flow rate on the main road upstream of the on-ramp $q_{in}$ are high enough, an GP should occur over a wide range of these flow rates. This theoretical prediction is confirmed by empirical observations of GP emergence discussed in Sect. 12.2.

Considering this GP emergence at bottlenecks due to on-ramps, we have seen that on the one hand, traffic demand behavior as a function of time is very similar on different days and years (Fig. 12.6).

If the flow rates $q_{in}$ and $q_{on}$ decrease over time, then from the congested pattern diagram we can expect some evolution of an initial GP at an isolated effectual bottleneck. Firstly, strong congestion in the GP should be replaced by the weak congestion condition. Later, when $q_{on}$ further decreases, rather than the GP one of the SPs can appear upstream of the on-ramp before free flow returns at the bottleneck. Such congested pattern evolution is indeed observed in empirical observations, as will be shown in this chapter [218].

Another aim of this chapter is to study transformations between various congested patterns [218]. The pattern transformation can also be predicted based on a consideration of the theoretical diagram of congested patterns (Fig. 7.13). This pattern transformation can be observed at high flow rate $q_{in}$ on the main road upstream of an isolated bottleneck under the weak congestion condition. In this case, we can expect from the diagram of congested patterns that a small change in traffic demand can lead to a complex transformation between WSPs, MSPs, and DGP s over time. The complex empirical pattern transformation is often observed at bottlenecks due to off-ramps where the weak congestion condition is more likely to be found.
13.2 Evolution of General Patterns at On-Ramp Bottlenecks

13.2.1 Transformation of General Pattern into Synchronized Flow Pattern

When the flow rate \( q_{\text{eff-on}} \) to the effectual on-ramp at the bottleneck at D6 on the freeway A5-South (Sect. 9.2.3) decreases below some value, an GP can transform into an SP (Fig. 13.1). This transformation occurs when the flow rate \( q_{\text{eff-on}} \) is related to the flow rate interval denoted by \( q_{\text{trans}} \) in Table 12.2.

Upstream of the SP free flow is realized (D1, Fig. 13.1c), i.e., this SP is localized at the bottleneck: an LSP occurs in this case. In comparison with an GP, in the LSP the flow rate is considerably higher than the limit flow rate \( q_{\text{lim}} \) (D5, Fig. 13.2). Thus, in the LSP the weak congestion condition is realized.

13.2.2 Alternation of Free Flow and Synchronized Flow in Congested Patterns

When the flow rate to the effectual on-ramp \( q_{\text{eff-on}} \) further decreases, local regions of free flow appear at D5 and D4. These free flow regions spatially alternate with local regions of synchronized flow at D3 (Fig. 13.1d). However, synchronized flow is self-sustaining at the bottleneck (D6). In addition, wide moving jams emerge in the farthest upstream region of synchronized flow within the pattern (D2 and D1 in Fig. 13.1d). Thus, this pattern can be considered a variant of an GP. Within this “alternating GP,” local regions of free flow spatially alternate with local regions of synchronized flow (labeled “AGP” in Fig. 13.1d).

The appearance of free flow at D5 and D4 can be explained by a return S→F transition within the initial LSP when the flow rate \( q_{\text{eff-on}} \) decreases (Fig. 13.2). However, because synchronized flow at the bottleneck (D6) is still self-sustaining, the discharge flow rate from the AGP remains at approximately the same level as those in the former LSP (Fig. 13.2). Thus, it can be assumed that the downstream front of the AGP is located at the bottleneck (D6), like the downstream front of the initial GP and the LSP.

There can also be another interpretation of the phenomenon of the appearance of the alternation of free flow and synchronized flow in congested patterns. The AGP can be considered as consisting of two different patterns: (i) an LSP that is localized in the vicinity of the bottleneck (D6, Fig. 13.1d) and (ii) an MSP; when this MSP is far from the bottleneck (D3–D1), wide moving jams emerge in synchronized flow of the MSP, i.e., the MSP transforms into an GP away from bottlenecks (Sect. 12.4).\(^1\)

\(^1\) This GP away from bottlenecks is qualitatively similar to the GP away from bottlenecks that is shown in Fig. 10.12. However, in the latter case, first a
Fig. 13.1. Evolution of an GP first into an LSP and later into an GP where a spatial alternation of free flow and synchronized flow occurs (labeled “AGP”) due to a decrease in the flow rate to the effectual on-ramp at the bottleneck at D6 on a section of the freeway A5-South (Fig. 2.1). (a) Overview. (b–d) Vehicle speed at different detectors within the GP (b), the LSP (c), and the AGP (d). Taken from [218]
On other days, the evolution of GPs when the flow rate $q_{\text{eff-on}}$ decreases can show qualitatively similar pictures as those in Figs. 13.1 and 13.2. However, sometimes instead of an AGP an SP where local regions of free flow spatially alternate with local regions of synchronized flow occurs. Such a congested pattern is called an “alternating SP” (ASP for short). The ASP can also be interpreted as consisting of two different patterns: (i) an LSP at the bottleneck, and (ii) an MSP where no wide moving jams occur later.

Sometimes, an initial GP transforms into an LSP that dissolves later. In some other cases, an appearance of free flow within synchronized flow of an initial GP or an SP leads to MSP occurrence only.

### 13.2.3 Hysteresis Effects Due to Pattern Formation and Dissolution

While an $F \rightarrow S$ transition at the effectual bottleneck is accompanied by the fall in vehicle speed (Sect. 10.2), a return $S \rightarrow F$ transition is accompanied by the jump in speed. Therefore, these two first-order phase transitions cause the well-known hysteresis effect (Fig. 13.3a,b).

In the case shown in Fig. 13.3a,b, there are two hysteresis effects. The first hysteresis effect is not due to pattern dissolution. This hysteresis effect is caused by the appearance of a wide moving jam between $D_6$ and $D_5$ (Fig. 12.9a). This jam leads to a return $S \rightarrow F$ transition at $D_6$ without GP dissolution (Sect. 12.2.7). However, this hysteresis effect is an exceptional case.

Usually an $S \rightarrow F$ transition occurs due to congested pattern dissolution. One of the scenarios of this dissolution is shown in Fig. 12.9a: a wave of return $S \rightarrow F$ transitions starts upstream of a congested pattern and propagates downstream (up-arrows $F_2$, $D_3-D_5$) to the effectual bottleneck ($D_6$, up-arrow $F_2$ and arrow at 9:21 in Figs. 12.9a and 13.3b, respectively). As a synchronized flow region has occurred away from bottlenecks and later a wide moving jam emerges in that synchronized flow.
result, the congested pattern dissolves and free flow occurs at the bottleneck (Fig. 12.9a).

Hysteresis phenomena in the flow–density plane are also observed at freeway locations upstream of a bottleneck where a congested pattern exists (Fig. 13.3c). After synchronized flow has occurred at the bottleneck (arrow at $t = 06:27$ at D6, Fig. 12.1), the upstream front of synchronized flow propagates upstream. As a result, synchronized flow reaches the upstream detectors D5 (arrow $S$ at D5, Fig. 12.1). This F→S transition (arrow $S$, Fig. 13.3c) and a return S→F transition (arrow $F_2$) lead to the hysteresis effect in the flow–density plane. However, the onset of synchronized flow upstream of the bottleneck (D5) is a secondary effect caused by the upstream propagation of synchronized flow from the bottleneck (D6). Nevertheless, we see that the hysteresis effects at the bottleneck (Fig. 13.3a) and upstream of the bottleneck (Fig. 13.3c) can show qualitatively the same behavior in the flow–density plane. Thus, to find the reason for an F→S transition, a study of

Fig. 13.3. Hysteresis phenomena at an on-ramp bottleneck (a, b) and upstream of the bottleneck (c). (a, b) F→S transitions (arrows are related to $t = 06:27$ and $t = 08:43$) and a reverse S→F transitions (arrows are related to $t = 08:35$ and $t = 09:21$) at the bottleneck. Data corresponds to the left lane at D6. Arrows at $t = 06:27$, $t = 08:35$, $t = 09:21$, and $t = 08:43$ are related to up-arrows at $t = 06:27$ in Fig. 12.1 (D6), $F_1$, $S_2$, and $F_2$ in Fig. 12.9a (D6), respectively. (c) The onset of synchronized flow (arrow $S$) and the reverse S→F transition (arrow $F_2$) upstream of the bottleneck (D5). Arrows $S$ and $F_2$ are related to up-arrows $S$ in Fig. 12.1 (D5) and $F_2$ in Fig. 12.9a (D5), respectively. Data is averaged across freeway lanes (per lane) in the flow–density plane. Free flow: black points, synchronized flow: circles. Taken from [218]
the spatiotemporal dynamics of synchronized flow (Sect. 12.2) is necessary (Figs. 12.1 and 12.9).

When an AGP or an ASP occurs, i.e., an GP or an SP where local regions of free flow spatially alternate with local regions of synchronized flow, some different scenarios of congested pattern dissolution are possible. In particular, in an AGP the dissolution of synchronized flow at the bottleneck D6 and the dissolution of another synchronized flow upstream of the bottleneck at the locations D3–D1 begin almost simultaneously (up-arrows $F_1$ and $F_2$ at 10:31, D6 and D1 in Fig. 12.2). However, because synchronized flow at D3–D1 is extended over about 3 km, this synchronized flow dissolves later. This dissolution is due to the wave of $S\rightarrow F$ transitions that propagate downstream of the detectors D1 to the detectors D2 and then to the detectors D3 (up-arrow $F_2$ at D2 and D3). As a result, the dissolution finishes at D3 three minutes after it began at D1. There can be another explanation of this pattern dissolution. This explanation is based on the above assumption that the AGP in Fig. 13.1 consists of two patterns as discussed in Sect. 13.2.2. Thus, these two patterns dissolve independently of each other due to two different waves of return $S\rightarrow F$ transitions.

13.3 Transformations of Congested Patterns
Under Weak Congestion

Empirical observations show that the weak congestion condition is usual a case for different congested patterns at off-ramp bottlenecks. In particular, weak congestion occurs in the MSP shown in Fig. 10.6a and in all other congested patterns shown in Figs. 13.4 and 13.5. Under the weak congestion condition diverse transformations between different types of congested patterns are often observed over time.

For example, an GP at an off-ramp bottleneck can transform into one of the SPs: either an LSP, or an WSP, or else an MSP can occur. In addition, in both the GP and the SP (LSP or WSP) local regions of free flow, which spatially alternate with regions of synchronized flow, are often realized, i.e., an AGP and an ASP appear. An example is shown in Fig. 13.4 where an GP first transforms into an WSP with such an alternation of free flow and synchronized flow (labeled “WSP”) and then later once again an GP without this alternation occurs.

2 This GP has been discussed in Sect. 9.4.2 (Fig. 9.14).

3 “Speed drops” in Fig. 13.4, which can be seen at D18 and D19 during the time interval 15:40–16:30, do not transform into wide moving jams, i.e., these drops are associated with synchronized flow. This follows from the analysis of the propagation of these drops through the upstream bottleneck $B_{North\ 2}$ that has been made: there are no wide moving jams that have emerged in synchronized flow during the mentioned time interval between the bottlenecks $B_{North\ 1}$ and $B_{North\ 2}$. This conclusion can be seen in Fig. 9.16.
Fig. 13.4. Evolution of congested patterns at an isolated bottleneck due to the off-ramp D25-off on a section of the freeway A5-North (Fig. 2.2). Vehicle speed on the main road at different detectors. GP: the general pattern (left and right); WSP: the widening SP (middle). Taken from [218]
Fig. 13.5. Evolution of congested patterns upstream of the off-ramp D23-off on a section of the freeway A5-South (Fig. 2.1). Vehicle speed at different detectors. DGP: the dissolving general pattern; GP: the general pattern; SP: the synchronized flow pattern. Taken from [218]
However, in this WSP almost free flow conditions are at D23–D21, i.e., this WSP can also be interpreted as a congested pattern whose downstream front has moved upstream from the initial effective location of the bottleneck in the vicinity of D23 to another effective location of the bottleneck at the detectors D20 (Fig. 13.4, the time interval 15:40–16:30 and Fig. 9.16). This interpretation is confirmed by the analysis made in Sect. 9.4.3. In the latter section, we have mentioned that in free flow between D23 and D20 small synchronized flow regions occur (Fig. 9.15a). The duration of this synchronized flow is sometimes only 1–2 minutes. Nevertheless we can also consider the part of this pattern as the alternation of free flow with synchronized flow.

At some bottlenecks due to off-ramps, GPs exist only during short time intervals: there are relatively frequent transformations between different congested patterns. Such a case is often realized upstream of the off-ramp D23-off on the freeway A5-South (Fig. 13.5). It can be seen that first an GP appears where only one wide moving jam is formed. Corresponding to the definition of dissolving GPs (DGP) (Sect. 7.5.2), this GP is an example of an DGP. The DGP transforms into an SP, then an GP appears once again.

An SP can transform spontaneously into an DGP or into an GP. An example of the transformation of an WSP into an GP is explained in Sect. 2.4.7 (Fig. 2.19). As we now know from the discussion in Sect. 12.5, this GP is associated with GPs under the weak congestion condition.

It can be seen in Fig. 13.6 that in these two different cases (Figs. 13.4 and 13.5), during pattern transformation the flow rate and percentage $\eta$ of vehicles that want to leave the main road to the off-ramp do not change appreciably over time. Thus, it can be assumed that at bottlenecks due to off-ramps, pattern transformation can easily occur, even for small changes in traffic demand and values $\eta$.

### 13.4 Discharge Flow Rate and Capacity Drop

The discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ is the flow rate in free flow formed in the congested pattern outflow. This free flow is measured downstream of the congested pattern at a bottleneck (see references in [30,64]).

The maximum and minimum values of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ during the time interval when a congested pattern existed at a bottleneck are shown both for one minute (Table 13.1) and for 10 minutes average data (Table 13.2). It can be seen that $q_{\text{out}}^{(\text{bottle})}$ can vary over a wide range $[q_{\text{out}}^{(\text{bottle})}, \min$, $q_{\text{out}}^{(\text{bottle})}, \max]$ [212]. It should be noted that the maximum value $q_{\text{out}}^{(\text{bottle})}, \max$ can appreciably exceed the flow rate in the outflow from a wide moving jam $q_{\text{out}}$ on the related day, whereas $q_{\text{out}}^{(\text{bottle})}, \min$ can be lower than $q_{\text{out}}$.

The discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ can depend on

(i) characteristics of the bottleneck
Fig. 13.6. Percentage $\eta$ of vehicles that want to leave the main road to the off-ramp (left) and total flow rates across the freeway (right) as functions of time on two different days. 10-min average data.

Table 13.1. The maximum total discharge flow rate $q_{\text{out, max}}^{(\text{bottle})}$ and the minimum total discharge flow rate $q_{\text{out, min}}^{(\text{bottle})}$ on three different days on a section of the freeway A5-South. Both flow rates are measured during the time interval when a congested pattern exists at the bottleneck at D6. 1-min average data. Taken from [218]
Table 13.2. The same parameters and characteristics as those in Table 13.1 but for 10-min average data. Taken from [218]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_{\text{out, max}}^{(\text{bottle})}$ [vehicles/h]</td>
<td>5940</td>
<td>5420</td>
<td>4880</td>
</tr>
<tr>
<td>$q_{\text{out, min}}^{(\text{bottle})}$ [vehicles/h]</td>
<td>5140</td>
<td>3640</td>
<td>3920</td>
</tr>
<tr>
<td>$q_{\text{out, max}}^{(\text{bottle})} / q_{\text{out}}$</td>
<td>1.19</td>
<td>1.2</td>
<td>1.16</td>
</tr>
<tr>
<td>$q_{\text{out, min}}^{(\text{bottle})} / q_{\text{out}}$</td>
<td>1.03</td>
<td>0.8</td>
<td>0.93</td>
</tr>
</tbody>
</table>

(ii) the type of congested pattern  
(iii) traffic demand  
(iv) traffic control parameters (weather, etc.)

The dependence of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ on traffic control parameters can be seen in Table 13.2. In this table, very different maximum values of the discharge flow rate $q_{\text{out, max}}^{(\text{bottle})}$ are found on three different days. These maximum values are related to the same bottleneck and to the same type of congested pattern – a GP under the strong congestion condition (Sect. 12.2). Also traffic demand (the flow rates $q_{\text{eff-on}}$ and $q_{\text{in}}$) related to these GPs are very similar on these days (Fig. 12.6). It has been found that the higher the flow rate in the wide moving jam outflow $q_{\text{out}}$, the higher the mean value of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ on the related day.

A typical dependence of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ on the congested pattern type and traffic demand at the bottleneck at D6 on the freeway A5-South is shown in Fig. 13.7a associated with patterns in Fig. 13.1. Firstly, a GP under the strong congestion condition occurs (Figs. 13.1a and 13.2). The discharge flow rate is relatively high because of the high flow rate to the effectual on-ramp $q_{\text{eff-on}}$. Over time the flow rate $q_{\text{eff-on}}$ begins to decrease (Fig. 12.6). Because the flow rate $q_{\text{pinch}}$ has only minor changes in the vicinity of the limit flow rate $q_{\text{lim}}^{(\text{pinch})}$ in the GP under the strong congestion condition, the average discharge flow rate (12.2) should also decrease corresponding to the change in the flow rate to the effectual on-ramp $q_{\text{eff-on}}$. This behavior is indeed observed in Fig. 13.7a. The discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ is considerably higher than the flow rate in the wide moving jam outflow $q_{\text{out}}$ shown by the dashed line in Fig. 13.7a.
Later, the flow rate $q_{\text{eff-on}}$ continues to decrease. Moreover, first the GP under the strong congestion condition transforms into a GP under the weak congestion condition (Fig. 12.10). Then an LSP and finally an AGP occur (Fig. 13.1a). The discharge flow rate decreases over time during these pattern transformations and the decrease in the flow rate $q_{\text{eff-on}}$ (Fig. 13.7a). The discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ can become lower than the flow rate in the wide moving jam outflow $q_{\text{out}}$ (Fig. 13.7a).

A different case is shown in Fig. 13.7b. This case is related to pattern transformation at the bottleneck due to the off-ramp shown in Fig. 13.5 and in Fig. 9.14 of Sect. 9.4.2. In contrast to the former case of the on-ramp bottleneck (Fig. 13.7a), we see that the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ does not change considerably over time. Furthermore, the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$
is slightly lower than the flow rate in the wide moving jam outflow $q_{\text{out}}$ shown by the dashed line in Fig. 13.7b.

This behavior of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ (Fig. 13.7b) can be explained as follows. In the case, first an WSP appears and later a GP under the weak congestion condition occurs (Sect. 12.5). When the GP occurs, the average discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ is approximately given by the formula (12.2). However, for the case of the bottleneck due to the off-ramp under consideration, $q_{\text{eff-on}} = 0$, i.e., rather than (12.2) we obtain

$$q_{\text{out}}^{(\text{bottle})} \approx q^{(\text{pinch})}.$$  

(13.1)

For continuous moving jam emergence in the pinch region of an GP, the average flow rate in the pinch region $q^{(\text{pinch})}$ in this GP must satisfy the condition

$$q^{(\text{pinch})} < q_{\text{out}}.$$  

(13.2)

The condition (13.2) has been derived in Sect. 7.6.2. The relations (13.2) and (13.1) can explain why the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ is slightly lower than the flow rate in the wide moving jam outflow $q_{\text{out}}$ and also why $q_{\text{out}}^{(\text{bottle})}$ does not change appreciably over time.

Note that for the GP in Fig. 9.14 the conditions (13.2) and (7.37) are only approximately satisfied. Strictly speaking, for this empirical GP $q^{(\text{pinch})}$ can be higher than $q_{\text{out}}$ during some time intervals. To explain this, we note that the intervals between wide moving jams in this GP are large (about 30 min and more; see the jams 1–4 in Fig. 9.14). As a result, the farthest downstream wide moving jam of the GP can pass the upstream on-ramp at D16 before a new moving jam emerges in the pinch region of the GP. For this reason, it can turn out that the inflow in the pinch region can be higher than $q_{\text{out}}$. This is due to vehicles that merge onto the main road at D16 from the on-ramp D15-on. These vehicles can increase the pinch region inflow rather than first dwelling in a wide moving jam of the GP.

The capacity drop [64] must be related to a difference between freeway capacity in free flow at a bottleneck and the discharge flow rate that gives the congested pattern capacity, i.e., the capacity of the congested bottleneck (Sect. 8.5).

Thus, for empirical measurements of the capacity drop $\delta q$ we can use the formula (8.30)$^4$:

$$\delta q = q_{\text{max}}^{(\text{free B})} - q_{\text{out}}^{(\text{bottle})}.$$  

(13.3)

The discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ can be a function of time (Fig. 13.7a). We can define the minimum empirical capacity drop through the formula$^5$

$^4$ In the theoretical formula (8.28), the value $q_{\text{max}}^{(\text{free})}$ is used that is related to the hypothetical case of a homogeneous road. For this reason, this formula cannot be used for the empirical capacity drop.

$^5$ We do not use the theoretical minimum capacity drop (8.29) because the maximum flow rate in synchronized flow is difficult to find empirically.
Empirical Pattern Transformation

\[ \delta q_{\text{min}} = q_{\text{max}}^{(\text{free B})} - q_{\text{out, max}}^{(\text{bottle})}. \]  

In this formula, it is assumed that \( q_{\text{max}}^{(\text{free B})} > q_{\text{out, max}}^{(\text{bottle})} \).

To discuss these definitions, we recall that for a bottleneck due to the on-ramp there are infinitely many maximum freeway capacities \( q_{\text{max}}^{(\text{free B})} \) (8.6). However, this is not the main problem for the determination of empirical freeway capacity.

The problem is that the maximum freeway capacity \( q_{\text{max}}^{(\text{free B})} \) can be found if the empirical dependence of the probability \( P_{FS}^{(B)} \) for speed breakdown (F→S transition) at a bottleneck as a function of traffic demand is known (Sect. 10.3.1). Specifically, the maximum freeway capacities \( q_{\text{max}}^{(\text{free B})} \) are found from the condition \( P_{FS}^{(B)} = 1 \) (10.6). To find the probability \( P_{FS}^{(B)} \) for the F→S transition at the bottleneck as a function of traffic demand, many different realizations (days) are required where F→S transitions with subsequent congested pattern formation occur at the bottleneck. However, we could see that the empirical discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \) can strongly depend on traffic control parameters (weather, etc.), i.e., the discharge flow rate can be very different on different days (Table 13.2). Thus, there is the following contradiction in the definitions of the capacity drop (13.3) and (13.4). On the one hand, the capacity drop must be related to only one realization where a congested pattern is formed at the bottleneck. On the other hand, to find the maximum capacity \( q_{\text{max}}^{(\text{free B})} \) one requires to use data measured on different days where F→S transitions occur.

Rather than the formula (13.3), one can use for measurements of capacity drop the difference between the pre-discharge flow rate \( q_{FS}^{(B)} \) and the maximum discharge flow rate

\[ \delta q_{1} = q_{FS}^{(B)} - q_{\text{out, max}}^{(\text{bottle})}. \]  

This is because the pre-discharge flow rate \( q_{FS}^{(B)} \) is equal to one of the freeway capacities in free flow at a bottleneck (Sect. 10.3.2). However, empirical studies show that this difference can sometimes be positive and sometimes negative on different days. This can be understood if we recall that the pre-discharge flow rate \( q_{FS}^{(B)} \) can vary over a wide range in which the conditions (10.11) are satisfied. In other words, in most of the different realizations (days) where F→S transitions have been observed at the bottleneck the pre-discharge flow rate \( q_{FS}^{(B)} \) is related to the condition \( P_{FS}^{(B)} < 1 \), i.e., this flow rate can be appreciably lower than the maximum freeway capacity \( q_{\text{max}}^{(\text{free B})} \).

Rather than capacity drop, one could try to measure the difference between the empirical maximum flow rate in free flow and the maximum discharge flow rate

\[ \delta q_{2} = q_{\text{max, emp}} - q_{\text{out, max}}^{(\text{bottle})}. \]  

Empirical studies show that this difference can sometimes be positive and sometimes negative on different days. This is probably because the empirical
maximum flow rate in free flow $q_{\text{max}}^{(\text{free, emp})}$ is also often related to the condition $P_{FS}^{(B)} < 1$, i.e., this flow rate can be appreciably lower than the maximum freeway capacity $q_{\text{max}}^{(\text{free B})}$.

13.5 Conclusions

(i) In accordance with the theoretical diagram of congested patterns at isolated bottlenecks in three-phase traffic theory (Chap. 7), after an GP has emerged at an on-ramp bottleneck and the flow rates $q_{\text{on}}$ and $q_{\text{in}}$ begin to decrease, then due to congested pattern evolution rather than the GP one of the SPs can appear upstream of the on-ramp before free flow resumes at the bottleneck.

(ii) During this GP evolution the strong congestion condition in the GP changes to the weak congestion condition before the GP transforms into an SP.

(iii) Under the weak congestion condition, congested pattern transformation is often observed. During this pattern transformation an WSP or an LSP can transform into an DGP and the latter can transform back into of one of the SPs over time. Spatial alternations of free flow and synchronized flow, i.e., the occurrence of an AGP or an ASP is often observed during complex pattern transformations.

(iv) The discharge flow rate from a congested bottleneck can depend on bottleneck characteristics, the type of congested pattern, traffic demand, and traffic control parameters (weather, etc.).
14 Empirical Complex Pattern Formation Caused by Peculiarities of Freeway Infrastructure

14.1 Introduction

As discussed in Sects. 2.4.8 and 9.1, if there are two or more adjacent effectual bottlenecks that are close to one another on a freeway section, then more complex congested patterns can be expected, in comparison with the possible patterns at an isolated effectual bottleneck. If some traffic phenomena at two adjacent effectual bottlenecks are considered below, then the effectual bottleneck upstream will be called “the upstream bottleneck” and the effectual bottleneck downstream will be called “the downstream bottleneck.”

Some effects associated with the existence of two or more adjacent effectual bottlenecks have already been discussed above. Examples are foreign wide moving jam propagation through a congested pattern (Sects. 2.4.9 and 11.3), an induced F→S transition caused either by wide moving jam propagation (Sect. 2.4.4) or by the catch effect associated with SP propagation (Sect. 2.4.6 and Chap. 10) [218]. In the latter two cases, a congested pattern (either a wide moving jam or an SP) first usually occurs due to the onset of congestion at the downstream bottleneck (wide moving jams or an SP can also occur away from bottlenecks). Due to the upstream propagation of the congested pattern, new effects of self-organization are realized at the upstream bottleneck. Here, either a new congested pattern emerges or some transformations are found within a congested pattern that already exists at the upstream bottleneck.

We can see in this chapter that at least the following different cases of the influence of freeway infrastructure on congested pattern features can be found [218]:

(i) A congested pattern occurs in which different spatiotemporal parts can be distinguished: each of these different parts is one of the congested patterns at the isolated bottlenecks (one of the SPs or the GPs) considered above.

(ii) When a congested pattern at the downstream bottleneck is an GP, then wide moving jams that emerge in this GP can propagate through the upstream bottleneck where another congested pattern has occurred. These wide moving jams are foreign wide moving jams for the congested pattern at the upstream bottleneck. If this pattern is also an GP, then
a complex congested pattern called “spatially separated GPs” occurs (Fig. 2.25 of Sect. 2.4.9). There are other cases, where at some of the bottlenecks GPs are formed and at the others SPs appear (see examples in Figs. 2.26 and 2.27).

(iii) In some cases, it turns out that the region of synchronized flow in this complex pattern affects two or more adjacent effectual bottlenecks. Such congested patterns have been called expanded congested patterns (EP for short) (Fig. 2.23 of Sect. 2.4.8).

(iv) A congested pattern appears that at the first glance seems to be one of the congested patterns at an isolated bottleneck. However, if a more precise analysis would be made one can find that the spatiotemporal structure of the pattern depends on some peculiarities of freeway infrastructure rather than on traffic demand upstream of the bottleneck only. In particular, we will see that a congested pattern shown in Fig. 4.2a (labeled “synchronized flow”) at the first glance seems to be an LSP at an isolated bottleneck. However, we find that in reality this pattern should be classified as a “shortened GP” under the strong congestion condition. In this “shortened GP,” due to some peculiarities of freeway infrastructure, all narrow moving jams dissolve at freeway locations upstream of the bottleneck during jam propagation before these narrow moving jams transform into wide moving jams.

We will also find that in some cases such dissolution of moving jams is caused by some kind of freeway bottleneck. These bottlenecks are upstream of an effectual bottleneck where a congested pattern has initially occurred.

14.2 Expanded Congested Pattern

14.2.1 Common Features

If two or more adjacent effectual bottlenecks exist close to one another on a freeway section, then an EP can occur where synchronized flow affects several adjacent effectual bottlenecks. We have already briefly discussed EPs in Sect. 2.4.8. Empirical observations show that in some cases two or more spatially separated pinch regions can occur in synchronized flow of the EP. Each of these pinch regions occurs upstream of the related effectual bottleneck.

In many other cases, where effectual bottlenecks are very close to one another EPs are observed where the pinch region in synchronized flow affects two or more adjacent effectual bottlenecks. The width (in the longitudinal direction) of this expanded pinch region can be much greater than the width of the pinch region occurring upstream of an isolated bottleneck. In these cases, narrow wide moving jams after they have emerged can remain to be narrow moving jams during jam propagation through the entire expanded pinch region without their transformation into wide moving jams. Naturally,
intermediate cases where some of the pinch regions are separated and the others cover several effectual bottlenecks are observed.

An EP can also have a spatiotemporal structure qualitatively similar to an GP at an isolated effectual bottleneck. However, empirical studies show that there are some important peculiarities of this EP. Firstly, because synchronized flow in the EP affects several adjacent effectual bottlenecks, the pinch region can be much longer than in the case of an isolated bottleneck. Secondly, many foreign wide moving jams, which have emerged within different pinch regions between different adjacent effectual bottlenecks, can propagate through the EP.

### 14.2.2 Example of Expanded Congested Pattern

As discussed in Sect. 12.5, upstream of the effectual bottleneck $B_{\text{North} 1}$ (Fig. 2.2) an GP is formed (Fig. 9.14). The upstream front of synchronized flow in this GP propagates upstream (up-arrows, D23–D17). After this front reaches the effectual bottleneck $B_{\text{North} 2}$ at the on-ramp (D16), an F→S transition at the latter bottleneck occurs (up-arrow in Fig. 9.14, D16). As a result, synchronized flow propagates subsequent upstream (up-arrows in Fig. 14.1, D15, D14) affecting both bottlenecks. Therefore, an EP occurs, whose overview is shown in Figs. 2.23 and 9.4.

In this EP, there are two spatially separated pinch regions in synchronized flow. The first pinch region has been formed between the bottleneck $B_{\text{North} 1}$ and the bottleneck $B_{\text{North} 2}$ (Fig. 9.14). In this pinch region, wide moving jams occur downstream of the bottleneck $B_{\text{North} 2}$ (Fig. 9.14, D17, up-arrows). These wide moving jams (down-arrows 1–4 in Figs. 9.14 and 14.1) that propagate upstream can be considered foreign wide moving jams for the second pinch region upstream of the bottleneck $B_{\text{North} 2}$ where other narrow moving jams emerge.

In the case shown in Fig. 14.1, strong congestion occurs in the pinch region (D15) upstream of the bottleneck $B_{\text{North} 2}$ (D16). As a result, the upstream propagation of the upstream front of synchronized flow is spatially limited. A region of wide moving jams is formed between D11 and D12, i.e., at approximately the same distance (about 4 km) from the effectual bottleneck at D16 as in the case of the effectual bottleneck at D6 on the freeway A5-South (e.g., Fig. 12.1).

However, these wide moving jams are not the reason for an F→S transition at the next upstream bottleneck $B_{\text{North} 3}$ (D7). This transition is earlier induced by the upstream propagation of a local region of synchronized flow that has originally appeared at D12 (up-arrow at 11:59). Therefore, the wide moving jams are also foreign jams for the congested pattern formed upstream of this third effectual bottleneck (D6 in Fig. 14.1). Upstream of this bottleneck another pinch region is formed where narrow moving jams emerge.

---

1 An overview of this effect is shown in Fig. 2.18.
Fig. 14.1. Part of an EP upstream of the bottleneck $B_{\text{North}2}$ at D16 on a section of the freeway A5-North (Fig. 2.2). Vehicle speed at different detectors. The downstream part of this EP (downstream of the bottleneck at D16) is shown in Fig. 9.14. An overview of the EP can be seen in Fig. 9.4. Taken from [218]
Thus, there are three different pinch regions in the final EP measured on this freeway section.

14.3 Dissolution of Moving Jams at Bottlenecks

14.3.1 Dynamics of Wide Moving Jam Outflow

There are bottlenecks that do not appear to be effectual ones, i.e., F→S transitions are not observed there. If a foreign moving jam that has initially occurred downstream of such a bottleneck propagates through this bottleneck either the width of the jam can decrease or the jam can dissolve.

We assume that the bottleneck is associated with a decrease in the number of freeway lanes from $m$ to $n$ in the direction of traffic flow ($n < m$). Then the flow rate $q_{\text{out}}$ in the outflow from a foreign wide moving jam increases when the jam propagates through the bottleneck corresponding to the ratio $m/n$ but the flow rate into the jam $q_{\text{in}}$ can remain the same.

An example of this foreign wide moving jam dissolution is shown in Fig. 14.2. In this case, a freeway bottleneck is related to a decrease in the freeway lanes from $m = 3$ to $n = 2$. Firstly, wide moving jams (marked by

---

![Diagram of freeway section](image)

**Fig. 14.2.** Dissolution of moving jams at bottlenecks. (a, b) Layout of the detector arrangement on a section of the freeway A1-Southwest between intersection “Dortmund-Unna” (upstream of D1) and intersection “Schwerte” (downstream of D9) in the direction of Cologne (Germany) (a) and vehicle speed at different detectors (b). Taken from [218]
down-arrows 1–3) are on a two-lane section of the freeway (D8). After the moving jams due to their upstream propagation reach a three-lane section of the freeway they begin to dissolve gradually (D7, D5). Finally, the moving jams have dissolved and free flow occurs at D3 upstream of the bottleneck.

A long off-ramp lane parallel to the other freeway lanes can also play the role of a bottleneck of this type. This case is realized on the freeway A5-South at D12, D13 (Fig. 2.1). Due to the off-ramp lane the effectual number of lanes decreases from $m = 4$ at D12 to $n = 3$ at D14 in the direction of traffic flow. Apparently for this reason the width of the foreign wide moving jam in Fig. 14.3 decreases when the jam propagates upstream in the vicinity

---

A5-South, June 23, 1998

---

Fig. 14.3. Decrease in the wide moving jam width at bottlenecks. Dependence of the duration of a wide moving jam due to jam propagation through a section of the freeway A5-South with a long off-ramp lane (D12 and D13, Fig. 2.1). Vehicle speed at different detectors. Taken from [218]
of D12 and D13 and the jam width again increases after the jam has passed the off-ramp. An overview of this wide moving jam is shown in Fig. 4.2a.

### 14.3.2 Localized Synchronized Flow Patterns Resulting from Moving Jam Dissolution

The influence of the long off-ramp lane (D12, D13) can also explain the dissolution of narrow moving jams shown in Fig. 14.4. These narrow moving jams emerge in the pinch region of a congested pattern (D14, Fig. 14.4) that appears upstream of the effectual bottleneck $B_2$ at D16 (Fig. 2.1).

![Graph showing vehicle speed at different detectors](image)

**Fig. 14.4.** Dissolution of narrow moving jams at bottlenecks. Formation of a “shortened GP” instead of an GP. Vehicle speed at different detectors. Taken from [218]

After this moving jam dissolution has finished, free flow occurs upstream of the pinch region at D12 (Fig. 14.4). Thus, a “shortened GP” under the strong congestion condition occurs at the on-ramp where the region of wide moving jams is not realized. This GP looks like an LSP in the overview of this “shortened GP” shown Fig. 4.2a where the “shortened GP” is labeled “synchronized flow.”

In contrast to the GPs and LSPs formed at isolated bottlenecks, the “shortened GP” that looks like an LSP (labeled “synchronized flow” in Fig. 4.2a) occurs only due to the mentioned peculiarities of the freeway
infrastructure upstream of the effectual bottleneck $B_2$ at D16 rather than due to nonlinear effects in traffic upstream of the bottleneck.

14.4 Conclusions

(i) When two or more adjacent effectual bottlenecks exist on a freeway section close to one another, complex empirical expanded congested patterns (EP) are observed. Synchronized flow in an EP affects these bottlenecks. This means that the downstream front of the EP is fixed at the farthest downstream of these effectual bottlenecks. Within this front vehicles accelerate from a lower speed in synchronized flow upstream of the front to a higher speed in free flow downstream of the front. Due to the upstream propagation of synchronized flow in the EP, the upstream front of this synchronized flow is upstream at least of one of the next upstream effectual bottlenecks.

(ii) The spatiotemporal structure of an EP can be very complex. This structure can often resemble the structure of one of the GPs. However, there can be several independent pinch regions in an EP. In each of these pinch regions narrow moving jams emerge. Some of these narrow moving jams can transform into wide moving jams. Thus, foreign wide moving jams that have emerged in the pinch region upstream of one of the downstream effectual bottlenecks can propagate through one of the upstream pinch regions in an EP. In this case, an EP is a complex “mixture” of several GPs with foreign wide moving jams propagating through the EP.

(iii) When effectual bottlenecks are very close to one another, EPs can be observed where rather than several spatially separated pinch regions an united pinch region is formed. The width of this pinch region (in the longitudinal direction) is usually greater than the one for each of separated pinch regions in an GP at an effectual isolated bottleneck.

(iv) Qualitative characteristics of the spatiotemporal structure of congested patterns can depend on peculiarities of the freeway infrastructure rather than on traffic demand upstream only.

(v) Freeway bottlenecks can cause wide moving jam dissolution when wide moving jams propagate through these bottlenecks. In this case, either an J→S transition or an J→F transition occurs.
15 Dependence of Empirical Fundamental Diagram on Congested Pattern Features

15.1 Introduction

As discussed in Sect. 2.3.5, the empirical fundamental diagram (the flow–density relationship) is a very important average characteristic of traffic flow. This characteristic is used in many engineering applications. Furthermore, the empirical fundamental diagram is an empirical basis for a theoretical fundamental diagram. The latter is used in a huge number of traffic flow theories and models within the scope of the fundamental diagram approach (Chap. 3).

To find the empirical fundamental diagram, the vehicle speed and flow rate are measured at a freeway location. When this empirical data is presented in the flow–density plane, there is a large spread in the data for congested traffic (e.g., [21,30,88]): any given density is related to many empirical points in the flow–density plane for different flow rates (different vehicle speeds). The empirical fundamental diagram (the empirical flow–density relationship) results from averaging these different empirical points to one average flow rate (one average speed) for each density.

15.1.1 Empirical Fundamental Diagram and Steady State Model Solutions

It must be noted that the empirical fundamental diagram does not exclude a two-dimensional (2D) region of steady states of synchronized flow in the flow–density plane within the scope of three-phase traffic theory (Fig. 4.4 of Sect. 4.3).

Let us explain this statement. The empirical fundamental diagram results from averaging aggregated empirical data. This empirical data averaging can be performed and used in engineering applications regardless of any traffic flow theory. In connection with traffic flow theories, the important point is whether the empirical fundamental diagram can be used as the basis for the fundamental diagram approach to traffic flow theory, i.e., as the basis for the hypothesis that hypothetical steady state solutions of a traffic flow mathematical model lie on a curve(s) in the flow–density plane. On the one hand, it is always possible and correct to average empirical data, to find the empirical
fundamental diagram. On the other hand, it is always a theoretical hypothesis whether theoretical steady states of traffic flow cover the curve(s) or a 2D-region in the flow–density plane: due to fluctuations and other complex dynamic spatiotemporal effects in congested traffic, hypothetical steady state model solutions cannot be found exactly in empirical data.

Thus, to decide which of two theoretical hypotheses (either theoretical steady states cover the curve(s) or a 2D-region in the flow–density plane) can be used for a correct description of traffic, we must compare theoretical features of phase transitions and congested patterns result from these two different approaches with empirical results. We have already mentioned that mathematical traffic flow models based on the fundamental diagram approach cannot show and predict important empirical features of phase transitions and congested traffic patterns (Sect. 3.3). In contrast to the fundamental diagram approach, three-phase traffic flow theory describes and predicts adequately all known empirical features of phase transitions and spatiotemporal congested patterns (Part III).

15.1.2 Two Branches of Empirical Fundamental Diagram

Corresponding to many empirical studies made in various countries, the empirical fundamental diagram consists of two isolated branches (curves). The branch for free flow and the branch for congested traffic (e.g., [21, 30, 37, 88, 151, 286]).

There are a huge number of different forms of the empirical fundamental diagrams that have been found on different freeways in various countries (see references in [20, 21, 30, 37, 151, 286]). In particular, the idea that certain features of the empirical fundamental diagram are affected by location within a congested pattern was discussed extensively in the late 1980s and early 1990s (see references in the book by May [21], the reviews by Hall et al. [30], Banks [37], and in the Highway Capacity Manual [286]). These discussions had to do with the question of whether the idea that the diagram was inverted-U shaped as opposed to inverted-V or reversed-λ was a result of insensitivity to the effects of entering and exiting traffic in congested traffic upstream of bottlenecks. Hall, Hurdle, and Banks [30] took note of the point and trace the idea that “the nature of the data acquired from a freeway depends on where they are taken with respect to bottlenecks and congestion” as far back as Edie and Foote [80].

Rather than a discussion of this diverse multitude of empirical fundamental diagrams at different freeway locations [20, 21, 30, 37, 151, 286] the main aim of this chapter is to discuss a possible correlation between the empirical fundamental diagrams and the spatiotemporal structure of congested patterns at freeway bottlenecks studied in Chaps. 9–14. One of the aims of this analysis is to explain the nature of some well-known empirical fundamental diagrams [20, 21, 30, 37, 151, 286], in particular why and when the diagram is inverted-U shaped, or inverted-V, or else reversed-λ. Another objective
of this chapter is to give a methodology for an explanation of other possible forms of empirical fundamental diagrams based on the spatiotemporal pattern classification (Sect. 2.4, Chaps. 9 and 14).

We could see in Chaps. 9 and 14 that at freeway bottlenecks, depending on the bottleneck type and traffic demand, qualitatively different spatiotemporal congested patterns can emerge. The spatiotemporal structure of some of these patterns can be very complex. In particular, there is a spatial dependence of mean values of traffic variables along a congested pattern upstream of a freeway bottleneck (see e.g., Fig. 12.5b). Thus, we can expect to find very different mean characteristics of these traffic variables at different freeway locations upstream of the bottleneck. This means that if average characteristics of freeway traffic are studied, these characteristics should depend on both the pattern type and on freeway location.

This general conclusion also concerns the flow–density relationship, i.e., the empirical fundamental diagram. In this chapter, we show that there is a certain qualitative scheme of spatial dependence of the empirical fundamental diagram on freeway location within congested patterns [469].

Note that for a study of the empirical fundamental diagram, traffic measurements on many different days are often aggregated and averaged (e.g., [21, 30, 37, 286]). However, on different days even at the same bottleneck, congested patterns with different features can occur. For this reason, to find a correlation between the spatial dependence of the empirical fundamental diagram and features of some congested pattern, we use data measured only for this congested pattern. In other words, branches of the empirical fundamental diagrams presented below in this chapter are only “portions” of empirical fundamental diagrams derived from average data aggregated on many different days. The study of these “portions” can help us to explain the well-known forms of fundamental diagrams and to develop a methodology, which establishes connection between empirical fundamental diagrams and spatiotemporal congested pattern features.

15.1.3 Line $J$ and Wide Moving Jam Outflow

To study features of the empirical fundamental diagram, we compare data in the flow–density plane related to synchronized flow within a congested pattern with the line $J$. The sense of this comparison is clear if we recall the definition of the line $J$ (Sects. 3.2.6 and 11.2.1) [166, 203].

The line $J$ is determined by the line slope and the left coordinates of the line in the flow–density plane. The line slope is equal to the velocity of the downstream jam front, $v_g$. The left coordinates (the vehicle density and flow rate) are associated with the jam outflow. After the line $J$ has been found, the point $(\rho_{\text{max}}, 0)$ of intersection of the line $J$ with the density axis gives the state within the wide moving jam with the jam density $\rho_{\text{max}}$ and speed $v_{\text{min}} = 0$. It should be noted that measurement points associated with a wide
moving jam, i.e., with states within the jam and within the jam fronts, do not necessarily lie on the line $J$ (a region labeled “Jam” in Fig. 15.1).

Recall that the mean value of the velocity of the downstream jam front is a characteristic value that does not depend on initial conditions and on traffic demand (Sect. 11.2). In other words, the propagation of the downstream jam front is a steady dynamic process, i.e., a dynamic process whose mean characteristics do not depend on time. The sense of the line $J$ is to represent this steady wide moving jam propagation in the flow–density plane rather than the line $J$ being a part of the empirical fundamental diagram.

However, there is an important connection between the line $J$ and the empirical fundamental diagram. This connection is related to the fact that the left coordinates of the line $J$ are associated with the outflow from a wide moving jam. Average traffic variables (density and flow rate) in this jam outflow are measured at some distance (about 0.5–2 km) downstream of the downstream jam front (a region labeled “Out” in Fig. 15.1) (see Sect. 11.2.1).\(^1\)

If free flow is formed in the jam outflow, the flow rate in this jam outflow $q_{out}$ and density $\rho_{min}$, which determine the left coordinates of the line $J$, are also characteristic parameters. In contrast, if synchronized flow is formed in the jam outflow, then the flow rate and density are not characteristic parameters. The flow rate is lower and the density in this synchronized flow is higher than the flow rate $q_{out}$ and density $\rho_{min}$, respectively. However, the

\(^1\) We assume in this explanation that a freeway location where the jam outflow is measured is not related to the pinch region of synchronized flow. This is because the measurement points associated with the pinch region can lie above the line $J$ in the flow–density plane (Sect. 12.2).
important feature of the line $J$ is that each coordinate on the line $J$ left of the point $(\rho_{\text{max}}, 0)$ in the flow–density plane is associated with the average flow rate and density in the jam outflow, i.e., downstream of the downstream jam front (the region labeled “Out” in Fig. 15.1). In other words, the average flow rate and density in the jam outflow give an empirical point that lies on the line $J$.

When synchronized flow occurs in the jam outflow, the density and flow rate in the jam outflow can be very different for different wide moving jams (Sect. 11.2.3) [205, 207]. As a result, these different jam outflows can give also different average points in the flow–density plane. On the one hand, corresponding to the definition of the line $J$, each of these points lies on the line $J$. On the other hand, if these points are measured at the same freeway location these average points are also related to the empirical fundamental diagram. Thus, the line $J$ should coincide with a part of the empirical fundamental diagram, if this part is exclusively determined by the measurement points associated with the outflows from different wide moving jams that are related to the same traffic control parameters (see footnote 2).

If a moving jam just transforms into a wide moving jam, then measured average data associated with the jam outflow should tend to the line $J$ in the flow–density plane [203]. This empirical result is related to the above conclusion that empirical average characteristics of the flow rate in the wide moving jam outflow are associated with the points in the flow–density plane that lie on the line $J$ [205, 207, 218]. This conclusion of the study of average data in [205, 207, 218] has recently been confirmed by an analysis of single vehicle data in [192].

Let us consider a congested pattern (an GP or an EP) where a sequence of wide moving jams is formed. We assume that between these jams synchronized flow is formed and that there is a detector on the freeway that measures traffic flow variables. During time intervals when these wide moving jams are not at the detector, average traffic variables measured at the detector are associated with the outflow from a wide moving jam that is upstream of the detector location. Corresponding to the line $J$ definition, average synchronized flow variables related to this jam outflow should give a point in the flow–density plane that lies on the line $J$, if no on- and off-ramps are between the detector and the jam (Sect. 11.2.3). Synchronized flow associated with different wide moving jam outflows can give different points on the line $J$ in the flow–density plane (see footnote 2).

In contrast, if an SP occurs, then there are no wide moving jams in synchronized flow of the SP. Average data associated with this synchronized flow does not necessarily lie on the line $J$. Depending on the vehicle speed in

\footnote{All these different points lie on the same line $J$ only if the jam outflows are related to the same traffic control parameters (weather, percentage of long vehicles, number of freeway lanes, etc.).}
this flow, we can expect qualitatively different flow–density relationships for congested traffic. Thus, we can expect the following features [469]:

(i) When wide moving jams emerge in synchronized flow, the empirical fundamental diagram should be related to the line \( J \).
(ii) When no wide moving jams emerge in synchronized flow, the empirical fundamental diagram for congested traffic should be different from the line \( J \).
(iii) Because wide moving jams emerge more likely in synchronized flow of higher density and lower speed (Chap. 12), the empirical fundamental diagram should asymptotically approach the line \( J \) at higher densities. In other words, at higher vehicle densities the empirical fundamental diagram should show the asymptotic behavior to the line \( J \) in the flow–density plane.

These expected features of the empirical fundamental diagram for congested traffic can explain the importance of the comparison of this diagram with the line \( J \) in the flow–density plane.

### 15.2 Empirical Fundamental Diagram and Line \( J \)

To study a possible correlation between the line \( J \) and the empirical fundamental diagram, we first consider an example of an WSP in Fig. 9.5. One should recall that the WSP in Fig. 9.5 can be considered a congested pattern at an isolated bottleneck due to the off-ramp D25-off only up to the detectors D17 (Sect. 9.3.1); the upstream detectors D16 are already related to the next upstream effectual bottleneck due to the on-ramp at D16 (Sect. 9.2.2). As a result of the upstream propagation of synchronized flow of the initial WSP upstream of this on-ramp, an EP is formed (Sect. 14.2). Synchronized flow in the EP covers the upstream bottleneck at the on-ramp D16. However, a part of the EP related to D17–D24, where no wide moving jams occur, can be considered an WSP upstream of the bottleneck due to the off-ramp D25-off.

#### 15.2.1 Asymptotic Behavior of Empirical Fundamental Diagrams

Firstly, we compare the line \( J \) with empirical data for the left freeway lane. Recall that on German three-lane freeways (Figs. 2.1 and 2.2) the left lane is the passing lane. This lane is usually not used by trucks and other long vehicles. “Aggressive” drivers often prefer this passing lane. This enables us to study effects approximately regardless of driver and vehicle characteristics.

Considering 1-min data for a part of the WSP (D19–D21), we can see a spread in empirical points for synchronized flow (Fig. 15.2). Such spread
Fig. 15.2. States of free flow (black points) and states of synchronized flow (circles) for the WSP in Fig. 9.5 (the detectors D21–D17) and at the detectors D16 in comparison with the line J (dashed line J). 1-min average data at different detectors in the left lane of the freeway A5-North (Fig. 2.2). Taken from [469]

of data in congested traffic is a well-known effect (e.g., [30, 88]). To see the features of empirical points clearly, we average these empirical points. There are several possibilities for this averaging. We use the following procedure: the density axis is divided in the flow–density plane into small density ranges. Each of these density ranges is equal to 3 vehicles/km. Then we average 1-min data shown in Fig. 15.2 for each of these density ranges separately. This means that all different empirical points in the density range are averaged to one average point. This point is related to the density value in the middle of the related density range. The result of this averaging for the left lane is shown in Fig. 15.3.
Recall that the lower the average vehicle speed, the greater the error in speed measurements (Sect. 2.3.1). For this reason, considering the average data, we do not use those densities $\rho = \frac{q}{v}$ (2.10) that are greater than 70 vehicles/km (Fig. 15.3).

Downstream of the WSP, at the detectors D24 only free flow is formed (Fig. 15.3). Considering the WSP in the upstream direction from the detectors D21 to the detectors D17, we can see the following.

A linear approximation of the average points of synchronized flow at the location D21 has a positive slope in the flow–density plane. The positive slope of the linear approximation of the synchronized flow states also remains for...
upstream locations (D20–D19 in Fig. 15.3). However, the slope decreases in the upstream direction.

At the location D18 a new effect occurs (Fig. 15.3): there should be two different linear approximations for average synchronized flow states. One approximation associated with lower densities has a positive slope. Another linear approximation associated with higher densities has a negative slope.

At the location D17 narrow moving jams emerge in synchronized flow. Apparently for this reason, empirical average synchronized flow states are described by only one linear approximation. This approximation is above the line $J$ and the approximation has a negative slope. This slope is slightly more negative than the slope of the line $J$. This corresponds to the conclusion of Sect. 12.2.1 that narrow moving jams can have a more negative velocity than the characteristic velocity $v_g$ of the downstream front of a wide moving jam. As mentioned above, upstream of D17 due to the upstream bottleneck $B_{\text{North 2}}$ at D16 the congested pattern is an EP, i.e., the pattern cannot be considered a congested pattern at the isolated bottleneck $B_{\text{North 1}}$ any more.

At the location D16 the dynamics of synchronized flow in this EP should essentially be determined by the upstream bottleneck $B_{\text{North 2}}$ due to the on-ramp D15-on. In this case, some narrow moving jams transform into wide moving jams. As a result, the linear approximation of empirical average synchronized flow states at D16 coincides with the line $J$ (Fig. 15.3).

An interpolation of the average free flow states and states of congested traffic (synchronized flow states plus states related to the fronts of wide moving jams) gives us the flow–density relationships, i.e., the empirical fundamental diagrams for the left lane shown in Fig. 15.4. In all cases we find the well-known result (e.g., [30,37,286]) that there are two branches of the empirical fundamental diagram: the branch $F$ for free flow and the branch $C$ for congested traffic. The branch $C$ results from averaging traffic states associated with both traffic phases, “synchronized flow” and “wide moving jam.” The branch for congested traffic $C$ exhibits, however, qualitative changes when the branch $C$ is measured at different freeway locations within the spatiotemporal congested pattern.

We have found the following qualitative dependence of the empirical fundamental diagram on spatial coordinate [469]:

(i) At freeway locations where the average speed in synchronized flow is higher, we find that the branch of the fundamental diagram for congested traffic $C$ has a positive slope in the flow–density plane, just like the branch for free flow $F$. The average slope of the branch for congested traffic measured at upstream locations decreases when the average synchronized flow speed decreases in the upstream direction.

(ii) At upstream freeway locations where the average speed in synchronized flow subsequent decreases, we find that the curve for congested traffic $C$ has a maximum in the flow–density plane: at lower densities the branch $C$ has a positive slope and at higher densities it has a negative slope.
At freeway locations where moving jams emerge in synchronized flow of lower vehicle speed, the branch for congested traffic at higher density exhibits the tendency towards the line $J$.

Finally, at locations where wide moving jams occur in synchronized flow, the branch for congested traffic $C$ at higher density is related to the line $J$ in the flow–density plane.

These results are correlated with conclusions that have been made above based on linear approximations in Fig. 15.3. Already from these results we see that the branch of congested traffic $C$ of the empirical fundamental diagram exhibits asymptotic behavior to the line $J$ when the average vehicle density increases: in the region of a congested pattern where wide moving jams are realized the branch $C$ at higher densities almost coincides with the line $J$. 

Fig. 15.4. Empirical fundamental diagram at different freeway locations for the left lane related to an approximation of free flow and synchronized flow states in Fig. 15.3. Branches $F$ and $C$ are related to free flow and congested traffic, respectively. Taken from [469]
This result is related to the physical sense and the definition of the line $J$: each average point of measured data associated with the wide moving jam outflow must lie on the line $J$ in the flow–density plane.

### 15.2.2 Influence of Different Vehicle Characteristics on Fundamental Diagrams

The above consideration is related to the left freeway lane where there are no long vehicles. In contrast, in the middle and right lanes the percentage of trucks and other long vehicles can be high.

*Qualitative results* about the dependence of the branch of congested traffic $C$ on the spatial coordinate discussed for the left freeway lane do not depend on various vehicle and driver characteristics in real traffic flow. However, there are some *quantitative* changes in characteristics of average synchronized flow states over all three lanes (Fig. 15.5) that are sometimes different from that

![Graphs showing various data points and lines](image)

**Fig. 15.5.** The same characteristics as those in Fig. 15.3, but for data averaged across all three freeway lanes. Taken from [469]
for the left lane (Fig. 15.3). These quantitative changes can also be seen if the empirical fundamental diagrams for the left lane (Fig. 15.4) and for data averaged across all three freeway lanes (Fig. 15.6) are compared.

**Fig. 15.6.** The same characteristics as those in Fig. 15.4, but for data averaged across all three freeway lanes. Data from Fig. 15.5. Taken from [469]

In particular, these differences are as follows [469]:

1. The maximum in the branch for congested traffic $C$ at the location D18 is a smooth one for the data averaged across all three lanes (Fig. 15.6) in comparison to the relatively sharp maximum for the left freeway lane (Fig. 15.4).
2. Whereas for the left lane the maximum flow rate in free flow has been higher than the maximum flow rate in the wide moving jam outflow $q_{out}$ at all detectors (Fig. 15.4), we find this result only at the locations D16–D18 for the data averaged across all three lanes (Fig. 15.6). Note that a
difference between flow–density relationships for different freeway lanes is also observed on US freeways (see review by Banks [37]).

(3) The branch $C$ for congested traffic at the location D16 is close to, but above the line $J$ for the data averaged across all three lanes (Fig. 15.6). This is probably because at the location D16 some of the moving jams do not exhibit the characteristic velocity $v_g$ for all three freeway lanes, i.e., they are still narrow moving jams in the middle and/or right lanes.

15.3 Dependence of Empirical Fundamental Diagram on Congested Pattern Type

In this section, we study the dependence of the empirical fundamental diagram on the congested pattern type at a freeway bottleneck [469]. We will see that in GPs and EPs the asymptotic behavior of the empirical fundamental diagram to the line $J$ found above is usually the case.

GP under Strong Congestion Condition

Let us first compare measured data at different locations within a GP under the strong congestion condition. Because the results for different GPs that occur on different days are very similar in all cases, we will restrict the analysis to only one GP that is shown in Fig. 12.5 (Sect. 12.2).

At the location D7, i.e., downstream of the bottleneck at D6 due to the on-ramp only free flow occurs (Fig. 15.7). At the bottleneck an GP emerges (Fig. 12.5); at the bottleneck at D6 synchronized flow is realized. Upstream of the bottleneck at the locations D5–D4 the pinch region occurs in synchronized flow of the GP where narrow moving jams emerge. Some of these narrow moving jams transform into wide moving jams at the locations D3–D2.

The average data (Fig. 15.8) and the related flow–density relationships (Fig. 15.9) show qualitatively the same results of the dependence of empirical fundamental diagrams on spatial coordinate discussed above in Sect. 15.2. In particular, at the freeway location D6 where the average speed in synchronized flow is higher we find that the branch of the fundamental diagram for congested traffic $C$ has a positive slope in the flow–density plane, just like the branch for free flow $F$. At upstream freeway locations D5–D4 where the average speed in synchronized flow subsequent decreases we find that the branch for congested traffic $C$ has a maximum in the flow–density plane. At freeway locations D3–D1 where wide moving jams occur the branch for congested traffic for higher density coincides with the line $J$ in the flow–density plane.

However, we also find some new peculiarities. Firstly, recall that for the WSP discussed in Sect. 15.2.1 there is a long section of the freeway between locations D21–D19 where the branch of the fundamental diagram for congested traffic $C$ has a positive slope in the flow–density plane. In contrast, in
the case of a GP under the strong congestion condition there is no such long freeway section: by location D5 just upstream of the bottleneck, the branch for congested traffic C has a maximum in the flow–density plane. At lower densities the slope of this branch is positive and at higher densities the slope is negative. This is associated with the pinch effect in synchronized flow of the GP. In the pinch region of the GP, narrow moving jams emerge propagating upstream. For this reason, at higher densities the slope of the flow–density relationship is negative. Investigations show that this slope can be more negative than the slope of the line J. This can be explained if we assume that narrow moving jams in the pinch region have a more negative velocity than the velocity of the downstream front of a wide moving jam. Furthermore, this part of the branch for congested traffic at higher density is above the line J.
This is related to results of Sect. 12.2.1 where we have discussed that states of synchronized flow in the pinch region at higher densities are above the line $J$ in the flow–density plane.

In the region of wide moving jams of the GP (the locations D3–D1), the branch of congested traffic of fundamental diagrams at higher densities lie on the line $J$. Thus, the line $J$ is an asymptote for the fundamental diagrams of traffic flow at higher densities for those freeway locations where the region of wide moving jams is formed within the GP.

This and other conclusions made for the left freeway lane are qualitatively valid for the other lanes as can be seen from Fig. 15.10, where the data is averaged across all freeway lanes. There are only some quantitative peculiarities in the latter case associated with the asymmetry between different
lanes (different vehicle parameters and driver characteristics). In particular, the difference between the maximum flow rate in free flow and the maximum flow rate in the wide moving jam outflow $q_{out}$ is lower than for the left lane at some locations.

**GP under Weak Congestion Condition**

If we now study measured data at different locations within a GP under the weak congestion condition shown in Fig. 9.14 (Sect. 12.5), then we find that the dependence of fundamental diagrams at different freeway locations displays intermediate features between the ones for the WSP (Fig. 15.4) and for the GP under the strong congestion condition (Fig. 15.9).
At the location D27, i.e., downstream of the bottleneck due to the off-ramp D25-off, only free flow occurs (Fig. 15.11). At the bottleneck an GP emerges (Fig. 9.14); upstream of the bottleneck at the locations D24–D23 synchronized flow occurs. In synchronized flow of the GP, wide moving jams emerge (D22–D19).

At freeway locations D24–D23 the branch of the fundamental diagram for congested traffic C has a positive slope in the flow–density plane, just like the branch for free flow F (Fig. 15.12). The line J is an asymptote for the fundamental diagrams of traffic flow at higher densities for those freeway locations D22–D19 where the region of wide moving jams is forming within the GP. The branches for congested traffic at the freeway locations D22–D19 exhibit a maximum point; at lower densities they have a positive slope.

On the one hand, it can be seen that in contrast to a GP under the strong congestion condition, for a GP under the weak congestion condition similar

---

**Fig. 15.10.** The same characteristics as those in Fig. 15.9, but for data averaged across all three freeway lanes. Taken from [469]
Fig. 15.11. States of free flow (black points) and states of synchronized flow (circles) for the GP in Fig. 9.14 in comparison with the line $J$ (dashed line $J$). 1-min average data at different detectors in the left lane of the freeway A5-North (Fig. 2.2). Taken from [469]

to that for an WSP there is a fairly long section of the freeway (D24–D23) where the branch of the fundamental diagram for congested traffic $C$ has a positive slope in the flow–density plane (Fig. 15.12). On the other hand, as in the GP under the strong congestion condition (Fig. 15.9) there is also a long freeway section (D22–D19) where wide moving jams occur in the GP and the line $J$ is an asymptote for the fundamental diagrams of traffic flow at higher densities.

**Expanded Patterns**

We find similar results for EPs. The line $J$ is often an asymptote for the fundamental diagrams of traffic flow at higher densities within an EP. This
Fig. 15.12. Empirical fundamental diagram at different freeway locations for the left freeway lane related to an approximation of free flow and synchronized flow states. Branches $F$ and $C$ are related to free flow and congested traffic, respectively. Taken from [469]

usually occurs at the freeway locations within EPs where wide moving jams propagate and the pinch region exists where narrow moving jams emerge. Recall that an EP often emerges if there are at least two adjacent effectual bottlenecks, the downstream and upstream bottlenecks, which are close to one another. At the downstream bottleneck a congested pattern should be formed. If the upstream front of synchronized flow in this pattern propagates upstream of the upstream bottleneck, then an EP emerges whose synchronized flow affects both of these bottlenecks. An example of an EP is the pattern in Fig. 9.5, where the synchronized flow in the initial WSP upstream of the downstream bottleneck (the off-ramp D25-off, Fig. 2.2) first reaches and then propagates upstream of the upstream bottleneck (the on-ramp at D16). Another example has been discussed in Sect. 14.2.
In the first example, wide moving jams emerge at the upstream bottleneck and therefore the line $J$ is an asymptote for the branch of congested traffic at this location (D16, Fig. 15.4). In the second example (Fig. 14.1), upstream of the bottleneck $B_{\text{North}}$ the pinch region is formed where narrow moving jams emerge. However, wide moving jams from the GP (Fig. 9.14) downstream of the bottleneck $B_{\text{North}}$ propagate through this pinch region upstream of D16 (Sect. 14.2). These wide moving jams are foreign wide moving jams for this pinch region.

It turns out that the empirical fundamental diagram for this pinch region shows the same features as those for the pinch region of an GP (D15, Fig. 15.13). This means that the compression of synchronized flow in the pinch region has more influence on the fundamental diagram than foreign wide moving jam propagation. However, already 1 km upstream of the pinch region, where the compression of synchronized flow decreases, the line $J$ is an asymptote for the branch of congested traffic of the empirical fundamental diagram (D14, Fig. 15.13). Although some of the narrow moving jams emerge in the pinch region upstream of the on-ramp bottleneck, they do not still transform into wide moving jams. This is associated with foreign wide moving jam propagation: in the outflow from these jams traffic flow is formed related to averaged points lying on the line $J$.

15.4 Explanation of Reversed-\(\lambda\), Inverted-V, and Inverted-U Empirical Fundamental Diagrams

We have already mentioned that there are many forms of the empirical fundamental diagram observed in various countries [21, 30, 35, 37, 88, 151, 286]. In particular, the most well-known are the reversed-\(\lambda\), the inverted-V, and inverted-U empirical fundamental diagrams (Fig. 15.14; e.g., [21, 30, 37, 80, 88, 286]). It must be noted that these empirical fundamental diagrams are usually related to averaging data aggregated on many different days. The
correlation between spatiotemporal features of congested patterns and the form of the branch for congested traffic on the empirical fundamental diagram discussed above enables us to explain the nature of these empirical fundamental diagrams.

Fig. 15.14. Qualitative forms of empirical fundamental diagrams [21, 30, 37, 80, 88]: (a) The reversed-\( \lambda \) fundamental diagram. (b) The inverted-V fundamental diagram. (c) The inverted-U fundamental diagram. The letters \( F \) and \( C \) mark the branches for free flow and congested traffic, respectively.

It can be assumed that the reversed-\( \lambda \) and the inverted-V fundamental diagrams are related to freeway locations upstream of bottlenecks where GPs and EPs usually occur. If these locations are within the region of wide moving jams, then the most contribution into the branch of congested traffic (the branches \( C \) in Figs. 15.14a,b) is made by flows associated with the outflows from these wide moving jams. In this case, average data for congested traffic lie on the line \( J \) in the flow–density plane.

The reversed-\( \lambda \) fundamental diagram (Fig. 15.14a) corresponds to the case when the maximum flow rate in free flow is considerably higher than the flow rate in the jam outflow (e.g., Fig. 15.10, D3–D1):

\[
q^{(\text{free, emp})}_{\text{max}} > q_{\text{out}}. \tag{15.1}
\]

In contrast, the inverted-V fundamental diagram (Fig. 15.14b) corresponds to the case when the maximum flow rate in free flow does not appreciably exceed the flow rate in the jam outflow (e.g., Fig. 15.6, D16, D17):

\[
q^{(\text{free, emp})}_{\text{max}} \approx q_{\text{out}}. \tag{15.2}
\]

For a bottleneck due to the on-ramp this is true even if the fundamental diagram is related to average data aggregated on many different days. This is because an GP is the most frequent type of congested pattern at the bottleneck due to the on-ramp. However, on different days the characteristic velocity of the downstream front of wide moving jams as well as other characteristic jam parameters can be different. This is because traffic control parameters (weather, etc.) can be very different on these days. For this reason, the branch for congested traffic (branches \( C \) in Figs. 15.14a,b) is related to some average line \( J \). This average line \( J \) represents average characteristics...
of the steady propagation of all wide moving jams whose outflows have been measured at the detector on these different days.

When traffic demand and the throughput from other adjacent upstream bottlenecks are high enough on these days, then the maximum flow rate in free flow can be considerably higher than the average flow rate in the latter jam outflows. Then the reversed-\(\lambda\) fundamental diagram should be observed. In the opposite case, the inverted-V fundamental diagram should be the case.

If the location of measurements of the empirical fundamental diagram is related to the pinch region in synchronized flow of GPs or EPs, then the inverted-U fundamental diagram should be observed (Fig. 15.14c). This is because in the middle range of density in congested traffic the main contribution to this fundamental diagram is caused by branches of congested traffic associated with synchronized flow states. In this case, there is a maximum point on the branch \(C\) and this branch can lie above the line \(J\) (e.g., Fig. 15.10, D4).

15.5 Conclusions

(i) The branch for congested traffic in the empirical fundamental diagram strongly depends both on the type of congested pattern and on the freeway location where the fundamental diagram is measured.

(ii) At freeway locations where the average speed in synchronized flow is higher the branch of the fundamental diagram for congested traffic has a positive slope in the flow–density plane.

(iii) The average positive slope of the branch for congested traffic can decrease when congested traffic is measured at upstream locations. This occurs if the average synchronized flow speed within a congested pattern decreases.

(iv) When the average speed in synchronized flow decreases in the upstream direction, the branch for congested traffic measured at some upstream location has a maximum in the flow–density plane: at lower densities this branch has a positive slope and at higher densities it has a negative slope.

(v) At freeway locations, where the average speed in synchronized flow further decreases, moving jams emerge in that synchronized flow. In this case, the branch for congested traffic exhibits a tendency towards the line \(J\).

(vi) At upstream locations within a congested pattern where wide moving jams occur the branch for congested traffic is related to the line \(J\) in the flow–density plane.

(vii) The branch of congested traffic of the empirical fundamental diagram has the asymptotic behavior to the line \(J\) at higher densities. This is because wide moving jams emerge more likely in synchronized flow of higher densities. In the region of a congested pattern, where wide
moving jams propagate, the branch $C$ almost coincides with the line $J$. This result is associated with the physical sense and the definition of the line $J$: each average point of measured data related to the wide moving jam outflow must lie on the line $J$ in the flow–density plane. This asymptotic behavior of the branch of congested traffic of empirical fundamental diagrams is important for GPs and EPs where wide moving jams occur.

(viii) Foreign wide moving jam propagation through a congested pattern can lead to the branch of congested traffic of the empirical fundamental diagram that asymptotically approaches the line $J$ at higher densities in the flow–density plane.

(ix) The above conclusions mean that the empirical fundamental diagram qualitatively depends on the freeway location where this diagram is measured. Secondly, at a given freeway location the empirical fundamental diagram qualitatively depends on the congested pattern type that is currently formed. The type of congested pattern at the same freeway bottleneck can depend on traffic demand and the initial conditions. Thus, at the given freeway location the qualitative form of the empirical fundamental diagram can also depend on traffic demand and the initial conditions.

(x) Correlation between spatiotemporal features of congested patterns and the form of the branch for congested traffic on the empirical fundamental diagram explains the nature of the reversed-\(\lambda\), the inverted-V, and inverted-U empirical fundamental diagrams.
Part III

Microscopic Three-Phase Traffic Theory
16 Microscopic Traffic Flow Models
for Spatiotemporal Congested Patterns

16.1 Introduction

Three-phase traffic theory, which is a qualitative theory, explains the main empirical features of phase transitions in traffic and of spatiotemporal congested patterns *without any mathematical traffic flow models* (Chaps. 4–8). This shows a general character of these theoretical results that should be regardless of methods and approaches of mathematical traffic flow modeling.

The main aim of this Part III is to show these qualitative theoretical results of the three-phase traffic theory based on a microscopic traffic flow modeling. This is possible because some microscopic models based on three-phase traffic theory have recently been developed [329–331].

Results presented in this Part III are based on simulations of the models [329–331] that can show and predict F→S→J transitions and other empirical features of spatiotemporal congested patterns discussed in Sect. 2.4 and Part II.

It should be noted that the models [329–331] are only a particular mathematical realization of the basic behavioral assumptions of the three-phase traffic theory (Sect. 8.6). Thus, it can be expected that in the future many other different model approaches based on three-phase traffic theory will be proposed and developed that can show and predict these empirical results (see footnote 4 of Sect. 4.3.4). For this reason, rather than to make a consideration of different solutions of the models [329–331] in this Part III only those results are presented that are necessary for explanation of main empirical congested pattern features. In this chapter, we discuss traffic flow models of [329–331] without discussion of results of simulations. The physics of phase transitions and traffic congested patterns will be considered separately in Chaps. 17–19.

The first microscopic three-phase traffic theory that can reproduce empirical F→S→J transitions (Chap. 10) and empirical spatiotemporal congested patterns was developed in 2002 by Kerner and Klenov [329]. They proposed a microscopic spatial continuum and discrete-time traffic flow model in which general rules of vehicle motion are in accordance with the fundamental hypothesis of the three-phase traffic theory [207–210]: steady state model solutions cover a two-dimensional (2D) region in the flow–density plane. The upper boundary of this 2D region is related to a safe driver speed. The low boundary of the 2D region of steady state model solution in the flow–density
plane is related to a synchronization distance. In addition, the model satisfies other behavioral assumptions of the three-phase traffic theory (Sect. 8.6), in particular for driver time delays. This enables us to reproduce empirical features of phase transitions and the congested pattern diagram at freeway bottlenecks of the three-phase traffic theory (Fig. 7.13).  

Some months later, Kerner, Klenov, and Wolf developed a cellular automata (CA) approach to three-phase traffic theory (the KKW CA models) [331]. The KKW CA models use the general rules of vehicle motion as introduced in [329]. However, the KKW CA models are much simpler than the spatial continuum models [329,330].

Recently Davis has proposed a new microscopic model based on three-phase traffic theory where the upper boundary of a 2D region of steady state model solutions in the flow–density plane is related to a desired driver speed [477]. The model of Davis can describe an F→S transition at on-ramps that is very similar to the F→S transition found earlier in [329,331]. Lee et al. [478] have proposed a CA model based on three-phase traffic theory. In this model, some hypotheses of the three-phase traffic theory postulated in [218], in particular, the hypotheses about the diagram of congested patterns at an on-ramp bottleneck and the pattern types have been used (Chap. 7). As a result, the model of Lee et al. exhibits some features of SPs and GPs, which have been found earlier in [329,331]. Jiang and Wu proposed a new CA traffic flow model in the context of three-phase traffic theory with a specific mathematical formulation for the fundamental hypothesis of this theory [479]. The model of Jiang and Wu [479] shows SPs and GPs as well as the diagram of congested patterns at an on-ramp bottleneck postulated in the three-phase traffic theory [218]. Features of these congested patterns and the pattern diagram are similar to those found earlier in [329,331].

It should be noted that both the spatial continuum and discrete-time models [329,330] and the KKW CA models [331] based on three-phase traffic theory have used some ideas of traffic flow models introduced and developed in the context of the fundamental diagram approach. In particular, the related theoretical ideas for the description of driver safety and security conditions, different driver time delays, mathematical description of model fluctuations, as well as lane changing rules developed by Gipps [17], Nagel, Schreckenberg, Schadschneider, Santen, and co-workers [364,373,374,378,426,429,481], Takayasu and Takayasu [428], Krauß et al. [377], and other groups are also very important elements of the models considered below (see for more detail the related references in [329-331]).

1 Recently some macroscopic traffic flow models have also been introduced that use one or several hypotheses of the three-phase traffic theory [207–210]. They include models by Klar, Greenberg, Rascler, Materne, and co-workers [472–475] and Colombo [476]. The development of a macroscopic three-phase traffic theory that can explain empirical F→S→J transitions (Chap. 10) and empirical congested patterns (different SPs and GPs, Chap. 9) is still a challenge for theoreticians.
To make the microscopic three-phase traffic flow theory more clear for readers who do not work in the field of mathematical traffic flow modeling, we view a simplified consideration of the models based on three-phase traffic theory below. In this consideration, some model peculiarities as discussed in the original papers [329–331] will be omitted.

Because the KKW cellular automata models are more simple ones, we will start with the consideration of “general rules of vehicle motion” of [329] in the application to the KKW CA models. Then we consider the spatial continuum and discrete-time model [330]. This model exhibits more realistic features of congested patterns. This enables us to explain both qualitative and quantitative features of empirical congested patterns discussed in Part II of this book.

16.2 Cellular Automata Approach to Three-Phase Traffic Theory

16.2.1 General Rules of Vehicle Motion

The starting point of the KKW CA models is the basic set of rules from the Kerner–Klenov model [329] that provides a 2D-region of steady states. These basic rules can be written for the KKW CA models in the form

\[
\begin{align*}
V_{n+1} &= \max(0, \min(V_{\text{free}}, v_{s,n}, v_{c,n})), \\
X_{n+1} &= x_n + v_{n+1}\tau, \\
v_{c,n} &= \begin{cases} 
V_n + a\tau & \text{for } x_{\ell,n} - x_n > D_n, \\
v_n + a\tau \text{sign}(v_{\ell,n} - v_n) & \text{for } x_{\ell,n} - x_n \leq D_n,
\end{cases}
\end{align*}
\]

(16.1)

(16.2)

where \(\text{sign}(x)\) is 1 for \(x > 0\), 0 for \(x = 0\) and -1 for \(x < 0\); \(v_n\) and \(x_n\) are the vehicle speed and the space coordinate of the vehicle front, respectively; the index \(n\) corresponds to the discrete time \(t = n\tau, n = 0, 1, 2, \ldots\); \(\tau\) is the time step; \(V_{\text{free}}\) is the maximum speed of vehicles in free flow that is assumed to be the same for all vehicles; the lower index \(c\) marks functions (or values) related to the vehicle in front of the vehicle at \(X_n\), the preceding vehicle, i.e., \(v_{\ell,n}\) and \(x_{\ell,n}\) are the speed and the space coordinate of the preceding vehicle, respectively; \(D_n = D(v_n)\) is a synchronization distance; \(a\) is a vehicle acceleration, which is assumed to be the same for all vehicles and independent of time; \(v_{s,n}\) is a safe speed that must not be exceeded in order to avoid collisions.

In general, the safe speed \(v_{s,n}\) depends on the space gap between vehicles \(g_n = x_{\ell,n} - x_n - d\) and on the speed of the preceding vehicle \(v_{\ell,n}\), where \(d\) is the vehicle length (assumed to be the same for all vehicles); the length \(d\) includes

\[2\] Because in the model the speed \(v_{\text{free}}\) does not depend on density we use a different designation in comparison with the designation \(v^{(\text{free})}\) of the speed in free flow in the three-phase traffic theory that is a function of density (curve \(F\) in Fig. 4.6).
the minimum space gap to the preceding vehicle within a wide moving jam. However, for the sake of comparison we choose the same expression for the safe speed as those in the standard Nagel–Schreckenberg CA model [364]:

\[ v_{s,n} = g_n / \tau . \] (16.3)

The crucial difference compared to previous CA models (e.g, [32, 364, 373, 374, 378, 426, 427, 429]) is that the acceleration behavior given by \( v_{c,n} \) (16.2) (the rule of “speed change”) depends on whether the preceding vehicle is within the synchronization distance \( D_n \) or further away [329]. The rules (16.2) [329,331] are some of the possible microscopic approaches to the modeling of a driver’s behavior when the driver approaches synchronized flow or the driver is within synchronized flow, i.e., these rules describe human expectation of the related local driving conditions. At sufficiently large distances from the preceding vehicle, the vehicle simply accelerates with acceleration \( a \). However, if the driver cannot pass the preceding vehicle, then within the synchronization distance the vehicle tends to adjust its speed to the preceding vehicle, i.e., it decelerates if it is faster, and accelerates if it is slower than the preceding vehicle.

This vehicle deceleration should not be confused with braking for safety purposes (i.e., in order not to exceed \( v_{s,n} \)). In practice the speed adjustment within the synchronization distance can often be achieved without braking simply as a result of the rolling friction of the wheels with the road. Thus, this deceleration is typically less than the braking capability of a vehicle. In accordance with (16.2), the speed change per time step due to either vehicle deceleration or acceleration within the synchronization distance is given by

\[ \Delta v_n = a \tau \text{sign}(v_{c,n} - v_n) . \] (16.4)

We want to emphasize that the rule (16.2) decouples the speed and the space gap between vehicles for dense traffic. This can be seen by assuming that vehicles drive behind each other with the same speed \( v \). According to (16.1)–(16.2), neither the speed nor the space gaps will change, provided all the distances are somewhere between the safe distance \( d + v \tau \) and the synchronization distance \( D(v) \geq d + v \tau \). There is neither a speed-dependent distance which individual drivers prefer, nor is there a distance-dependent optimal speed. This is the principal conceptual difference between three-phase traffic theory and the fundamental diagram approach, and this is the reason, why in three-phase traffic theory the steady states fill a two-dimensional region in the flow–density plane (Sect. 4.3 and item (i) of Sect. 8.6), while in the fundamental diagram approach they lie on a curve (on the theoretical fundamental diagram, Sect. 3.1).

### 16.2.2 Synchronization Distance

The conditions (16.1), (16.2) are the basis of the cellular automata models under consideration. It will be shown that this enables us different formulations
for fluctuations, acceleration, deceleration, and for the synchronization distance $D_n$ that all lead to qualitatively the same features of phase transitions and congested patterns as well as the same diagram of these patterns as those postulated in the three-phase traffic theory [205, 209, 217, 218] and in agreement with the model of Kerner and Klenov.

In particular, let us consider two different formulations for the dependence of the synchronization distance $D_n$ on vehicle speed. In the first formulation, the synchronization distance $D_n$ in (16.2) is a linear function of vehicle speed:

$$D(v_n) = d_1 + kv_n \tau .$$

(16.5)

In the second formulation, the synchronization distance $D_n$ in (16.2) is a nonlinear function of vehicle speed:

$$D(v_n) = d + v_n \tau + \phi v_n^2 / 2a .$$

(16.6)

In (16.5) and (16.6) $d_1$, $k$, and $\phi$ are positive constants. Both formulations lead to 2D regions of steady states in the flow–density plane.

16.2.3 Steady States

In the KKW models belonging to three-phase traffic theory with the basic structure (16.1), (16.2), a driver within the synchronization distance $D_n$ adapts his speed to the preceding vehicle without caring what the precise gap is, as long as it is safe. This explains why there is no unique flow–density relationship for steady states in the KKW CA models.

In steady states all accelerations must be zero. Then the time-index $n$ can be dropped in the above formulae. According to (16.1)–(16.2), there are two possibilities: either the synchronization distance $D(v) < g + d$ and the speed is $v = v_{\text{free}}$, or

$$D(v) \geq g + d \quad \text{and} \quad v = v_{\ell} \leq \min(v_{\text{free}}, v_{\ell}(g, v)) .$$

(16.7)

In steady states all speeds are equal. Equality holds in (16.7) only if the space gaps $g$ are all equal. Thus, we defined steady states as the model solutions in which all vehicles move with the same distances to one another and with the same time-independent speed, i.e., the steady state solutions are time-independent and homogeneous.

The density $\rho$ and the flow rate $q$ are related to the space gap $g$ and the speed $v$ by

$$\rho = 1/(x_{\ell} - x) = 1/(g + d), \quad q = \rho v = v / (g + d) .$$

(16.8)

Because $v$ and $g$ are integers in CA models, the steady states do not form a continuum in the flow–density plane as they do in the spatial continuum model considered below. However, the inequalities of (16.7) define a two-dimensional region in the flow–density plane in which steady states exist. It
Fig. 16.1. 2D regions supporting steady states in the flow–density plane for three of the KKW models. (a) KKW model with linear synchronization distance (16.5) and \(d_1 = d\). (b) KKW model with nonlinear dependence of the synchronization distance (16.6) on vehicle speed. (c) KKW model with linear synchronization distance (16.5) and \(d_1 < d\). Taken from [331]

is limited by three boundaries (Fig. 16.1), the upper line \(U\), the lower curve \(L\), and the left line \(F\).

The left boundary \(F\) is given by \(q = \rho v_{\text{free}}\). This is free flow, where the flow rate \(q\) is not restricted by safety requirements. On the upper boundary \(U\) the flow rate is determined by the safe speed \(v_s\). For example, inserting (16.3) in (16.8) it is given by

\[ q = \rho v_s = (1 - \rho d)/\tau . \]  

(16.9)

The lower boundary \(L\) is determined by the synchronization distance \(D\): a steady state with density \(\rho\) and a speed \(v < v_{\text{free}}\) requires that \(D(v) \geq 1/\rho\) with equality on the lower boundary \(L\). For example, using (16.5) with \(d_1 = d\) one obtains (Fig. 16.1a)

\[ q = (1 - \rho d)/k\tau . \]  

(16.10)

In the second model (16.6), the lower boundary \(L\) is a nonlinear curve (Fig.16.1b):

\[ q = \frac{\bar{\rho}}{\tau} \left( \sqrt{1 + \frac{2}{\bar{\rho}(1 - \rho d)} - 1} \right) , \]  

with \(\bar{\rho} = \frac{\rho_0^2 a}{\phi}\).  

(16.11)

This curve has the upper line \(U\) as a tangent at \(\rho = \rho_{\text{max}} = 1/d\). This enables us to reproduce qualitatively the diagram of congested patterns of three-phase traffic theory with simpler fluctuations than what is needed in
the case of the linear synchronization distance (16.5) with $d_1 = d$ [331]. However, if the parameter $d_1$ is chosen to be smaller than $d$ in (16.5), the line $L$ intersects the line $U$ before the jam density $\rho_{\text{max}}$ is reached (Fig. 16.1c). In this case, if the difference $d - d_1$ is chosen in a proper way, the fluctuations in the model with linear synchronization distance (16.5) can be as simple as for the nonlinear $D$ (16.6), in order to lead qualitatively to the same features.

16.2.4 Fluctuations of Acceleration and Deceleration in Cellular Automata Models

As in some other traffic flow models [17, 364, 373, 374, 378, 426, 429, 481] (see references in the reviews [32,33,35,36,38]), in the models under consideration, we use variables that are stochastic functions. At the first step, a preliminary vehicle speed of each vehicle $v_{n+1}$ is

$$v_{n+1} = v_{n+1}, \quad (16.12)$$

where $v_{n+1}$ is calculated based on the system of the dynamic equations (16.1)–(16.2). At the second step, a fluctuation $a\tau \chi_n$ (to be specified below) is added to the value $\tilde{v}_{n+1}$ calculated from the first step. Finally, the speed $v_{n+1}$ at the time $n + 1$ is calculated by

$$v_{n+1} = \max(0, \min(\nu_{\text{free}}, \tilde{v}_{n+1} + a\tau \chi_n, v_n + a\tau, v_{s,n})). \quad (16.13)$$

This means that the stochastic contribution $a\tau \chi_n$ may neither lead to a speed smaller than zero nor to a speed larger than what the deterministic acceleration $a$ would give, taking the limitations by $\nu_{\text{free}}$ and $v_{s,n}$ into account.

We implement the fluctuation $\chi_n$ in (16.13) as

$$\chi_n = \begin{cases} -1 & \text{if } r < p_b, \\ 1 & \text{if } p_b \leq r < p_b + p_a \\ 0 & \text{otherwise,} \end{cases} \quad (16.14)$$

where $r = \text{rand}(0,1)$ denotes a random number uniformly distributed between 0 and 1. $p_a + p_b \leq 1$ must be satisfied.

The probability of random deceleration $p_b$ in (16.14) is taken as a decreasing function of vehicle speed $v_n$:

$$p_b(v_n) = \begin{cases} p_{b0} & \text{if } v_n = 0 \\ p_{b1} & \text{if } v_n > 0, \end{cases} \quad (16.15)$$

where $p_{b1}$ and $p_{b0} > p_{b1}$ are constants. This corresponds to the behavioral assumptions of item (viii)–(xi) of Sect. 8.6. In particular, at $v_n = 0$ the condition (16.15) is related to the slow-to-start rules first introduced by Takayasu and Takayasu [428] and by Barlovic et al. [378]: vehicles accelerate at the downstream front of a wide moving jam with the mean time delay.
\[ \tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p_{b0}}. \]  

This provides wide moving jam propagation through free and synchronized flows while maintaining the same velocity \( v_g \) of the downstream jam front. At the downstream front of synchronized flow (this front separates synchronized flow upstream and free flow downstream) vehicles accelerate with the mean time delay

\[ \tau_{\text{del}, \text{syn}}^{(a)} = \frac{\tau}{1 - p_{b1}}. \]  

Because \( p_{b0} \) is taken higher than \( p_{b1} \), we obtain

\[ \tau_{\text{del}}^{(a)} < \tau_{\text{del}, \text{syn}}^{(a)}. \]  

The probability of random acceleration \( p_a \) in (16.14) is also taken as a decreasing function of vehicle speed \( v_n \):

\[ p_a(v_n) = \begin{cases} 
  p_{a1} & \text{if } v_n < v_p \\
  p_{a2} & \text{if } v_n \geq v_p,
\end{cases} \]  

where \( v_p, p_{a1}, \) and \( p_{a2} < p_{a1} \) are constants. This simulates the effect that a vehicle moving at a low speed in dense flow tends to approach the preceding vehicle (item (xii) of Sect. 8.6). Indeed, according to (16.1)-(16.2), (16.13), (16.14), if the probability \( p_a \) of acceleration is high, the effect of adapting of the vehicle speed to the speed of the preceding vehicle is weak: with the probability \( p_a \) the vehicle does not reduce its speed. The vehicle can do this until it reaches the safe space gap \( v_n \tau \). Note that the tendency to minimize the space gap at a low speed can lead in particular to the pinch effect in synchronized traffic flow, i.e., to the self-compression of synchronized flow at lower vehicle speeds with spontaneous moving jam emergence (Sect. 12.2).

The tendency to minimize the space gap at low speeds is automatically built into the KKW CA models if the lower boundary \( L \) approaches the upper line \( U \) as in Figs. 16.1b,c, because then vehicle deceleration is determined by safety conditions at a small speed. Indeed it turns out that the speed dependence (16.19) of \( p_a \) is not required in these cases for realistic modeling. Therefore, we choose the probability \( p_a \) of acceleration in (16.14) to be a constant in the model variants of Figs. 16.1b,c.

As mentioned in Sects. 4.3.3 and 4.3.4, random fluctuations of acceleration and deceleration destroy steady states of synchronized flow: in reality, after an F\( \rightarrow \)S transition has occurred in an initial free flow rather than steady states some dynamic spatiotemporal synchronized flow states appear \cite{329,331}. However, these complex dynamic synchronized flow states can be close to steady states of synchronized flow. These dynamic states exhibit features of steady states of synchronized flow discussed in Chaps. 5–8.
16.2.5 Boundary Conditions and Model of On-Ramp

The usual choice of the time step $\tau$ in CA models of traffic has been used: $\tau = 1$ s [364]. Like in [426] we use a small-scale discretization of space: the length of cells is chosen equal to $\delta x = 0.5$ m. This leads to a speed discretization in units of $\delta v = 1.8$ km/h. Hence two vehicles are only considered as moving with different speeds, if this difference is equal to (or larger) than $\delta v$.

For simulations of congested patterns on a homogeneous single-lane road, cyclic boundary conditions have been used. The single-lane homogeneous road has the length 60000 cells (30 km). We checked that all qualitative results remain the same, if open boundary conditions are used and the road is long enough.

For all simulations of congested patterns at an on-ramp bottleneck a single-lane road of 100 km length (200000 cells) with open boundary conditions is used. The reference point $x = 0$ is placed at the distance 20 km from the end of the road, so that it begins at the coordinate $x = x_b = -80$ km. The on-ramp starts at the point $x = 16$ km (32000 cells) and its merging region was 0.3 km long (600 cells).

At the beginning of the road new vehicles are generated one after another at the time
\[ t^{(k')} = \tau \left\lceil k' \frac{\tau_{in}}{\tau} \rightceil, \quad k' = 1, 2, \ldots, \] (16.20)
where $\tau_{in} = 1/q_{in}$, $q_{in}$ is the flow rate in the incoming boundary flow per lane, $\lceil z \rceil$ denotes the nearest integer greater than or equal to $z$. A new vehicle appears on the road at the time (16.20) only if the distance from the beginning of the road ($x = x_b$) to the position $x = x_{\ell,n}$ of the farthest upstream vehicle on the road is not smaller than a safe distance $\ell_{safe}$, i.e.,
\[ x_{\ell,n} - x_b \geq \ell_{safe}, \] (16.21)
where $\ell_{safe} = v_{\ell,n} \tau + d$, $n = t^{(k')}/\tau$. Otherwise, the latter condition is checked at time $(n + 1)\tau$ that is the next one to time $t^{(k')}$ (16.20), and so on, until the condition (16.21) is satisfied. Then a new vehicle appears on the road. After this occurs, the number $k'$ in (16.20) is increased by 1.

The speed $v_n$ and coordinate $x_n$ of the new vehicle are set to
\[ v_n = v_{\ell,n}, \quad x_n = \max(x_b, x_{\ell,n} - \max([v_n \tau_{in}], \ell_{safe})), \] (16.22)
where $\lfloor z \rfloor$ denotes the integer part of $z$. The flow rate $q_{in}$ is chosen to have the value $v_{free}\tau_{in}$ integer. In the initial state ($n = 0$), all vehicles have the maximum speed $v_n = v_{free}$ and they are positioned at space intervals $x_{\ell,n} - x_n = v_{free}\tau_{in}$.

When the vehicle reaches the end of the road, it is removed from the road. Before this, the farthest downstream vehicle maintains its speed.

For simulation of the on-ramp bottleneck two consecutive vehicles on the main road are chosen randomly within the on-ramp merging region of length
The coordinates of these vehicles are denoted by $x_n^+$ and $x_n^-$, respectively, where $x_n^+ > x_n^-$. A vehicle that merges onto the main road is placed at the midpoint between them at the coordinate $x_n^{(m)} = \frac{(x_n^+ + x_n^- + \delta x)}{2}$. As in [405], this merging vehicle takes the speed of the preceding vehicle $v_n^+$. The vehicle merges onto the main road only if the space gap $x_n^+ - x_n^- - d$ between the two vehicles on the main road exceeds some value $g_{on}^{(\text{min})} = \lambda v_n^+ + d$, i.e.,

$$x_n^+ - x_n^- - d > \lambda v_n^+ + d,$$  \hspace{1cm} (16.23)

where $\lambda$ is a parameter. If the condition (16.23) is not satisfied, the next pair of consecutive vehicles within the merging region on the main road is chosen randomly, and so on. If the condition (16.23) is not satisfied for all pairs of vehicles, the vehicle does not merge onto the main road at time step $n$ and it tries to do it at the next time step according to the same procedure. New vehicles at the on-ramp are generated at time $t^{(k')} = T_{k'\text{on}}/T_{\text{on}}$, $k' = 1, 2, \ldots$, where $T_{\text{on}} = 1/q_{\text{on}}$, $q_{\text{on}}$ is the flow rate to the on-ramp.

**16.2.6 Summary of Model Equations and Parameters**

We consider in this book results of simulations with the KKW CA model related to Fig. 16.1a only (results of simulations of other KKW CA models can be found in [331]). For this reason, we summarize the equations that describe this model in Table 16.1. In addition, we provide two tables containing a list of symbols (Tables 16.2 and 16.3), and typical values for model parameters (Table 16.4).

**16.3 Continuum in Space Model Approach to Three-Phase Traffic Theory**

**16.3.1 Vehicle Motion Rules**

In a spatial continuum and discrete-time single-lane model of Kerner and Klenov, the general rules of vehicle motion are [329]:

$$v_{n+1} = \max(0, \min(v_{\text{free}}, v_{c,n}, v_{s,n})),$$ \hspace{1cm} (16.24)

$$x_{n+1} = x_n + v_{n+1}\tau,$$ \hspace{1cm} (16.25)

$$v_{c,n} = \begin{cases} v_n + \Delta_n & \text{at } x_{\ell,n} - x_n \leq D_n \\ v_n + a_n\tau & \text{at } x_{\ell,n} - x_n > D_n \end{cases},$$ \hspace{1cm} (16.26)

where

$$\Delta_n = \max(-b_n\tau, \min(a_n\tau, v_{\ell,n} - v_n)).$$ \hspace{1cm} (16.27)

As in the KKW CA models of Sect. 16.2, in (16.24)–(16.27), $v_n$ and $x_n$ are the speed and space coordinate of the vehicle; the index $n$ corresponds to the
Table 16.1. KKW CA model

<table>
<thead>
<tr>
<th>Dynamic part of KKW model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{v}<em>{n+1} = \max(0, \min(v</em>{\text{free}}, v_{s,n}, v_{c,n}))$, $g_n = x_{\ell,n} - x_n - d$,</td>
</tr>
<tr>
<td>$v_{s,n} = g_n/\tau$, $v_{c,n} = \begin{cases} v_n + a\tau &amp; \text{for } g_n &gt; D_n - d, \ v_n + a\tau \text{sgn}(v_{\ell,n} - v_n) &amp; \text{for } g_n \leq D_n - d, \end{cases}$</td>
</tr>
<tr>
<td>$v_{\text{free}}$, $d$, $\tau$, and $a$ are constants.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stochastic part of KKW model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{n+1} = \max(0, \min(\tilde{v}<em>{n+1} + a\tau \chi_n, v_n + a\tau, v</em>{\text{free}}, v_{s,n}))$, $x_{n+1} = x_n + v_{n+1}\tau$,</td>
</tr>
<tr>
<td>$\chi_n = \begin{cases} -1 &amp; \text{if } r &lt; p_b, \ 1 &amp; \text{if } p_b \leq r &lt; p_b + p_a, \ 0 &amp; \text{otherwise}. \end{cases}$</td>
</tr>
</tbody>
</table>

Specifications of synchronization distance $D_n$ and noise $\chi_n$

$D_n = d + kv_n\tau$, $p_b(v_n) = \begin{cases} p_{b0} & \text{if } v_n = 0 \\ p_{b1} & \text{if } v_n > 0 \end{cases}$, $p_a(v_n) = \begin{cases} p_{a1} & \text{if } v_n < v_p \\ p_{a2} & \text{if } v_n \geq v_p \end{cases}$, constant parameters: $k$, $p_{b0}$, $p_{b1}$, $p_{a1}$, $p_{a2}$, $v_p$.

Model of on-ramp

Condition for vehicle merging: $x_n^+ - x_n^- - d > \lambda v_n^+ + d$, Parameters of new vehicle: $x_n^{(m)} = \lfloor (x_n^- + x_n^+ + 1)/2 \rfloor$, $v_n = v_n^+$, $\lambda$ is a constant.

discrete time $t = n\tau$, $n = 0, 1, 2, \ldots$; $\tau$ is the time step; $v_{\text{free}}$ is the maximum speed in free flow that is constant; $v_{s,n}$ is the safe speed; the lower index $\ell$ marks values and variables related to the preceding vehicle; all vehicles have the same length $d$. The vehicle length $d$ includes the minimum space gap between vehicles within a wide moving jam. Values $a_n \geq 0$ and $b_n \geq 0$ in (16.26), (16.27) restrict changes in speed per time step when the vehicle accelerates or adjusts the speed to that of the preceding vehicle.

16.3.2 Speed Adaptation Effect Within Synchronization Distance

Equations (16.26), (16.27) describe the adaptation of the vehicle speed to the speed of the preceding vehicle, i.e., the speed adaptation effect in synchronized flow (Sect. 4.3.2). This vehicle speed adaptation takes place when the vehicle cannot pass the preceding vehicle, within the synchronization distance $D_n$: at $x_{\ell,n} - x_n \leq D_n$ the vehicle tends to adjust its speed to the preceding vehicle. This means that the vehicle decelerates if $v_n > v_{\ell,n}$, and accelerates if $v_n < v_{\ell,n}$ [329]. This speed adaptation effect within the synchronization distance is related to the fundamental hypothesis of the three-phase traffic theory: hypothetical steady states of synchronized flow cover a 2D region in the flow–density (Fig. 16.1a).
Table 16.2. List of symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>time discretization interval</td>
</tr>
<tr>
<td>$\delta x$</td>
<td>space discretization length</td>
</tr>
<tr>
<td>$\delta v = \delta x / \tau$</td>
<td>speed discretization unit</td>
</tr>
<tr>
<td>$n = 0, 1, 2, \ldots$</td>
<td>number of time steps</td>
</tr>
<tr>
<td>$\bar{v}_n$</td>
<td>vehicle speed at time step $n$ without fluctuating part</td>
</tr>
<tr>
<td>$v_n$</td>
<td>vehicle speed at time step $n$</td>
</tr>
<tr>
<td>$v_{\ell,n}$</td>
<td>speed of preceding vehicle at time step $n$</td>
</tr>
<tr>
<td>$v_{s,n}$</td>
<td>safe speed at time step $n$</td>
</tr>
<tr>
<td>$v_{\text{free}}$</td>
<td>maximum speed (free flow)</td>
</tr>
<tr>
<td>$x_n$</td>
<td>vehicle position at time step $n$</td>
</tr>
<tr>
<td>$x_{\ell,n}$</td>
<td>position of preceding vehicle at time step $n$</td>
</tr>
<tr>
<td>$d$</td>
<td>vehicle length</td>
</tr>
<tr>
<td>$g_n = x_{\ell,n} - x_n - d$</td>
<td>space gap (front to end distance) at time step $n$</td>
</tr>
<tr>
<td>$D_n$</td>
<td>synchronization distance at time step $n$</td>
</tr>
<tr>
<td>$a$</td>
<td>vehicle acceleration</td>
</tr>
<tr>
<td>$\chi_n$</td>
<td>speed fluctuation at time step $n$,</td>
</tr>
<tr>
<td>$p_a$</td>
<td>probability of random acceleration</td>
</tr>
<tr>
<td>$p_b$</td>
<td>probability of random deceleration</td>
</tr>
<tr>
<td>$r$</td>
<td>random number uniformly distributed between 0 and 1</td>
</tr>
<tr>
<td>$q_0$</td>
<td>maximum flow rate in steady states</td>
</tr>
<tr>
<td>$q_{\text{max}}$</td>
<td>maximum flow rate in free flow ((v = v_{\text{free}}))</td>
</tr>
<tr>
<td>$\rho_{\text{max}} = 1/d$</td>
<td>density within wide moving jam</td>
</tr>
<tr>
<td>$v_g$</td>
<td>velocity of downstream front of wide moving jam</td>
</tr>
<tr>
<td>$q_{\text{out}}$</td>
<td>flow rate in free flow formed by wide moving jam outflow</td>
</tr>
<tr>
<td>$\rho_{\text{min}}$</td>
<td>density in free flow relative to flow rate $q_{\text{out}}$</td>
</tr>
<tr>
<td>$q'_{\text{lim}}$</td>
<td>limiting flow rate in pinch region of general pattern</td>
</tr>
<tr>
<td>$q'_{\text{out}}$</td>
<td>flow rate in free flow formed by congestion pattern outflow downstream of bottleneck</td>
</tr>
<tr>
<td>$q_{\text{in}}$</td>
<td>flow rate of incoming flow onto main road</td>
</tr>
</tbody>
</table>

In the general rules (16.24)–(16.27), the synchronization distance $D_n$ depends on the vehicle speed $v_n$ and on the speed of the preceding vehicle $v_{\ell,n}$

$$D_n = D(v_n, v_{\ell,n}) \, ,$$

(16.28)

where the function $D(u, w)$ is chosen as

$$D(u, w) = d + \max(0, k\tau u + \phi_0 a^{-1} u(u - w)) \, ,$$

(16.29)

$k > 1$ and $\phi_0$ are constants, $a$ is the maximum acceleration. If $v_n = v_{\ell,n}$, the synchronization distance $D_n$ is $d + kv_n\tau$. This corresponds to a fixed time gap $k\tau$. If $v_n > v_{\ell,n}$, the distance $D_n$ increases and vice versa.
16.3 Continuum in Space Model Approach

Table 16.3. List of symbols (model of on-ramp)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(q_{\text{on}})</td>
<td>flow rate to on-ramp</td>
</tr>
<tr>
<td>(L_m)</td>
<td>length of on-ramp merging region</td>
</tr>
<tr>
<td>(x^+_n, x^-_n)</td>
<td>coordinates of preceding and trailing vehicles in pair of consecutive vehicles within merging region on main road at time step (n), respectively</td>
</tr>
<tr>
<td>(v^+_n)</td>
<td>speed of preceding vehicle in pair of consecutive vehicles within merging region on main road at time step (n)</td>
</tr>
<tr>
<td>(g_{\text{on}}^{(\text{min})})</td>
<td>minimum space gap between two vehicles within merging region on main road at which new vehicle can merge onto main road between them</td>
</tr>
<tr>
<td>(x^{(m)}_n)</td>
<td>coordinate of midpoint between preceding and trailing vehicles within merging region</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>parameter related to (g_{\text{on}}^{(\text{min})})</td>
</tr>
</tbody>
</table>

Table 16.4. Model parameters and characteristic values

<table>
<thead>
<tr>
<th>Discretization units</th>
<th>(\tau = 1 \text{ s}, \delta x = 0.5 \text{ m}, \delta v = \delta x/\tau = 1.8 \text{ km/h}, a = \delta v/\tau = 0.5 \text{ m/s}^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model parameters</td>
<td>(v_{\text{free}} = 108 \text{ km/h} = 60 \delta v, d = 7.5 m = 15 \delta x, k = 2.55, p_{b0} = 0.425, p_{b1} = 0.04, p_{a1} = 0.2, p_{a2} = 0.052, v_p = 50.4 \text{ km/h} = 28 \delta v, \lambda = 0.55\tau)</td>
</tr>
<tr>
<td>Model results</td>
<td>(v_g = -15.5 \text{ km/h}, q_{\text{out}} = 1810 \text{ vehicles/h}, \rho_{\text{min}} = 16.76 \text{ vehicles/km}, q_0 = 2880 \text{ vehicles/h}, \tilde{q}^{(\text{free})} \approx 2400 \text{ vehicles/h}, \tilde{q}^{(\text{pinch})} \approx 1150 \text{ vehicles/h})</td>
</tr>
</tbody>
</table>

16.3.3 Motion State Model for Random Acceleration and Deceleration

Random Acceleration and Deceleration

As in Sect. 16.2.4, at the first step, the preliminary speed of each vehicle is \(\tilde{v}_{n+1} = v_{n+1}\), where \(v_{n+1}\) is calculated based on (16.24), (16.26), (16.27). At the second step, a noise component \(\xi_n\) is added to the calculated speed \(\tilde{v}_{n+1}\) and then the final value of the speed \(v_{n+1}\) at time step \(n + 1\) is found from the condition introduced in [331]

\[
v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}_{n+1} + \xi_n, v_n + a\tau, v_{\text{s,n}})). \tag{16.30}
\]

In the motion state model [330], random deceleration and acceleration are applied depending on whether the vehicle decelerates or accelerates, or else
maintains its speed:

\[
\xi_n = \begin{cases} 
-\xi_b & \text{if } S_{n+1} = -1 \\
\xi_a & \text{if } S_{n+1} = 1 \\
0 & \text{if } S_{n+1} = 0 , 
\end{cases}
\]  

(16.31)

where \( \xi_b \) and \( \xi_a \) are random sources for deceleration and acceleration, respectively; \( S \) in (16.31) denotes the state of motion (\( S_{n+1} = -1 \) represents deceleration, \( S_{n+1} = 1 \) acceleration, and \( S_{n+1} = 0 \) motion at nearly constant speed)

\[
S_{n+1} = \begin{cases} 
-1 & \text{if } \ddot{v}_{n+1} < v_n - \delta \\
1 & \text{if } \ddot{v}_{n+1} > v_n + \delta \\
0 & \text{otherwise } , 
\end{cases}
\]  

(16.32)

where \( \delta \) is a constant (\( \delta \ll a\tau \)).

The random components in (16.31) are “impulsive”:

\[
\xi_b = a\tau \Theta(p_b - r) 
\]  

(16.33)

and

\[
\xi_a = a\tau \Theta(p_a - r) .
\]  

(16.34)

In (16.33), (16.34), \( p_b \) and \( p_a \) are probabilities of random deceleration and acceleration, respectively. \( r = \text{rand}(0,1) \), i.e., this is an independent random value uniformly distributed between 0 and 1. The function \( \Theta(z) \) has the following meaning: \( \Theta(z) = 0 \) at \( z < 0 \) and \( \Theta(z) = 1 \) at \( z \geq 0 \).

**Random Time Delays**

To simulate a time delay either in vehicle acceleration or in vehicle deceleration, \( a_n \) and \( b_n \) in (16.27) are taken as the following stochastic functions

\[
a_n = a\Theta(P_0 - r_1) ,
\]  

(16.35)

\[
b_n = a\Theta(P_1 - r_1) ,
\]  

(16.36)

\[
P_0 = \begin{cases} 
p_0 & \text{if } S_n \neq 1 \\
1 & \text{if } S_n = 1 ,
\end{cases}
\]  

(16.37)

\[
P_1 = \begin{cases} 
p_1 & \text{if } S_n \neq -1 \\
p_2 & \text{if } S_n = -1 ,
\end{cases}
\]  

(16.38)

where \( r_1 = \text{rand}(0,1) \), i.e., this is a random number uniformly distributed between 0 and 1; values \( p_0, p_1, \) and \( p_2 \) are determined in Sect. 16.3.6. The physical sense of the functions \( P_0 \) and \( P_1 \) in (16.37) and (16.38) is as follows. The function \( P_0 \) in (16.37) determines the probability \( \psi_a \) of a random time delay in vehicle acceleration at time step \( n + 1 \) corresponding to

\[
\psi_a = 1 - P_0 ,
\]  

(16.39)

whereas the function \( P_1 \) (16.38) determines the probability \( \psi_b \) of a random time delay in vehicle deceleration at time step \( n + 1 \) corresponding to

\[
\psi_b = 1 - P_1 .
\]  

(16.40)
16.3.4 Safe Speed

In the model, the safe speed \( \nu_{s,n} \) in (16.24) is chosen in the form

\[
\nu_{s,n} = \min\left(\nu_{n}^{(\text{safe})}, g_{n}/\tau + \nu_{\ell}^{(a)}\right).
\]  

(16.41)

In all simulations presented in this book, the function \( \nu_{n}^{(\text{safe})} = \nu_{n}^{(\text{safe})}(g_{n}, \nu_{\ell,n}) \) in (16.41) is related to the safe speed in the model by Krauß et al. [377] that belongs to the Gipps-class models [17]. In (16.41), \( g_{n} = x_{\ell,n} - x_{n} - d \) is the space gap; \( \nu_{\ell}^{(a)} \) is an “anticipation” speed of the preceding vehicle at the next time step that is considered below.

In the model by Krauß et al. [377], the safe speed \( \nu_{n}^{(\text{safe})}(g_{n}, \nu_{\ell,n}) \) is found from the safety condition for vehicle motion first proposed by Gipps [17]:

\[
\nu^{(\text{safe})} + X_{d}(\nu^{(\text{safe})}) \leq g_{n} + X_{d}(\nu_{\ell,n}),
\]

(16.42)

where \( \tau^{(\text{safe})} \) is a safe time gap that can be individual for drivers, \( X_{d}(u) \) is the braking distance that should be passed by the vehicle moving first with the speed \( u \) before the vehicle can come to a stop. The condition (16.42) when it is the equality enables us to find the safe speed \( \nu^{(\text{safe})} \) as a function of the space gap \( g_{n} \) and speed \( \nu_{\ell,n} \) provided \( X_{d}(u) \) is a known function. In the case when the vehicle brakes with a constant deceleration \( b \), the change in vehicle speed for each time step is \(-b\tau\) except the last time step before the vehicle comes to a stop. At the last time step the vehicle decreases its speed at the value \( b\tau\beta \) where \( \beta \) is the fractional part of \( u/b\tau \). According to formula (16.25) for the displacement of the vehicle for one time step, the braking distance \( X_{d}(u) \) is [377]

\[
X_{d}(u) = \tau ((u - b\tau) + (u - 2b\tau) + \ldots + \beta b\tau).
\]

(16.43)

Formula (16.43) can be written as [377]

\[
X_{d}(u) = b\tau^{2} \left( \alpha \beta + \frac{\alpha (\alpha - 1)}{2} \right),
\]

(16.44)

where \( \alpha \) is the integer part of \( u/b\tau \). The safe speed \( \nu^{(\text{safe})} \) as a solution of the equation (16.42) at the distance \( X_{d}(u) \) given by (16.44) and at \( \tau^{(\text{safe})} = \tau \) has been found in [377]

\[
\nu^{(\text{safe})}(g_{n}, \nu_{\ell,n}) = b\tau (\alpha_{\text{safe}} + \beta_{\text{safe}}),
\]

(16.45)

where

\[
\alpha_{\text{safe}} = \left\lfloor \frac{1}{2} \frac{X_{d}(\nu_{\ell,n}) + g_{n}}{b\tau^{2}} + \frac{1}{4} - \frac{1}{2} \right\rfloor,
\]

(16.46)
[z] denotes the integer part of a real number z.

The safe speed in the model by Krauß et al. [377] provides collisionless motion of vehicles if \( g_n \geq v_{\ell, n, \tau} \) [480]. In the model under consideration, it is assumed that in some cases, mainly due to lane changing or merging of vehicles onto the main road within the merging region of bottlenecks, the space gap \( g_n \) can become less than \( v_{\ell, n, \tau} \). In these critical situations, the collisionless motion of vehicles in the model is a result of the second term in (16.41) in which some prediction \( (v_{\ell}^{(a)}) \) of the speed of the preceding vehicle at the next time step is used. The related “anticipation” speed \( v_{\ell}^{(a)} \) is chosen as

\[
v_{\ell}^{(a)} = \max \left(0, \min \left(v_{\ell, n}^{(\text{safe})} - a\tau, v_{\ell, n} - a\tau, g_{\ell, n}/\tau\right)\right),
\]  

where \( v_{\ell, n}^{(\text{safe})} \) is the safe speed (16.45)–(16.47) for the preceding vehicle, \( g_{\ell, n} \) is the space gap in front of the preceding vehicle. Simulations have shown that formulae (16.41), (16.45)–(16.48) lead to collisionless vehicle motion over a wide range of parameters of the merging region (Sect. 16.3.9) and for chosen lane changing rules (Sect. 16.3.8).

### 16.3.5 2D Region of Steady States

As in the KKW CA models, steady states of the spatial continuum and discrete-time single-lane traffic flow model cover a 2D region in the flow-density plane [329]. Since in a steady state all vehicles have the same time-independent speed \( v \) and the same time-independent space gap \( g \), from (16.24), (16.26), (16.27) we obtain that in steady state model solutions either the space gap \( g > D(v) - d \) and the speed is \( v = v_{\text{free}} \), or the space gap \( g \) and the speed \( v \) are determined by (16.7). The latter formulae define a 2D region limited by three boundaries in the flow-density plane: the upper line \( U \), the lower curve \( L \), and the left line \( F \) (Fig. 16.1a). Similar to the KKW CA model, the left boundary \( F \) corresponds to the condition \( v = v_{\text{free}} \), the upper boundary \( U \) is determined by the safe speed \( v_s \), and the lower line \( L \) is related to the synchronization distance \( D \). Moreover, in the spatial continuum model under consideration the boundaries \( F, U, \) and \( L \) are determined by the same formulae as those in the KKW CA model (Sect. 16.2). Indeed, from the condition \( v = v_{\text{free}} \) we obtain that the formula \( q = \rho v_{\text{free}} \) applies for the line \( F \). In steady states, the synchronization distance \( D \) (16.28) is

\[
D(v) = d + kv\tau.
\]

This formula is identical to (16.5) at \( d_1 = d \). Hence, the boundary \( L \) in the flow-density plane is given by (16.10). At the upper boundary \( U \), the speed \( v \) is equal to the safe speed: \( v = v_s(g, v) \). To extract the relationship between \( v \) and \( g \) at the line \( U \) from the latter formula, we first consider
at \( v = v^\text{(safe)} = v_{\text{safe}} \) and \( \tau^\text{(safe)} = \tau \). In this case, from (16.42) we obtain that \( v^\text{(safe)} = g/\tau \). Inserting the latter expression in (16.41) and taking into account that \( v^\text{(a)}(v) \geq 0 \) (16.48), one finds the safe speed \( v_s = g/\tau \) at the boundary \( U \). The same expression follows from (16.3) for the safe speed in the CA model. Thus, in both cases at the boundary \( U \) we have \( v = g/\tau \). As a result, this boundary in the flow–density plane is given by the same formula (16.9).

Recall that in the KKW CA models where \( v \) and \( g \) are integers, steady state model solutions correspond to separate points within a 2D region in the flow–density plane. In contrast, in the spatial continuum traffic flow model, the steady states do form a continuum restricted by the boundaries \( F, U \), and \( L \) in the flow–density plane (Fig. 16.1a).

### 16.3.6 Physics of Driver Time Delays

#### Time Delay in Vehicle Acceleration

Formulae (16.35), (16.37) together with (16.24), (16.26), (16.27) simulate a time delay in vehicle acceleration that takes place if the vehicle does not accelerate \( (S_n = 1) \) at time step \( n \). After the time delay the vehicle accelerates. The mean value of this random time delay is

\[
\tau^\text{(a)}(v) = \frac{\tau}{p_0(v)}. \tag{16.50}
\]

Formula (16.50) is valid if the vehicle speed and hence the probability \( p_0(v) \) are not functions of time. Otherwise, (16.50) is only an estimation of the mean time delay in vehicle acceleration. The mean time delay \( \tau^\text{(a)}(v) \) (16.50) is assumed to be maximum for vehicles at a standstill, i.e., at \( v = 0 \). This is achieved by the choice of the probability \( p_0 \) as an increasing function \( p_0(v) \) of speed. The existence of the random time delay in vehicle acceleration and a dependence of the mean time delay \( \tau^\text{(a)}(v) \) on speed are related to behavioral assumptions of the three-phase traffic theory formulated in item (viii)-(xi) of Sect. \( 8.6 \).

The mean time delay in vehicle acceleration (16.50) explains the physics of formation of the downstream fronts of an MSP (Sect. 7.4.2) and of a wide moving jam (Sect. 3.2.6). At \( v = 0 \) formulae in (16.35), (16.37) together with (16.24), (16.26), (16.27) simulate the slow-to-start rules [378, 428]: vehicles accelerate at the downstream front of a wide moving jam with a mean time delay

\[
\tau^\text{(a)} = \tau^\text{(a)}(0) = \frac{\tau}{p_0(0)}. \tag{16.51}
\]

This gives the velocity of the downstream jam front

\[
v_g = -\frac{p_0(0)d}{\tau} \tag{16.52}
\]
corresponding to (3.10):

\[ v_g = -\frac{1}{\rho_{\text{max}} \tau_{\text{del}}^{(a)}} \]  

(16.53)

where \( \rho_{\text{max}} = d^{-1} \) is the density within a wide moving jam. In synchronized flow, (16.35), (16.37) together with (16.24), (16.26), (16.27) simulate the time delay in vehicle acceleration at the downstream front of synchronized flow:

\[ \tau_{\text{del, syn}}^{(a)} = \tau_{\text{del}}^{(a)} \left( v^{(\text{syn})} \right) = \frac{\tau}{p_0 \left( v^{(\text{syn})} \right)} , \]  

(16.54)

where \( v^{(\text{syn})} \) is a synchronized flow speed just upstream of the downstream front of synchronized flow.

**Time Delay in Vehicle Deceleration**

Let us consider the mean time delay in vehicle deceleration, \( \tau_{\text{del}}^{(d)} \), associated with the case when the preceding vehicle is slower than the vehicle and the vehicle is within the synchronization distance:

\[ v_n > v_{\ell,n} \quad \text{and} \quad x_{\ell,n} - x_n \leq D_n . \]  

(16.55)

Three different vehicle classes should be distinguished: (1) Vehicles satisfy the condition

\[ v_n > v_{s,n} . \]  

(16.56)

Then the mean time delay in vehicle deceleration is equal to the time step:

\[ \tau_{\text{del, 1}}^{(d)} = \tau . \]  

(16.57)

(2) Vehicles satisfy the condition

\[ v_n \leq v_{s,n} . \]  

(16.58)

The time delay in deceleration is simulated by (16.36), (16.38) together with (16.24), (16.26), (16.27). The vehicle that is not decelerating at time step \( n \) \((S_n \neq -1)\) starts to decelerate at time step \( n + 1 \) with probability \( p_1 \). The mean time delay in deceleration for vehicles of class (2) is

\[ \tau_{\text{del, 2}}^{(d)} = \frac{\tau}{p_1} . \]  

(16.59)

Because \( p_1 < 1 \), in the model \( \tau_{\text{del, 1}}^{(d)} < \tau_{\text{del, 2}}^{(d)} \). (3) The vehicle speed satisfies the condition (16.58). However, at the end of time interval \( \Delta t = m, m = 1, 2, \ldots \), the vehicle speed exceeds the safe speed, i.e., \( v_{n+m} > v_{s,n+m} \). In this case, the vehicle decelerates at time step \( m + 1 \). If this vehicle deceleration occurs *before* the vehicle decelerates in accordance with (16.24), (16.26),
(16.27), (16.36), and (16.38), the mean time delay in deceleration of this vehicles class, \( \tau_{\text{del}, 3} \), satisfies the condition\(^3\)

\[
\tau_{\text{del}, 1} < \tau_{\text{del}, 3} < \tau_{\text{del}, 2}.
\]

(16.60)

The mean time delay \( \tau_{\text{del}}^{(d)} \) in vehicle deceleration corresponds to an averaging of time delays in vehicle deceleration over all vehicles.

Note that if a vehicle decelerates at time step \( n \) \( (S_n = -1) \), the vehicle continues to decelerate at time step \( n + 1 \) with probability \( p_2 \) and stops the deceleration with probability \( 1 - p_2 \). The latter case corresponds to the situation that a driver approaches the preceding vehicle to minimize the space gap. This effect is more significant for vehicles moving at low speed in dense flow that enables us to simulate the pinch effect, i.e., the self-compression and narrow moving jam emergence in synchronized flow. This is related to the behavioral assumption of item (xii) of Sect. 8.6. Therefore, the probability \( p_2 \) is taken as an increasing function of speed.

The time delay in vehicle deceleration makes an influence on the formation of the upstream front of congested patterns. In particular, the velocity of the upstream front of MSPs can depend considerably on \( \tau_{\text{del}}^{(d)} \).

However, it must be noted that in the models under consideration the time delay \( \tau_{\text{del}}^{(d)} \) is not responsible for moving jam emergence. This is because in the model the time delay in vehicle deceleration \( \tau_{\text{del}}^{(d)} \) does not lead to the well-known vehicle over-deceleration effect (see Sect. 16.3.7).

Comparison with Driver Time Delays in KKW Model

Let us consider the connection of driver time delays in vehicle acceleration and in deceleration in the spatial continuum and discrete-time model and the KKW CA model.

The driver delay in acceleration in the KKW CA model is described as in the CA models in the context of the fundamental diagram approach [364, 378]. In this case, the driver delay in acceleration is modeled due to the vehicle deceleration noise. To explain this, let us assume that a vehicle in a CA model should accelerate according to dynamic motion rules (without randomization). Then the new vehicle speed at the next time step \( n + 1 \) before applying any randomization is \( \tilde{v}_{n+1} = v_n + a \tau \). After applying the random deceleration, the value \(-a \tau\) is added to \( \tilde{v}_{n+1} \) (16.13) with probability \( p_b \) (16.15). The vehicle speed is then either

\[
v_{n+1} = v_n \quad \text{with probability} \quad p_b
\]

\[
v_{n+1} = v_n - a \tau \quad \text{with probability} \quad p_d
\]

\[
v_{n+1} = v_n - a \tau \quad \text{with probability} \quad 1 - p_b - p_d
\]

\[\text{(16.61)}\]

\(^3\) We can conclude from this discussion that the formula \( (16.59) \) is only an approximation for \( \tau_{\text{del}}^{(d)} \). This formula yields the mean time delay \( \tau_{\text{del}, 2}^{(d)} \) only when the number of vehicles of class (3) tends to zero.
or

\[ v_{n+1} = v_n + a\tau \quad \text{with probability} \quad 1 - p_b . \]  

(16.62)

In the first case (16.61), the vehicle speed maintains as it was at time step \( n \). In the second case (16.62), the vehicle accelerates. Let us denote the probability of the random driver time delay in vehicle acceleration by \( \psi_{a, \text{KKW}} \).

Corresponding to (16.61), in the KKW model \( \psi_{a, \text{KKW}} \) is simply the probability of random deceleration \( p_b \) (16.15):

\[ \psi_{a, \text{KKW}} = p_b . \]  

(16.63)

Corresponding to (16.62), in the KKW model the probability that the vehicle starts to accelerate at time step \( n+1 \) is

\[ p_{0, \text{KKW}} = 1 - p_b . \]  

(16.64)

Note that the probability \( p_b \) (16.15) is a function of vehicle speed. Thus, similar to (16.50) for the spatial continuum model, the mean time delay \( \tau_{\text{del}}^{(a)}(v) \) in vehicle acceleration in the KKW CA model is

\[ \tau_{\text{del}}^{(a)}(v) = \frac{\tau}{p_{0, \text{KKW}}(v)} . \]  

(16.65)

If the vehicle is at a standstill, i.e., \( v = 0 \), in the KKW CA model the formula (16.65) determines the mean time delay in vehicle acceleration at the downstream front of a wide moving jam:

\[ \tau_{\text{del}}^{(a)} = \tau_{\text{del}}^{(a)}(0) = \frac{\tau}{p_{0, \text{KKW}}(0)} . \]  

(16.66)

Thus, both the KKW model and the spatial continuum model lead to the same expression for the mean time delay in vehicle acceleration at the downstream front of the wide moving jam (compare (16.66) with (16.51)). However, in the KKW CA model this mean time delay (16.66) is determined by the probability \( p_b(0) \) of random deceleration, because in (16.66) we have

\[ p_{0, \text{KKW}}(0) = 1 - p_b(0) , \]  

(16.67)

where (16.64) and \( p_b(0) = p_b(0) \) (16.15) are taken into account. In other words, (16.66) coincides with (16.16) if in (16.66) we use (16.67).

The driver delay in vehicle deceleration in the KKW CA models is modeled by acceleration noise. Let us consider a case in which the vehicle speed \( v_n \) is lower than the safe speed and is higher than the speed of the preceding vehicle. If the distance \( x_{1,n} - x_n \) to the preceding vehicle is less than the synchronization distance \( D_n \), then according to the dynamic vehicle motion rules (16.1), (16.2) at the next time step \( n+1 \) the vehicle has to decelerate. For this reason, before applying the noise the vehicle speed is

\[ \tilde{v}_{n+1} = v_n - a\tau \]  

(16.68)
at the time step $n+1$. After applying a random acceleration, the value $a\tau$ is added to $\tilde{v}_{n+1}$ (16.13) with probability $p_a$ (16.19), and the vehicle speed is either

$$v_{n+1} = v_n \text{ with probability } p_a$$

or

$$v_{n+1} = v_n - a\tau \text{ with probability } 1 - p_a.$$  (16.70)

In the first case (16.69), the vehicle speed maintains as it was at time step $n$. In the second case (16.70), the vehicle decelerates. We denote the probability of the random time delay in vehicle deceleration by $\psi_{b, KKW}$. Corresponding to (16.69), in the KKW model $\psi_{b, KKW}$ is simply the probability of random acceleration $p_a$ (16.19):

$$\psi_{b, KKW} = p_a.$$  (16.71)

Corresponding to (16.70), in the KKW model the probability that the vehicle starts to decelerate at time step $n+1$ is

$$p_{1, KKW} = 1 - p_a.$$  (16.72)

Note that the probability $p_a$ (16.19) is a function of speed. Thus, similar to (16.59) for the spatial continuum model, the mean time delay $\tau_{del}$ in vehicle deceleration in the KKW model is

$$\tau_{del, 2}(v) = \frac{\tau}{p_{1, KKW}(v)}.$$  (16.73)

However, in the KKW model this mean time delay (16.73) is determined by the probability $p_a(v)$ of random acceleration, i.e., (16.73) can be rewritten as

$$\tau_{del, 2}(v) = \frac{\tau}{1 - p_a(v)},$$  (16.74)

where (16.72) is taken into account.

16.3.7 Over-Acceleration and Over-Deceleration Effects

Competition of Over-Acceleration and Speed Adaptation

The random acceleration component $\xi_a$ (16.31), (16.34) describes a random additional increase in speed when the vehicle accelerates. This random event that occurs with the probability $p_a$ simulates the vehicle over-acceleration effect (Sect. 5.2.6).

As discussed in Sect. 5.2.6 and in item (v) of Sect. 8.6, a competition of the vehicle over-acceleration and the adaptation of the vehicle speed to the speed of the preceding vehicle is responsible for an F$\rightarrow$S transition and the related Z-shaped dependence of vehicle speed on density. Thus, in the model, the simulation of the vehicle over-acceleration effect is made through
the use of random vehicle fluctuations. This formulation describes the vehicle over-acceleration as a “collective effect” that occurs on average in traffic flow.

For this reason, qualitative features of phase transitions and congested patterns are the same in the models of a single-lane road\(^4\) and a two-lane road (Chaps. 17–19).

In the KKW CA models (Sect. 16.2), the case \(\chi_n = 1\) in (16.14) describes a random additional increase in speed. One of the aims of this random vehicle acceleration is to simulate the vehicle over-acceleration effect. Thus, as in the spatial continuum model, in the KKW CA models the vehicle over-acceleration effect is also simulated through the use of fluctuations, i.e., as a “collective effect” that occurs on average in traffic flow.

**Over-Deceleration Effect**

Due to the vehicle over-deceleration effect, after the preceding vehicles has begun to decelerate unexpectedly, the vehicle reduces the speed harder than this is required to avoid collisions (Sect. 3.2.4). As in the Nagel–Schreckenberg model [364], in the model under consideration the vehicle over-deceleration effect is simulated through the use of model fluctuations, i.e., as a “collective effect” that occurs on average in traffic flow.

When the vehicle has decelerated at time step \(n + 1\), i.e., in (16.31) \(S_{n+1} = -1\), then in accordance to (16.30)–(16.33) with probability \(p_b\) there is an additional decrease in vehicle speed that is equal to \(-a\tau\). This random decrease in speed simulates over-deceleration of the vehicle.

This random effect of the vehicle over-deceleration is responsible for moving jam emergence both in the model under consideration and in the KKW CA-model of Sect. 16.2.

**16.3.8 Lane Changing Rules**

Lane changing rules in a two-lane model are based on the well-known incentive and security conditions (see Nagel et al. [429]). However, these conditions should be adjusted to take the effect of the synchronization distance into account.

The following incentive conditions for lane changing from the right lane to the left lane (\(R \rightarrow L\)) and a return change from the left lane to the right lane (\(L \rightarrow R\)) have been used in the model:

\[
R \rightarrow L : v^+_n \geq v_{\ell,n} + \delta_1 \text{ and } v_n \geq v_{\ell,n} ,
\]

\[(16.75)\]

\(^4\) In this model approach, *dynamic* states of synchronized flow, i.e., the states that occur as a result of an F→S transition at each given flow rate are separated by a finite gap in speed from states with higher vehicle speeds. These higher speed states will be considered as states of free flow. This explains the sense of free flow in the model of a single-lane road.
The security conditions for lane changing are given by the inequalities:

\[ g_n^+ > \min(v_n \tau, D_n^+ - d), \quad (16.77) \]
\[ g_n^- > \min(v_n \tau, D_n^- - d), \quad (16.78) \]

where

\[ D_n^+ = D(v_n, v_n^+), \quad (16.79) \]
\[ D_n^- = D(v_n^-, v_n), \quad (16.80) \]

the function \( D(u, w) \) is given by (16.29); \( \delta_1 \geq 0 \) is a constant. Superscripts + and – in variables, parameters, and functions denote the preceding vehicle and the trailing vehicle in the “target” (neighboring) lane, respectively. The target lane is the lane into which the vehicle wants to change. For example, \( g_n^+ \) is the space gap between the vehicle that wants to change the lane and the preceding vehicle in the target lane; \( g_n^- \) is the space gap between the vehicle that wants to change the lane and the trailing vehicle in the target lane.

Similar to lane changing rules in other models \[429\], the speed \( v_n^+ \) or the speed \( v_{\ell,n} \) in (16.75), (16.76) is set to \( \infty \) if the space gap \( g_n^+ \) or the space gap \( g_n^- \) exceeds a given look-ahead distance \( L_a \), respectively. The functions \( D_n^+ \), \( D_n^- \) in (16.77), (16.78) facilitate the synchronization of vehicle speed across the lanes if there are a large difference between speeds in different lanes.

If the conditions (16.75)-(16.78) are satisfied, then as in Rickert et al. \[481\], in this model the vehicle changes lanes with probability \( p_c \).

### 16.3.9 Boundary Conditions and Models of Bottlenecks

#### Boundary and Initial Conditions

In the model, open boundary conditions are applied. At the beginning of the road new vehicles are generated one after another in each of the lanes of the road at time moments (16.20), where \( \tau_{in} = 1/q_{in} \), \( q_{in} \) is the flow rate in the incoming boundary flow per freeway lane. A new vehicle appears in a lane of the road at time (16.20) only if the distance from the beginning of the road \( (x = x_b) \) to the position \( x = x_{\ell,n} \) of the farthest upstream vehicle in the lane is not smaller than the safe distance \( \ell_{safe} = v_{\ell,n} \tau + d \), i.e., the condition (16.21) should be satisfied, where \( n = t^{(k')} / \tau \). Otherwise, the condition (16.21) is checked at time \( (n + 1)\tau \) that is the next one to time \( t^{(k')} \) (16.20), and so on, until the condition (16.21) is satisfied. Then a new vehicle appears on the road. After this occurs, the number \( k' \) in (16.20) is increased by 1. The speed \( v_n \) and the coordinate \( x_n \) of the new vehicle are

\[ v_n = v_{\ell,n}, \]
\[ x_n = \max(x_b, x_{\ell,n} - \max(v_n \tau_{in}, \ell_{safe})). \quad (16.81) \]
In the initial state \( n = 0 \), all vehicles have the maximum speed \( v_n = v_{\text{free}} \) and they are positioned at space intervals \( x_{\ell,n} - x_n = v_{\text{free}} \tau_{\text{in}} \).

After a vehicle has reached the end of the road it is removed; before this, the farthest downstream vehicle maintains its speed and lane. For the vehicle following the farthest downstream one, the “anticipation” speed \( v^{(a)}_{\ell} \) in (16.41) is equal to the speed of the farthest downstream vehicle.

**Models of Bottlenecks**

A bottleneck due to the on-ramp, a merge bottleneck, where two lanes are reduced to one lane, and a bottleneck due to the off-ramp (Fig. 16.2) are considered.

(a) The on-ramp bottleneck consists of two parts (Fig. 16.2a):

(i) The merging region of length \( L_m \) where vehicles can merge onto the main road from the on-ramp lane.

(ii) A part of the on-ramp lane of length \( L_r \) upstream of the merging region where vehicles move in accordance with the model (16.24)–(16.38), (16.41), (16.48). The maximal speed of vehicles is \( v_{\text{free}} = v_{\text{free on}} \).

At the beginning of the merging region of the on-ramp lane \( (x = x^{(b)}_{\text{on}}) \) the flow rate to the on-ramp \( q_{\text{on}} \) is given as in \( q_{\text{in}} \).

At the merge bottleneck (Fig. 16.2b) within the merging region of length \( L_c \) upstream of the merge point \( x = x_M \) vehicles have to change from the right lane to the left lane.

The off-ramp bottleneck consists of two parts (Fig. 16.2c):
(i) A merging region of length \( L_m \) where vehicle can merge from the main road onto the off-ramp lane.

(ii) A part of the off-ramp lane of length \( L_r \) downstream of the merging region where vehicles move in accordance with the model (16.24)–(16.38), (16.41), (16.48). The maximal speed of vehicles is \( \dot{v}_{\text{free}} = v_{\text{free off}} \).

Within a second merging region of length \( L_m + L_c \), which is on the main road \( (x_{\text{off}}^{(s)} \leq x \leq x_{\text{off}}^{(b)} \) in Fig. 16.2c), vehicles going to the off-ramp have to change from the left lane to the right lane of the main road. The flow rate of vehicles that go to the off-ramp is given as a percentage \( \eta \) of the flow rate \( q_{\text{in}}' \). For this purpose, at the beginning of the road a vehicle is given an “attribute,” which marks this vehicle as a vehicle going to the off-ramp if

\[
\rho_{\text{off}} < \eta/100\% ,
\]

where \( \rho_{\text{off}} = \text{rand}(0,1) \).

Vehicle Motion Rules in Merging Region

There are some peculiarities of vehicle motion for vehicles that move within the merging regions of these bottlenecks (Fig. 16.2a–c). This is related only to those vehicles that intend to merge onto the target lane. These vehicles move either in the on-ramp lane, or in the right lane of the merge bottleneck, or else these are the vehicles that want to leave the main road to the off-ramp. These peculiarities are as follows.

Each of these vehicles moves in accordance with the model rules (16.24)–(16.38), (16.41), (16.48). However, in (16.26)–(16.28) the coordinate \( x_{\ell,n} \) and the speed \( v_{\ell,n} \) of the preceding vehicle are replaced by values \( x_{n}^{+} \) and \( \dot{v}_{n}^{+} \), respectively where

\[
\dot{v}_{n}^{+} = \max \left(0, \min \left( \dot{v}_{\text{free}}, \dot{v}_{n}^{+} + \Delta v_{r}^{(2)} \right) \right) ,
\]

\( x_{n}^{+} \) and \( v_{n}^{+} \) are the coordinate and the speed of the preceding vehicle in the target lane, \( \Delta v_{r}^{(2)} \) is a constant. However, the safe speed in the vehicle motion equations (16.24) is further determined by the safe speed \( v_{s,n} \) (16.41) related to the preceding vehicle in the lane where the vehicle currently moves.

Models of Vehicle Merging

The lane changing rules of Sect. 16.3.8 are used beyond all bottleneck merging regions. The following rules for vehicle merging within the merging regions are assumed to be the same for all three types of bottlenecks.

The rule (*): a speed \( \dot{v}_{n} \) is calculated corresponding to

\[
\dot{v}_{n} = \min \left( \dot{v}_{n}^{+}, v_{n} + \Delta v_{r}^{(1)} \right) ,
\]
where $\Delta v^{(1)}_n$ is a constant that describes the maximum possible increase in speed after merging. Then the speed $\dot{v}_n$ is used instead of $v_n$ in the security lane changing rules (16.77), (16.78). If these conditions are satisfied, then the vehicle merges within the merging region of a bottleneck. In this case, the vehicle speed $v_n$ is set to $\dot{v}_n$ (16.84) and the vehicle spatial coordinate in the new lane does not change in comparison with the spatial coordinate in the old lane.

If the rule (*) is not satisfied, the rule (**) is applied: the space gap

$$x_n^+ + x_n^- - d$$

(16.85)

between two neighboring vehicles in the target lane should exceed some value

$$g_{on}^{(\text{min})} = \lambda v_n^+ + d,$$

(16.86)

where $\lambda$ is a parameter. In addition, the condition that the vehicle passes the midpoint

$$x_n^{(m)} = (x_n^+ + x_n^-)/2$$

(16.87)

between two neighboring vehicles in the target lane for time step $n$ should be satisfied, i.e.,

$$x_{n-1} < x_n^{(m)} \text{ and } x_n \geq x_n^{(m)}$$

or

$$x_{n-1} \geq x_n^{(m)} \text{ and } x_n < x_n^{(m)}.$$

(16.88)

After merging, the coordinate of the merging vehicle is set to $x_n = x_n^{(m)}$ and the vehicle speed $v_n$ is set to $\dot{v}_n$ (16.84).

If neither the rule (*) nor the rule (**) is satisfied, the vehicle does not merge onto the target lane. If the vehicle cannot merge onto the main road, the vehicle moves in the on-ramp lane until it comes to a stop at the end of the merging region, $x = x_{\text{mer}}^{(e)}$. The end of the merging region is located at the point $x_{\text{mer}}^{(e)} = x_{\text{on}}^{(e)}$ for the on-ramp bottleneck (Fig. 16.2a), at $x_{\text{mer}}^{(e)} = x_{\text{M}}$ for the merge bottleneck (Fig. 16.2b), and at $x_{\text{mer}}^{(e)} = x_{\text{off}}^{(e)}$ for the off-ramp bottleneck (Fig. 16.2c). To describe the motion of the vehicle, which is the closest one to the end of the merging region $x = x_{\text{mer}}^{(e)}$, in (16.24) instead of (16.41), (16.48) the following formula is used for calculation of the safe speed $v_{s,n}$:

$$v_{s,n} = v^{(\text{safe})}(g_n^{(e)}, 0).$$

(16.89)

In (16.89), the safe speed $v^{(\text{safe})}$ is given by (16.45)–(16.47), and the space gap $g_n^{(e)}$ is the distance to the end of the merging region:

$$g_n^{(e)} = x_{\text{mer}}^{(e)} - x_n.$$

(16.90)
16.3.10 Summary of Model Equations and Parameters

In tables presented below, we summarize equations that describe the spatial continuum and discrete-time model (Tables 16.5 and 16.6), symbols (Tables 16.7–16.10), and values for model parameters (Table 16.11) used in simulations of phase transitions and spatiotemporal congested patterns presented in this book [330].

Discussion of Traffic Flow Model Parameters

In the KKW CA model (Sect. 16.2) and in the spatial continuum model (Sect. 16.3) all vehicles are identical. This is certainly an idealization of real traffic flow where there are vehicles of different classes and of different lengths. Furthermore, drivers have different driving strategies. The individual parameters of drivers and vehicles can be easily integrated into the traffic flow models under consideration [332] (see Chap. 20). In this case, different individual attributes of drivers and vehicles, such as the vehicle length $d$, the maximum speed $v_{\text{free}}$, the mean time delay $\tau^{(a)}(v)$, and other characteristics can be set for different vehicles. These attributes can be given to a vehicle at the beginning of the road in the same manner as has been done for vehicles going to the off-ramp (Sect. 16.3.9).

However, simulations of the traffic flow model based on three-phase traffic theory, when vehicles with different individual parameters are used [332], show that all main qualitative fundamental features of the congested pattern diagram and of congested patterns remain approximately the same as those in the case when identical vehicles are assumed. For this reason and for the sake of simplicity, in Chaps. 17–19 results of a study of the traffic flow model with identical vehicles are considered. A discussion of some specific pattern features associated with different driver behavioral characteristics and different vehicle parameters in heterogeneous traffic flow appears in Chap. 20.

Traffic flow models should be validated based on empirical data. In some cases, the model parameters are chosen in the way that empirical distributions of time gaps between vehicles measured at some road locations under different traffic flow conditions should be in agreement with those simulated in the model. In other cases, empirical data for vehicle speeds and accelerations as functions of time for several vehicles moving on the road one after another are used for the validation of model parameters.

The model parameters discussed in Sects. 16.2 and 16.3 can also be calibrated in the same manner. However, the aim of this book is to explain the physics of phase transitions and spatiotemporal congested traffic patterns observed in real traffic. For this reason, the main criterion for model validity is that the model should be able to describe the main features of spatiotemporal congested patterns empirically observed in real traffic. As shown in Sect. 3.3, traffic flow models based on the fundamental diagram approach cannot describe and predict most of fundamental empirical congested traffic pattern...
**Table 16.5.** Spatial continuum microscopic single-lane model based on three-phase traffic theory [330]

<table>
<thead>
<tr>
<th>Dynamic part of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{n+1} = \max(0, \min(v_{\text{free}}, v_{s,n}, v_{c,n}))$,</td>
</tr>
</tbody>
</table>
| $v_{c,n} = \begin{cases} 
v_n + \Delta_n & \text{at } x_{\ell,n} - x_n \leq D_n, \\
v_n + a_n \tau & \text{at } x_{\ell,n} - x_n > D_n, 
\end{cases}$ |
| $\Delta_n = \max(-b_n \tau, \min(a_n \tau, v_{\ell,n} - v_n))$, |
| $g_n = x_{\ell,n} - x_n - d$, |
| $v_{\text{free}}, \, d, \, \text{and } \tau$ are constants. |

<table>
<thead>
<tr>
<th>Stochastic part of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}<em>{n+1} + \xi_n, v_n + a \tau, v</em>{s,n}))$,</td>
</tr>
<tr>
<td>$x_{n+1} = x_n + v_{n+1} \tau$,</td>
</tr>
<tr>
<td>$a$ is a constant.</td>
</tr>
</tbody>
</table>

**Stochastic time delay of acceleration and deceleration:**

- $a_n = a \Theta(P_0 - r_1)$,  $b_n = a \Theta(P_1 - r_1)$,
- $P_0 = \begin{cases} p_0 & \text{if } S_n \neq 1 \\
1 & \text{if } S_n = 1, \end{cases}$  $P_1 = \begin{cases} p_1 & \text{if } S_n \neq -1 \\
p_2 & \text{if } S_n = -1, \end{cases}$
- $p_0 = p_0(v)$, $p_2 = p_2(v)$ are functions of speed, $r_1 = \text{rand}(0,1)$, $p_1$ is constant.

**State of motion:**

- $S_{n+1} = \begin{cases} -1 & \text{if } \tilde{v}_{n+1} < v_n - \delta \\
1 & \text{if } \tilde{v}_{n+1} > v_n + \delta \\
0 & \text{otherwise}, \end{cases}$
- $\delta$ is a constant ($\delta \ll a \tau$).

**Random acceleration and deceleration:**

- $\xi_n = \begin{cases} -\xi_b & \text{if } S_{n+1} = -1 \\
\xi_a & \text{if } S_{n+1} = 1 \\
0 & \text{if } S_{n+1} = 0, \end{cases}$

**Specifications of synchronization distance $D_n$ and noise $\xi_n$**

- $D_n = d + \max(0, k v_n \tau + \phi_0 a^{-1} v_n (v_n - v_{\ell,n}))$
- $\xi_a = a \tau \Theta(p_a - r)$,  $\xi_b = a \tau \Theta(p_b - r)$,
- $r = \text{rand}(0,1)$, $\phi_0, \, k, \, p_a, \, p_b$ are constants.
### Lane changing rules in two-lane model

| Right → Left | \( v_{n+}^+ \geq v_{\ell,n} + \delta_1 \) and \( v_n \geq v_{\ell,n} \), |
| Left → Right | \( v_{n+}^+ > v_{\ell,n} + \delta_1 \) or \( v_{n+}^+ > v_n + \delta_1 \), |
| \( g_n^+ > \min(v_n \tau, D_n^+ - d) \), \( g_n^- > \min(v_n^- \tau, D_n^- - d) \), |
| \( r_c < p_c \), |
| \( v_n^+ \Rightarrow \infty \) if \( g_n^+ > L_a \), \( v_{\ell,n} \Rightarrow \infty \) if \( g_n > L_a \), |
| \( D_n^+ = D(v_n, v_{n+}^+) \), \( D_n^- = D(v_n, v_n) \), \( r_c = \text{rand}(0, 1) \), |
| \( \delta_1, p_c \) and \( L_a \) are constants. |

### Rules for vehicle merging

**The rule (\(*\))**:

\[
\begin{align*}
  g_n^+ &= \min(v_n \tau, D_n^+ - d), \\
  g_n^- &= \min(v_n^- \tau, D_n^- - d), \\
  \hat{v}_n &= \min(v_n^+, v_n + \Delta v_r^{(1)}),
\end{align*}
\]

\( \Delta v_r^{(1)} > 0 \) is a constant.

**The rule (**\(*\))**:

\[
\begin{align*}
  x_n^+ - x_n^- - d &> \lambda v_n^+ + d, \\
  x_{n-1} < x_{n-1}^{(m)} \text{ and } x_n \geq x_n^{(m)} \text{ or } x_{n-1} \geq x_{n-1}^{(m)} \text{ and } x_n < x_n^{(m)}, \\
  x_n^{(m)} &= (x_n^+ + x_n^-)/2,
\end{align*}
\]

\( \lambda \) is a parameter.

### Parameters after vehicle merging:

\( v_n = \hat{v}_n \),

under the rule (\(*\)): \( x_n \) maintains the same,

under the rule (**\(*\))**: \( x_n = x_n^{(m)} \).

### Speed adaptation before vehicle merging:

\[
\begin{align*}
  v_{c,n} &= \begin{cases} 
    v_n + \Delta_n^+ & \text{at } x_n^+ - x_n \leq D(v_n, \hat{v}_n^+), \\
    v_n + a_n \tau & \text{at } x_n^+ - x_n > D(v_n, \hat{v}_n^+),
  \end{cases} \\
  \Delta_n^+ &= \max(-b_n \tau, \min(a_n \tau, \hat{v}_n^+ - v_n)), \\
  \hat{v}_n^+ &= \max(0, \min(v_{\text{free}}, v_n^+ + \Delta v_r^{(2)})),
\end{align*}
\]

\( \Delta v_r^{(2)} \) is a constant.
features. As we see below, the models discussed in Sects. 16.2 and 16.3 enable us to explain and predict phase transitions in traffic flow and main empirical congested pattern features at different freeway bottlenecks. For this reason, the parameters of the KKW CA model (Sect. 16.2) and of the spatial continuum and discrete-time traffic flow model (Sect. 16.3) are chosen to have the simulation parameters of spatiotemporal congested patterns that are close to the empirical ones.
### Table 16.8. List of symbols (two-lane model: lane changing rules)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{n}^{+}$</td>
<td>coordinate of preceding vehicle in target (neighboring) lane at time step $n$</td>
</tr>
<tr>
<td>$x_{n}^{-}$</td>
<td>coordinate of trailing vehicle in target lane at time step $n$</td>
</tr>
<tr>
<td>$v_{n}^{+}$</td>
<td>speed of preceding vehicle in target lane at time step $n$</td>
</tr>
<tr>
<td>$v_{n}^{-}$</td>
<td>speed of trailing vehicle in target lane at time step $n$</td>
</tr>
<tr>
<td>$g_{n}^{+} = x_{n}^{+} - x_{n} - d$</td>
<td>space gap to preceding vehicle in target lane at time step $n$</td>
</tr>
<tr>
<td>$g_{n}^{-} = x_{n} - x_{n}^{-} - d$</td>
<td>space gap to trailing vehicle in target lane at time step $n$</td>
</tr>
<tr>
<td>$L_{a}$</td>
<td>look-ahead distance for lane changing rules</td>
</tr>
<tr>
<td>$p_{c}$</td>
<td>probability for lane changing</td>
</tr>
<tr>
<td>$\delta_{1}$</td>
<td>parameter in lane changing rules</td>
</tr>
</tbody>
</table>

### Table 16.9. List of symbols (bottlenecks)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_{on}$</td>
<td>flow rate to on-ramp</td>
</tr>
<tr>
<td>$\eta$</td>
<td>percentage of vehicles going to off-ramp</td>
</tr>
<tr>
<td>$v_{on}^\text{free}$</td>
<td>maximum speed in on-ramp lane</td>
</tr>
<tr>
<td>$v_{off}^\text{free}$</td>
<td>maximum speed in off-ramp lane</td>
</tr>
<tr>
<td>$t_{0}$</td>
<td>time at which on-ramp inflow is switched</td>
</tr>
<tr>
<td>$L_{m}$</td>
<td>length of merging region for on- and off-ramps</td>
</tr>
<tr>
<td>$L_{r}$</td>
<td>length of on-ramp lane or off-ramp lane</td>
</tr>
<tr>
<td>$L_{c}$</td>
<td>length of region on main road where vehicles change lane upstream of merge bottleneck or upstream of off-ramp</td>
</tr>
<tr>
<td>$g_{on}^{\text{(min)}}$</td>
<td>minimum space gap for vehicle merging between vehicles in target (neighboring) lane</td>
</tr>
<tr>
<td>$x_{n}^{(m)}$</td>
<td>coordinate of midpoint between preceding vehicle and trailing vehicle in target lane</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>parameter in merging rule (**)</td>
</tr>
<tr>
<td>$\Delta v_{r}^{(1)}$</td>
<td>maximum possible increase in vehicle speed after vehicle merging</td>
</tr>
<tr>
<td>$\Delta v_{r}^{(2)}$</td>
<td>parameter of speed adaptation before vehicle merging</td>
</tr>
</tbody>
</table>
Table 16.10. List of symbols (pattern parameters)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_{\text{max}} )</td>
<td>maximum flow rate in free flow (( v = v_{\text{free}} ))</td>
</tr>
<tr>
<td>( q_0 )</td>
<td>maximum flow rate in steady states</td>
</tr>
<tr>
<td>( \rho_{\text{max}} )</td>
<td>maximum density in free flow related to ( q_{\text{max}}^{(\text{free})} )</td>
</tr>
<tr>
<td>( \rho_{\text{max}} = 1/d )</td>
<td>density within wide moving jam</td>
</tr>
<tr>
<td>( v_g )</td>
<td>velocity of downstream front of wide moving jam</td>
</tr>
<tr>
<td>( q_{\text{out}} )</td>
<td>flow rate in free flow in wide moving jam outflow</td>
</tr>
<tr>
<td>( \rho_{\text{min}} )</td>
<td>density in free flow relative to flow rate ( q_{\text{out}} )</td>
</tr>
<tr>
<td>( q_{\text{lim}} )</td>
<td>limit flow rate in pinch region of general pattern</td>
</tr>
<tr>
<td>( q_{\text{out}} )</td>
<td>outflow from congestion pattern at bottleneck</td>
</tr>
<tr>
<td>( q_{\text{th}} )</td>
<td>threshold flow rate for MSP excitation in free flow</td>
</tr>
<tr>
<td>( \rho_{\text{th}} )</td>
<td>threshold density in free flow related to flow rate ( q_{\text{th}} )</td>
</tr>
</tbody>
</table>

Table 16.11. Model parameters and characteristic values

| Single-lane model parameters | \( \tau = 1 \text{s}, d = 7.5 \text{m}, \)
| \( v_{\text{free}} = 30 \text{m/s} = 108 \text{km/h}, \)
| \( a = 0.5 \text{m/s}^2, b = 1 \text{m/s}^2, \)
| \( D_n = 3v_n \tau + 7.5 \text{ in m (for } v_n = v_{t,n}, v_n \text{ in m/s)}, \)
| \( v_{s,n} = g_n/\tau \text{ m/s (for } v_n = v_{t,n}) (g_n \text{ in m}), \)
| \( k = 3, \phi_0 = 1, \delta = 0.01, \)
| \( p_1 = 0.3, p_0 = 0.17, p_b = 0.1, \)
| \( p_0(v) = 0.575 + 0.125 \min(1, v/10) (v \text{ in m/s}), \)
| \( p_2(v) = 0.48 + 0.32\Theta(v - 15) (v \text{ in m/s}). \)
| Lane changing parameters | \( \delta_1 = 1 \text{ m/s}, \)
| \( L_a = 150 \text{ m}, \)
| \( p_c = 0.2. \)
| Model parameters for bottlenecks | \( \lambda = 0.75\tau (\lambda = 0.6\tau \text{ for off-ramp lane}), \)
| \( v_{\text{free on}} = 80 \text{km/h}, v_{\text{free off}} = 90 \text{km/h}, \)
| \( \Delta v^{(1)}_t = 10 \text{m/s}, \)
| \( \Delta v^{(2)}_t = 5 \text{m/s at on-ramp} \)
| \( \Delta v^{(2)}_t = -2 \text{m/s at off-ramp} \)
| and at merge bottleneck |
| Model results | \( v_g = -15.5 \text{km/h}, q_{\text{out}} = 1810 \text{vehicles/h}, \)
| \( \rho_{\text{min}} = 16.76 \text{vehicles/km}, q_0 = 2880 \text{vehicles/h}, \)
| \( q_{\text{free}}^{(\text{max})} \approx 2430 \text{vehicles/h}, \)
| \( q_{\text{free}}^{(\text{max})} \approx 22.5 \text{vehicles/km}, \)
| \( q_{\text{th}} \approx 2200 \text{vehicles/h}, \)
| \( \rho_{\text{th}} \approx 20.4 \text{vehicles/km}. \)
16.4 Conclusions

(i) In microscopic models based on three-phase traffic theory, a two-dimensional region of steady state model solutions is formed by the introduction of a “synchronization distance.”

(ii) The synchronization distance is a dynamic model characteristic that depends on time-dependent vehicle speeds. The synchronization distance determines vehicle acceleration behavior (the rule of “speed change”) depending on whether the preceding vehicle is within a “synchronization distance” or further away. At a sufficiently large distance from the preceding vehicle, the vehicle simply accelerates. However, if the vehicle cannot pass the preceding vehicle, then within the synchronization distance the vehicle tends to adjust its speed to the preceding vehicle, i.e., it decelerates if it is faster, and accelerates if it is slower than the preceding vehicle.

(iii) The well-known (from different microscopic traffic flow models in the context of the fundamental diagram approach) safety conditions, driver time delays, stochastic behavior of fluctuations, and lane changing rules should be adjusted to satisfy dynamic vehicle motion rules associated with the introduction of the synchronization distance.
17 Microscopic Theory of Phase Transitions in Freeway Traffic

17.1 Introduction

The aim of this chapter is an analysis of phase transitions based on a microscopic three-phase traffic theory. As discussed in Sect. 2.4 and Chap. 10, in empirical observations moving jams do not emerge spontaneously in free flow, i.e., an $F \rightarrow J$ transition is not observed. Instead of this phase transition, an $F \rightarrow S$ transition, i.e., the spontaneous emergence of a local region of synchronized flow, is responsible for the onset of congestion in free flow. Only later, and usually at another freeway location can an $S \rightarrow J$ transition, i.e., moving jam emergence in synchronized flow, occur spontaneously. We will see in this chapter that the microscopic three-phase traffic theory can explain and predict $F \rightarrow S \rightarrow J$ transitions observed in real traffic flow.

Firstly, we will consider an $F \rightarrow S$ transition on a hypothetical homogeneous road, i.e., a road without bottlenecks. Such consideration of the $F \rightarrow S$ transition gives us the opportunity to see some intrinsic nonlinear features of this phase transition. Secondly, we will consider $F \rightarrow S$ transitions at different freeway bottlenecks. We will find that, on the one hand, the features of an $F \rightarrow S$ transition are qualitatively the same as those on the homogeneous road. On the other hand, these features are qualitatively the same as those observed in empirical observations of the breakdown phenomenon at bottlenecks (Chap. 10).  

Finally, we will present results of simulations of an $S \rightarrow J$ transition and of $F \rightarrow S \rightarrow J$ transitions.

Theory of phase transitions in traffic flow presented below is based on the papers [329, 330]. These results, on the other hand, confirm all conclusions of the three-phase traffic theory discussed in Chaps. 5 and 6. In particular, we will find double Z-characteristics of traffic flow associated with $F \rightarrow S \rightarrow J$ transitions. These double Z-characteristics consist of a Z-characteristic for an $F \rightarrow S$ transition and of a Z-characteristic for an $S \rightarrow J$ transition. On the other hand, the microscopic theory of phase transitions enables us to explain and predict empirical features of these phase transitions discussed in Sect. 2.4, Chaps. 10 and 12.
17.2 Microscopic Theory of Breakdown Phenomenon (F→S Transition)

17.2.1 Homogeneous Road

Nucleation of Synchronized Flow

Numerical simulations of the spatial continuum and discrete-time model of Sect. 16.3 show that in an initial free flow a local region of synchronized flow can occur (Fig. 17.1a). After this F→S transition has occurred, dependent on the initial density in free flow and on perturbation characteristics, various states of synchronized flow can be realized.

Fig. 17.1. F→S and F→J transitions. (a) MSP (moving synchronized flow pattern) that occurs due to an F→S transition. (b) Dependencies of the critical amplitude of local perturbations on density for an F→J transition (curve F_J), and for an F→S transition (curve F_S); Δv_{cr} = Δv_{cr}^{(FS)} for the F→S transition and Δv_{cr} = Δv_{cr}^{(FJ)} for the F→J transition. (c) Steady states of the spatial continuum model together with the line J and synchronized flow states S within the MSP in (a). Taken from [329]
An F→S transition (Fig. 17.1a) can occur if the flow rate (vehicle density) in an initial free flow is in the range

$$q_{th} \leq q_{in} \leq q_{max}^{(free)} \quad \left( \rho_{th} \leq \rho_{in} \leq \rho_{max}^{(free)} \right), \tag{17.1}$$

where $\rho_{in} = q_{in}/v_{free}$.

The F→S transition is a first-order phase transition. There is a nucleation effect that governs this phase transition: if a local perturbation occurs in the density range (17.1), then this perturbation grows and leads to the F→S transition if the amplitude of the perturbation exceeds some critical value $\Delta v_{cr}^{(FS)}$ (Fig. 17.1b). Otherwise, the perturbation gradually dissolves and the initial free flow is recovered. A numerical simulation of this decay for an initial perturbation whose amplitude is lower than the critical amplitude is shown in Fig. 5.11.

The critical amplitude of the local perturbation $\Delta v_{cr}^{(FS)}$ is a decreasing function of density in the initial free flow (curve $F_S$ in Fig. 17.1b). The critical amplitude $\Delta v_{cr}^{(FS)}$ reaches its maximum value at the threshold density for the F→S transition, $\rho_{th}$, and is zero at the maximum density in free flow $\rho_{max}^{(free)}$. If the density is lower than $\rho_{th}$, no F→S transition can be induced by a time-limited local perturbation.

After the F→S transition has occurred, an MSP emerges. This MSP is shown in Fig. 17.1a. The small circles $S$ in Fig. 17.1c, which are related to synchronized flow within the MSP, show that this synchronized flow approximately corresponds to “homogeneous-in-speed” synchronized flow, where the speed is constant (in this case $v \approx 55$ km/h) and the density varies in space and time.

Wide moving jams do not occur spontaneously at arbitrary free-flow density. In free flow, above the threshold density $\rho_{min}$ for an F→J transition, wide moving jams can only be excited if an initial local perturbation of very large amplitude is applied. This amplitude should be greater than the critical amplitude of a local perturbation for an F→J transition (curve $F_J$ in Fig. 17.1b). The density $\rho_{min}$ is associated with the flow rate $q_{out}$ in the outflow of a wide moving jam for the case when free flow is formed downstream of the wide moving jam.

**Threshold Point**

At the threshold point for an F→S transition ($\rho_{th}$, $q_{th}$) (Fig. 17.2a) the downstream front velocity of an MSP $v_{down}$ is equal to the velocity of the MSP upstream front $v_{up}$: $v_{down} = v_{up}$. At $\rho_{in} > \rho_{th}$ ($q_{in} > q_{th}$) we have $|v_{down}| < |v_{up}|$ (Fig. 17.2a), i.e., the MSP width (in the longitudinal direction) increases over time (Fig. 17.1a).
In contrast, at $\rho_{in} < \rho_{th}$ ($q_{in} < q_{th}$) we have $|v_{down}| > |v_{up}|$ (Fig. 17.2a), i.e., the MSP width decreases and the MSP dissolves (Fig. 17.2b). These model results confirm formula (5.29) and the physical explanation of the threshold point for an F→S transition [330] discussed in Sect. 5.2.7. Thus, within the density range (17.1) the nucleation of synchronized flow in an initial free flow can occur, rather than moving jams emerging.

S→F Transition

If the density in a region of synchronized flow is relatively low, then model fluctuations can destroy this initial synchronized flow: due to a spontaneous local S→F transition in the region of synchronized flow a local region of free flow occurs.

This effect is due to spontaneous occurrence and propagation of the upstream front of the region of free flow. Within this upstream front drivers accelerate from synchronized flow to a higher speed in free flow: the upstream front of free flow spatially separates free flow downstream of the front and synchronized flow upstream of the front. This upstream front of free flow initially appears due to over-acceleration of one of the drivers in synchronized flow. The following driver accelerates also, and so on.

However, within the downstream front of the region of free flow, which separates free flow upstream and synchronized flow downstream, drivers must slow down to a lower speed in synchronized flow. The S→F transition occurs only if the absolute value of the velocity of the downstream front of free flow is lower than the absolute value of the velocity of the upstream front of free flow. As a result, the local region of free flow is spatially widening over time.

In simulations of this effect, many such widening local regions of free flow can occur almost simultaneously in an initial synchronized flow. In this case, different local regions of free flow can merge over time within synchronized flow. This causes the substitution of free flow for synchronized flow on the entire road.
Z-Shaped Speed–Flow Characteristic

There is a Z-shaped speed–flow characteristic (Fig. 17.3a) associated with these metastability and nucleation effects with respect to a first-order F→S transition.

![Graph](image)

**Fig. 17.3.** Z-characteristic for an F→S transition on a homogeneous road. (a) Z-shaped speed–flow characteristic. (b) Speed–density characteristic where \( v_{\text{free}} \) and \( v^{(\text{FS})} \) have the same meaning as those in (a); the curve \( J \) is associated with the line \( J \) in Fig. 17.1a. In (a) \( q_{\text{in}} \) is a control (external) parameter; the speed \( v_{\text{free}} \) corresponds to free flow; the critical branch \( v_{\text{cr}}^{(\text{FS})} \) gives the speed within the critical local perturbation required for the F→S transition as a function of \( q_{\text{in}} \); \( v^{(\text{syn})} \) is a synchronized flow speed inside a 2D region of dynamic synchronized flow states (dashed region) that occur due to the F→S transition. Taken from [330]

The Z-characteristic consists of the branch of free flow \( v_{\text{free}} \), of a 2D region of synchronized flow states \( v^{(\text{syn})} \), and of the critical branch \( v_{\text{cr}}^{(\text{FS})} \). The critical branch gives the speed within critical perturbations, i.e., it determines the critical amplitude of local perturbations in free flow

\[
\Delta v_{\text{cr}}^{(\text{FS})} = v_{\text{free}} - v_{\text{cr}}^{(\text{FS})} .
\]  

(17.2)

When the amplitude of an external local perturbation exceeds the critical amplitude (17.2), an F→S transition (down-arrow in Fig. 17.3a) occurs. As a result of the F→S transition, an MSP emerges (Fig. 17.1a) whose width continuously increases over time. Otherwise (up-arrow in Fig. 17.3a), the initial perturbation decays (see a simulation of this decay in Fig. 5.11). At \( q = q_{\text{max}}^{(\text{free})} \) an F→S transition (dotted down-arrow in Fig. 17.3a) occurs even at a very small amplitude of perturbations.

In Fig. 17.3a, the 2D region of states of synchronized flow corresponds to *dynamic synchronized flow states* within the MSP rather than steady states of synchronized flow. This leads to a difference between the Z-characteristic in Fig. 17.3a and the Z-characteristic in Fig. 5.2b derived in the three-phase
traffic theory: in the latter case, a 2D region of synchronized flow is associated with steady states of synchronized flow. This difference has already been explained in Sect. 4.3.4: steady states of synchronized flow are only hypothetical states that cannot be found in reality. This is because these states are destroyed by fluctuations and dynamic effects in real synchronized flow. As a result, rather than steady states some dynamic states of synchronized flow occur in the model (the states $v^{(syn)}$ in Fig. 17.3a) after an $F \rightarrow S$ transition has occurred. However, these dynamic synchronized flow states exhibit the main features of steady states of synchronized flow postulated in the hypotheses of the three-phase traffic theory (in particular, characteristics of the steady state stability and phase transitions in traffic flow). For this reason, qualitative features of all Z-characteristics discussed in the three-phase traffic theory (Chaps. 5 and 6) remain for all dynamic synchronized flow states discussed in this chapter.

Let us consider the vehicle speed as a function of density within these dynamic synchronized flow states, i.e., the density within the MSP (Fig. 17.3b). It can be seen that there is a gap in density between states of free flow $v_{\text{free}}$ and dynamic states of synchronized flow $v^{(syn)}$ within the MSP in the speed–density plane (Fig. 17.3b).

**Physics of Speed Breakdown**

An $F \rightarrow S$ transition and the related Z-shaped speed–flow characteristic can be explained by a competition between a tendency towards free flow due to over-acceleration, which is simulated by the acceleration noise in (16.31), and a tendency towards synchronized flow due to the adaptation of the vehicle speed to the speed of the preceding vehicle. This adaptation is simulated with (16.26), (16.27) due to the synchronization distance $D_n$ and the deceleration noise in (16.31).

The over-acceleration is harder at a higher vehicle speed, i.e., at a lower density (the flow rate is approximately the same in free and synchronized flows (see Fig. 5.10b,c)). This causes an $S \rightarrow F$ transition and the MSP dissolves (see Fig. 5.11). In contrast, the tendency towards speed adaptation is harder at lower speeds, i.e., at higher density. As a result, synchronized flow is self-sustaining within the MSP (Fig. 17.1c).

**17.2.2 Breakdown Phenomenon at On-Ramp Bottlenecks**

In the models of Chap. 16, rather than moving jams an $F \rightarrow S$ transition occurs spontaneously at an on-ramp bottleneck if the flow rate to the on-ramp $q_{\text{on}}$ gradually increases from zero and the flow rate on the main road upstream of the on-ramp $q_{\text{in}}$ is high enough [329,331].
Deterministic Perturbation

The numerical simulations of the spatial continuum model of Sect. 16.3 show the following features [329,330]. The average speed in states of free flow on the main road in the vicinity of the bottleneck \( v_{\text{free}}^{(B)} \) is lower than the speed in free flow on a homogeneous road: \( v_{\text{free}}^{(B)} < v_{\text{free}} \). The free flow states \( v_{\text{free}}^{(B)} \) are dynamic states caused by an influence of the bottleneck on free flow. The speed \( v_{\text{free}}^{(B)} \) decreases when \( q_{\text{on}} \) increases (Fig. 17.4).

This behavior of free flow is related to the occurrence of an internal localized perturbation (labeled “localized perturbation” in Fig. 17.5a) on the main road in the vicinity of the on-ramp caused by vehicle merging. The perturbation amplitude oscillates over time: the perturbation consists of permanent (“deterministic”) and random components.

Spontaneous Nucleation of SP

At a given flow rate \( q_{\text{in}} \) in an initial free flow in a range of the flow rate \( q_{\text{on}} \) (Fig. 17.4)

\[
q_{\text{on}}^{(\text{th})} \leq q_{\text{on}} \leq q_{\text{on}}^{(\text{determin., FS})}
\]

(17.3)

![Fig. 17.4. Z-shaped dependence of speed on the flow rate to the on-ramp \( q_{\text{on}} \) (a control parameter) at a fixed flow rate on the main road upstream of the bottleneck \( q_{\text{in}} \). The speed \( v_{\text{free}}^{(B)} \) is related to free flow on the main road in the vicinity of the on-ramp; the critical branch \( v_{\text{cr, FS}}^{(B)} \) gives the speed within the critical local perturbation required for an F→S transition as a function of \( q_{\text{on}} \); \( v_{\text{syn}}^{(B)} \) is associated with dynamic synchronized flow speeds on the main road in the vicinity of the on-ramp. Taken from [330]](image-url)
due to the self-growth of an internal perturbation a *spontaneous* nucleation of synchronized flow can occur: free flow is metastable with respect to a first-order spontaneous $F \rightarrow S$ transition leading to SP emergence (Fig. 17.5a).

SP nucleation occurs after a time delay $T_{FS}^{(B)}$ (Fig. 17.5b). The metastability and nucleation effects are related to the competition between over-acceleration and speed adaptation in synchronized flow considered above for a homogeneous road.

**Z-Characteristic at Bottlenecks**

The effect of SP nucleation is associated with a Z-shaped speed–flow characteristic (Fig. 17.4).

This Z-shaped dependence of speed on the flow rate $q_{on}$ at a given flow rate $q_{in}$ consists of the branch of free flow $v_{free}^{(B)}$, of a 2D region of dynamic synchronized flow speeds $v_{syn}^{(B)}$ (dashed region), and of the critical branch $v_{cr, FS}^{(B)}$ that determines the critical amplitude

$$ \Delta v_{cr, FS}^{(B)} = v_{free}^{(B)} - v_{cr, FS}^{(B)} $$

(17.4)

of local perturbations; if the perturbation amplitude exceeds the critical amplitude (17.4), an $F \rightarrow S$ transition occurs (down-arrow in Fig. 17.4). Otherwise (up-arrow in Fig. 17.4), perturbations decay.

If the density in dynamic free flow states $v_{free}^{(B)}$ and the density in dynamic synchronized flow states $v_{syn}^{(B)}$ are calculated, then we find a gap in density between dynamic free flow and dynamic synchronized flow states in the speed–density plane (Fig. 17.6).
Random Time Delay for Speed Breakdown

The time delay $T_{FS}^{(B)}$ (Fig. 17.5b) is a random value. In different realizations, at the same $q_{on}$ and $q_{in}$ usually different $T_{FS}^{(B)}$ are found. The lower $q_{on}$ and/or $q_{in}$ are, the higher the mean time delay $T_{FS}^{(B, mean)}$ is.

At a higher $q_{in}$ during the time delay $T_{FS}^{(B)}$ many perturbations can occur spontaneously, grow and then decay on the main road in the vicinity of the on-ramp before the perturbation appears whose growth leads to an F$\rightarrow$S transition.

The critical perturbation amplitude of a local perturbation for the F$\rightarrow$S transition (17.4) and therefore the mean time delay $T_{FS}^{(B, mean)}$ decreases when $q_{on}$ increases. In the vicinity of some

$$q_{on} = q_{on}^{(determ, FS)}$$

(17.5)

an F$\rightarrow$S transition occurs after $T_{FS}^{(B)} \approx 1\text{–}2$ min. This value is comparable with the duration of the F$\rightarrow$S transition. Thus, the exact value $q_{on} = q_{on}^{(determ, FS)}$ cannot be found (small dashed parts of $v_{free}(q_{on})$ and $v_{cr, FS}(q_{on})$). At $q_{on} \geq q_{on}^{(determ, FS)}$ a deterministic F$\rightarrow$S transition is expected (dotted down-arrow in Fig. 17.4) (see Sect. 5.3.2).

Induced Speed Breakdown

External local perturbations can lead to an induced F$\rightarrow$S transition and to SP emergence (Fig. 17.7).
This induced speed breakdown occurs when the amplitude of an external perturbation is greater than $\Delta v_{cr}^{(b)}$ \((17.4)\) (Fig. 17.7a). Otherwise, the perturbation decays and free flow is recovered on the main road near the on-ramp (Fig. 17.7b).

### 17.3 Moving Jam Emergence and Double Z-Shaped Characteristics of Traffic Flow

#### 17.3.1 F→J Transition on Homogeneous Road

Simulations show that the mean velocity of the downstream front of a wide moving jam $v_g$ is the characteristic, i.e., unique, predictable, and reproducible parameter that is constant for given model parameters. This velocity together with the threshold point $(\rho_{min}, q_{out})$ determines the characteristic line $J$ for the downstream front of a wide moving jam (the line $J$ in Figs. 17.1c and 17.8a).

If the initial density in free flow is only slightly greater than $\rho_{min}$ (dotted part of the curve $F_1$ in Fig. 17.1b), after the maximum possible amplitude of the critical perturbation leading to a stop of a vehicle is chosen, it is necessary to maintain this stop for some time further (about 2–3 minutes at the density in free flow that is equal to $\rho_{min}$) for an F→J transition in free flow. In each case, the critical amplitude of a local perturbation required for the F→J transition is considerably greater than the critical amplitude of the perturbation required for an F→S transition. Thus, regardless of the initial density an F→J transition does not occur spontaneously in free flow. Instead, the F→S transition governs phase transitions in free flow (Sect. 17.2.1).
17.3 Moving Jam Emergence

17.3.2 S→J Transition on Homogeneous Road

It has been found that in accordance with the three-phase traffic theory (Sect. 6.3), the line \( J \) determines the threshold of moving jam excitation in synchronized flow.

Numerical simulations of the models of Chap. 16 show that all densities in steady states of synchronized flow related to the line \( J \) are threshold densities with respect to jam formation (S→J transition) (e.g., \( \rho_{\text{min} \, 1}^{(\text{syn})} \) and \( \rho_{\text{min} \, 2}^{(\text{syn})} \) are the threshold densities for the speeds \( v_1^{(\text{syn})} = 72 \text{ km/h} \) and \( v_2^{(\text{syn})} = 51 \text{ km/h} \), respectively, Fig. 17.8).

At the same speed, the higher the density, the less the critical amplitude \( \Delta v_{\text{cr}}^{(\text{SJ})} \) of a local perturbation for an S→J transition (curves \( S_j^{(1)} \) and \( S_j^{(2)} \) for the speeds \( v_1^{(\text{syn})} \) and \( v_2^{(\text{syn})} \), respectively, Fig. 17.8b). The critical amplitude \( \Delta v_{\text{cr}}^{(\text{SJ})} \) for the S→J transition reaches its maximum value at the threshold density (\( \rho_{\text{min} \, 1}^{(\text{syn})} \) and \( \rho_{\text{min} \, 2}^{(\text{syn})} \) in Fig. 17.8b). At the same difference between the initial density and threshold density, the lower the initial speed, the less the critical amplitude \( \Delta v_{\text{cr}}^{(\text{SJ})} \) (compare curves \( S_j^{(1)} \) and \( S_j^{(2)} \)).

The latter result enables us a simulation of F→S→J transitions (see Fig. 6.12). If due to an F→S transition, which is caused by the local perturbation applied at some location, synchronized flow of a relatively low vehicle speed occurs, then an S→J transition occurs later spontaneously at another road location.

---

**Fig. 17.8.** S→J transition. (a) Steady states of the model (2D region) in the flow–density plane together with the line \( J \) and two lines for constant synchronized flow speeds \( v_1^{(\text{syn})} \) and \( v_2^{(\text{syn})} \). (b) Density dependencies of the critical amplitude of local perturbations for an S→J transition (curves \( S_j^{(1)} \) and \( S_j^{(2)} \)) associated with the speeds \( v_1^{(\text{syn})} \) and \( v_2^{(\text{syn})} \) in (a), respectively. Taken from [329]
**Z-Characteristic for S→J Transition**

For a homogeneous road, there is a Z-shaped speed-flow characteristic (Fig. 17.9), which is related to the metastability of synchronized flow states \( v^{(\text{syn})} \) and to the *spontaneous* nucleation effect in this flow leading to a first-order S→J transition.

![Fig. 17.9. Z-Characteristic for an S→J transition on a homogeneous road. The critical branch \( v_{\text{cr}}^{(\text{SJ})} \) gives the speed within the critical local perturbation required for an S→J transition as a function of \( q_{\text{in}} \). The line \( v_{\text{min}} = 0 \) is the speed within a wide moving jam. The states of synchronized flow \( v^{(\text{syn})} \) are taken from Fig. 17.3a. Taken from [330] (17.6)

The Z-characteristic consists of the states \( v^{(\text{syn})} \) (dashed 2D region), of the critical branch \( v_{\text{cr}}^{(\text{SJ})} \) that determines the critical amplitude of local perturbations

\[
\Delta v_{\text{cr}}^{(\text{SJ})} = v^{(\text{syn})} - v_{\text{cr}}^{(\text{SJ})},
\]

and of the line \( v_{\text{min}} = 0 \) for the speed within a wide moving jam. The nucleation of an S→J transition occurs in the flow rate range

\[
q_{\text{th}} \leq q_{\text{in}} \leq q_{\text{in}}^{(\text{cr}, \text{SJ})}
\]

if the perturbation amplitude in synchronized flow exceeds the critical amplitude (17.6) (down-arrow in Fig. 17.9). Otherwise, the perturbation decays (up-arrow).

There is a random time delay \( T_{\text{SJ}} \) for an S→J transition. In the vicinity of \( q_{\text{in}} = q_{\text{in}}^{(\text{cr}, \text{SJ})} \) fluctuations cause the S→J transition during \( T_{\text{SJ}} \approx 5-10 \text{ min} \) that is comparable with the duration of the S→J transition; the exact point \( q_{\text{in}} = q_{\text{in}}^{(\text{cr}, \text{SJ})} \) cannot be found (small dashed parts of \( v^{(\text{syn})} \) and \( v_{\text{cr}}^{(\text{SJ})} \)).
Double Z-Characteristic

The Z-characteristic for an $F \rightarrow S$ transition (Fig. 17.3) together with the Z-characteristic for an $S \rightarrow J$ transition (Fig. 17.9) form a double Z-shape speed-density characteristic associated with $F \rightarrow S \rightarrow J$ transitions (Fig. 17.10).

![Double Z-characteristic diagram](image)

**Fig. 17.10.** Double Z-characteristic for $F \rightarrow S \rightarrow J$ transitions on a homogeneous road. The states $v_{\text{free}}$, $v_{\text{cr}}^{(FS)}$, and $v_{\text{cr}}^{(\text{syn})}$ are taken from Fig. 17.3a and the states $v_{\text{cr}}^{(SJ)}$ and $v_{\min} = 0$ are taken from Fig. 17.9. Taken from [330]

17.3.3 Moving Jam Emergence in Synchronized Flow Upstream of Bottlenecks

Model simulations show that in the case of a freeway bottleneck, first synchronized flow occurs at the bottleneck ($F \rightarrow S$ transition) and only later can wide moving jams emerge spontaneously at other freeway locations upstream of the bottleneck in that synchronized flow. Moving jams do not emerge spontaneously at the bottleneck. Thus, moving jams emerge spontaneously due to $F \rightarrow S \rightarrow J$ transitions only.

For an on-ramp bottleneck, there is also a Z-shaped speed-flow characteristic (Fig. 17.11). This Z-characteristic consists of dynamic states for synchronized flow $v_{\text{syn}}^{(B)}$ (dashed 2D region), of the critical branch $v_{\text{cr}, SJ}^{(B)}$ that determines the critical amplitude of local perturbations.

Different dynamic 2D states of synchronized flow are related to different critical perturbation amplitudes. Therefore, there should be an infinite number of different critical branches $v_{\text{cr}, SJ}^{(B)}$ for an $S \rightarrow J$ transition. The critical branches in Figs. 17.11, 17.12, and 17.13 are an average of these different critical branches.
Fig. 17.11. Z-characteristic for an S→J transition at on-ramp bottlenecks. The critical branch \( v_{cr, SJ}^{(B)} \) gives the speed within the critical local perturbation required for the S→J transition as a function of \( q_{on} \). States of synchronized flow \( v_{syn}^{(B)} \) are taken from Fig. 17.4. Taken from [330]

\[
\Delta v_{cr, SJ}^{(B)} = v_{syn}^{(B)} - v_{cr, SJ}^{(B)}, \quad (17.8)
\]

and of the line \( v_{min} = 0 \) for the speed within a wide moving jam.

The Z-characteristic shows the metastability of synchronized flow states \( v_{syn}^{(B)} \) and the related spontaneous nucleation effect leading to an S→J transition and to GP emergence. This occurs after a random time delay \( T_{SJ} \) in the range

\[
q_{on}^{(th, SJ)} \leq q_{on} \leq q_{on}^{(cr, SJ)}, \quad (17.9)
\]

if the perturbation amplitude exceeds the critical amplitude (17.8) (down-arrow, Fig. 17.11). Otherwise, the perturbation decays (up-arrow). At \( q_{on} \approx q_{on}^{(cr, SJ)} \), the time delay \( T_{SJ} \approx 10 \text{ min} \), which is comparable with the duration of an S→J transition (shown by dashed curves related to states \( v_{syn}^{(B)} \) and \( v_{cr, SJ}^{(B)} \)). For this reason, the exact point \( q_{on} = q_{on}^{(cr, SJ)} \) cannot be found.

**Double Z-Characteristic at On-Ramp Bottlenecks**

The Z-characteristic for an F→S transition (Fig. 17.4) *together with* the Z-characteristic for an S→J transition (Fig. 17.11) form a *double Z-shape speed-density characteristic* at the on-ramp bottleneck (Fig. 17.12).

The double Z-characteristic is associated with spontaneous F→S→J transitions and GP emergence. This is due to two different spontaneous nucleation effects in free flow \( v_{free}^{(B)} \) and synchronized flow \( v_{syn}^{(B)} \), respectively.
17.3 Moving Jam Emergence

\[ q_{in} = \text{const} \]

Fig. 17.12. Double Z-characteristic for F→S→J transitions at on-ramp bottlenecks. The states \( v_{\text{free}}^{(B)} \), \( v_{\text{cr,FS}}^{(B)} \), and \( v_{\text{syn}}^{(B)} \) are taken from Fig. 17.4 and the states \( v_{\text{cr, SJ}}^{(B)} \) are taken from Fig. 17.11. Taken from [330]

Fig. 17.13. Simplified double Z-characteristics at an on-ramp bottleneck for different ranges of the flow rate \( q_{in} \). Curve 1 for \( q_{in} > q_{out, th} \); curve 2 for \( q_{in} = q_{th} \); curve 3 for \( q_{out} < q_{in} < q_{th} \); curve 4 for \( q_{in} < q_{out} \). Taken from [330]
The double Z-characteristic depends on \( q_{in} \) (curves 1–4, Fig. 17.13). At

\[
q_{in} \geq q_{th} \text{ and } q_{on} = 0
\]  

(17.10)

there are gaps \( \Delta_S \) and \( \Delta_J \) in speed on the double Z-characteristic (Fig. 17.13, curve 1). The gap \( \Delta_S \) is related to the speed gap between states \( v_{cr}^{(FS)} \) and \( v_{syn} \) in Fig. 17.3a. The gap \( \Delta_J \) is related to the speed gap between states \( v_{cr}^{(SJ)} \) and \( v_{min} = 0 \) in Fig. 17.11. At \( q_{in} < q_{th} \) the gaps do not appear (curves 3, 4 in Fig. 17.13).

These simulation results confirm general conclusions of the three-phase traffic theory (Sect. 6.4).

### 17.4 Conclusions

(i) The introduction of the synchronization distance in the microscopic models based on three-phase traffic theory enables us to satisfy the main features of empirical phase transitions in freeway traffic discussed in Chap. 10. Rather than wide moving jams, synchronized flow can occur spontaneously in an initial free flow. In free flow, wide moving jams can occur only due to the sequence of F \( \rightarrow \) S \( \rightarrow \) J transitions. In other words, first an F \( \rightarrow \) S transition occurs spontaneously in an initial free flow. Only later, and at other freeway locations can an S \( \rightarrow \) J transition occur in that synchronized flow.

(ii) In accordance with the three-phase traffic theory (Chap. 5), there is a Z-characteristic for an F \( \rightarrow \) S transition. The Z-shaped speed–density characteristics are associated with the spontaneous nucleation of synchronized flow in an initial free flow either on a hypothetical homogeneous road or at a freeway bottleneck.

(iii) In accordance with the three-phase traffic theory (Chap. 6), there is a Z-characteristic for an S \( \rightarrow \) J transition. These Z-shaped speed–density characteristics are associated with the spontaneous nucleation of wide moving jams in synchronized flow either on a homogeneous road or upstream of the freeway bottleneck.

(iv) The sequence of F \( \rightarrow \) S \( \rightarrow \) J transitions are described by double Z-characteristics of freeway traffic. The sequence of F \( \rightarrow \) S \( \rightarrow \) J transitions are responsible for the spontaneous emergence of GPs in free flow.
18 Congested Patterns at Isolated Bottlenecks

18.1 Introduction

Due to phase transitions at an isolated bottleneck discussed in Sect. 17.2.2 and Sect. 17.3.3, a congested pattern can occur upstream of a bottleneck. In the three-phase traffic theory (Chap. 7) and in empirical observations (Chaps. 10 and 12), if only an F $\rightarrow$ S transition occurs, then one of the SPs should appear. When the sequence of F $\rightarrow$ S $\rightarrow$ J transitions occur, then one of the GPs should emerge.

These conclusions of the three-phase traffic theory and of freeway traffic observations are fully confirmed by numerical investigations of both the microscopic CA model and the spatial continuum and discrete-time model based on three-phase traffic theory [329–331]. In addition, based on these models a microscopic theory of the pinch effect in synchronized flow [329, 331] as well as a theory of weak and strong congestion [330] have been developed. Results of these theories are in accordance with empirical results [208, 213]. Furthermore, based on model simulations we can find possible regions in the diagram of congested patterns where different types of congested patterns can coexist. The latter means that there should be metastable regions of traffic demand where depending on initial local perturbation at the same traffic demand one of various types of congested patterns emerge. These nonlinear effects of congested pattern emergence and metastability at different freeway bottlenecks will be the aims of this chapter.

Firstly, we consider the diagram of congested patterns at isolated on-ramp bottlenecks. Here, we discuss various types of SPs, certain microscopic features of synchronized flow within these congested patterns, and a theory of the pinch effect in synchronized flow that leads to GP emergence. Secondly, strong and weak congestion conditions in congested patterns are discussed. Thirdly, we study the pattern metastability effects at on-ramp bottlenecks. Further, we will study congested patterns, which can occur at a merge bottleneck associated with a reduction of freeway lanes in the flow direction. We will find that only strong congestion can occur in the pinch region of an GP at the merge bottleneck. In contrast to this feature of the merge bottleneck, in the next consideration we will find that only weak congestion can occur in the pinch region of an GP at an off-ramp bottleneck. Finally, we will analyze results of a numerical study of congested pattern capacity at isolated
on-ramp bottlenecks. A consideration made in this chapter is based on results of [221, 329, 330].

18.2 Diagram of Congested Patterns at Isolated On-Ramp Bottlenecks

18.2.1 Synchronized Flow Patterns

Numerical examples of the diagram of different congested patterns, i.e., the regions of spontaneous occurrence of the patterns in the flow–flow plane whose coordinates are $q_{in}$ and $q_{on}$, have already been considered in Sect. 7.5. It has been mentioned that these diagrams are qualitatively similar to the diagram of congested patterns at an on-ramp bottleneck postulated in the three-phase traffic theory (Fig. 7.13).

In particular, there are two main boundaries in this diagram, $R^{(B)}_S$ and $S^{(B)}_W$ (Fig. 18.1a). Below and left of the boundary $R^{(B)}_S$ free flow occurs. Between the boundaries $R^{(B)}_S$ and $S^{(B)}_W$ different SPs emerge on the main road upstream of the on-ramp (Figs. 18.1b–d). There are three types of SPs: (1) widening synchronized flow pattern (WSP), (2) localized synchronized flow pattern (LSP), and (3) moving synchronized flow pattern (MSP). Right of the boundary $S^{(B)}_W$ wide moving jams occur spontaneously in synchronized flow, i.e., different GPs appear (Figs. 18.1e–g).

SPs exhibit the following common characteristic features (Fig. 18.2):

(i) The downstream front of an WSP is fixed at the on-ramp bottleneck (Fig. 18.1b). The upstream front of the WSP is continuously widening upstream. The WSP occurs above the boundary $W$ in the diagram in Fig. 18.1a.

(ii) Below the boundary $W$ an LSP occurs. As in WSP, the downstream front of the LSP is fixed at the on-ramp bottleneck. However, the upstream front of the LSP is localized on the main road at some distance $L_{LSP}$ upstream of the on-ramp bottleneck (Fig. 18.1c).

(iii) At higher $q_{in}$ and a very low $q_{on}$ an MSP can occur rather than WSP. The MSP occurs spontaneously on the main road at the on-ramp bottleneck because at very low $q_{on}$ the downstream front of synchronized flow can depart from the on-ramp bottleneck. The MSP begins to propagate on the main road as an independent localized structure (Fig. 18.1d), as on a homogeneous road (Fig. 7.4). A sequence of MSPs can also occur on the main road at the on-ramp bottleneck (Fig. 18.3).

(iv) The flow rate in an SP is often only slightly lower than the initial flow rate in free flow (Figs. 18.2a,b,e).1

1 This is a very important feature of SPs that will be used for congested pattern control at freeway bottlenecks (Sect. 23.3.2).
Fig. 18.1. Diagram of congested patterns at an isolated on-ramp bottleneck (a) and the related congested patterns (b–g). (b–d) SPs and (e–g) GPs. (b) WSP. (c) LSP. (d) MSP. (e) GP at \( q_{in} > q_{out} \). (f) GP at \( q_{in} < q_{out} \). (g) DGP. Taken from [330]
Fig. 18.2. Synchronized flow patterns. (a, b, e) Vehicle speed (left) and flow rate (right) in the SP shown in Figs. 18.1b–d, respectively. (a) WSP. (b) LSP. (c) Vehicle density in an LSP over time. (d) Width of an LSP (in the longitudinal direction) $L_{\text{LSP}}$ over time. (e) MSP. Data in the flow–density plane in (f) is shown for the WSP (a) and in (g) for the LSP (b). The dashed line in (c) is related to $\rho_{\text{max}}^{(\text{free})}$. Taken from [330]
Fig. 18.3. Sequence of MSPs. (a) Vehicle speed on the main road in space and time. (b, c) Speed (b) and flow rate (c) at two different locations on the main road as functions of time. Taken from [329]

(v) The vehicle density in an SP can often be approximately the same or even lower than $\rho_{\text{max}}^{(\text{free})}$ (Figs. 18.2c,f–g).

(vi) The width of the LSP $L_{\text{LSP}}$ (in the longitudinal direction), i.e., the distance of the upstream LSP front from the effective location of the on-ramp bottleneck, depends on time, and it can exhibit complex oscillations with large amplitude (Fig. 18.2d, where $2 \text{ km} \leq L_{\text{LSP}} \leq 4 \text{ km}$).

(vii) The mean width of the LSP $L_{\text{LSP}}^{(\text{mean})}$ can depend hardly on $q_{\text{in}}$ and $q_{\text{on}}$. In numerical simulations [330], $L_{\text{LSP}}^{(\text{mean})}$ varies between 0.5 km and 10 km.

At $q_{\text{on}}$ and $q_{\text{in}}$ associated with a neighborhood of the boundary $F_{S}^{(B)}$, regions of free flow sometimes occur within WSPs and LSPs. These patterns are called “SP with alternations of free and synchronized flow” (alternating SP or ASP for short).\footnote{An example of an ASP will be shown in Fig. 19.4c.} The appearance of an ASP can be explained by an S→F transition within an SP. There are two different cases of ASP emergence:

(i) Free flow regions occur on the main road at some distance upstream of the on-ramp; at the bottleneck synchronized flow is self-sustaining. Often regions of free flow appear only for a finite time, i.e., free flow is replaced by synchronized flow over time.
(ii) Free flow regions occur on the main road for short time intervals randomly at the on-ramp bottleneck. Free flow replaces partially synchronized flow upstream. In this case, there is some analogy with the occurrence of an MSP. However, in contrast with the case of the MSP, free flow returns on the main road at the on-ramp bottleneck for much shorter time intervals than the duration of synchronized flow on the main road at the on-ramp bottleneck.

Both single-lane and two-lane models of Sect. 16.3 show qualitatively the same diagram of congested patterns and congested patterns at the bottleneck (Figs. 7.14 and 18.1) [329,332]. This is because the over-acceleration is described in the model as a “collective effect” that occurs on average in traffic flow (Sect. 16.3.7). In an example of simulations of an WSP on a two-lane road, qualitatively the same WSP appears spontaneously in the left and right lanes of the road (Fig. 18.4a,b). Distributions of speed and flow rate in both lanes are almost identical (Fig. 18.4c,d). This is an effect of identical vehicle parameters and driver behavioral characteristics used for all vehicles in the model. If different driver behavioral characteristics and different vehicle parameters are used in the two-lane model, then some new specific effects are observed (Chap. 20). However, different driver behavioral characteristics and different vehicle parameters do not qualitatively change the main fundamental features of spatiotemporal congested patterns discussed in this chapter.

18.2.2 Single Vehicle Characteristics in Synchronized Flow

It is obvious that already small amplitude random fluctuations in synchronized flow destroy hypothetical steady states of synchronized flow (Sect. 4.3.3): these model steady states are only related to hypothetical unperturbed and noiseless vehicle motion that does not occur in reality. Besides fluctuations there are dynamic effects in synchronized flow that destroy steady states. Examples of these dynamic effects are different driver time delays. These time delays lead to vehicle speed differences. There is a competition of these time delay effects with the speed adaptation effect in synchronized flow that attracts vehicles to a region of small speed differences. Thus, rather than steady states some dynamic spatiotemporal synchronized flow states appear after an SP occurs [205,329,331]. Examples of these dynamic states of synchronized flow are dependencies of speed and space gap of time for two different vehicles moving through an WSP that occurs at an on-ramp bottleneck (Figs. 18.4 and 18.5a–d).

Features of dynamic states of synchronized flow found in simulations (Fig. 18.5) are correlated with empirical single vehicle data in real traffic flow (e.g., [88,483,484]). In particular, as in empirical data [88,483] these dynamic states cover a 2D region in the space gap–speed plane (Fig. 18.5e,f). In synchronized flow, as in empirical results [484] the vehicle acceleration $\alpha_n$ and also the vehicle speed difference $\delta v_n = v_{t,n} - v_n$ (the difference between
the vehicle speed and the speed of the preceding vehicle) as functions of time exhibit random jumps and drops in the vicinity of $\alpha_n = 0$ and $\delta v_n = 0$, respectively (Fig. 18.6).

In accordance with an empirical data analysis made by Wagner and Lubashevsky [484], model frequency distributions $p_{\delta v_n}(\delta v)$ of speed differences between vehicles in synchronized flow have very sharp maximum at $\delta v = 0$ (Fig. 18.7) [332]. The above behavior of frequency distributions $p_{\delta v_n}(\delta v)$ is valid for different vehicle speeds and space gaps in synchronized flow (Fig. 18.7). This attraction of vehicles in synchronized flow to a region with small speed differences is associated with the speed adaptation effect within the synchronization distance. This speed adaptation effect in synchronized flow is a fundamental feature of three-phase traffic theory (Sects. 4.3.2 and 16.3.2).

Simulations show that features of phase transitions and congested pattern formation in these complicated dynamic synchronized flow states (Figs. 18.5–18.7) are qualitatively the same as those in hypothetical steady states of synchronized flow postulated in the three-phase traffic theory (Chaps. 5–8).
Fig. 18.5. Single vehicle characteristics in synchronized flow in the WSP shown in Fig. 18.4a. (a–d) Speed (a, c) and space gap (b, d) as functions of time for two vehicles moving through the WSP. (e, f) Points in the space gap–speed plane related to (a, c) and (b, d). Figures (a, c, e) are for vehicle 1 and figures (b, d, f) are for vehicle 2. Vehicle 1 and the related preceding vehicle move in the right lane only, vehicle 2 starts to move in the left lane, then it changes to the right lane at \( t^{(1)} = 48 \) min; the related preceding vehicle changes from the right lane to the left lane at \( t^{(2)} = 49 \) min. The lines \( F, U, \) and \( L \) in (e, f) are the boundaries for steady states in the space gap–speed plane related to the corresponding boundaries for the steady states in the flow–density plane in Fig. 16.1a. Taken from [332]

This confirms the assumption of Sect. 4.3.4 that if dynamic synchronized flow states of a traffic flow model exhibit features of steady states of synchronized flow with respect to phase transitions and congested pattern formation postulated in the three-phase traffic theory, then such traffic flow models can also explain and predict some empirical features of spatiotemporal congested patterns.
Fig. 18.6. Speed difference $\delta v_n = v_{t,n} - v_n$ (a, b) and vehicle acceleration $a_n$ (c, d) for vehicle 1 (a, c) and vehicle 2 (b, d) related to time intervals marked by arrows in Fig. 18.5a–d. Taken from [332]

Fig. 18.7. Frequency distributions $p_{\delta v}$ of speed difference $\delta v$ between two following one another vehicles related to different space gap intervals and different vehicle speed intervals in synchronized flow within the WSP shown in Fig. 18.4. Space gaps between vehicles are: 15–25 m (curves 1), 25–35 m (curves 2), and 35–45 m (curves 3). Taken from [332]
18.2.3 Maximum Freeway Capacities and Limit Point in Diagram

In accordance with a theory of freeway capacity in free flow at a bottleneck discussed in Sect. 8.3, numerical simulations show that there are an infinite number of maximum freeway capacities \( q_{\max}^{(\text{free B})} \) in free flow at an on-ramp bottleneck. These maximum capacities are related to the boundary \( F_{S}^{(B)} \):

\[
q_{\max}^{(\text{free B})} = q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}) \mid _{p_{S}^{(B)}},
\]

where

\[
q_{\text{sum}} = q_{\text{on}} + q_{\text{in}}
\]

is the flow rate in free flow on the main road downstream of the on-ramp under a free flow condition at the bottleneck.

For the spatial continuum model of Sect. 16.3 the maximum freeway capacity can be a weak function of \( q_{\text{on}} \) with a minimum (Fig. 8.4a). For the KKW CA model (Sect. 16.2) the maximum freeway capacity can be a strong decreasing function of \( q_{\text{on}} \) with a saturation at higher \( q_{\text{on}} \) (Fig. 8.4b).

At the limit point in the diagram of congested patterns related to \( q_{\text{on}} \rightarrow 0 \) (but \( q_{\text{on}} \neq 0 \)) (Fig. 18.1a), the maximum capacity

\[
q_{\max}^{(\text{free B})} \mid _{q_{\text{on}} \rightarrow 0} = q_{\max, \lim}^{(\text{free B})} \quad \text{(but } q_{\text{on}} \neq 0). \tag{18.3}
\]

The maximum capacity \( q_{\max, \lim}^{(\text{free B})} \) is lower than the maximum flow rate \( q_{\max}^{(\text{free})} \) for a homogeneous road:

\[
q_{\max, \lim}^{(\text{free B})} < q_{\max}^{(\text{free})}. \tag{18.4}
\]

We have already explained this result in the three-phase traffic theory by additional random local perturbations on the main road in the vicinity of the on-ramp that should occur when \( q_{\text{on}} \rightarrow 0 \) (Sect. 7.4.1). This qualitative theory is also confirmed by the microscopic three-phase traffic theory under consideration.

As \( q_{\text{on}} \rightarrow 0 \), i.e., in the vicinity of the flow rate \( q_{\text{sum}} = q_{\max, \lim}^{(\text{free B})} \) (18.3), a local short-time fluctuation can play the role of a nucleus for spontaneous MSP emergence at the on-ramp bottleneck. After an MSP has emerged, it leaves the on-ramp bottleneck and begins to propagate on the main road as an independent localized structure (Fig. 18.1d), as on a homogeneous road (Fig. 7.4).

18.2.4 Pinch Effect in General Patterns

Right of the line \( G \) and right of the boundary \( S_{I}^{(B)} \) an GP occurs where the pinch region of synchronized flow continuously exists and a sequence of wide
moving jams emerges (Figs. 18.1e,f). The pinch region occurs as a result of the pinch effect in synchronized flow. The pinch effect is a self-compression of synchronized flow. In a microscopic theory of the pinch effect [329], it has been shown that in accordance with the three-phase traffic theory (Sect. 7.6.2) and empirical results (Sect. 12.2), the pinch effect exhibits the following features:

(i) Due to the pinch effect the density is high and the average speed is low in the pinch region of synchronized flow. However, when the pinch effect occurs the flow rate does not necessarily decrease. Even if the flow rate decreases due to the pinch effect the relative value of this decrease is considerably lower than the decrease in average speed.

(ii) In the pinch region of an GP, narrow moving jams emerge spontaneously and grow (Figs. 18.8 and 18.9a, \( x = 15.8 \) km and \( x = 14.5 \) km).

(iii) Points related to the pinch region lie above the line \( J \) in the flow–density plane (Fig. 18.9b).

(iv) The velocity of narrow moving jams in the pinch region is more negative than the characteristic velocity \( v_g \) of the downstream front of a wide moving jam: the slope of the line \( J \) is equal to \( v_g = -15.5 \text{ km/h} \).

The upstream front and the downstream front of a narrow moving jam are shown in Fig. 18.10 by the lines \( J^{(\text{up})} \) and \( J^{(\text{down})} \), respectively. The slope of the line \( J^{(\text{up})} \) is equal to the average velocity of the upstream front \( v_{\text{narrow}} \) of the narrow moving jam. The slope of the line...
Fig. 18.9. Data in the flow–density plane for the GP shown in Fig. 18.8. (a) One minute average data of virtual detectors. (b) Average data in the pinch region of the GP. In (a) data related to moving jams is labeled ⊗. Taken from [330]

\( J_{\text{narrow}}^{(\text{down})} \) is equal to the average velocity of the downstream front \( v_{\text{narrow}}^{(\text{down})} \) of the narrow moving jam.\(^3\)

(v) Corresponding to Fig. 18.10d, \( |v_{\text{narrow}}^{(\text{up})}| > |v_{\text{narrow}}^{(\text{down})}| \), i.e., the jam width increases over time. The velocities \( v_{\text{narrow}}^{(\text{up})} \) and \( v_{\text{narrow}}^{(\text{down})} \) are functions of time.

(vi) When a narrow moving jam transforms into a wide moving jam, \( v_{\text{narrow}}^{(\text{down})} \rightarrow v_{\text{g}} \) and the line \( J_{\text{narrow}}^{(\text{down})} \) tends to the line \( J \). The location of this \( S \rightarrow J \) transition is related to the upstream boundary of the pinch region of the GP. Different narrow moving jams can transform into

\(^3\) The velocities \( v_{\text{narrow}}^{(\text{up})} \approx -23.5 \text{ km/h} \) and \( v_{\text{narrow}}^{(\text{down})} \approx -17.3 \text{ km/h} \). These front velocities have been estimated based on the Stokes shock-wave formula (3.5) where data shown in Fig. 18.10d is used.
Fig. 18.10. Characteristics of narrow moving jams in the pinch region of the GP (a) taken from Fig. 18.1e. (b–d) Speed (b) and flow rate (c) as functions of space, and the points in the flow–density plane (d) at \( t = 40\min 50\sec \) for the narrow moving jam labeled “jam 3” in (a). The points in (d) are related to the space intervals marked in (b, c). Each of these points corresponds to data averaged over 10 vehicles. Taken from [332]

wide moving jams at different locations. For this reason, the upstream boundary of the pinch region can exhibit complicated oscillations over time.

(vii) After a narrow moving jam has transformed into a wide moving jam, this wide moving jam can suppress the growth of the downstream narrow moving jam (Fig. 18.8, \( x = 13\km \)). In this case, the mean time \( T_j \) between narrow moving jams emerging in the pinch region is appreciably lower than the mean time \( T_j^{(\text{wide})} \) between the downstream fronts of wide moving jams (in Fig. 18.8, \( T_j \approx 6\min \) at \( x = 14.5\km \) and \( T_j^{(\text{wide})} \approx 30\min \) at \( x = 8\km \)). However, this jam suppression effect occurs only if the narrow moving jam is close to the downstream wide moving jam front.

(viii) Distances between different narrow moving jams in the pinch region can be very different to one another. The average distance between initial narrow moving jams in the pinch region depends on average speed in the pinch region (see Sect. 18.3). Let the distance between the initial narrow moving jams emerging in the pinch region \( R_{\text{narrow}} \) be low, i.e., the frequency of narrow jam emergence in the pinch region

\[
  f_{\text{narrow}} = \frac{1}{T_j}
\]  

(18.5)
is high. Then only some of the initial narrow moving jams transform into wide moving jams; the other narrow moving jams either dissolve or merge with other moving jams. With model parameters (Sect. 16.3.10) this case is related to $R_{\text{narrow}} < 2.5 \text{ km}$ and the related frequency $f_{\text{narrow}} > 0.1 \text{ min}^{-1}$. If in contrast $R_{\text{narrow}}$ is relatively high, i.e., $f_{\text{narrow}}$ is low, then almost every narrow jam transforms into a wide moving jam (see an GP in Fig. 18.11a).

These numerical results are in accordance with empirical results of Chap. 12 and they confirm the qualitative theory of the pinch effect discussed in Sect. 7.6.2.

**18.2.5 Peculiarities of General Patterns**

**Dissolving GP**

There is a region in the diagram right of the boundary $S_j^{(B)}$ where a dissolving GP (DGP) occurs (Fig. 18.1g).

In a DGP, after a wide moving jam has been formed this wide moving jam suppresses the pinch region in the initial GP. The suppression effect can occur when the initial flow rate $q_{\text{in}}$ is high enough and the flow rate to the on-ramp $q_{\text{on}}$ is lower than some characteristic value that determines the position of the boundary $G$ in the diagram of congested patterns in Fig. 18.1a. To understand the physics of the DGP, let us consider flow rates $q_{\text{in}}$ that satisfy the condition

$$q_{\text{in}} > q_{\text{out}}.$$  \hspace{1cm} (18.6)

After the first wide moving jam has been formed in the GP, the flow rate in the jam outflow cannot be higher than $q_{\text{out}}$. This means that rather than an initial high flow rate $q_{\text{in}}$ (18.6) a lower flow rate related to the jam outflow determines the inflow in the pinch region of the GP. At a low enough flow rate to the on-ramp $q_{\text{on}}$ (left of the boundary $G$ in the diagram of congested patterns in Fig. 18.1a), the pinch region cannot exist at this lower inflow into the pinch region: the pinch region dissolves.

As a result of this dissolution effect, either free flow or an LSP appears on the main road at the on-ramp bottleneck while the wide moving jam moves on the main road upstream of the on-ramp (Fig. 18.1g). In the diagram of congested patterns at the on-ramp bottleneck between the boundary $S_j^{(B)}$ and the line $G$, an DGP emerges. From the above consideration of the physics of an DGP based on the analysis of the condition (18.6), one can conclude that the boundary $G$, which separates DGPs and GPs in the diagram of congested patterns, should intersect the boundary $S_j^{(B)}$ at the point $q_{\text{in}} = q_{\text{out}}$. This is correct if hysteresis effects are not taken into account. However, the hysteresis effect of GP formation (Sect. 18.5) leads to the result that the line $G$ does not necessarily intersect the boundary $S_j^{(B)}$ at the point $q_{\text{in}} = q_{\text{out}}$ as is shown in Fig. 18.1a.
Fig. 18.11. Weak congestion and strong congestion in an GP at an on-ramp bottleneck. (a, b) Speed in space and time in a GP under the weak congestion condition on the main road (a) and in the on-ramp lane (b). (c) Vehicle speed (left) and flow rate (right) related to (b). (d, e) Speed in space and time in a GP under the strong congestion condition on the main road (d) and in the on-ramp lane (e). (f) Vehicle speed (left) and flow rate (right) related to (e). (g) Spatial distributions of speed (left) and flow rate (right) in the on-ramp lane at $t = 27$ min related to (e). Taken from [330]
Dissolution of Wide Moving Jams in GP

If $q_{in} > q_{out}$, then the width of the farthest upstream wide moving jam is a gradually increasing function of time (Fig. 18.1e).

If in contrast $q_{in} < q_{out}$, the farthest upstream wide moving jam gradually dissolves (see the dissolution of the first three wide moving jams in an GP in Fig. 18.1f). However, simulations show that even if $q_{in} < q_{out}$, the region of wide moving jams can expand in the upstream direction.

GP of Types (1) and (2)

The empirical result about the existence of two types (1) and (2) of GP (Sect. 9.4) is also confirmed in the microscopic three-phase traffic theory.

In many cases, an GP consists of the pinch region in synchronized flow and the region of wide moving jams. The boundary, which spatially separates these two regions of the GP, is the upstream front of synchronized flow where an $S \rightarrow J$ transition occurs. In the GP of type (1), the width (in the longitudinal direction) of synchronized flow in the GP $L_{syn}^{(pinch)}$, i.e., the distance between the downstream front of synchronized flow in the GP (located at the effective location of the on-ramp bottleneck) and the upstream front of synchronized flow is spatially limited.

In contrast, there is a case of GP of type (2), where the width of synchronized flow in the GP $L_{syn}$ can be a continuously increasing function over time. This can occur above the boundary $W$ in Fig. 18.1a when $q_{on}$ is gradually increased. Firstly, an WSP occurs, then an DGP appears, and finally an GP is realized. In this case, the pinch region and moving jam emergence in the DGP and the GP occur within the initial WSP. It can turn out that the upstream front of the initial WSP propagates upstream quicker than the upstream front of the wide moving jam (Fig. 18.1e), i.e., the upstream front of the entire GP can be determined by the upstream front of the initial WSP, rather than by the farthest upstream wide moving jam. Upstream of this jam synchronized flow associated with the initial WSP exists and propagates upstream.

18.3 Weak and Strong Congestion in General Patterns

18.3.1 Criteria for Strong and Weak Congestion

It has been found that there is a special flow rate to the on-ramp

$$q_{on} = q_{on}^{(strong)}.$$  (18.7)
If the condition

\[ q_{on} < q_{on}^{\text{(strong)}} \]  

(18.8)
is satisfied, then the average flow rate \( q^{\text{(pinch)}} \) in the pinch region of an GP is a decreasing function of traffic demand, i.e., of \( q_{on} \) (Fig. 18.12a).

The average vehicle speed \( v^{\text{(pinch)}} \) in the pinch region displays the same behavior (Fig. 18.12b). The averaging time interval for \( q^{\text{(pinch)}} \) and \( v^{\text{(pinch)}} \) is chosen to be longer than \( T_j \). Also the mean width of the pinch region, \( L_{\text{syn}}^{\text{(mean)}} \), the mean distance between narrow moving jams in the pinch region, \( R_{\text{narrow}} \), and the related value \( T_j \) are decreasing functions of \( q_{on} \) (Figs. 18.12c–e). Consequently, the frequency \( f_{\text{narrow}} \) (18.5) is an increasing function of \( q_{on} \).

In empirical investigations (Sect. 12.3), this case is called “weak congestion” (an GP in Figs. 18.11a–c).

During weak congestion conditions, the flow rate \( q^{\text{(on)}} \) of vehicles, merging onto the main road from the on-ramp, is almost equal to \( q_{on} \) (Fig. 18.12f).

Fig. 18.12. Parameters of GPs. \( q^{\text{(pinch)}} \) (a), \( v^{\text{(pinch)}} \) (b), \( L_{\text{syn}}^{\text{(mean)}} \) (c), \( R_{\text{narrow}}^{\text{(mean)}} \) (d), \( T_j \) (e), \( q^{\text{(on)}} \) (f), and \( v^{\text{(on)}} \) (g) as functions of \( q_{on} \). Taken from [330]
In this case, the vehicle speed in the on-ramp lane is approximately equal to free flow speed in the on-ramp lane: \( v^{(on)} \approx v^{\text{free on}} \) (Fig. 18.12g; see also Figs. 18.11b,c; \( v^{(on)} \) and \( q^{(on)} \) are related to figures labeled “16 km (on-ramp)”).

**Strong Congestion**

In contrast, when \( q_{on} \geq q_{on}^{(\text{strong})} \),

\[
q_{on} \geq q_{on}^{(\text{strong})}, \tag{18.9}
\]

the mean characteristics of the pinch region and of narrow moving jam emergence no longer depend on traffic demand, i.e., on \( q_{on} \) any more.

The mean values \( q^{(\text{pinch})} \), \( v^{(\text{pinch})} \), \( L_{\text{syn}}^{(\text{mean})} \), \( R_{\text{narrow}}^{(\text{mean})} \), and \( T_{J} \) as functions of \( q^{(on)} \) are saturated and reach some limit (minimum) values:

\[
q^{(\text{pinch})} = q_{\text{lim}}^{(\text{pinch})}, \quad v^{(\text{pinch})} = v_{\text{lim}}^{(\text{pinch})}, \quad L_{\text{syn}}^{(\text{mean})} = L_{\text{syn, lim}}^{(\text{mean})}, \tag{18.10}
\]

\[
R_{\text{narrow}}^{(\text{mean})} = R_{\text{narrow, lim}}^{(\text{mean})}, \quad T_{J} = T_{J, \text{lim}}. \tag{18.11}
\]

(Figs. 18.12a–e). Correspondingly, \( f_{\text{narrow}} \) (18.5) reaches a maximum

\[
f_{\text{narrow}} = f_{\text{max}} = 1/T_{J, \text{lim}}. \tag{18.12}
\]

In empirical observations, this case has been called “strong congestion” (Sect. 12.3). Note that \( L_{\text{syn}}^{(\text{pinch})}(t) \) changes in the vicinity of \( L_{\text{syn, lim}}^{(\text{mean})} \) (Fig. 18.13). This is because transformations of different narrow moving jams into wide moving jams occur at different locations.

If \( q_{on} \) appreciably exceeds \( q_{on}^{(\text{strong})} \), then there is a saturation effect in the boundary \( S_{J}^{(B)} \). This effect is realized at a considerably lower flow rate \( q_{in} = q_{\text{lim}}^{(\text{pinch})} \) than the flow rate \( q_{out} \) (Fig. 18.1a). Only when

\[
q_{in} < q_{\text{lim}}^{(\text{pinch})}, \tag{18.13}
\]

![Fig. 18.13. Width of synchronized flow \( L_{\text{syn}}^{(\text{pinch})} \) (in the longitudinal direction) in a GP under the strong congestion condition as a function of time. Taken from [330]](image-url)
an GP does not exist (see also Sect. 18.5). Thus, the GP can exist only if

$$q_{\text{in}} \geq q_{\text{lim}}^{(\text{pinch)}}.$$  \hspace{1cm} (18.14)

It is found that there is a limit flow rate to the on-ramp

$$q_{\text{on}} = q_{\text{on}}^{(G)}.$$ \hspace{1cm} (18.15)

(see \(q^{(\text{pinch)}}(q_{\text{on}})\) in Fig. 18.12a). This flow rate to the on-ramp gives the line \(G\) in the diagram (Fig. 18.1a). The flow rate (18.15) is related to the condition

$$q^{(\text{pinch)}}|_{q_{\text{on}}=q_{\text{on}}^{(G)}} = q_{\text{out}}.$$ \hspace{1cm} (18.16)

The GP exists only if

$$q^{(\text{pinch)}} \leq q_{\text{out}} \text{ at } (18.17)$$

$$q_{\text{on}} \geq q_{\text{on}}^{(G)}.$$ \hspace{1cm} (18.18)

At

$$q_{\text{on}} < q_{\text{on}}^{(G)}$$ \hspace{1cm} (18.19)

we have

$$q^{(\text{pinch)}} > q_{\text{out}}.$$ \hspace{1cm} (18.20)

In this case, the pinch region dissolves and the GP cannot exist.

To explain the condition (18.17), note that the average flow rate in the pinch region \(q^{(\text{pinch)}}\) is equal to the average flow rate in the region of wide moving jams. The latter cannot exceed \(q_{\text{out}}\): the flow rate \(q_{\text{out}}\) is the maximum flow rate in the wide moving jam outflow that is achieved when free flow is formed in the jam outflow.

### 18.3.2 Strong Congestion Features

The following features of strong congestion in an GP have been found in the microscopic three-phase traffic flow theory [330]:

(i) The mean time between the downstream fronts of wide moving jams \(T_{J}^{(\text{wide})}\) under strong congestion is less (Fig. 18.11d) than under weak congestion (Fig. 18.11a). Because the mean duration of wide moving jams, \(\tau_{J}\), is similar in the two cases (\(\tau_{J} \approx 2\text{ min}\) in Figs. 18.11d and 18.11a) the mean time between wide moving jams, \(T_{\text{int}}^{(\text{wide})} = T_{J}^{(\text{wide})} - \tau_{J}\), is very different: the value \(T_{\text{int}}^{(\text{wide})}\) under strong congestion can be considerably less than under weak congestion.

(ii) The limit characteristics \(q_{\text{lim}}^{(\text{pinch)}}\), \(v_{\text{lim}}^{(\text{pinch)}}\), and \(T_{J, \text{lim}}\) are decreasing functions of length of the merging region \(L_{m}\) whereas the discharge flow rate \(q_{\text{out}}^{(\text{bottle})}\) is independent of \(L_{m}\) (Figs. 18.14a–d). \(q_{\text{out}}^{(\text{bottle})}\) is the flow rate determined on the main road downstream of the on-ramp where free flow is realized.
(iii) There is a correlation between features of wide moving jams and $q^{(\text{pinch})}$. Under the strong congestion condition

$$q^{(\text{pinch})} = q_{\text{lim}}^{(\text{pinch})}. \quad (18.21)$$

From (18.21) taking into account (7.37) and (7.40), we obtain

$$q_{\text{lim}}^{(\text{pinch})}/q_{\text{out}} = T_{\text{int}}^{(\text{wide})}/T_{J}^{(\text{wide})} < 1. \quad (18.22)$$

For the example of the GP in Fig. 18.11d, $q_{\text{lim}}^{(\text{pinch})} = 1470$ vehicles/h, therefore $q_{\text{out}}/q_{\text{lim}}^{(\text{pinch})} = 1.23$.

(iv) The greater the flow rate $q_{\text{out}}$, the greater the flow rate $q_{\text{lim}}^{(\text{pinch})}$. This dependence is approximately linear (Fig. 18.15a). $q_{\text{out}}^{(\text{bottle})}$ also increases with $q_{\text{out}}$ (Fig. 18.15b). However, this dependence is much weaker than the former one. Note that the condition $q_{\text{lim}} + q_{\text{on}} > q_{\text{out}}^{(\text{bottle})}$ is satisfied. Thus, $q_{\text{out}}^{(\text{bottle})}$ gives the congested pattern capacity (Sect. 8.5).
(v) Each maximum freeway capacity in free flow $q_{\text{max}}^{(\text{free B})}$ is considerably greater than $q_{\text{lim}}^{(\text{pinch})}$:

$$q_{\text{max}}^{(\text{free B})} > q_{\text{lim}}^{(\text{pinch})}.$$  \hspace{1cm} (18.23)

For chosen model parameters (Sect. 16.3.10) $q_{\text{max,lim}}^{(\text{free B})}/q_{\text{lim}}^{(\text{pinch})} = 1.63$.

(vi) A GP under the strong congestion condition also occurs in the on-ramp lane. In this case, the flow rate of vehicles that can merge onto the main road from the on-ramp, $q_{\text{on}}^{(\text{on})}$, is saturated and reaches a limit value $q_{\text{on}}^{(\text{on})} = q_{\text{lim}}^{(\text{on})}$ (Fig. 18.12f). A further increase in $q_{\text{on}}$, at the beginning of the on-ramp lane (at $x = x_{\text{on}}^{(b)} = 14$ km, Fig. 16.2a), does not lead to an increase in $q_{\text{on}}^{(\text{on})}$ any more. The dependence of $v_{\text{on}}^{(\text{on})}$ on $q_{\text{on}}$ exhibits a sharp decrease when the transition to strong congestion occurs in the on-ramp lane (Fig. 18.12g). If $q_{\text{on}}$ is initially set much higher than $q_{\text{lim}}^{(\text{on})}$, the abrupt decrease in speed and flow rate occurs first in the vicinity of the merging region of the on-ramp (Fig. 18.11f). Then the front separating lower and higher speeds (and flow rates) in the on-ramp lane propagates upstream to $x = x_{\text{on}}^{(b)}$ (Fig. 18.11e). Accordingly, a pinch region is formed in the on-ramp lane. Growing narrow moving jams emerge in this pinch region and finally some of these jams transform into wide moving jams (Fig. 18.11g). In contrast to the main road, where narrow moving jams transform into wide moving jams at some distance (about 1.5–2 km in Fig. 18.8) upstream from the on-ramp, some wide moving jams in the on-ramp lane even occur within the merging region: some of the vehicles that move in synchronized flow under the strong congestion condition in the on-ramp lane have to stop before they can merge onto the main road. Moreover, the transition from weak to strong congestion occurs at a slightly higher flow rate $q_{\text{on}}$ than the characteristic flow rate $q_{\text{on}}^{(\text{strong})}$ (18.7) (Fig. 18.12g).

18.4 Evolution of Congested Patterns at On-Ramp Bottlenecks

To simulate pattern evolution, the flow rates $q_{\text{on}}(t)$ and $q_{\text{in}}(t)$ from the empirical study of pattern evolution (Sect. 13.2) are used (Fig. 18.16a). In accordance with these empirical results, we found that during the increase in $q_{\text{on}}(t)$ and $q_{\text{in}}(t)$ first an F—S transition occurs spontaneously on the main road at the on-ramp bottleneck at $t = 6:12$ and then synchronized flow appears on the main road upstream of the on-ramp (Figs. 18.16b,c at $t = 6:20$). Because the flow rate $q_{\text{on}}(t)$ further increases sharply (Fig. 18.16a) the pinch effect is realized in this synchronized flow leading to GP emergence (Fig. 18.16c, $t = 9:12$). The maximum of $q_{\text{on}}(t)$ exceeds $q_{\text{on}}^{(\text{strong})}$ appreciably. Thus, as in the empirical Fig. 12.1, within the GP the strong congestion condition is realized.
Fig. 18.16. Evolution of congested patterns at an on-ramp bottleneck. (a) $q_{on}(t)$ (top) and $q_{in}(t)$ (bottom) related to the effective flow rate labeled “eff-on” and to the flow rate at the detectors D1, respectively taken from empirical data in Fig. 12.6. (b) Pattern evolution. (c) Speed distributions. $t_j$ in (a) is the time at which the first wide moving jam reaches the detectors D1 in the empirical data shown in Fig. 13.1. At $t \geq t_j$ $q_{in}$ is chosen to be $q_{in} = q_{in}(t_j)$. Taken from [330]
Over time the flow rate $q_{on}$ begins to decrease (Fig. 18.16a). As in the empirical Fig. 13.1, the transition to weak congestion in the GP occurs (Fig. 18.17). During further decrease in $q_{on}$ (Fig. 18.16a) the GP transforms into an LSP (Fig. 18.16c, $t = 9:40$). The vehicle speed in the LSP is higher than in the pinch region of the GP. During the further evolution of synchronized flow in the LSP the pinch effect in this synchronized flow occurs and a wide moving jam emerges in this synchronized flow: the LSP transforms into an DGP (Fig. 18.16c, $t = 10:26$). As in the empirical Fig. 13.1, this pinch effect occurs on the main road at some distance (about 3 km) upstream from the on-ramp. After the wide moving jam of the DGP is far from the on-ramp bottleneck, a new LSP appears on the main road (Fig. 18.16c, $t = 11:00$). The LSP disappears and free flow remains on the main road at the on-ramp bottleneck when $q_{on}$ further decreases.

\[ k_{\text{in}} \geq k_{\text{out}} \]

\[ \begin{align*}
\text{weak} & \quad \text{strong} \\
\text{LSP} & \quad \text{DGP}
\end{align*} \]

\[ \begin{align*}
80 & \quad 40 \\
2000 & \quad 0
\end{align*} \]

Fig. 18.17. Evolution of congested patterns at an on-ramp bottleneck. Speed (left) and flow rate (right) on the main road at a virtual detector in the pinch region of the GP shown in Fig. 18.16. Taken from [330]

### 18.5 Hysteresis and Nucleation Effects by Pattern Formation at On-Ramp Bottlenecks

#### 18.5.1 Threshold Boundary for Synchronized Flow Patterns

The boundary $F^{(B)}_S$ in the diagram of congested patterns (Fig. 18.1a) corresponds to a spontaneous $F \rightarrow S$ transition. This phase transition occurs at the on-ramp bottleneck at the boundary $F^{(B)}_S$ during a given time interval $T_{ob}$ ($T_{ob} = 30$ min in Figs. 18.1a and 18.18a) with probability

\[ P_{FS}^{(B)} = 1 \quad \text{(18.24)} \]

It is found that left of the boundary $F^{(B)}_S$, i.e., in the free flow region, an $F \rightarrow S$ transition on the main road at the on-ramp bottleneck can nevertheless also occur spontaneously with probability...
Fig. 18.18. Metastable regions in the diagram of congested patterns at an on-ramp bottleneck. (a) Metastable states (hatched regions) where either free flow or an SP on the main road can exist and be excited. (b) Metastable states (hatched regions) where an GP (right of line $G$) and an DGP (left of line $G$) on the main road can exist and be excited. (c) The general diagram related to Fig. 18.1a where the regions of metastable states are shown. Taken from [330]
This result is explained by a Z-shaped function in Fig. 17.4. For a given \( q_{\text{in}} < q_{\text{th}} \) the threshold flow rate \( q_{\text{th}}^{(\text{th})} \) at this Z-characteristic is related to a point on the threshold boundary \( F_{\text{th}}^{(B)} \) in Fig. 18.18a. Below and left of the threshold boundary \( F_{\text{th}}^{(B)} \) the probability for an \( F \rightarrow S \) transition at the on-ramp bottleneck

\[
P_{FS}^{(B)} = 0 .
\]

In accordance with a theory of freeway capacity in free flow at a bottleneck discussed in Sect. 8.3, numerical simulations show that the threshold boundary \( F_{\text{th}}^{(B)} \) is associated with an infinite number of minimum freeway capacities. These minimum freeway capacities are equal to the threshold flow rates \( q_{\text{th}}^{(B)} \) (8.13). If the flow rate in free flow downstream of the bottleneck \( q_{\text{sum}} \) is related to the condition (8.14), then this flow rate is lower than freeway capacity. In other words, simulations show that there are the infinity of freeway capacities \( q_{C}^{(B)} \) in free flow at a freeway bottleneck associated with the infinite number of points \((q_{\text{on}}, q_{\text{in}})\) on and between the boundaries \( F_{S}^{(B)} \) and \( F_{\text{th}}^{(B)} \) in the diagram of congested patterns at the bottleneck (Fig. 18.1a). This means that at a given value of the flow rate \( q_{\text{in}} \) the freeway capacity \( q_{C}^{(B)}(q_{\text{on}}, q_{\text{in}}) \) in free flow at the bottleneck satisfies the condition (8.19).

In numerical simulations, the threshold boundary \( F_{\text{th}}^{(B)} \) can be estimated in the following way. The probability for an \( F \rightarrow S \) transition \( P_{FS}^{(B)} \) is calculated as defined in Sect. 8.3.2. However, we start calculations of the probability \( P_{FS}^{(B)} \) for an \( F \rightarrow S \) transition at low enough values \( q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}) \) that in all \( N_{FS} \) realizations no \( F \rightarrow S \) transitions occur at the on-ramp bottleneck, i.e., \( P_{FS}^{(B)} = 0 \). Then the flow rate \( q_{\text{sum}} \) increases. The threshold flow rates \( q_{\text{th}}^{(B)} \) at the threshold boundary \( F_{\text{th}}^{(B)} \) are approximately related to the maximum flow rates \( q_{\text{sum}} \) at which the condition

\[
P_{FS}^{(B)} = 0 \mid q_{\text{sum}}(q_{\text{on}}, q_{\text{in}}) = q_{\text{th}}^{(B)}
\]

is still satisfied.

Thus, below and left of \( F_{\text{th}}^{(B)} \) in the plane \((q_{\text{on}}, q_{\text{in}})\) no long-lived SPs can persist. Between the boundaries \( F_{\text{th}}^{(B)} \) and \( F_{S}^{(B)} \) (dashed region in Fig. 18.18a) depending on initial conditions either free flow or one of the SPs can exist. This region in the diagram is a metastable region of free flow with respect to an \( F \rightarrow S \) transition and to SP formation.

An SP can persist later for a long time. In some cases, depending on \( q_{\text{on}} \) and \( q_{\text{in}} \) the SP can either dissolve or transform into an GP. There is hysteresis related to the metastable states (Figs. 18.19a,b): after the SP first occurs spontaneously, right of the boundary \( F_{S}^{(B)} \), the SP can also further exist when one of the flow rates \( q_{\text{on}} \) or \( q_{\text{in}} \) (or both of them) decreases and
the point \((q_{on}, q_{in})\) is left of the boundary \(F_{S}^{(B)}\). The SP disappears near the threshold boundary \(F_{th}^{(B)}\). At \(q_{on} \to 0\) the flow rate \(q_{in} \to q_{th}\) (Sect. 5.2.4) at the threshold boundary \(F_{th}^{(B)}\).

In some interval of low values of \(q_{on}\), the flow rate \(q_{in}\) at \(F_{th}^{(B)}\) remains approximately constant: \(q_{in} \approx q_{th}\). There is a boundary (the vertical boundary labeled \(M\) in Fig. 18.18a) that separates the region of MSPs (left of the boundary \(M\)) and WSPs (right). Left of the boundary \(M\) and between the
boundaries $F^{(B)}_{th}$ and $F^{(B)}_S$ an MSP can be excited by applying a short-time perturbation of $q_{on}$. At a lower $q_{on}$ a single MSP occurs (Fig. 18.1d). If $q_{on}$ increases ($q_{on}$ is still related to a point in the diagram left of the boundary $M$) first an WSP occurs spontaneously. However, within this WSP local regions of free flow spatially alternating with initial synchronized flow appear spontaneously over time. This alternating SP (ASP) finally transforms into a sequence of MSPs. In contrast, between the boundaries $F^{(B)}_{th}$ and $F^{(B)}_S$ and right of the boundary $M$ (and also above the boundary $W$) rather than an MSP a widening SP (WSP) can be excited or occur on the main road at the on-ramp bottleneck.

### 18.5.2 Threshold Boundary for General Patterns

The hysteresis effect and metastable states also appear when an GP emerges. At the boundary $S^{(B)}_J$ an S→J transition occurs during a given time interval ($T_{ob} = 60$ min in Figs. 18.1a and 18.18b) with probability\(^4\)

\[
P_{SJ} = 1. \quad (18.28)
\]

However, left of $S^{(B)}_J$ the S→J transition can also occur with probability

\[
P_{SJ} < 1. \quad (18.29)
\]

This result is associated with the Z-shaped characteristic for an S→J transition (Fig. 17.11).

For a given $q_{in} < q_{out}$ the flow rate $q^{(th, SJ)}$ at this Z-characteristic is related to a point at a threshold boundary $J^{(B)}_S$ in the diagram. Below and left of this threshold boundary $J^{(B)}_S$ (Fig. 18.18b) the probability for an S→J transition

\[
P_{SJ} = 0. \quad (18.30)
\]

Between $S^{(B)}_J$ and $J^{(B)}_S$ (dashed regions in Fig. 18.18b) free flow and an SP are metastable with respect to GP formation, i.e., the GP can either occur spontaneously during a long enough time interval or be excited.

If an GP has occurred spontaneously right of the boundary $S^{(B)}_J$, this GP can further exist if one of the flow rates $q_{in}$ and $q_{on}$ (or the both) decrease so that the point in the diagram is displaced into the region between the boundaries $S^{(B)}_J$ and $J^{(B)}_S$. Below and left of $J^{(B)}_S$ the GP dissolves.

The boundary $J^{(B)}_S$ consists of a horizontal line

\[
q_{in} = q_{out} \text{ at } q_{on} \leq q^{(G)}_{on} \quad (18.31)
\]

\(^4\) In numerical simulations, a large but finite number of different realizations $N_{SJ}$ are made. The exact value of the probability $P_{SJ}$ of an S→J transition can only be found as $N_{SJ} \to \infty$. 
and a curve determined by the condition

\[ q_{in} = q^{(\text{pinch})}(q_{on}) \text{ at } q_{on} > q^{(G)}_{on}, \quad (18.32) \]

where the dependence \( q^{(\text{pinch})}(q_{on}) \) is shown in Fig. 18.12a. These two branches merge at a point \( q_{on} = q^{(G)}_{on} \) (18.16). At that point the boundary \( J_{S}^{(B)} \) intersects the vertical boundary \( G \) that separates DGPs and GPs in the diagram (Fig. 18.18b). Thus, between the boundaries \( J_{S}^{(B)} \) and \( S_{J}^{(B)} \) and left of the boundary \( G \) metastable states exist where DGPs can persist and be excited.

In contrast, the region between \( J_{S}^{(B)} \) and \( S_{J}^{(B)} \) and right of the boundary \( G \) is the metastable region with respect to GP formation. For example, if a GP under the strong congestion condition has occurred spontaneously near the boundary \( S_{J}^{(B)} \) at \( q_{on} > q^{(\text{strong})}_{on} \) and \( q_{in} > q^{(\text{pinch})}_{in} \), the real flow rate of vehicles merging onto the main road can decrease: \( q^{(on)}_{in} < q_{in} \). However, if a related effective point in the diagram is shifted to the metastable region between \( J_{S}^{(B)} \) and \( S_{J}^{(B)} \), the GP will not dissolve.

To explain the boundary \( J_{S}^{(B)} \), recall that if the condition (18.18) is not valid, i.e., at \( q_{on} < q^{(G)}_{on} \) the pinch region and therefore an GP cannot continuously exist. However, if \( q_{in} > q_{\text{out}} \) a single wide moving jam, i.e., an DGP can persist. In contrast, at \( q_{in} < q_{\text{out}} \) the single wide moving jam cannot be formed and therefore an DGP cannot appear. This explains (18.31). When the condition (18.18) is valid, an GP can occur if the inflow to the GP \( q_{in} \) is greater than \( q^{(\text{pinch})}_{in} \). In contrast, at \( q_{in} < q^{(\text{pinch})} \) the pinch region and the GP dissolve. Thus, the condition (18.32) indeed determines \( J_{S}^{(B)} \) at \( q_{on} \geq q^{(G)}_{on} \).

### 18.5.3 Overlap of Different Metastable Regions and Multiple Pattern Excitation

The metastable states with respect to SP formation (dashed region in Fig. 18.18a) and to DGP and GP formation (dashed region in Fig. 18.18b) partially overlap one another (Fig. 18.18c). In these overlapping metastable states, depending on initial conditions either free flow or one of the SPs or else one of the GPs (or DGPs) can occur.

In Fig. 18.19c, depending on the amplitude of an initial short-time perturbation applied to \( q_{on} \) either free flow (Fig. 18.19d) or an LSP (Fig. 18.19e), or else an GP (Fig. 18.19f) is excited. The stability of these three traffic states is markedly different. For example, whereas free flow (Fig. 18.19d) and the GP (Fig. 18.19f) remain stable for 120 min in 10 different calculated realizations, the LSP has transformed spontaneously into an GP in 6 of 10 realizations.
18.6 Strong Congestion at Merge Bottlenecks

18.6.1 Comparison of General Patterns at Merge Bottleneck and at On-Ramp Bottleneck

At an on-ramp bottleneck when traffic demand \( q_{in} \) and/or \( q_{on} \) varies weak congestion in the pinch region of an GP can transform into strong congestion, where pinch region characteristics do not depend on traffic demand any more (Sect. 18.3). It is found that at a merge bottleneck (Fig. 16.2b), regardless of traffic demand \( q_{in} \), only strong congestion occurs in the pinch region of an GP (Figs. 18.20a,b).

To explain this result, note that for the case of the on-ramp bottleneck the main road at \( x < x_{on}^{(e)} \) in Fig. 16.2a can be considered as the left lane of the merge bottleneck at \( x < x_M \) in Fig. 16.2b and the on-ramp lane as the right lane of the merge bottleneck. In each of the lanes of the merge bottleneck at \( x = -L/2 \) the flow rate is equal to \( q_{in} \) (Fig. 16.2b). If now on the on-ramp the related condition

\[
q_{on} = q_{in}
\]

is used, then a GP under the strong congestion condition is formed (Fig. 18.1a).

However, at the merge bottleneck at \( x < x_M^{(s)} \) vehicles can change lane between the left and right lanes (Fig. 16.2b). In contrast, at the on-ramp bottleneck at \( x < x_{on} \) the main road and the on-ramp lane are separate roads (Fig. 16.2a). This can explain the following differences between GPs at the merge bottleneck and the on-ramp bottleneck:

(i) In contrast to a GP under the strong congestion condition on the main road at the on-ramp bottleneck, the average speed between moving jams in an GP at the merge bottleneck has a saturation to a limit (maximum) speed \( v_{\text{max}}^{(M)} \) that satisfies the condition

\[
v_{\text{max}}^{(M)} < v_{\text{free}}.
\]

The actual speed \( v^{(M)} \) between the moving jams shows only small amplitude oscillations around \( v_{\text{max}}^{(M)} \) (Fig. 18.20b).

(ii) The limit flow rate per freeway lane at the merge bottleneck \( q_{\text{lim}, M}^{(\text{pinch})} \) is considerably lower than \( q_{\text{lim}}^{(\text{pinch})} \) at the on-ramp bottleneck.

(iii) For the same model parameters the frequency of narrow moving jam emergence \( f_{\text{max}}^{(M)} \) is higher and the mean time between the jams, \( T_J^{(M)} = 1/f_{\text{max}}^{(M)} \), is smaller at the merge bottleneck than the related values at the bottleneck due to the on-ramp.

(iv) The mean time \( T_J^{(\text{wide} M)} \) between the downstream fronts of wide moving jams in an GP at the merge bottleneck are smaller than the related value \( T_J^{(\text{wide} \text{on})} \) at the bottleneck due to the on-ramp. Nevertheless, after speed saturation has occurred between the narrow-looking moving jams in the
GP at the merge bottleneck, the downstream jam fronts propagate at the characteristic velocity $v_g$ through other upstream bottlenecks and through other states of traffic flow while maintaining the velocity of the downstream front $v_g$ (Sect. 19.3.1), i.e., the jams are wide moving jams.

### 18.6.2 Diagram of Congested Patterns

Pattern emergence at the merge bottleneck exhibits the following pattern diagram (Fig. 18.21). If

$$q_{in} \geq q_{max}^{(B)} ,$$

(18.35)
Fig. 18.21. Diagram of patterns at a merge bottleneck. Taken from [330]

an GP occurs spontaneously during a given time interval $T_{ob}$ ($T_{ob} = 30 \text{ min}$ in Fig. 18.21). The maximum flow rate

$$q_{\text{in}}^{(B)} = \frac{q_{\text{in}}^{(\text{free})}}{2}.$$  

(18.36)

The factor $1/2$ appears in this formula because at the merge bottleneck the two-lane road transforms into a single-lane road at $x \geq x_M$.

There is a time delay $T_{FS}$ for GP formation after traffic flow is switched. Many local perturbations can appear spontaneously, grow and then decay at the merge bottleneck before a perturbation appears whose growth indeed leads to GP formation (Fig. 18.21).

The flow rate $q_{\text{max}, M}^{(B)}$ is the maximum freeway capacity of free flow at the merge bottleneck. At $q_{\text{in}} = q_{\text{max}, M}^{(B)}$ the probability $P_{FS}^{(B)}$ for speed breakdown at the merge bottleneck during the time interval $T_{ob}$ is equal to 1. The threshold flow rate $q_{\text{th}, M}^{(B)}$ is the minimum freeway capacity. The maximum and minimum capacities depend on the time interval $T_{ob}$. There are an infinite number of freeway capacities $q_{\text{C}, M}^{(B)}$ of free flow at the merge bottleneck, which satisfy the conditions

$$q_{\text{th}, M}^{(B)} \leq q_{\text{C}, M}^{(B)} \leq q_{\text{max}, M}^{(B)}.$$  

(18.37)

If

$$q_{\text{th}, M}^{(B)} \leq q_{\text{in}} < q_{\text{max}, M}^{(B)}$$  

(18.38)

free flow is metastable. In this case, GPs exhibit the following features:

(i) At

$$q_{\text{lim}, M}^{(\text{pinch})} \leq q_{\text{in}} < q_{\text{max}, M}^{(B)}$$  

(18.39)

an GP can be excited by a high amplitude short-time external perturbation.
(ii) At
\[ q_{\text{th}, \text{M}}^{(B)} \leq q_{\text{in}} \leq q_{\lim, \text{M}}^{(\text{pinch})} \]  \hspace{0.5cm} (18.40)

rather than an GP only an LSP can be excited.

At
\[ q_{\text{in}} < q_{\text{th}, \text{M}}^{(B)} \]  \hspace{0.5cm} (18.41)

free flow is stable, i.e., regardless of the initial amplitude of a short-time local perturbation no congested patterns can occur at the merge bottleneck.

18.7 Weak Congestion at Off-Ramp Bottlenecks

18.7.1 Diagram of Congested Patterns

In contrast to the on-ramp and merge bottlenecks, only weak congestion can occur in the pinch region of an GP at an off-ramp bottleneck (Fig. 16.2c). This is true for all values
\[ \eta < 100\% \]  \hspace{0.5cm} (18.42)

Recall that \( \eta \) is the percentage of vehicles that want to leave the main road via the off-ramp at a flow rate upstream of the off-ramp \( q_{\text{in}} \).

In an GP, wide moving jams are often formed on the main road at a considerable distance (more than 5 km in Fig. 18.22) upstream of the off-ramp. The downstream front of an SP and an GP is located at some distance (about 1–1.5 km in Fig. 18.22) upstream of the off-ramp. In addition, the speed synchronization across the both freeway lanes occurs only at some finite distance upstream of \( x = x_{\text{off}} \) (Fig. 16.2c).

The pattern diagram at the off-ramp bottleneck, i.e., the regions of pattern formation in the \( (\eta, q_{\text{in}}) \) plane (Fig. 18.22a) qualitatively resembles the diagram at the on-ramp bottleneck (Fig. 18.1a).

There are different SPs (MSP, WSP, and LSP), DGPs and GPs (Figs. 18.22b–g). However, there is no saturation of the boundary \( S_{\text{j}}^{(B)} \) at a higher \( \eta \). This is associated with the mentioned fact that in GPs only the weak congestion condition occurs at all \( \eta < 100\% \).

Indeed, all pinch region characteristics in GPs as well as the flow rate \( q_{\text{off}} \) of vehicles that really leave the main road via the off-ramp do depend on \( \eta \) (Fig. 18.23).

18.7.2 Comparison of Pattern Features at Various Bottlenecks

At the same model parameters the values \( q^{(\text{pinch})}, v^{(\text{pinch})}, \) and \( T_{\text{j}} \) in an GP at an off-ramp bottleneck at high \( \eta \) can be considerably lower (Fig. 18.23)
Fig. 18.22. Congested patterns at an isolated off-ramp bottleneck. (a) Diagram of congested patterns. (b–d) GPs. (e–g) SPs. (b) GP at a high $\eta$. (c) GP at a low $\eta$. (d) DGP. (e) WSP. (f) LSP. (g) MSP. Taken from [330]
Fig. 18.23. Features of congested patterns at an off-ramp bottleneck. Average characteristics of GPs at the off-ramp bottleneck as functions of \( \eta \). Taken from [330]

than \( q^{(\text{pinch})} \), \( v^{(\text{pinch})} \), and \( T_{\text{lim}} \), respectively in a GP under the strong congestion condition at an on-ramp bottleneck (Figs. 18.12a,b,g). The off-ramp bottleneck forces vehicles at higher \( \eta \) to much stronger traffic compression.

There are some differences between an SP at on- and off-ramp bottlenecks:

1. The downstream front of an SP at the off-ramp is located at some distance upstream from the off-ramp. This is in accordance with the empirical results of Sect. 12.5.
2. The effect of the speed synchronization across both freeway lanes is realized only at some finite distance upstream of the coordinate \( x = x_{\text{off}} \) where the off-ramp begins.

This has a simple explanation. Vehicles that do not want to leave the main road remain the left lane. In the left lane the vehicle density decreases in space coordinate within the merging region on the main road. Thus, these vehicles can accelerate to the speed in free flow. However, at some distance upstream of the off-ramp there is a considerable decrease in vehicle speed in the right lane due to congested pattern formation. This decrease forces vehicles still in the left lane that want to leave via the off-ramp to slow down before they change to the right lane. The latter forces all vehicles to slow down, too. Therefore, speeds synchronize in the two lanes.

This explanation also leads to another qualitative conclusion. Let us assume that there are many more than two lanes on a freeway in the vicinity of the off-ramp. Then vehicles in the leftmost lane are almost undisturbed by vehicles that leave the freeway via the off-ramp. Thus, in this case, synchronization of vehicle speeds across the freeway does not necessarily occur in lanes far from the right lane. Nevertheless SP and GP in lanes close to the right lane can be expected. This possible two-dimensional spatial effect is, however, a case for special study.
At $\eta = 100\%$ the off-ramp bottleneck transforms into the merge bottleneck. Thus, we find qualitatively the same GP features at $\eta = 100\%$ as for the merge bottleneck.

18.8 Congested Pattern Capacity at On-Ramp Bottlenecks

Here we illustrate a qualitative theory of congested pattern capacity (Sect. 8.5) based on a numerical simulation of the KKW cellular automata microscopic traffic flow model based on three-phase traffic theory (Sect. 16.2).

18.8.1 Transformations of Congested Patterns at On-Ramp Bottlenecks

To study the congested pattern capacity, in the diagram of congested patterns at on-ramp bottlenecks three lines (labeled “line 1,” “line 2,” and “line 3”) are shown (Fig. 18.24). Different congested patterns have been studied when the flow rates $q_{on}$ (line 1 and line 2) or the flow rate $q_{in}$ (line 3) are increasing along the related lines.

The transformation of congested patterns along line 1 is related to a given high flow rate in free flow on the main road upstream of the on-ramp $q_{in}$ when the flow rate to the on-ramp $q_{on}$ is increased from lower to higher values (Fig. 18.25). Right of the boundary $F_s^{(B)}$ an WSP occurs on the main road (Fig. 18.25a). If the flow rate to the on-ramp is only slightly increased, then the WSP remains (Fig. 18.25b). However, the vehicle speed in this WSP decreases on average in comparison with the speed in the initial WSP in Fig. 18.25a.

Right of the boundary $S_j^{(B)}$ at the high flow rate $q_{in}$, which is related to line 1, an DGP occurs on the main road (Fig. 18.25c). After the first wide moving jam has emerged in synchronized flow, the flow rate on the main road upstream of the on-ramp decreases because it is now determined by the jam outflow. The maximum flow rate in the wide moving jam outflow is reached when free flow is formed downstream of the jam ($q_{out} = 1810 \text{ vehicles/h}$). When due to upstream jam propagation the wide moving jam is far from the on-ramp an effective flow rate upstream of the on-ramp is determined by $q_{out}$, i.e., the effective flow rate on the main road upstream of the on-ramp is equal to $q_{in}^{(eff)} = q_{out} = 1810 \text{ vehicles/h}$. However, this flow rate is considerably lower than the initial flow rate $q_{in} = 2400 \text{ vehicles/h}$. As a result, no wide moving jams can emerge on the main road upstream of the on-ramp any more: the DGP occurs that consists of the only one wide moving jam propagating upstream and an SP localized on the main road at the on-ramp bottleneck. At the mentioned effective flow rate $q_{in}^{(eff)} = 1810 \text{ vehicles/h}$ and
the flow rate to the on-ramp $q_{on} = 105$ vehicles/h this SP at the on-ramp bottleneck is an WSP (Figs. 18.24 and 18.25c).

If the flow rate $q_{on}$ further increases, then the pinch region in synchronized flow on the main road upstream of the on-ramp bottleneck appears where narrow moving jams emerge continuously. Some of these jams transform into wide moving jams, leading to GP formation (Fig. 18.25d). This occurs right of the boundary $G$ where GPs are realized. When the flow rate $q_{on}$ further increases, the GP does not transform into another pattern: it remains an GP at any possible flow rate $q_{on}$ (Fig. 18.25e,f).

The transformation of congested patterns along line 2 in Fig. 18.24 is related to a given low flow rate $q_{in}$ in free flow on the main road upstream of the on-ramp bottleneck when the flow rate to the on-ramp $q_{on}$ increases from a relatively low initial value (Fig. 18.26). In this case, first an LSP occurs (Fig. 18.26a,b). If the flow rate $q_{on}$ increases, the LSP transforms into GPs (Fig. 18.26c–e). However, whereas for the GPs related to line 1 the condition $q_{in} > q_{out}$ is satisfied, in the case of line 2 we have $q_{in} < q_{out}$ (Fig. 18.24). As a result, the farthest upstream wide moving jam in the GPs dissolves over time. This indeed occurs for the GPs related to line 2. For this reason, the width (in the longitudinal direction) of the GPs increases more slowly over time in comparison to the width of the GPs related to line 1.
18.8 Congested Pattern Capacity

Fig. 18.25. Evolution of vehicle speed on the main road in space and in time at the given $q_{in} = 2400$ vehicles/h for different $q_{on}$ related to line 1 in Fig. 18.24. 
(a, b) WSPs. (c) DGP. (d–f) GPs. The flow rate $q_{on}$ is: (a) (40), (b) (60), (c) (105), (d) (150), (e) (200), (f) (500) vehicles/h. The on-ramp is at the location $x = 16$ km. 
Taken from [221]

The transformation of congested patterns along line 3 in Fig. 18.24 is related to a given flow rate to the on-ramp $q_{on}$ when the flow rate in free flow on the main road upstream of the on-ramp $q_{in}$ increases from a relatively low initial value (Fig. 18.27). Because at this flow rate line 3 intersects the boundary $F_{s}^{(B)}$ above the boundary $W$ (that separates WSPs and LSPs in the diagram of congested patterns in Fig. 18.24), an WSP occurs (Fig. 18.27a). The WSP remains in a wide range of the flow rate $q_{in}$ when this flow rate increases (Fig. 18.27b,c). However, the average speed in the WSP decreases, whereas the flow rate $q_{in}$ increases. Finally, if the flow rate $q_{in}$ increases and line 3 intersects the boundary $S_{j}^{(B)}$. As a result, an GP occurs (Fig. 18.27d).
Fig. 18.26. Evolution of vehicle speed on the main road in space and in time at the given $q_{in} = 1255$ vehicles/h for different $q_{on}$ related to line 2 in Fig. 18.24. (a, b) LSPs. (c-e) CPs. The flow rate $q_{on}$ is: (a) (550), (b) (630), (c) (700), (d) (850), (e) (1000) vehicles/h. Taken from [221]

This GP transforms into the GP shown in Fig. 18.25e when the flow rate $q_{in}$ further increases.

### 18.8.2 Temporal Evolution of Discharge Flow Rate

According to simulations of congested patterns that results have been discussed above, the flow rate to the on-ramp $q_{on}$ has been switched only after the time $t = t_0$. Let us denote the flow rate on the main road far downstream of the on-ramp where free flow occurs by $q_{down}$. During the time $0 \leq t < t_0$ free flow occurs on the main road at the on-ramp bottleneck and

$$q_{down} = q_{sum} = q_{in} \quad \text{at} \quad 0 \leq t < t_0 . \quad (18.43)$$
18.8 Congested Pattern Capacity

At $t > t_0$, i.e., after the on-ramp inflow has been switched, the flow rate $q_{\text{down}}$ should increase because $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ at $t \geq t_0$.

However, at $t \geq t_0$ a congested pattern has begun to form on the main road upstream of the on-ramp bottleneck. Thus, at $t > t_0$ the flow rate $q_{\text{down}}$ is determined by the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$:

$$q_{\text{down}} = q_{\text{out}}^{(\text{bottle})} \quad \text{at} \quad t > t_0 . \quad (18.44)$$

Numerical simulations enables us to study the temporal evolution of the flow rate $q_{\text{down}}$ by the formation of each of the congested patterns on the main road upstream of the on-ramp (Figs. 18.28–18.30).

This evolution for congested patterns that appear along line 1 is shown in Fig. 18.28. It can be seen that although at $t = t_0$ the inflow from the on-ramp is switched, i.e., additional vehicles merge onto the main road from the on-ramp, the discharge flow rate is lower than the flow rate $q_{\text{sum}} = q_{\text{in}}$ at $0 \leq t < t_0$. This is the result of congested pattern formation: any congested pattern along line 1 reduces flow rate on the main road just upstream of the on-ramp. This decrease is greater than the increase in the flow rate of vehicles merging onto the main road from the on-ramp.

The flow rate $q_{\text{down}}$ (18.44) is considerably lower than $q_{\text{sum}}$ (18.43) when an GP occurs at the bottleneck. When an WSP shown in Fig. 18.25a occurs, then the flow rate $q_{\text{down}}$ (18.44) is only slightly lower than the flow rate $q_{\text{down}}$ (18.43) (Fig. 18.28a). This means that SPs can be more favorable than...
Fig. 18.28. Temporal evolution of the flow rate $q_{\text{down}}$ downstream of an on-ramp bottleneck during pattern formation at the on-ramp bottleneck for different congested patterns related to line 1 in Fig. 18.24. (a–f) are related to the congested patterns with the same letters (a–f) in Fig. 18.25. $t_0 = 8$ min. Taken from [221]
GPs in terms of the discharge volume from the congested pattern, i.e., in terms of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ (18.44) and of the vehicle time delay due to congestion. The latter is because the average vehicle speed within the WSP (Fig. 18.25a) is considerably higher than the average speed in the pinch region of the GP (Fig. 18.25d–f). These conclusions of the theory of congested patterns at bottlenecks are used in methods of congested pattern control at bottlenecks (Sect. 23.3.2).

A different situation is realized for congested patterns that appear along line 2 (Fig. 18.29). In this case, the discharge flow rate is greater than the flow rate $q_{\text{sum}} = q_{\text{in}}$ at $0 \leq t < t_0$: all congested patterns along line 2 also leads to a decrease in the flow rate on the main road just upstream of the on-ramp. However, this decrease is less than the increase in flow rate of vehicles merging onto the main road from the on-ramp.

Fig. 18.29. Temporal evolution of the flow rate downstream of the on-ramp bottleneck $q_{\text{down}}$ during pattern formation at an on-ramp bottleneck for different congested patterns related to line 2 in Fig. 18.24. (a–e) are related to congested patterns with the same letters (a–e) in Fig. 18.26. $t_0 = 8$ min. Taken from [221]
Fig. 18.30. Temporal evolution of the flow rate downstream of the on-ramp bottleneck \(q_{\text{down}}\) during pattern formation at the on-ramp for different congested patterns related to the line 3 in Fig. 18.24. (a–d) are related to congested patterns with the same letters (a–d) in Fig. 18.27. \(t_0 = 8\) min. Taken from [221]

An intermediate case has been found for congested patterns that appear along line 3 (Fig. 18.30), i.e., when the initial flow rate upstream of the on-ramp \(q_{\text{in}}\) changes. Firstly, there is a slight increase in the flow rate \(q_{\text{down}} = q_{\text{out}}^{(\text{bottle})}\) due to congested pattern formation, in comparison with the initial flow rate \(q_{\text{down}} = q_{\text{in}}\) (Fig. 18.30a). For a greater flow rate \(q_{\text{in}}\) there is almost no change in the flow rate \(q_{\text{down}}\) after congested patterns have been formed (Figs. 18.30b,c). When the flow rate \(q_{\text{in}}\) further increases, the discharge flow rate is lower than the initial flow rate \(q_{\text{down}} = q_{\text{in}}\) at \(0 \leq t < t_0\), as it was for congested patterns along line 1 (Fig. 18.30d), i.e., the flow rate \(q_{\text{down}}\) decreases during congested pattern formation.

### 18.8.3 Dependence of Congested Pattern Capacity on On-Ramp Inflow

The congested pattern capacity \(q_{\text{cong}}^{(B)}\) (8.27) has been calculated via 60-min averaging of the discharge flow rate \(q_{\text{out}}^{(\text{bottle})}\) after the related congested pattern has been formed. It has been found that the congested pattern capacity \(q_{\text{cong}}^{(B)}\) depends heavily on the type of congested patterns and on pattern parameters.

Indeed, along line 1 (Fig. 18.31a) the congested pattern capacity \(q_{\text{cong}}^{(B)}\) is a decreasing function of the flow rate to the on-ramp \(q_{\text{on}}\).
18.8 Congested Pattern Capacity

Fig. 18.31. Dependence of the average congested pattern capacity on pattern type and pattern parameters. (a, b) Congested pattern capacity as a function of flow rate $q_{on}$ related to line 1 (a) and line 2 (b) in Fig. 18.24. (c) Congested pattern capacity as a function of flow rate $q_{in}$ related to line 3 in Fig. 18.24. Black points correspond to congested patterns for lines 1–3 (Figs. 18.25–18.27), respectively. Dashed lines show the flow rate in the wide moving jam outflow $q_{out}$. Taken from [221]

At a lower flow rate $q_{on}$ an WSP occurs (Fig. 18.25a). The congested pattern capacity $q_{cong}^{(B)}$ is a relatively high value for the WSP. This capacity is compatible with the maximum capacity in free flow at the bottleneck. This result is associated with a high discharge flow rate $q_{out}^{(bottle)}$ downstream of the bottleneck when the WSP appears (Fig. 18.28a). This result has already been discussed when congested patterns along line 1 have been considered (Fig. 18.25).

When the flow rate $q_{on}$ increases, the congested pattern capacity $q_{cong}^{(B)}$ abruptly decreases. This is related to the pinch effect in synchronized flow that occurs when $q_{on}$ increases. Due to the pinch effect, i.e., a compression of synchronized flow with narrow moving jam formation, the flow rate in synchronized flow on the main road upstream of the on-ramp decreases (Fig. 18.31a). The decrease in the latter flow rate is appreciably higher than the increase in flow rate at the on-ramp. For this reason, the congested pattern capacity $q_{cong}^{(B)}$ decreases although $q_{on}$ increases. However, there is a saturation of the decrease in congested pattern capacity $q_{cong}^{(B)}$ when $q_{on}$ further increases.

The congested pattern capacity at lower $q_{on}$, when the WSPs shown in Figs. 18.25a,b are realized, is greater than the flow rate in the wide moving jam outflow, $q_{out}$ (dashed line in Fig. 18.31a). At higher $q_{on}$ the congested
pattern capacity becomes less than $q_{\text{out}}$. This occurs when GPs shown in Figs. 18.25e,f are formed at the bottleneck.

By calculation of the congested pattern capacity the condition (8.26) has been satisfied. The only exception is the calculation of the capacity along line 2; in this case, first an LSP appears and therefore the condition (8.31) has been used. The congested pattern capacity associated with this LSP is determined by the maximum discharge flow rate for which the LSP still exists upstream of the on-ramp. It has been found that the latter condition is only satisfied for the LSP at some $q_{\text{on}}$ (in the example in Fig. 18.31b this flow rate is approximately 650 vehicles/h). This has been taken into account in Fig. 18.31b. The dashed curve is related to the discharge flow rate from the LSP when the congested pattern capacity is not reached, and the solid curve corresponds to the congested pattern capacity. It can be seen that the congested pattern capacity decreases when the flow rate $q_{\text{on}}$ increases. However, in comparison with line 1 the decrease in the congested pattern capacity along line 2 is considerably smaller given the same increase in flow rate $q_{\text{on}}$.

Along line 3 (Fig. 18.31c) the congested pattern capacity $q_{\text{cong}}^{(B)}$ is a non-monotonic function of the flow rate $q_{\text{in}}$: there is a maximum point on this dependence. It turns out that at the same flow rate to the on-ramp $q_{\text{on}}$ the flow rate in synchronized flow in an WSP on the main road upstream of the on-ramp increases during the evolution of the WSP with increasing in flow rate $q_{\text{in}}$. The flow rate in synchronized flow of the WSP at the maximum point in Fig. 18.31c is less than the flow rate in the wide moving jam outflow $q_{\text{out}}$ (dashed line in Fig. 18.31c). When the flow rate $q_{\text{in}}$ further increases and therefore line 3 intersects the boundary $S_{j}^{(B)}$, wide moving jams begin to form in synchronized flow of the initial WSP and an GP forms. The flow rate in the wide moving jam outflow cannot exceed $q_{\text{out}}$. As a result, the flow rate through synchronized flow of the GP happens to be appreciably lower than in the WSP associated with the maximum point of the congested pattern capacity in Fig. 18.31c. Thus, the congested pattern capacity decreases when the WSP transforms into an GP. This explains the maximum point in the congested pattern capacity as a function of the flow rate $q_{\text{in}}$.

### 18.9 Conclusions

(i) The microscopic three-phase traffic theory can explain the main empirical features of spatiotemporal traffic patterns at isolated freeway bottlenecks discussed in Chaps. 9–13.

(ii) In accordance with empirical results (Chap. 12) and the hypotheses of the three-phase traffic theory (Chap. 6), neither a moving jam nor a sequence of moving jams emerge spontaneously in free flow at a freeway bottleneck, if in the vicinity of the bottleneck the vehicle density gradually increases from low density to higher densities. Rather than moving
jams, SP occurs first at the bottleneck. There can be three types of SPs: WSP, MSP, and LSP. Alternation of free and synchronized flow can occur within an SP. Wide moving jams can occur spontaneously only in synchronized flow. This leads to GP formation.

(iii) The “weak congestion” and “strong congestion” cases must be distinguished. In accordance with empirical features of weak and strong congestion (Chap. 13), in the microscopic three-phase traffic theory we found that

1. When the flow rate to the on-ramp slowly decreases, strong congestion in the pinch region of an GP changes to weak congestion. This leads to complex spatiotemporal pattern evolution; in particular, an GP can transform into one of the SPs or an DGP.

2. Under strong congestion, the average flow rate within the pinch region of an GP and characteristics of moving jam emergence do not depend on traffic demand. There is a correlation between the wide moving jam outflow and the flow rate limit in synchronized flow within an GP.

3. Under weak congestion, the characteristics of moving jam emergence in an GP do depend on traffic demand.

4. Strong congestion often occurs in GPs at on-ramp bottlenecks whereas in GPs at off-ramp bottlenecks only weak congestion occurs.

5. Under weak congestion, diverse transformations can occur between different congested patterns. In particular, one SP can transform into another or into an DGP, which later can subsequently transform into one of the SPs, and so on.

(iv) At a merge bottleneck (a freeway bottleneck due to a reduction in the number of freeway lanes), at lower flow rate an LSP can be induced and at higher flow rate an GP can occur. Only strong congestion can occur in the pinch region of an GP at the merge bottleneck.

(v) At an on-ramp bottleneck, there are diverse regions with metastable states of free flow, while at the same flow rates, depending on the initial conditions, either free flow, one of the SPs, or one of the GPs can be formed.

(vi) At an off-ramp bottleneck, the diagram of congested patterns, i.e., the regions of spontaneous occurrence of patterns in the plane whose coordinates are $(\eta, q_{in})$ $(\eta$ is the percentage of vehicles that want to leave the main road at the off-ramp) is qualitatively similar to the diagram at the on-ramp bottleneck. However, only weak congestion can occur in GPs at the off-ramp bottleneck at $\eta < 100\%$. 
19 Complex Congested Pattern Interaction and Transformation

19.1 Introduction

A real freeway has many different bottlenecks. Sometimes two or more adjacent effectual bottlenecks are very close to one another. In this case, we have seen that very complex expanded congested patterns (EP) (Sect. 14.2), the catch effect (Sect. 10.4.2), and other complex spatiotemporal phenomena are observed.

In this chapter, we will see that a microscopic three-phase traffic theory can explain the main features of these diverse empirical results. In addition, the theory also predicts some new effects whose empirical study is the case of future traffic flow observations.

We will consider here complex congested patterns that occur when two spatially separated adjacent effectual bottlenecks exist on a freeway. We call the effectual bottleneck that is upstream the “upstream bottleneck” and the effectual bottleneck that is downstream the “downstream bottleneck.” When both bottlenecks are due to on-ramps, the upstream bottleneck will be designated the on-ramp ‘U’ with coordinate $x_{on} = x_{on}^{(up)}$ and flow rate $q_{on} = q_{on}^{(up)}$, and the downstream bottleneck will be designated the on-ramp ‘D’ with $x_{on} = x_{on}^{(down)}$ and $q_{on} = q_{on}^{(down)}$.

Accordingly, we call congestion that emerges upstream of the downstream bottleneck “downstream congestion” or “downstream congested pattern” and congestion that emerges upstream of the upstream bottleneck as “upstream congestion” or “upstream congested pattern.”

We will discuss the following effects found in this microscopic theory.

(i) Induced F→S transitions caused by upstream congested pattern propagation. A congested pattern, which occurs at the downstream bottleneck, can induce another congested pattern at the upstream bottleneck where free flow is realized before.

(ii) The catch effect. The catch of synchronized flow at the upstream bottleneck with the subsequent formation of a new congested pattern at this bottleneck.

(iii) An influence of foreign wide moving jam propagation on a congested pattern at the upstream bottleneck. These foreign wide moving jams occur
downstream of the congested pattern within an GP at the downstream bottleneck.

(iv) An occurrence of expanded congested patterns (EP) where synchronized flow, which has initially appeared at the downstream bottleneck, covers the upstream bottleneck.

(v) An intensification of downstream congestion due to the onset of upstream congestion.

Results of this chapter are based on a microscopic traffic flow theory developed in \[330\].

19.2 Catch Effect and Induced Congested Pattern Formation

19.2.1 Induced Pattern Emergence

Firstly, we consider a congested pattern that has earlier occurred at the downstream bottleneck. The upstream front of this pattern further propagates continuously upstream. This congested pattern can induce another congested pattern at the upstream bottleneck where free flow is realized before.

Wide Moving Jam Propagation

Let us first consider induced pattern formation at the upstream bottleneck that is caused by wide moving jam propagation through the upstream bottleneck. We assume that this wide moving jam has initially occurred downstream of the upstream bottleneck. In this case, in accordance with empirical observations (Sect. 11.3), regardless of the pattern type induced at the upstream bottleneck, the wide moving jam propagates further upstream while maintaining the downstream jam front velocity \(v_g\).

In Figs. 19.1a,b, a wide moving jam has been induced at the downstream bottleneck (at the on-ramp \('D') through the use of a short-time local perturbation in the flow rate \(q_{on}^{(down)}\). Firstly, free flow is at the upstream bottleneck (at the on-ramp \('U'\). The wide moving jam propagates upstream and reaches the on-ramp \('U'\). Then depending on the flow rate \(q_{on}^{(up)}\), wide moving jam propagation through the on-ramp \('U'\) causes either LSP formation (Fig. 19.1a) or GP formation (Fig. 19.1b). The wide moving jam maintains the velocity \(v_g\) of the downstream jam front (for a comparison see empirical Fig. 4.2a).

After the wide moving jam is far upstream of the on-ramp \('U'\) the flow rate \(q_{out}\) plays the role of the flow rate \(q_{in}\) in free flow upstream of the upstream bottleneck (the on-ramp \('U') where \(q_{in} = q_{out}\). In this case, at the chosen model parameters there is no WSPs and no MSPs (Fig. 18.1a), i.e., either an LSP or an GP can only occur at the on-ramp \('U'\).
Fig. 19.1. Induced pattern formation and the catch effect. Speed on the main road in space and time. (a, b) Wide moving jams induce an LSP (a) and an GP (b). (c–e) MSPs induce an LSP (c), an WSP (d), and an GP (e). (f–h) WSPs induce an LSP (f), an WSP (g), and an GP (h). The initial flow rates to the on-ramp \('U'\) and on the road upstream of the on-ramp \('U'\) \((q_{on}^{(up)}, q_{in})\) are: (a) (320, 1800), (b) (450, 1756), (c) (400, 1800), (d) (200, 2118), (e) (440, 1800), (f) (300, 1756), (g) (150, 2034), and (h) (240, 200) vehicles/h. In (a–e) the flow rate to the on-ramp \('D'\) \(q_{on}^{(down)} = 60\) vehicles/h. In (f–h) \(q_{on}^{(down)}\) are 170 (f–g) and 180 (h) vehicles/h. Taken from [330]
SP upstream Propagation

Let us now discuss induced pattern formation at the upstream bottleneck caused by upstream propagation of an SP. We assume that this SP has initially occurred at the downstream bottleneck. In contrast with the above case, the initial SP is usually caught at the upstream bottleneck rather than propagating further upstream. This is regardless of the pattern type induced at the upstream bottleneck. Following the related effect in empirical observations (Sect. 10.4.2), this effect is called the catch effect.

In Fig. 19.1c, an MSP that has been induced at the on-ramp 'D' first propagates upstream. At the on-ramp 'U', the initial MSP causes an F→S transition. The MSP does not propagate further through the on-ramp 'U'. The initial MSP is caught by the on-ramp 'U'. Instead of the MSP an LSP occurs at the on-ramp 'U' (see empirical Fig. 10.6a). Two other examples of the catch effect are shown in Figs. 19.1d,e. An initial MSP is caught at the on-ramp 'U' in both cases. Instead of this MSP either an WSP (Fig. 19.1d) or an GP (Fig. 19.1e) are induced at the on-ramp 'U'.

If an WSP occurs at the on-ramp 'D' (Figs. 19.1f,g,h), then due to the catch of the WSP at the on-ramp 'U' either an LSP (Fig. 19.1f), or other WSP (Fig. 19.1g), or else one of the GPs (in Fig. 19.1h an DGP is shown) can be induced. Note that the induced WSP (Fig. 19.1g) exhibits different characteristics than the initial WSP. In particular, the absolute value of the negative velocity of the upstream front of the induced WSP (and the induced DGP in Fig. 19.1h) is considerably lower than that for the initial WSP.

All these effects and nonlinear features discussed above are also realized if the downstream bottleneck is the off-ramp bottleneck rather than the on-ramp 'D'.

19.3 Complex Congested Patterns and Pattern Interaction

19.3.1 Foreign Wide Moving Jams

In empirical observations, wide moving jams that have first emerged at the downstream bottleneck and then propagate through another congested pattern, which has earlier occurred at the upstream bottleneck, are called foreign wide moving jams (Sect. 11.3). Foreign wide moving jam propagation can be shown both in the fundamental diagram approach (see Knospe et al. [427]; Chap. 3) and in the three-phase traffic theory [331]. Probably, foreign jam propagation and the free flow metastability with respect to an F→J transition first predicted in [367] are the only two effects that can be shown within the scope of the fundamental diagram approach in accordance with empirical investigations.

Foreign moving jam propagation can lead to the following nonlinear effects of pattern interaction and transformation [330].
Induced Emergence and Suppression of SP

In Fig. 19.2a, first the upstream front of synchronized flow in an GP that emerges at the on-ramp 'D' is caught at the on-ramp 'U' and induces an WSP, as in Fig. 19.1g. However, after the first foreign wide moving jam of the GP has propagated through the on-ramp 'U', the WSP at the on-ramp 'U' is suppressed and free flow appears at the on-ramp 'U' (Fig. 19.2a).

![Fig. 19.2. Foreign wide moving jam propagation. Speed on the main road in space and time. In (e, f) the merge bottleneck is labeled “M”](image)

These effects are related to a decrease in the flow rate upstream of the on-ramp 'U' due to foreign moving jam propagation. In an example shown in Fig. 19.2a the initial flow rate $q_{in} = 2034$ vehicles/h upstream of the on-ramp 'U' (at which the WSP can exist) decreases to the flow rate $q_{out} =$...
1810 vehicles/h in the foreign wide moving jam outflow. At the flow rate $q_{in} = 1810$ vehicles/h the WSP cannot exist (Fig. 18.1a).

Another effect is shown in Fig. 19.2b. When foreign wide moving jams propagate through an LSP at the on-ramp 'U', the jams change the LSP width (in the longitudinal direction) and the speed within the LSP only.

**Induced Emergence of GP**

Let us assume that the flow rate $q_{in}^{(up)}$ and the flow rate $q_{in}$ on the main road upstream of the on-ramp 'U' are related to GP nucleation (Sect. 18.5.2). Then after the first (farthest upstream) foreign wide moving jam from another GP at the on-ramp 'D' propagates through the on-ramp 'U', the induced emergence of an GP occurs at the on-ramp 'U'. However, because in this case $q_{in} < q_{out}$, this foreign jam dissolves very quickly (Fig. 19.2c). After the GP upstream of the on-ramp 'U' has been formed, other foreign wide moving jams can nevertheless propagate further upstream before they dissolve. This is a similar effect to the GP at $q_{in} < q_{out}$ in Fig. 18.1f (Sect. 18.2.4).

**United Sequence of Wide Moving Jams**

If GPs are formed at both on-ramp bottlenecks, then narrow moving jams that emerge in the pinch region of the upstream GP at the on-ramp 'U' can be suppressed by foreign wide moving jams from the GP that has occurred at the on-ramp 'D' (Figs. 19.2c,d). Each of the foreign wide moving jams suppresses the growth of those narrow moving jams that are close enough to the downstream front of the foreign wide moving jam. Only narrow moving jams that are far from the foreign wide moving jam can grow and transform into a wide moving jam. Finally, an united sequence of wide moving jams occurs upstream of the on-ramp 'U'. This sequence consists of former foreign wide moving jams and new wide moving jams formed in the GP at the on-ramp 'U'.

**Dynamics of Foreign Wide Moving Jam Propagation from GP at Merge Bottleneck**

In Figs. 19.2e,f, the downstream bottleneck is a merge bottleneck and the upstream bottleneck is an on-ramp bottleneck. Moving jams that have emerged in an GP at the merge bottleneck propagate through an LSP that has occurred spontaneously at the on-ramp bottleneck (Fig. 19.2e). The foreign moving jams suppress the LSP. These jams fully determine the traffic dynamics on the main road upstream of the on-ramp bottleneck. The foreign moving jams propagate on the main road through the on-ramp bottleneck while maintaining the downstream jam velocity $v_g$. Thus, the moving jams in the GP at the merge bottleneck are indeed wide moving jams (Sect. 18.6).
In Fig. 19.2f, foreign wide moving jams from an GP at the merge bottleneck propagate on the main road through another GP at the on-ramp bottleneck. These foreign wide moving jams suppress the growth of all narrow moving jams that would emerge in the pinch region of the GP at the on-ramp bottleneck if there were no foreign wide moving jams. The foreign wide moving jams from the GP at the merge bottleneck fully determine the traffic dynamics at the on-ramp bottleneck.

19.3.2 Expanded Congested Patterns

Synchronized flow at a downstream bottleneck can first reach an upstream bottleneck and then propagate upstream of the upstream bottleneck. In empirical observations (Sect. 14.2), such a congested pattern is called the expanded congested pattern (EP) [218]. The downstream front of synchronized flow is fixed at the effective location of the downstream bottleneck and the upstream front of the synchronized flow is upstream of the upstream bottleneck; synchronized flow in an EP affects both bottlenecks. The following types of EP features have been found [330].

Nonlinear Interactions of Initial Congested Patterns by EP Formation

(i) An EP can consist of any spatiotemporal combinations between WSPs, LSPs, GPs, DGPs, and foreign wide moving jams. There are nonlinear interactions between these patterns in the EP. These interactions can change their features in comparison with pattern features at isolated bottlenecks (Sects. 18.2–18.7). In particular, an EP can consist of two interacting GPs formed on the main road upstream of each of two adjacent effectual bottlenecks. In Fig. 19.3a, two separated pinch regions occur in the EP. Wide moving jams that emerge in the GP at the on-ramp ‘D’ are foreign wide moving jams for the GP at the on-ramp ‘U’. In the example, synchronized flow affects both on-ramp bottlenecks because the distance between them \( L_1 = 4 \text{ km} \) is not too large.

(ii) In contrast, if the on-ramps are far enough away from one another, an EP is not formed. Instead, two different GPs occur so that free flow exists between wide moving jams downstream of the on-ramp ‘U’ (Figs. 19.2c,d where \( L_1 = 8 \text{ km} \)).

(iii) An EP can consist of an WSP at the on-ramp ‘D’ and different patterns at the on-ramp ‘U’: an WSP (Fig. 19.1g), or an LSP (Fig. 19.1f), or an DGP (Fig. 19.1h), or else an GP (Figs. 19.3b,c).

The WSP at the on-ramp ‘D’ can influence the patterns upstream because of two effects:

1. The WSP can cause pattern emergence at the on-ramp ‘U’ (catch effect) (Figs. 19.1f–h and Fig 19.3c), and
Fig. 19.3. Expanded congested patterns (EPs). Speed on the main road in space and time. (a) GP with two separated pinch regions. (b, c) EPs that consist of an GP at the on-ramp ‘U’ and an WSP at the on-ramp ‘D’. (d, e) Dissolution of an EP as a result of a suppression of an WSP at the on-ramp ‘U’ due to the upstream propagation of an GP (d) or an DGP (e) at the on-ramp ‘D’. (f) GP with the pinch region at the on-ramp ‘D’. (g) EP with the pinch region affecting both on-ramp bottlenecks. (h) $L_{\text{syn}}^{(\text{pinch})}(t)$ for the GP in (f) (triangles) and for the EP in (g) (circles); the mean value $L_{\text{syn}}^{(\text{mean})}$ is shown as a dotted line for (f) and a dashed line for (g). Taken from [330]
(2) the synchronized flow in the WSP can restrict the flow rate and speed in the outflow of congested patterns at the on-ramp 'U'. The latter pattern (either an LSP, or an WSP, or an DGP, or else an GP, Figs. 19.1f,g,h and Figs. 19.3b,c) influences the WSP at the on-ramp 'D', because the inflow to the WSP is the outflow from the upstream pattern.

(iv) If a GP under the strong congestion condition occurs at the on-ramp 'U' due to the upstream propagation of the upstream front of an WSP that has emerged at the downstream bottleneck (Fig. 19.3c), parameters of the GP depend only very slightly on the existence of the initial WSP at the downstream bottleneck. However, the features of the initial WSP at the on-ramp 'D' can be changed considerably after the GP has appeared. In contrast with the case of the WSP at an isolated on-ramp, the average speed within the WSP becomes a stronger falling function of distance in the upstream direction.

(v) An EP can also occur due to the interaction of an WSP at the on-ramp 'U' with either an GP (Fig. 19.3d) or an DGP (Fig. 19.3e) at the on-ramp 'D'. In the first case, rather than an EP, an GP affecting both on-ramp bottlenecks occurs later (Fig. 19.3d): the initial WSP at the on-ramp 'U' disappears. Thus, like the GP in Fig. 19.2e, the GP in Fig. 19.3d totally determines the traffic dynamics near the on-ramp 'U'. In the second case (Fig. 19.3e), after the wide moving jam from the DGP has reached the on-ramp 'U', the EP disappears, and either an LSP or free flow occurs at each the two on-ramp bottlenecks. Sometimes, however, an EP like the one in Fig. 19.1f can occur.

**Degree of Strong Congestion**

It is found that a GP under the strong congestion condition makes the major influence on an EP resulting from the interaction of two patterns at the downstream and upstream bottlenecks. Let us make the following definition of the degree of strong congestion. The lower the limit flow rate \( q_{\text{lim}}^{\text{pinch}} \) (per a freeway lane) in the GP, the higher the degree of strong congestion. It is found that the greater the degree of strong congestion in an GP, the greater the influence of this GP on EP characteristics. An example is shown in Fig. 19.2f, where an GP at the merge bottleneck exerts greater influence on the EP characteristics than an GP at the on-ramp bottleneck.

**United Pinch Region in EP**

If the distance between two adjacent effectual bottlenecks (due to the on-ramp 'U' and the on-ramp 'D') is commensurate with the width of the pinch region (in the longitudinal direction) in the GP at the isolated on-ramp bottleneck (in Fig. 19.3f where \( q_{\text{on}}^{\text{up}} = 0 \), \( L_{\text{syn}}^{\text{(mean)}} \approx 1.9 \text{ km} \)), then at \( q_{\text{on}}^{\text{up}} > 0 \) an GP...
occurs in which the pinch region affects both on-ramps, i.e., an EP occurs (Fig. 19.3g).

The width of the united pinch region in the EP is greater than the one for an GP at the isolated on-ramp bottleneck. In an example shown in Figs. 19.3h,g, \( L_{\text{syn}}^{(\text{mean})} \approx 2.7 \text{ km} \).

19.4 Intensification of Downstream Congestion Due to Upstream Congestion

It can be expected that upstream congestion (congested pattern formation at an upstream bottleneck) should tend to a reduction in downstream congestion. This is because at the same traffic demand the flow rate in free flow downstream of the congested bottleneck (discharge flow rate) is usually lower than the initial flow rate in free flow at the same freeway location before a congested pattern at the upstream bottleneck has been formed.

Nevertheless in some cases, we will see that rather than a reduction in downstream congestion, an intensification of downstream congestion can occur due to upstream congestion. This is the result of complex nonlinear interactions among congested patterns occurring at different spatially separated adjacent effectual bottlenecks.

The initial flow rate in free flow upstream of the on-ramp \('D'\)

\[
q_{\text{in}}^{(\text{down})} = q_{\text{in}} + q_{\text{on}}^{(\text{up})}
\]  

(19.1)

in each of the left panels of Fig. 19.4 is equal to the flow rate \( q_{\text{in}}^{(\text{down})} \) (19.1) in each of the associated right panels of Fig. 19.4. However, in Figs. 19.4a,c,e,g the flow rate

\[
q_{\text{on}}^{(\text{up})} = 0
\]

(19.2)

whereas in Figs. 19.4b,d,f,h the flow rate

\[
q_{\text{on}}^{(\text{up})} > 0
\]

(19.3)

The equality of the flow rate \( q_{\text{in}}^{(\text{down})} \) (19.1) in Figs. 19.4a,b is realized due to higher flow rate \( q_{\text{in}} \) in free flow upstream of the upstream bottleneck in Fig. 19.4a in comparison with the flow rate \( q_{\text{in}} \) in Fig. 19.4b. This is also valid for Figs. 19.4c,d, 19.4e,f, and 19.4g,h, respectively. Depending on the flow rate \( q_{\text{on}}^{(\text{up})} \) and on the congested pattern type at the downstream bottleneck, the following intensification phenomena for the initial downstream congestion are found.

Transformation of DGP into GP

An initial DGP at the on-ramp \('D'\), where only one wide moving jam emerges (Fig. 19.4a), transforms into an GP at this bottleneck where an uninterrupted sequence of wide moving jams emerges (Fig. 19.4b).
Fig. 19.4. Intensification of upstream congestion due to downstream congestion. Speed on the main road in space and time. (a, b) DGP (a) transforms into an GP (b). (c, d) In an WSP (c) the average speed decreases and a moving jam emerges (d). (e, f) WSP (e) transforms into an DGP (f). (g, h) In an GP (g) the frequency of moving jam emergence increases (h). Taken from [330]
To explain this, note that after the wide moving jam in the DGP (Fig. 19.4a) emerges, the current flow rate upstream of the on-ramp ‘D’, \(q_{\text{in}}^{(\text{down})}\), has dropped to \(q_{\text{out}}\): \[q_{\text{in}}^{(\text{down})} = q_{\text{out}}.\] (19.4)

This decrease in the flow rate \(q_{\text{in}}^{(\text{down})}\) occurs because vehicles merging onto the main road from the on-ramp ‘U’ must slow down when they reach the wide moving jam. In other words, when the wide moving jam is between the on-ramps ‘D’ and ‘U’ the flow rate \(q_{\text{on}}^{(\text{up})}\) does not lead to an increase in the flow rate \(q_{\text{in}}^{(\text{down})}\): due to the flow rate \(q_{\text{on}}^{(\text{up})}\) only the jam inflow increases. In the example, \(q_{\text{out}} = 1810\) vehicles/h. At the flow rate \(q_{\text{in}}^{(\text{down})} = 1810\) vehicles/h the GP cannot exist upstream of the on-ramp ‘D’ at the chosen flow rate to the on-ramp ‘D’, \(q_{\text{on}}^{(\text{down})}\) (Fig. 19.4a).

In Fig. 19.4b, an LSP is induced at the on-ramp ‘U’ after the wide moving jam from an DGP (downstream congestion) has reached the on-ramp ‘U’. After the wide moving jam has passed the on-ramp ‘U’, there are no wide moving jams between the on-ramps ‘D’ and ‘U’. For this reason, vehicles merging onto the main road from the on-ramp ‘U’ cause an increase in the flow rate \(q_{\text{in}}^{(\text{down})}\) to the value \(q_{\text{in}}^{(\text{down})} = 2150\) vehicles/h. This flow rate is much higher than 1810 vehicles/h. This leads to subsequent wide moving jam emergence at the on-ramp ‘D’, and so on. As a result, an GP is formed at the on-ramp ‘D’ (Fig. 19.4b).

The mean period of moving jam emergence in the GP is determined by the time of wide moving jam propagation to the on-ramp ‘U’. Before a wide moving jam has passed the on-ramp ‘U’ the condition (19.4) is satisfied, and no moving jam can emerge at the given flow rate \(q_{\text{on}}^{(\text{down})}\) at the on-ramp ‘D’.

### Decrease in Speed within WSP

The onset of upstream congestion can lead to a decrease in average speed within an initial WSP (Figs. 19.4c,e) at the downstream bottleneck.

This intensification of downstream congestion sometimes occurs with moving jam emergence in the initial WSP (Figs. 19.4d,f), i.e., it is accompanied by transformation of this WSP into an DGP (Fig. 19.4f) or an GP. The speed within an GP at the on-ramp ‘U’ is low. This can lead to a decrease in speed and to jam emergence in the initial WSP at the on-ramp ‘D’ (Fig. 19.4d).

### Increase in Frequency of Moving Jam Emergence

The onset of upstream congestion can lead to the intensification of downstream congestion that results in an increase in the mean frequency of wide
moving jam emergence in an initial GP at the downstream bottleneck (the on-ramp 'D') (Fig. 19.4g). In this GP, weak congestion is realized.

This effect is similar to the transformation of an DGP into an GP discussed above. After the first wide moving jam has emerged in the GP (Fig. 19.4g), the current flow rate upstream of the on-ramp 'D', \( q_{in}^{(down)} \), has dropped to \( q_{out}^{(J)} \leq q_{out} \) (7.16). When the wide moving jam is between the on-ramps 'D' and 'U', vehicles merging onto the main road from the on-ramp 'U' increase the jam inflow only. At the flow rate \( q_{in}^{(down)} = q_{out}^{(J)} \) the frequency of wide moving jam emergence in the GP upstream of the on-ramp 'D' is relatively low at the chosen flow rate \( q_{on}^{(down)} \) to the on-ramp 'D' (Fig. 19.4g).

In contrast, in Fig. 19.4h after the wide moving jam has passed the on-ramp 'U', first there are no wide moving jams between the on-ramps 'D' and 'U'. For this reason, vehicles merging onto the main road from the on-ramp 'U' cause an increase in the flow rate \( q_{in}^{(down)} \) upstream of the pinch region of the GP. This increase in the flow rate in the inflow of the pinch region of the GP causes the emergence of a new wide moving jam in the GP during a considerably shorter time interval than the mean time between wide moving jams in the GP shown in Fig. 19.4g. As a result, the mean frequency of wide moving jam emergence in the GP at the on-ramp 'D' increases (Fig. 19.4h).

The intensification of downstream congestion due to upstream congestion has been observed when weak congestion occurs at the downstream bottleneck. To explain this result, let us recall that under weak congestion conditions congested pattern characteristics depend considerably on values of traffic variables in the pinch region of GPs (see Sect. 18.3 and Fig. 18.12). In particular, the lower the average speed in the pinch region of a GP under the weak congestion condition, the higher the frequency of narrow moving jam emergence in the pinch region of the GP, \( f_{narrow} \) (18.5). Thus, the frequency of narrow moving jam emergence can change when parameters of the GP under weak congestion change. In contrast, under strong congestion the frequency \( f_{narrow} \) reaches the maximum possible value (18.12) that does not change at given control parameters of traffic (weather, road conditions, etc.). These results are related to empirical observations where diverse transformations among varied congested patterns occur at on- and off-ramp bottlenecks under weak congestion (Sect. 13.3).

19.5 Conclusions

(i) The microscopic three-phase traffic theory explains the empirical catch effect (Sects. 2.4.6 and 10.4.2), empirical phenomena at an upstream bottleneck due to foreign wide moving jam propagation (Sects. 2.4.9 and 11.3), and a diverse variety of complex empirical EPs (Sects. 2.4.8 and 14.2). In accordance with empirical results, we have found:
(1) If two adjacent effectual bottlenecks are close to one another, a diverse variety of complex EPs in which synchronized flow affects these bottlenecks can occur.

(2) An SP that has emerged at a downstream bottleneck can be caught at an upstream bottleneck (catch effect).

(3) The upstream propagation of an MSP, an WSP, or a foreign wide moving jam produced at the downstream bottleneck can induce diverse patterns at the upstream bottleneck.

(4) A foreign wide moving jam that has emerged at the downstream bottleneck propagating through the pinch region of an GP emerging at the upstream bottleneck can prevent a narrow moving jam growing in the pinch region. Due to foreign wide moving jam propagation, an united sequence of wide moving jams appear. This sequence consists of foreign wide moving jams and wide moving jams that have emerged in the GP at the upstream bottleneck.

(ii) An GP with the highest degree of strong congestion exerts the major influence on EPs features.

(iii) Due to the onset of congestion at the upstream bottleneck, congestion at the downstream bottleneck can intensify.
20 Spatiotemporal Patterns in Heterogeneous Traffic Flow

20.1 Introduction

In this chapter, a microscopic three-phase traffic theory for heterogeneous traffic flow with various driver behavioral characteristics and vehicle parameters developed in [332] is discussed. Results of this theory are compared with features of patterns in traffic flow with identical vehicles presented in Chaps. 17 and 18.

We will find that different driver behavioral characteristics and different vehicle parameters lead to a well-known lane specific behavior in free traffic flow: fast vehicles use mostly the left (passing) lane of a freeway whereas slow and long vehicles use mostly the right freeway lane (e.g., [20]). This is the vehicle lane separation effect in free flow. As a result, the average speed in free flow in the left lane is higher than the average speed in the right lane. The vehicle lane separation effect is realized when the percentage of slow and/or long vehicles in heterogeneous traffic flow is a relatively high value in comparison with the percentage of fast vehicles.

It will be shown that different driver characteristics and different vehicle parameters can change certain quantitative spatiotemporal congested pattern parameters and conditions for pattern emergence discussed in Chaps. 17 and 18 for traffic flow with identical vehicles. There can also be some secondary and specific effects in congested traffic caused by these different driver behavioral characteristics and different vehicle parameters. However, simulations show qualitatively the same fundamental features of congested patterns as those in the case when all drivers and all vehicles have the same characteristics and parameters. In particular, we will find that all specific pattern properties gradually disappear when differences between driver behavioral characteristics and vehicle parameters in traffic flow decrease. This gradual transformation from heterogeneous traffic flow to traffic flow with identical vehicles does not qualitatively change fundamental traffic pattern features. This means that specific pattern properties associated with different driver behavioral characteristics and vehicle parameters are secondary effects in comparison with fundamental traffic pattern features (see Sect. 2.5).
20.2 Microscopic Two-Lane Model for Heterogeneous Traffic Flow with Various Driver Behavioral Characteristics and Vehicle Parameters

We consider heterogeneous traffic flow where there are three types of vehicles: “fast vehicles,” “slow vehicles,” and “long vehicles.” Fast and slow vehicles have the same vehicle length that is lower than the length of long vehicles. The maximum vehicle speed in free flow of fast vehicles is higher than the one for slow and long vehicles. There are also other model parameters and variables that are different for different drivers and vehicles. In the model, fast, slow, and long vehicles are specified by a vehicle identifier $j$. The vehicle identifier is $j = 1$ for fast vehicles, it is $j = 2$ for slow vehicles, and it is $j = 3$ for long vehicles. All model parameters and variables that are chosen in the model different for the three types of vehicles are marked by superscripts $(j)$ where $j = 1, 2, 3$. The percentages of fast vehicles $\eta^{(1)}$, slow vehicles $\eta^{(2)}$, and long vehicles $\eta^{(3)}$ in heterogeneous flow satisfy the obvious condition:

$$\eta^{(1)} + \eta^{(2)} + \eta^{(3)} = 100\% .$$  \hspace{1cm} (20.1)

To take into account different driver behavioral characteristics and different vehicle parameters in heterogeneous traffic flow, the following changes have been made in the model for identical vehicles of Sect. 16.3.

20.2.1 Single-Lane Model

Steady States and Vehicle Motion

Equations (16.24)–(16.27) are also the general rules of vehicle motion in heterogeneous flow. However, in heterogeneous flow, the maximum vehicle speed $v_{\text{free}}$ in (16.24) is

$$v_{\text{free}} = v_{\text{free}}^{(j)}, \quad j = 1, 2, 3 ,$$  \hspace{1cm} (20.2)

where $v_{\text{free}}^{(1)}$, $v_{\text{free}}^{(2)}$, and $v_{\text{free}}^{(3)}$ are constant values.\(^1\)

In (16.26), the synchronization distance $D_n$ has been rewritten in the form

$$D_n = d_\ell + G(v_n, v_{\ell,n}) ,$$  \hspace{1cm} (20.3)

where the function $G(u, w)$ is

$$G(u, w) = \max(0, k\tau u + \phi_0 a^{-1} u(u - w)) ,$$  \hspace{1cm} (20.4)

$$k = k^{(j)}, \quad j = 1, 2, 3 ,$$  \hspace{1cm} (20.5)

$k^{(1)}$, $k^{(2)}$, and $k^{(3)}$ are constants, $d_\ell$ is the length of the preceding vehicle

\(^1\) In free flow, besides different maximum vehicle speeds very different time gaps between vehicles are observed (e.g., [193, 195]). It can be assumed that these and many other empirical statistical features of free flow are associated with various driver characteristics and vehicles parameters. However, a description of statistical features of free flow is beyond the scope of this book (Chap. 1).
that can be different from the length of the vehicle \( d \) where

\[
d = d^{(j)}, \quad j = 1, 2, 3.
\]  

Two-dimensional regions of steady-state model solutions in the flow–density plane for traffic flows in which either all vehicles are fast vehicles, or all vehicles are slow vehicles, or else all vehicles are long vehicles are shown in Figs. 20.1a,b,c, respectively.

![Diagram of steady-state model solutions](image)

**Fig. 20.1.** Steady-state model solutions and the lines \( J \) for the following cases. (a) All vehicles are fast vehicles. (b) All vehicles are slow vehicles. (c) All vehicles are long vehicles. The flow rate in the outflow from a wide moving jam and the downstream jam front velocity are \( q_{\text{out}}^{(1)} = 1900 \text{ vehicles/h} \) and \( v_{g}^{(1)} = -16.2 \text{ km/h} \) for fast vehicles, \( q_{\text{out}}^{(2)} = 1510 \text{ vehicles/h} \) and \( v_{g}^{(2)} = -13 \text{ km/h} \) for slow vehicles, \( q_{\text{out}}^{(3)} = 1130 \text{ vehicles/h} \) and \( v_{g}^{(3)} = -24.5 \text{ km/h} \) for long vehicles, respectively. Taken from [332]

### Fluctuations

Random acceleration and deceleration are described by (16.30)–(16.38). However, in heterogeneous flow, the probabilities \( p_b \) and \( p_a \) in (16.33) and (16.34) are

\[
p_b = p_b^{(j)}, \quad j = 1, 2, 3.
\]

\[
p_a = p_a^{(j)}, \quad j = 1, 2, 3.
\]
Moreover, in (16.37)
\[ p_0(v) = p_0^{(j)}(v), \ j = 1, 2, 3, \]  
\[ p_0^{(1)}(v) > p_0^{(2)}(v) > p_0^{(3)}(v). \]  
The mean time delays in vehicle acceleration
\[ \tau_{\text{del}}^{(a)}(v) = \tau_{\text{del}}^{(a, j)}(v), \ j = 1, 2, 3 \]  
are
\[ \tau_{\text{del}}^{(a, j)}(v) = \frac{\tau}{p_0^{(j)}(v)}, \ j = 1, 2, 3, \]  
\[ \tau_{\text{del}}^{(a, 1)}(v) < \tau_{\text{del}}^{(a, 2)}(v) < \tau_{\text{del}}^{(a, 3)}(v). \]  
Corresponding to (20.13), fast vehicles have a shorter time delay than the related time delays for slow and long vehicles, i.e., fast vehicles prefer a more aggressive drive.

**Safe Speed**

In (16.41) for the safe speed as in all other related formulae of Sect. 16.3, the space gap \( g_n \) in heterogeneous flow has been rewritten in the form
\[ g_n = x_{\ell,n} - x_n - d_{\ell}. \]  

**20.2.2 Two-Lane Model**

As in Sect. 16.3.8, lane changing rules in a two-lane model of heterogeneous flow are based on incentive and security conditions. However, these conditions should be adjusted for heterogeneous flow. The following incentive conditions for lane changing from the right lane to the left (passing) lane \((R \rightarrow L)\) and a return change from the left lane to the right lane \((L \rightarrow R)\) have been used in the model of heterogeneous flow:
\[ R \rightarrow L: \ v^+_n \geq v_{\ell,n} + \delta_1 \text{ and } v_n \geq v_{\ell,n}, \]  
\[ L \rightarrow R: \ v^+_n > v_{\ell,n} + \delta_2 \text{ or } v^+_n > v_n + \delta_2, \]  
where
\[ \delta_1 < \delta_2 \text{ for fast vehicles }, \]  
\[ \delta_2 < \delta_1 \text{ for slow and long vehicles }, \]  
\( \delta_1 \geq 0, \delta_2 \geq 0 \) are constants.

It is assumed that if the vehicle speed in the right lane is high enough, slow and long vehicles moving in the left lane are usually forced to change to the right lane, whereas slow and long vehicles moving in the right lane retain
20.2 Microscopic Two-Lane Model for Heterogeneous Traffic

in this lane. To simulate this effect, the following incentive conditions for slow and long vehicles have been applied. For lane changing from the left lane to the right lane \((L \rightarrow R)\) for slow and long vehicles the incentive condition is (20.16) or

\[
L \rightarrow R: \quad v_n^+ > v_{\text{free}}^{(j)} - \delta_0, \; j = 2, 3 ,
\]

where \(\delta_0 > 0\) is a constant. For lane changing from the right lane to the left lane \((R \rightarrow L)\) for slow and long vehicles the incentive conditions are (20.15) and

\[
R \rightarrow L: \quad v_{\ell, n} \leq v_{\text{free}}^{(j)} - \delta_0, \; j = 2, 3 .
\]

The security conditions for lane changing are given by the inequalities:

\[
g_n^+ > \min(v_n \tau, \; G^+_n) ,
\]

\[
g_n^- > \min(v_n \tau, \; G^-_n) .
\]

Here

\[
G^+_n = G(v_n, v_n^+) ,
\]

\[
G^-_n = G(v_n^-, v_n) ,
\]

where the function \(G(u, w)\) is given by (20.4).

If the incentive and security conditions are satisfied, then as in Rickert et al. [481] in this model the vehicle changes the lane with the probability \(p_c < 1\) \((p_c = 1\) if (20.19) is satisfied).

Slow and long vehicles can change from the left lane to the right lane even if the security conditions (20.21), (20.22) are not satisfied. This lane changing is realized if the above incentive condition (20.19) is satisfied and the following two security conditions are satisfied: (i) The gap between two neighboring vehicles in the right lane on the main road exceeds some value \(g^{(\text{min})}\), i.e.,

\[
x_n^+ - x_n^- - d^+ > g^{(\text{min})} .
\]

(ii) The vehicle passes the point

\[
x_n^{(m)} = (x_n^+ + x_n^- + d - d^+)/2
\]

between two vehicles in the target lane of the main road for time step \(n\), i.e.,

\[
x_{n-1} < x_{n-1}^{(m)} \quad \text{and} \quad x_n \geq x_n^{(m)}
\]

or

\[
x_{n-1} \geq x_{n-1}^{(m)} \quad \text{and} \quad x_n < x_n^{(m)} .
\]

In (20.25)

\[
g^{(\text{min})} = \lambda v_n^+ + d ,
\]

\[
\lambda = \lambda^{(j)}, \; j = 2, 3 ,
\]
\(\lambda^{(2)}\) and \(\lambda^{(3)}\) are constants. If, in accordance with (20.19), (20.25)–(20.28), a vehicle changes lane, then the vehicle coordinate in the target lane is set to \(x_n = x^{(m)}_n\). These security conditions facilitate lane changing for slow and long vehicles from the left lane to the right lane when the security rules (20.21), (20.22) are not satisfied.

Two slow and/or long vehicles moving side by side in the left and right lanes hinder fast vehicles accelerating in free flow on a two-lane road [482]. To avoid this effect, it is assumed that either a slow vehicle or a long vehicle in the left lane, which should change lane, can move with a higher maximum speed in free flow before the vehicle changes to the right lane. For this purpose, when the condition (20.19) is satisfied, the maximum speed \(v_{\text{free}}\) in (20.2) for slow or long vehicles in the left lane is equal to

\[
v_{\text{free}}^{(2, \text{left})} > v^{(2)}_{\text{free}}
\]

and

\[
v_{\text{free}}^{(3, \text{left})} > v^{(3)}_{\text{free}}
\]

for slow and long vehicles, respectively; \(v_{\text{free}}^{(2, \text{left})}\) and \(v_{\text{free}}^{(3, \text{left})}\) are constants.

### 20.2.3 Boundary, Initial Conditions, and Model of Bottleneck

In the model, open boundary conditions of Sect. 16.3.9 are used; however, in the related formulae the length of the preceding vehicle \(d\) should be replaced by \(d_\ell\).

In the initial state \((n = 0)\), all vehicles have a related maximum speed \(v_n = v^{(j)}_{\text{free}}, \ j = 1, 2, 3\) and they are positioned at space intervals \(x_{\ell,n} - x_n = v_{\text{free}}^{(j)}\tau_{in}, \ j = 1, 2, 3\).

In the model, the vehicle identifier \(j\) and the related vehicle parameters are ascribed to the vehicle as its individual “attributes” when the vehicle is generated at the beginning of the road.

In the model, there are at least two different possibilities to generate vehicles of various types at the beginning of the road: (1) Fast, slow, and long vehicles are randomly generated in the left and right lanes with the rates related to chosen values of the flow rate \(q_{\text{in}}\) and the percentages \(\eta^{(1)}\), \(\eta^{(2)}\), and \(\eta^{(3)}\). (2) Fast vehicles are preferably generated in the left lane whereas slow and long vehicles are preferably generated in the right lane. In the case (2), only \(\max(0, 2\eta^{(1)} - 100\%)\) of fast vehicles are randomly generated in the right lane, whereas only \(\max(0, 2\eta^{(2)} + 2\eta^{(3)} - 100\%)\) of slow and long vehicles are randomly generated in the left lane.

In the model of the on-ramp bottleneck (Sect. 16.3.9), the rule (***) has been rewritten as follows. The space gap between two neighboring vehicles in the right lane on the main road exceeds some value \(g_{\text{on}}^{(\text{min})}\), i.e.,

\[
x^{+}_n - x^{-}_n - d^{+} > g_{\text{on}}^{(\text{min})},
\]

\((20.32)\)
20.3 Patterns in Flow with Different Driver Characteristics

\begin{equation}
\theta^{(\text{on})} = \lambda^{(\text{on})} v_n^+ + d, \tag{20.33}
\end{equation}

\(\lambda^{(\text{on})}\) is a constant. In addition, the condition (20.27) that the vehicle passes the point \(x_n^{(\text{m})}\) (20.26) between two neighboring vehicles in the target lane of the main road for time step \(n\) should be satisfied. In this case, the coordinate of the merging vehicle is set to \(x_n = x_n^{(\text{m})}\) and the vehicle speed \(v_n\) is changed in accordance with (16.84).

20.2.4 Simulation Parameters

In simulations of patterns at a homogeneous road, the length of the road is \(L = 20\text{ km}\). In simulations of patterns at the on-ramp bottleneck, the length of the main road is \(L = 40\text{ km}\), the point \(x = 0\) is at the distance \(L/2 = 20\text{ km}\) from the end of the road. The beginning of the main road is at \(x_b = -L/2 = -20\text{ km}\). \(v_{\text{free}}^{(1)} = 33.3\text{ m/s (120 km/h)}\), \(v_{\text{free}}^{(2)} = v_{\text{free}}^{(3)} = 25\text{ m/s (90 km/h)}\), \(d^{(1)} = 7.5\text{ m}\), \(d^{(2)} = 17\text{ m}\), \(k^{(1)} = k^{(2)} = 3\), \(k^{(3)} = 4\), \(p_{a}^{(3)} = 0.3\), \(p_{b}^{(1)} = p_{b}^{(2)} = 0.17\), \(p_{b}^{(3)} = 0.2\), \(p_{b}^{(1)} = p_{b}^{(2)} = 0.1\). \(x_{\text{on}} = 10\text{ km}\). The probabilities \(p^{(j)}(v), j = 1, 2, 3, \text{ in (20.9)}\) are \(p^{(1)}(v) = 0.6 + 0.17 \min (1, v/10)\), \(p^{(2)}(v) = 1 - 1.3(1 - p^{(1)}(v))\), \(p^{(3)}(v) = 1 - 1.5(1 - p^{(1)}(v))\).

Lane changing parameters are \(\delta_1 = 1\text{ m/s}, \delta_2 = 3.5\text{ m/s}\) for fast vehicles; \(\delta_1 = 3.5\text{ m/s}, \delta_2 = 1\text{ m/s}\) for slow and long vehicles; \(\delta_0 = 6\text{ m/s}\). \(\lambda^{(2)} = \lambda^{(3)} = 0.87\); \(v_{\text{free}}^{(1, \text{left})} = 28.5\text{ m/s}\) and \(v_{\text{free}}^{(3, \text{left})} = 27.5\text{ m/s}\); \(\lambda^{(\text{on})} = 0.72\tau\). Other parameters are the same as those in the model of identical vehicles of Sect. 16.3.

20.3 Patterns in Heterogeneous Traffic Flow with Different Driver Behavioral Characteristics

20.3.1 Vehicle Separation Effect in Free Flow

Numerical simulations show that when, in heterogeneous traffic flow, percentages of fast and slow vehicles are comparable with one another, then there is the well-known lane specific behavior in free flow. Fast vehicles try to change to the left lane (passing lane), whereas slow vehicles are required to move in the right lane. As a result, the average speed in the left lane is higher than in the right lane (Fig. 20.2).

At the beginning of the road the difference between the average speeds in the left and right lanes is low. During further moving on the road fast vehicles change to the left lane. In contrast, slow vehicles have to change to the right lane. This lane changing leads to lane separation of fast and slow vehicles: beginning at some road location all fast vehicles move with their maximum speed \(v_{\text{free}}^{(1)}\) in the left lane and all slow vehicles move with their
maximum speed $v_{\text{free}}^{(2)}$ in the right lane. Because $v_{\text{free}}^{(1)} > v_{\text{free}}^{(2)}$ the speed in the left (passing) lane is higher than the speed in the right lane. The lane separation effect can explain empirical free flow where the average speed in the passing lane is higher than in other freeway lanes (Fig. 2.3a–d).

### 20.3.2 Onset of Congestion in Free Flow on Homogeneous Road

As in flow with identical vehicles (Sect. 17.2), an F→S transition can occur in heterogeneous flow. The F→S transition is realized when the flow rate is within the range (17.1) and a local perturbation appears, whose amplitude $\Delta v_{\text{initial}}^{(\text{pert})}$ exceeds some critical amplitude $\Delta v_{\text{cr}}^{(FS)}$.

However, there is a peculiarity of the F→S transition in heterogeneous traffic flow. If a local perturbation initially appears in only one of the freeway
lanes, then the critical amplitude of the local perturbation $\Delta v_{cr}^{(FS)}$ depends on which of the lanes the perturbation has initially occurred. At the same flow rate $q_{in}$ in free flow the critical amplitude of a local perturbation in the right lane (curve 1 in Fig. 20.3a) is lower than the critical amplitude of the perturbation in the left lane (curve 2 in Fig. 20.3a). In particular, at the critical point $q_{in}^{(free)}$ the critical amplitude of the local perturbation $\Delta v_{cr}^{(FS, left)}$ for the left lane (Fig. 20.3a, curve 2) is higher than zero whereas the critical amplitude of the local perturbation $\Delta v_{cr}^{(FS, right)}$ for the right lane is zero (curve 1).

As in flow with identical vehicles, the threshold point $(\rho_{th}, q_{th})$ is related to the condition (5.29). In the vicinity of the threshold point, the duration of the initial perturbation $T_{initial}^{(pert)}$ (Fig. 20.3b) should be increased for $F \rightarrow S$ transition occurrence (dashed parts of curves 1 and 2 in Fig. 20.3a) in comparison with a constant duration $T_{initial}^{(pert)} = 10$ s used for the solid parts of curves 1 and 2 in Fig. 20.3a. On the dashed parts of curves 1 and 2 in Fig. 20.3a the duration of the initial perturbation $T_{initial}^{(pert)}$ increases when the flow rate $q_{in}$ decreases.

If $v_{down}, v_{up} < 0$ and $|v_{down}| > |v_{up}|$, then initial perturbations dissolve over time (Fig. 20.3c). At the threshold point where the condition (5.29) is satisfied, an MSP can be excited by applying a local perturbation. The downstream and upstream fronts of the MSP move with the same mean velocity (Fig. 20.3d). However, at the threshold point model fluctuations in synchronized flow of the MSP can lead, over time, to a return $S \rightarrow F$ transition. Consequently, the MSP disappears (Fig. 20.3d). When $q_{in} > q_{th}$ and $\Delta v_{initial}^{(pert)} \geq \Delta v_{cr}^{(FS)}$, an $F \rightarrow S$ transition occurs and an MSP emerges where $|v_{down}| < |v_{up}|$ (Fig. 20.3e). The width of this MSP (in the longitudinal direction) increases continuously over time.

However, if in the vicinity of the threshold point $(\rho_{th}, q_{th})$ the amplitude of the initial perturbation $\Delta v_{initial}^{(pert)}$ exceeds $\Delta v_{cr}^{(FS)}$ appreciably, then instead of the MSP a wide moving jam can be formed in synchronized flow of the incipient MSP. This $F \rightarrow J$ transition occurs if the amplitude of a local perturbation $\Delta v_{initial}^{(pert)}$ in free flow exceeds the critical amplitude $\Delta v_{cr}^{(FJ)}$. In this case, first synchronized flow begins to form in free flow and later a wide moving jam emerges spontaneously in that synchronized flow. Thus, the $F \rightarrow J$ transition is a sequence of $F \rightarrow S \rightarrow J$ transitions.

If a local perturbation initially appears in only one of the freeway lanes, then the critical amplitude of the local perturbation $\Delta v_{cr}^{(FJ)}$ for an $F \rightarrow J$ transition depends on which of the lanes the perturbation has occurred. At the same flow rate $q_{in}$ the critical amplitude of a local perturbation in the right lane (curve 3 in Fig. 20.3a) is lower than the critical amplitude of the perturbation in the left lane (curve 4 in Fig. 20.3a). As in flow with identical vehicles (Sect. 17.2), the critical amplitude of the perturbation required for an $F \rightarrow S$ transition is always lower than the one required for an $F \rightarrow J$ transition.
Fig. 20.3. Phase transitions in heterogeneous flow on a homogeneous road. (a) Dependencies of critical amplitude of a local perturbation for MSP excitation ($\Delta v_{cr} = \Delta v_{cr,FS}$; curves 1 and 2) and for wide moving jam excitation ($\Delta v_{cr} = \Delta v_{cr,FJ}$; curves 3 and 4) on the flow rate $q_{in}$. (b) Speed of the vehicle that is the source of a local perturbation as a function of time. (c–e) Evolution of the initial local perturbation in space and time at $q_{in} < q_{th}$ (c), $q_{in} = q_{th}$ (d), and $q_{in} > q_{th}$ at $\Delta v_{initial}^{(pert)} > \Delta v_{cr}^{(FS)}$ (e). In (a) curves 1, 3 are related to initial perturbations in the right lane; curves 2, 4 are related to initial perturbations in the left lane. $T_{initial}^{(pert)} = 10$ s for the solid parts of curves 1, 2 and for curves 3, 4 in (a). In (c–e) the inflow rate $q_{in}$ is: (c) 2117, (d) 2145, (e) 2200 vehicles/h. $\eta^{(1)} = 50\%$, $\eta^{(2)} = 50\%$. Taken from [332].
20.3.3 Lane Asymmetric Emergence of Moving Synchronized Flow Patterns

Let us assume that the flow rate $q_{in} > q_{th}$. We consider the further development of an initial local perturbation that occurs only in the right lane. The amplitude of this perturbation is assumed to be slightly higher than the critical amplitude for this lane, $\Delta v_{cr}^{(FS, \text{right})}$. In this case, an $F \rightarrow S$ transition occurs first only in the right lane. As a result, an MSP appears in the right lane (Fig. 20.4b) whereas in the left lane vehicles move with the free flow speed (Fig. 20.4a).

Because the speed in the right lane sharply decreases within the MSP, some of the slow vehicles try to change the lane. As a result, over time the speed in the left lane also decreases. This leads to the self-formation of an MSP in both lanes. This is due to the effect of synchronization of the vehicle speed between different lanes in synchronized flow.

This lane asymmetric MSP emergence is caused by different driver behavioral characteristics in heterogeneous traffic flow. However, after the effect of the speed synchronization between different lanes has occurred, qualitative features and spatiotemporal structure of the MSP (Fig. 20.4c,d; $x = 5 \text{ km}$ and $x = 0 \text{ km}$) are the same as those for the MSP in flow with identical vehicles (Fig. 5.10). The vehicle separation within the downstream front of the MSP can lead to a change in the velocity of this front over time.

If an initial local perturbation appears in the left lane only and the amplitude of this perturbation is greater than $\Delta v_{cr}^{(FS, \text{left})}$, then the speed synchronization effect occurs very quickly (Fig. 20.5). In this case, there is almost no time delay in MSP formation in the left and right lanes (Fig. 20.5a,b). When the speed begins to decrease in the left lane, fast vehicles in the region of the lower speed try to change to the right lane where the speed is higher. This synchronizes the average speeds between the lanes and leads to MSP emergence in both lanes (Fig. 20.5c,d). Also in this case, after the MSP is formed in both lanes features of this MSP are qualitatively the same as those for the MSP in flow with identical vehicles.

20.3.4 Congested Patterns at On-Ramp Bottlenecks

The diagram of congested patterns in traffic flow with fast and slow vehicles at an on-ramp bottleneck (Fig. 20.6a) is qualitatively similar to the diagram for identical vehicles (Fig. 18.1a). The same three types of synchronized flow patterns (WSP, LSP, and MSP, Fig. 20.6b–d) can occur spontaneously between the boundaries $F^{(B)}_S$ and $S^{(B)}_J$ in Fig. 20.6a. Right of the boundary $G$ and right of the boundary $S^{(B)}_J$ an GP occurs. In the GP, the pinch region continuously exists where narrow moving jams emerge spontaneously. Some of these narrow moving jams transform into wide moving jams. As a result, a sequence of wide moving jams propagating upstream appears (Fig. 20.7a,b).
Fig. 20.4. MSP on a homogeneous road excited by a local perturbation in the right lane. (a, b) Speed distributions in time and space in the left (a) and right (b) lanes. (c, d) Vehicle speed (c) and of flow rate (d); one minute average data of virtual detectors. Curves 1 and 2 are related to the left and right lanes, respectively. $q_{in} = 2230$ vehicles/h. The initial local perturbation with amplitude $\Delta v_{initial}^{(pert)} = 5$ m/s and $T_{initial}^{(pert)} = 1$ s is applied at $t_0 = 7$ min at the location $x = 11$ km. $\eta^{(1)} = 50\%$, $\eta^{(2)} = 50\%$. Taken from [332]

Right of the boundary $S_J^{(B)}$ and left of the boundary $G$, a dissolving GP (DGP) occurs (Fig. 20.7c) where the pinch region is dissolved after wide moving jam formation.

Within SPs and GPs the vehicle speed is appreciably synchronized across the lanes (Fig. 20.8a,c,e,g). However, at the downstream front of SPs and GPs when vehicles accelerate from synchronized flow to free flow downstream the effect of the vehicle separation is realized (Sect. 20.3.1). For this reason, in
20.3 Patterns in Flow with Different Driver Characteristics

Fig. 20.5. MSP on a homogeneous road excited by a local perturbation in the left lane. (a, b) Speed distributions in time and space in the left (a) and right (b) lanes. (c, d) Vehicle speed (c) and flow rate (d); one minute average data of virtual detectors. Curves 1 and 2 are related to the left and right lanes, respectively. $\Delta t_{\text{initial}}^{(\text{pert})} = 20 \text{ m/s}$. Other parameters are the same as those in Fig. 20.4. Taken from [332]

free flow downstream of SPs and GPs there is a large difference between the average vehicle speeds in the left and right lanes (Fig. 20.8a,e,g).

At a high flow rate $q_{\text{in}}$ and a low flow rate $q_{\text{on}}$ MSPs can occur spontaneously at the bottleneck (Fig. 20.6d). Often an MSP appears first only in the right lane whereas in the left lane vehicles move with the free flow speed. The physics of this lane asymmetric MSP emergence is the same as those discussed above for the case of asymmetric MSP emergence on a homogeneous road. Over time some of the slow vehicles moving in the right lane change
Fig. 20.6. Diagram of congested patterns at an on-ramp bottleneck in heterogeneous traffic flow with fast and slow vehicles (a) and speed on the main road in space and time within SPs (b–d). (b) WSP. (c) LSP. (d) MSP. In (b–d) the flow rates \( q_{on}, q_{in} \) are: (b) \((200, 2195)\), (c) \((400, 1895)\), and (d) \((35, 2235)\) vehicles/h. \( \eta^{(1)} = 50\% \), \( \eta^{(2)} = 50\% \). In (a) criteria for the boundaries \( F^{(B)}_S \) and \( S^{(B)}_J \) are the same as those in Fig. 18.1. \( q_{max, lim} \approx 2250 \) vehicles/h. Taken from [332]
20.3 Patterns in Flow with Different Driver Characteristics

Fig. 20.7. GPs related to the diagram in Fig. 20.6a. Speed on the main road in space and time. (a) GP at $q_{in} > q_{out}$. (b) GP at $q_{in} < q_{out}$. (c) DGP. The flow rates $(q_{on}, q_{in})$ are: (a) (850, 2180), (b) (1250, 1600), and (c) (360, 2195) vehicles/h. Taken from [332]

to the left lane where the speed is higher. As a result, after a time delay an MSP occurs in both lanes.

This time delay in MSP formation in the left lane is a random effect. In other realizations at the same initial conditions, it can turn out that an MSP appears almost simultaneously in both lanes. To explain this, note that during MSP formation in the right lane there are some fast vehicles that merge onto the right lane on the main road from the on-ramp. These fast vehicles have to move with a lower speed in the right lane within the MSP. When these fast vehicles change to the left lane, then due to their low initial speed they decrease the speed in the left lane. This leads to MSP emergence in both lanes.

There is also the effect of the lane asymmetric emergence of wide moving jams in the pinch region of an GP: a moving jam often occurs first only in the right lane. Only after a time delay the moving jam also appears in the
Fig. 20.8. Space distributions of vehicle speed (a, c, e, g) and flow rate (b, d, f, h) at given times for an WSP (a, b), an LSP (c, d), an MSP (e, f), and an GP (g, h). Figures (a, b) correspond to the WSP shown in Fig. 20.6b, (c, d) to the LSP in Fig. 20.6c, (e, f) to the MSP in Fig. 20.6d, and (g, h) to the GP in Fig. 20.7a. Curves 1 and 2 are related to the left and right lanes, respectively. Taken from [332]
left lane. This is due to the synchronization of average vehicle speeds between lanes.

Within the “synchronized flow” and “wide moving jam” phases, i.e., in congested traffic, there is no complete separation of fast and slow vehicles between the left and right lanes, respectively. In synchronized flow, fast vehicles can change to the right lane and slow vehicles can change to the left lane (Fig. 20.9). This lane changing depends on the current difference between vehicle speeds in these lanes: fast and slow vehicles change to the lane where the speed is currently higher. This decreases the difference between the average vehicle speeds in the left and right lanes, i.e., this intensifies the speed synchronization effect.

The lower the average speed in synchronized flow, the higher the percentage of fast vehicles that move in the right lane. In particular, in SPs on average only about 5%-10% of fast vehicles move in the right lane and about 5%-10% of slow vehicles move in the left lane. In GPs where the average speed is lower than within SPs, we obtain that about 20% of fast vehicles move in the right lane and about 20%-30% of slow vehicles move in the left lane (Fig. 20.9).

For the chosen model parameters (Sect. 20.2.4) the limit flow rate in the pinch region of a GP under the strong congestion condition \( q_{\text{lim}}^{(\text{pinch})} \approx 1540 \text{ vehicles/h} \). This limit flow rate is close to the value \( q_{\text{lim}}^{(\text{pinch})} = 1500 \text{ vehicles/h} \) in flow of identical vehicles. However, the maximum flow rate in free flow at the on-ramp \( q_{\text{max, lim}}^{(\text{free B})} = 2250 \text{ vehicles/h} \) in this heterogeneous traffic flow is appreciably lower than the related value for flow of identical vehicles \( q_{\text{max, lim}}^{(\text{free B})} = 2400 \text{ vehicles/h} \). This decrease in \( q_{\text{max, lim}}^{(\text{free B})} \) is because the maximum speed of slow vehicles \( v_{\text{free}}^{(2)} \) is lower than the maximum speed of identical vehicles \( v_{\text{free}} \) and the time delay in vehicle acceleration for slow vehicles \( \tau_{\text{del}}^{(a, 2)} \) is greater than this time delay \( \tau_{\text{del}}^{(a)} \) for identical vehicles (Sect. 16.3.6).

Simulations show that the flow rate in the outflow from a wide moving jam in heterogeneous free flow \( q_{\text{out}} \approx 1725 \text{ vehicles/h} \) is less than the related flow rate \( q_{\text{out}} = 1810 \text{ vehicles/h} \) in flow with identical vehicles (Sect. 18.2).

### 20.3.5 Wide Moving Jam Propagation

Simulations show that in heterogeneous flow wide moving jam propagation exhibits some specific features discussed in Sect. 20.6.1. There are time intervals when the downstream jam front in the left lane moves with a more negative velocity than the velocity of the downstream jam front in the right lane (Fig. 20.10, \( t = 84 \text{ min} \)). This is because within a wide moving jam the fraction of fast vehicles in the left lane is greater than the one in the right lane. As a result, within these time intervals locations of the downstream jam fronts in different lanes do not coincide to one another at lower speeds (the distance between the fronts is denoted by \( \ell_{\text{down}}^{(J)} \) in Fig. 20.10,
Fig. 20.9. Percentages of fast (curves $f$) and slow vehicles (curves $sl$) in the left and right lanes. (a) WSP. (b) LSP. (c) MSP. (d) GP. Figures (a–c) are related to the SPs shown in Fig. 20.6b–d, respectively. One minute average data of virtual detectors. Figure (d) corresponds to the GP in Fig. 20.7a. Taken from [332]
However, within other time intervals differences between locations of the downstream jam fronts in the left and right lanes are not observed (Fig. 20.10, \( t = 83 \) and 85 min). This behavior can be explained by a competition of the lane changing effect and the effect of different time delays in acceleration of fast and slow vehicles within the downstream front of the wide moving jam. As a result, the downstream jam front moves on average with the same mean characteristic velocity in the left and right lanes. This mean velocity does not change over time. This velocity is also the same for different wide moving jams. In general, the mean velocity \( v_g \) of the downstream front of a wide moving jam depends on the relations between the percentages of fast vehicles \( \eta^{(1)} \), slow vehicles \( \eta^{(2)} \), and long vehicles \( \eta^{(3)} \) in heterogeneous flow (see also Sect. 20.6). For the chosen model parameters in Fig. 20.6, the velocity \( v_g = -14.8 \) km/h satisfies the condition \(|v_g| < |v_g^{(1)}|\).

---

**Fig. 20.10.** Wide moving jam propagation in heterogeneous flow with fast and slow vehicles. Space distributions of vehicle speed (a) and flow rate (b) for a wide moving jam that is the 3rd jam in the GP in Fig. 20.7a. Curves 1 and 2 are related to the left and right lanes, respectively. Taken from [332]

It should be noted that after vehicles have accelerated from a wide moving jam to free flow downstream of the jam, the separation of slow and fast vehicles occurs (Sect. 20.3.1). As a result, the vehicle speed in the left lane is higher than the speed in the right lane (Fig. 20.10). However, when synchronized flow is realized between wide moving jams, some of fast vehicles move in the right lane and some of slow vehicles move in the left lane in
this synchronized flow (see spatial speed distributions between moving jams shown in Fig. 20.8g).

Let us emphasize a qualitative difference between the propagation of the downstream front of a wide moving jam and the propagation of the downstream front of an MSP. It can be seen in Figs. 20.5a,b and 20.6d that the velocities of the downstream front of these MSPs are not some constant values. For two different MSPs in Figs. 20.5a,b and 20.6d the related mean velocities are different as well. Moreover, the mean velocities of the downstream fronts of the MSPs strongly depend on time when these MSPs propagate on the road. Thus, the mean velocity of the downstream front of an MSP is not a characteristic parameter. This velocity can be a function of initial conditions and time. Furthermore, this velocity can be different for different MSPs.

This behavior of MSPs is qualitatively different from wide moving jam propagation. In the latter case, the mean velocity of the downstream front of a wide moving jam is a characteristic parameter that does not depend on time and that is the same for different wide moving jams. This conclusion is true both for traffic flow with identical vehicles and for heterogeneous traffic flow.

Besides specific effects discussed above there are other specific effects in heterogeneous traffic flow with different driver behavioral characteristics. These effects are a partial destroying of speed synchronization between different lanes at higher speeds of synchronized flow and the occurrence of spatially extended regions of free flow recovering. The free flow recovering occurs when vehicles accelerate to free flow at the downstream front of congested patterns at the bottleneck. These two effects will be considered in the next section, where traffic patterns in heterogeneous traffic flow with different vehicle parameters will be discussed. This is because when heterogeneous traffic flow with very different vehicle lengths is considered these effects make a considerable quantitative influence on pattern parameters.

20.4 Patterns in Heterogeneous Traffic Flow with Different Vehicle Parameters

When heterogeneous traffic flow consists of fast vehicles and slower long vehicles, then an F→S transition and congested pattern features are qualitatively similar to those discussed above for heterogeneous traffic flow with different driver behavioral characteristics where all vehicles have the same length (Sect. 20.3). In particular, when traffic flow consists of 50% vehicles and 50% long vehicles there are qualitatively the same first-order F→S transitions leading to the onset of congestion in initial free flow, the same types of congested patterns, and the same pattern diagram at the on-ramp bottleneck (Figs. 20.11 and 20.12) as those found in traffic flow with identical vehicles (Sect. 18.2) and in heterogeneous flow with fast and slow vehicles.
Fig. 20.11. Diagram of congested patterns at an on-ramp bottleneck in heterogeneous flow with fast and long vehicles (a) and speed on the main road in space and time within SPs (b–d). (b) WSP. (c) LSP. (d) MSP. In (b–d) the flow rates \((q_{\text{on}}, q_{\text{in}})\) are: (b) \((150, 1740)\), (c) \((400, 1470)\), and (d) \((15, 1740)\) vehicles/h. \(\eta^{(1)} = 50\%\), \(\eta^{(3)} = 50\%\). \(q_{\text{max, lim}} \approx 1740\) vehicles/h. Taken from [332]
Fig. 20.12. GPs related to the diagram in Fig. 20.11a. (a) GP at $q_{in} > q_{out}$. (b) GP at $q_{in} < q_{out}$. (c) DGP. The flow rates $(q_{on}, q_{in})$ are: (a) $(1000, 1714)$, (b) $(1150, 1310)$, and (c) $(360, 1714)$ vehicles/h. Taken from [332]

(20.4.1) Peculiarity of Wide Moving Jam Propagation

There is, however, a difference in the propagation of a wide moving jam in the case under consideration. There are time intervals, when for lower vehicle speeds, the downstream jam front in the right lane is upstream of the downstream jam front in the left lane (Fig. 20.14, $t = 80$ min). To understand this jam propagation behavior, note that if traffic flow consists of long vehicles only, then the velocity of the downstream jam front is
Although $\tau^{(a, 1)}_{\text{del}}(0) < \tau^{(a, 3)}_{\text{del}}(0)$, nevertheless due to the condition $d^{(1)} < d^{(3)}$, we obtain

$$| v^{(1)}_g | < | v^{(3)}_g | .$$

Because the fraction of long vehicles in the right lane is greater than the fraction of fast vehicles, the velocity of the downstream jam front in the right lane is more negative than the one in the left lane. This explains the mentioned difference in downstream jam front propagation in the right and left lanes. However, there are other time intervals when the jam front is at the same location in both freeway lanes (Fig. 20.14, $t = 79, 81$ and $82$ min). This is due to a competition of the effect of different velocities (20.35) with
the lane changing effect. As a result, the downstream jam front moves on average at the same mean characteristic velocity in the left and right lanes. This mean velocity does not change over time. This velocity is also the same for different wide moving jams (see also Sect. 20.6.1). For the chosen model parameters in Fig. 20.11, the velocity $v_g = -21.9 \text{ km/h}$ satisfies the condition $|v_{g1}| < |v_g| < |v_{g3}|$.

In synchronized flow, where the average speed is lower than $v_{\text{free}}^{(3)}$ long vehicles can change to the left lane and fast vehicles can change to the right lane. Fast and long vehicles change to the lane where the speed is currently higher.
20.4 Patterns in Flow with Different Vehicle Parameters

20.4.2 Partial Destroying of Speed Synchronization

It has been found that in SPs of higher vehicle speeds, the average speed in the left lane is appreciably higher than the one in the right lane (Fig. 20.15a,c,e). Moreover, there are high amplitude oscillations of the speed over time in the left lane. There are also time intervals when the vehicle speed in the left lane becomes higher than the maximum vehicle speed of long vehicles. This partial destroying of speed synchronization occurs because at a higher speed conditions for lane changing of long vehicles from the right lane to the left lane are more difficult to satisfy.

To explain this, note that when in synchronized flow the vehicle speed in the left lane is higher than the speed in the right lane, long vehicles try to change to the left lane. However, there should be large space gaps in synchronized flow in the left lane for this lane changing. This is because of the large length of long vehicles. As a result, conditions for lane changing of long vehicles to the left lane are seldom satisfied. For this reason, fast vehicles can accelerate more frequently in the left lane. This leads to increase in average speed in the left lane. However, some long vehicles nevertheless can change to the left lane in synchronized flow. This decreases the average speed in the left lane and maintains the “synchronized flow” phase in both lanes. The lower the average speed in synchronized flow, the lower the difference between average speeds in the left and right lanes (Fig. 20.15a,c,e).

A destroying of speed synchronization in the left and right lanes can also occur during wide moving jam formation in an GP (Fig. 20.13, \( t = 87 \) and 88 min). In this case, there can be a random time delay in moving jam emergence in the left lane after the moving jam has emerged in the right lane.

20.4.3 Extension of Free Flow Recovering and Vehicle Separation

In traffic flow with identical vehicles, vehicles accelerate to free flow at the downstream front of synchronized flow that is fixed at a bottleneck. In heterogeneous traffic flow, this free flow recovering is accompanied by the vehicle separation effect in free flow (Sect. 20.3.1).

In addition, in heterogeneous flows spatially extended regions of free flow recovering appear. Downstream of a congested bottleneck spatially extended regions are formed where the average speed is lower than the one in free flow. These regions of free flow recovering propagate downstream of the congested bottleneck (Figs. 20.6b, 20.11b, c, and 20.12). The occurrence of spatially extended regions of free flow recovering can be one of the possible explanations of empirical results, where downstream of a congested bottleneck, spatially extended regions of lower average speed have been found [59]. In particular, in heterogeneous flow with 50% fast vehicles and 50% slow long vehicles these spatially extended regions of free flow recovering are relatively large (Fig. 20.12).
Fig. 20.15. Space distributions of vehicle speed (a, c, e, g) and flow rate (b, d, f, h) for the WSP in Fig. 20.11 (a, b), the LSP in Fig. 20.11c (c, d), the MSP in Fig. 20.11d (e, f), and the GP in Fig. 20.12a (g, h). Curves 1 and 2 are related to the left and right lanes, respectively. Taken from [332]
The occurrence of spatially extended regions of free flow recovering can be explained by a blocking effect (Fig. 20.13). A long vehicle moving in the left lane at its maximum speed $v_{\text{free}}(3, \text{left})$ forces all upstream moving fast vehicles to move with this speed. Only after the long vehicle has changed to the right lane fast vehicles can accelerate to their maximum speed $v_{\text{free}}^{(1)}$.

20.5 Weak Heterogeneous Flow

When the percentage of slow vehicles and/or long vehicles gradually decreases, special features of free and congested traffic in heterogeneous flow discussed above gradually disappear. None of these special effects in heterogeneous flows are a threshold effect. At a low enough percentage of slow vehicles and/or long vehicles features of free and congested traffic are the same as those in traffic flow with identical vehicles. An illustration of this conclusion is made in Fig. 20.16a,b where WSP emergence at the bottleneck is shown.

20.5.1 Spontaneous Onset of Congestion Away from Bottlenecks

However, when the percentage of slow vehicles and/or long vehicles is low some additional effect in such a weak heterogeneous traffic flow has been found. This effect is the occurrence of large amplitude fluctuations in free flow. These large fluctuations can lead to a random emergence of MSPs away from bottlenecks when the density in free flow is high enough. In Fig. 20.16c, an WSP occurs spontaneously at a bottleneck. However, due to large amplitude fluctuations in free flow another MSP emerges spontaneously on the main road away from the bottleneck (Fig. 20.16c). This MSP propagates through free flow on the main road regardless of the WSP at the bottleneck. It turns out that the amplitude of fluctuations in free flow away from bottlenecks is of the same order of magnitude as those at the bottleneck. This explains MSP emergence away from the bottleneck.

At a higher flow rate $q_{\text{in}}$ and a lower flow rate $q_{\text{on}}$ MSPs can appear spontaneously independently of one another: an MSP can occur away from the on-ramp bottleneck and another MSP appears on the main road at the on-ramp bottleneck almost simultaneously.

To understand the physics of this effect, we consider heterogeneous flow with fast and long vehicles. When the percentage of long vehicles is high enough there is separation of fast and long vehicles in free flow (Sect. 20.3.1). When the percentage of long vehicles decreases, the average space gap between long vehicles in the right lane increases. If the space gap between long vehicles is high enough, then some of fast vehicles change to the right lane, i.e., there is no complete vehicle separation in free flow. A fast vehicle moves
Emergence of SPs in weak heterogeneous traffic flow at an on-ramp bottleneck. (a, c) Speed distributions in time and space in the left and right lanes of the main road. (b) One minute average data of virtual detectors for an WSP shown in (a). In (c) the arrow shows an MSP that has emerged spontaneously away from the bottleneck. The flow rates \((q_{on}, q_{in})\) are: (a) \((300, 2235)\), (b) \((300, 2310)\) vehicles/h. \(\eta^{(5)} = 3\%\), \(\eta^{(1)} = 97\%\). Taken from [332]
in the right lane with its maximum speed before the vehicle approaches the long preceding vehicle. After the fast vehicle has reached this long preceding vehicle, the fast vehicle changes to the left lane to overtake the long vehicle. This lane changing of fast vehicles causes large amplitude fluctuations in free flow. These fluctuations lead to spontaneous MSP emergence when density in free flow is high enough. The same effects occur also in heterogeneous traffic flow with fast and slow vehicles of the same vehicle length at a low percentage of slow vehicles. These results can be one of the explanations of empirical observations of an F→S transition away from bottlenecks (Sect. 10.6).

20.5.2 Lane Asymmetric Free Flow Distributions

In addition to large amplitude fluctuations, at a lower percentage of slow and/or long vehicles another effect in free flow is also realized, which is well-known in both empirical observations and traffic flow modeling (see references in [357, 432]) (Fig. 20.17).

![Figure 20.17](image)

Fig. 20.17. Vehicle speed (a) and flow rate (b) in the left lane (curves 1) and in the right lane (curves 2) in free flow as functions of the incoming flow rate $q_{in}$ on a homogeneous road. $\eta^{(3)} = 3\%$, $\eta^{(1)} = 97\%$. Taken from [332]

At a very low flow rate $q_{in}$ the flow rate in the right lane (curve 1 in Fig. 20.17b) exceeds the flow rate in the left lane (curve 2). If the flow rate $q_{in}$ increases, then the flow rate in the left lane becomes greater than the flow rate in the right lane. In weak heterogeneous flow with fast and long vehicles ($\eta^{(1)} \gg \eta^{(3)}$), this effect is related to a decrease in the mean distance between long vehicles in the right lane when the total flow rate increases. As a result, fast vehicles should change more frequently to the left lane to facilitate passing. Therefore, beginning at some density in the right lane the flow rate of fast vehicles that move in the left lane increases in a greater extent than the increase in total flow rate on the road. This can be one of the possible explanations of the related well-known empirical result shown in Fig. 2.3f.

In the vicinity of the critical point for an F→S transition, the flow rates in both lanes approach one another. At this critical point the average speed in
the lanes is approximately equal to the maximum speed $v_{\text{free}}^{(3)}$ of long vehicles in free flow (Fig. 20.17).

## 20.6 Characteristics of Congested Pattern Propagation in Heterogeneous Traffic Flow

In this section, we derive some general characteristics of wide moving jams and MSP propagation in heterogeneous flow that consists of vehicles of $K$ different types moving on a multilane (one-way) road with $M$ lanes. The percentage of vehicles of type $i$ is $\eta^{(i)}$, where $i = 1, 2, \ldots, K$ and $\sum_{i=1}^{K} \eta^{(i)} = 100\%$.

### 20.6.1 Velocity of Downstream Jam Front

Corresponding to (3.10), in traffic flow with vehicles of only one type $i$, the velocity of the downstream front of a wide moving jam is

$$v_g^{(i)} = -\frac{d^{(i)}}{\tau_{\text{del}}^{(a, i)}(0)} , \quad (20.36)$$

where $d^{(i)}$ is the vehicle length, $\tau_{\text{del}}^{(a, i)}(0)$ is the mean delay time in vehicle acceleration at the downstream front of a wide moving jam. This delay time is related to the vehicle speed $v = 0$ within the jam. The velocity $v_g^{(i)}$ determines the slope of the line $J$ in the flow-density plane. For example, the line $J$ is related to the line $J^{(1)}$ in Fig. 20.1a for fast vehicles, to the line $J^{(2)}$ in Fig. 20.1b for slow vehicles, and to the line $J^{(3)}$ in Fig. 20.1c for long vehicles.

In heterogeneous traffic flow, during some time intervals the downstream jam front can move with more negative velocity in one of the freeway lanes than in the other lanes. This is because vehicles of different types can be distributed non-uniformly between the lanes. To explain this, we consider an example of flow where there are vehicles of two types, fast and slow vehicles. We assume that the percentage of fast vehicles is greater in the left lane than in the right lane. Because of the condition (20.13), i.e.,

$$\tau_{\text{del}}^{(a, 1)}(0) < \tau_{\text{del}}^{(a, 2)}(0) , \quad (20.37)$$

and corresponding to (20.36), the velocity of the downstream jam front should be more negative in the left lane than in the right lane. However, this effect is compensated on average by lane changing of vehicles to the lane where the velocity of the downstream jam front is more negative. In the example of fast and slow vehicles, vehicles at the downstream jam front in the right lane change to the left lane when the current location of the downstream jam front in the left lane is upstream of the front location in the right lane. As
a result of this lane changing effect, the downstream front of a wide moving jam moves on the road on average with the same mean velocity $v_g$. This conclusion is confirmed by numerical simulations discussed above.

In order to find this mean velocity $v_g$, we assume that during a time interval $T_f$ the downstream jam front propagates a distance $X_f$. The average velocity of the front is

$$\tilde{v}_g = -\frac{X_f}{T_f}. \tag{20.38}$$

During the time interval $T_f$ there are $N^{(i)}$ vehicles of each type $i$ ($i = 1, 2, \ldots, K$) that have accelerated from a standstill within the jam at the downstream jam front. The total number of these vehicles is

$$N = \sum_{i=1}^{K} N^{(i)}. \tag{20.39}$$

The displacement of the front $X_f$ can be expressed in terms of distances $d^{(i)}$ between vehicle fronts within the jam:

$$X_f = M^{-1} \sum_{i=1}^{K} N^{(i)} d^{(i)}. \tag{20.40}$$

The time interval $T_f$ is

$$T_f = M^{-1} \sum_{i=1}^{K} N^{(i)} \tau_{\text{del}}^{(a, i)}(0). \tag{20.41}$$

Taking into account (20.40) and (20.41), $\tilde{v}_g$ (20.38) can be written as follows

$$\tilde{v}_g = -\frac{\sum_{i=1}^{K} \tilde{\eta}^{(i)} d^{(i)}}{\sum_{i=1}^{K} \tilde{\eta}^{(i)} \tau_{\text{del}}^{(a, i)}(0)}, \tag{20.42}$$

where $\tilde{\eta}^{(i)} = (N^{(i)}/N)100\%$. If the time interval $T_f$ increases, values $\tilde{\eta}^{(i)}$, $i = 1, \ldots, K$ (20.42) tend to the percentages $\eta^{(i)}$, $i = 1, \ldots, K$, respectively, and the velocity $\tilde{v}_g$ (20.42) to the mean velocity $v_g$:

$$v_g = -\frac{\bar{d}}{\bar{\tau}_{\text{del}}^{(a)}(0)}, \tag{20.43}$$

where $\bar{d}$ and $\bar{\tau}_{\text{del}}^{(a)}(0)$ are the mean vehicle length and the mean time delay averaged over all types of vehicles, respectively:

$$\bar{d} = \sum_{i=1}^{K} \eta^{(i)} d^{(i)}/100, \tag{20.44}$$

$$\bar{\tau}_{\text{del}}^{(a)}(0) = \sum_{i=1}^{K} \eta^{(i)} \tau_{\text{del}}^{(a, i)}(0)/100. \tag{20.45}$$
For parameters of heterogeneous flow associated with Fig. 20.7, formula (20.43) yields the velocity $v_g = -14.4 \text{ km/h}$, whereas the mean velocity $v_g$ found in numerical simulations is $-14.8 \text{ km/h}$. In heterogeneous flow consisting of fast and long vehicles (Fig. 20.12), the corresponding values are $v_g = -21.2 \text{ km/h}$ (20.43) and $v_g = -21.9 \text{ km/h}$. The difference between analytical and numerical values of the velocity $v_g$ can be explained by blanks that occur within wide moving jams in numerical simulations (Figs. 20.7 and 20.12). As a result, the average distance between vehicle fronts within the wide moving jams in numerical simulations is slightly greater than the analytical value $\bar{d}$ (20.44) used in (20.43). This effect is more essential in heterogeneous flow of fast and long vehicles in which the blanks are usually larger. Recall that blanks often appear within wide moving jams in empirical observations (Chap. 11).

### 20.6.2 Flow Rate in Jam Outflow

The mean velocity $v_g$ of the downstream front of a wide moving can be written in the well-known form

$$v_g = -\frac{q_{\text{out}}^{(i)}}{\rho_{\text{max}}^{(i)} - \rho_{\text{min}}^{(i)}} , \quad (20.46)$$

where $q_{\text{out}}^{(i)}$ is the flow rate of vehicles of type $i$ in the outflow from a wide moving jam when free flow is formed in the jam outflow,

$$\rho_{\text{max}}^{(i)} = \frac{\eta^{(i)}}{100\bar{d}} , \quad (20.47)$$

$$\rho_{\text{min}}^{(i)} = \frac{q_{\text{out}}^{(i)}}{v_{\text{max}}^{(i)}} , \quad (20.48)$$

$v_{\text{max}}^{(i)}$ is the average speed of vehicles of type $i$ in heterogeneous free flow in the jam outflow, $v_{\text{max}}^{(i)} \leq v_{\text{free}}^{(i)}$, $v_{\text{free}}^{(i)}$ is the speed in free flow with vehicles of type $i$ only. From (20.43) and (20.46)–(20.48), we obtain

$$q_{\text{out}}^{(i)} = \frac{\eta^{(i)}}{100\left(\bar{r}_{\text{del}}^{(a)}(0) + \bar{d}/v_{\text{max}}^{(i)}\right)} . \quad (20.49)$$

This formula can also be written as

$$q_{\text{out}}^{(i)} = \frac{1}{\bar{r}_{\text{del}}^{(a)}(0)} \left(\frac{\eta^{(i)}}{100} - \frac{\rho_{\text{min}}^{(i)}}{\rho_{\text{max}}^{(i)}}\right) . \quad (20.50)$$

The flow rate in the jam outflow $q_{\text{out}}$ is the sum of the flow rates $q_{\text{out}}^{(i)}$, $i = 1, \ldots, K$ for all $K$ types of vehicles. Thus, from (20.49) and (20.50) we obtain

$$q_{\text{out}} = \sum_{i=1}^{K} \frac{\eta^{(i)}}{100\left(\bar{r}_{\text{del}}^{(a)}(0) + \bar{d}/v_{\text{max}}^{(i)}\right)} \quad (20.51)$$
and

\[ q_{\text{out}} = \frac{1}{\tau_{\text{del}}^{(a)}(0)} \left( 1 - \frac{\rho_{\text{min}}}{\rho_{\text{max}}} \right) \], \quad (20.52)

respectively.

In heterogeneous flow, the flow rate \( q \) and density \( \rho \) on the line \( J \) are related to the equation

\[ q = \frac{1}{\tau_{\text{del}}^{(a)}(0)} \left( 1 - \frac{\rho}{\rho_{\text{max}}} \right) \]. \quad (20.53)

In (20.51)–(20.53), the density \( \rho_{\text{min}} \) is

\[ \rho_{\text{min}} = \sum_{i=1}^{K} \rho_{\text{min}}^{(i)} \], \quad (20.54)

\( \rho_{\text{max}} = 1/\bar{d} \) is the average density within a wide moving jam, \( \tau_{\text{del}}^{(a)}(0) \) is given by (20.45).

### 20.6.3 Velocity of Downstream Front of Moving Synchronized Flow Patterns

Let us compare mean velocities of the downstream fronts of a wide moving jam and an MSP in heterogeneous flow. Whereas the mean velocity \( v_g \) (20.43) of the downstream front of a wide moving jam is a characteristic parameter that does not depend on initial conditions, the mean velocity of the downstream front of the MSP is not a characteristic parameter.

To show this, we use (20.43) to obtain the mean velocity \( v_{\text{down}} \) of the downstream MSP front in heterogeneous flow. We assume that all vehicles within an MSP move with the same vehicle speed \( v^{(\text{syn})} \). In a spatial coordinate system moving with the velocity \( v^{(\text{syn})} \), vehicles within the MSP do not move. The average distance between these vehicles is \( \Delta x = \bar{g} + \bar{d} \), where \( \bar{g} \) is the average space gap between vehicles within the MSP. The mean velocity \( v_{\text{down}} \) can be found if in (20.43) the mean vehicle length \( \bar{d} \) and the mean time delay \( \tau_{\text{del}}^{(a)}(0) \) are replaced by \( \Delta x \) and \( \tau_{\text{del}, \text{syn}}^{(a)} \), respectively. Then in the motionless coordinate system

\[ v_{\text{down}} = v^{(\text{syn})} - \frac{\bar{g} + \bar{d}}{\tau_{\text{del}, \text{syn}}^{(a)}} \], \quad (20.55)

where \( \tau_{\text{del}, \text{syn}}^{(a)} \) is the average time delay in vehicle acceleration associated with vehicles that accelerate from synchronized flow within the MSP to free flow downstream of the MSP; \( \tau_{\text{del}, \text{syn}}^{(a)} \) is a function of speed \( v^{(\text{syn})} \). Note that (20.55) is a generalization of (7.6) for the case of heterogeneous flow.
Values \( v^{(\text{syn})}, \overline{T}^{(a)}_{\text{del, syn}}, \) and \( \overline{y} \) in (20.55) can change within an MSP over time and they can also be different for different MSPs. Thus, in accordance with numerical results discussed above (Sect. 20.3.5) the mean velocity \( v_{\text{down}} \) (20.55) is not a characteristic parameter.

### 20.7 Conclusions

(i) Different driver behavioral characteristics and different vehicle parameters lead to the well-known lane specific behavior in free traffic flow: fast vehicles use mostly the left (passing) lane whereas slow and long vehicles use mostly the right freeway lane. This is the vehicle lane separation effect in free flow. As a result, the average speed in free flow in the left lane is higher than the average speed in the right lane. The vehicle lane separation effect is realized when the percentage of slow and/or long vehicles is compatible with the percentage of fast vehicles. When the percentage of slow and/or long vehicles is low enough, fast vehicles also move in the right lane. In this case, the vehicle lane separation effect does not occur when the flow rate is high enough.

(ii) In weak heterogeneous flow in which the percentage of slow and/or long vehicles is low, large amplitude fluctuations can appear in free flow associated with passing of fast vehicles when they approach slow and/or long vehicles in the right lane. At a high enough flow rate in free flow these large amplitude fluctuations can be comparable with fluctuations at freeway bottlenecks. This can cause an \( F \rightarrow S \) transition away from bottlenecks.

(iii) There are some spatiotemporal traffic pattern features that are qualitatively the same in traffic flow with identical vehicles (Chaps. 17 and 18) and in heterogeneous traffic flow. These fundamental pattern features are:

(a) \( F \rightarrow S \) transitions and \( S \rightarrow J \) transitions in traffic flow.

(b) Types of congested patterns at a freeway bottleneck (WSP, MSP, LSP, GP, DGP).

(c) The diagram of congested patterns at the bottleneck at different traffic demand.

(iv) There are also some specific features of congested patterns at freeway bottlenecks in heterogeneous traffic flow:

(a) The critical amplitude of a local perturbation required for an \( F \rightarrow S \) transition in the left (passing) lane can be greater than the critical amplitude in the right lane. As a result, there can be a random time delay in MSP occurrence in the left lane after the MSP has appeared in the right lane.

(b) There can be a random time delay in moving jam emergence in the left lane after the moving jam has emerged in the right lane in the
pinch region of an GP. During jam propagation in the right lane the moving jam appears also in the left lane.

(c) If a wide moving jam propagates on a road, there can be some time intervals when locations of the downstream jam fronts in the left and right lanes are different to one another. However, there is also an opposite effect of synchronization of these locations. As a result, the downstream jam front propagates on average with the same mean velocity in both lanes. At chosen parameters of heterogeneous flow this characteristic velocity is the same for different wide moving jams.

(d) There can be a partial destroying of the speed synchronization between different lanes at higher speeds of synchronized flow.

(e) Downstream of a congested bottleneck spatially extended regions of free flow recovering can appear when vehicles accelerate to free flow at the downstream front of a congested pattern at the bottleneck.

(v) All specific pattern properties of item (iv) gradually disappear when differences between driver behavioral characteristics and vehicle parameters in traffic flow decrease. This gradual transformation from heterogeneous traffic flow to traffic flow with identical vehicles does not qualitatively change the fundamental traffic pattern features of item (iii). This means that these specific pattern properties associated with different driver behavioral characteristics and vehicle parameters are secondary effects in comparison with the aforementioned fundamental traffic pattern features.
Part IV

Engineering Applications
21 ASDA and FOTO Models of Spatiotemporal Pattern Dynamics based on Local Traffic Flow Measurements

21.1 Introduction

One of the approaches and trials in traffic technology to estimate, track, and predict traffic patterns is an application of either microscopic (in particular, car-following models), mesoscopic, or macroscopic traffic flow models that calculate spatiotemporal distributions of vehicle speed and density in freeway networks (see e.g., [20, 25, 27, 31, 279, 291, 303, 316, 485–488, 490–492, 520, 521], reviews of traffic flow models [33, 35, 36, 38]). Traffic data measured at some locations on a freeway can be used as a kind of “boundary conditions” for these mathematical traffic flow models. Unfortunately, practical online applications of this approach have some basic problems. One of the problems is the need to validate certain model parameters. These parameters are dependent on the infrastructure, weather, and other environmental conditions. It is hard to find and adapt a set of model parameters that are valid for real traffic flow under all possible totally different conditions. However, this is not the main problem for the online application of these model approaches.

The problem is that for the same traffic data measured at freeway locations (the same “boundary conditions” for these models) there can be various spatiotemporal congested patterns. This is associated with the probabilistic nature of congested traffic patterns (Sect. 2.4.10). This can be one of the principal reasons why the mentioned model approaches are still seldom used for practical applications on freeways. This is also confirmed by simulations of models based on three-phase traffic theory that can show and predict empirical spatiotemporal features of congested patterns at bottlenecks. As discussed in Sect. 18.5, congested patterns can be metastable against short-time local perturbations in traffic. For this reason, depending on initial conditions and fluctuations in traffic flow different types of congested patterns can occur at a bottleneck at the same traffic demand and traffic conditions.

Therefore, for reconstruction, tracking, and prediction of the congested pattern dynamics, rather than a calculation of complicated spatiotemporal distributions of traffic flow variables (like vehicle speed and/or density), simpler models should be developed and used. These models should perform a direct spatiotemporal reconstruction, tracking, and prediction of the traffic phases based on local traffic measurements. The FOTO and ASDA models introduced by the author in 1996–1998 exhibit these features.
Results of the application of the FOTO and ASDA models on different freeways are in accordance with empirical spatiotemporal congested pattern features [226, 228, 230–232].

In this chapter, two models, FOTO (Forecasting of Traffic Objects) and ASDA (Automatische Staudynamikanalyse: Automatic Tracking of Moving Jams), will be considered [226, 228–235, 237–239]. The ASDA and FOTO models are devoted to the automatic recognition and tracking of congested spatiotemporal traffic patterns on freeways. The models are based on a spatiotemporal traffic phase classification made in three-phase traffic theory. The FOTO model identifies traffic phases and tracks synchronized flow. The ASDA model is devoted to the tracking of moving jam propagation. The general approach and the different extensions of the FOTO and ASDA models will be explained. It should be noted that the FOTO and ASDA models perform without any validation of model parameters in different environmental and traffic conditions.

21.2 Identification of Traffic Phases

The main features of the FOTO and ASDA models are as follows. In the FOTO model based on local measurements of traffic (detectors D1 and D2 in Fig. 21.1) the local identification of the traffic phases is performed first. Secondly, the FOTO and ASDA models perform the recognition of moving fronts of wide moving jams and of fronts of synchronized flow at detectors. These fronts define the spatial size and location of the related “synchronized flow” object and “wide moving jam” object. Finally, the FOTO and ASDA models track these object fronts in time and space.

This tracking is related to the online description of spatiotemporal features of the traffic objects. In other words, in the FOTO and ASDA models, after the recognition of the “synchronized flow” and “wide moving jam” objects at detector locations these objects are tracked as macroscopic single objects, i.e., only the determination of object attributes (locations of object fronts, object widths (in the longitudinal direction), velocities of the object

1 To explain the term “traffic object,” note that traffic states in the “synchronized flow” and “wide moving jam” traffic objects are associated with the “synchronized flow” and “wide moving jam” traffic phases, respectively. Furthermore, each of the traffic objects has the following attributes:

(i) object width (in the longitudinal direction);
(ii) locations of the downstream and upstream object fronts;
(iii) velocities of the downstream and upstream object fronts;
(iv) object lifetime.

The FOTO and ASDA models reconstruct and predict these object attributes as functions of time. Additional attributes of the traffic objects are the average vehicle speed, average density, and average flow rate within the objects.
21.2 Identification of Traffic Phases

Traffic phases are identified using the FOTO and ASDA models. These phases are:

2. Tracking of traffic phases

Fig. 21.1. Illustration of the FOTO and ASDA model approach. Taken from [235]

Recall that a wide moving jam is a localized upstream moving traffic pattern separated upstream and downstream by the associated fronts \( x_{\text{up}}^{\text{jam}} \), \( x_{\text{down}}^{\text{jam}} \) (Fig. 21.1), where the vehicle speed, flow rate, and density spatially change abruptly. Within the jam each vehicle is in a stop at least for a finite time that is large enough that there is no influence of traffic flow upstream of the jam on traffic flow downstream of the jam and vice versa. Synchronized flow can also be considered a localized structure separated upstream and downstream by the associated fronts \( x_{\text{up}}^{\text{syn}} \), \( x_{\text{down}}^{\text{syn}} \) (Fig. 21.1), where the vehicle speed changes abruptly. However, in contrast to a wide moving jam, in synchronized flow the flow rate can remain almost constant within these fronts.

Figure 21.1 and Fig. 21.2 illustrate the approach: due to local measurements of traffic the recognition of the traffic phases is performed first. Secondly, the initial fronts of wide moving jams \( x_{\text{up}}^{\text{jam}} \), \( x_{\text{down}}^{\text{jam}} \) and of synchronized flow \( x_{\text{up}}^{\text{syn}} \), \( x_{\text{down}}^{\text{syn}} \) are determined. These fronts define spatial sizes
and locations of the associated “synchronized flow” and “wide moving jam” objects in congested traffic. Finally, based on the ASDA and FOTO models the tracking and forecasting of these objects fronts in time and space is calculated, i.e., positions of all fronts \( x_{up}^{(jam)} \), \( x_{down}^{(jam)} \), \( x_{up}^{(syn)} \), and \( x_{down}^{(syn)} \) as functions of time are found. Note that for traffic forecasting, historical time series at least for flow rates are necessary. After the recognition of the mentioned traffic objects in congested traffic, these objects are tracked and predicted as macroscopic single objects. In particular, the determination of locations and widths of the related objects at any time is performed.

Using the average vehicle speed and density within the traffic objects (that are known from empirical studies and current traffic measurements) one can calculate other traffic characteristics, e.g., trip travel times and/or vehicle trajectories.

### 21.3 Determination of Traffic Phases with FOTO Model

The recognition of congested patterns with the FOTO model will be limited here to traffic data measured at stationary detectors on freeways. One of the simplest ways is the classification of the traffic phases based on a comparison of measured flow rates and vehicle speeds in different traffic states. In this section, we discuss results and limitations\(^2\) of this traffic phase classification.

\(^2\) Recall that for an accurate classification of the “synchronized flow” and “wide moving jam” phases in congested traffic, the objective criteria for these traffic phases should be used (Sect. 4.2.1). However, to apply these objective criteria, measurements of traffic variables in space and time are necessary. Thus, when distances between detectors are too large, only some approximate classification of the traffic phases is possible. An evaluation of an approximate online classification of the traffic phases performed by the FOTO model is considered in Sect. 22.2 and Sect. B.2.
For practical application, it is convenient to perform the classification of the traffic phases through the use of a fuzzy inference system.

### 21.3.1 Fuzzy Rules for FOTO Model

Current traffic measurements for each detector \( q(t) \) and \( v(t) \) are considered in a set of fuzzy rules. The fact that the flow rate in synchronized flow is usually much higher than the flow rate within wide moving jams is taken into account, as are other empirical features of the “synchronized flow” and “wide moving jam” phases. With the current flow rate classified by fuzzy logic into the values “low” and “high” and the vehicle speed classified by fuzzy logic into the values “low,” “medium,” and “high,” the following fuzzy rules are implemented (Fig. 21.3):

**Rule 1:** If the vehicle speed is “high,” the traffic phase is “free flow.”

**Rule 2:** If the vehicle speed is “medium,” the traffic phase is “synchronized flow.”

**Rule 3:** If the vehicle speed is “low” and the flow rate is “high,” the traffic phase is “synchronized flow.”

![Diagram](image_url)

*Fig. 21.3. Illustration of the FOTO model. Fuzzification of input variables with the basic rules. (a) Vehicle speed. (b) Flow rate. The degree of membership is determined for each fuzzy set by the membership function (see curves “low,” “medium,” and “high” in (a) and “low” and “high” in (b)). Dashed lines show the degree of membership for an example with \( v = 28 \text{ km/h} \), \( q = 980 \text{ vehicles/h} \). Taken from [235]*
Rule 4: If both the vehicle speed and the flow rate are “low,” the traffic phase is “wide moving jam.”

Numerical values of the membership functions are found based on empirical data measured on the freeway A5 in Germany (Figs. 2.1 and 2.2). It turns out that the same values can be used for all detectors on the freeway A5 near Frankfurt and for the rest of the freeways of the German Federal State of Hessen, as well as for freeways near Magdeburg and Munich. It can be expected that these numerical values can be influenced by traffic regulations (overtaking laws, variable speed limits, etc.) and peculiarities of the infrastructure (roadworks, narrow lanes, road gradient, etc.). Therefore, the numerical values have to be adapted for example for US freeways or inner-urban expressways with fixed speed limits. In addition, an accident or a sudden road blockage can lead to the necessity of adapted membership functions because the infrastructure changes.

Note that the traffic phases recognized by the fuzzy rules are not numerical values. These phases are also not scalable among each other. For example, the “free flow” phase cannot be weighted numerically in comparison with the “wide moving jam” phase. Therefore, the defuzzification is not performed with the standard COG (center of gravity) or MOM (mean of maxima) methods. Instead, the traffic state with the highest degree of membership for traffic states is used (rule1–rule4) after processing all fuzzy rules. Let us demonstrate the use of these fuzzy sets for one example shown in Fig. 21.4 [238].

Figure 21.4 and Table 21.1 show results of a decision about the degree of membership for the measured values of vehicle speed and flow rate, the degree of membership for the four fuzzy rules, and the classified traffic state at detectors D5 for times between 8:45 and 9:10 am. The values for flow rates \( q \) (in vehicles/h per lane) and vehicle speed \( v \) (in km/h) are fuzzified according to the fuzzy membership functions in Fig. 21.3. The membership values for flow rates (“low” and “high”) and speeds (“low,” “medium,” and “high”) are used for each minute in the four fuzzy rules. The resulting traffic phase with

<table>
<thead>
<tr>
<th>Time</th>
<th>q</th>
<th>v</th>
<th>Flow</th>
<th>Flow</th>
<th>Speed</th>
<th>Speed</th>
<th>Speed</th>
<th>Rule 1</th>
<th>Rule 2</th>
<th>Rule 3</th>
<th>Rule 4</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:45</td>
<td>1260</td>
<td>30</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>syn</td>
</tr>
<tr>
<td>08:46</td>
<td>1100</td>
<td>24</td>
<td>0.125</td>
<td>0.875</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
<td>0.2</td>
<td>0.8</td>
<td>0.125</td>
<td>0</td>
<td>syn</td>
</tr>
<tr>
<td>...</td>
<td>09:03</td>
<td>220</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>jam</td>
</tr>
<tr>
<td>09:04</td>
<td>980</td>
<td>28</td>
<td>0.275</td>
<td>0.725</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
<td>0.275</td>
<td>0</td>
<td>syn</td>
</tr>
<tr>
<td>...</td>
<td>09:09</td>
<td>1260</td>
<td>64</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>09:10</td>
<td>1460</td>
<td>72</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>free</td>
</tr>
</tbody>
</table>
Fig. 21.4. Traffic phase determination with the FOTO model. (a) Part of a section of the freeway A5-South. (b) Average speed (left) and flow rate (right) at the detectors D5, D4, and D2 on March 11, 2002 (left axis in each diagram), together with the corresponding decision of the FOTO model on the current traffic phase (right axis in each diagram). Taken from [238]
the maximum degree of membership of the four fuzzy rules (rule1–rule4) can be found in the rightmost column of Table 21.1 (syn: synchronized flow, jam: wide moving jam, free: free traffic flow). The evaluation of this application will be made in Sect. 22.2. Note that low speeds are measured at detectors D5 at $t = 8:46$ and $t = 9:04$. Despite these jam-like low speeds, the traffic state is classified as synchronized flow because the flow rate is still high. However, in some cases the distinction between synchronized flow and wide moving jams based on the four fuzzy rules mentioned above is not accurate. To cope with such special measurement constellations, an extended set of fuzzy rules has been developed. The approach and an empirical example for the utility of the extended rules can be found in Appendix B.

21.4 Tracking Moving Jams with ASDA: Simplified Discussion

After a wide moving jam has first been detected by the FOTO model (see time $t_0$ below), the ASDA model is used to track and predict the wide moving jam at all times between detectors even when the jam cannot be measured. In an application of the model, the following results were obtained:

(i) The movement of a wide moving jam can be tracked and predicted at any time, even if the moving jam is between detectors on a road section.
(ii) Vehicle trip times can be predicted according to location and the jam width (in the longitudinal direction).

Further features of the ASDA model include aspects of freeway infrastructure such as lane-changing or the integration of on- and off-ramps (Appendix B).

For an overview of the ASDA model, consider Fig. 21.5, where the dynamic movement of a wide moving jam is sketched on a freeway section at different times.

The sketch for the ASDA model (Fig. 21.5) shows a measurement infrastructure with two detectors ($Q_0$, $Q_n$) on a road section. After a moving jam has been observed in cyclic measured data at $Q_n$ at time $t_0$ by the FOTO model, the ASDA model starts to calculate continuously the positions of the upstream jam front, $x_{up}^{(jam)}$. After the downstream jam front is registered at the detector $Q_n$ at a later time $t_1$ ($t_1 > t_0$), the ASDA model starts to calculate continuously the positions of the downstream jam front $x_{down}^{(jam)}$ and consequently the jam width $L_J$. This can be done cyclically even if none of the detectors is within or close to the moving jam. For this calculation, the flow rates $q_0$ and $q_n$ and the speeds $v_0$ and $v_n$ measured at the detectors $Q_0$ and $Q_n$ are required.

The movement and the tracking of a moving jam can be determined from the equation for the position of the upstream jam front (see Fig. 21.5)
21.4 Tracking Moving Jams with ASDA

Fig. 21.5. Illustration of the ASDA model. Moving jam recognition with two detectors

\[ x_{\text{up}}^{(\text{jam})}(t) = \int_{t_0}^{t} v_{\text{up}}^{(\text{jam})}(t) \, dt \approx -\int_{t_0}^{t} \frac{q_{0}(t) - q_{\text{min}}^{(\text{jam})}(t)}{\rho_{\text{max}}^{(\text{jam})}(t) - (q_{0}(t)/v_{0}(t))} \, dt, \quad t \geq t_0, \]  
\[ x_{\text{down}}^{(\text{jam})}(t) = \int_{t_1}^{t} v_{\text{down}}^{(\text{jam})}(t) \, dt \approx -\int_{t_1}^{t} \frac{q_{\text{out}}^{(\text{jam})}(t) - q_{\text{min}}^{(\text{jam})}(t)}{\rho_{\text{max}}^{(\text{jam})}(t) - \left( q_{\text{out}}^{(\text{jam})}(t)/v_{\text{max}}^{(\text{jam})}(t) \right) } \, dt, \quad t \geq t_1, \]  
\[ L_J(t) = x_{\text{down}}^{(\text{jam})}(t) - x_{\text{up}}^{(\text{jam})}(t), \quad t \geq t_1. \]

In (21.1)–(21.3), \( t_0 \) is the time when the moving jam is detected at the downstream detector \( Q_n \) (Fig. 21.5); the time \( t_1 \) determines the appearance of the downstream jam front at the detector \( Q_n \), i.e., the first time the local traffic phase is no longer classified as a wide moving jam by the FOTO model; \( v_{\text{down}}^{(\text{jam})}(t) \) and \( v_{\text{up}}^{(\text{jam})}(t) \) are the velocities of the downstream and upstream fronts of the jam, respectively, which are calculated based on the Stokes shock-wave formula (3.5); \( q_{0}(t) \) and \( v_{0}(t) \) are the flow rate and average vehicle speed measured at the upstream border of the section (Fig. 21.5, \( x = -L \)); \( q_{\text{out}}^{(\text{jam})}(t) = q_{n}(t) \mid_{t > t_1}, \) \( v_{\text{max}}^{(\text{jam})}(t) = v_{n}(t) \mid_{t > t_1} \); \( q_{n}(t) \) and \( v_{n}(t) \) are the flow rate and average vehicle speed measured at the downstream border of the
section (Fig. 21.5, $x = 0$); $q_{\text{min}}^{(\text{jam})}(t)$ is the flow rate within the moving jam; $\rho_{\text{max}}^{(\text{jam})}(t)$ is the density within the moving jam.\(^3\)

The parameter $\rho_{\text{max}}^{(\text{jam})}(t)$ can be calculated from traffic data measured at the detectors via the following formula:

$$\rho_{\text{max}}^{(\text{jam})}(t) = \frac{1000}{d_{\text{PC}}S_{\text{PC}}(t) + d_{\text{long}}(1 - S_{\text{PC}}(t))}, \quad (21.4)$$

where $d_{\text{PC}}$ is the average length of passenger cars (PC) including an average distance between vehicles within the moving jam (e.g., $d_{\text{PC}} = 7$ m), and $d_{\text{long}}$ is the average length of long vehicles (heavy goods vehicles) including an average distance between vehicles within the moving jam (e.g., $d_{\text{long}} = 17$ m); $S_{\text{PC}}$ is the fraction of all vehicles that are PC and $1 - S_{\text{PC}}$ is the fraction that are long vehicles; the unit of $d_{\text{PC}}$ and $d_{\text{long}}$ is [m], the unit of the density $\rho_{\text{max}}^{(\text{jam})}$ is [vehicles/km]. The time dependence $S_{\text{PC}}(t)$ in (21.4) can be determined via local detector measurements. $S_{\text{PC}}(t)$ is usually almost regardless of the detector location if there are no on- and off-ramps between detectors. In online application of the ASDA model, the detector that is upstream of a moving jam (detector $Q_0$ in Fig. 21.5) is used for the measurement of $S_{\text{PC}}(t)$.

The flow rate $q_{\text{min}}^{(\text{jam})}$ in (21.1), (21.2) can be set to zero:

$$q_{\text{min}}^{(\text{jam})} = 0. \quad (21.5)$$

The use of the approximate formula (21.5) shows good results in online application of the ASDA model.

Note that the fact that the downstream front of a wide moving jam propagates upstream with a constant mean characteristic velocity $v_g$ is also used in the ASDA model. If a wide moving jam has been detected but measurements of flow rate and vehicle speed downstream of the jam cannot be made, the characteristic velocity $v_g$ found via earlier measurements is used to track the downstream jam front:

$$v_{\text{down}}^{(\text{jam})} = v_g. \quad (21.6)$$

This is also used when a sequence of moving jams occurs and no measurements of flow rate and speed between the jams are possible. In this case, the velocity of the downstream jam front $v_{\text{down}}^{(\text{jam})}(t)$ in (21.2) between the jams is set to $v_g$.

Altogether the proposed model for tracking moving jams and predicting vehicle trip times consists of (21.1)–(21.5), where in an application traffic data can be measured by detectors (double loop detectors), and the FOTO

---

\(^3\) Note that parameters associated with moving jams ($v_{\text{down}}^{(\text{jam})}(t), v_{\text{up}}^{(\text{jam})}(t), v_{\text{max}}^{(\text{jam})}(t)$, etc.) in (21.1)–(21.3) are not necessarily equal to the related mean values. For this reason, we use in this chapter and in Appendix B other designations in comparison with the designations of the related parameters of wide moving jams used above in Parts I–III.
model is used as the incident detection algorithm to determine $t_0$ and $t_1$. The parameters $d_{PC}$ and $d_{HGV}$ are not dependent on local situation. Therefore, they can be chosen as constants. Consequently, the ASDA model has no parameters to be validated under different environmental conditions.

### 21.4.1 Tracking Synchronized Flow with FOTO Model

Similar to the jam-tracking ASDA model, the positions of the upstream front $x_{up}^{(syn)}(t)$ and the downstream front $x_{down}^{(syn)}(t)$ of “synchronized flow” objects are tracked after synchronized flow is first detected by the FOTO model. It is taken into account that the downstream front of synchronized traffic $x_{down}^{(syn)}$ is usually spatially fixed at the effective location of a bottleneck (e.g., near an on-ramp) on a freeway section while the upstream front is influenced by the upstream supply of vehicles.

The best results of synchronized flow tracking with the FOTO model can be achieved if a detector is close to the effective location of the bottleneck where the downstream front $x_{down}^{(syn)}$ of synchronized flow is fixed. In contrast to the downstream front, the upstream front of synchronized flow begins to move upstream after an $F\to S$ transition has occurred at the bottleneck.

Let us consider a detector $B$ upstream of an effectual bottleneck where synchronized flow is first detected at time $t_{syn,B}$ due to its upstream propagation from the effectual bottleneck. Let us further consider an upstream detector $A$ where traffic is still in the “free flow” phase at time $t > t_{syn,B}$ when traffic at detector $B$ is still in the “synchronized flow” phase (Fig. 21.6a).

It is clear that the upstream front of synchronized flow is somewhere between detectors $A$ and $B$ at time $t$. The flow rate upstream of synchronized flow is computed by taking into account the incoming flow from the detector $A$, $q_A(t)$, and also taking into account flow rates to the on- and off-ramps. To calculate the upstream front location of synchronized flow at time $t$, $x_{up}^{(syn)}(t)$, we set the origin $x = 0$ at the location of the detector $B$. With distance $D$ between detectors $A$ and $B$, the location of detector $A$ is $x = -D$. Since the upstream front of synchronized flow $x_{up}^{(syn)}(t)$ is somewhere between detectors $A$ and $B$, the following inequality holds:

$$-D < x_{up}^{(syn)}(t) < 0. \quad (21.7)$$

To estimate the position $x_{up}^{(syn)}(t)$, two different approaches have been developed and evaluated:

(i) An ASDA-like approach (Sect. 21.4.2).
(ii) A cumulative flow rate approach (Sect. 21.4.3).
Fig. 21.6. Tracking synchronized flow with the FOTO model. (a) Sketch of synchronized flow tracking. (b) Velocity of the upstream front of synchronized flow calculated through the formula (21.8). (c) Location of the calculated upstream front of synchronized flow for the ASDA-like approach (21.8) (dashed curve) and for the cumulative flow rate approach (21.12), (21.13) (solid curve). Black points in (c) are associated with measurements of the first occurrence of synchronized flow at the detectors. In (b, c) data from January 28, 2002; arrangement of detectors is shown in Fig. 21.4a. Taken from [238]
21.4 Tracking Moving Jams with ASDA

21.4.2 ASDA-Like Approach to Tracking Synchronized Flow

Let us assume that synchronized flow at detector \( B \) has been registered at \( t = t_{\text{syn},B} \) (Fig. 21.6a). At \( t > t_{\text{syn},B} \) the position \( x_{\text{up}}^{(\text{syn})} \) of the front of synchronized flow downstream and free flow upstream between any two neighbouring detectors \( A \) and \( B \) (detectors \( A \) and \( B \) characterized by flow rates \( q_A \) and \( q_B \), respectively) can be calculated by the ASDA-like formula:

\[
x_{\text{up}}^{(\text{syn})}(t) = \int_{t_{\text{syn},B}}^{t} v_{\text{up}}^{(\text{syn})}(t) \, dt = \int_{t_{\text{syn},B}}^{t} \frac{q_A^*(t) - q_B(t)}{\rho_A(t) - \rho_B(t)} \, dt, \quad t \geq t_{\text{syn},B},
\]

(21.8)

where

\[
q_A^* = q_A + q_{\text{on}} - q_{\text{off}},
\]

(21.9)

\[
\rho_A = \frac{q_A^*}{v_A}, \quad \rho_B = \frac{q_B}{v_B}.
\]

(21.10)

Recall that in tracking wide moving jams with the ASDA model, it has been shown empirically that the upstream front position of a wide moving jam \( x_{\text{up}}^{(\text{jam})}(t) \) (21.1) is in close agreement with the real jam front position found from measurements of jam propagation through different detectors. However, in tracking the front of synchronized flow with the formula (21.8), we found that the calculated front position \( x_{\text{up}}^{(\text{syn})}(t) \) sometimes shows very poor correspondence with measurements of this front at detectors.

An example is shown in Figs. 21.6b,c. The difference in the front positions calculated via (21.8) (dashed curve in Fig. 21.6c) and found through measurements (black points in Fig. 21.6c) is sometimes up to 100%. To explain this, note that the mean empirical velocity of the upstream synchronized flow front is \( \bar{v}_{\text{up}}^{(\text{syn})} = -7.9 \text{ km/h} \), however, the mean calculated velocity via (21.8) is \( \bar{v}_{\text{up}}^{(\text{syn})} = -11 \text{ km/h} \).

The reason for this behavior can be as follows. The formula (21.8) assumes that flow rates and vehicle densities measured at location \( A \) approximate those in free flow directly upstream of the location of the front \( x_{\text{up}}^{(\text{syn})} \). Moreover, this formula assumes that measurements of these traffic variables at detector \( B \) are nearly the same as these variables are directly downstream of the synchronized flow upstream front location \( x_{\text{up}}^{(\text{syn})} \). Our empirical investigations show that for free flow upstream of the front location \( x_{\text{up}}^{(\text{syn})} \) the assumption of (21.8) is nearly valid. However, downstream of the location \( x_{\text{up}}^{(\text{syn})} \) of the upstream front of synchronized flow, i.e., in synchronized flow, (21.8) leads often to an incorrect result. This is associated with features of synchronized flow, which usually exhibits a very large variation in vehicle density over the road even if the flow rate remains almost constant. The error in this ASDA-like model decreases with smaller distances between detectors and shorter measurement intervals. An important advantage of this
approach is that it works without any parameter validation through the use of measured traffic data only.

21.4.3 Cumulative Flow Rate Approach to Tracking Synchronized Flow

Empirical observations have shown that there is a nearly linear correspondence between the upstream front position \( x_{\text{up}}^{(\text{syn})} \) of synchronized flow (Fig. 21.6c, black points) and variations in the cumulative number of vehicles between detectors A and B that occurred after the time \( t = t_{\text{syn,B}} \). This change is the difference \( \Delta M_{\text{total}} \) of the cumulative flow rate passing detector B and the cumulative flow rate into the region of synchronized flow that is influenced by the flow rate on the main road measured at detector A and the on- and off-ramps between detectors A and B:

\[
\Delta M_{\text{total}}(t) = \int_{t_{\text{syn,B}}}^{t} q_B(t) \, dt - \int_{t_{\text{syn,B}}}^{t} q_A^*(t) \, dt, \quad t > t_{\text{syn,B}}. \tag{21.11}
\]

The flow rates in (21.11) are the total flow rates across the road. \( \Delta M_{\text{total}}(t) \) is the net number of vehicles leaving the road section between detectors A and B. When the number of vehicles between the detectors at \( t > t_{\text{syn,B}} \) is greater than the one at the time \( t = t_{\text{syn,B}} \), the value \( \Delta M_{\text{total}} \) is negative.

To make the FOTO formulae for the determination of \( x_{\text{up}}^{(\text{syn})} \) independent of the number of lanes \( n \), \( \Delta M_{\text{total}} \) is normalized to the traffic volume per lane, \( \Delta M_{\text{lane}} \), which will be referred to as \( \Delta M \):

\[
\Delta M(t) = \Delta M_{\text{lane}}(t) = \frac{\Delta M_{\text{total}}(t)}{n}. \tag{21.12}
\]

The correspondence between \( x_{\text{up}}^{(\text{syn})} \) \( (x_{\text{up}}^{(\text{syn})} < 0) \) and \( \Delta M \) has been found empirically to be approximately a linear function:

\[
x_{\text{up}}^{(\text{syn})}(t) = \mu \Delta M(t), \tag{21.13}
\]

where \( \mu \) is a constant, \( x_{\text{up}}^{(\text{syn})} \) is the upstream synchronized flow front position relative to detector B (in meters), the unit of \( \Delta M \) is [vehicles/lane]. However, this is an average behavior of the upstream front of synchronized flow. When measurements of \( \Delta M(t) \) are inserted into (21.13), the calculated front position \( x_{\text{up}}^{(\text{syn})} \) (21.13) can sometimes contradict condition (21.7). Recall that we have assumed that at detector A the “free flow” phase and at detector B the “synchronized flow” phase are measured (Fig. 21.6a). For this reason, the calculated value \( x_{\text{up}}^{(\text{syn})} \) (21.13) is adjusted according to (21.7), i.e., if the calculated value \( x_{\text{up}}^{(\text{syn})} > 0 \), it is set to zero, and if the calculated value \( x_{\text{up}}^{(\text{syn})} \leq -D \), it is set to \(-D + \epsilon_d\), where \( \epsilon_d \) is a small value, \( \epsilon_d \ll D \) (we used \( \epsilon_d = 1 \text{ m} \)).
For the bottleneck near detectors D6 (Fig. 21.4a) it has empirically been found that the mean value $\mu = 33 \text{ m/vehicles}$ can be used in (21.13). For other different effectual bottlenecks on sections of the freeway A5 a variation of $28 < \mu < 40 \text{ m/vehicles}$ has been found. Nevertheless, we used the same value $\mu = 33 \text{ m/vehicles}$ for all different effectual bottlenecks in the online application of the FOTO model on the freeway A5. This produces good results in tracking synchronized flow because the possible error in the position of the upstream front of synchronized flow is about 15%. However, the constant $\mu$ in (21.13) might be different for effectual bottlenecks on other freeways. Thus, in these cases the constant $\mu$ in (21.13) should be found for each effectual bottleneck based on historical empirical data.

An example of the application of (21.13) to tracking synchronized flow is shown in Fig. 21.6c (solid curve). The example was done with a reduced infrastructure where only data from D1, D5, and D5-off were used. The times when synchronized flow has been measured at the detectors, D1, D2, D3, D4, and D5, are shown as a reference with black circles. In comparison to the positions of the upstream front of synchronized flow calculated by the ASDA-like formula (dashed curve in Fig. 21.6c; error 100% at 6:56: the upstream front of synchronized flow, which is supposed to be at $x = -4.0 \text{ km}$, is in reality at $x \approx -2.0 \text{ km}$), the positions of this front from the cumulative flow rate formula (21.13) match the positions of the upstream front of synchronized flow measured at the detectors (black points in Fig. 21.6c) much better. The cumulative flow rate approach enables us the determination of the upstream front of synchronized flow in accordance with real synchronized flow propagation. However, in this case an additional parameter $\mu$ in (21.13) is required that is valid for a specific effectual bottleneck.

21.5 Conclusions

(i) The FOTO and ASDA models, which are based on three-phase traffic theory, perform without any validation of model parameters in different environmental and infrastructure situations. For this reason, the models are suitable for online application in traffic control centers for all kinds of traffic management and control.

(ii) The FOTO and ASDA models make it possible to reconstruct and to track the main spatiotemporal features of congested patterns in freeway traffic, in particular at freeway bottlenecks. The inputs to the models are local measurements of traffic variables. The output of the models is spatiotemporal pattern recognition and tracking between detectors where no measurements can be made.

(iii) The FOTO and ASDA models reconstruct and track the “synchronized flow” and “wide moving jam” phases between detectors on freeways with high quality results.
22 Spatiotemporal Pattern Recognition, Tracking, and Prediction

22.1 Introduction

Recognition, tracking, and prediction of traffic are necessary for almost all traffic engineering applications, in particular for vehicle routing guidance systems, traffic control, assignment and management (e.g., [21,495,517–520,522–524,526–534]). In all these and other engineering applications a knowledge of current spatiotemporal congested patterns and their prediction are of a great importance. The main aim of this chapter is a discussion of results of methods for recognition, tracking, and prediction of spatiotemporal congested patterns at freeway bottlenecks.

Firstly, we will consider results of the online application of the FOTO and ASDA models at the Tee (traffic control center) of the German Federal State of Hessen, on other German freeways and on a freeway in the USA for recognition and tracking of spatiotemporal congested patterns. Secondly, we will briefly discuss some other engineering methods that can be used for recognition of predictable pattern features. These methods belong to the method of “floating car data” (FCD), where measurements of traffic are made by “probe vehicles” in traffic flow. Thirdly, we will consider a possible application of the method of historical time series for the creation of historical prediction of spatiotemporal congested patterns in traffic networks. A possible matching of this historical database with current measurements can improve traffic prediction considerably. Finally, we discuss a method for efficient traffic prediction in urban areas.

22.2 FOTO and ASDA Application for Congested Pattern Recognition and Tracking

22.2.1 Validation of FOTO and ASDA Models at Traffic Control Center of German Federal State of Hessen

Both the FOTO and ASDA models are integrated into the application FOTOwin. For operation at the TCC of the German Federal State of Hessen the data exchange is performed with a central information distributor, which manages the data of the detectors. The models were installed first on the A5
freeway near the variable message signs equipment between the intersections “Westkreuz Frankfurt” and “Anschlussstelle Friedberg” in both directions. A more detailed discussion of the implementation of the FOTOwin software can be found in [237].

FOTOwin has a feature, which allows it to show the “synchronized flow” and “wide moving jam” objects over a longer time interval and to investigate the spatiotemporal movement of the objects. A graph of synchronized flow and wide moving jams in space and time gives the history of the traffic objects (Fig. 22.1). The x-axis is the time and the y-axis is the location on the road. In the representation, “wide moving jam” objects are shown in black and “synchronized flow” objects are shown in gray. The infrastructure is shown with horizontal lines for the detector locations. Using the graph of synchronized flow and moving jams in space and time, the movement of the traffic objects can directly be seen.

It should be noted that in each case the reference for a correct or false classification of the “synchronized flow” and “wide moving jam” phases with the FOTO model is made based on the following approach. We compare the traffic phases reconstructed by the FOTO model with results of the identification of synchronized flow and wide moving jams in congested traffic based on the objective criteria of Sect. 4.2. Naturally, the latter is only possible with high quality results if detector distances are small (about or less than 1000 m). Most results of the FOTO and ASDA application below will be presented for sections of the A5 freeway (Figs. 2.1 and 2.2). On these freeway sections, there are about about 30 detectors along approximately 30 km lengths, i.e., the mean detector distance is 1 km.

For the evaluation of the FOTO and ASDA models, the application is tested in considerably reduced configurations: traffic data of several detectors have been omitted. In the reduced configuration, the FOTO and ASDA models (Figs. 22.1c,d) reproduce wide moving jams and synchronized flow that correlate well with Figs. 22.1a,b where all detectors have been used. In particular, a possible dissolution of the “wide moving jam” and “synchronized flow” objects, which occurs between detectors, can be predicted, i.e., even when the objects cannot be measured. An example is shown in Fig. 22.1a, where the wide moving jam labeled “jaml” dissolves at \( t = 9:44 \). This moving jam dissolution is predicted by means of jam tracking with the ASDA model in the reduced detector configuration at \( t = 9:41 \) (Fig. 22.1c), i.e., when the moving jam cannot be measured by the detectors.

As discussed above, when location of the upstream front of synchronized flow cannot be measured, this location can be predicted via synchronized flow tracking with the FOTO model. An example is shown in Fig. 22.1d where through this tracking a limitation of the upstream propagation of the “synchronized flow” object is predicted. In this case, the upstream front of synchronized flow is predicted to be located between 10 and 15 km (in agreement with the result in Fig. 22.1b where the full detector configuration is
Fig. 22.1. Evaluation of FOTO and ASDA application. Overview of spatiotemporal congested patterns. (a) Congested patterns in space and time on the freeway A5-South (data from March 17, 1997) where all 31 available detectors are used. (b) Congested patterns in space and time on the freeway A5-North (data from August 11, 2000) where all 30 available detectors are used. (c) Congested patterns in space and time related to (a), however, for a reduced detector configuration, where instead of 31 detectors 5 detectors are used. (d) Congested patterns in space and time related to (b), however, for a reduced detector configuration, where instead of 30 detectors 10 detectors are used. Taken from [238]

used) whereas measurements between these freeway locations are not used in the FOTO model in Fig. 22.1d. A statistical evaluation of various reduced detector configurations is described in Sect. B.2.

22.2.2 Application of FOTO and ASDA Models on Other Freeways in Germany and USA

The FOTO and ASDA models have been installed over the entire freeway network of the German Federal State of Hessen with about 800 detectors and 850 km of a freeway network. The models have to cope with different
detector infrastructures (distances between detectors up to 10 km), different measurement intervals (1 and 5 minutes) and with different kinds and mixtures of local, regional, long-distance, and leisure traffic. Figure 22.2a illustrates FOTO and ASDA model results in the Rhein/Main-area: locations of wide moving jams (in black) and regions of synchronized flow (in white) are marked on different freeway sections.

Figure 22.2b shows a graph of synchronized flow and moving jams in space and time on the freeway A67-North between AS Lorsch (44 km) and Darmstädter Kreuz (31 km) within the network in Fig. 22.2a that has a larger percentage of long-distance traffic than freeway A5 located near the city of Frankfurt. In an example of FOTO and ASDA application (Fig. 22.2b), three wide moving jams propagating upstream can be seen on the 13 km stretch with four detectors at 3–5 km spacing. Note that because the detectors are in this case far from bottlenecks (only a detector A67/106N is close to on- and off-ramps), on the two-lane (in each direction) freeway, the regions of synchronized flow that emerge at those bottlenecks cannot be detected and tracked. On the other hand, wide moving jams are tracked with the ASDA model at relatively large detector spacing of 3–5 km. The characteristic velocity of downstream jam front propagation here is $-16$ km/h at a maximum observation spacing of 8.1 km between the detectors A67/103N and A67/107N. This correlates well with results on freeway sections of the A5 freeway near Frankfurt (Figs. 2.1 and 2.2).

Additional examples of FOTO and ASDA application on freeways near Magdeburg and near Munich are shown in Figs. 22.3-22.8. The results of ASDA and FOTO application in these examples are qualitatively the same as results of online application to freeways at the TCC of the German Federal State of Hessen (Figs. 22.1 and 22.2).

The FOTO and ASDA models have been applied to the congested pattern reconstruction, tracking, and prediction on freeways in California. Two examples that show common features of this application can be seen in Figs. 22.9 and 22.10. On freeway sections where congested patterns occur, there are usually many potential adjacent bottlenecks close to one another. For this reason, a variety of complex EPs often appear (see footnote 4 of Sect. 2.4.8). All EPs that we could observe in traffic data of the PeMS (Freeway Performance Measurement System) database in the USA (California), usually consist of two traffic phases of congested traffic, “synchronized flow” and “wide moving jam.” Qualitative features of these traffic phases are the same as those in complex congested patterns observed on German freeways. In other words, congested patterns on American freeways can also be explained by three-phase traffic theory. Thus, the results of FOTO and ASDA application on American freeways (Figs. 22.9 and 22.10) are qualitatively the same as those for various freeways in Germany (Figs. 22.1, and 22.4-22.8).

It has been shown that the FOTO and ASDA models are reliable without any validation of model parameters in all situations. Most traffic objects in
Fig. 22.2. Results of FOTO and ASDA application in the Rhein/Main area on several freeways. (a) Scheme of the area with results of data from April, 22, 2002 at 8:00. (b) Overview of congested patterns on the freeway A67-North (data from April 22, 2002). Taken from [238]
Congested traffic (wide moving jams and synchronized flow) can be reconstructed and tracked to freeway infrastructures with sparser detection. The dissolution of the “wide moving jam” and “synchronized flow” objects, which occurs between detectors, can be predicted, i.e., even when the objects cannot be measured. The choice of optimal detector locations with regard to locations of effectual bottlenecks yields significant savings in roadside infrastructure investment. In a four month field trial we found that with only 23% of all given measurements on the freeway A5 approximately 60% of relevant information can be reconstructed.

The application of the FOTO and ASDA models fully confirms three-phase traffic theory that is the theoretical basis for this application approach to freeway congested pattern traffic recognition, tracking, and prediction.

The online description of current traffic states and the tracking of “wide moving jam” and “synchronized flow” objects based on stationary measurements performed by the FOTO and ASDA models gives successful opportunities to reconstruct and to track realistic traffic patterns for traffic control centers as basic information for all kinds of traffic management.

### 22.3 Spatiotemporal Pattern Prediction

#### 22.3.1 Historical Time Series

The very first empirical studies of freeway traffic showed that the time dependence of traffic demand possesses some predictable features (see, e.g., [21, 286, 498–502, 504, 505, 534]). These features depend on characteristics of the day, for example whether data is associated with a working day or a weekend, or whether there is a football match on that day.

If all related characteristics of a day are known, then there will be a certain set of the dependencies of flow rates over time that are relevant for a freeway section on that day. Each of these dependencies is one of the historical time series for flow rates.

Thus, based on measurements of traffic flow rates on different days, months, and years, and if characteristics of these days are also known, a
22.3 Spatiotemporal Pattern Prediction

Fig. 22.4. Evaluation of FOTO and ASDA application on freeways near Magdeburg. Moving jams and synchronized flow in space and time. (a) All available detectors are used. (b) Reduced detector infrastructure. Data from September 13, 2002. Taken from [238]

database for different historical time series can be created (e.g., [497–504, 534]). There are at least two parameters of the database:

(i) the day characteristics
(ii) the probability for occurrence of each of the time series of traffic variables for days with the same characteristics
Fig. 22.5. Layouts of sections of freeways A9-North (a) and A9-South (b) near Munich. Taken from [238]

Fig. 22.6. Results and evaluation of FOTO and ASDA application on freeways near Munich. Moving jams and synchronized flow in space and time. (a) All available detectors are used. (b) Reduced detector infrastructure. The freeway A9-North, data from June 13, 2002. Taken from [238]
22.3 Spatiotemporal Pattern Prediction

Fig. 22.7. Results of FOTO and ASDA application on freeways near Munich. Complex spatiotemporal traffic pattern on the freeway A9-South, data from June 24, 2002. Taken from [238]

Fig. 22.8. Results of FOTO and ASDA application on the freeway A99 near Munich, Germany. Propagation of a sequence of wide moving jams through complex states of synchronized flow, data from June 22, 2002. Taken from [496]
Fig. 22.9. Application of the FOTO and ASDA models on the Interstate freeway I405-South, Orange County near Los Angeles, California, the USA. (a) Layout of a section between freeway exits “Bolsa Chica” and “Euclid Street.” (b) Propagation of a sequence of wide moving jams through complex states of synchronized flow of an expanded congested pattern (EP). A part of this figure is shown in Fig. 1.4. Taken from [236]
Thus, each of the historical time series in this database exhibits day characteristics and the probability for historical time series occurrence. This procedure can also be used for historical time series of average vehicle speed on different links of a traffic network.

If the flow rates and average speed are known, one can also estimate the average vehicle density, link vehicle trip (travel) times, and other traffic variables on links of a traffic network. Thus, besides flow rates and average vehicle speed, vehicle density and link travel times in the traffic network are traffic variables for which databases of historical time series are often made.

Because the probability distribution for different time series is known, the time series with the highest probability is usually chosen for traffic prediction.

### Matching of Time Series with Current Measurements

The choice of the most probable of the time series is made only if there are no current traffic measurements. If there are current measurements of traffic variables in some traffic network locations, then one can apply a matching of current measurements with time series from a database.
Due to this matching, time series related to the current data should be chosen. This best matched time series is not necessarily related to the highest probability in the database for different time series on the day. This enables us to increase the probability of traffic prediction considerably.

It should be noted that the form of time dependencies of traffic variables in historical time series is often very similar on different days that have the same characteristics (see an example of very similar flow rates to the effectual on-ramp at D6 on the freeway A5-South on three different years, labeled “eff-on” in Fig. 12.6).

However, instants of characteristic changes in these time dependencies (e.g., the time at which the flow rate begins to increase abruptly) are usually randomly shifted with respect to one another. The onset of congestion is also a random effect, because it is associated with a first-order F→S transition that has a probabilistic nature (Sects. 2.4 and 10.3). Congested pattern emergence usually changes characteristic features of the time dependencies of flow rates downstream of a congested pattern. In addition, at the same traffic demand various congested patterns with very different parameters can be formed (Part II). This is associated with the metastability of congested pattern formation. At given the same traffic demand, depending on initial conditions and on random perturbations in traffic, qualitatively different congested patterns can emerge. These probabilistic effects lead to an additional randomization of instants of characteristic changes in the time dependencies of flow rate (an example is flow rates at D6 in Fig. 12.6), average vehicle speed, link travel times, and other traffic variables in the related historical time series. The matching of time series with current measurements enables us to reduce the errors in traffic prediction that are due to these random effects.

The development of such databases for various time series and the matching of time series with current measurements are now standard methods for all traffic applications and traffic management strategies (e.g., [497–504, 534]).

**Floating Car Data**

Both historical time series of traffic variables in a traffic network and current measurements can also be made if there are no detectors on a freeway section. For this purpose floating car data (FCD) can be used (e.g., [506–516, 534–536]).

FCD is a traffic technology that has been used for years to measure link travel times in traffic networks. Each FCD vehicle is equipped with an FCD device. The device can measure vehicle speed and location in a traffic network. Based on this data, the FCD device calculates the travel time on the link of the traffic network that the vehicle has just left.

If the current link travel time is greater than some given travel time, it will be sent to a traffic center. The travel center, based on matching of the current FCD measurements with a database of time series for link travel
times, makes a prediction of the link travel times. This prediction can further be used to calculate the best (quickest) individual vehicle trip in the traffic network.

This traffic navigation can also be used for traffic assignment and other management strategies (see references in [517–520, 522–526]), for example with the goal of minimizing all link traffic times in a traffic network.

**Probe Data**

To measure current travel times in a traffic network, another method called “phone probe data” can be used (e.g., [537–540]). In this method, a vehicle is equipped with a mobile phone. The current location of the mobile phone and time at which the mobile phone is at that location can be found. This data is sent through a mobile phone provider to a traffic center. The traffic center calculates the current travel time on the link of the traffic network that the mobile phone (and therefore the vehicle) has just left.

**22.3.2 Database of Reproducible and Predictable Spatiotemporal Pattern Features**

Empirical studies of spatiotemporal congested patterns at freeway bottlenecks considered in Part II of the present book, based on data measured during 1995–2003, have shown that the patterns exhibit a number of reproducible and predictable spatiotemporal features [218] (see also Sect. 2.4.10). This conclusion is also confirmed by application of the ASDA and FOTO models to the various freeways in Germany and the USA discussed above.

These reproducible and predictable spatiotemporal pattern features are as follows.

(i) Types of congested patterns that appear spontaneously at a given effectual freeway bottleneck or types of patterns on a freeway section where several closely spaced adjacent bottlenecks exist. This feature determines one of the SPs, or one of the GPs, or else one of the EPs that occur with the highest probability at a given traffic demand.

(ii) In many cases, for a given time dependence of traffic demand, a certain type either of pattern evolution or transformation between different congested patterns is observed with the highest probability.

(iii) The probability distribution of a certain pattern occurrence that depends on traffic demand can be found. This gives the probability of the type of congested pattern at a given effectual freeway bottleneck or the type of congested pattern at a freeway section where several closely spaced adjacent effectual bottlenecks exist.

(iv) If this pattern is an GP, then characteristics of the GP such as whether the weak congestion condition or the strong congestion condition is realized within the GP, the mean width (in the longitudinal direction) of
the pinch region in the GP, as well as the average speed and density in the pinch region in the GP can be found, which the GP should have with the highest probability at a certain bottleneck.

(v) If this pattern is an EP, then characteristics of the EP such as whether there is only one or several separated pinch regions in the EP, whether foreign wide moving jam propagation is expected, as well as an approximate location of the upstream front of the farthest upstream pinch region in the EP can be predicted.

(vi) If there are bottlenecks where moving jams dissolve (Sect. 14.3), then special types of congested patterns like an GP where the region of wide moving jams dissolves either fully or partially (this pattern can resemble an LSP) can be predicted for some freeway sections.

Thus, based on measurements of these spatiotemporal congested pattern features on many different days, we can create a database for spatiotemporal congested patterns [231].

In this database, each set of spatiotemporal pattern features is related to the day that has the same characteristics (such as a weekend, working day, events relevant to traffic, and so on). If there are several different sets of spatiotemporal pattern features for the days with the same characteristics, one must find a probability for each such spatiotemporal pattern feature. This permits the determination of the probability distribution for spatiotemporal pattern features.

If current measurements are available, then the following matching of data of the database with these current measurements can be used for reliable prediction of spatiotemporal congested pattern features for a given freeway section.

Spatiotemporal congested pattern characteristics can be found both based on FCD [545] and probe phone technologies [542–544].

In particular, an FCD vehicle can determine the location of the downstream front of a congested pattern at an effectual bottleneck. Recall that within the downstream front of the congested pattern (WSP, LSP, GP, or EP) the vehicle accelerates from the “synchronized flow” phase to the “free flow” phase. Thus, there is a characteristic spatiotemporal behavior of vehicle speed while the vehicle moves through synchronized flow within the pattern and then accelerates to free flow speed inside the downstream front of synchronized flow.

Due to current measurements of vehicle speed and location made by an FCD vehicle, the determination of location of the downstream front of the congested pattern with FCD vehicles is also possible if effectual bottlenecks in a freeway network are not known. The latter is often the case when there are not enough detectors for this purpose on roads in a traffic network. Based on FCD measurements on different days, a traffic center can determine all locations of effectual bottlenecks in the freeway network. To find the effective locations of these effectual bottlenecks, one requires to compare the
FCD-determined locations of the downstream fronts of the congested patterns on different days. If there is a freeway location where the downstream front has been measured on many different days, then that location is one of the effective locations of effectual freeway bottlenecks.

Current spatiotemporal characteristic features of congested patterns can also be determined by FCD vehicles due to continuous speed measurements when an FCD vehicle moves through a congested pattern. These pattern characteristics can be used for two purposes:

(i) A historical database of effectual bottlenecks in a freeway network should first be produced. This is related to empirical results that the case whether a freeway bottleneck is an effectual bottleneck also depends on traffic demand. The latter is a function of time and day characteristics. Thus, the probability distribution function (whose variables are day characteristics, events such as a football match, weather, and so on) for effectual bottlenecks and their effective locations can be found and stored in this database.

(ii) A historical database for different spatiotemporal characteristic features of congested patterns associated with effectual bottlenecks in the historical database of effectual bottlenecks should be created. In the database of spatiotemporal characteristics of congested patterns, the types of patterns (WSP, LSP, GP, or EP), pattern parameters and scenarios for time dependence of these parameters can be stored. Examples of these pattern parameters are:

1. The mean width of the pinch region in an GP.
2. The average speed in the pinch region of the GP.
3. The mean frequency of narrow moving jams emerging in the pinch region of the GP.
4. The mean velocity $v_g$ of the downstream front of wide moving jams, the flow rate $q_{out}$ and the density $\rho_{min}$ in the jam outflow when in this outflow free flow is formed.
5. The mean duration of wide moving jams and the mean time between wide moving jams.
6. The mean velocity and the time dependence of location of the upstream front of a pattern (for WSP, GP, and EP).
7. The mean lifetime of the pattern.
8. Possible transformation scenarios between different patterns over time after the pattern has emerged.
9. Time dependence of locations of the downstream and upstream fronts of wide moving jams.
10. Different average characteristics associated with a congested pattern, like travel time through each part of the pattern.

Some of this information can also be presented in the form of a graph of synchronized flow and moving jams in space and time (Fig. 22.11),
where a predicted vehicle trajectory is shown that should occur if the vehicle moves through the congested pattern.

(iii) Current measurements can be used to match the most suitable set of pattern characteristics from the database. This enables us to predict subsequent pattern evolution and pattern transformation. These spatiotemporal pattern characteristics are stored in the database.

FCD-based pattern recognition and measurement can also be used to estimate traffic demand upstream of a congested pattern. This, in turn, enables us to make reliable prediction of pattern lifetime. This is because the current flow rate upstream of the pattern can be matched with the historical database of flow rate to predict the development of this flow rate.

For example, if a congested pattern is an GP or an EP, the upstream flow rate \( q_{\text{in}} \) can be estimated if there are at least two FCD vehicles. These FCD vehicles should approach the upstream front of the farthest upstream wide moving jam in the GP or the EP at different times, \( t_1 \) and \( t_2 \) \((t_2 > t_1)\). Then the flow rate \( q_{\text{in}} \) can be estimated from the formula

\[
\begin{align*}
q_{\text{in}} &= \frac{q_1 + q_2}{2} \\
&= \frac{q(t_1) + q(t_2)}{2}
\end{align*}
\]
22.3 Spatiotemporal Pattern Prediction

\[ q_{in} = \left[ 1 + \frac{(\tau_2 - \tau_1)}{(t_2 - t_1)} \right] q_{out} \]  \hspace{1cm} (22.1)

In (22.1), \( q_{out} \) is the characteristic flow rate in the wide moving jam outflow; \( \tau_1 \) and \( \tau_2 \) are respectively waiting times of the first and the second FCD vehicles within the farthest upstream wide moving jam before vehicles can accelerate at the downstream jam front.

Traffic pattern reconstruction can be more accurate when FCD methods are combined with the ASDA and FOTO models [541].

Spatiotemporal pattern features at freeway bottlenecks can also be reconstructed based on probe phone data technology [542].

The database for spatiotemporal congested patterns can also be used for traffic prediction based on the ASDA and FOTO models. In this case, a spatiotemporal congested pattern recognized through the use of the ASDA and FOTO models at the current time is matched with historical spatiotemporal congested patterns from the database (Fig. 22.12). Based on this matching a more appropriate congested pattern from the database is chosen for traffic prediction. This pattern can be used for traffic prediction. There is also another way for traffic prediction. Based on the chosen congested pattern from the database a choice of time series of flow rate and vehicle speed can be found associated with this historical spatiotemporal congested pattern. Then these time series are used in the ASDA and FOTO calculation of traffic prediction based on current measurements.

Fig. 22.12. Simplified qualitative scheme for traffic prediction based on the ASDA and FOTO models [231, 236]
**22.3.3 Vehicle Onboard Autonomous Spatiotemporal Congested Pattern Prediction**

Predicted spatiotemporal congested patterns can be used to enhance vehicle safety and improve vehicle navigation. However, if it is not possible of leaving the congested section of a freeway, results of this prediction can be used to increase safety and comfort when driving through the congested pattern.

Predicted spatiotemporal congested pattern characteristics can be sent to a vehicle from a traffic center. The traffic center can make this prediction based on matching of current measurements from detectors and from FCD vehicles with historical data for spatiotemporal congested pattern characteristics, as discussed above.

There is also another way for a vehicle to have such predicted data: the vehicle can make traffic prediction *autonomously*, i.e., without any traffic center [546–550].

Information about the dependence of vehicle speed on time, the vehicle location in a traffic network, and time gaps and speeds of other vehicles in the vehicle vicinity (if the latter information is available) can also be used in a vehicle for vehicle onboard autonomous spatiotemporal congested pattern recognition and prediction. This traffic prediction can be performed if there is a database for spatiotemporal congested pattern characteristics in the vehicle (Fig 22.13). This database should be stored in the vehicle (e.g., on a CD-ROM). The simplest case is when a driver usually uses only a part of a traffic network. In this case, the database should be stored only for the part of the traffic network most used by the driver.

Current time dependence of vehicle speed, current vehicle location in the traffic network, and historical locations of effectual bottlenecks from the database are the necessary inputs for this vehicle-based spatiotemporal congested pattern recognition and prediction.

Using this and other available input information, the vehicle automatically matches potential congested patterns in the database to the current time dependence of vehicle speed and location. Matching is carried out if the vehicle speed becomes lower than the minimum vehicle speed in free flow for the current section of the traffic network where the vehicle moves. This speed should be one of the traffic parameters in the database. Due to this matching, the most suitable spatiotemporal congested pattern characteristics are chosen that the vehicle must meet in the future if the vehicle follows the route. This prediction of spatiotemporal pattern characteristics can be used for traffic adaptive automatic cruise control and for other driver assistance systems in vehicles.

This vehicle based spatiotemporal congested pattern recognition and prediction can be combined with microscopic traffic simulations within the scope of three-phase traffic theory. A possible scenario of microscopic simulations is as follows. One of the vehicles in a microscopic traffic flow model is considered the vehicle whose measurements are used. The dependence of measured
vehicle speed on time and the vehicle location, as well as the location of an effectual bottleneck downstream of the vehicle, are used for congested pattern simulation at the effectual bottleneck. During these simulations, traffic demand in the vicinity of the bottleneck (the flow rates $q_{in}$ and $q_{on}$ for the case of an effectual bottleneck due to an on-ramp) is automatically calculated by the model. This traffic demand estimation is based on a comparison of the simulated time dependence of vehicle speed with the measured speed: these speeds should almost coincide to one another. A spatiotemporal congested pattern found in simulations can be matched to spatiotemporal congested patterns from the database (Fig. 22.13), to make a more reliable pattern choice. This matching can improve the pattern choice through the use of traffic flow models based on three-phase traffic theory because in this case a
model can predict spatiotemporal congested patterns at effectual bottlenecks in accordance with empirical results (Part III).

Vehicle-based spatiotemporal congested pattern recognition and prediction can be further improved if there are several communicating vehicles equipped with the above prediction system. In this case, the matching of the most suitable spatiotemporal congested pattern characteristics can be more accurate if this matching is based on measurements from two or more vehicles that move on the same freeway.

22.4 Traffic Analysis and Prediction in Urban Areas

Typical urban areas consist of two qualitatively different kinds of traffic networks:

(i) Freeway networks, in which dynamic traffic phenomena are associated with spatiotemporal effects determined by intrinsic traffic features, i.e., by driver interactions in traffic discussed above in this book.
(ii) City networks, in which dynamic traffic phenomena are fully determined by light signals and other traffic regulations at road intersections. In this case, traffic phenomena are associated with queue formation at road intersections (e.g., [19, 28, 31, 244–291, 316]).

Thus, one of the possible methodologies for traffic prediction in urban areas can be based on a separate consideration of traffic dynamics in these two different kinds of traffic networks.

In freeway networks of an urban area, methods for traffic prediction discussed above in this book can be applied (Sect. 22.3). However, in city networks of the urban area, another method should be used that is more adequate for prediction of queue formation at road intersections. The inflows into the entire traffic network at a chosen boundary of the urban area can be found based on historical time series as discussed in Sect. 22.3.1. The inflows into each of the freeway networks of the urban area are the outflows from other neighboring city networks of this area and vice versa. This ensures adequate traffic prediction over the entire traffic network of the urban area. To illustrate a possible way for traffic prediction in city networks of the urban area, in the remainder of this section a macroscopic model called “UTA” (Urban Traffic Analysis) introduced in [324] will be discussed.

22.4.1 Model for Traffic Prediction in City Networks

It is well-known that traffic phenomena in freeway traffic and in city traffic are qualitatively different (e.g., see references in [21, 31, 35]). As discussed in this book, spatiotemporal traffic patterns occurring in freeway traffic can be considered self-organized patterns. In contrast, in city networks, the traffic
dynamics is determined by light signals at road intersections: traffic flow is interrupted during the red phase there. As a result, in contrast to self-organized spatiotemporal congested patterns associated with driver interaction effects in freeway traffic, in city networks spatiotemporal congested patterns are associated with queue formation at road intersections. The queue dynamics in city traffic is governed mostly by light signal operations and other traffic regulations at road intersections rather than driver interaction effects.

The UTA model briefly discussed below is based on the following well-known results (e.g., [19, 28, 31, 244–291, 316]; see also references in [552]):

(i) The traffic dynamics at signalized intersections in urban areas can approximately be considered based on queuing theory where the dynamics of lengths of waiting queues at road intersections should be calculated.

(ii) On a time scale that is considerably greater than the whole period of the light signal at a road intersection $T_{\text{light}}$ the dynamics of a waiting queue length is important for the purpose of traffic control and for other applications.

(iii) There are two different cases of queue formation at a road intersection: (1) “undersaturation” and (2) “oversaturation.” In the undersaturation condition, a vehicle queue formed during the red phase $T_R$ of the light signal at a road intersection dissolves fully during the green phase $T_G$ of the light signal (Fig. 22.14a, dashed curves). In contrast, in the oversaturation condition, the queue dissolves only partially during the green phase of the light signal (Fig. 22.14b, dashed curves).

In the UTA model, a well-known method of an analysis of average flow rates at a road intersection over a time interval that is longer than the period of the light signal is applied. This method is used in many adaptive light signal control methods at signalized road intersections (see e.g., [274]).

In the UTA model [324, 551], effective macroscopic traffic variables (flow rates, waiting queue lengths, number of vehicles in the network links, etc.) are considered. For example, the effective number of vehicles in a queue $N_q(t)$ (Fig. 22.14a,b, solid curves $N_q(t)$) is associated with the real number of vehicles in the queue $N_q^{(ac)}(t)$ (Fig. 22.14a,b, dashed curves $N_q^{(ac)}(t)$) at a road intersection corresponding to the formula:

$$N_q(t) = T^{-1} \int_{t-T/2}^{t+T/2} N_q^{(ac)}(t') \, dt' , \quad (22.2)$$

where

$$T > T_{\text{light}} = T_R + T_G . \quad (22.3)$$

All other traffic variables in a city traffic network are also effective traffic variables associated with those changes in city traffic in which the time scale is considerable greater than the duration of the period of the light signal
Fig. 22.14. Explanation of the UTA model. (a, b) Qualitative time dependencies of the real number of vehicles $N_q^{(ac)}$ within a queue (dashed curves) and of the effective number of vehicles within the queue $N_q$ (solid curves) associated with a lane group on a link at a signalized road intersection in a city network for the undersaturation condition (a) and for the oversaturation condition (b). (c) Sketch of a possible road intersection in the network with four exiting links $i = 1, 2, 3, 4$ and four entering links $j = 1, 2, 3, 4$ each of them has two lane groups (i.e., two directions) $m = 1, 2$ and $k = 1, 2$, respectively (each of these lane groups consists of one lane). Arrows within the intersection in (c) illustrate some of the possible flows within the intersection

This approximation enables us to forecast traffic variables in very large urban traffic networks more easily and quicker than a calculation of real traffic variables, which can be very complex functions of time due to traffic interruption during red phases at road intersections. In particular, in contrast to a real queue outflow rate $q_{out, q}^{(ac)}$, which is a discontinuous function of time, the effective outflow rate $q_{out, q}$...
22.4 Traffic Prediction in Urban Areas

\[
q_{out}, q(t) = T^{-1} \int_{t-T/2}^{t+T/2} q_{out}, q(t') dt'
\]  

(22.4)

is a continuous function of time. Thus, in the UTA model, there is no interruption of effective traffic flow at road intersections in a city network. Duration of the red and green phases influence only the related effective flow rates in this effective continuous traffic flow in the city network.

In the UTA model, there can be different lane groups on a link of the network associated with possible different directions for vehicles beyond the link. Within one lane group on the link there can be one or more lanes.

Based on given inflow rates at the boundary of a city network, the UTA model calculates an effective queue length \( L_q \) associated with a lane group on a link of the network and other related values and variables, for example the effective number of vehicles in the queue \( N_q \), the effective number of vehicles \( N \), the effective time delay in the queue \( t_q \), the effective time delay \( t_{free} \) for vehicles moving outside the queue (from the beginning of the link to the beginning of the queue), the average travel time across the lane group on the link \( t_{tr} \), the effective inflow rate \( q_{in} \) into the lane group on the link (at the beginning of the link), the effective inflow rate into the queue \( q_{in,q} \) associated with the lane group on the link, the effective lane group outflow \( q_{out, q} \), and all other macroscopic effective traffic variables in the network as functions of time.

In the UTA model, the following vehicle balance equations and relations for each lane group on each link of the network are used:

1. The balance equation for the effective number of vehicles \( N \) within a lane group on a link

\[
\frac{dN}{dt} = q_{in} - q_{out, q} .
\]  

(22.5)

2. In the case of undersaturation, the effective number of vehicles \( N_q \) in the queue is

\[
N_q = q_{in} q^{t_q} ,
\]  

(22.6)

where

\[
t_q = t_{q, \text{min}} + \frac{N_q}{q_{sat} n} ,
\]  

(22.7)

\[
t_{q, \text{min}} = \frac{T_{R}^2}{2T_{\text{light}}} ,
\]  

(22.8)

\( q_{sat} \) is the saturation flow rate per lane, i.e., the maximum flow rate in the related queue outflow per lane; \( n \) is the number of lanes within the lane group on the link. From (22.6) and (22.7) we obtain
In the case of oversaturation, i.e., when the effective number of vehicles in the queue $N_q$ exceeds the effective number of vehicles, which can leave the queue during one green phase, the vehicle balance equation for the effective number of vehicles within the queue $N_q$ associated with the lane group on the link is

$$\frac{dN_q}{dt} = q_{in}, q - q_{out}, q,$$

the effective vehicle waiting time $t_q$ within the queue is found from the equation

$$N_q(t_{free}) = \int_{t_{free}}^{t_{free}+t_q} q_{out}, q(t)dt.$$  

(4) The effective travel time $t_{free}$ is found from the equation

$$L - L_q(t_{free}) = \int_{0}^{t_{free}} v_{free}^{(city)}(\rho(t))dt,$$

the travel time associated with the lane group on the link is

$$t_{tr} = t_{free} + t_q,$$

the average vehicle density outside the queue in (22.12) is

$$\rho(t) = \frac{N(t) - N_q(t)}{(L - L_q(t))n},$$

$L$ is the length of the related lane group on the link, $L_q$ is the effective length of the queue associated with the lane group on the link (in the limiting case $L_q = L$, we obtain $t_{free} = 0$); $v_{free}^{(city)}$ is the effective vehicle speed outside the queue that is a function of the density $\rho$ (22.14),

$$L_q = N_q\bar{d}/n,$$

$\bar{d}$ is the mean front-to-front distance between vehicles within the queue; the effective flow rate $q_{in}, q$ is

$$q_{in}, q = n\rho v_{free}^{(city)}.$$

(5) Furthermore, the UTA model defines the transition from undersaturation to oversaturation by the condition:
\[ q_{\text{in}, q} > q_{\text{out}, q} \]  \hspace{1cm} (22.17)

The transition from oversaturation to undersaturation is defined by the conditions:

\[ q_{\text{in}, q} \leq q_{\text{out}, q} \quad \text{and} \quad N_q \leq q_{\text{in}, q t_q} \],  \hspace{1cm} (22.18)

where \( t_q \) is given by (22.9).

In the UTA model, all values, parameters, and variables in (22.5)-(22.18) as well as \( T_R \) and \( T_G \) can be different for different links and for possible different lane groups on each of the links of the network. Thus, these equations and formulae are applied separately for each lane group on each link of the network.

The effective continuous inflow rate \( q^{(i,m)}_{\text{in}} \) and the effective continuous outflow rate \( q^{(j,k)}_{\text{out}, q} \) associated with different lane groups and links at a road intersection \( p \) (with \( p = 1, 2, \ldots, P \), where \( P \) is the number of road intersections in the network) (Fig. 22.14c) are found from the following balance equations:

\[ q^{(i,m)}_{\text{in}} = \sum_{j,k} \alpha^{(i,m)}_{j,k} q^{(j,k)}_{\text{out}, q} \],  \hspace{1cm} (22.19)

where \( \alpha^{(i,m)}_{j,k} \) are turning ratio coefficients (from a link \( j \), lane group \( k \) to a link \( i \), lane group \( m \)) within the intersection \( p \) of the network that satisfy the obvious condition

\[ \sum_{i,m} \alpha^{(i,m)}_{j,k} = 1 \]  \hspace{1cm} (22.20)

The effective continuous inflow rate \( q^{(i,m)}_{\text{in}} \) and the effective continuous outflow rate \( q^{(j,k)}_{\text{out}, q} \) can be limited. This occurs when queue lengths are almost as long as or equal to the related link lengths for some of the lane groups on links at the intersection \( p \). In the latter case, complex spatiotemporal phenomena of the occurrence and propagation of congestion in a city network can be realized. The following relations take into account these effects in the city network:

\[ q^{(i,m)}_{\text{in}} = \min \left( \frac{N^{(i,m)}_{\text{max}} - N^{(i,m)}}{\Delta t}, \frac{q^{(i,m)}_{\text{in}}}{\bar{q}_{\text{in}}} \right) \]  \hspace{1cm} (22.21)

\[ q^{(j,k)}_{\text{out}, q} = \bar{q}_{\text{out}, q} \sum_{i,m} \alpha^{(i,m)}_{j,k} q^{(i,m)}_{\text{in}} \]  \hspace{1cm} (22.22)

\[ 1 \] The number of lanes \( n^{(i,m)} \) (\( n^{(j,k)} \)) can be different for different lane groups on a link and different links. The mean front-to-front distance \( d^{(i,m)} \) (\( d^{(j,k)} \)) between vehicles within a queue can depend on the percentage of long vehicles within the queue that can be different for different lane groups on the link.
where

\[ q_{\text{in}}^{(i,m)} = \sum_{j,k} \alpha_{j,k}^{(i,m)} q_{\text{out}, q}^{(j,k)} \]

\[ q_{\text{out}, q}^{(j,k)} = \begin{cases} \min(q_{\text{in}}, q_{\text{out}, \text{max}}^{(j,k)}) & \text{for undersaturation} \\ \min(q_{\text{out}, \text{max}}, N_{\text{q}}^{(j,k)})/\Delta t & \text{for oversaturation} \end{cases} \]

\[ q_{\text{out}, \text{max}}^{(j,k)} = n_{\text{sat}}^{(j,k)} q_{\text{sat}} T_{G}^{(j,k)}/T_{\text{light}} \]

\[ T_{\text{light}}^{(j,k)} = T_{G}^{(j,k)} + T_{R}^{(j,k)} \]

\( \Delta t \) is a given discrete time interval;

\[ N_{\text{max}}^{(i,m)} = n_{(i,m)} L^{(i,m)}/d^{(i,m)} \]

is the maximum possible number of vehicles associated with a lane group on a link. The maximum number of vehicles (22.27) is achieved in the limiting case

\[ L_{q}^{(i,m)} = L^{(i,m)} \]

In (22.19)–(22.28), the index \( i \) (with \( i = 1, \ldots, I(p) \), where \( I(p) \geq 1 \)) is related to a link exiting the intersection \( p \) of the network, the index \( m \) (with \( m = 1, \ldots, M^{(i)} \), where \( M^{(i)} \geq 1 \)) is related to one of the lane groups \( M^{(i)} \) on the link \( i \); the index \( j \) (with \( j = 1, \ldots, J(p) \), where \( J(p) \geq 1 \)) is related to a link entering the intersection \( p \), the index \( k \) (with \( k = 1, \ldots, K^{(j)} \), where \( K^{(j)} \geq 1 \)) is related to one of the lane groups \( K^{(j)} \) on the link \( j \). The relations (22.19)–(22.28) should be applied to each of the \( P \) road intersections.

The flow rates \( q_{\text{in}} \) at the network boundary are given based on historical time series. These flow rates can also be outflows from neighboring freeway networks. The flow rates \( q_{\text{out}, q} \) at the network boundary can be inflows onto neighboring freeway networks.

The calculation time with the UTA model can be about several thousand times shorter than the one with microscopic traffic flow models. This is because of two reasons: (i) instead of the calculation of the behavior of each vehicle only numbers of vehicles on links of a city network are considered in the UTA model; (ii) the time scale used in the UTA model is considerably greater than the one in microscopic models because the UTA model goes well beyond the duration of the period of light signals at road intersections.

A travel time prediction \( t_{\text{prog}} = t_{\text{tr}} = t_{\text{free}} + t_{q} \) is performed using a virtual vehicle that has the average vehicle speed \( v_{\text{free}}^{(\text{city})} \) outside the queue and the average vehicle speed

\[ v_{q} = q_{\text{out}, q} d/n \]

within the queue.

Additional simulations of virtual single vehicles, which move within the network in accordance with the mentioned average speed calculated by the
UTA model [325], can be used for a proof of strategies of traffic control and traffic assignment.

In addition, the UTA model can use a correlation between green phases ("green wave") in the network. The model can also include road intersections where rather than light signals other traffic regulation rules are used. In the latter case, $T_R$ and $T_G$ are the duration of some effective red and green phases, respectively. Accordingly, the effective cycle time associated with traffic regulation rules, which correspond to a lane group (direction) on a link at a road intersection, is $T_{eff} = T_R + T_G$. For this effective cycle time $T_{eff}$ the condition (22.3) is also valid, i.e., in (22.2) the value $T$ corresponds to the condition $T > T_{eff}$ [324,551].

The UTA model has the following features:

- **Input parameters:**
  description of the network, traffic light signal programs, historical time series of saturation flow rates, percentages of long vehicles, turning ratios, and incoming flow rates at the boundary of the network.

- **Results of the UTA model:**
  predicted lengths of queues (number of vehicles in queues), travel times associated with links of the network and with any route through the network, waiting times within queues.

In contrast to other model approaches, the UTA model requires parameters, which can be measured directly (no "origin–destination" matrices are required). The macroscopic UTA model is efficient with regard to computational costs because of the concept of effective continuous in space and time variables and parameters. In the UTA model, traffic prediction for about 60 min can be performed during about several seconds for large networks consisting of tens of thousands of links [553]. Thus, about hundred or more cycles of traffic prediction can be repeated in the entire network during the time interval between two successive measurements (the latter is usually 2–10 min). This feature can be used to prove various strategies for vehicle routing guidance systems, traffic management, traffic control, and traffic assignment in the network.

### 22.5 Conclusions

(i) Results obtained with FOTO and ASDA for empirical data measured on various German and American freeways show that for each effectual bottleneck or set of several adjacent effectual bottlenecks, which are close to one another, the spatiotemporal structure of congested patterns exhibits predictable, i.e., characteristic, unique, and reproducible features, such as the types of patterns that are frequently formed. These features can be approximately the same on different days. They can also remain over a wide range of flow rates (traffic demand) at which patterns exist.
These results can be used for congested pattern prediction at freeway bottlenecks.

(ii) The evaluation and online application of the FOTO and ASDA models at the TCC (traffic control center) of the German Federal State of Hessen show great promise for spatiotemporal congested pattern recognition and tracking in all kinds of freeway traffic management and control.

(iii) The characteristic spatiotemporal features of congested patterns can be used to develop a historical database for the basic predictable spatiotemporal structures typical of each effectual bottleneck or for each set of several adjacent effectual bottlenecks on freeways. These historical spatiotemporal series can later be used to match current local measurements and predict the subsequent development of spatiotemporal congested pattern structure. This can improve driver safety in freeway traffic, and to make freeway traffic management and control effective.

(iv) One of the ways for efficient traffic prediction in urban areas is a separate analysis of neighboring freeway and city networks. This is because the congested pattern dynamics in freeway networks associated with self-organized pattern formation due to various driver interaction effects is qualitatively different from those in city traffic. The traffic dynamics in city traffic is governed mostly by light signal operations and other traffic regulations at road intersections rather than driver interaction effects.
23 Control of Spatiotemporal Congested Patterns

23.1 Introduction

Traffic management and control are some of the most important applications of traffic science. There are a huge number of publications and many regular scientific conferences devoted to these subjects (see e.g., references in [21, 27, 286, 486, 495]).

In this chapter, we will use empirical features of spatiotemporal congested patterns discussed in Part II of this book for a theoretical evaluation of some engineering methods of freeway traffic control. This theoretical evaluation of spatiotemporal congested pattern control at freeway bottlenecks will be made using the microscopic three-phase traffic theory of Part III. In particular, we will consider the application of the well-known methods of ramp metering and of automatic cruise control (ACC) for control of spatiotemporal congested patterns. The aim of this pattern control is to find strategies to reduce congested pattern occurrence or to dissolve already existing congested patterns.

However, in this book we will only consider some examples of the application of three-phase traffic theory to freeway traffic control problems. In other words, rather than giving a review of various freeway traffic control methods, the purpose of this chapter is to show the necessity of adequate spatiotemporal congested pattern analysis for freeway traffic control. We will see that a very complex nonlinear spatiotemporal dynamics of congested patterns can occur when congested pattern characteristics are altered by the application of one of these control methods. Therefore, a careful choice of control strategy should be made to achieve the desired control result.

In particular, we compare a standard “free flow control approach” to feedback on-ramp metering whose basic idea is to maintain free flow condition at an on-ramp bottleneck (e.g., [48, 72, 74, 555–573]; see references in [569, 570, 574]) with a “congested pattern control approach” introduced in [575, 576]. In the congested pattern control approach, congestion at the bottleneck is allowed to set in. However, a congested pattern that emerges at the bottleneck should not propagate upstream; rather than propagating upstream the congested pattern should be localized on the main road in a small neighborhood of the bottleneck.
Before we start with this theoretical analysis of the influence of different control strategies on spatiotemporal congested freeway patterns we will briefly discuss possible scenarios for freeway traffic management and control.

### 23.2 Scenarios for Traffic Management and Control

Empirical spatiotemporal congested pattern features at freeway bottlenecks discussed in Part II and the FOTO and ASDA models (Chap. 21) enable us to distinguish features of congested patterns that fundamentally can be influenced in an effective way. Possible management strategies to prevent congested pattern formation or to destroy existing congested patterns can include the following stages (Fig. 23.1):

1. **measurements of traffic**
2. **recognition and prediction of phase transitions in free flow and of congested patterns**
3. **recognition of features of congested patterns that can effectively be influenced with goal of reduction or preventing congested pattern formation**
4. **recognition and prediction of locations where congested patterns initially emerge**
5. **methods of freeway traffic control with aims of reduction or dissolution of congested patterns**
6. **possible instruments for traffic control:**
   - (i) dynamic route guidance
   - (ii) vehicle-vehicle communication
   - (iii) driver assistance systems
   - (iv) ramp metering

![Fig. 23.1. Possible stages of management strategies to reduce congested pattern formation or to destroy existing congested patterns. Taken from [554]](image)

(i) Measurements of traffic.
(ii) Recognition and prediction of phase transitions in free flow and of congested patterns.
(iii) Recognition of features of congested patterns that can effectively be influenced with the goal to prevent congested pattern emergence or to destroy existing patterns.
Applications of methods of traffic flow control with the aims:

1. Prevention of the emergence of congested patterns.
2. Reduction or dissolution of existing congested patterns.

These methods should have an effect on features of congested patterns corresponding to item (iii). In particular, they should affect the dissolution of wide moving jams and spatiotemporal patterns of synchronized flow. This can be possible, if for example, a sensible spatiotemporal combination of a reduction of inflows into congested patterns and an increase in outflows from these patterns were applied. Possible instruments for such methods of traffic control could be dynamic route guidance, vehicle-vehicle communication, driver assistance systems, and ramp metering.

23.3 Spatiotemporal Pattern Control Through Ramp Metering

Ramp metering is a very efficient method of freeway traffic control that is used on freeways in various countries (e.g., [48, 72, 74, 555–573]; see references in [569, 570, 574]).

Below we consider results of a microscopic simulation of spatiotemporal pattern control through ramp metering. This theoretical analysis [575, 576] is based on empirical spatiotemporal pattern features at freeway bottlenecks (Part II) and on the microscopic spatial continuum and discrete-time two-lane model of Sect. 16.3. This model is based on three-phase traffic theory.

On-ramp inflow control at a bottleneck due to the on-ramp should operate when traffic demand in the vicinity of the bottleneck is high enough. This is because at high enough traffic demand it can be expected that speed breakdown with subsequent pattern formation can occur spontaneously at the bottleneck. We compare two qualitatively different approaches to on-ramp inflow control:

(i) **Free flow control approach.** The basic idea of this approach is to maintain free flow conditions at a bottleneck due to the on-ramp. Free flow conditions at the bottleneck should be maintained at the maximum possible throughput in free flow downstream of the bottleneck (e.g., [563–568]; see references in [569, 570, 574]). In other words, the onset of congestion and congested pattern formation at the bottleneck should be prevented by the application of this approach.

There are many different methods based on this approach (see e.g., references in [563–570, 574]). The basic strategy is carried out through automatic control of the flow rate of vehicles that can merge onto the main road from the on-ramp. In some of the methods based on this approach, using detectors for feedback control, traffic variables downstream of the on-ramp merging region are measured (Fig. 23.2a) (e.g., [565–567, 569]).
Results of these measurements are used as feedback for on-ramp inflow control, controlling light signal operation in the on-ramp lane(s). Depending on measurements of current traffic variable, the on-ramp inflow onto the main road from the on-ramp is either restricted or not.

(ii) **Congested pattern control approach.** In this approach, congestion at the bottleneck is allowed to set in [575,576]. The basic idea is to maintain congestion conditions at an on-ramp bottleneck to the minimum possible level. A congested pattern at the bottleneck should not propagate upstream: rather than propagating upstream the congested pattern should be localized at the bottleneck within a relatively small stretch of the main road. As a result, the decrease in vehicle speed within this congested pattern in comparison with free flow conditions should be relatively small. This approach will be called “automatic on-ramp control of congested patterns” (ANCONA for short). Benefits of the ANCONA approach in comparison with the free flow control approach will be considered below. The ANCONA approach should ensure that congestion at the bottleneck does not propagate far upstream of the bottleneck. To achieve this with the ANCONA approach, the detectors for feedback control (Fig. 23.2b) should register congested patterns at the bottleneck. After the congestion condition has been registered, the on-ramp inflow is automatically reduced via light signal operation in the on-ramp lane(s). This leads to localization of the congested pattern on the main road in a small neighborhood of the bottleneck. From results of Sect. 9.2 we know that correct recognition of congested patterns at a bottleneck is only possible if
feedback control detectors are upstream of the effective location of the bottleneck (Fig. 23.2b). Recall that the effective location of the bottleneck is the freeway location in the vicinity of the bottleneck where the downstream front of synchronized flow in the congested pattern is spatially fixed. Within this downstream front vehicles accelerate from synchronized flow upstream of the front to free flow downstream of the front.

### 23.3.1 Free Flow Control Approach

First we consider simulations of ramp metering methods based on the idea of the free flow control approach to maintain free flow conditions upstream of an on-ramp bottleneck (e.g., [563–568]). Free flow conditions should be related to the maximum possible flow rate in free flow downstream of the bottleneck. This basic strategy is carried out through an automatic reduction of the flow rate $q_{\text{on}}^{(\text{cont})}$ of vehicles that can merge onto the main road from the on-ramp in comparison with the flow rate to the on-ramp $q_{\text{on}}$ (Fig. 23.2a).

To achieve this goal, a feedback control method is used (e.g., [563–568, 574]). In some of the methods based on this approach, traffic flow variables are measured downstream of the bottleneck, specifically at the location downstream of the end of the merging region of the on-ramp (Fig. 23.2a). These traffic variables are related to free flow at a flow rate downstream of the bottleneck $q_{\text{sum}}^{(\text{cont})}$:

$$q_{\text{sum}}^{(\text{cont})} = q_{\text{on}}^{(\text{cont})} + q_{\text{in}}.$$  \hspace{1cm} (23.1)

In (23.1), $q_{\text{in}}$ is the flow rate in free flow on the main road upstream of the bottleneck. When traffic demand is low enough,

$$q_{\text{on}}^{(\text{cont})} = q_{\text{on}}.$$  \hspace{1cm} (23.2)

In this case, there should be no on-ramp inflow control. When traffic demand increases, on-ramp inflow control starts to perform. To explain the basic idea

---

1 In Sect. 9.2 it has been noted that the effective location of the bottleneck can be upstream or downstream of the bottleneck. This means that the feedback control detectors can also be upstream or downstream of the bottleneck. The lower the distance between the detectors for feedback control and effective location of the bottleneck, the narrower (in the longitudinal direction) the congested pattern that remains at the bottleneck after the application of the ANCONA method. Thus, the detectors should be upstream but as close as possible to the effective location of the bottleneck. However, in Sect. 9.2 it has also been mentioned that the effective location of the bottleneck can be different for different patterns and on different days (realizations) for the same bottleneck. Therefore, successfully apply the ANCONA method under various congestion conditions, the detectors should be at some “safe” distance upstream of the most probable effective location of the bottleneck. This location can be found by studying the congested pattern on different days.
of on-ramp inflow control, we assume that the on-ramp inflow is reduced when the flow rate downstream of the bottleneck $q_{\text{sum}}^{\text{(cont)}}$ is greater than some optimal flow rate in free flow. We denote this optimal flow rate by $q_{\text{sum}}^{\text{(opt)}}$. One possible on-ramp inflow control rule can be

$$q_{\text{on}}^{\text{(cont)}}(t_k) = q_{\text{on}}^{\text{(cont)}}(t_{k-1}) + K_q \left( q_{\text{sum}}^{\text{(opt)}} - q_{\text{sum}}^{\text{(cont)}}(t_k) \right),$$

where $K_q$ is a given constant, $K_q > 0$; $t_k = T_{av}k$, $k = 1, 2, \ldots$; $T_{av}$ is the averaging time interval for data measured by feedback control detectors in Fig. 23.2a (e.g., $T_{av} = 1$ min). Due to on-ramp inflow control, on average the flow rate of vehicles that can merge onto the main road from the on-ramp is less than the flow rate to the on-ramp:

$$q_{\text{on}}^{\text{(cont)}} < q_{\text{on}}^{\text{(on)}}.$$  

The condition (23.4) is enforced by a light signal in the on-ramp lane(s) (Fig. 23.2a). Thus, vehicles must wait at the light signal before merging onto the main road. To make the inflow from the on-ramp reasonably smooth, many real ramp-metering installations operate by the rule that one vehicle is allowed onto the main road per cycle of the light signal. We will use this common rule of a light signal operation: in all simulations of on-ramp inflow control methods presented below, the light signal in the on-ramp lane(s) switches from red to green at time $t_k$ such that

$$t_k \geq T_m, \text{ where } T_m = T_{m-1} + \frac{3600}{q_{\text{on}}^{\text{(cont)}}(t_k)} \text{ [sec], } m = 1, 2, \ldots,$$

where the closest value of $t_k$ to $T_m$ is chosen. The light signal in the on-ramp lane(s) switches from green to red when a vehicle passes the light signal.

If appropriate values of the optimal flow rate $q_{\text{sum}}^{\text{(opt)}}$ and the constant $K_q$ in (23.3) are chosen, then free flow conditions should be maintained at the bottleneck. In this case, there should also be no congested patterns upstream of the bottleneck.

### Control Strategy ALINEA

In some real ramp-metering installations, rather than the optimal flow rate $q_{\text{sum}}^{\text{(opt)}}$ the optimal occupancy (for the definition of a traffic variable "occupancy" see below) is usually measured downstream of an on-ramp bottleneck for feedback on-ramp inflow control.

An example of this approach is the ALINEA (Asservissement Linéaire d’Entrée Autoroutière) method proposed by Papageorgiou et al. [564–567]. Because this method is widely used in real on-ramp installations in various countries, we will consider it in more detail below. In the ALINEA method, instead of the control rule (23.3) the following rule is used [564–567]:
23.3 Spatiotemporal Pattern Control Through Ramp Metering

\[ q^{(\text{cont})}(t_k) = q^{(\text{cont})}(t_{k-1}) + K_o \left( o^{(\text{opt})}_{\text{sum}} - o^{(\text{cont})}_{\text{sum}}(t_k) \right) , \]  

(23.6)

where \( K_o \) is a given constant; \( t_k = T_{av}k, \ k = 1, 2, \ldots \); \( o^{(\text{opt})}_{\text{sum}} \) is the optimal occupancy; \( o^{(\text{cont})}_{\text{sum}}(t_k) \) is the occupancy measured at the feedback control detectors (Fig. 23.2a). This occupancy is averaged over the time interval

\[ t_{k-1} < t \leq t_k . \]  

(23.7)

When traffic demand is low enough, the condition (23.2) is satisfied. In this case, there should be no on-ramp inflow control. When traffic demand increases, on-ramp inflow control starts to perform. Due to on-ramp inflow control, the flow rate of vehicles that can merge onto the main road from the on-ramp is lower than the flow rate to the on-ramp, i.e., the condition (23.4) is satisfied.

Note that the traffic variable occupancy is (e.g., [21])

\[ o = \frac{T_{\text{veh}}}{T_{av}} \times 100\% , \]  

(23.8)

where \( T_{\text{veh}} \) is the sum of the time intervals when detectors have measured vehicles during the time interval \( T_{av} \). As mentioned in Sect. 2.3.1, an induction loop registers a vehicle \( i \) moving on the freeway, producing a pulse of electric current that begins at some time \( t_{i,b} \) when the vehicle reaches the induction loop, and it ending some time later \( t_{i,f} \) when the vehicle leaves the induction loop (2.1). Thus, the time \( T_{\text{veh}} \) is equal to the sum of the duration of current pulses for all vehicles measured by the detector during time interval \( T_{av} \):

\[ T_{\text{veh}} = \sum_{i=1}^{\Delta N_{av}} \Delta t_i , \]  

(23.9)

where \( \Delta t_i \) is defined via (2.1), \( \Delta N_{av} \) is the number of vehicles that have passed the detector during time interval \( T_{av} \). The occupancy is related to the vehicle density in accordance with the formula

\[ \rho = 100^{-1} \frac{o}{d_{av} + L_{det}} , \]  

(23.10)

where \( d_{av} \) is the average vehicle length during the time interval \( T_{av} \), and \( L_{det} \) is the length of the detector that is constant.

For a numerical analysis of the on-ramp inflow control strategies we will use the two-lane model of Sect. 16.3. In this two-lane model, all vehicles have the same length. In this case, we can assume that \( d_{av} + L_{det} = d \), where \( d \) is the vehicle length that includes the minimum space gap between two successive vehicles stopped in a wide moving jam.

For free flow conditions at the bottleneck we have

\[ q = v_{\text{free}} \rho = \frac{o v_{\text{free}}}{100d} , \]  

(23.11)
where \( v_{\text{free}} \) and \( d \) are constants (Sect. 16.3.1). From (23.11) we see that the control rules (23.3) and (23.6) for free flow conditions are equivalent when all vehicles have the same parameters and

\[
q_{\text{sum}}^{(\text{opt})} = o_{\text{sum}}^{(\text{opt})} \frac{v_{\text{free}}}{100d}.
\]  

(23.12)

It must be noted that (23.11) and (23.12) are valid only when the flow rate is an increasing function of occupancy, specifically for free flow conditions at the bottleneck.

**Features of On-Ramp Inflow Control with ALINEA Method**

In empirical observations, the most frequent type of congested pattern at an isolated bottleneck due to the on-ramp is a general pattern (GP) (Sects. 2.4.7 and 9.4).

Let us assume that an GP emerges at an on-ramp bottleneck where first there is no on-ramp inflow control (Fig. 23.3a). In Fig. 23.3, the flow rate \( q_{\text{in}} \) is constant but the flow rate to the on-ramp \( q_{\text{on}} \) is a function of time (Fig. 23.3b):

\[
q_{\text{on}}(t) = \begin{cases} 
0 & \text{for } t < t_0, \\
200 + 700(t - t_0)/3600 \text{ vehicles/h} & \text{for } t_0 \leq t < t^{(1)}, \\
900 \text{ vehicles/h} & \text{for } t \geq t^{(1)}.
\end{cases}
\]  

(23.13)

In (23.13), as in the microscopic model of Sect. 16.3, the discrete time \( t = n\tau \) is used, \( n = 0, 1, 2, \ldots; \tau \) is the time step; \( t_0 = 7 \text{ min}, t^{(1)} = 67 \text{ min}.^2 \)

After the GP has emerged upstream of the bottleneck, the discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \) at \( t \geq t^{(1)} \) is 2100 vehicles/h per lane (Fig. 23.3e). This flow rate is appreciably lower than the maximum flow rate \( q_{\text{max}}^{(\text{free})} \approx 2430 \text{ vehicles/h per lane on a homogeneous road that corresponds to the chosen model parameters (see Table 16.11 of Sect. 16.3). Recall that the discharge flow rate } q_{\text{out}}^{(\text{bottle})} \text{ is related to the outflow from a congested pattern at a bottleneck. This is precisely the flow rate in free flow on the main road downstream of the congested bottleneck.}^3 \text{ The onset of the GP leads to an increase in travel}

\(^2\) The maximum flow rate \( q_{\text{on}} = 900 \text{ vehicles/h} \) in (23.13) is chosen to be considerably lower than the maximum capacity of free flow (without light signal operation) in the on-ramp lane. Simulations show that when the flow rate \( q_{\text{in}} \) is small enough, no congestion occurs in the on-ramp lane without light signal operation at the flow rate \( q_{\text{on}} = 900 \text{ vehicles/h} \), specifically vehicles move at their maximum possible speed in free flow in the on-ramp lane.

\(^3\) As the discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \), the flow rate \( q_{\text{sum}}^{(\text{cont})} \) (23.1) is also the flow rate in free flow downstream of the bottleneck. However, the flow rate \( q_{\text{sum}}^{(\text{cont})} \) is associated with free flow conditions at the bottleneck. In contrast, the flow rate \( q_{\text{out}}^{(\text{bottle})} \) corresponds to the outflow rate from the congested bottleneck, i.e., when a congested pattern is at the bottleneck.
Fig. 23.3. GP characteristics without on-ramp inflow control. (a) Vehicle speed on the main road in space and time. (b) Flow rate to the on-ramp $q_{on}$ as a function of time (23.13). (c) Travel time on the main road as a function of time. (d) Travel time in the on-ramp lane as a function of time. (e) Discharge flow rate $q_{out}^{(bottle)}$ as a function of time. (f) Flow rate $q_{on}^{(on)}$ as a function of time. $q_{in} = 1946$ vehicles/h.

The length of the road is 20 km, the beginning of the merging region of the on-ramp is located at $x_{on} = 10$ km, the length of merging region of the on-ramp is $L_m = 300$ m. The length of the road to the on-ramp is $L_r = 3$ km, the travel time in the on-ramp lane is measured over a 3 km long stretch of the on-ramp between $x_{on}^{(on)} = 7$ km and $x_{on} = 10$ km, the travel time on the main road is measured over a 10 km long stretch between $x = 1$ km and $x = 11$ km. Taken from [576]
time on the main road (Fig. 23.3c) and in the on-ramp lane (Fig. 23.3d). The increase in travel time in the on-ramp lane is associated with the occurrence of an GP also in the on-ramp lane. This effect has been considered in Sect. 18.3.2. After the GP forms in the on-ramp lane, the flow rate of vehicles that can merge onto the main road \( q^{(on)} \) (Fig. 23.3f) is lower than the flow rate \( q_{on} \) (Fig. 23.3b). This decrease in flow rate \( q^{(on)} \) occurs without any on-ramp control due to the onset of congestion in the on-ramp lane.

![Graphs and diagrams showing vehicle speed, flow rate, and travel time on the main road and on-ramp lane.](image)

**Fig. 23.4.** Suppression of GP emergence due to on-ramp inflow control with the ALINEA method. (a) Vehicle speed on the main road in space and time. (b) Flow rate to the on-ramp \( q_{on} \) as a function of time (23.13). (c) Travel time on the main road as a function of time. (d) Travel time in the on-ramp lane as a function of time. (e) Flow rate \( q_{sum}^{(cont)} \) as a function of time. (f) Flow rate \( q_{on}^{(cont)} \) as a function of time. In (23.6) \( T_{av} = 1 \) min, \( K_o = 70 \) vehicles/h, \( o_{sum}^{(opt)} = 14.8\% \) \( q_{sum}^{(opt)} = 2130 \) vehicles/h per lane). The light signal is at the location in the on-ramp lane 100 m upstream of the beginning of the merging region of the on-ramp \( x = x_{on} = 10 \) km. The location of the detectors for feedback control is \( x_{1}^{(det)} = 11 \) km. Other parameters are the same as those in Fig. 23.3. Taken from [576]
Due to on-ramp inflow control with the ALINEA method, the GP does not appear if under the same parameters as those in Fig. 23.3 the flow rate to the on-ramp $q_{on}$ increases over time (Fig. 23.4).\(^4\) This effect of the control ALINEA method (23.6) is associated with a decrease in the flow rate $q_{on}^{(cont)}$ of vehicles merging onto the main road from the on-ramp (Fig. 23.4f) in comparison with the higher flow rate $q_{on}$ (Fig. 23.4b). This decrease in flow rate $q_{on}^{(cont)}$ is due to light signal operation in the on-ramp lane. In numerical simulations, on-ramp inflow control related to (23.6) starts to perform after the condition $o_{sum}^{(cont)} > o_{sum}^{(opt)}$ is satisfied for the first time at some time $t_k = t_b$. At this time the flow rate $q_{on}^{(cont)}(t_b)$ is set to $q_{on}(t_b)$, specifically, to the flow rate measured at $t_k = t_b$ at a detector in the on-ramp lane at $x = x_{on} = 10$ km.

Simulations show that the flow rate $q_{on}^{(cont)}$ is maintained by the light signal to have the occupancy $o_{sum}^{(cont)}$ downstream of the bottleneck at the location of the detectors for feedback control approximately equals the chosen optimal occupancy $o_{sum}^{(opt)}$, i.e., $o_{sum}^{(cont)} \approx o_{sum}^{(opt)}$.

For choosing the optimal location of the detectors for feedback control in the ALINEA method two different detector locations have been compared. The first is $x_{1}^{(det)} = 11$ km, i.e., 700 m downstream of the end of the merging region of the on-ramp (Fig. 23.4). The second detector location is $x_{2}^{(det)} = 10.4$ km, i.e., 100 m downstream of the end of the merging region of the on-ramp. It should be noted that the downstream front of synchronized flow in the GP in Fig. 23.3a is fixed at $x_{bottle}^{(det)} \approx 10.45-10.5$ km, i.e., 150-200 m downstream of the end of the merging region of the on-ramp. The location of this front determines the effective location of the bottleneck $x_{eff}^{(bottle)}$. Thus, for the first case $x_{1}^{(det)} > x_{eff}^{(bottle)}$, i.e., the feedback control detectors are downstream of the effective location of the bottleneck.\(^5\)

In contrast, for the second case $x_{2}^{(det)} < x_{eff}^{(bottle)}$, i.e., the feedback control detectors are upstream of the effective location of the bottleneck. In the latter case, when the feedback control detectors are upstream of the effective location of the bottleneck when the GP is just emerging, synchronized flow first occurs at the detectors for feedback control. We can expect that the ALINEA method should suppress this synchronized flow. This is because the occupancy in synchronized flow should be much greater than the optimal occupancy associated with free flow.

\(^4\) Because in this model there are no long vehicles, the optimal occupancy $o_{sum}^{(opt)}$ (23.6) is chosen lower than the optimal occupancy that is usually used in real ramp-metering installations (the latter is about 18–32%). However, the optimal flow rate $q_{sum}^{(opt)}$ in (23.12) is related to real empirical values.

\(^5\) Simulations show that the conclusions about GP suppression via on-ramp inflow control with the ALINEA method made above for the detector location $x_{1}^{(det)}$ are also valid for other locations $x_{(det)}^{(det)}$ of the detectors for feedback control in the range $x_{eff}^{(bottle)} < x_{(det)}^{(det)} < x_{1}^{(det)}$. 

This suppression effect is indeed realized. When the detectors for feedback control are at the location \( x^{(\text{det})}_{2} \), then the ALINEA method (23.6) can maintain free flow conditions at the bottleneck. As a result, in this case approximately the same vehicle speed distribution as those shown in Fig. 23.4a is realized. However, from Fig. 23.5a it can be seen that the flow rate \( q_{\text{on}}^{\text{(cont)}} \) begins to decrease earlier when detector location \( x^{(\text{det})}_{2} \) is used (curve 2). Moreover, the travel time in the on-ramp lane is greater, when detector location \( x^{(\text{det})}_{2} \) is used (Fig. 23.5b). This means that in the ALINEA method the detector location \( x^{(\text{det})}_{1} \) downstream of the effective location of the bottleneck is more conducive to free flow conditions at the bottleneck at the maximum possible throughput downstream of the bottleneck.

Fig. 23.5. Comparison of GP suppression via on-ramp inflow control with the ALINEA method at two different locations \( x^{(\text{det})}_{1} = 11 \text{ km} \) and \( x^{(\text{det})}_{2} = 10.4 \text{ km} \) of the detectors for feedback control. (a) Flow rates \( q_{\text{on}}^{\text{(cont)}} \) as a function of time. (b) Travel times in the on-ramp lane as functions of time. Curves 1 for \( x^{(\text{det})}_{1} \), curves 2 for \( x^{(\text{det})}_{2} \). Other parameters are the same as those in Fig. 23.4. Taken from [576]

We can conclude that benefits of the ALINEA method (23.6) are as follows:

(i) If the appropriate optimal occupancy is chosen, congested pattern occurrence can be prevented by a reduction of on-ramp inflow (Fig. 23.4a).

(ii) There is no increase in travel time on the main road over time (Fig. 23.4c).

However, from Fig. 23.4d we can see a disadvantage of this control method: the travel time of vehicles in the on-ramp lane is a abruptly increasing function of time. This is because vehicles in the on-ramp lane must wait at the light signal before they merge onto the main road from the on-ramp. This waiting time increases over time because the flow rate to the on-ramp \( q_{\text{on}} \) is higher than the flow rate of vehicles allowed to merge onto the main road \( q_{\text{on}}^{\text{(cont)}} \).

Thus, the methods under consideration maintain free flow conditions at an on-ramp bottleneck. However, this results in an increase of congestion on other sections of a traffic network. There are two consequences of this effect:
(a) The onset of congestion in a city network if the main road is connected to the city network via this on-ramp.

(b) The onset of congestion upstream of a bottleneck due to the off-ramp on another freeway of the related freeway network. This occurs if the on-ramp lane is also the off-ramp lane of the latter freeway in this freeway network.

One can try to solve this problem by increasing the optimal occupancy. The result of this increase is shown in Fig. 23.6. It can be seen that even a relatively small increase in optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ leads to the spontaneous occurrence of an GP. This GP remains further upstream of the bottleneck. This result of numerical simulations of the ALINEA method can be explained as follows:

1. Let us consider an GP that has already appeared at the bottleneck. If the optimal location of detectors for feedback control is chosen downstream of the effective location of the bottleneck, then the detectors are in free flow. The flow rate in this free flow is equal to the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ from the GP. The discharge flow rate is usually lower than the flow rate $q_{\text{sum}}^{(\text{opt})}$ associated with the optimal occupancy $o_{\text{sum}}^{(\text{opt})}$. For this reason, the ALINEA control rule (23.6) does not lead to GP dissolution: the flow rate $q_{\text{on}}^{(\text{cont})}$ is not sufficiently reduced to prevent GP emergence (Fig. 23.6f). Even if the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ is slightly higher than $q_{\text{sum}}^{(\text{opt})}$ the GP does not dissolve. This corresponds to the empirical result that for GP dissolution a considerable decrease in the flow rate $q_{\text{on}}^{(\text{cont})}$ and/or the flow rate $q_{\text{in}}$ is required (Chap. 13).

2. The result in item (1) is associated with another nonlinear effect of GP formation. This is the metastability effect in GP formation: when an GP has formed at the bottleneck at some flow rate $q_{\text{on}}^{(\text{cont})}$, it can survive at the bottleneck even if the flow rate $q_{\text{on}}^{(\text{cont})}$ is reduced. Only when the flow rate $q_{\text{on}}^{(\text{cont})}$ is reduced below some threshold flow rate, can the GP dissolve. This metastability effect in GP emergence and existence is observed both in empirical results (Sect. 13.2) and in a theory of GPs (Sect. 18.5.2).

The above metastability effect in GP emergence and existence leads to another disadvantage of the methods under consideration. If a short-term increase in the flow rate $q_{\text{on}}$ occurs, then due to the time delay of the ALINEA control method a short-term increase in the flow rate $q_{\text{on}}^{(\text{cont})}$ appears. This pulse in the flow rate plays the role of a nucleus for GP emergence at the bottleneck. Let us choose the optimal detector location $x_{1}^{(\text{det})} = 11$ km and the optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ in the control rule (23.6) to be the same as those in Fig. 23.4. In the latter case, the ALINEA method can prevent GP emergence. In contrast, if at the same optimal occupancy a pulse in the flow rate $q_{\text{on}}$ appears, then an GP is induced by this pulse at the bottleneck (Fig. 23.7a,b).
Fig. 23.6. GP emergence under on-ramp inflow control with the ALINEA method. (a) Vehicle speed on the main road in space and time. (b) Flow rate to the on-ramp \(q_{\text{on}}\) as a function of time (23.13). (c) Travel time on the main road as a function of time. (d) Travel time in the on-ramp lane as a function of time. (e) Flow rate \(q_{\text{out}}^{(\text{bottle})}\) as a function of time. (f) Flow rate \(q_{\text{on}}^{(\text{cont})}\) as a function of time. In (23.6) \(\phi_{\text{sum}}^{(\text{opt})} = 15.5\%\) \((q_{\text{sum}}^{(\text{opt})} = 2230\) vehicles/h per lane). Other parameters are the same as those in Figs. 23.3 and 23.4. \(x_1^{(\text{det})} = 11\) km. Taken from [576]

This is due the above metastability effect in GP formation. Simulation results show that when the flow rate to the on-ramp very slowly increases over time (Fig. 23.4b), the ALINEA method can prevent GP emergence at the bottleneck. If, however, a pulse in the flow rate to the on-ramp occurs (Fig. 23.7b), then with all the parameters of the control method, including optimal occupancy, remaining the same, the ALINEA control method cannot
23.3 Spatiotemporal Pattern Control Through Ramp Metering

Fig. 23.7. GP emergence under on-ramp inflow control with the ALINEA method due to a short-term pulse in the on-ramp inflow. (a) Vehicle speed on the main road in space and time. (b) Flow rate to the on-ramp \( q_{on} \) as a function of time. (c) Travel time on the main road as a function of time. (d) Travel time in the on-ramp lane as a function of time. (e) Flow rate \( q_{out}^{(bottle)} \) as a function of time. (f) Flow rate \( q_{on}^{(cont)} \) as a function of time. In (b) a pulse of the flow rate to the on-ramp is applied to the function \( q_{on}(t) \) (23.13): the flow rate is increased to 1050 vehicles/h during 8 min. Other parameters are the same as those in Figs. 23.3 and 23.4. Taken from [576]
After the GP has emerged, this method at the chosen simulation parameters does not lead to GP dissolution. This is explained by the metastability effect of GP emergence and existence.

23.3.2 Congested Pattern Control Approach

To overcome the above disadvantages of the ALINEA control method, which are also characteristic for other methods based on the free flow control approach, the ANCONA approach has been introduced [575, 576]. In the ANCONA approach, congested patterns are allowed to occur at the bottleneck. However, when a congested pattern is forming at the bottleneck the free flow rate of vehicles, which can merge onto the main road from the on-ramp, is reduced to some flow rate \( q^{(\text{cont})}_{\text{on}} \) that is lower than the flow rate to the on-ramp \( q_{\text{on}} \). As a result, the congested pattern is localized within a small region on the main road in the neighborhood of the bottleneck. This means that the congested pattern does not propagate further upstream of the on-ramp bottleneck.

It turns out that SPs can be more favorable than GPs in terms of the discharge volume from a congested pattern (Sect. 18.8) and of the vehicle time delay due to congestion. This is used in the ANCONA approach. After the onset of congestion has occurred spontaneously at an on-ramp bottleneck, through the use of the ANCONA approach a localized SP (LSP) with the maximum possible average speed in synchronized flow should be formed at the bottleneck.

Benefits of the ANCONA approach are considerably greater throughput and considerably shorter travel times in the on-ramp lane(s) than in the ALINEA control method.

The ANCONA approach [575, 576] is carried out due to feedback from a location on the freeway where congestion conditions have been registered. The onset of congestion is registered based on loop detector traffic measurements. As mentioned above, these detectors for feedback control should be upstream of the effective location of the bottleneck.

In the ANCONA approach, there is no on-ramp control as long as free flow conditions are measured at the bottleneck. On-ramp inflow control is first realized only after the onset of congestion has occurred spontaneously at the bottleneck. In other words, feedback control is performed when the average speed \( v^{(\text{det})} \) measured at the detectors for feedback control (Fig. 23.2b) is equal to or drops below a chosen “congestion speed” \( v_{\text{cong}} \):

\[
   v^{(\text{det})} \leq v_{\text{cong}} .
\]  

When this condition is satisfied, the flow rate \( q^{(\text{cont})}_{\text{on}} \) is reduced via light signal operation. The decrease in this flow rate should lead to an increase in the

---

6 Qualitatively the same effect of GP emergence is realized if a short-term increase in the flow rate \( q_{\text{in}} \) occurs on the main road.
speed \( v^{(\text{det})} \) above the congestion speed \( v_{\text{cong}} \):

\[
v^{(\text{det})} > v_{\text{cong}} .
\]  

When the condition (23.15) is satisfied, higher flow rate \( q_{\text{on}}^{(\text{cont})} \) is allowed via light signal operation. If under this higher flow rate \( q_{\text{on}}^{(\text{cont})} \) the onset of congestion occurs at the bottleneck once more, an incipient congested pattern begins to propagate upstream of the bottleneck. As a result, the speed at the detector for feedback control decreases. Therefore, the condition (23.14) is satisfied once more. This leads to a new decrease in the flow rate \( q_{\text{on}}^{(\text{cont})} \), and so on.

Thus, the basic idea of this approach is to allow the congestion condition at the bottleneck at the minimum possible level of traffic congestion. The aims of the ANCONA approach are as follows:

1. Achievement of higher throughput in the whole traffic network under very high traffic demand when congestion has to occur somewhere in the traffic network. This is realized due to a more homogeneous distribution of congestion among all bottlenecks in the network.
2. Maintenance of shorter travel time of vehicles at the light signal in the on-ramp lane(s) and in the entire network.
3. Prevention of upstream propagation of a congested pattern that occurs at the bottleneck due to the metastability of congested pattern formation.

One possible realization of the ANCONA approach can be formulated as follows [576]:

\[
q_{\text{on}}^{(\text{cont})}(t_k) = \begin{cases} 
q_{\text{on}1}(v^{(\text{det})}(t_k)) & \text{if } v^{(\text{det})}(t_k) \leq v_{\text{cong}} \\
q_{\text{on}2}(\Omega(t_k)) & \text{if } v^{(\text{det})}(t_k) > v_{\text{cong}} 
\end{cases},
\]  

where \( t_k = kT_{\text{av}}, \ k = 1, 2, \ldots \); \( v^{(\text{det})}(t_k) \) is the vehicle speed measured at the detector that is averaged during the time interval \( t_{k-1} < t \leq t_k \); \( q_{\text{on}1}(v^{(\text{det})}) \) is an increasing function of the speed \( v^{(\text{det})} \); \( q_{\text{on}2}(\Omega) \) is an increasing function of the value \( \Omega \) that is a function of time \( t_k \), \( \Omega \) is the number of the time intervals of the data averaging where the condition (23.15) has been satisfied without interruption. If the condition (23.15) is satisfied and \( q_{\text{on}2}(\Omega(t_k)) \geq q_{\text{on}} \), then the light signal in the on-ramp lane(s) is permanently switched to the green phase.\(^7\)

\(^7\) Note that instead of the congestion speed \( v_{\text{cong}} \), a chosen “congestion occupancy,” \( o_{\text{cong}} \), can be used for the detection of a congested pattern. In this case, (23.14)–(23.16) should be replaced by

\[
o^{(\text{det})} \geq o_{\text{cong}} ,
\]

and

\[
o^{(\text{det})} < o_{\text{cong}} ,
\]
To explain the function $q_{\text{on1}}(v^{(\text{det})})$, note that the lower the measured speed in a congested pattern $v^{(\text{det})}$, the lower the flow rate $q_{\text{on1}}^{(\text{cont})}$ should be. For this reason, an increasing function $q_{\text{on1}}(v^{(\text{det})})$ in (23.16) should be chosen.

To explain the function $q_{\text{on2}}(\Omega(t_k))$, we assume that during the time interval

$$(k - \Omega)T_{\text{av}} < t \leq kT_{\text{av}},$$

where $\Omega \geq 1$, the condition (23.15) has been satisfied. This means that free flow has been measured at the detectors for feedback control during the time interval (23.20). The longer the time interval of free flow conditions (23.15) at the detectors, the higher the flow rate $q_{\text{on1}}^{(\text{cont})}$ allowed. This is reflected in the increasing function $q_{\text{on2}}(\Omega(t_k))$.

We can see that in the ANCONA approach through the use of on-ramp inflow control there can be two nonlinear transitions between two qualitatively different traffic states (Fig. 23.2b). One of these states is related to the “synchronized flow” phase. The second state is related to the “free flow” phase. Synchronized flow occurs due to a spontaneous $F \rightarrow S$ transition at the bottleneck. In contrast, an $S \rightarrow F$ transition is induced by light signal operation when the condition (23.14) is satisfied. When synchronized flow does not occur during a long enough time interval, there is no on-ramp control any more.

In numerical simulations of this approach, for simplification we assume that $q_{\text{in}}$ is a given constant and the flow rate $q_{\text{on}}(t)$ is not a decreasing function of time. This enables us to use the following simplified on-ramp inflow control rules:

$$q_{\text{on}}^{(\text{cont})}(t_k) = \begin{cases} q_{\text{on1}} & \text{if } v^{(\text{det})}(t_k) \leq v_{\text{cong}} \\ q_{\text{on2}} & \text{if } v^{(\text{det})}(t_k) > v_{\text{cong}} \end{cases},$$

where $q_{\text{on1}}$ and $q_{\text{on2}}$ are constants ($q_{\text{on2}} > q_{\text{on1}}$).  

In Fig. 23.8, the same flow rate $q_{\text{in}}$ and dependence of the flow rate $q_{\text{on}}$ on time as those in Fig. 23.3 are used. It can be seen that due to the ANCONA

$$q_{\text{on}}^{(\text{cont})}(t_k) = \begin{cases} q_{\text{on1}}(o^{(\text{det})}(t_k)) & \text{if } o^{(\text{det})}(t_k) \geq o_{\text{cong}} \\ q_{\text{on2}}(\Omega(t_k)) & \text{if } o^{(\text{det})}(t_k) < o_{\text{cong}} \end{cases},$$

respectively. In (23.17)–(23.19), $o^{(\text{det})}$ is the average occupancy measured at the detectors for feedback control (Fig. 23.2b), $q_{\text{on1}}(o^{(\text{det})})$ is a decreasing function of $o^{(\text{det})}$.

Note that if instead of the congestion speed $v_{\text{cong}}$, a chosen congestion occupancy $o_{\text{cong}}$ is used for the detection of a congested pattern (see (23.17)–(23.19)), then the control rules (23.21) should be replaced by

$$q_{\text{on}}^{(\text{cont})}(t_k) = \begin{cases} q_{\text{on1}} & \text{if } o^{(\text{det})}(t_k) \geq o_{\text{cong}} \\ q_{\text{on2}} & \text{if } o^{(\text{det})}(t_k) < o_{\text{cong}} \end{cases}. $$

8 Note that if instead of the congestion speed $v_{\text{cong}}$, a chosen congestion occupancy $o_{\text{cong}}$ is used for the detection of a congested pattern (see (23.17)–(23.19)), then the control rules (23.21) should be replaced by
23.3 Spatiotemporal Pattern Control Through Ramp Metering

Fig. 23.8. GP suppression due to application of the ANCONA method. (a) Vehicle speed on the main road in space and time. (b) Flow rate to the on-ramp $q_{on}$ as a function of time (23.13). (c) Travel time on the main road as a function of time. (d) Travel time in the on-ramp lane as a function of time. (e) Speed (left) and flow rate (right) as functions of time. (f) Flow rate $q_{out}^{(bottle)}$ as a function of time. (g) Flow rate $q_{on}^{(cont)}$ as a function of time. The control method (23.21) is used where $v_{cong} = 85$ km/h, $q_{on1} = 360$ vehicles/h, and $q_{on2} = 600$ vehicles/h, $T_{av} = 1$ min. The detectors for feedback control are located at $x_{3}^{(det)} = 9.8$ km on the main road, i.e., 200 m upstream of the beginning of the on-ramp merging region $x_{on} = 10$ km (Fig. 23.2b). Other parameters are the same as those in Figs. 23.3 and 23.4. Taken from [576]
method (23.21) the GP that arose in Fig. 23.3 at the same parameters has not been formed. Instead of the GP only an LSP appears on the main road at the on-ramp bottleneck. Synchronized flow in this LSP is related to the average speed that is only slightly lower than 85 km/h (Fig. 23.8e). It is important that the congestion region is localized on a small stretch of the freeway in the vicinity of the bottleneck: there is no further upstream propagation of congestion after the LSP has been formed at the bottleneck (Fig. 23.8a).

The congested pattern in the ANCONA method (23.21) is localized due to the automatic switching from a higher flow rate \( q_{on}^{(cont)} \) to a lower flow rate \( q_{on}^{(cont)} \) (Fig. 23.8g). The higher flow rate is allowed when the congested pattern almost dissolves, specifically when the condition (23.15) is satisfied. The lower flow rate is related to congested pattern growth, specifically when the condition (23.14) is satisfied. This control method ensures both the spatial localization of congestion and the highest possible throughput.

Simulations show that the ANCONA model (23.21) is not very sensitive to specific values of the model parameters \( v_{cong}, q_{on1}, \) and \( q_{on2} \). However, to avoid moving jam emergence in synchronized flow of an LSP on the main road at the bottleneck, the speed, which is realized within the LSP at chosen model parameters, should not be very low.

### 23.3.3 Comparison of Free Flow and Congested Pattern Control Approaches

A comparison of the two approaches presented in Figs. 23.4 and 23.8 is shown in Fig. 23.9. It can be seen that the LSP, which is maintained at the bottleneck in the ANCONA method, results in a very small increase in travel time on the main road in the ANCONA method in comparison with the ALINEA method (Fig. 23.9c). The difference between these travel times on the main road is so small as to be almost indistinguishable (curves 1 and 2).

Although the congested pattern remains at the bottleneck in the ANCONA method (Fig. 23.9a), nevertheless at the same flow rate \( q_{in} \) and the same time dependence of the flow rate \( q_{on} \) we find a higher flow rate downstream of the bottleneck in the ANCONA method than in the ALINEA method. In the ALINEA method, the average flow rate \( q_{sum}^{(cont)} = 2130 \) vehicles/h and in the ANCONA method the average flow rate \( q_{out}^{(bottle)} = 2190 \) vehicles/h. This is because the flow rate \( q_{on}^{(cont)} \) in the ALINEA method (Fig. 23.9b) is 360 vehicles/h whereas the mean flow rate \( q_{on}^{(cont)} \) in the ANCONA method (Fig. 23.9a) is 470 vehicles/h. As a result, the travel time of vehicles in the on-ramp lane(s) is considerably greater in the ALINEA method (curve 2 in Fig. 23.9d) than in the ANCONA method (curve 1).

Another benefit of the ANCONA method in comparison with the ALINEA method can be seen in Fig. 23.10. When a pulse of the flow rate to the on-ramp appears, the ALINEA method cannot prevent the formation and dissolution of the GP at the bottleneck (Fig. 23.7). In contrast, the ANCONA method
leads to the dissolution of this GP at the same flow rate \( q_{in} \) and the same time dependence of the flow rate \( q_{on} \) (Fig. 23.10). The ANCONA method enables us to suppress other types of congested patterns propagating upstream that can occur due to the metastability of traffic flow with respect to congested pattern formation at the bottleneck. Rather than a congested pattern, which propagates continuously upstream, an LSP remains at the bottleneck, whose width (in the longitudinal direction) is spatially limited. As a result, the influence of this congested pattern at the on-ramp bottleneck on the travel time on the main road is negligible.

We have mentioned that if the location of the detectors for feedback control in the ALINEA method is chosen upstream of the effective location of the bottleneck \( (x_2^{(det)} = 10.4 \text{ km}) \), this method can suppress congestion at the bottleneck. However, this detector location is not optimal. The throughput is lower than for the downstream detector location \( x_1^{(det)} = 11 \text{ km} \) (Fig. 23.5). Nevertheless, let us compare the ANCONA method with the ALINEA method where the non-optimal detector location \( x_2^{(det)} = 10.4 \text{ km} \) is used. This comparison is interesting because when \( x_2^{(det)} = 10.4 \text{ km} \), the
ALINEA method suppresses an GP that occurs due to a pulse in the flow rate to the on-ramp. This GP remains when the ALINEA method with optimal detector location $x_1^{(det)} = 11 \text{ km}$ is used (Fig. 23.7).

However, if the non-optimal detector location $x_2^{(det)} = 10.4 \text{ km}$ is used, which is upstream of the effective location of the bottleneck, the ALINEA method maintains free flow conditions at the bottleneck. Results of comparison of the ALINEA method in this case with the ANCONA method are shown in Fig. 23.11. It can be seen that in the ANCONA method (curve 1 in Fig. 23.11a) the flow rate $q_{on}^{(cont)}$ is higher than in the ALINEA method (curve 2). The average flow rate $q_{on}^{(cont)}$ in the ANCONA method is 400 vehicles/h, whereas it is 300 vehicles/h in the ALINEA method.
Accordingly, the travel time in the on-ramp lane (waiting time at the light signal) in the ANCONA method (curve 1 in Fig. 23.11b) is considerably smaller than the travel time in the ALINEA method (curve 2).

We can summarize the comparison of these two approaches to on-ramp inflow control. Benefits of the ANCONA method in comparison with the ALINEA method are:

(a) higher throughputs on the main road and in the on-ramp lane(s);
(b) considerably lower vehicle waiting times at the light signal in the on-ramp lane(s);
(c) upstream propagation of congestion on the main road does not occur even if a congested pattern occurs at the bottleneck: the congested pattern is spatially localized at the bottleneck.

These benefits of the ANCONA method can be explained based on the diagram of congested patterns (Fig. 8.1a) and theory of freeway capacity at an on-ramp bottleneck (Chap. 8). However, we should take into account that under on-ramp metering, in the diagram in Fig. 8.1a and formulae of Chap. 8

\[ q_{on} = q_{on}^{(cont)} \] \[ q_{sum} = q_{sum}^{(cont)} \]

Firstly, recall that there are an infinite number of maximum freeway capacities \( q_{max}^{(free)} \) (8.6) and an infinite number of minimum freeway capacities \( q_{th}^{(B)} \) (8.13). At each of the maximum freeway capacities, the probability \( P_{FS}^{(B)} \) for speed breakdown at the bottleneck during a given time interval \( T_{ob} \) is 1 (8.7). The maximum freeway capacities are associated with points \((q_{on}^{(cont)}, q_{in})\) on the boundary \( F_{S}^{(B)} \) in the diagram of congested patterns (Fig. 8.1a). The minimum freeway capacities are associated with points \((q_{on}^{(cont)}, q_{in})\) on the threshold boundary \( F_{th}^{(B)} \) in this diagram.
The ALINEA method and other models in the context of the free flow control approach can maintain reliably free flow at the bottleneck if the flow rate $q_{\text{sum}}^{(\text{cont})}$ is less than minimum freeway capacities, i.e., when the condition (8.14) is satisfied. To understand this statement, note that in this case the probability for speed breakdown at the bottleneck during the time interval $T_{\text{ob}}$ is $P_{F_S}^{(B)} = 0$, i.e., free flow is stable against the onset of congestion at the bottleneck; corresponding points $(q_{\text{on}}^{(\text{cont})}, q_{\text{in}})$ in the diagram are left and below of the threshold boundary $F_{\text{th}}^{(B)}$ (Fig. 8.1a). However, at each given flow rate $q_{\text{in}}$ the minimum freeway capacity is considerably smaller than the maximum freeway capacity at the bottleneck. In particular, from the empirical example shown in Fig. 10.3 we can see that for $T_{\text{ob}} = T_{\text{av}} = 1 \text{ min}$ (see (10.9)) $q_{\text{th}}^{(B)} \approx 2200 \text{ vehicles/h}$ and $q_{\text{free}}^{(B)} \approx 3200 \text{ vehicles/h}$.

Different points $(q_{\text{on}}^{(\text{cont})}, q_{\text{in}})$ between the boundaries $F_S^{(B)}$ and $F_{\text{th}}^{(B)}$ in the diagram of congested patterns are related to the infinity of freeway capacities for which the probability for speed breakdown at the bottleneck during the time interval $T_{\text{ob}}$ is in the range $0 < P_{F_S}^{(B)} < 1$ (Sect. 8.3.3). In the latter case, free flow at the bottleneck is metastable with respect to the onset of congestion. Under application of the ALINEA method and other models in the context of the free flow control approach, the onset of congestion at the bottleneck is nevertheless possible: a short-time perturbation in free flow in the vicinity of the bottleneck can cause congested pattern emergence. This metastability of free flow explains GP emergence at the bottleneck under on-ramp metering via the ALINEA method (Fig. 23.7).

In contrast with the ALINEA method and other models in the context of the free flow control approach, in the ANCONA method speed breakdown is allowed to set in at the bottleneck. Therefore, in the ANCONA method the maximum value of the flow rate $q_{\text{sum}}^{(\text{cont})}$ can be greater than maximum freeway capacities. For this reason, the ANCONA method shows a greater mean throughput than the ALINEA method. Moreover, in the ANCONA method during some time intervals, points $(q_{\text{on}}^{(\text{cont})}, q_{\text{in}})$ in the diagram can even lie right of the boundary $S_J^{(B)}$ where GPs should appear spontaneously without feedback on-ramp control: feedback on-ramp control of the ANCONA method prevents GP formation at the bottleneck regardless of the occurrence of short-time perturbations in traffic flow. These features of the ANCONA method can explain the aforementioned benefits of the congested pattern control approach in comparison with the free flow control approach.

### 23.3.4 Comparison of Different Control Rules in Congested Pattern Control Approach

Let us consider a hypothetical application of the control rule (23.6) of the ALINEA method, where the detectors for feedback control are upstream of the bottleneck as in the ANCONA method (Fig. 23.2b; the location of the
detectors at $x_{3}^{\text{det}} = 9.8 \text{ km})$. In this case, the control rule (23.6) of the ALINEA method can change the on-ramp inflow only after the onset of congestion at the bottleneck has occurred. This is because when the detectors for feedback control are upstream of the on-ramp (Fig. 23.2b), the occupancy $o_{\text{sum}}^{\text{(cont)}}$ in (23.6) measured by the detectors in free flow is related to a constant flow rate on the main road $q_{\text{in}}$. When free flow is at the bottleneck then the measured occupancy $o_{\text{sum}}^{\text{(cont)}}$ does not depend on the flow rate of vehicles that merge onto the main road from the on-ramp $q_{\text{on}}^{\text{(cont)}}$. Thus, under free flow conditions at the bottleneck we obtain $q_{\text{on}}^{\text{(cont)}} = q_{\text{on}}$.

In this hypothetical application of the ALINEA method, on-ramp inflow control is realized when a congested pattern has occurred at the bottleneck. To explain this, note that only in this case the occupancy at the upstream detectors for feedback control can be higher than a chosen optimal occupancy $o_{\text{sum}}^{\text{(opt)}}$ in the rule (23.6). This is due to the upstream propagation of congestion. In this case, through the use of the rule (23.6) the flow rate $q_{\text{on}}^{\text{(cont)}}$ can be lower than the flow rate $q_{\text{on}}$.

In other words, this hypothetical application of the ALINEA method is associated with one of the possible realizations of the congested pattern control approach. The aim of a consideration of this hypothetical application is to make a comparison of the control rule (23.6) of the ALINEA method with the control rules (23.21) of the ANCONA method when the same upstream detectors for feedback control are used in the different control rules (23.6) and (23.21).

The result of this comparison is shown in Figs. 23.12 and 23.13 for two different values of the optimal occupancy, $o_{\text{sum}}^{\text{(opt)}}$, in the control rule (23.6).

In Fig. 23.12, the same optimal occupancy is used in the control rule (23.6) of the ALINEA method as those in Fig. 23.4. We found the following results: (i) The flow rate $q_{\text{on}}^{\text{(cont)}}$ in the ANCONA method (curve 1 in Fig. 23.12c) is higher than in the control rule (23.6) of the ALINEA method (curve 2); the average flow rate $q_{\text{on}}^{\text{(cont)}}$ in the ANCONA method is 430 vehicles/h, whereas it is 205 vehicles/h in the control rule (23.6) of the ALINEA method. (ii) Accordingly, the travel time in the on-ramp lane (waiting time at the light signal) in the ANCONA method (curve 1 in Fig. 23.12d) is considerably lower than the travel time in this hypothetical application of the ALINEA method (curve 2).

In Fig. 23.13b the optimal occupancy in the control rule (23.6) of the ALINEA method is increased considerably in comparison with the case shown in Fig. 23.12b. The chosen optimal occupancy $o_{\text{sum}}^{\text{(opt)}} = 18\%$ in the control rule (23.6) of the ALINEA method is now related to congested traffic rather than to free flow conditions on the main road. This is because the optimal density $\rho_{\text{sum}}^{\text{(opt)}} = 24 \text{ vehicles/km}$ related to the optimal occupancy $o_{\text{sum}}^{\text{(opt)}} = 18\%$ is higher than the maximum possible density in free flow $\rho_{\text{free}}^{\text{(max)}} = 22.5 \text{ vehicles/h}$ at the chosen model parameters (see Table 16.11 of
Fig. 23.12. Comparison of different control rules (23.6) and (23.21) in the congested pattern control approach. (a, b) Vehicle speed on the main road in space and time for the ANCONA control rule (23.21) (a) and ALINEA control rule (23.6) (b). (c) Flow rates $q_{on}^{(cont)}$ as functions of time. (d) Travel times in the on-ramp lane as functions of time. Curves 1 and 2 are related to the ANCONA (23.21) and ALINEA control rule (23.6), respectively. The location of the detectors for feedback control is $x_{3}^{(det)} = 9.8$ km (Fig. 23.2b). Other parameters for the ALINEA method are the same as those in Fig. 23.4. Parameters for the ANCONA method are the same as those in Fig. 23.10

Sect. 16.3). However, as in Figs. 23.11 and 23.12 in the ANCONA method (curve 1 in Fig. 23.13c) the average flow rate $q_{on}^{(cont)}$ is higher than in the case when the control rule (23.6) of the ALINEA method is used (curve 2). The average flow rate $q_{on}^{(cont)}$ in the ANCONA method is 430 vehicles/h, whereas it is 340 vehicles/h in the control rule (23.6) of the ALINEA method. Accordingly, the travel time in the on-ramp lane in the ANCONA method (curve 1 in Fig. 23.12d) is smaller than the travel time when the control rule (23.6) of the ALINEA method is used (curve 2).

We can conclude that the control rule (23.21) of the ANCONA method is more favorable than the control rule (23.6) of the ALINEA method even if the same upstream detectors for feedback control are used in these control rules (Fig. 23.2b).
23.4 Dissolution of Congested Patterns

The above ANCONA approach enables us to reduce congestion at an on-ramp bottleneck. This is due to a decrease in the flow rate of vehicles that can merge onto the main road from the on-ramp $q_{on}^{(cont)}$. However, to dissolve an GP at the bottleneck, a relatively large decrease in the flow rate $q_{on}^{(cont)}$ should be made. This reduction of the flow rate $q_{on}^{(cont)}$ can lead to an increase in congestion in other parts of a traffic network. To reduce this negative effect, flow rates $q_{on}$ and $q_{in}$ can both be reduced. The flow rate $q_{in}$ can also be reduced by one of the automatic on-ramp inflow control methods. However, in this case, this control should also be applied to an upstream bottleneck in the traffic network.

Here we consider simulations of congested pattern dissolution at an on-ramp bottleneck when both flow rates $q_{on}$ and $q_{in}$ are reduced after an GP has formed at the bottleneck [575]. In simulations, the flow rate $q_{on}^{(cont)}$ of vehicles that can merge onto the main road from the on-ramp is lower than $q_{on}$ due to light signal operation in the on-ramp lane(s) in accordance with
However, for simplicity we do not use feedback on-ramp control. The aim of this analysis is to compare different combinations of reduced flow rates $q_{on}$ and $q_{in}$ that leads to GP dissolution at the bottleneck.

If a congested pattern has already occurred at a freeway bottleneck, then a decrease in the flow rate $q_{in}$ or $q_{on}$ can obviously lead to pattern dissolution.

Let us consider a GP under the strong congestion condition at an on-ramp bottleneck (Fig. 23.14a). Now after the first moving jam in the incipient GP has been recognized on the main road at some distance upstream of the on-ramp, the flow rate $q_{in}$ is decreased (Figs. 23.14b–f).

We can see that the decrease in the flow rate $q_{in}$, however, does not lead to GP dissolution if the flow rate $q_{in}$ remains higher than the limit flow rate in the pinch region of the GP, $q_{lim}^{(pinch)}$ (Fig. 23.14b–d). Only when

$$q_{in} < q_{lim}^{(pinch)}$$

(23.23)

does the GP dissolve (Figs. 23.14e,f). This means that a relatively large decrease in flow rate upstream of the GP is necessary for GP dissolution. Note that the flow rate $q_{lim}^{(pinch)}$ has been explained in Sect. 18.3.

The physical meaning of this result is associated with the nature of the pinch effect and wide moving jam emergence in the GP. The average flow rate in the pinch region of the GP under the strong congestion condition is $q_{lim}^{(pinch)}$. This flow rate is lower than the maximum flow rate in the wide moving jam outflow $q_{out}$. If the flow rate $q_{in}$ becomes lower than $q_{out}$, then the farthest upstream wide moving jam in the GP dissolves first (Figs. 23.14b–d). This wide moving jam has, however, almost no influence on the pinch region of the GP, where narrow moving jams continuously continue to emerge even at

$$q_{in} < q_{out}$$

(23.24)

if the condition

$$q_{in} \geq q_{lim}^{(pinch)}$$

(23.25)

is satisfied. Only when the opposite condition (23.23) is valid, does the GP dissolve (Sect. 18.3).

A more efficient dissolution of the GP can be achieved via on-ramp inflow control. In Figs. 23.14g,h, on-ramp inflow control is applied. We see

---

9 The flow rate $q_{in}$ is decreased from the initial value $q_{in} = q_{in}^{(1)}$ to $q_{in} = q_{in}^{(2)} < q_{in}^{(1)}$ after the formation of the first moving jam has been detected. The criterion for moving jam formation is as follows. The 1-min average vehicle speed at the detector located on the main road 2.5 km upstream of the beginning of the merging region of the on-ramp ($x_{on} = 16$ km) drops below the level 20 km/h. In Figs. 23.14b–f, $q_{in}^{(2)}$ is 1580 (b), 1530 (c), 1475 (d), 1150 (e), and 950 (f) vehicles/h.

10 In Fig. 23.14g, the flow rate $q_{in}$ is not changed, but after the criterion for moving jam formation has been satisfied (footnote 9) the on-ramp flow rate decreases via light signal operation in the on-ramp lane. As a result, the flow rate of vehicles
Fig. 23.14. Dissolution of an GP in (a). Vehicle speed on the main road. (b-f) Variation of the incoming flow rate $q_{in}$. (g) Variation of the on-ramp flow rate $q_{on}$. (h) Variation of both $q_{in}$ and $q_{on}$. In (a–h) $q_{in}^{(1)} = 2000$ vehicles/h, $q_{on} = 750$ vehicles/h, $q_{in}^{(pinch)}$ $\approx 1475$ vehicles/h. The on-ramp inflow is switched at $t = t_0 = 8$ min. Results of numerical simulations of the model of Sect. 16.3. Taken from [575]
that a decrease in the flow rate $q_{\text{on}}^{\text{(cont)}}$ from 750 vehicles/h to 300 vehicles/h leads to the dissolution of the GP even if the flow rate $q_{\text{in}}$ does not change (Fig. 23.14g). If, in addition, the flow rate $q_{\text{in}}$ decreases, then the congested pattern fully dissolves at the bottleneck (Fig. 23.14h), even though the flow rate $q_{\text{in}}$ still satisfies the condition (23.25). The considered effect of on-ramp inflow control for GP dissolution can only be reached when the flow rate $q_{\text{on}}^{\text{(cont)}}$ satisfies the condition:

$$q_{\text{on}}^{\text{(cont)}} < q_{\text{on}}^{(G)}.$$  

(23.26)

Here $q_{\text{on}}^{(G)}$ is some characteristic value of the flow rate $q_{\text{on}}$ in the diagram of congested patterns at the on-ramp, as discussed in Sect. 18.5.

The dissolution of an WSP at the bottleneck (Fig. 23.15a) can be reached by a smaller decrease in the flow rate $q_{\text{in}}$ even without on-ramp control (Figs. 23.15b–d). WSP upstream propagation becomes slower (Fig. 23.15b). Then upstream WSP propagation is interrupted (Fig. 23.15c) until the WSP has fully dissolved (Fig. 23.15d) whereupon the flow rate $q_{\text{in}}$ decreases.

23.5 Prevention of Induced Congestion

Here we consider the dissolution of congested patterns that result from induced congestion [575]. Induced congestion can be realized when there are two different adjacent effectual bottlenecks on a freeway section. As in Sect. 19.2.1, bottleneck downstream is called the downstream bottleneck. The bottleneck upstream is called the upstream bottleneck. We assume that both bottlenecks are due to on-ramps. The downstream bottleneck is designated “the on-ramp ‘D’.” The upstream bottleneck is designated “the on-ramp ‘U’” (Fig. 23.16). The coordinates of the beginning of the merging regions of the on-ramps ‘D’ and ‘U’ and the flow rate to these on-ramps are denoted by $x_{\text{on}}^{(\text{down})}$, $x_{\text{on}}^{(\text{up})}$ and $q_{\text{on}}^{(\text{down})}$, $q_{\text{on}}^{(\text{up})}$, respectively. We will also assume that the flow rate in free flow $q_{\text{in}}$ upstream of the on-ramp ‘U’ is given.

We assume that a congested pattern occurs at the on-ramp ‘D’. This congested pattern propagates upstream (Fig. 23.16a,c,e,g). When this congested pattern reaches the on-ramp ‘U’, this pattern causes congested pattern formation at this upstream bottleneck. This effect is called induced pattern formation at the upstream bottleneck. The term “induced” reflects the effect of the emergence of the congested pattern at the upstream bottleneck due
to some “external” reason. This external reason is the upstream propagation of the congested pattern that has initially occurred at the downstream bottleneck. Empirical and theoretical features of this induced congested pattern effect have been considered in Sects. 2.4.4, 10.4, and Sect. 19.2.1, respectively.

In particular, empirical observations show (Sects. 2.4.4 and 10.4) that there are two possible cases of induced pattern formation:

(i) A wide moving jam occurs in an GP at the downstream bottleneck. Due to jam propagation upstream, the wide moving jam reaches an upstream freeway bottleneck. Then an F→S transition can be induced at this bottleneck by moving jam propagation through the upstream bottleneck. The induced F→S transition leads to congested pattern emergence. Even after the wide moving jam is far upstream of the upstream bottleneck, the congested pattern remains at the upstream bottleneck.

(ii) A region of synchronized flow is formed at the downstream bottleneck. Synchronized flow propagates upstream. When synchronized flow reaches the upstream bottleneck, this synchronized flow is caught at the upstream bottleneck (catch effect). The catch effect causes an induced F→S
Fig. 23.16. Initial congested patterns (left) and their dissolution (right) on a road with the upstream (‘U’) and downstream (‘D’) on-ramp bottlenecks due to a variation in on-ramp flow rates. Vehicle speed on the main road. (a, b) GP induced by an WSP (a) and its dissolution (b). (c, d) GP induced by an MSP (c) and its dissolution (d). (e, f) GP covering both on-ramp bottlenecks (e) and its dissolution (f). (g, h) WSP covering both on-ramp bottlenecks (g) and its dissolution (h). Flow rates on the main road, at the on-ramp ‘U’, and the on-ramp ‘D’ ($q_{in}$, $q_{up}$, $q_{down}$) are (1846, 620, 320) (a, b), (1846, 640, 30) (c, d), (1450, 900, 900) (e, f), (2000, 250, 300) (g, h) vehicles/h. ($x_{up}$, $x_{down}$) are (10, 16) km. The on-ramp inflows are switched at $t = t_0 = 8$ min. Results of numerical simulations of the model of Sect. 16.3. Taken from [575]
transition at the upstream bottleneck. This leads to congested pattern emergence at this bottleneck.

Here we restrict a discussion to the case of item (ii), i.e., when induced pattern formation at the upstream bottleneck is associated with the upstream propagation of synchronized flow that has initially emerged at the downstream bottleneck (Fig. 23.16a,c,e,g). This induced effect is possible because an F→S transition is a first-order phase transition: in free flow at the bottleneck the nucleation of the F→S transition is possible. This nucleation is induced by synchronized flow propagation.

We discuss four examples of this induced pattern formation. In the first case, an WSP emerges at the on-ramp 'D'. The upstream front of synchronized flow in the WSP propagates upstream. When synchronized flow reaches the on-ramp 'U', another congested pattern is induced by this synchronized flow at the on-ramp 'U'. This pattern is an GP (Fig. 23.16a).

In the second case, an MSP occurs at the on-ramp 'D'. The MSP propagates upstream. When the MSP reaches the on-ramp 'U', an GP is induced by the MSP at the on-ramp 'U' (Fig. 23.16c).

In the third case, an GP of type (2) occurs at the on-ramp 'D'. Recall that the upstream front of the GP of type (2) is related to the upstream front of synchronized flow in this GP (Sect. 9.4.2). The upstream front of synchronized flow of the GP propagates upstream. When synchronized flow reaches the on-ramp 'U', another GP is induced by this synchronized flow at the on-ramp 'U' (Fig. 23.16e).

In the last case, an WSP emerges at the on-ramp 'D'. The upstream front of synchronized flow in the WSP propagates upstream. When synchronized flow reaches the on-ramp 'U', another SP is induced by this synchronized flow at the on-ramp 'U'. This pattern is an WSP (Fig. 23.16g).

In simulations of congested pattern dissolution (Fig. 23.16b,d,f,h), $q^{(\text{cont, down})}_{\text{on}}$ and $q^{(\text{cont, up})}_{\text{on}}$ are the flow rates of vehicles that can merge onto the main road from the on-ramps 'D' and 'U', respectively. These flow rates can be lower than $q^{(\text{down})}_{\text{on}}$ and $q^{(\text{up})}_{\text{on}}$, correspondingly. This is due to light signal operation in the on-ramp lane(s) of the on-ramps 'D' and 'U', which is made in accordance with the formula (23.5). However, for simplicity we do not use feedback on-ramp inflow control. The aim of this analysis is to compare different combinations of the reduction of the flow rates $q^{(\text{cont, down})}_{\text{on}}$ and $q^{(\text{cont, up})}_{\text{on}}$ that lead to the dissolution of congested patterns at the bottlenecks.

In the cases shown in Fig. 23.16a,c,g, after the occurrence of the related SP has been detected, on-ramp inflow control (without feedback) is applied. The flow rate of vehicles that can merge onto the main road from the

---

11 The criterion for SP detection is as follows. The 1-min average vehicle speed at the detector located on the main road 200 m upstream of the beginning of the on-ramp lane 'D' ($x_{\text{on}}^{\text{down}} = 16 \text{ km}$) drops below the speed level $v_{\text{con}} = 85 \text{ km/h}$.
on-ramp 'U' \( q_{\text{on}}^{(\text{cont, up})} \) is reduced below the flow rate to this on-ramp \( q_{\text{on}}^{(\text{up})} \). This is carried out by the light signal operation on the on-ramp 'U'. As a result, the average flow rate \( q_{\text{on}}^{(\text{cont, up})} \) decreases.\(^{12}\) This leads to congested pattern control at both bottlenecks. Rather than initial congested patterns that cover both bottlenecks in Figs. 23.16a,c,g the congested patterns in the corresponding Figs. 23.16b,d,h almost dissolve after the flow rate \( q_{\text{on}}^{(\text{cont, up})} \) is reduced.

In the case shown in Fig. 23.16e, after the occurrence of a moving jam upstream of the on-ramp 'D' has been detected,\(^{13}\) the light signals in both the on-ramp lane 'U' and the on-ramp lane 'D' are switched. As a result, the average flow rates \( q_{\text{on}}^{(\text{cont, down})} \) and \( q_{\text{on}}^{(\text{cont, up})} \) onto the main road from the on-ramps 'D' and 'U' decrease to 250 and 400 vehicles/h, respectively. As a result, both GPs at the on-ramps 'U' and 'D' dissolve (Fig. 23.16f).

### 23.6 Influence of Automatic Cruise Control on Congested Patterns

#### 23.6.1 Model of Automatic Cruise Control

Automatic cruise control (ACC) is one of the most important ways of enhancing driver comfort and safety in freeway traffic (e.g., [535,577–591]). An ACC vehicle measures the space gap \( g_n = x_{\ell,n} - x_n - d \) and the relative speed \( v_{\ell,n} - v_n \) (Fig. 23.17; we use the same designations of values as those in Sect. 16.3). The ACC vehicle calculates the current time gap between the ACC vehicle and the preceding vehicle: \( \tau_n^{(\text{net})} = g_n / v_n \).

![Fig. 23.17. Model of the ACC vehicle](image)

In real ACC systems, two or more ranges of vehicle speed are usually chosen where different dynamic rules for the ACC vehicle are used. For simplicity of further analysis we will discuss a hypothetical ACC system where\(^{12}\) The average flow rate \( q_{\text{on}}^{(\text{cont, up})} \) in Figs. 23.16b,d,h is 200 (b), 250 (d), and 50 (h) vehicles/h, respectively.

\(^{13}\) The criterion for moving jam detection is the same as those in footnote 9 of Sect. 23.4.
there is only one dynamic rule for the ACC vehicle in the whole possible range of vehicle speed. In most known ACC systems, at least in one of the speed ranges the dynamic behavior of the ACC vehicle can be approximately described by the equation (e.g., [579–586]):

\[
a^{(\text{ACC})}_{n} = K_1 \left( g_n - v_n \tau^{(\text{ACC})}_{d} \right) + K_2 (v_{\ell,n} - v_n) .
\]  

(23.27)

In (23.27) \(a^{(\text{ACC})}_{n}\) is the acceleration of the ACC vehicle. \(\tau^{(\text{ACC})}_{d}\) is a desired time gap that is a given parameter of the ACC vehicle. The desired time gap \(\tau^{(\text{ACC})}_{d}\) is usually set by the driver of the ACC vehicle. In (23.27), as in the microscopic model of Sect. 16.3, the discrete time \(t = n\tau\) is used, \(n = 0, 1, 2, \ldots\) \(\tau\) is the time step. \(K_1\) and \(K_2\) are coefficients of ACC adaptation. These coefficients describe the dynamic adaptation of the ACC vehicle when either the space gap is different from \(v_n \tau^{(\text{ACC})}_{d}\):

\[
g_n \neq v_n \tau^{(\text{ACC})}_{d} 
\]

(23.28)

or the vehicle speed is different from the speed of the preceding vehicle:

\[
v_n \neq v_{\ell,n} .
\]

(23.29)

If in contrast

\[
v_n = v_{\ell,n}
\]

(23.30)

and the condition

\[
g_n = v_n \tau^{(\text{ACC})}_{d}
\]

(23.31)

is satisfied, from (23.27) we obtain that

\[
a^{(\text{ACC})}_{n} = 0 ,
\]

(23.32)

i.e., the ACC vehicle moves with a time-independent speed.

The physics of the dynamic equation for the ACC vehicle (23.27) is as follows. It can be seen that the current time gap \(\tau^{(\text{net})}_{n} = g_n/v_n\) in (23.27) is compared with the desired time gap \(\tau^{(\text{ACC})}_{d}\). If \(\tau^{(\text{net})}_{n} > \tau^{(\text{ACC})}_{d}\), then the ACC vehicle automatically accelerates to reduce the time gap to the desired value \(\tau^{(\text{ACC})}_{d}\). If \(\tau^{(\text{net})}_{n} < \tau^{(\text{ACC})}_{d}\), then the ACC vehicle decelerates to increase the time gap. Moreover, the acceleration and deceleration of the ACC vehicle depend on the current difference between the speed of the ACC vehicle and the preceding vehicle. If the preceding vehicle has higher speed than the ACC vehicle, i.e., when \(v_{\ell,n} > v_n\), the ACC vehicle accelerates. Otherwise, if \(v_{\ell,n} < v_n\) the ACC vehicle decelerates.

Here, based on the microscopic three-phase traffic theory of Part III we study a possible influence of ACC systems on spatiotemporal congested patterns at a freeway bottleneck due to the on-ramp [575,589].
In simulations of this ACC vehicle influence, there are vehicles that have no ACC system and ACC vehicles. Vehicles that have no ACC system move in accordance with the microscopic model of Sect. 16.3. The ACC vehicles are randomly distributed on the road between other vehicles that have no ACC system. The ACC vehicles move in accordance with (23.27) where, in addition, the following formulae are used to simulate the motion of the ACC vehicles:

\[ v_{c,n}^{(\text{ACC})} = v_n + \tau \max \left( -b_{\text{ACC}}, \min \left( a_n^{(\text{ACC})}, a_{\text{ACC}} \right) \right), \tag{23.33} \]

\[ v_{n+1} = \max \left( 0, \min \left( v_{\text{free}}, v_{c,n}^{(\text{ACC})}, v_{s,n} \right) \right). \tag{23.34} \]

The formula (23.33) limits the acceleration and deceleration of the ACC vehicles, where \( a_{\text{ACC}} \) and \( b_{\text{ACC}} \) are the maximum possible acceleration and deceleration of the ACC vehicle, respectively. Owing to the formula (23.34), the speed of the ACC vehicle \( v_{n+1} \) at time step \( n + 1 \) is limited by the maximum speed in free flow \( v_{\text{free}} \) and by the safe speed \( v_{s,n} \) to avoid collisions between vehicles. The safe speed \( v_{s,n} \) is the same as those in the microscopic model of Sect. 16.3.

### 23.6.2 Automatic Cruise Control with Quick Dynamic Adaptation

To study the influence of ACC vehicles on congested patterns, we consider the model of a freeway with a bottleneck due to the on-ramp (Fig. 16.2a). We assume that at the chosen flow rate upstream of the on-ramp \( q_{\text{in}} \) and the flow rate to the on-ramp \( q_{\text{on}} \) a general pattern (GP) occurs at the bottleneck, when there are no ACC vehicles on the freeway (Fig. 23.18a).

ACC vehicles can decrease the amplitude of moving jams in the initial GP if the coefficients of ACC adaptation \( K_1 \) and \( K_2 \) in the dynamic equation (23.27) are large enough. In this case, the ACC vehicle quickly reacts to changes in time gap and speed difference to the preceding vehicle. The suppression of moving jams in the initial GP due to ACC vehicles can be seen in Figs. 23.18b–d. Here, at the same \( q_{\text{in}} \) and \( q_{\text{on}} \) as those in Fig. 23.18a the percentage of ACC vehicles, \( \gamma \), in the flow rates \( q_{\text{in}} \) and \( q_{\text{on}} \) increases from \( \gamma = 20\% \) in Fig. 23.18b to \( \gamma = 37\% \) in Fig. 23.18d. There is some critical \( \gamma_{\text{cr}} \). When this critical percentage of ACC vehicles in traffic flow is reached, there are almost no moving jams in the congested pattern on the main road upstream of the on-ramp. If the percentage of ACC vehicles further increases, no new moving jams occur in the congested pattern. Thus, ACC vehicles can prevent moving jam emergence.

However, the discharge flow rate \( q_{\text{out}}^{(\text{bottle})} \), i.e., the flow rate on the main road downstream of the on-ramp in free flow is a continuously decreasing function of the percentage of ACC vehicles \( \gamma \) (Fig. 23.18e). This is because the vehicle speed upstream of the on-ramp in the congested pattern decreases
Fig. 23.18. Influence of ACC vehicles with high coefficients of the ACC adaptation on an GP at a bottleneck due to the on-ramp. (a–d) Vehicle speed on the main road. (a) No ACC vehicles, \( \gamma = 0 \). (b) \( \gamma = 20\% \). (c) \( \gamma = 30\% \). (d) \( \gamma = 37\% \). (e, f, g) Discharge flow rate (e), the flow rate within the congested pattern on the main road just upstream of the on-ramp (f) and travel time (g) as functions of \( \gamma \). \( K_1 = 0.1 \text{s}^{-2} \), \( K_2 = 0.55 \text{s}^{-1} \). \( \gamma_{(\text{ACC})} = 1.8 \text{sec} \). \( a_{\text{ACC}} = b_{\text{ACC}} = 2 \text{m/s}^2 \). \( \gamma_{(\text{cr})} = 35\% \). \( q_{\text{on}} = 600 \), \( q_{\text{in}} = 1730 \text{vehicles/h} \). Travel time in (g) is related to vehicle motion between \( x = 0 \) and the location \( x = 15 \text{km} \), which is 1 km downstream of the on-ramp; the vehicle starts at the time at which the upstream front of the congested pattern reaches the point \( x = 3 \text{km} \). Taken from [575, 589]
when $\gamma$ increases. As a result, the flow rate $q_{\text{con}}$ within the congested pattern upstream of the on-ramp also decreases (Fig. 23.18f). Thus, ACC vehicles with large enough coefficients of ACC adaptation prevent moving jams emergence, i.e., a more comfortable and safer driving experience is possible. However, these ACC vehicles cannot prevent or dissolve traffic congestion. They even lead to some decrease in discharge flow rate $q_{\text{bottle}}$ from the congested pattern at the bottleneck. The travel time $T_{\text{tr}}$ also increases when the percentage of ACC vehicles $\gamma$ increases (Fig. 23.18g). The function $T_{\text{tr}}(\gamma)$ has a maximum at $\gamma = \gamma_{\text{cr}}$, when all moving jams of the initial GP disappear. When the percentage of the ACC vehicles $\gamma$ further increases, the travel time decreases. However, the travel time remains much greater than the travel time in traffic flow without ACC vehicles.

### 23.6.3 Automatic Cruise Control with Slow Dynamic Adaptation

Another case occurs when an ACC vehicle reacts slowly to changes in time gap and speed difference to the preceding vehicle, i.e., the coefficients of ACC adaptation $K_1$ and $K_2$ in the dynamic equation (23.27) are small. Then ACC vehicles can lead to the onset of congestion at the bottleneck even if there is no congestion in traffic flow without ACC vehicles.

In Fig. 23.19a, where there are no ACC vehicles, free flow occurs at the bottleneck at the chosen flow rates $q_{\text{in}}$ and $q_{\text{on}}$. There is some critical percentage of ACC vehicles, $\gamma_{\text{cr}}$. At the same $q_{\text{in}}$ and $q_{\text{on}}$, if ACC vehicles appear on the freeway, then at a low enough percentage of ACC vehicles

$$\gamma < \gamma_{\text{cr}}$$

(23.35)

there is no influence of ACC vehicles. Under the condition (23.35), free flow remains at the bottleneck. If in contrast,

$$\gamma > \gamma_{\text{cr}}$$

(23.36)

ACC vehicles induce the onset of congestion, i.e., an F→S transition occurs at the bottleneck (Fig. 23.19b). As a result of the onset of congestion, an GP appears at the bottleneck. The higher the percentage of ACC vehicles $\gamma$, the higher the frequency of moving jam emergence in the GP (Figs. 23.19c,d).

We can see from these two examples that ACC vehicles can influence congested pattern features qualitatively. However, this occurs only if the percentage of ACC vehicles $\gamma$ is greater than some critical value $\gamma_{\text{cr}}$. This critical value depends strongly on ACC vehicle dynamics. It can be assumed that if an ACC vehicle exhibits qualitatively different dynamic features than those described by (23.27), quantitatively different results for ACC vehicle influence on congested patterns at freeway bottlenecks can be expected. Thus, there is great potential for the development of new assistance systems in
Fig. 23.19. Influence of ACC vehicles with low coefficients of ACC adaptation on the onset of congestion at a bottleneck due to the on-ramp. Vehicle speed on the main road. (a) No ACC vehicles, $\gamma = 0$. (b) $\gamma = 5\%$. (c) $\gamma = 8\%$. (d) $\gamma = 20\%$. 

$K_1 = 0.03 \text{ s}^{-2}$, $K_2 = 0.18 \text{ s}^{-1}$. $\tau_{d}^{(\text{ACC})} = 1.8 \text{ sec}$. $a_{\text{ACC}} = b_{\text{ACC}} = 2 \text{ m/s}^2$. $\gamma_{\text{cr}} = 4\%$. 

$q_{\text{on}} = 400$, $q_{\text{in}} = 1730 \text{ vehicles/h}$. Taken from [575,589]

vehicles. These assistance systems can improve traffic comfort and increase traffic safety considerably due to their influence on congested pattern features.

23.7 Conclusions

(i) Very complex nonlinear spatiotemporal dynamics of congested patterns can occur, when congested pattern characteristics are changed by application of one of the methods of traffic control. Therefore, to find the efficiency of methods of traffic control, assignment, and management, a possible influence of the chosen strategy on spatiotemporal features of congested patterns at freeway bottlenecks should first be studied.

(ii) The on-ramp control strategies based on the ramp metering with feedback are efficient methods for increasing the flow rate on a freeway, reducing traffic congestion at freeway bottlenecks, and reducing individual vehicle travel time. The ANCONA control approach where congestion at a bottleneck is allowed to set in shows the following benefits
in comparison with the ALINEA method that maintains free flow at the bottleneck:

(a) Higher throughputs on the main road and on-ramp.
(b) Considerably lower vehicle waiting times at the light signal on the on-ramp.
(c) The upstream propagation of congestion does not occur even if a congested pattern occurs at the bottleneck due to a short-term perturbation in traffic flow: the congested pattern is spatially localized in the vicinity of the bottleneck.

(iii) Changes in ramp inflow at the bottleneck where a congested pattern has occurred and in flow rate on the main road upstream of a congested pattern can lead to qualitatively different consequences for the spatiotemporal pattern dynamics and pattern dissolution.

(iv) Automatic cruise control (ACC) and other assistance systems in vehicles can be an useful method for control of spatiotemporal congested pattern features at freeway bottlenecks.

(v) Traffic control based on autonomous ACC functions in vehicles can improve the efficiency of the system. Wide moving jams can be suppressed by ACC vehicles, and therefore traffic flow can be harmonized and stabilized in the synchronized mode.

(vi) At certain parameters and percentages of ACC vehicles, traffic flow can be influenced in a negative direction. The ACC vehicles can possibly lead to traffic breakdowns and congested pattern occurrence at bottlenecks.

(vii) Different kinds of vehicle assistance systems can be investigated in simulations based on three-phase traffic theory concerning their influence on the efficiency of a traffic control system. This is because a microscopic simulation environment based on three-phase traffic theory is in accordance with empirical spatiotemporal features of congested patterns. Thus, this is a high-value instrument in a study of different freeway traffic control strategies and traffic assistance systems before they are introduced to the market.
Empirical spatiotemporal features of congested freeway traffic from German, American, and Canadian freeways discussed in this book show that the physics of traffic flow is proving to be much richer than most traffic researchers working in the field had expected. Empirical observations of real traffic patterns have revealed that in congested traffic there are two qualitatively different traffic phases, “synchronized flow” and “wide moving jam,” and revealed a large number of nonlinear features. Apparently for this reason, earlier traffic flow theories and models cannot adequately explain and therefore not predict most of these empirical spatiotemporal congested pattern features.

On the one hand, some of these features are qualitatively similar to those observed in other nonlinear non-equilibrium physical, chemical, and biological spatial distributed systems. On the other hand, there are three qualitatively different phases in real freeway traffic – “free flow,” “synchronized flow,” and “wide moving jam” – that exhibit some very specific features.

Three-phase traffic theory along with mathematical microscopic models based on this theory discussed in this book enable us to explain most of these empirical spatiotemporal freeway traffic phenomena. The next challenge for theoreticians in this field is to develop new theoretical concepts for a macroscopic mathematical description of these features.

There are also many “black areas” in empirical features of freeway traffic that should be understood in the future. In empirical observations, synchronized flow exhibits very complex spatiotemporal dynamic behavior. In particular, the behavior and the role of fluctuations in comparison with deterministic effects and also features of critical fluctuations in synchronized flow leading to moving jam emergence have not been sufficiently understood.

Thus, these and many other unsolved problems associated with synchronized flow features are a very interesting and important field of the future empirical and theoretical investigations.

The FOTO and ASDA models based on three-phase traffic theory have shown very good results of their online application for the reconstruction and tracking of real spatiotemporal congested traffic patterns at the traffic control center of the German Federal State of Hessen. Empirical features of phase transitions and of spatiotemporal congested patterns are also the basis
of concepts and strategies for traffic control and management discussed in this book.

The application of the FOTO and ASDA models at the traffic control center of the German Federal State of Hessen over the last three years and also the study of congested patterns on other different freeways in Germany and the USA presented in this book show that empirical congested traffic patterns exhibit various reproducible and predictable spatiotemporal features. These features can and should be used for the development of new applicable and efficient concepts for freeway traffic control, assignment, and management with the aim to reduce congestion in a traffic network or to dissolve congested patterns. The next challenge for traffic engineers is to develop such concepts and to prove them in practical applications.

Different kinds of vehicle assistance systems can be investigated in simulations based on three-phase traffic theory concerning their influence on comfortable and safe driving, as well as on the efficiency of traffic control systems. This is because a microscopic simulation environment based on three-phase traffic theory is in accordance with empirical spatiotemporal features of congested patterns. Thus, this is a high-value instrument in a study of different freeway traffic control strategies and traffic assistance systems before they are introduced to the market.
A Terms and Definitions

A.1 Traffic States, Parameters, and Variables

Traffic state: A state that is characterized by a certain set of statistical properties of traffic variables and by a certain set of traffic “control” parameters.

Traffic variable: This is traffic characteristic that can show complex spatiotemporal behavior even if traffic control parameters are spatially homogeneous and time-independent. Examples of traffic variables are:

- flow rate, $q$ (vehicles per hour),
- vehicle speed, $v$ (km per hour),
- vehicle time gap (sec),
- vehicle space gap (m),
- vehicle density, $\rho$ (vehicles per km).

Traffic control parameter: A “control” parameter is regardless of the spatiotemporal behavior of traffic variables in traffic. However, a change in a traffic control parameter can lead to a considerable change in traffic variables. Examples of traffic control parameters are weather, other road conditions, and vehicle characteristics.

A.2 Traffic Phases

Traffic phase: A traffic phase is a set of traffic states considered in space and time that exhibit some specific empirical spatiotemporal features. These features are specific (unique) for each traffic phase.

Three-phase traffic theory: Corresponding to reproducible empirical studies of freeway traffic related to investigations of spatiotemporal traffic measurements on different freeways over many different days and years [203, 207, 208, 210, 218], in the three-phase traffic theory it is assumed that besides the “free flow” traffic phase there are two different traffic phases in congested traffic, “synchronized flow” and “wide moving jam.” This theory explains the complexity of traffic phenomena based on phase transitions among these three traffic phases and nonlinear spatiotemporal features of these traffic phases.
Objective criteria for the traffic phases in congested traffic: The objective (empirical) criteria to distinguish between the “synchronized flow” phase and the “wide moving jam” phase are associated with the following spatiotemporal empirical features of these different traffic phases.

The “wide moving jam” traffic phase is a moving jam\(^1\) that maintains the velocity of the downstream front of the jam during jam propagation. The downstream front velocity is maintained as long as a moving jam is a wide moving jam even when the wide moving jam propagates through any other (sometimes very complicated) traffic states and through freeway bottlenecks (Fig. 1.2).

The “synchronized flow” phase. In contrast to a wide moving jam, the downstream front of synchronized flow does not maintain the velocity of the front. In particular, the downstream front of synchronized flow is usually fixed at a freeway bottleneck (Fig. 1.2). Within this downstream front vehicles accelerate from lower vehicle speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front.

A front between two different traffic states or two different traffic phases is a region in traffic where a spatial transition from one state of traffic or one traffic phase upstream to another state of traffic or traffic phase downstream occurs. The front between two different traffic states separates two different states of the same traffic phase. The front between two different traffic phases separates two different traffic phases.

A.3 Phase Transitions

Local phase transition in traffic: A transition from an initial traffic phase to another traffic phase that occurs at some location on a freeway.

Spontaneous local phase transition: A local phase transition caused by an internal local disturbance in traffic flow. This internal local disturbance can be associated with a deterministic local perturbation at a freeway bottleneck or to the occurrence of a random local perturbation in traffic flow. An example of a spontaneous phase transition from the “free flow” phase to the “synchronized flow” phase is shown in Figs. 2.10 and 2.11. This phase transition is labeled “F→S transition” in these figures.

Induced phase transition: A local phase transition in traffic flow is caused by an external short-time local disturbance in traffic flow. This external disturbance can be associated with the propagation of a moving spatiotemporal traffic pattern that has initially occurred at a different freeway location in comparison with the location of the induced F→S transition. An example of an induced phase transition from the “free flow” phase to the “synchronized flow” phase due to wide moving jam propagation through the location of

\(^1\) The term “moving jam” is defined in Chap. 1.
A freeway bottleneck is shown in Fig. 1.2a. This phase transition is labeled “induced $F \rightarrow S$ transition” in this figure.

Types of phase transitions in traffic. Phase transitions can be one of the following types:

(i) A phase transition from the “free flow” phase to the “synchronized flow” phase. This phase transition is called the $F \rightarrow S$ transition.

(ii) A phase transition from the “free flow” phase to the “wide moving jam” phase. This phase transition is called the $F \rightarrow J$ transition.

(iii) A phase transition from the “synchronized flow” phase to the “wide moving jam” phase. This phase transition is called the $S \rightarrow J$ transition.

(iv) A phase transition from the “synchronized flow” phase to the “free flow” phase. This phase transition is called the $S \rightarrow F$ transition.

(v) A phase transition from the “wide moving jam” phase to the “free flow” phase. This phase transition is called the $J \rightarrow F$ transition.

(vi) A phase transition from the “wide moving jam” phase to the “synchronized flow” phase. This phase transition is called the $J \rightarrow S$ transition.

(vii) A sequence of two phase transitions, first a phase transition from the “free flow” phase to the “synchronized flow” phase and later, and usually at a different freeway location a phase transition from the “synchronized flow” phase to the “wide moving jam” phase. This sequence of phase transitions is called the sequence of $F \rightarrow S \rightarrow J$ transitions.

Breakdown phenomenon at a freeway bottleneck (Fig. 2.11): The breakdown phenomenon is an abrupt drop in average vehicle speed in the vicinity of a freeway bottleneck (Fig. 2.11a). In three-phase traffic theory, the well-known breakdown phenomenon at the bottleneck is associated with an $F \rightarrow S$ transition at the bottleneck. An example of the breakdown phenomenon with synchronized flow pattern formation is shown in Fig. 2.10. Firstly, an $F \rightarrow S$ transition has occurred leading to traffic congestion. The downstream front of this congested traffic, where vehicles accelerate from congestion to free flow, is fixed at the bottleneck. Thus, corresponding to the objective criteria for the traffic phases in congested traffic, this congested traffic is associated with the “synchronized flow” phase.

A.4 Bottleneck Characteristics

An effectual freeway bottleneck is a freeway bottleneck at which the onset of congestion can occur spontaneously and then a spatiotemporal congested pattern is formed upstream of a freeway bottleneck on different days. A spontaneous $F \rightarrow S$ transition leading to traffic congestion occurs most often at freeway bottlenecks. An example is shown in Fig. 2.10, where after the $F \rightarrow S$ has occurred, synchronized flow is formed upstream of a bottleneck due to the on-ramp.
An isolated effectual freeway bottleneck (or an isolated bottleneck for short) is a freeway bottleneck that is far away from other effectual adjacent freeway bottlenecks. The influence of all other bottlenecks on both the onset of congestion and on subsequent pattern formation upstream of the isolated bottleneck can be neglected at an isolated bottleneck. An example of the breakdown phenomenon at an isolated bottleneck is shown in Fig. 2.10.

The effective location of an effectual bottleneck is the freeway location in the vicinity of the effectual isolated bottleneck where the downstream front of synchronized flow is spatially fixed. Within the downstream front of synchronized flow vehicles accelerate from synchronized flow upstream of the front to free flow downstream of the front.

A.5 Congested Patterns at Bottlenecks

Due to an F→S transition, a congested pattern occurs at an isolated freeway bottleneck. There are two main types of congested patterns [218]:

Synchronized flow pattern (SP): An SP is a congested pattern that consists of synchronized flow only, i.e., no wide moving jams emerge in the SP (Fig. 2.14). An F→S transition is responsible for SP emergence in an initial free flow.

General pattern (GP): An GP is a congested pattern that consists of synchronized flow upstream of the effectual bottleneck and wide moving jams that emerge spontaneously in that synchronized flow. Thus, in the GP there are two traffic phases of congested traffic, the “synchronized flow” and “wide moving jam” phases (Fig. 2.21). F→S→J transitions are responsible for GP emergence in an initial free flow.

There are diverse SPs and GPs that can be formed at bottlenecks. If a bottleneck is due to the on-ramp, different SPs and GPs occur spontaneously, depending on both the flow rate $q_{in}$ in an initial free flow on the main road upstream of the on-ramp, and the flow rate $q_{on}$ to the on-ramp. This leads to a diagram of congested patterns in the flow–flow plane where pattern emergence on the main road upstream of the bottleneck is presented as a function of the flow rates $q_{in}$ and $q_{on}$ (Fig. 2.22b).

A characteristic parameter of a congested pattern is a unique, coherent, predictable, and reproducible congested pattern parameter that does not depend on the initial conditions in traffic. A characteristic parameter is time-independent under fixed traffic control parameters. It can depend on traffic control parameters (weather, other road conditions, average driver and vehicle characteristics, and so on). An example of a characteristic parameter is the mean velocity of the downstream front of a wide moving jam (in Fig. 1.2a, the mean velocity of the downstream jam front is constant for more than one hour of jam propagation on a freeway section).

However, this is only true while the upstream boundary of a congested pattern reaches an upstream effectual adjacent bottleneck.
A.6 Local Perturbations

A local perturbation in traffic is a spatially localized disturbance in traffic flow, i.e., a local change in a traffic variable(s).

A deterministic local perturbation is caused by a permanent local disturbance of traffic flow. The usual reason for the permanent local disturbance of traffic flow is a freeway bottleneck.

A random local perturbation (local fluctuation) is a spatially localized time-dependent disturbance in traffic flow caused by a random event. For example, an unexpected lane change by one of the drivers can lead to a random local perturbation in the form of a decrease in the vehicle speed and the related local increase in vehicle density in some spatially limited region upstream of this driver.

A critical local perturbation is a local perturbation whose amplitude equals the critical amplitude required for a local phase transition in traffic flow. In addition, a critical local perturbation should exhibit some critical spatial form.

The amplitude of a local perturbation associated with traffic variable is the maximum difference between the value of this variable within the local perturbation and the value of this traffic variable before the local perturbation in the traffic state occurs.

Critical amplitude of a local perturbation: If the amplitude of a local perturbation exceeds the critical amplitude, then this perturbation grows and leads to a local phase transition in an initial state of traffic. Otherwise, if the amplitude of the perturbation is less than the critical amplitude, no phase transition due to this perturbation can occur.

A.7 Critical and Threshold Traffic Variables

Critical value of a traffic variable (e.g., critical density or critical flow rate): The value of a traffic variable at which the probability for a local phase transition reaches 1 (Fig. 2.12). Critical values of a traffic variable can be very different for the various aforementioned phase transitions in traffic.

The probability for a phase transition is the probability for a phase transition during a given time interval $T_{ob}$ and a given freeway section length. In the case of an F→S transition at an effectual freeway bottleneck, there is a sharp maximum in spatial dependence of the probability density for this transition in the vicinity of the bottleneck (Fig. 5.12): the F→S transition occurs with the highest probability at the location of this maximum. The probability density is the probability for the phase transition per unit length of road during the time interval $T_{ob}$. An example of dependence of the probability for an F→S transition on flow rate at a bottleneck due to the on-ramp is shown in Fig. 2.12.
Threshold value of a traffic variable (e.g., threshold density or threshold flow rate): The threshold value of a traffic variable is the value of the traffic variable at which the critical amplitude of the critical perturbation for one of the possible phase transitions in traffic reaches the maximum possible value for this phase transition. This means that at the threshold value of the traffic variable, this phase transition can occur only if a local perturbation occurs in traffic flow, whose amplitude exceeds this maximum critical amplitude. Threshold traffic variables can be very different for the various phase transitions in traffic.

The critical range of a traffic variable is the range between the threshold value of a traffic variable and the critical value of the traffic variable. In this critical traffic variable range a local phase transition can occur. However, the phase transition will not necessarily occur. The phase transition occurs in this critical range only with some probability when in an initial traffic phase a local perturbation amplitude exceeds the critical perturbation amplitude. Examples of critical ranges of the flow rate for an F→S transition (for various averaging time intervals $T_{av}$ for the flow rate) are shown in Fig. 2.12.

The critical amplitude of the critical local perturbation is usually a function of traffic variables and/or traffic control parameters. The critical amplitude of a local perturbation reaches zero at the critical value of a traffic variable. The critical amplitude of a local perturbation reaches the maximum value at the threshold value of a traffic variable.

Time delay of a phase transition: There is a delay between the time when a traffic variable is in the critical range for a phase transition and the time when this phase transition occurs. The time delay is a very complex random characteristic of the phase transition. In different realizations, the time delay can differ at the same values of traffic variables and traffic control parameters.

A.8 Some Features of Phase Transitions and Traffic State Stability

Hysteresis effect: An effect where first a local phase transition occurs from an initial traffic phase to another traffic phase, and later a reverse phase transition occurs to the initial traffic phase (Fig. 2.15). As a result of the hysteresis effect, a hysteresis loop is realized when both transitions are studied, for example in the flow–density plane or in the speed–density plane.

Stable traffic state with respect to a local phase transition: A traffic state where regardless of the amplitude of a local short-time disturbance in a traffic state the local phase transition does not occur. The probability of a local phase transition in a stable traffic state is zero. In particular, all values of flow rate below the threshold flow rate are related to stable traffic states in traffic flow. However, a traffic state that is stable with respect to one type of phase transition can be a metastable state with respect to another phase transition. This is one of the most complex features of traffic flow.
Metastable traffic state: A traffic state of one of the traffic phases in which random fluctuations (perturbations) of small enough amplitude fail to grow. However, if the amplitude of a local perturbation exceeds the critical amplitude of this perturbation, then this local perturbation grows. This growth leads to a local phase transition to another traffic phase.

Nucleation effect: The effect of the occurrence of a local perturbation in traffic whose amplitude exceeds the critical amplitude. The subsequent growth of this perturbation ("nucleus" for the phase transition) leads to a local phase transition in this traffic state.

Nucleus for a local phase transition: A local perturbation in traffic flow whose amplitude either equals or exceeds the critical amplitude for a local phase transition in traffic flow.

First-order local phase transition: A local phase transition that occurs in a metastable state of traffic flow. A first-order phase transition is accompanied by a precipitous local change in traffic variables. A first-order spontaneous local phase transition results from spontaneous nucleation in a metastable state of traffic flow, i.e., from the occurrence of a nucleus for the phase transition. The hysteresis effect is a common attribute of a first-order local phase transition.

Self-organization in traffic flow: Each of the above spontaneous and induced phase transitions leading to congested pattern formation and effects of pattern transformation are examples of self-organization in traffic. The induced F→S transition shown in Fig. 1.2a and the spontaneous F→S transition shown in Fig. 2.11 are examples of self-organization in traffic flow. The existence of a characteristic parameter of wide moving jam propagation, i.e., jam propagation with a constant mean velocity of the downstream jam front is another example of self-organization.
B ASDA and FOTO Models for Practical Applications

B.1 ASDA Model for Several Road Detectors

The aim of this Appendix is to consider some extended formulations the FOTO and ASDA models used in online application of these models at the traffic control center of the German Federal State of Hessen [238].

In real applications, several adjacent detectors in succession exist on freeways (Fig. B.1). In such a case, corresponding to (21.1) and (21.2), locations of the upstream and downstream fronts of a wide moving jam are found from the equation for the location of the upstream jam front

\[ x^{(\text{jam})}_{\text{up}}(t) = L_{i+1} + \int_{t_0^{(i+1)}}^{t} v^{(\text{jam})}_{\text{up}}(t) dt \approx \]

\[ \approx L_{i+1} - \int_{t_0^{(i+1)}}^{t} \frac{q_0^{(i)}(t) - q_{\text{min}}^{(\text{jam})}(t)}{\rho_{\text{max}}^{(\text{jam})}(t) - \left( \frac{q_0^{(i)}(t)}{v_0^{(i)}(t)} \right)} dt, \quad t \geq t_0^{(i+1)} \]  

(B.1)

and the equation for location of the downstream jam front

\[ x^{(\text{jam})}_{\text{down}}(t) = L_j + \int_{t_0^{(j)}}^{t} v^{(\text{jam})}_{\text{down}}(t) dt \approx \]

\[ \approx L_j - \int_{t_1^{(j)}}^{t} \frac{q_{\text{out}}^{(\text{jam}, j)}(t) - q_{\text{min}}^{(\text{jam})}(t)}{\rho_{\text{max}}^{(\text{jam}, j)}(t) - \left( \frac{q_{\text{out}}^{(\text{jam}, j)}(t)}{v_{\text{max}}^{(\text{jam}, j)}(t)} \right)} dt, \quad t \geq t_1^{(j)}, \]  

(B.2)

where indexes \( i \) in (B.1) and \( j \) in (B.2) for detectors, whose values at time \( t \) have to be used, increase in the direction of flow, i.e., corresponding to the direction of the x axis: e.g., \( L_m > L_n \) for \( m > n \); \( L_{i+1}, L_j \) are coordinates of the corresponding detectors; \( t_0^{(i+1)} \) indicates the time when the upstream jam front is measured at the detector \( i+1 \); \( t_1^{(j)} \) indicates the time when the downstream jam front is measured at the detector \( j \); \( q_0^{(i)}(t) \) and \( v_0^{(i)}(t) \) are the flow rate and average vehicle speed measured at the detector \( i \) upstream.
of the moving jam; $q_{\text{out}}^{(\text{jam}, j)}(t)$ and $v_{\text{max}}^{(\text{jam}, j)}(t)$ are the flow rate and average vehicle speed in the moving jam outflow measured at the detector $j$; the parameter $\rho_{\text{max}}^{(\text{jam})}$ can be calculated directly from measured data via (21.4); in online application of the ASDA model, the flow rate $q_{\text{min}}^{(\text{jam})}$ is determined via (21.5). The time-dependent width of the jam $L_J$ is given by (21.3).

For the choice of the correct detector $i$, whose measured data should be used in (B.1), the following algorithm can be applied. This algorithm will be illustrated for the case shown in Fig. B.1 where the upstream front of a wide moving jam has just been registered at detector $i + 1 = k$.

In the beginning, detector $i = k - 1$ is used in (B.1) and the downstream jam front is not tracked since it has not reached detector $k$ yet.

Firstly, we will define the times at which the next detector upstream has to be used in (B.1) for the calculation of the velocity of the upstream jam front. If the following inequality becomes true

$$x^{(\text{jam})}(t) < L_{\text{up}} - k - 1,$$

measurements of detector with index $i = k - 2$ (Fig. B.1) have to be used in (B.1). On the other hand, the upstream jam front is at detector $i = k - 1$ at the time $t_{k-1}$ that is found from the condition

$$x^{(\text{jam})}(t_{k-1}) = L_{k-1} \quad \text{with} \quad x^{(\text{jam})}(t_{0}) = L_{k} + \int_{t_{0}}^{t_{k-1}} v_{up}^{(\text{jam})}(t) \, dt.$$

Note that there is always a difference $\Delta$ between the calculated coordinate of the upstream jam front $x_{\text{up}}^{(\text{jam})}$ (B.1) and a real coordinate. If, for example, (B.3) is still false for the calculated upstream jam front at time $t_{k-1}$, when this front is already registered at detector $k - 1$, the calculated front location $x_{\text{up}}^{(\text{jam})}(t_{k-1})$ is definitely too high. In this case, the algorithm sets $x_{\text{up}}^{(\text{jam})}(t_{k-1})$ to $L_{k-1}$. Similarly, the calculated location $x_{\text{up}}^{(\text{jam})}(t)$ is definitely too low if $x_{\text{up}}^{(\text{jam})}(t) \leq L_{k-1}$, but the upstream jam front has not yet been detected by detector $k - 1$ at time $t$. Then $x_{\text{up}}^{(\text{jam})}(t)$ is set to $L_{k-1} + \epsilon_d$ where $\epsilon_d$ is a small given value (we used $\epsilon_d = 1$ m).
Accordingly, the choice of the next upstream detector in (B.1) does not depend on the calculated location of the upstream jam front, but on registration times of the upstream jam front at detectors: detector \( i = k - m - 1 \) is used in (B.1) as soon as the upstream jam front has been registered by detector \( k - m \) at time \( t = t_0^{(k-m)} \).

A similar algorithm is applied to choose the detector to be used for the location of the downstream jam front in (B.2). This formula can first be applied at time \( t = t_1^{(k)} \) when the downstream jam front is registered at detector \( k \). As soon as the downstream jam front is registered at the next upstream detector \( k - 1 \) at time \( t = t_1^{(k-1)} \), data of detector \( j = k - 1 \) is used in (B.2).

In a similar manner to the remark of the difference between the real location and the calculated location of the upstream jam front described above, the calculated location of the downstream jam front is corrected according to registration times of the downstream jam front at the detectors: if the calculated location \( x_{\text{down}}^{(\text{jam})} > L_{k-p} \), but the downstream jam front has already been registered at detector \( k - p \), \( x_{\text{down}}^{(\text{jam})} \) is set to \( L_{k-p} - \epsilon_d \). If the calculated location \( x_{\text{down}}^{(\text{jam})} < L_{k-p} \), but the downstream jam front has not yet been registered at detector \( k - p \), \( x_{\text{down}}^{(\text{jam})} \) is set to \( L_{k-p} \).

The case of one wide moving jam detected between the detectors \( i = k - 1 \) and \( i = k \) has been discussed above (Fig. B.1). If after the first moving jam (as often happens in traffic flow), further wide moving jams appear and these wide moving jams are spatially close together with small distances so that there is no detector in between, the ASDA model has to be extended. In this case, results of empirical investigations of characteristic features of wide moving jams can be applied. In particular, the velocity of the downstream front of a wide moving jam is a characteristic parameter. If this velocity was recently measured for a specific wide moving jam, it is used to predict the velocity of the downstream front of wide moving jams while this velocity cannot be measured. For the tracking of a wide moving jam downstream of the first one, it can also be assumed that the velocity of the upstream front of this jam equals the known characteristic velocity of the downstream front of the first moving jam.

### B.1.1 Extensions of ASDA for On-Ramps, Off-Ramps, and Changing of Number of Freeway Lanes

#### Upstream of Moving Jam

In this and the next section all flow rates are related to the flow rate per freeway lane.

In the vicinity of an on-ramp, if the upstream jam front has passed the detector \( i + 1 \) on the main road downstream of the on-ramp (Fig. B.2a), then the flow rate \( q_0^{(i)}(t) \) in (B.1) should be extended by the flow rate \( q_{\text{on}}/n \) distributed on all \( n \) freeway lanes.
Fig. B.2. Extensions of the ASDA model
In the vicinity of an off-ramp, if the upstream jam front has passed the detector \( i + 1 \) on the main road downstream of the off-ramp (Fig. B.2b), the contribution of the flow rate \( q_{\text{off}}/n \) has to be subtracted from the flow rate \( q_0^{(i)} \) in (B.1). The flow rates \( q_{\text{on}} \) and \( q_{\text{off}} \) are the flow rates to the on-ramp and to the off-ramp, respectively.

Let us assume that downstream of the detector \( i \) and upstream of a moving jam the number of freeway lanes on the main road reduces in the direction of traffic flow from \( m \) lanes to \( n \) lanes \((m > n)\) (Fig. B.2c). In this case, the flow rate \( q_0^{(i)} \) in (B.1) has to be applied in relation to the number of lanes, \( m/n \).

Thus, in a general case, when there are the on-ramp, the off-ramp, and the reduction of lanes, the equation (B.1) reads as follows:

\[
x_{\text{jam}}^{(i+1)}(t) = L_{i+1} + \int_{t_0^{(i+1)}}^{t} v_{\text{jam}}^{(i+1)}(t) dt \approx \nonumber \]

\[
\approx L_{i+1} - \int_{t_0^{(i+1)}}^{t} \frac{q_0^{*}(i)(t) - q_{\text{min}}^{(i)}(t)}{\rho_{\text{max}}^{(i)}} dt, \quad t \geq t_0^{(i+1)}, \quad (B.5) \]

where

\[
q_0^{*}(i)(t) = \frac{m}{n} q_0^{(i)}(t) + \frac{q_{\text{on}}(t)}{n} - \frac{q_{\text{off}}(t)}{n}, \quad i = 1, 2, \ldots \quad (B.6) \]

### B.1.2 Extensions of ASDA for On-Ramps, Off-Ramps, and Changing of Number of Freeway Lanes Downstream of Moving Jam

In the vicinity of an on-ramp, when the downstream jam front due to upstream jam propagation on the main road has passed the on-ramp (Fig. B.2d), then the flow rate \( q_{\text{on}}/n \) has to be subtracted from the flow rate \( q_{\text{out}}^{(\text{jam}, j)} \) in (B.2).

In the vicinity of an off-ramp, when the downstream jam front has passed the off-ramp (Fig. B.2e), the flow rate \( q_{\text{out}}^{(\text{jam}, j)} \) in (B.2) should be extended by the flow rate \( q_{\text{off}}/n \).

If upstream of detector \( j \) and downstream of a moving jam the number of freeway lanes increases in the direction of traffic flow from \( n \) lanes to \( m \) lanes \((m > n)\) (Fig. B.2f), then the flow rate \( q_{\text{out}}^{(\text{jam}, j)} \) in (B.2) has to be applied in relation to the number of lanes, \( m/n \).

Thus, when there are the on-ramp, the off-ramp, and the reduction of lanes, the equation (B.2) reads as follows:

\[
x_{\text{jam}}^{(i)}(t) = L_{j} + \int_{t_0^{(i)}}^{t} v_{\text{jam}}^{(i)}(t) dt \approx \nonumber \]
\[ L_j = \int_{t_{ij}}^{t} \frac{q_{\text{out}}^{(j)}(t) - q_{\text{in}}^{(j)}(t)}{\rho_{\text{max}}^{(j)}(t)} dt, \quad t \geq t_{ij}^{(j)}, \quad (B.7) \]

where

\[ q_{\text{out}}^{(j)}(t) = \frac{m}{n} q_{\text{out}}^{(j)}(t) - \frac{q_{\text{on}}(t)}{n} + \frac{q_{\text{off}}(t)}{n}, \quad j = 1, 2, \ldots. \quad (B.8) \]

### B.1.3 FOTO Model for Several Road Detectors

So far, we have discussed the tracking of the upstream front of synchronized flow between a detector A (Fig. 21.6) where traffic is still in the “free flow” phase and a detector B where synchronized flow has initially been detected. Once the upstream front of synchronized flow has reached the detector A at a time \( t = t_{\text{syn A}} \), the further propagation of synchronized flow can be tracked by monitoring the traffic volume since \( t = t_{\text{syn A}} \) between a next upstream detector C and the detector A. With the new detector distance \( D^* \) between the detectors C and A, the same method as for the tracking of the upstream synchronized flow front between the detectors A and B is used. However, this method can only be used as long as traffic at the detector A is in the “synchronized flow” phase, i.e., no wide moving jam occurred at the detector A.

When the upstream supply of vehicles flowing into the region of synchronized flow from the main road decreases, the upstream front of synchronized flow can begin to move downstream until it reaches the effectual bottleneck and synchronized flow dissolves. In the course of this process, it can turn out that the upstream front of synchronized flow moves downstream across the detector A at time \( t = t_{\text{free A}} \) but traffic is still in the “synchronized flow” phase at the next downstream detector B. In this case, the same method is used as that for the tracking of the upstream front of the expanding synchronized flow between the detectors A and B (Sect. 21.4.3). The difference is that in (21.11) the time \( t_{\text{free, B}} \) is replaced by \( t_{\text{free, A}} \), and the resulting coordinate of the upstream front of synchronized flow \( x_{\text{up}}^{(\text{syn})} \) (21.13) is replaced by

\[ x_{\text{up}}^{(\text{syn})}(t) = -D + \mu \Delta M(t). \quad (B.9) \]

### B.1.4 Extended Rules for FOTO Model

The basic set of fuzzy rules discussed above in Sect. 21.3.1 enables us an accurate classification of the traffic phases in almost all cases. For the recognition of wide moving jams, those rules use the fact that both the flow rate and average vehicle speed are at a low level within a wide moving jam.

However, there are cases when the flow rate drops abruptly within a moving jam, but remains at a relatively high value. If this flow rate remains above
800 vehicles/h, it has a higher degree of membership in the fuzzy set “high” than in the fuzzy set “low” with the basic fuzzy rules (Fig. 21.3). Figure B.3 shows such an example. The drop in average vehicle speed and flow rate at detector D3 at 8:29 is classified as “synchronized flow” by the basic rules because the flow rate does not get lower than 880 vehicles/h at 8:29. The correct classification would be “wide moving jam”: this wide moving jam propagates upstream and reaches D2 at 8:35 where it is correctly classified even with the basic fuzzy rules.

This problem cannot be solved by simply using different membership functions for the flow rate because flow rates of 880 vehicles/h can also occur in synchronized flow. Empirical analyses suggest that there is a certain range of vehicle speeds and flow rates where a distinction between “synchronized flow” and “wide moving jam” phases can be made by considering recent changes in time of flow rate and average vehicle speed in addition to the current values of flow rate and speed. This was realized with an extended set of 13 fuzzy rules that classifies the wide moving jam in Fig. B.3 correctly.

The range of vehicle speeds where the value change of speed has to be considered is between the fuzzy sets “low” and “medium” in the basic fuzzy rules. Therefore, these fuzzy sets have been split into the three new sets...
“very low,” “low,” and “medium” in the extended fuzzy logic (Fig. B.4a). The fuzzy set “high” remains the same as those in the basic fuzzy logic (Figs. 21.3a and B.4a). The set “low” in the new fuzzy logic is the range of vehicle speeds where the value change of speed is considered. Accordingly, the flow rate is classified in the new sets “very low,” “low,” and “high” in the extended fuzzy logic, where “low” is the range where the value change of flow rate is considered (Fig. B.4b). In the case when the vehicle speed is “low” and the flow rate is “low,” it is taken into account that the upstream front of a wide moving jam is accompanied by a sharp drop in both vehicle speed and flow rate, whereas the downstream jam front is accompanied by an abrupt rise in vehicle speed and flow rate at a detector. The relative changes in flow rate and vehicle speed are used as an additional input to the fuzzy logic circuit. To account for fluctuations in the vehicle speed and flow rate measured at detectors, the maximum differences in flow rate and in speed within a higher time interval before the last measurement is considered rather than taking just these differences between the last two measurements.

Usually a discrete sequence of measurements is available for the FOTO model. Let us further consider two such consecutive times \( t_0 \) and \( t^* \) \((t_0 > t^*)\) within an interval \( \Delta t \). If the flow rate has dropped between the time \( t^* \) and the time \( t_0 \), i.e., \( q(t_0) < q(t^*) \), the difference between the measurement, \( q(t_0) \) at the time \( t_0 \), and the maximum value measured during the whole time \( \Delta t \), \( \max_t(q(t)) \), is considered as a value “diff” in Fig. B.5a. If the flow rate has
Fig. B.5. Difference in flow rate “diff” (down-arrows for “diff”<0 and up-arrows for “diff”>0) when there are three measurements within the interval Δt. (a) Flow rate has dropped in the last interval (situations F1 and F2). (b) Flow rate has risen in the last interval (F3 and F4). Taken from [238]

risen between $t^*$ and $t_0$, the difference between the measurement $q(t_0)$ and the minimum flow rate $\text{min}_t(q(t))$ measured during the whole time, $\Delta t$, is considered as a value “diff” in Fig. B.5b.

The input variable $d_q(t_0)$ to the fuzzy-logic circuit is the change of the flow rate (diff) relative to $\text{max}_t(q(t))$ if the flow rate has dropped and relative to $q(t_0)$ if the flow rate has risen, i.e.,

$$d_q(t) = \frac{q(t_0) - \text{max}_t(q(t))}{\text{max}_t(q(t))},$$  \hspace{1cm} (B.10)

if $q(t_0) \leq q(t^*)$ and

$$d_q(t) = \frac{q(t_0) - \text{min}_t(q(t))}{q(t_0)},$$  \hspace{1cm} (B.11)

if $q(t_0) > q(t^*)$.

The same method is used to calculate the relative change of vehicle speed, $d_v(t_0)$:

$$d_v(t) = \frac{v(t_0) - \text{max}_t(v(t))}{\text{max}_t(v(t))},$$  \hspace{1cm} (B.12)

if $v(t_0) \leq v(t^*)$ and

$$d_v(t) = \frac{v(t_0) - \text{min}_t(v(t))}{v(t_0)},$$  \hspace{1cm} (B.13)

if $v(t_0) > v(t^*)$.

Another input variable to the extended fuzzy-logic circuit is the ratio

$$c(t_0) = \frac{d_v(t_0)}{d_q(t_0)}$$  \hspace{1cm} (B.14)

of the relative change of vehicle speed to the relative change of flow rate. In the case of negative $d_v(t_0)$ and negative $d_q(t_0)$ (a potential phase transition either from free flow or from synchronized flow to a wide moving jam), a low value of $c(t_0)$ indicates that the drop in flow rate was relatively stronger
than the drop in vehicle speed. Empirical analyses have shown that this is a sign that the phase transition to a wide moving jam did indeed occur whereas a high value of $c(t_0)$ indicates that it is less likely that the transition has occurred. The membership functions of the additional variables $d_v(t_0)$, $d_q(t_0)$, and $c(t_0)$ (Fig. B.6) have been chosen empirically by comparing the quality of results achieved with different functions.

In the case of a possible transition from a wide moving jam to another traffic phase (free flow or synchronized flow), i.e., when both $d_v(t_0)$ and $d_q(t_0)$ are positive, a similar analysis concerning $c(t_0)$ has been made. If the rise in vehicle speed is relatively sharper than the rise in flow rate, i.e., if $c(t_0)$ is high, traffic is likely to have remained in the “wide moving jam” phase while a transition to another traffic phase is likely to have occurred if $c(t_0)$ is low.

As a result, the whole set of fuzzy rules for the online determination of traffic phases based on local detector measurements consists of 13 rules [238]:

(1) If the vehicle speed is “high,” the traffic phase is “free flow.”

![Fuzzification of the input variables in the FOTO model with the extended fuzzy logic. (a) $d_v(t_0)$. (b) $d_q(t_0)$. (c) $c(t_0)$]
B.2 Evaluation of Reduced Detector Configurations

(2) If the vehicle speed is “medium,” the traffic phase is “synchronized flow.”

(3) If the vehicle speed is not “high” and the flow rate is “high,” the traffic phase is “synchronized flow.”

(4) If the vehicle speed is “very low,” the traffic phase is “wide moving jam.”

(5) If the vehicle speed is “low” and the flow rate is “very low,” the traffic phase is “wide moving jam.”

(6) If the previous (i.e., at the last time interval) traffic phase was not “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ and $d_q(t_0)$ are “negative” and $c(t_0)$ is “low,” then the traffic phase is “wide moving jam.”

(7) If the previous traffic phase was not “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ is “negative” and $d_q(t_0)$ is “negative” and $c(t_0)$ is “high,” then the traffic phase is “synchronized flow.”

(8) If the previous traffic phase was not “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ is not “negative,” then the traffic phase is “synchronized flow.”

(9) If the previous traffic phase was not “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_q(t_0)$ is not “negative,” then the traffic phase is “synchronized flow.”

(10) If the previous traffic phase was “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ is “positive” and $d_q(t_0)$ is “positive” and $c(t_0)$ is “low,” then the traffic phase is “wide moving jam.”

(11) If the previous traffic phase was “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ is “positive” and $d_q(t_0)$ is “positive” and $c(t_0)$ is “high,” then the traffic phase is “wide moving jam.”

(12) If the previous traffic phase was “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ is not “positive,” then the traffic phase is “wide moving jam.”

(13) If the previous traffic phase was “wide moving jam” and the vehicle speed is “low” and the flow rate is “low” and $d_v(t_0)$ is not “positive,” then the traffic phase is “wide moving jam.”

B.2 Statistical Evaluation of Different Reduced Detector Configurations

For the online field trial made from June 13, 2000 to October 22, 2000 results of the statistical evaluation of the FOTO and ASDA models under different reduced detector configurations have been performed in [493]. This evaluation
has shown a very high practical relevance of the FOTO and ASDA approach for the recognition and tracking of spatiotemporal congested patterns on freeways.

Here we discuss results of FOTO and ASDA applications for different reduced detector configurations [238].

To find out about cost-efficient measurement infrastructures that still lead to good results with the FOTO and ASDA models, the following approach has been chosen. The full detector configuration at the A5-North (the detector configuration “Full” in Fig. B.7, A5-North) is reduced to only about 23% of the detectors (77% of the measurements have been omitted for the FOTO and ASDA models). These omitted detectors have been used for the evaluation of the results performed with reduced detector configurations.

The remaining 23% of detectors are chosen differently, i.e., two different reduced detector configurations have been studied:

**A5-North: Full**

![Diagram of A5-North: Full](image1)

**A5-North: Min1**

![Diagram of A5-North: Min1](image2)

**A5-North: Min2**

![Diagram of A5-North: Min2](image3)

**Fig. B.7.** Different detector configurations: the configuration “Full” with all available 30 detectors on the main road, the configuration “Min1” with five main road detectors and detectors at on- and off-ramps, and the configuration “Min2” with eight main road detectors. Taken from [238]
B.2 Evaluation of Reduced Detector Configurations

(1) detectors between freeway intersections (between potential effectual bottlenecks) plus detectors at on- and off-ramps have been used (the “Min1” detector infrastructure in Fig. B.7);

(2) only main road detectors located close to the intersections without detectors at ramps have been chosen (the reduced “Min2” detector infrastructure in Fig. B.7). Both configurations “Min1” and “Min2” are oriented at complete balancing of vehicle numbers, i.e., there are no large detector distances with undetected on- and/or off-ramps in between.

All the following analyses use 15,300 empirical objects (about 4,800 “wide moving jam” objects and about 10,500 “synchronized flow” objects) that have occurred on the section of the freeway A5-North from June 13, 2000 to October 22, 2000. The mean lifetime of the “wide moving jam” object is about nine minutes whereas the mean lifetime of the “synchronized flow” object is about six minutes.

It is important to note that for both configurations “Min1” and “Min2” the detection rate for wide moving jams is close to 50% of the case of the detector configuration “Full,” while for all “synchronized flow” objects the detector configuration “Min2” detects almost 42% of the detector configuration “Full” and the detector configuration “Min1” only 26%. In other words, choosing the detector configuration in a more optimal way (“Min2”), in the cited example, only 23% of all detectors are sufficient to detect and track more than 40% of both kinds of traffic objects.

The detection rate for wide moving jams increases slightly from about 50% to 61% in the detector configuration “Min2” if only long living “wide moving jam” objects (lifetime is longer than five minutes) are considered. The detector configuration “Min1” shows a similar increase.

This effect is more significant for “synchronized flow” objects: the detection rate in the detector configuration “Min2” increases from about 42% to 64% if only long living “synchronized flow” objects are considered (lifetime is longer than five minutes). In the detector configuration “Min1,” however, this detection rate increases only from 26% to 33%. Thus, the detector configuration “Min2” detects about 2/3 of long living “synchronized flow” and “wide moving jam” objects (lifetime is longer than five minutes) detected and tracked by the detector configuration “Full”.

For many applications (e.g., routing) a time delay due to traffic congestion is an important characteristic. This time delay is calculated by contributions of each single traffic object with its separate width (in the longitudinal direction) and average speed for the entire stretch of the A5 freeway (Fig. B.7) at each single minute cycle. For the 4-month field trial the detector configurations “Min1” and “Min2” detect almost 2/3 of the time delays due to wide moving jam occurrence. For “synchronized flow” objects the detector configuration “Min1” shows only one third of the total (for the configuration “Full”) time delays due to synchronized flow occurrence, whereas the detector configuration “Min2” is still at two thirds of the total time delays for
“synchronized flow” objects. Thus, only 23% of the detectors estimate up to two thirds of the full measurement infrastructure, but the advantage of the detector configuration “Min2” over the configuration “Min1” is smaller for “wide moving jams” objects and bigger for “synchronized flow” objects.

An illustration of these conclusions for data from October 19, 2000 on the A5-North freeway is shown in Fig. B.8. The configuration “Min2” enables us to detect “synchronized flow” objects within to higher accuracy than that for the configuration “Min1.” For this reason, the travel time in the configuration “Min2” comes closer to the travel time in the configuration “Full” (Fig. B.8d). After 17:00 both reduced configurations “Min1” and “Min2” are less accurate in comparison with the configuration “Full” because at 12–15 km new wide moving jams occur that are not detected. In free flow, the travel time would be about 15 min (at 120 km/h), i.e., in congested traffic in Fig. B.8 it is approximately twice as much as in free flow.
References

25. R. Wiedemann: Simulation des Verkehrsflusses (University of Karlsruhe, Karlsruhe 1974)
27. M. Cremer: Der Verkehrsfluss auf Schnellstrassen (Springer, Berlin 1979)
47. A.D. May, P. Athol, W. Parker, J.B. Rudden: Highway Research Record 21, 48–70 (1963)
61. B.N. Persaud, S. Yagar, R. Brownlee: Transportation Research Record 1634, 64–69 (1998)
63. F.L. Hall, D. Barrow: Transportation Research Record 1194, 55–65 (1988)
70. J.H. Banks: Transportation Research Record 1287, 20–28 (1990)
75. P. Hsu, J.H. Banks: Transportation Research Record 1398, 17–23 (1993)
76. J.H. Banks: Transportation Research Record 1510, 1–10 (1995)
84. J. Treiterer: Transportation Research 1, 231–251 (1967)
93. P. Athol: Highway Research Record 72, 58–87 (1965)
94. P. Athol: Highway Research Record 72, 137–155 (1965)
120. M.E. Goolsby: Highway Research Record 349 (1971)
134. D. Branston: Transportation Science 10, 125–148 (1976)
150. C.R. Bennett, C.M. Dunn: Transportation Research Record 1510, 70–75 (1995)
159. V.F. Hurdle, M.I. Merlo, D. Robertson: Transportation Research Record 1591, 7–13 (1997)
160. R.T. Luttinen: Transportation Research Record 1365, 92–97 (1992)
161. F.L. Hall, W. Brilon: Transportation Research Record 1457, 35–42 (1994)
201. Y. Kim, H. Keller: Straßenverkehrstechnik, No. 9, 433–442 (2001)
References

211. B.S. Kerner: Physics World 12 (8), 25–30 (August 1999)
215. B.S. Kerner: Networks and Spatial Economics, 1, 35 (2001)
222. B.S. Kerner: ‘Three-Phase Traffic Theory’. In: [45]


251. J.A. Miller, R. Rothery: Transportation Science 1, 81–94 (1967)
256. N.H. Gartner: Highway Research Record 445, 12–23 (1973)
262. W.B. Cronje: Transportation Research Record 905, 93–95 (1983)

299. M. Ben-Akiva, M. Bierlaire: ‘Discrete Choice Methods and Their Applications to Short Term Travel Decisions’. In: Transportation Science Handbook (Kluwer 1999)


References


333. R. Thom: Structural Stability and Morphogenesis (Benjamin, Reading, MA 1975)


361. A. Reuscher: Z. Österr. Ing.-Archit.-Ver. 95, 73–77 (1950)
References

451. S.P. Hoogendoorn: Multiclass Continuum Modelling of Multilane Traffic Flow (Delft University of Technology, Delft 1999)
453. S.P. Hoogendoorn, P.H.L. Bovy: Networks and Spatial Econ. 1, 137–166 (2001)
454. S.P. Hoogendoorn, P.H.L. Bovy: ‘Short-term prediction of traffic flow conditions in a multilane multiclass network’. In: [41] pp. 625–651
455. J.P. Lebacque, J.B. Lesort: ‘Macroscopic traffic flow models: A question of order’. In: [40] pp. 3–25
456. J.P. Lebacque: ‘A two phase extension of the LWR model based on the boundedness of traffic acceleration’. In: [40] pp. 697–718
457. C.F. Daganzo: ‘The lagged cell-transmission model’. In: [40] pp. 81–104
459. H.M. Zhang: Networks and Spatial Econ. 1, 9–33 (2001)
526. M.G.H. Bell, Y. Iida: Transportation Network Analysis (Wiley & Sons, Chichester 1997)


552. N. Rouphail, A. Tarko, J. Li: ‘Traffic flow at signalized intersections’, In: [31], chapter 9


References

555. B.L. Allen, G.F. Newell: Transportation Science 10 No. 3 (1976)


Index

alternating general pattern, 350
alternating SP, 352, 453
amplitude of local perturbation, 637
automatic cruise control, 624
autosoliton, 15
average flow rate in pinch region, 325
averaging time interval for traffic variables, 274
boundary for spontaneous SP emergence at bottleneck, 182
boundary in diagram of congested patterns for spontaneous wide moving jam emergence in synchronized flow, 182
breakdown flow, 278
breakdown phenomenon, 2, 32, 108, 635
capacity drop, 2, 232, 361
catch effect, 37, 285, 495
cellular automata traffic flow model, 401
characteristic feature of wide moving jams, 28
characteristic parameter, 15, 636
characteristic spatiotemporal features of traffic phases, 27
collective flow, 68
components of time delay of F→S transition, 140
congested bottleneck, 278
congested pattern capacity, 232
congested pattern control approach, 594
congested pattern dissolution, 352
congested pattern evolution, 349
congested pattern transformation, 42, 349, 354
control parameters of traffic, 30
cooperative driving, 7
critical amplitude of deterministic perturbation, 131
critical amplitude of local perturbation, 637
critical branch on Z-characteristic for F→S transition at bottleneck, 136
critical local perturbation, 108, 637
critical point of free flow, 106
critical point of free flow for deterministic F→S transition at bottleneck, 130
critical range of traffic variable, 638
critical spatial form of local perturbation, 108
critical traffic variable, 637
definition of congested traffic, 22
definition of freeway capacity at bottleneck, 219
degree of strong congestion, 503
density in free flow within deterministic perturbation at bottleneck, 127
density in traffic flow formed by wide moving jam outflow, 201
dependence of empirical fundamental diagram on spatiotemporal structure of congested pattern, 385
desired vehicle speed, 401
deterministic F→S transition, 131
deterministic local perturbation, 34, 105, 124, 637
diagram of congested patterns at bottleneck, 46
discharge flow rate, 190, 279, 330, 357
discharge of congestion, 279
dissolution of moving jams at bottlenecks, 204, 369, 576
dissolving GP, 198, 357, 462
double Z-characteristic of F→S→J transitions at bottleneck, 176
double Z-shaped density dependence of speed, 161
downstream front of moving jam, 2
downstream bottleneck, 38, 365, 495
downstream congested pattern, 495
downstream congestion, 495
downstream front of congested pattern, 179
downstream front of synchronized flow, 28, 242
dynamic synchronized flow states, 101, 437, 456
effective location of bottleneck, 179, 243, 636
effectual bottleneck, 32, 179, 242, 635
empirical fundamental diagram, 373
empirical limit point of free flow, 18
equilibrium states, 63
expanded congested pattern, 40, 46, 366, 496

F→J transition, 41, 635
F→S transition, 32, 34, 635
F→S→J transitions, 45, 635
first-order local phase transition, 108, 639
floating car data, 574
flow rate in traffic flow formed by wide moving jam outflow, 201
flow–density relationship, 373
foreign wide moving jam, 316, 495, 498
free flow control approach, 593
free flow traffic phase, 22, 27
freeway bottleneck, 2, 635
freeway capacity in free flow at freeway bottleneck, 226
frequency of moving jam emergence, 45
front between different traffic phases, 634
front between different traffic states, 66, 634
fundamental diagram approach, 64
fundamental empirical congested pattern features, 59
general pattern (GP), 42, 636
general pattern emergence, 329
GP of type (1), 195, 259
GP of type (2), 195, 259
GP under strong congestion condition, 339
GP under weak congestion condition, 339
heterogeneous traffic flow, 509
historical time series, 568
homogeneous model states, 63
homogeneous-in-speed states of synchronized flow, 296
hysteresis effect, 36, 108, 638
hysteresis loop, 36, 108, 638
identification of traffic phases, 548
incentive lane changing conditions, 420
induced F→S transition, 34, 495
induced phase transition, 634
induction loop detectors, 16
infinity of freeway capacities at bottleneck, 226, 473
intensification of downstream congestion, 504
interruption of traffic flow by wide moving jam, 203
isolated effectual bottleneck, 179, 247, 635
J→F transition, 635
J→S transition, 635
jam suppression effect, 51, 209, 317, 323, 329, 461
lane asymmetric emergence of congested patterns, 519
lane separation effect, 516
limiting (minimum) flow rate in pinch region, 325
line J, 28, 71
line J and empirical fundamental diagram, 375, 376
local deterministic perturbation, 127
local perturbation, 637
local phase transition, 634
localized SP, 255
location of effectual bottleneck, 242

matching of traffic data, 573
maximum density at empirical limit
point of free flow, 18
maximum flow rate at empirical limit
point of free flow, 18
maximum freeway capacities at
bottleneck, 219
mean duration of wide moving jams, 208
mean time between downstream fronts
of wide moving jams, 208
mean time between narrow moving
jams, 208
merger of moving jams, 323
metastability effect, 111
metastable state of free flow with
respect to F→S transition, 114
metastable state of free flow with
respect to wide moving jam
emergence, 31, 69
metastable traffic state, 638
methodology of empirical congested
pattern study, 58
minimum freeway capacities at
bottleneck, 224
model for automatic feedback on-ramp
metering, 596
model for automatic identification of
traffic phases and tracking of
synchronized flow (FOTO), 548
model for automatic on-ramp control of
congested patterns, 607
model for automatic tracking of moving
jams (ASDA), 548
model for traffic prediction in city
networks, 583
moving jam, 2
moving jam emergence in synchronized
flow, 205
moving SP, 256

narrow moving jam, 43, 76, 194
nucleation effect, 111, 639
nucleation-interruption effect, 175
nucleus for local phase transition, 639

objective criteria for traffic phases, 3,
27, 89, 633
occupancy, 597
onset of congestion, 32, 34
optimal occupancy, 596

phone probe data, 575
physics of deterministic onset of
congestion at bottleneck, 133
physics of moving jam emergence in
synchronized flow, 174
physics of onset of congestion, 119
physics of onset of congestion at
bottleneck, 139
physics of S→F transition, 119
pinch effect, 43, 194, 205, 260, 321
pinch region, 43, 194, 260, 322, 459
pre-discharge flow rate, 278
prediction of spatiotemporal congested
pattern, 563
probabilistic freeway capacity, 218
probabilistic freeway capacity at
bottleneck, 218
probabilistic nature of breakdown
phenomenon, 32
probability density for phase transition,
637
probability for F→S transition, 110
probability for F→S transition at
bottleneck, 220
probability for F→J transition, 150
probability for phase transition, 637
probability for S→J transition, 156
probability of breakdown phenomenon,
32
probability of passing, 115

queuing theories, 84

random local perturbation, 637
region of wide moving jams of GP, 194,
196
region of wide moving jams of GP of
type (1), 196
reproducible empirical spatiotemporal
congested pattern features, 5

S→F transition, 36, 119, 635
S→J transition, 45, 194, 635
safe speed, 401
saturation effect in narrow moving jam emergence, 333
security lane changing conditions, 421
self-compression of synchronized flow, 207, 260
self-organization in traffic flow, 639
shortened GP, 371
space gap between vehicles, 16
spatially separated general patterns, 51
speed in free flow within deterministic perturbation at bottleneck, 127
speed adaptation effect in synchronized flow, 100
speed correlation function, 335
spontaneous breakdown phenomenon, 34
spontaneous F→S transition, 34, 109
spontaneous phase transition, 634
stable high density synchronized flow states, 156
stable traffic state, 638
stationary model states, 63
steady states, 64, 95, 403
stop-and-go traffic, 46
strong congestion, 337
synchronization distance, 100, 401, 431
synchronized flow, 90, 634
synchronized flow pattern (SP), 35, 636
synchronized flow traffic phase, 27
synergetics, 6, 14

three-phase traffic theory, 3, 633
threshold boundary for speed breakdown at bottlenecks, 223
threshold densities for S→J transition, 158
threshold density, 111
threshold density in free flow for excitation and persistence of wide moving jam, 73
threshold flow rate, 111
threshold flow rate in free flow for excitation and persistence of wide moving jam, 73
threshold point for F→S transition, 111
threshold point in free flow for excitation and persistence of wide moving jam, 73

threshold traffic variables, 637
time delay of phase transition, 638
time delay in vehicle acceleration, 72, 415
time delay in vehicle deceleration, 416
time delay of deterministic F→S transition, 141
time interval for observing traffic flow, 110, 218
tracking of traffic phases, 548
traffic control center, 563
traffic control parameter, 89, 633
traffic phase, 27, 633
traffic state, 27, 633
traffic variable, 27, 633
types of phase transitions, 635

united pinch region, 504
united sequence of wide moving jams, 51, 318
upstream front of moving jam, 2
upstream bottleneck, 38, 365, 495
upstream congested pattern, 495
upstream congestion, 495
upstream front of synchronized flow, 35, 191

vehicle onboard autonomous traffic prediction, 580
vehicle over-acceleration, 119, 235, 419
vehicle over-deceleration, 69, 174, 236, 420
vehicle speed adaptation, 119

weak congestion, 339
wide moving jam, 90, 634
wide moving jam traffic phase, 27
widening SP, 250
width of pinch region, 210
width of synchronized flow, 35
width of wide moving jam, 31, 202

Z-shaped characteristic for F→S transition, 114
Z-shaped characteristic for S→J transition, 171