UNIVERSITY OF CALIFORNIA,
IRVINE

Impacts of Capacity Drop on Freeway Control

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Civil Engineering

by

Felipe Augusto de Souza

Dissertation Committee:
Associate Professor Wenlong Jin, Chair
Professor R. Jayakrishnan
Associate Professor Solmaz S. Kia

2018

DEDICATION

To my parents, siblings and grandparents.

TABLE OF CONTENTS

Page

LIST OF FIGURES **vi**

LIST OF TABLES **ix**

ACKNOWLEDGMENTS **x**

CURRICULUM VITAE **xi**

ABSTRACT OF THE DISSERTATION **xiii**

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Research Objectives . . . 5
- 1.3 Research outline . . . 7

2 Review of Traffic Flow and Capacity Drop **9**
- 2.1 Merge Bottleneck and the Capacity Drop Phenomenon . . . 10
- 2.2 Review of Traffic Flow Models . . . 15
  - 2.2.1 Microscopic Traffic Flow Models . . . 16
  - 2.2.2 Macroscopic Traffic Flow Models . . . 17
- 2.3 Capacity Drop Models . . . 28

3 Review of Freeway Control Methods **33**
- 3.1 Control Theory Concepts . . . 34
- 3.2 Review of Ramp Metering Algorithms . . . 43
  - 3.2.1 Local Control Algorithms . . . 44
  - 3.2.2 Coordinated Control Algorithms . . . 49

4 Impacts of capacity drop on equilibrium states of freeway corridors **58**
- 4.1 Introduction . . . 58
- 4.2 Equilibrium state definition and its Properties . . . 61
- 4.3 Traffic Flow Model . . . 66
  - 4.3.1 Link Transmission Model . . . 66
  - 4.3.2 Computation of Flows at Nodes . . . 68
  - 4.3.3 Properties of Equilibrium States in a Link and in a block . . . 70
- 4.4 Equilibrium, Stability and Reachability Without Capacity Drop . . . 77

4.4.1 Model Analysis . . . 77
4.4.2 Numerical Experiments . . . 87
4.5 Equilibrium, Stability and Reachability With Capacity Drop . . . 90
4.5.1 Model Analysis . . . 90
4.5.2 Numerical Experiments . . . 97
4.6 Conclusions . . . 100

5 Reachability and Stability of A Local Ramp Metering System 103
5.1 Introduction . . . 103
5.2 System Description and Model . . . 107
5.3 Why Ramp Metering? The Impact of the Capacity Drop on The Delay . . . 112
5.4 When Ramp Metering is effective? The Equilibrium States and Reachability Property . . . 113
5.4.1 The equilibrium states and their Behavior . . . 114
5.4.2 Reachability with dynamic metering rates . . . 117
5.5 How lead the system to the desired state? Closed-Loop Analysis . . . 121
5.5.1 PI-ALINEA . . . 121
5.5.2 Set-Point Specification and Equilibrium States . . . 122
5.5.3 Closed Loop Response and Stability . . . 123
5.5.4 Numerical Examples . . . 127
5.6 Validation of the Results in Cell Transmission Model . . . 129
5.6.1 Verification of Stability . . . 132
5.6.2 Reachability . . . 133
5.7 Conclusion . . . 134
5.8 Appendix - Stability Region Derivation . . . 136
5.8.1 No Capacity Drop Effect . . . 141

6 Integrating a Smith Predictor into Control of Freeways 146
6.1 The System Model . . . 149
6.2 Closed Loop Analysis . . . 152
6.2.1 Control Strategy . . . 153
6.2.2 The Dead-Time in the Control Loop . . . 154
6.2.3 The Smith Predictor . . . 156
6.3 Simulation Experiments . . . 162
6.3.1 Stability Region . . . 162
6.3.2 Robustness . . . 164
6.3.3 Performance . . . 165
6.4 Conclusion . . . 168

7 Practical Aspects and Validation of Results in Microsimulation Models 170
7.1 Simulation Model . . . 172
7.2 Calibration Procedure and Results . . . 177
7.3 Conclusion . . . 183

8 Concluding Remarks **184**
8.1 Summary 185
8.2 Conclusions 186
8.3 Open Questions for Future Research 188

Bibliography **191**

LIST OF FIGURES

Page

2.1 Schematic of the bottleneck at I405-N with Jeffrey Road in Irvine, CA . . . 10
2.2 Schematic of the bottleneck at I405-N with Jeffrey Road in Irvine, CA . . . 12
2.3 Observed data at location. Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depict upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130$K vphpl and graph (f) depicts the upstream an estimation of the upstream demand (dashed) in comparison to the flow on the upstream detector 13
2.4 Schematic of car-following models and variables 17
2.5 Fundamental diagram viewed as (a) flow-density and (b) speed-spacing relationships 18
2.6 Density profiles at different times in a homogeneous road. 19
2.7 Schematic of example scenario and fundamental diagram. 22
2.8 Density solution at different times. 23
2.9 Cell transmission model schematic at a particular time step 25
2.10 A fundamental diagram and its associated demand, $D(k)$ and supply, $S(k)$. 27
2.11 The fundamental diagram with projected densities with capacity drop. . . . 30
2.12 Modified demand function for bounded acceleration. Obtained from [71] . . . 32

3.1 A schematic representation of a control loop with a process, represented by G, and a controller, represented by C, a reference signal, and input and output disturbances. 35

4.1 Schematic of freeway with alternating on- and off-ramps. 61
4.2 Schematic of a building block and variables associated to boundary flows computation 68
4.3 Representation of state based on the active bottlenecks with $M = 2$ 83
4.4 Cumulative outflow for each of experiment with no-capacity drop. 89
4.5 Cumulative outflow for each case considering the capacity drop. 98
4.6 Queues and metering rates for each of the four cases. At bottom left the total outflow and bottom right the oblique cumulative curves. 101

5.1 Schematic of merge bottleneck and the model variables. 107
5.2 Fundamental diagram of upstream/downstream and merging segments. . . . 110
5.3 Arrival (continuous) and departure (dashed) cumulative curves. The outflow (departure) could be either at capacity, C, or at dropped capacity, $C(1 - \Delta)$. 113

5.4 Possible system transitions . . . 115
5.5 Bifurcation diagram: the set of equilibrium states for varying demands. . . . 118
5.6 Phase diagram $x(t)$x$v(t)$ with initial condition in $Y(0) = [0, v_1]^T$. 128
5.7 In (a) and (b) blue line is $K_i = \frac{15}{24}$ and dashed lines $K_i = \frac{17}{24}$. The stable case, blue line, it goes to $Y_1^\star$, while the unstable case it keeps oscillating. Note in the phase plane (a), it follows the same trajectory multiple times as it can be observed in (b) and the dashed lines became continuous. In (c) and (d) blue line corresponds to $K_i = \frac{1}{25}$, and dashed to $K_i = \frac{1}{23}$. Both follow similar trajectory, but when $K_i = \frac{1}{25}$, $Y_2^\star$ is a equilibrium point, whereas for $K_i = \frac{1}{23}$ it is not and goes to $Y_1^\star$ instead. . . . 143
5.8 $v_0 < v^\star$ . . . 144
5.9 $v_0 > v^\star$ . . . 144
5.10 Oscillating Trajectories for (a) $v_0 < v^\star$ and (b) $v_0 > v^\star$. . . . 144
5.11 Analytical stability region using LQM (straight lines) and dots are the stability through CTM simulation, blue for $\Delta = 10\%$ and red for $\Delta = 0\%$. . . . 144
5.12 Density (top left), demands and metering rates (top right), ramp and out fluxes (bottom left), and on-ramp queue (bottom right). . . . 145
5.13 Path that the system can follow depending on eigenvalues nature and initial condition. The blue lines represent complex eigenvalues, black real and negative, and red saddle ($K_i > 0$ and $K_p < 0$). The shaded area represents the region where the system with real and negative eigenvalues switches to the another state. . . . 145

6.1 The merge bottleneck: A freeway merging with an on-ramp with a downstream lane drop bottleneck. . . . 148
6.2 The triangular and a general traffic fundamental diagram, the relationship between flow and density. . . . 150
6.3 Proposed control structure with Smith Predictor and output filter to improve robustness as in [101]. It is assumed all blocks in discrete time or discretized. 156
6.4 Comparison between the stability region of SP-ALINEA with different α and PI-ALINEA for different segment lengths, L. . . . 163
6.5 Stability of the system for different case of controller and model uncertainties. The dashed lines is the PI-ALINEA for L ranging from $600m$ to $1500m$. The values of α from 0.92 to 0.98 is the SP-ALINEA with internal model with different free flow speeds. . . . 165
6.6 System response and metering rate when speed is overestimated. With $F(z) = 1$ for $t < 1000s$ and a second order filter $F(z) = \frac{1}{(0.9z+0.1)^2}$ for $t \geq 1000$. . . . 166
6.7 System response to disturbance without and with dead time compensation ($\alpha = 0.96$). Even with modeling errors, the response of the SP-ALINEA is faster than PI-ALINEA. . . . 167

7.1 Notation of a subject vehicle driving in vicinity of its neighbor vehicles. . . . 172
7.2 Study site with position of the loop detectors . . . 177

7.3 Observed data at location. Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depict upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130$K vphpl and graph (f) depicts the upstream demand used in the simulation. 179
7.4 Trade-off curve (Pareto Frontier) between occupancy (OE) and cumulative flow error 180
7.5 Results for calibration minimizing only one of the three objectives considered. 181
7.6 Calibration results vs. actual data. Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depicts upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130K$ vphpl and graph (f) depicts the density flow relationship with orange and blue denoting the upstream and downstream location respectively. 182
7.7 Simulation results vs. actual data for validation (Feb-1-2012). Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depicts upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130K$ vphpl and graph (f) depicts the density flow relationship with orange and blue denoting the upstream and downstream location respectively. 182

LIST OF TABLES

Page

4.1 Notation 62
4.2 Classification, condition and stability with respect to perturbation in demand and state for the different types of equilibrium. U refers to uncongested, C refer to congested, OC to over-critical and C to critical, $C^{i+1,-} = C^{i+1}(1 - \Delta_{i+1})$. Proofs are on appendix. 94

5.1 Notation 108
5.2 Demands 133
5.3 Performance Metrics for each case 134

6.1 Total time spent for SP-ALINEA, PI-ALINEA and No-Control. 167

ACKNOWLEDGMENTS

I am thankful to a lot of people that really helped me a lot during my academic steps.

Definitely Wenlong Jin, my advisor, helped and challenged me a lot during the last four years. We had several discussions that made me develop as researcher. I am really thankful for that.

Though as not as close before, I have always close to me three persons - Werner Kraus Jr., Rodrigo, and Eduardo Camponogara - that are always there to discuss and offer me meaningful feedback. This first time in a different university made me value much more the period at UFSC.

I have been lucky to have funding from Brazil through Science Without Border Program. It really helped me to focus on my research since the first day. I also want to thank the UCI staff and the ITS staff.

In ITS I had the privilege to work among very nice people. Among them, I would like to mention Daniel Rodriguez and Sarah Hernandez for being friendly and caring about the new students. Also each student of Professor Jin including Anupam, Qijian, Xuting, an Qinglong. In the last few weeks Pratik and Irene (she attended my practice presentation 4 times and providing very helpful feedback at every session!) had helped me a lot. Another person of great help is Suman. It is always hard to mention names one by one, nonetheless I will mention ITS friends that helped me in different ways: Karina, Danny, Koti, Lu, and Mariana. A very special thanks to Marjan to have supported me in several different ways in the last three years.

Ph.D. can be stressful at times and several friends have made my life smoother throughout this period. I want to mention Luiz Fernando, Aline, Pedro, Elise, Elisa, Julian, Maria, Mitra and many more. I also want to mention some friends in Brazil that gave me a lot of support during these years: Crici, Shrek, Diego, Andre, Klauss, Thiago, Kurupira, among others.

Definitely, I had a lot of family support to complete the Ph.D. They supported unconditionally since much before I joined my undergraduate studies. In a country that unfortunately cannot provide opportunities to its population, I had the privilege to take a riskier path during my life. I am thankful for that though I wish my country fellows to have the opportunity to pursue high level education, trying on entrepreneurship or developing whatever project they want to invest their time. Of course the help of my family goes much further that. A great of source of happiness is talking to my nieces when I miss them! I thank you all.

CURRICULUM VITAE

Felipe Augusto de Souza

EDUCATION

Doctor of Philosophy in Civil Engineering **2018**
University of California, Irvine *Irvine, CA*

Master of Science in Systems Engineering **2012**
Universidade Federal de Santa Catarina *Florianopolis, Brazil*

Bachelor of Science in Control and Automation Engineering **2008**
Universidade Federal de Santa Catarina *Florianopolis, Brazil*

RESEARCH AND PROFESSIONAL EXPERIENCE

Graduate Research Assistant **2014–2018**
University of California, Irvine *Irvine, California*

Systems Engineer **2012-2014**
Brascontrol *Barueri, Brazil*

Chief Technology Officer **2010-2012**
ATTA Trafego *Barueri, Brazil*

Graduate Research Assistant **2008–2010**
Universidade Federal de Santa Catarina *Florianopolis, Brazil*

REFEREED JOURNAL PUBLICATIONS

Distributed MPC for urban traffic networks: A simulation based performance analysis **2015**
Optimal Control Applications and Methods

Cost effective real-time traffic signal control using the TUC strategy **2010**
IEEE Intelligent Transportation Systems Magazine (Second author)

REFEREED CONFERENCE PUBLICATIONS

Integrating a Smith Predictor into Ramp Metering Control of Freeways **2017**
TRB Annual Meeting

System Performance and Controller Design of the PI-ALINEA Ramp Metering Scheme **2016**
TRB Annual Meeting

Distributed model predictive control applied to urban traffic networks: Implementation, experimentation, and analysis **2010**
IEEE International Conference on Automation Science and Engineering

SOFTWARE

pythonsimtraffic
Python packages with implementation of the models used on this dissertation. Not yet released.

PyTrans `https://pytrans.github.io/`
Implementation and tutorials of algorithms related to transportation

ABSTRACT OF THE DISSERTATION

Impacts of Capacity Drop on Freeway Control

By

Felipe Augusto de Souza

Doctor of Philosophy in Civil Engineering

University of California, Irvine, 2018

Associate Professor Wenlong Jin, Chair

An unfortunate feature of freeway traffic flow at merge bottlenecks is the capacity drop (CD) phenomenon. It refers to a drop in the bottleneck outflow when a queue forms upstream to that bottleneck compared to the outflow observed before the formation of the queue. While its causes and exact mechanism are still open questions, this research concerns in the impacts of CD and how to mitigate them.

The distinct features of CD in a freeway corridor are assessed based on the behavior of equilibrium states in a model capable of replicating CD. The impacts are unveiled by comparing the system properties with and without the CD. The main finding is that the highest outflow occurs under uncongested equilibrium; however, it may not be reachable depending on the demands and initial conditions.

The local ramp metering control is investigated into more details. CD imposes a hysteresis on the system response with respect to the demand level. Also, we analyze the system in closed loop considering ALINEA, a well-known control algorithm. We establish the stability range with respect to parameters which is a necessary requirement for the controller to be effective. Further, we propose an extension of ALINEA to enlarge the stability range mitigating a performance loss that occurs when the on-ramp and the bottleneck are far apart.

Essential aspects of ramp metering are better captured with microscopic models; however, there were few evidences that such models can replicates CD. To that end, we propose a parameter calibration procedure that ensures the underlying model properly captures CD. The approach is tested with loop detector data from a merge bottleneck in which the CD is consistently observed.

All results with different approaches point to the direction that the existence of CD imposes additional challenges on the system control. Fortunately, in most cases the effects of CD can be mitigated with a properly designed control strategy, such as the ones tested and proposed in this research.

Chapter 1

Introduction

The freeway is jammed and it backed
up for miles
This car is an oven and baking is wild
Nothing is ever the way it should be
What we deserve we don't get, you see

Iron Maiden (Man On the Edge)

1.1 Background

Across the transportation field, a key property of the various facilities is capacity. The Highway Capacity Manual (2010) defines capacity as "the maximum hourly rate at which persons or vehicles reasonably can be expected to traverse a point or a uniform section of a lane or roadway during a given time period under prevailing roadway, traffic, and control conditions" [93, Section 4-17]. Though some aspects are narrowed down for each application, this definition is open to some arbitrariness such as "reasonably expect", "prevailing conditions" and the period in which we can safely infer a "hourly rate". The ubiquitous usage of

the term capacity in transportation may transmit the notion that capacity is a deterministic and easily measurable variable.

Traffic engineering manuals, such as the Highway Capacity Manual, generally assumes that the system will serve at capacity whenever the demand into the system is higher or equal than capacity (e.g., see [93, 10-44]). Definitely, it is a fair assumption in several applications in transportation. For example, if a toll booth can process two vehicles per minute, the discharge rate will be exactly that if the incoming demand exceeds that level. Similar reasoning is applicable to gates at ports, recharging stations for electric vehicles, boarding and alighting of passengers at transit systems, to name a few.

It has long has been assumed the same behavior of capacity in urban traffic flow. Though arguably this concept is applicable in some cases, the application at freeway bottleneck is, at least, nuanced [1]. The capacity drop phenomenon phenomenon challenges this notion. This phenomenon is a consistent drop in the bottleneck outflow when queues are formed just upstream of that bottleneck compared to the outflow observed before the formation of that queue. In simple words "congestion causes more congestion". This may look intuitive in traffic, but it is not so obvious when we make analogies for the aforementioned cases in which the basic behavior of capacity applies. Let's consider again the aforementioned example of the tool booth that can discharges 2 vehicles per minute. If during a long period the demand is almost constant at 1.9 vehicles per minute in average, the system could experience some small queues due to the randomness of the arrivals, but the queues would not grow indefinitely as the facility can discharge more vehicles than the average arrival rate whenever there are queued vehicles. The analogy of the capacity drop in this case would be the operator processing only 1.8 vehicles per minute, as oppose as to 2 vehicles prior the formation of the queue, whenever the queue exceeds 10 vehicles. In that case, the system

[1]The very definition of bottleneck is related to capacity: a point in which the capacity is smaller than the region immediately upstream making making it the potential places to trigger a congestion. Typical bottlenecks at freeways are lane drops, but it can also occur at tunnel entrances and sags

can operates at full capacity for some period; however, as soon as a queue higher than 10 vehicles forms, the discharge flow rate would drop to a level smaller than the arrival flow rate and the queue grows indefinitely.

Luckily, this intriguing behavior does not happen at toll booths, but unfortunately it does happen at freeway bottlenecks. The majority of the research related to freeway traffic flow does not consider the existence of such phenomenon. For instance, the Highway Capacity Manual 2010 [93] briefly mentions the existence of the capacity drop at freeways bottlenecks, but has not changed the quantitative methods for freeways accordingly.

Two publications in early 1990's [10] and [46] are commonly referred as the founding publications around this topic. Both have reported a small drop (3-4%) in the flow in a freeway bottleneck after the onset of queue just upstream to the bottleneck under study. However, it was not the first time that similar fact was reported; this was a subject for discussion for 30 years [45, Section 2-14] and several research had found confounding conclusions regarding the existence of the capacity drop. For instance, in a publication of the same research group as one of the aforementioned publications ([46]), studied the same location one year before and claimed the capacity drop as a long-standing question as they describe that "Some studies have suggested that after a queue forms there is a drop in the maximum flow possible through a bottleneck. Wattleworth ([133], in 1963) discusses three such studies". Nonetheless, they later claim their study does not support this view as "Figures 3-6 indicate that, although there is a clear drop in speeds, there is no easily discernible drop in the flow rates at the time the queue forms. (...) Hence, a significant flow reduction under such circumstances cannot be deduced from these data." While they do not necessarily deny its existence, they say that Wattleworth [133] "provides a convincing explanation of why this result arises in some studies, whereas others continue to show a capacity reduction". In fact, by then the methods in which the data was processed and analyzed varied from study to study. Some of the differences were the placement of the detector, upstream or downstream to the bottle-

neck, the data sample time (30 seconds, 5 minutes, etc), and how to infer that the bottleneck was queued.

Later, in 1999, it was stated again the drop in the flow after the bottleneck become queued [17]. Unlike the previous studies in which was based on a time-series of flow, they used transformed cumulative flow curves [19] from both upstream and downstream detectors instead. The transformed cumulative flow curve turns the visualization and quantification of the capacity drop easier, perhaps closing the discussion surrounding the existence of the drop in the queued flow. They used data from the same location as the aforementioned study ([46]) and found consistent drop in the flows of around 9% as oppose to 3-4%. Arguably, the main contribution of the study is the method based on transformed cumulative curve solving the issue of inconsistent results regarding the capacity drop. Since then, several other reports have used such curve to confirm that capacity drop happens at various other locations. Nonetheless, the complete understanding of its mechanism and how we can mitigate its effects are still open questions. That perhaps explains why traffic manuals still have not yet adapted despite these findings.

This dissertation is devoted to the second question: how can we mitigate the effects of the capacity drop phenomenon? The implication of such phenomenon to freeway control, more specifically ramp metering, was object of Banks publication just following his previous study [10]. In [9] it was discussed whether the "Two-Capacity" (capacity drop) phenomenon is a basis for ramp metering. The author conclusion is "that is unlikely". Around the same time, field-test in Paris, France they stated that the application of ALINEA showed performance improvements [44] through ramp metering. Once again repeating the pattern of confounding conclusions around the same issue, even though both studies had reached a similar difference in the outflow (3%). Further implementation of ALINEA had also showed modest outflow increase in other sites [108].

Though these studies have suggested improvements of ramp metering control, the question

whether ramp metering can recover from the capacity drop was still a doubt. In [18] it was shown that the ramp metering control could recover a bottleneck from a congested state, discharging flow below its capacity, to a uncongested state discharging at capacity. That study used a simple strategy - set the most restrictive metering rate at the on-ramp once the freeway becomes congested - in order to confirm their hypothesis - which was in fact confirmed. Therefore, it has been shown that through ramp metering it is possible to recover higher outflows in a merge bottleneck.

Given that is possible to achieve higher outflows through ramp metering, the following question is what is the best strategy to do so. This is the central question in which the research documented here concerns. Though there is plenty of literature on ramp metering, it still is limited studies addressing the capacity drop phenomenon specifically. That is not surprising considering the sinuous path on the understanding of such phenomenon.

Nonetheless, there have been recent developments that make this research timely. Namely, there are traffic flow models that reproduces the capacity phenomenon. That enables the study of the impact of capacity drop on freeway through assessing analytical properties of the model such as stability and reachability as well as analyzing performance of ramp metering control through simulation based on that models. This research aims to reveal fundamental characteristics of freeway control related to capacity drop based on analysis of traffic flow models that reproduces the capacity drop phenomenon.

1.2 Research Objectives

The goal of this research is to identify the aspects of capacity drop that are relevant to different applications of freeway control strategy and how strategies should be designed in light of such phenomenon. To that end, we focus on fundamental control theory properties

such as stability, reachability and equilibrium state analysis. The objective is split into:

1. **Develop an analytical framework that considers both control variables and traffic flow models capable of reproducing the capacity drop phenomenon**

   with a traffic flow model that takes into account control variable, such as metering rate, and reproduces the capacity drop phenomenon, it is possible to infer about the interaction between the control strategy and the capacity drop phenomenon. The investigation can uncover aspects related to the (asymptotic) equilibrium behavior such as uniqueness, stability and reachability of these states.

2. **Propose and analyze control strategies able to appropriately handle the capacity drop phenomenon**

   For local ramp metering (i.e., a single bottleneck), the capacity drop phenomenon is the key aspect affecting the total delay experienced by the travelers. The follow question can potentially be answered: (i) why an on-ramp should be metered; (ii) if so, when the on-ramp should be metered; (iii) how the ramp should be metered; and, finally, (iv) at what conditions will the controller work.

3. **Validate the results in microscopic models**

   Assess whether the results of the aforementioned studies, based on macroscopic model, also holds in microscopic models. The confidence of the results in the previous analysis increases if the same result holds in different models. To that end, first it is necessary how microsimulation models can model the capacity drop for later analyzing it under ramp metering control.

1.3 Research outline

This dissertation is organized into eight chapters. This introduction had presented the background of this research as well as its specific goals. This background is further extended on chapter 2 in which relevant literature to this research is presented. This research is at intersection between control theory and traffic flow theory, a sub-section is devoted for each of these topics. Within traffic flow theory, it is presented the traffic flow models that is the foundation for the latter analysis. Also, the capacity drop phenomenon is further detailed and models able to reproduce it are discussed. In control methods, basic concepts and properties of control system is presented. Also, it is also presented selected control approaches that have been applied into freeway control.

Chapter 3 presents an analysis of the equilibrium states of a freeway based on the continuous link transmission model combined with the phenomenological capacity drop model as in [65]. The equilibrium states are identified and its behavior is characterized with respect to reachability, stability, and their relationship with performance metrics. By enabling or not the capacity drop model extension, the specific features of the capacity drop phenomenon are identified. The mathematical program for identifying the optimal equilibrium is derived for both cases and a method is provided for the case in which capacity drop cannot be avoided at all bottlenecks of a corridor. The results are validated with numerical experiments.

Chapter 4 presents the dynamic behavior of a single merge bottleneck controlled by PI-ALINEA [132]. The dynamics of the bottleneck is described based on an switched ordinary differential equation approximation of the Lighthill-Witham[86]-Richards [113] (LWR) model. Important characteristics of the system is disclosed such as the hysteresis imposed by the capacity drop and close solution for reachability and closed-loop stability. The results are validated in a discretization of the LWR model, the cell transmission model [27].

Chapter 5 proposes a control approach based control strategy suitable for the local ramp

metering problem, especially when the distance between the on-ramp and lane drop is long. In that scenario, the elapsed time between a control action is performed and the time it interferes the system, referred as to dead time, is long making the control design more challenging. A Smith Predictor, a long-standing approach in control theory for systems with long dead-time, is integrated into the controller to overcome such limitation. The introduction of the Smith Predictor provide two advantages: (i) it enlarges the stability range of the controller, which is analytically derived; and (ii) the system response is faster. Numerical experiments confirm the results.

Chapter 6 presents a calibration of microscopic model study in a merge bottleneck based on loop detector data. Morning-peak data from a merge bottleneck at I-405N was used to calibrate the various input parameters of the car-following lane-changing model considered. Gipps [37] car-following model and the lane-changing model used was by Hidas [53]. The results show the calibrated model is able to reproduce the capacity drop phenomenon with high accuracy over its different aspects.

Chapter 7 the research is summarized. The current challenges and potential directions for future work are discussed.

Chapter 2

Review of Traffic Flow and Capacity Drop

> It's like a shockwave to your brain. A voice that makes you go insane.
>
> Myrath (Shockwave)

In this chapter a brief review of aspects related to capacity drop and traffic flow models are presented. In the first section, the basic building block of this research - a merge bottleneck - is introduced. The main features of the traffic around a merge bottleneck including the empirical evidence of the capacity drop, are presented. Later, traffic flow models that can be used to describe the dynamics on the merge bottleneck are presented. Lastly, models specific to capacity drop is presented.

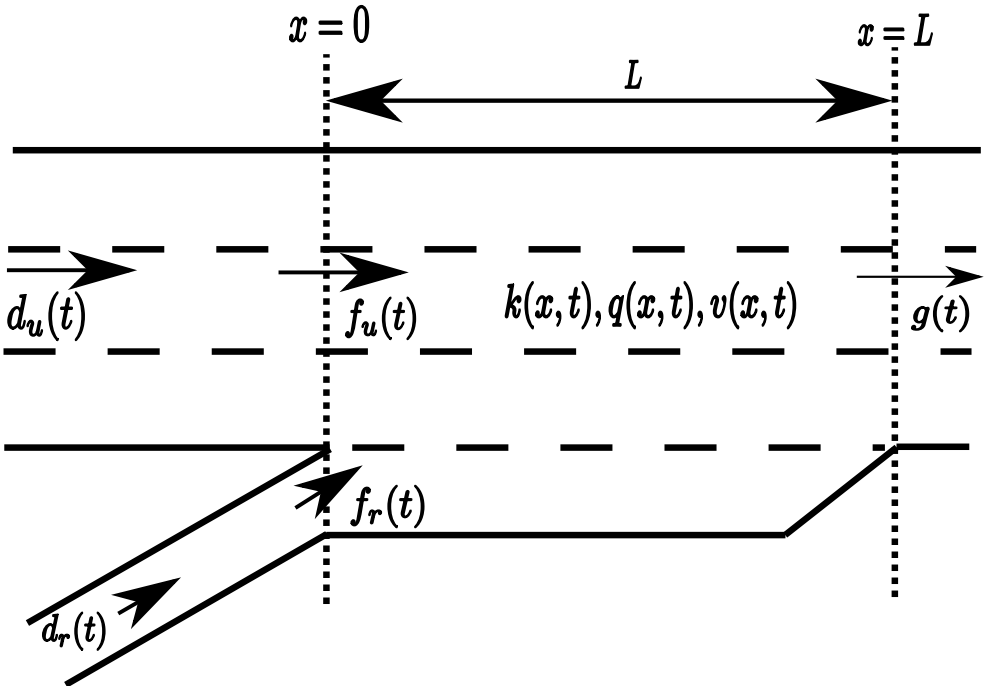


Figure 2.1: Schematic of the bottleneck at I405-N with Jeffrey Road in Irvine, CA

2.1 Merge Bottleneck and the Capacity Drop Phenomenon

The Figure 2.1 presents a schematic of a merge bottleneck or merging segment, as the Highway Capacity Manual [93] refers to. It contains a merge between the mainline freeway and on-ramp, and a bottleneck on the reduction of number lanes (lane drop). There might be other types of connection (freeway-freeway, etc) and other types of bottleneck (sag, tunnel, etc), but in this research we always assume a scenario exactly as depicted in the figure.

The freeway in its upstream section has N lanes which becomes $N + 1$ with the on-ramp acceleration lane. Further downstream the number of lanes reduces again to N. The length of the merge bottleneck is referred as L which we consider as the distance between the on-ramp and the lane drop. The position, x, grows in the direction of traffic stream and time as t.

The arrival demand coming from the upstream is denoted as $d_u(t)$ defined as number of vehicles per unit time. Similarly, the demand from the on-ramp is denoted by $d_r(t)$. Depending

on the traffic conditions inside the merge, part of the demand may be served with delays. The flows from on-ramp and mainline that in fact could join the merge zone are denoted as $f_r(t)$ and $f_u(t)$, respectively. The outflow is denoted as $g(t)$.

Objective measurements are in general derived from three key variables: mean speed, flow rate and density. Speed is related to the distance traveled per unit time by a single vehicle or a group of vehicle. Later the concept of mean speed is further detailed. Flow rate is the number of vehicles that pass a specific point per unit time. Density is the number of vehicles per unit time in a give section. All these variables can change in space and time and are tied through the following identity:

$$q(x,t) = k(x,t)v(x,t), \tag{2.1}$$

where q is the flow rate, k the density and v the space mean speed.

The adjective bottleneck to this specific scenario because it is a potential location to start congestion. Vehicles from the freeway needs to leave the acceleration lane as soon as they join the freeway due to the eventual end of the lane. If the incoming demand is high, the gap for changing the lanes become small in such way that vehicles at freeway right-most lane are forced to slow down if a vehicle in the acceleration lane perform a lane changing just ahead of them. That turns the right lane congested and the situation spreads to the middle and right lanes [18] as vehicles through similar mechanism, though the lane changing maneuvers happens for obtaining speed advantage as oppose to a mandatory lane-changing due to imminent lane end.

This instant in which the merge becomes congested is often referred to as *flow breakdown* [111]. Note however that flow breakdown is not necessarily a reduction in the flow rate of the

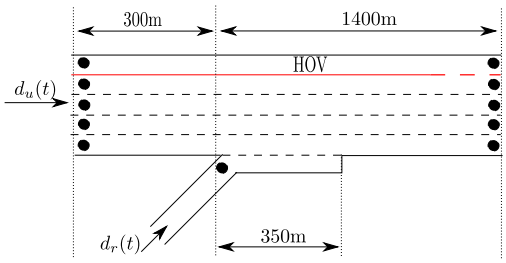


Figure 2.2: Schematic of the bottleneck at I405-N with Jeffrey Road in Irvine, CA

bottleneck. The term probably is coined after the break on the flow stream to a relatively homogeneous when congested to a specific location in which the flow is slower and more turbulent.

When the traffic is congested, it is qualitatively clear what happens with each of the variable in the right hand side of Equation 2.1. When uncongested, the mean speed is high and the density is low; while it is the opposite when congested. The key question is what happens with the flow? Or more specific to the merge bottleneck case: when congestion starts at the merge bottleneck, does it reduce the maximum flow through the lane drop ($x = L$ in Figure 2.1)? This is the question that was surrounding researchers for about 30 years [45].

As it was an empirical question, I will discuss about the capacity drop phenomenon based on collected data from a merge bottleneck rather than a mathematical description of the phenomenon. The location in which the data was collected is at I-405N with Jeffrey Road in Irvine, CA. A schematic of that location is depicted in Figure 7.2 with the location of on-ramp, upstream and downstream loop detectors. Each loop detector provides occupancy and flow rate (i.e., counts on the 30 seconds period) with 30 seconds sample time. Occupancy is share of time in which there was a vehicle on the top of detector during the period. With some assumptions, there is a direct relationship between occupancy and density. The important for now is that higher occupancy means higher density.

The data of the morning peak of April-19-2012, obtained through California Performance and Measurement System (PeMS) [118], is depicted in Figure 7.3. The High-Occupancy-Lane

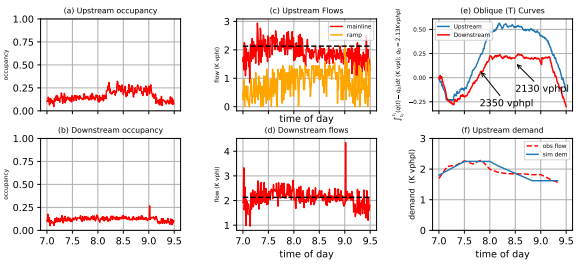


Figure 2.3: Observed data at location. Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depict upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130$K vphpl and graph (f) depicts the upstream an estimation of the upstream demand (dashed) in comparison to the flow on the upstream detector

data, though also available, are disregarded. The left graphs (a-b) show the occupancies upstream (top) and downstream (bottom). The observed counts are on the middle graphs. The upstream counts (c) include mainline and on-ramp. The downstream counts are according to the middle bottom graph (d). Note that the downstream is always uncongested as the downstream occupancy is almost constant throughout the period; however, high occupancies were observed at the upstream detector between 8:00 - 9:00 AM and decreased around 9:00 AM. It means congestion started at the bottleneck, reached the upstream detector, and dissipated when the upstream demand ceased with the congestion at the upstream detector being eliminated just after. The dashed line on the upstream and downstream flows (c-d) refers to the average flow at the downstream between 8:03 - 9:08 AM (q_0). Therefore, the flow breakdown happens around 8:20 AM, or at least, this is the time the queues grow until the upstream detector. The following question is whether there is any difference in the outflow curve following the onset of congestion.

Assisted by the the baseline flow, q_0, at graph (d) perhaps one could infer a flow slightly higher between 7:30AM and 8:00AM compared to the next 30 minutes period. It is not easy to discern, however, if that was due to an inherent variability of counts or a consistent decrease of the outflow.

However, it becomes clearer on the top right graph (e) in which the transformed cumulative

curve [19] (T-curve) is plotted. The T-curve, or transformed cumulative curve, is the area between the outflow and a baseline outflow, q_0. The baseline outflow is arbitrary and we chose it as the average flow during the congested period. This curve has positive slope whenever the flow is greater than q_0 and negative otherwise. Similarly to cumulative curves, the vertical distance between two T-curves is associated to number vehicles between the points and horizontal difference related to travel times.

Clearly, the outflow was smaller during the congested period compared to the 20 minutes interval before the drop in the downstream flow, a decrease from 2350 vphpl to 2130 vphpl (9.5%), as annotated on the graph. The vertical distance between the two curves increases between 7AM and 8AM on the congestion build up and decreases around 9AM when the congestion dissipates. This is smaller outflow following the onset of congestion is the capacity drop phenomenon [10, 46, 17].

Note how the phenomenon become clearer with the use of the T-curve compared the outflow time series. This may explain the long standing question regarding the capacity drop phenomenon until 1990's as the T-curve had not been proposed as a method of analysis. The use of T-curve is opportune because that circumvents some key issues related to loop detectors: (i) the traffic measurements are naturally noisy which turns the analysis more intricate; and (ii) the loop detector reports measurements from a specific point, but the traffic has a spatial dimension not easily captured by the occupancy data.

About the second issue, one would assume congestion starts around 8:20 AM based on the upstream occupancy. However, the flow drops down to q_0 just at 8:00AM. If one had identified 8:20AM as the initiation of congestion and compared the downstream flow in the 20 minutes preceding and following that instant, would have found no difference. With the T-Curve we can easily observe changes in the outflow by the variations in the slope. In addition, it is possible to infer queue formation by the vertical displacement between the upstream and downstream curve.

The capacity drop phenomenon was presented from a empirical point of view. In the next subsections a literature review of traffic flow models is presented with special attention on the aspects related to the behavior of the merge bottleneck and the capacity drop phenomenon.

2.2 Review of Traffic Flow Models

There are several approaches to model the dynamics of traffic flow. Nonetheless, most of model can be classified into one of the following categories [73]:

- **Microscopic**: refers to models in which each driver or pair vehicle-driver is described individually.
- **Macroscopic**: refers to models in which the dynamic is described based on aggregated variables (such as average density, flow) as oppose to track vehicles individually;
- **Mesoscopic**: refers to models in which coexist elements of microscopic and macroscopic models;
- **submicroscopic**: refers to models in which sub-items of the pair vehicle-driver is explicitly modeled such as throttle position and psychologically speed perception;
- **network level**: this approach is derived of macroscopic models in the sense that deals with aggregated variables, but the smallest element is an area (a set of roads) instead of a road or a section of road.

For different applications some type models may be more suitable than others. In particular to our basic building block, the merge bottleneck, both macroscopic and microscopic models can be applied. Following, each of this type of models are reviewed with higher attention to macroscopic models as most of this research was based in such models.

2.2.1 Microscopic Traffic Flow Models

Microscopic models describe the traffic with the vehicle as a basic unit. A given pair driver-vehicle takes decisions, such as acceleration and lane changing, based on the position, speed and acceleration of vehicles on its surrounding.

The main advantage of microscopic models is that the individual vehicle representation provides a detailed representation of traffic flow. The outputs of such models may include acceleration, speeds, position and lanes of each vehicle undertook throghout all period of study. Another advantage of macroscopic models is the simplicity in which heterogeneity between drivers and vehicles can be incorporated such as different vehicle class (passenger cars, trucks, buses,etc) and different driving behaviors within a class, such as different desired speeds and reaction times.

Nevertheless, there are disadvantages on using such models as they require more parameters to be defined and they are computationally more expensive so that is harder to apply to large networks. Also, it is not a straightforward task to model the human component of the driving behavior. Specifically, the lane changing models are a known weakness of microscopic models [141].

Though there might be more components, to study the merge bottleneck the microscopic models are combination of car-following and lane-changing models. Consider the subject vehicle in Figure 2.4 located at the right lane following another vehicle. The car-following component models the kinematic variables (acceleration, speed, position) of the subject vehicle based on the same variables of the leader vehicle. The lane-changing component models when and how the vehicles perform a lane changing based on the current leader and the potential leader and follower kinematic variables in the adjacent lane.

The first car-following models were proposed in the 1950's with Pipes [112] models. The

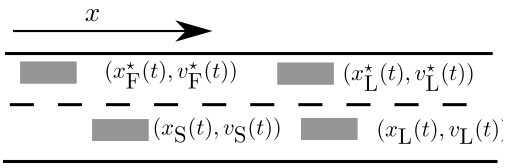


Figure 2.4: Schematic of car-following models and variables

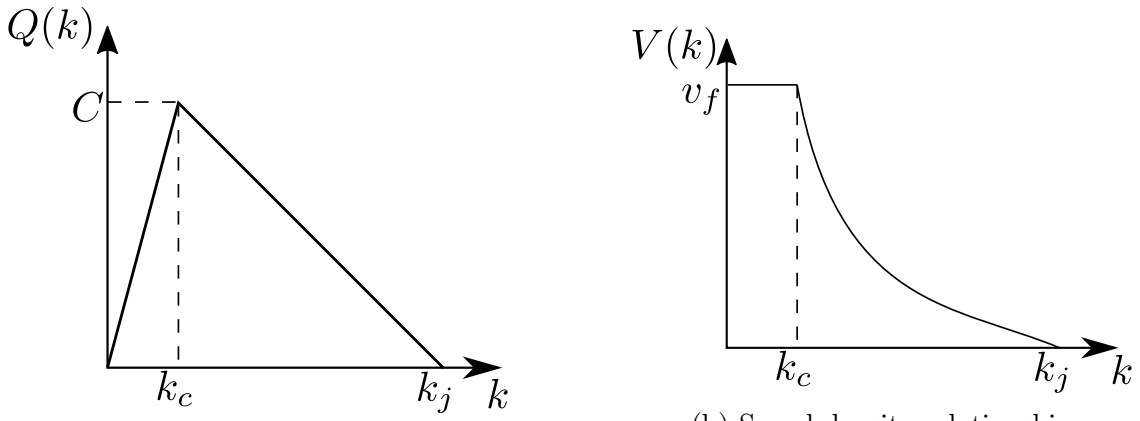
latter model is a class of models where based on principle in which drivers accelerate or brakes based on the relative speed and position with respect to the leader vehicle. This class of model was popular until the 1970's [12]. Later, the safe-distance models in which are based on the principle in which vehicles drivers adjust speeds so as to not collide with the leader. Among them, the Gipps [37] model is one of the most used car-following model.

Lane changing model development started relatively later than car-following model. The lane changing models in general model two sequential steps: (i) deciding to perform a lane changing and (ii) undertaking the actions to perform that decision. There are two popular class of models for structuring decisions. First is rule-based or Gipps-type in which the decision to change lanes are based on a set of rules. The driver considers two basic aspects: driving at desired speed and keeping in the correct lane to follow the intended route [141]. Relevant work includes [38, 138, 52]. Second is utility-based in which the various aspects of lane-changing are converted into an utility function and drivers takes the decision that maximizes its utility. Relevant work includes [3, 127].

2.2.2 Macroscopic Traffic Flow Models

In macroscopic models, the traffic is described based on aggregated variables either varying in time or in both space and time as oppose to tracking each vehicle individually. A significant share of the theory related to traffic flow model is inspired in fluid mechanics.

Though inspired in fluid mechanics, one distinct feature of traffic flow is the speed-density relationship. This relationship, first proposed by Greenshields [43], postulates that the



(a) Flow-density relationship

(b) Speed-density relationship

Figure 2.5: Fundamental diagram viewed as (a) flow-density and (b) speed-spacing relationships

average mean speed varies as function of the density (amount of vehicles per unit distance). From the driver point of view, it means that drivers adjust their speeds according to the spacing they experience with their respective leaders. As density, speed, and flow are tied by (2.1) this relationship can be seen as speed-density relationship or, as commonly used, flow-density relationship. In Figure 2.5 an example of fundamental diagram is depicted as flow-density (a) and speed-density relationship (b). In general it is assumed that speed is a non-decreasing function of density. The speed for density equal zero is referred to as free flow speed; similarly, the density in which yields zero speed is referred to as jam density. On these extremes, the flow is zero either because there is no car (zero density) or because the cars do not move (zero speed). Capacity, C, in the fundamental diagram is defined as the maximum outflow; the density that yields capacity flow is the critical density, k_c.

Like in fluid mechanics, the system is governed by a partial differential equation due to mass conservation and the flow identity (Equation (2.1)). Consider a homogeneous road with length L in which is given the densities profile at time t_1 and t_2 as depicted in Figure 2.6. The number of vehicles on the interval $x_1 \leq x \leq x_2$ is defined by $N(t)$ in which we can obtain through the integration of the density over space:

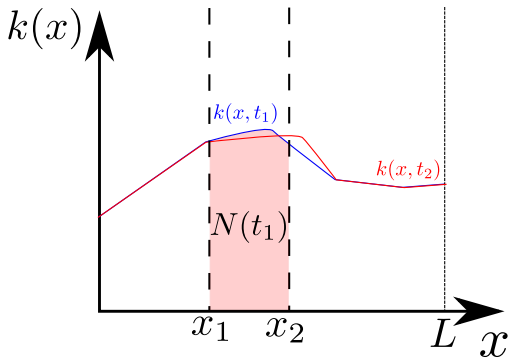


Figure 2.6: Density profiles at different times in a homogeneous road.

$$N(t) = \int_{x_1}^{x_2} k(x,t)dx. \tag{2.2}$$

Figure 2.6 depicts $N(t_1)$. Note by the density profiles that $N(t_2)$ is smaller than $N(t_1)$. For mass conservation principle, this reduction of number of vehicles means that the outflow was higher than the inflow during the period. Following the sample principle, we can describe the relationship at time t as:

$$\frac{d}{dt}N(t) = q(x_1,t) - q(x_2,t), \tag{2.3}$$

Substituting (2.2) and (2.3):

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} k(x,t)dx = q(x_1,t) - q(x_2,t) \tag{2.4}$$

Taking the limit of $x_1 \to x_2$ and using the mean value theorem on the left hand side:

$$\begin{aligned}
&\frac{\partial}{\partial t}k(x,t)(x_2 - x_1) = q(x_1,t) - q(x_2,t) \\
&\frac{\partial}{\partial t}k(x,t)(x_2 - x_1) = (x_2 - x_1)\frac{q(x_1,t) - q(x_2,t)}{x_2 - x_1} \\
&\frac{\partial}{\partial t}k(x,t) = -\frac{\partial}{\partial x}q(x,t) \\
&\frac{\partial}{\partial t}k(x,t) + \frac{\partial}{\partial x}q(x,t) = 0
\end{aligned} \tag{2.5}$$

Therefore traffic flow models are governed by a partial differential equation (2.5) which means that a variation of density in a point is an outcome of the inflow and outflow at that point. One of earliest model proposed in the literature based on the assumptions of fluid flow and mass conservation was the Lightwhill-Whitham -Richards (LWR) model back in the 1950's. It is the base of a plethora of models proposed ever since.

The Lighthill-Whitham[135]-Richards [113] describes the evolution of flow and density in space and time. Density, $k(x,t)$, is the state-variable and flow $q(x,t)$ is assumed to be a function of the density:

$$q(x,t) = Q(k(x,t)), \tag{2.6}$$

where $Q(k)$ is a flow-density relationship often referred as fundamental diagram as depicted in Figure 2.5 (a). This relationship is the unique aspect of traffic compared to fluids mechanics and it is derived from the assumption that drivers adjusts speed according to the spacing to its leading vehicle. With the fundamental diagram assumption the mass conservation

becomes:

$$\frac{\partial}{\partial t}k(x,t) + \frac{\partial}{\partial x}Q(k(x,t)) = 0, \tag{2.7}$$

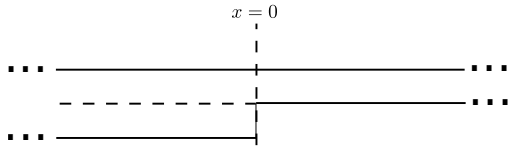
Equation 2.7 is able to describe the traffic dynamics in a homogeneous link (that is, the fundamental diagram is the same throughout the link). Giving boundary condition, in which in this case is $k(x, t = 0)$, $q(0, t)$ and $q(L, t)$ and solving (2.7), we obtain $k(x, t)$. Flows, $q(x, t)$, can be obtained through the fundamental diagram and speed by (2.1).

In order to apply this model to obtain future density and flows, it is also necessary to provide boundary conditions which includes the initial density $k(x, 0)$ and the boundary flows, $q(0, t)$ and $q(L, t)$.

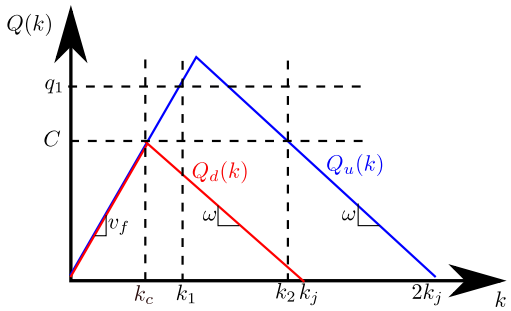
Several features of the LWR model resembles the characteristics of traffic empirically observed. Namely, it models in both time and space the initiation and dissipation of congestion through shock-waves. The solution of the LWR model leads to a piece-wise smooth densities as $t \to \infty$ with discontinuities between the pieces referred to as shock. These boundaries or shocks can travel backwards or forwards depending the density and flow upstream and downstream to that shock.

The basic study scenario of this research presents all these features and also will bring us the attention of the necessity of additional assumptions for cases in which the solution is not unique.

Let's consider a stretch of road with infinite length having 2 lanes for $x < 0$ and 1 lane for $x > 0$ as depicted in Figure 2.7a. For both stretches triangular fundamental diagram are considered as depicted in Figure 2.7b denoted as $Q_u(k)$ and $Q_d(k)$ for the upstream and



(a) Schematic of the case with infinite length freeway with a lane drop from 2 to 1 lane at $x = 0$.



(b) Triangular fundamental for upstream and downstream areas.

Figure 2.7: Schematic of example scenario and fundamental diagram.

downstream sections respectively. The initial density is the following:

$$\begin{aligned} k(x,0) &= k_1 \quad x \leq 0 \\ k(x,0) &= k_c \quad x > 0 \end{aligned} \tag{2.8}$$

At any point, x, except $x = 0$, the flow induced by the density is the same at x^+ and x^-, and the solution for this case is trivial: $k(x) = k_1 = k(0)$. We can verify this from the flow balance. This solution leads to flow $Q(k)$ at all points an therefore $\frac{\partial}{\partial x}q$ is zero and therefore the density does not change in time. At $x = 0$, however, the road characteristics and the initial conditions are different. Also, observe that $Q_u(k_1) = q_u > C$ which means the upstream flow is higher than the capacity of downstream stretch. It is a reasonable assumption that the flow at the boundary will be the maximum possible which in that case would be $q(0,t) = q_d = C$. This flow corresponds to $k(x) = k_c$ in the downstream section which is exactly equal to the initial condition. At $x = 0$ the flow just upstream is q_u and downstream is C and by mass conservation (Eq. (2.5)) the density must increase as $\frac{\partial}{\partial x}q(x,t)_{x=0} < 0$. In the upstream stretch, the density higher than k_1 in which $Q^-(k) = C$ is k_2. Therefore, there will be a small portion of the upstream section experiencing $k(x) = k_2$. As time goes by, there will be more vehicles coming at rate q_1, but the flow at $x = 0$ is

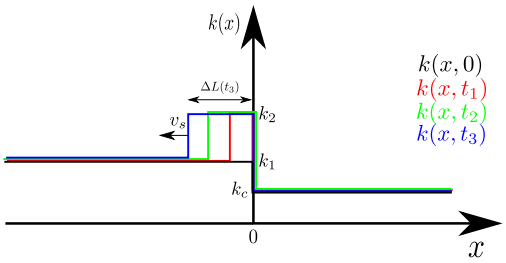


Figure 2.8: Density solution at different times.

constrained by the downstream capacity, C. Therefore, the length of this congested portion should increase over time. In Figure 2.8 the shapes of density at times $t = 0 < t_1 < t_2 < t_3$.

There is a discontinuity (shock) in the density profile in which the density shifts from k_1 to k_2. As one can see in the density profile in Figure 2.8, the discontinuity is moving backwards. We can determine the speed in which this shock travels based on the mass conservation principle. Considering a section $\Delta L(t)$ which starts at $x = -\Delta L$, for $t = t_3$ and ends at $x = 0$, as depicted in the Figure 2.8, where the interface between the discontinuities lies on this section. Let's consider the portion between $x = x_1 < \Delta L(t) \leq 0$. With that assumptions, we know that at $x = 0$ the flow is $q(0,t) = q_d = C$ and at $q(x_1,t) = q_u$. The position of the shock is denoted as $-x_1$. The number of vehicles on that area, $N(t)$, which by integrating densities:

$$N(t) = k_1(\Delta L(t) - x_1) + k_2 \Delta L(t)), \tag{2.9}$$

and as we know the boundary flow, $N(t)$ evolve as:

$$N(t) = N(0) + (q_u - q_d)t, \tag{2.10}$$

as the boundary of shock-wave remains the same as the shock-wave propagate backwards

(i.e., k_1 for $x \leq \Delta L(t)$ and k_2 for $x > x_1(t)$), the speed of the shock-wave will be constant. Denoting v_s the speed of the shock-wave we can write $\Delta L(t)$ as:

$$\Delta L(t) = -v_s t, \tag{2.11}$$

Combining (2.9),(2.10) and (2.11), the only possible value for v_s is:

$$v_s = \frac{q_d - q_u}{k_d - k_u} = \frac{\Delta q}{\Delta k}, \tag{2.12}$$

where $\Delta q = q_d - q_u$ and $\Delta k = k_d - k_u$. Therefore in this example, the congested region will grow with speed v_s. This expression for the shock-wave speed is referred to as Rankine-Hugoniot condition for conservation of mass.

In LWR model "queues" are modeled in space and time as one can observe in the density solution (Figure 2.8). That is, the congested area grows as long as the incoming flow exceeds the downstream capacity. Queue here is a loose concept as the vehicles actually never stops; rather, they travel through a portion with significantly lower speeds.

The dissipation of congestion work in a similar manner and the equation (2.12) still holds. If the upstream flow and density decreases below the downstream value, the speed, v_s, will be greater than zero and the length of the congested area decreases.

Along this example, I mentioned that it is reasonable that the flow at $x = 0$ should reduce to downstream capacity (1 lane) as the incoming flow is higher than that value. This was suggested by Lighthill and Witham [86] in their original paper. However, for different values of flow, say $q(0, t) = \frac{1}{2}C$, we could construct a density solution in which the mass conservation

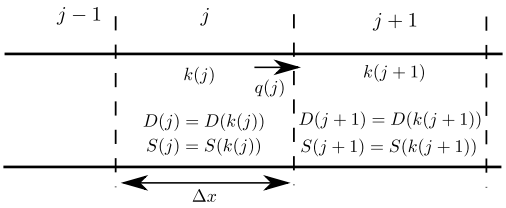


Figure 2.9: Cell transmission model schematic at a particular time step

equation and the fundamental diagram holds at all points, except at the discontinuities. Therefore, the solution we obtained is a weak solution of the partial differential equation.

As it is a physical system, we expect that we can find an unique solution that represents the dynamic of the system. For particular case, entropy conditions should be prescribed in order to obtain an unique and physical solution of the problem [5]. A common assumption is to choose the maximum flow possible that does not violate any constraint of the model. This is the implicit assumption in the previous example. The demand and supply concept proposed in [27] (as sending and receiving flow) and [78] provides a method to determine the boundary flow based on this assumption in which is briefly presented next.

The concept of demand and supply [78, 28] was first proposed in the cell transmission model [27] which is a discretization of the LWR model in both space and time. Nevertheless, the concept was extended to only-time discretization [139] and to continuous time formulations [61, 48]. Particularly here we are interested on the flow computation at the boundaries rather than the particular discretization scheme.

Let's assume the space is discretized into cell of equal length, Δx, indexed as $j = 1, ..., n_j$ where $x = j\Delta x$, where Δx. Similarly, the time is discretized into steps $i = 1, ..., n_i$ where $t = i\Delta t$. The density at cell j at discrete step i is denoted by $k(j, i)$. This scenario in a given time is depicted in the Figure 2.9. The goal is to compute the flow from the upstream to the downstream cell.

Demand is the maximum flow the upstream cell can send to the downstream cell regardless of

the state of the downstream cell. There are two physical constraints for that: (i) the number of vehicles cannot exceed the number of vehicles inside the cell, otherwise a negative density would be observed; and (ii) the flow cannot exceed capacity. For supply the principle is similar in which the constraints are: (i) adding the received vehicles and the vehicles already in the cell cannot be higher than the jam density; and (ii) the flow cannot exceed the capacity of the downstream cell. Demand, supplies and boundary flow are computed as:

$$\begin{aligned} D(j,i) &= Q(\min\{k(j,i), k_c\}) \\ S(j,i) &= Q(\max\{k(j,i), k_c\}) \\ q(j,i) &= \min\{D(j,i), S(j+1,i)\} \end{aligned} \tag{2.13}$$

With the same principle, the flow at boundary at any boundary. For example, the flow at $x + \Delta x$ would be computed and the density at downstream cell would be updated as:

$$k(j, i+1) = k(j,i) + \frac{\Delta t}{\Delta x}(q(j-1,i) - q(j,i)) \tag{2.14}$$

where Δx and Δt respect the Courant-Friedrich-Lewy condition [26, 25]. A graphical representation of demand and supply can be noticed from the fundamental diagram as in Figure 2.10 Demand is the increasing section of the fundamental diagram; analogously, supply is the decreasing section of the fundamental diagram.

The computation through demand and supply was presented here as a time-space discretization of two homogeneous cells. Nonetheless, the concept has been extended to networks as in [28], to inhomogeneous roads and in continuous time shown to serve as entropy conditions to obtain unique solutions. It is not detailed here, nonetheless the intuition to continuous

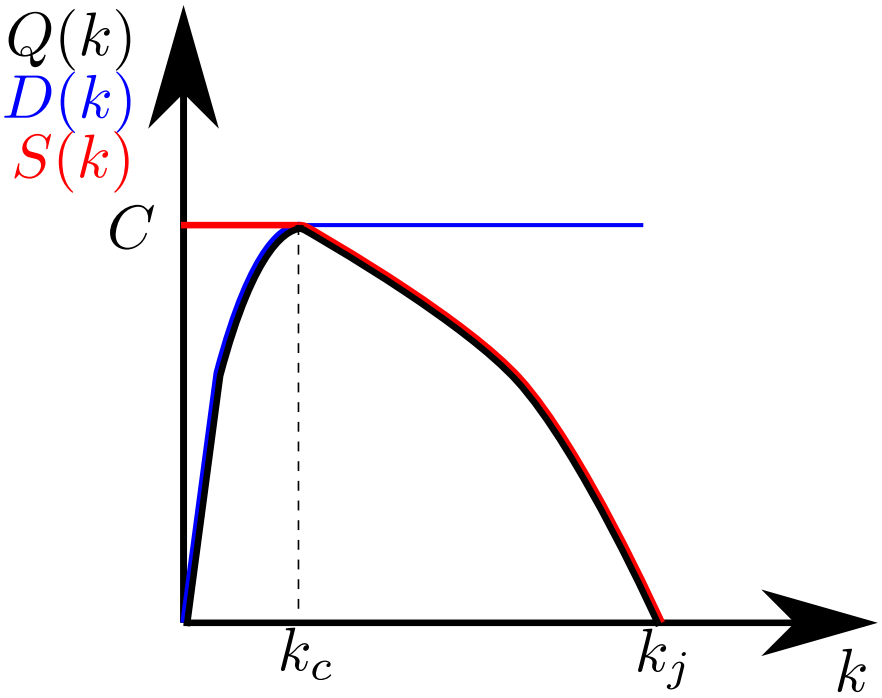


Figure 2.10: A fundamental diagram and its associated demand, $D(k)$ and supply, $S(k)$.

can be captured as the above equations with $\Delta x \to 0$ and $\Delta t \to 0$. Along the same line, henceforth when I refer to the flow at x, x^+ and x^- refers to point at limit approaching x from downstream ($x^+ > x$) and upstream of x ($x^- < x$).

Going back to the example of Figure 2.7a, we can apply the demand and supply at $x = 0$. The downstream area ($x > 0$) is always uncongested and therefore the supply is always the downstream capacity, C. The upstream area $t = 0$ would lead to demand $q_u > C$ at $t = 0$ and to the maximum demand (upstream capacity, approximately $2C$). At any case, applying the min operator to compute the boundary flow, would find $q(0, t = 0) = C$. That is exactly the same result we have found in the previous sub-section.

The merge bottleneck example shows one of common weakness of the LWR model [107]. In a situation like the previous example, the flow at $x = 0$ reduces to the downstream capacity. However, by the empirical evidence of the capacity drop, we know the outflow is smaller than the downstream capacity when its upstream section is congested. Second, several extensions

of the LWR model aiming to capture the capacity drop phenomenon is based on demand and supply concept. I present some of the LWR model extensions of capacity drop in the next section.

2.3 Capacity Drop Models

Since the capacity drop phenomenon is a fact, it is imperative that traffic flow models replicate such phenomenon when modeling a similar situation. Second order models (e.g., [95, 8]) arguably can replicate the effects of capacity drop in a lane drop scenario such as the one depicted in Figure 2.7a as pointed out in [107], which also points out this specific case as a deficiency of first-order (LWR) model. However, several extensions of the LWR model have been proposed to circumvent this deficiency and I introduce some of them here.

Such extensions, in general, model the merging segment (or just a lane drop) as a specific link with slightly changed dynamics in order to capture the outcomes of the capacity drop. Some of the models captures the capacity drop phenomenon as an outcome of the combination of the basic physics governing the system associated to driving behaviors including lane-changing and bounded acceleration. Others models are conceptual and model the outcomes based on the relevant state variables based on reasonable assumptions. [1]

Throughout this section, the models are being used to compute flow at point x where the lane drop is located. In addition, C is the downstream capacity.

[1] The distinction into conceptual "empirical" and physical model is not common in traffic. This distinction is based on hydrologic models to compute flow at rivers as in [99] which is very similar to traffic flow.

Phenomenological Capacity Drop Model

Proposed in [65], it is a conceptual model in which the capacity drop is replicated based on a simple modification in the conditions in which capacity drop is triggered rather than modeling the phenomenon itself based on the macroscopic variables. It is based on demand-supply concepts, but difference lies on the computation of the boundary flows:

$$q(x,t) = \begin{cases} D(x^-,t), & D(x^-,t) \leq S(x^+,t) \\ \min\{S(x^+,t), C^-\}, & D(x^-,t) > S(x^+,t), \end{cases} \tag{2.15}$$

where C^- is the flow observed when the capacity drop is present in which is referred to here as congested capacity. Often the value the congested capacity is obtained from the capacity-drop ratio, defined as:

$$\Delta = 1 - \frac{C^-}{C}. \tag{2.16}$$

The congested capacity is exogenous to the model and should be calibrated for each case. Note that it does not change the fundamental diagram at any point between $x = 0$ and $x = L$; it actually changes the boundary conditions. Figure 2.11 depicts these two capacities and specific densities that can help to better understand what is the impact of such phenomenon on the LWR model.

Assuming the downstream area is uncongested, the downstream supply is the downstream capacity (i.e., $S(x^-,t) = C$). If the upstream demand exceeds the downstream supply there will be unserved vehicles and the merging segment become congested. In our previous example applying the LWR model, the outflow would be capacity. However, in this case computing flows through (2.15) leads to $q(x,t) = C^- < C$. The downstream section remains

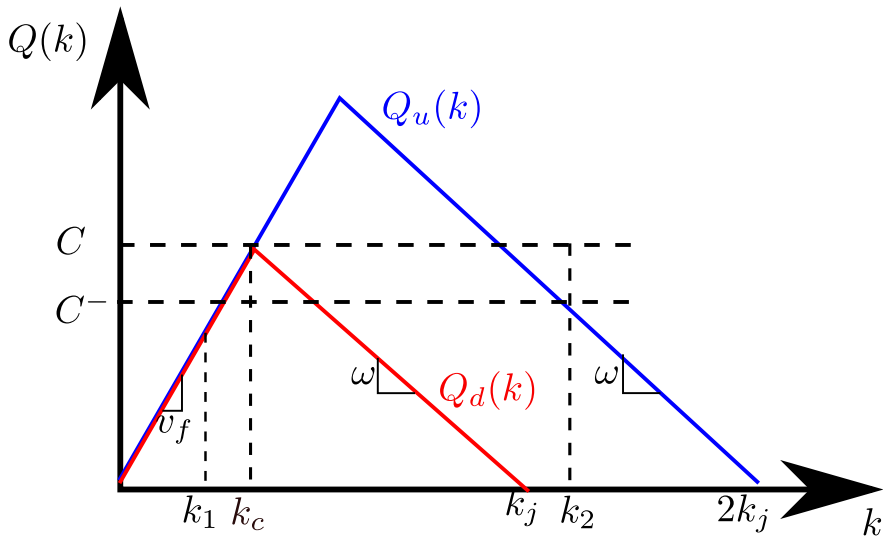


Figure 2.11: The fundamental diagram with projected densities with capacity drop.

uncongested with the density, k_1 associated to flow C^- on the uncongested branch of the fundamental diagram, as depicted in Figure 2.11. Upstream become congested and therefore the density that emerges on that area is the density that yields C^- in the congested branch of the fundamental diagram. This density value is depicted as k_2 in Figure 2.11. Compared to the LWR model, the outflow is smaller and the congestion is more severe as measured by a higher density upstream to the bottleneck.

Dissipating the congestion also becomes more difficult. To decrease the number of vehicles inside the merge is zone is necessary an upstream arrival $d(x^-, t) < C^-$. It shows the baffling effect of the capacity drop: it requires arrivals higher than downstream capacity to trigger the capacity drop; however, it requires a sharp reduction on the arrival rate to clear the congestion.

Modified Demand Models

A number of models attempt to capture the effects of capacity drop by changing the demand function instead of changing the boundary flows. Though a specific model does not introduce

a driving behavior to justify the modified demand function, others derived derive the demand based on bounded acceleration and lane changing behaivor.

In [98] they propose a discontinuous demand function that resembles Equation 2.15 of the phenomenological capacity drop model. The difference on being in the demand function as oppose to a change in the boundary flow change the behavior in few cases. Specifically, it changes when the capacity drop is triggered when the supply is between capacity and congested capacity.

In [115] they propose a linear decreasing demand function when the upstream density exceeds the critical density in order to capture the outcomes of the capacity drop. Despite they mention the existence of bounded acceleration, they do not derive the model based on a specific driving behavior for that study.

In [79] a "two-phase" model was proposed to introduce bounded acceleration into LWR model. That is in fact a limitation of the LWR model [81] as in a situation like the previous examples, when a vehicle crosses a "shock" (discontinuity) from a congested to uncongested area (i.e., larger density to low density) it is implicit assumed that vehicles can accordingly change speeds instantaneously. The two-phase model addresses this aspect by explicitly modeling the bounded acceleration. In the two-phase model, drivers cannot exceed a maximum acceleration, denoted by A, which is in the order of 2m/s^2. In a recent study [71] they integrated the constant bounded acceleration into demand-supply framework. The Figure 2.12 was obtained from that study.

Though demand decreases as larger is the upstream density, in the onset of congestion (i.e., transition from uncongested to congested) the upstream density will converge to an equilibrium value and the congested capacity is the flow yielded by the equilibrium density. Decreasing the maximum acceleration rate leads to a decrease of the congested capacity [122].

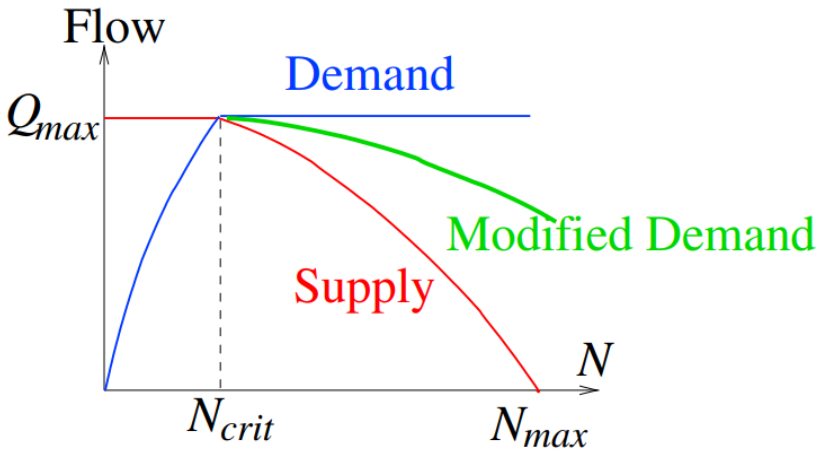


Figure 2.12: Modified demand function for bounded acceleration. Obtained from [71]

.

In [122] extends the demand-supply framework for bounded acceleration to different microscopic acceleration models. It is proposed a framework in which microscopic acceleration models are transformed into macroscopic demand functions based on two different mechanisms, instantaneous (within cell) reaction time and (ii) based on the assumption the information to accelerate travels at finite speed.

The use of different demand and supply function does not limit to bounded acceleration. In [121] the authors combined both lane changing effects and bounded acceleration to capture the capacity drop phenomenon. An import aspect of this model is also modeling a decrease in the outflow due to higher flows from on-ramps which induces more lane changes based on the model proposed in [57].

Chapter 3

Review of Freeway Control Methods

Let it out, take control
Find your way towards the place you belong

Angra (Storm of Emotions)

In the Chapter 2 details of the traffic flow dynamics were reviewed with special attention to the capacity drop phenomenon at merge bottlenecks. Nevertheless it was not mentioned specifically how we can mitigate the effects of capacity drop and how to reduce delays in the mentioned examples. There are ways in which management systems can interfere in the traffic dynamics and thereby improving performance.

The most traditional way to interfere in the traffic flow at freeways perhaps is ramp metering which consists in limiting the on-ramp flow by basically forcing vehicles to experience queues on on-ramps instead of in the mainline freeway. Nonetheless ramp metering is not the only one, other strategies include variable speed limits [4] and cooperative merging [96].

In this chapter we review the management strategies in the literature. As several approaches

in the literature as well as this research are based on control theory methods, a brief review of control theory is first presented. Later, a review of relevant freeway control methods is presented.

3.1 Control Theory Concepts

The first definition related to control theory is related to system concept. A dynamic system has states and outputs that change over time [6] which may be impacted by an external system through the system inputs. The merge bottleneck of Figure 2.1 fits exactly in this definition. The output of this system can be the outflow, the state the density in the merging segment, and the input the metering rate on the on-ramp.

A control system is a set of three components that combined can properly achieve its goals: control logic, sensor, and actuator. The sensor is the element that provides a quantitative measurement of the state or the output of the system. In the case of the merge bottleneck, it can be a loop detector that measures occupancy and flows at a specific location. The actuator is the element that affects the system. In the case of ramp metering, the actuator is the signal that informs when the next vehicle can merge in the freeway. Varying the gap between vehicles we can allow more or less vehicles into the merging segment therefore changing the flow. The control logic is the strategy that defines how the control system impacts the system - through its actuators - based on the current system state - provided by its sensors.

Formally, the information provided by the sensor we refer as the vector $\mathbf{y}(t)$ which may contain one or more outputs (measurements) of the system. The information provided to the actuator, the control action, is denoted as $\mathbf{u}(t)$ which again can be one or several. The control action is closely related to the system inputs. The control logic is denoted

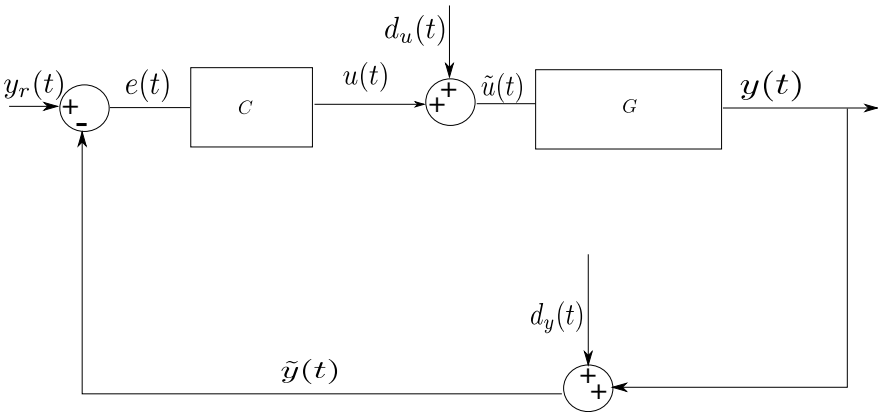


Figure 3.1: A schematic representation of a control loop with a process, represented by G, and a controller, represented by C, a reference signal, and input and output disturbances.

by $\mathbf{u}(t) = f(\mathbf{y}(t), t)$ which maps how the measurements are translated to control actions. Finally, the system dynamics is denoted as $\mathbf{y}(t) = \mathbf{g}(\mathbf{u}(t), t)$.

A first distinction in control is whether a system is under open- or closed-loop control. Open-loop control refers to a controller in which its logic does not depend on the current measurements of the system, that is, $\mathbf{u} = f(t)$. Conversely, a closed loop control refers to systems in which the control logic does consider the current measurements. The term feedback refers to using information of the system (i.e., the measurements) to change its input. It is called closed-loop due to dynamic dependency that arises when the controller uses information of the system output. Consider a schematic in Figure 3.1. The control signal $\mathbf{u}$ depends in the output $\mathbf{y}$. The output itself depends on the control input through the system dynamic creating this closed-loop.

The figure depicts 3.1 a basic block diagram of a control system. The system is represented by the block G, taking as input $\tilde{u}(t)$ and output $y(t)$; the controller takes as input $e(t)$ and computes output $u(t)$. There could be input disturbance, that changes signal $u(t)$ to $\tilde{u}(t)$ and output disturbance that transforms $y(t)$ to $\tilde{y}(t)$. The signal y_r is the set-point, that is, the desired value for $y(t)$. The output $y(t)$ is also referred to as manipulated variable. The signal $u(t)$ is called the controlled variable.

Disturbances, $d_u(t)$, are other inputs that might interfere in the system. It could be a variable that was not explicitly considered in the model, but may affect the system. In the case of a merge bottleneck, it is known that weather conditions [47] changes the system dynamics, but due to various reasons the models does not take into account that fact. The weather influence in this case can be treated as a disturbance. The output disturbance is often some uncertainty or errors on the measurements. For example, in the case of a traditional ramp metering system, the loop detector provides occupancies and flows which may contain errors due to discretization or miscounts. The output disturbance is denoted as $d_y(t)$.

The system or plant, G, can be described in different forms, usually as a differential equation. For most of the applications, the controller is a linear differential equation in the form:

$$\sum_{0}^{N_c} c_n u^{(n_c)} = \sum_{0}^{M_c} d_m e^{(m_c)} + C \quad n_c = 0, 1, ..., N_c, m = 0, 1..., M_c, \tag{3.1}$$

where a_n and b_m are real numbers, and $x^{(i)}$ denotes the i-th derivative of variable x. The advantage to model the system as Equation (3.1) is that it is an ordinary differential equation (ODE) which solution method as well as closed solutions are known.

If the system cannot be represented as a linear differential equation, the dynamic can be represented as some function g such as $y(t) = g(y(t), u(t))$ where y and u in the arguments refer to the function y and u, not only its value at time t. These systems are harder to deal with as there is no general solution to the system response. A basic procedure is to find the equilibrium points of the system, that is, the output y that the system reaches when a constant input u is applied. We can linearize the system dynamics around an equilibrium

point (x_0, y_0) as $g(.) = y_0 + \sum_0^n \frac{\partial g}{y}^{(i)} \delta y^{(i)} + \sum_0^n \frac{\partial g}{u^{(m)}} \delta u^{(m)}(t)$.

$$\sum_0^N a_n \delta y^{(n)}(t) = \sum_0^M b_m \delta u^{(m)}(t) \quad n_c = 0, 1, ..., N, m = 0, 1..., M, \tag{3.2}$$

where coefficients $a_n = \frac{\partial g}{\partial y^{(n)}}$ and $b_m = \frac{\partial g}{\partial u^{(m)}}$. Therefore, only is taken into account a shift from the equilibrium point and $u(t) = u_0 + \delta u(t)$ and $y(t) = y_0 + \delta y(t)$. From now on, is is considered the controller and the system is either linear or linearized around an equilibrium point.

A linear system also has an associated transfer function in the frequency domain. A system with a single input and output can be described as $\frac{y}{u} = \frac{n(s)}{d(s)}$ where $n(s)$ and $d(s)$ are polynomials.

The transfer function on frequency domain is a continuous representation that depends on the current value of input and outputs and their derivatives. There is also the discrete representation when control actions and measurements are updated at times kT_s, where T_s is the sample time and k an integer and the dynamic equation becomes is $y(k) = f(y(k-1), ..., y(k-n), u(k), ..., u(k-m+1))$. All the analysis and properties derived for continuous time, especially regarding linear systems, have its counterpart for discrete-time.

For the system described in Equation 3.2 it becomes:

$$\frac{y}{u} = \frac{n_g(s)}{d_g(s)} = \frac{b^M s^M + b^{M-1} s^{M-1} ... + b^0}{a^N s^N + a^{N-1} s^{N-1} ... + a^0} \tag{3.3}$$

The controller is $C(s) = \frac{n_c(s)}{d_c(s)}$ and the system is $G(s) = \frac{n_g(s)}{d_g(s)}$. Two relationships can be

derived from the loop depicted in Figure 3.1. The reference to output response:

$$\frac{y(s)}{y_r(s)} = \frac{n_c(s)n_g(s)}{n_c(s)n_g(s) + d_c(s)d_g(s)}, \tag{3.4}$$

and the disturbance to output:

$$\frac{y(s)}{d_u(s)} = \frac{n_g(s)}{n_c(s)n_g(s) + d_c(s)d_g(s)} \tag{3.5}$$

Note that the denominator is the same for both cases and the set of roots, λ_i, determine the nature of the response. A root λ_i will have an associated term $\alpha_i e^{\lambda_i t}$ [1] in the response.

The basic goals of a controller is to achieve a desired dynamic (reference to output and disturbance to output) and robustness to model uncertainties [6]. It can be further divided in the following list which appears in control system textbooks as [6, 104], but sometimes with different terminology:

1. stability: the system is stable if considering two close initial conditions, Y_a and Y_b. For any $b_1 > 0$ there is a $b_2 > 0$ such that: $|Y_a - Y_b| < b_1 \implies |Y(t, Y_a) - Y(t, Y_b)| < b_2$ for all $t > 0$ [6]. It means that if the initial conditions are close, the system response will follow similar trajectories. For linear system it reduces to have the real part of the roots, λi, lower than zero. Real part negative ensures that the associated exponential term goes to zero as $t \to \infty$.

2. tracking performance: refers to how close the system output remains to a desired output or reference output $y_r(t)$. Often the signal $e(t) = y_r(t) - y(t)$ is called tracking

[1] It changes for repeated roots, but it is not detailed this case here.

error [6]. It is referred in this document as steady-state tracking or only tracking when $e = 0$ as $t \to \infty$.

3. robustness to model uncertainty: refers to the capability of the system to keep reasonable performance when the dynamic of the system is not exactly the same as modeled. Lets assume the controller was designed based on a nominal model dynamics, $G(s)$. However, in practice the dynamic is slightly different, denoted as $\tilde{G}(s)$. The controller is robust if it ensures stability and steady-state tracking even when the system dynamics, $\tilde{G}(s)$ is different from the nominal dynamic, $G(s)$

4. disturbance rejection: related to whether the controller ensures tracking performance in presence of disturbances, that is, when $d_u(t) \neq 0$ is not zero.

Stability is necessary for achieving all others. The other goals might not be achieved for a controller given the system dynamic and they are somehow conflicting. A good tracking performance might be achieved at expense of lower robustness to model uncertainty, for example.

The performance of a ramp metering algorithm in terms of traffic flow variables, for example minimizing delay or maximizing throughput will depend on these items. The performance in terms of throughput and delay is related to tracking performance. As there are inherent stochastic components on the model. For example, the relationship between flow and density is assumed to be fixed in traffic, but actually is scattered which leads to the necessity of robustness. The upstream demand, which is uncontrolled, can change over time and it is desirable to ensure performance even when it changes. Therefore it is worth investigating these properties as all of them have impact in the overall performance.

Modeling the controller and the system as transfer functions are useful for systems with a single input and output. The system can have multiple inputs and outputs. For example,

one case of multiple inputs and outputs is the control of multiple on-ramps based on the information of several detectors.

For this case, the system has a state-vector, $\mathbf{x}$ containing n states. The input also becomes a vector, $\mathbf{u}$ with m components. The measurements are represented in the vector $\mathbf{y}$ that has p elements. The representation is the following [6]:

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= A\mathbf{x} + B\mathbf{u} \\ \mathbf{y} &= C\mathbf{x} + D\mathbf{u}, \end{aligned} \tag{3.6}$$

while in discrete time:

$$\begin{aligned} \mathbf{x}(k+1) &= A\mathbf{x}(k) + B\mathbf{u}(k) \\ \mathbf{y}(k) &= C\mathbf{x}(k) + D\mathbf{u}(k) \end{aligned} \tag{3.7}$$

In state-space representation it is possible deal with multiple variables at same time. While for single input and output system the analysis is done based on the root of the closed loop polynomial, in state-space the analysis is similar but looking to the eigenvalues. A common scheme in continuous time is to have $C = I$, and set $\mathbf{u} = -K\mathbf{x}$ the dynamic reduces to $\frac{d\mathbf{x}}{dt} = (A - BK)\mathbf{x}$ and performance and stability are analyzed based on the eigenvalues of matrix $A - BK$.

All concepts are the same or extended to a general case in the state-space representation. However, there are two properties that can be established in this representation that are trivial for SISO systems: reachability/controllability and observability.

Reachability and controllability are often interchangeable, it is followed here the definition in [6]. Reachability is whether an arbitrary state $\mathbf{x_1}$ can be be achieved through an arbitrary $\mathbf{u}(t)$, $0 \leq t \leq \tau$, with a given initial condition $\mathbf{x_0}$; whereas controllability it is related to reach the origin, $\mathbf{x_1} = 0$ from $\mathbf{x_0}$. Both are equivalent for linear and unconstrained systems.

Let $W_r = \begin{bmatrix} B & AB & ... & A^{n-1}B \end{bmatrix}$, it is possible to reach any point in the state space with the matrix W_r has n linear independent columns. As it can reach any point in the state-space (reachability), it can reach the origin (controllability).

Note that reachability is whether is possible to achieve a state, but it is not guaranteed that it will remain there. Those are the equilibrium points that is defined as [6]:

$$Eq = \{\mathbf{x_e} : A\mathbf{x_e} + B\mathbf{u_e} = 0\}. \tag{3.8}$$

That is, the set of points, $\mathbf{x_e}$, that can be reached with $d\mathbf{x}/dt = 0$.

Another property is observability which is related to reconstruct $\mathbf{x}$ based on measurements $\mathbf{y}$ and the control inputs $\mathbf{u}$. Sometimes it is not possible to measure all states, but based on the system dynamic it might be possible to estimate the ones that are not measured. For linear systems, testing observability is similar to reachability. Let $W_o = \begin{bmatrix} C & CA & ... & CA^{n-1} \end{bmatrix}^T$, it is observable when the matrix W_o has n linearly independent rows.

PID control is the most used feedback control in engineering system [6]. The PID controller computes the control signal u based on the error signal $e(t) = y_r(t) - y(t)$. The control signal

u has three different terms, proportional to current, integral and derivative of $e(t)$:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{de(t)}{dt}, \tag{3.9}$$

where parameter K_p is the proportional gain, K_i the integral gain and K_d derivative gain. The set-point y_r should also be defined, but it can change at any moment whereas, traditionally, the gains are fixed or rarely change.

If for a given K_p, K_i, and K_d the system in closed loop is stable, all the properties are achieved with the PID controller. Tracking is achieved as it yields a constant $u(t)$ when $e(t) = 0$ and therefore $y(t) = y_r(t)$. With the same reasoning, it can be shown that any constant disturbance $d_u(t) = d_u$ will be rejected and eventually $e(t) = 0$. Also, as it does not depend explicitly on the model, it is also robust as long as it is stable, at least for linear or linearized systems.

The facts that it is simple and yet able to guarantee basic performance requirements are some the reasons to be widely applied and studied. The control problem becomes choosing K_p, K_i and K_d that ensures, at least, stability. Usually there is a region in which it is stable, choosing the exact values then depend on the requirements of the specific application. Large gains in general lead to faster response, but usually it is less robust to model uncertainty.

Although it has several advantages, there are also disadvantages that can prevent PID controllers to be used in specific applications:

1. it is not straightforward to guarantee stability for non-linear systems;
2. oftentimes stability is ensured, but at expense of a slow response to disturbances;
3. slow responses in systems with dead-time [103];

4. often there might exist operational constraints such as maximum and minimum values on u, but operations constraints are not explicitly considered in PID controllers.

When one or some of these disadvantages are crucial, other techniques might be considered to improve performance such as adaptive control, SWARM and model predictive control. The first two are similar. Adaptive control [7] tries to obtain approximate the a time-dependent and local dynamic based on the past inputs and outputs of the system. As the parameters are obtained, K_p, K_i and K_d are changed according to the current dynamic. It leads to good performance when the system in fact are time-varying or the dynamic is non-linear, but a time-varying linear dynamic is a good approximation of the system. SWARM [106] has a similar scheme where the current dynamic is approximated based on updated measurements.

Model Predictive control (MPC) is a class of technique in which an explicit model of the system is used to obtain the control actions [14] unlike the PID controller. The rationale of the system is to use the knowledge of the system dynamics to obtain control actions that optimizes the system performance. One common example of model predictive control is the rolling-horizon approach which has been applied into some freeway control methods which will be presented in the next section.

3.2 Review of Ramp Metering Algorithms

Several ramp metering algorithms have been proposed using different control techniques. Here some relevant algorithms is presented divided in local and coordinated control. The latter refers to control strategies intended to control a single bottleneck while the latter refers to strategies intended to multiple bottlenecks.

3.2.1 Local Control Algorithms

For the two control strategies presented here it is considered a merge bottleneck as Figure 2.1 where the local controller is setting the metering rate $r(t)$ defined as the number of vehicles allowed into the freeway per unit time.

The Demand-Capacity Algorithm

The demand-capacity was a quite-popular strategy in United States [109], though there is no recent literature regarding this strategy. The reason to be presented is that it can provide some intuition concerning the aforementioned aspects of control systems such as model uncertainty and controller robustness.

The goal of the Demand-Capacity algorithm is to balance upstream demand with downstream capacity through the following control law:

$$r(t) = \begin{cases} C - d_u(t) & \text{if } k(0,t) \leq k_c \\ r_{min} & \text{if } k(0,t) > k_c, \end{cases} \tag{3.10}$$

where C refers to downstream capacity, k_c the downstream critical density and r_{min} is the minimum metering rate, the most restrictive metering rate, which is applied whenever the demand exceeds capacity. The measurement at $x = 0$ is provided by a loop detector at that location. The rationale of the strategy is straightforward: drive the system to discharge at capacity when uncongested; when congested, limit as much as possible vehicles to enter the freeway to turn the freeway uncongested again.

We can qualitatively analyze the behavior of a local ramp-metering system when controlled

by the demand-capacity algorithm assuming LWR model dynamics combined with the phenomenological capacity drop model [65]. In the first analysis, we also assume that we have a perfectly calibrated fundamental diagram and therefore we know the true value of C and k_c. With all these assumptions, the control law (3.10) ensures that the sum of the on-ramp and upstream demand never exceeds capacity. Therefore, the system remains uncongested as long as the initial condition was uncongested. The control law also ensures the system discharges at capacity or the total demand which is also desirable. Therefore, as long as an unexpected event happens to the system, the demand-capacity would perform reasonably well.

However, when the system experience some congestion the response may not be appropriate. The congestion starts at $x = L$ while the detector is located at $x = 0$. It means that the congestion is only sensed when the congestion reaches the upstream detector. Therefore, the system operates as if it was uncongested even though there is some congestion on the merging segment which is undesirable. Even if congestion reaches $x = 0$, at that moment the minimum metering rate would be applied which may mitigate the congestion momentarily. If through $r = r_{min}$ the congestion can be mitigated, it would return to the uncongested mode as soon as at $x = 0$ become uncongested while there might be congestion between $x = 0$ and $x = L$. Therefore, the control would return to the uncongested mode again, before clearing all the congestion in the merging segment.

One could argue that it is a minor drawback as it would never occur as the system is initially uncongested. The fact is that the system has an inherent stochasticity. It is unavoidable that at some point the system will discharge slightly less than capacity for few minutes which is enough time to build up a small congestion. With some congestion the outflow can further reduce due to capacity drop phenomenon. There is no mechanism on this algorithm to prevent this situation.

We can also analyze the system when there is over or underestimation of the downstream

capacity. Lets consider the overestimation case in which is assumed that the actual downstream capacity is C and the demand-capacity algorithm considers a capacity $\tilde{C}$ such as $\tilde{C} > C$. In that case the demand-capacity allows more vehicles than downstream capacity. This excess of vehicles, probably small, will start from the downstream boundary, at $x = L$ and will slowly propagate backwards until $x = 0$. At this moment the density will be higher than critical and r_{min} is applied. Then, the demand will be lower than downstream capacity and the shock wave propagates forward and the density at the upstream detector fall below the critical before the congestion has been dissipated. Again, as the capacity is underestimated, it allows more vehicle and the shock wave start to propagate back again and this cycle continues as long as the upstream and ramp demand is high enough. Therefore this strategy is not robust if downstream capacity is underestimated. Case it is overestimated, it might not use some available capacity which is also not desirable.

In summary there is no clear feedback mechanism in the Demand-Capacity algorithm; in [109] it is defined as a feed-forward, instead of feedback, algorithm. Feed-forward means that the controller compensates for a measurable disturbance of the system, in which in this case is variations in the upstream demand. The feed-forward may or may not be combined with a feedback control.

The relationship between this algorithm with this dissertation is the following. It is an important example of an algorithm that starts from a sound principle - avoid the inflow to exceed capacity in order to avoid congestion - and works properly in a nominal case, but it does not keep good performance when it does not operate in ideal conditions which in this case are: system initially congested and not robust to slight changes in the parameter. We could reach this conclusion by analyzing the system response considering the system dynamics and the control law which is one of the goals of this research.

ALINEA

ALINEA proposed in [91] is a method based on PID controllers from control theory, instead of a switched control-law as the demand-capacity algorithm. In that regard it inherits the advantages of the PID controllers. Another key difference with respect to the demand-capacity algorithm is the manipulated variable, $y(t)$. Instead of considering measurements of upstream flow, ALINEA use the information of the occupancy [2] just downstream to the bottleneck.

The original ALINEA is an I-controller. Here it is considered the PI-ALINEA extension[132], where the PI-controller, a special case of PID with $K_d = 0$. The error signal $e(t)$ is the difference between the current density $k(t)$ and the target density $k_o(t)$:

$$e(t) = k_o(t) - k(t). \tag{3.11}$$

In addition, the control signal $r(t)$ is bounded, leading to:

$$\begin{aligned} r(t) = K_p(k_o - k(t)) + \int_0^t (k_o - k(\tau))d\tau \\ r_{min} \leq r(t) \leq C_r, \end{aligned} \tag{3.12}$$

where C_r is the on-ramp capacity.

Thus to design a PI-ALINEA, we need to determine the following parameters: the coefficients K_p and K_i, the target density, $k_o(t)$, and the minimum metering rate r_{min}.

The target occupancy is $k_o = k_c$ or close to it so keeping the mainline close to capacity and still uncongested. The rationale behind the PI-Controller is that the integral terms ensures that $r(t)$ will be constant only if $e(t) = 0$ and therefore $k_o = k_c$. Lower values of $k(t)$ leads

[2] Occupancy is the percentage of time in which a traffic detector senses the presence of a vehicle in a period of time. It is closely related to density.

the system to increase the metering rate. Similarly, it decreases the metering rate if density is higher than critical.

For the theoretical point of view, ALINEA has one disadvantage. A sudden increase in the upstream demand will be detected only after reaching the downstream boundary which may cause a queue to form and thereby increasing the occupancy. Only at this point the controller will start to decrease $r(t)$. The response time depends on K_p and K_i. On the other hand, ALINEA uses only a single traffic model parameter: the critical density. It is able to accordingly clear a formed queue only by trying to push the system to the critical density.

Nonetheless, it is possible that the value of k_o is not exactly equal to the critical density. In the case of ALINEA if one overestimates k_c can lead to an erratic behavior. The reason is that a density slightly higher than k_c is not an equilibrium point assuming that LWR model on the lane drop scenario. As it becomes congested, $k = k_1$ and the PI-Controller will decrease the metering rate until it becomes uncongested again. Then, it will allow more vehicles again and the cycle will repeat over time. The difference to Demand-Capacity it responds as the congestion starts and completely clears the congestion while in the Demand-Capacity it clears the congestion on the upstream boundary, but increases again the metering rate with, possibly, a congestion inside the merge zone.

With the capacity drop phenomenon the impact of this switching behavior is greater. That is, probably, the reason a slightly under critical target occupancy yields better results in field deployments [108].

A study of ramp metering using the link queue model and considering ALINEA as the ramp metering algorithm is presented in Chapter 5.

3.2.2 Coordinated Control Algorithms

ZONE and Stratified Algorithm

The Minnesota Department of Transportation (Mn DOT) has been applying ramp metering since early 1970s [137]. The ZONE metering strategy had been in operation for years. After some public complaining, the Mn DOT conducted an 8-weeks shutdown in order to assess the effectiveness of the ramp metering strategy that had been in operation. The conclusions were that the ZONE strategy had been providing overall benefits, but it was true that delays on on-ramps were large. It lead to the development of a new algorithm that keeps main features of the ZONE, but is able to overcome some of the issues. Both are briefly presented here based on [137].

The ZONE algorithm divides the controlled section of the freeway into zones. It is defined as a region where upstream works in free-flow and a downstream bottleneck. The goal is to balance in- and out-fluxes in this zone and therefore keeping a constant density in those areas. The conservation equation is the following:

$$M + F + A + U = X + B + S, \tag{3.13}$$

where M is the total local-access ramp volume, F is total freeway-to-freeway volume to be controlled, A is the measured upstream mainline volume, U the total measured non metered entrance ramp volume, X total exit ramp volumes, B downstream bottleneck capacity, and S the spare capacity representing "space available" within the zone.

The goal is to regulate flows at all zones. The metering rate has impact in the zone and also in downstream zones. It is computed the on-ramp flux M in order to balance a flow within

the zone and another to match a system goal (some downstream zone). The most restrictive of both is applied. The flow M is then converted to the flow for each on-ramp inside the zone.

In order to avoid large delays on on-ramps, a different mechanism was introduce to avoid too restrictive metering rates for long periods.

The stratified algorithm follows the same principles. The main differences are that ramps can be inside different overlapping zones and still follows the concept that ramp flows should be split in the on-ramps inside a zone and the most restrictive is applied. Also, there is also the layer concept. A zone is grouped with more zones as the layer level increases and the level of coordination decreases as the layer increases. The coordination goal is to balance volume to downstream layers.

To avoid long queues at on-ramp, the minimum release rate was introduced. Queues at on-ramp are estimated based on detectors at on-ramp, one upstream and one downstream. The minimum release rate is proportional to the queue length therefore if queue grows the metering rates increase as well.

In both algorithms, from the control and model point of view it uses simple concepts. The method to compute metering rate is similar to a P-Controller, but it is non linear because it is computed a metering rate for an on-ramp several times and the minimum is applied. As the goal is to balance inflow and outflow, the consequence is a constant density and therefore it might not be able to clear an already formed congestion similarly to demand-capacity algorithm.

Also, all the relationships are based in steady state flows. On-ramps flows impact downstream zones, but at different times, but it is not taken into account. This effect is better modeled in model based approaches.

Model Predictive Control Algorithms

In the last decade, several ramp metering algorithms that can be classified as model predictive control. Most approaches addressing freeway control is based on macroscopic models similar to the presented in Section 2.2. For related problems, such as traffic signal control, both macroscopic and microscopic models are being used. Rhodes [97], for example, is predictive but would be classified as microscopic. Similar approaches are [13] and goes back to OPAC [35]. While macroscopic, the TUC-MPC [2], a similar in [88], and [77]. Common to all of them, the dynamic is based on aggregated flow in times in the order of one cycle.

In similar problems to ramp metering, such as variable speed limit, model predictive control is also been used. In [51] the METANET model is used; using the same model, [15] also address coordinated control through variable speed limit. There also work addressing both variable speed limit and ramp metering [90, 16, 50].

Some work addressing the ramp metering problem are presented together with two recent ones that address the same problem, but considers the capacity drop in the following subsections.

Advanced Motorway Optimal Control The Advanced Motorway Optimal Control first proposed in [75] which analyzed the open-loop optimal given an initial condition. More recently [110], the same model was integrated into ramp metering control as an upper layer and ALINEA was used for local control, an hierarchical control approach.

The prediction model is based on METANET tool [95] which is based on the Payne-Witham

model with few modifications, the segment i of link m is described as:

$$\begin{aligned}
\rho_{m,i}(k+1) &= \rho_{m,i} + \frac{T}{L_m \Lambda_m}[q_{m,i-1} - q_{m,i}(k)] \\
q_{m,i}(k) &= \rho_{m,i}(k) v_{m,i}(k) \Lambda_m \\
v_{m,i}(k+1) &= v_{m,i}(k) + \frac{T}{\tau}(V[\rho_{m,i}(k)] - v_{m,i}(k)) + \\
&\quad \frac{T}{L_m}[v_{m,i-1}(k) - v_{m,i}(k)] v_{m,i}(k) - \frac{\nu T}{\tau L_m} \frac{\rho_{m,i+1}(k) - \rho_{m,i}(k)}{\rho_{m,i}(k) + \kappa} \\
V[\rho_{m,i}(k)] &= v_{f,m} exp[-\frac{1}{\alpha_m}(\frac{\rho_{m,i}(k)}{\rho_{cr,m}})^{\alpha_m}]
\end{aligned} \tag{3.14}$$

where $\rho_{m,i}$ (in models was used k, here k is the time step index) is the density of the segment i of link m, $\rho_{cr,m}$ is the critical density,$v_{f,m}$ is the free-flow speed, α_m a parameter of the fundamental diagram assumed on link m, τ a time constant, ν an anticipation constant, $q_{m,i}$ the flow leaving segment i to segment $i+1$, $v_{m,i}$ is the space mean speed within the segment, and T the time-step. On the third equation, $v_{m,i}$ two terms were added in order to consider speed decrease on merges with lane-drop.

Input links, on-ramps, receives a demand d_o and forward into a specific segment of the freeway network. The outflow of an on-ramp is q_0 depends in the traffic condition of the segment (m, i) and the metering rate and the input link has a queue ω_0, following the following conservation equation:

$$\begin{aligned}
\omega_0(k+1) &= \omega_0(k) + T[d_0(k) - q_0(k)] \\
q_o(k) &= r_0(k) \hat{q}_0(k),
\end{aligned} \tag{3.15}$$

where $\hat{q_0}$ is the non-metered flow, determined as:

$$q_0\hat{(k)} = \min(d_0(k) + \omega_0(k)/T, Q_0 \min(1, \frac{\rho_{max} - \rho_{m,i}}{\rho_{max} - \rho_{cr,m}})), \tag{3.16}$$

where Q_o is the freeway capacity. The flow through on-ramps is the minimum between the demand (i.e., the queue added the arriving vehicles) and the traffic condition on the freeway segment. Also, note the metering rate, r_0, is the ratio between allowed flow and non metered flow, ranging from a pre-defined minimum, r_{min}, to 1.

A node model is also introduced at the junctions to determine each flow on such segments, but it is not detailed in this review.

Therefore, combining all the equation it is possible to compute the state variables $\rho_{m,i}(k+1)$, $v_{im}(k+1)$, $\omega_{oj}(k+1)$ from the values of those variables at time-step, demands, and metering rates at k and recursively to $k+2$, ..., $k+K_p$ where K_p is the horizon. The goal is to minimize total time spent on the freeway and at the on-ramps:

$$J = T \sum_{k=1}^{K_p-1} \sum_{\forall m} \sum_{\forall i} \rho_{m,i}(k) L_m \Lambda_m + T \sum_{k=1}^{K_p-1} \sum_{\forall o} w_o(k) + \sum_{k=2}^{K_p-1} j_1(r(k), r(k-1)) + \sum_{k=1}^{K_p-1} j_2(w(k)), \tag{3.17}$$

where the first and second terms are the total time spent on the freeway and on-ramps; j_1 and j_2 are penalty functions to avoid oscillations on the metering rate and queues longer than a pre-defined values on on-ramps respectively.

The output of the problem provides flows $q_{m,i}$ at each segment of the network. At the junctions, ALINEA is used to control the on-ramps, but with target density based on density

that yields the desired out-flow. It might occur on this approach, a target density higher than critical. If the predicted flow is lower than capacity, a small change on the metering rate assures that the metering rate will be higher than the on-ramp demand.

In this approach the optimization problem is solved less often than the local controllers sample time applying the ALINEA algorithm. This allows for larger computational time while still controlling the on-ramps. Hierarchical control is a common architecture in control systems [116].

Even though it was not compared with other coordinated algorithms, the results were better than applying only ALINEA on on-ramp without the upper layer. The same research group had previously tested coordinated ramp metering based on state-feedback, METALINE [108], addressing the same freeway, but there is no a comparison between both in the paper.

There are few points worthing mentioning from the control point of view. First, it uses a second-order model in which properties equilibrium points and their stability are not trivial. To guarantee stability in model predictive control, the basic three methods are based on either push the system into their equilibrium points through a equality, a terminal set constraint or a terminal cost. Clearly, none of them was adopted. Therefore, in this approach stability is not guaranteed and also not discussed. Second, even though it was recognized that demands and parameters are stochastic, a systematic analysis on the variation of demands and parameters were not conducted.

The lack of stability can be seen as a major drawback, but it is a common situation with non-linear systems, as it becomes harder to obtain proof of stability. As the local controllers are stable and the objective function has a penalty term to avoid oscillations on the metering rates, it probably leads to a stable behavior. On the other hand, a given sequence of control action can be optimal, but not stable. This is one of the first observations in the theory of optimal control: "optimality does not imply stability" [67], even though regarding a different

scheme.

Optimal Ramp Metering using Asymmetric Cell Transmission Model Proposed in [41] use as a model the asymmetric cell transmission model. An Optimization problem is solved given initial condition and demands traffic dynamic for a given period.

The problem of obtaining set of optimal metering rates based on known demand and freeway parameters is object of study since 1960's. In [134] a linear programming was proposed considering demands origin and destination (for each on-ramp it is known the ratio of vehicles that leave on all downstream off-ramps). Back then, it was not considered an initial state (density) on the freeway, but considered capacities. An advance on the same direction was given in [140] which a dynamic model was introduced and optimality conditions were provided for some specific cases, mainly constant demands (or "uniform congestion") and also went further by providing an optimal given a set of known demands at each on-ramp. In [41] the problem is conceptually the same: finding an optimal set of metering rates given known dynamic freeway model, CTM in this case and also has shown. As cell transmission model is non-linear due to the min operator, they also provided a simplification in which is possible to solve a convex optimization problem that yields close (but not equal) results.

The main difference between the asymmetric and classic cell transmission model is the treatment of merges that only has "asymmetric connections" where minor branches feeds a major branch. It is not clear from the manuscript, but cells may have different lengths. In this case Δt should be chosen accordingly based on the smallest cell length. Each cell has at most one on-ramp and off-ramp and if it has both, the on-ramp must be upstream of the off-ramp.

The cell transmission model is modeled as optimization problem. The focus in this summary is in what it differs or extends from what was already presented in Section 2.2. It is assumed split ratios, β_i in the interval $[0, 1]$ at the off-ramp of segment i that keeps at the freeway

to the segment $i+1$; the ratio $\bar{\beta}_i = 1 - \beta_i$ leaves through the off-ramp. The metering rate changes supply and demand scaled by a parameter γ. Both together lead to a slightly different supply and demand functions:

On-ramps are modeled as queues in which evolves based on the arrival demand on-ramps and the flow into the freeway. Constraints on metering rates (maximum and minimum) and queue (maximum) are also present.

Two terms are considered in the objective function, the Total Time Spent (TTS) and Total Distance Traveled (TTD):

$$\begin{aligned} TTT &= \sum_j \sum_i k_i(j) + \sum_j \sum_i w_i(j) \\ TTD &= \sum_j \sum_i q_i(j) + \sum_j \sum_i r_i(j), \\ J &= TTT - \mu TTD \end{aligned} \tag{3.18}$$

where $w_i(j)$ is the queue on-ramp i in time-step j. They show that TTD is a prescribed value under fixed metering rates and demands. The objective function favors larger travel distances with smaller travel times. They ensure in the optimization problem a "cool down" period appended to the end of the optimization time-window in which all demands are set to zero.

They solve three optimization problems, the full non-linear cell transmission model with the stated assumptions, one without queue length bounds and simplified on-ramp flows and a third with the queue bounds and simplified on-ramp flows. It is shown that solving the simplified problems it is possible to achieve similar performance.

Regarding the control, the main conclusion is that the maximum queue on-ramps imposes

an additional challenge and in that case it might not be possible to keep all the freeway segments uncongested.

It is provided proofs about the solution of non-linear problems and its relaxations opening possibilities to solve a linear model instead of a non-linear problem. As one of the challenges of Model Predictive Control is the computational time to obtain a solution, it is an important issue addressed. Simplifying (or adapting the model to fit into a specific structure) the optimization problem or stopping the algorithm before it reaches the optimal solution was shown to be a reasonable strategy and almost optimal in control systems[129]. Probably the same happens on freeway control given the uncertainties as demands, split ratios, measurements errors which solving very accurately a given optimization problem might not yield advantages comparing to an approximate solution.

Chapter 4

Impacts of capacity drop on equilibrium states of freeway corridors

The killing fields, the grinding wheels
Crushed by equilibrium

Iron Maiden (Out of the Silent Planet

4.1 Introduction

Congestion on freeways is a common experience on metropolitan areas around the world and leads to longer travel times, higher fuel consumption and air pollutant emissions. Ramp metering is one of the possible methods to mitigate congestion and its impacts as corroborated by both field deployments [108, 84] and simulation studies [55, 41, 110]. A well designed ramp metering control strategy can help to avoid two mechanisms that arise in congested traffic [109]:

(i) Queue spill back: this relates to the impact of congestion initiated on a given bottleneck

on vehicles exiting upstream of that bottleneck. Congestion on freeway often starts at a bottleneck, often a lane drop bottleneck after the merge with an on-ramp. If the total flow coming from the mainline freeway and on-ramp is higher than the bottleneck capacity, congestion will start at the lane drop bottleneck and propagate backwards. If the congestion reaches one or more upstream off-ramps, vehicles upstream to that off-ramps will have to slow down due to the congestion ahead. Thus the vehicles leaving at that off-ramp experience delay even though they will not travel through the bottleneck. Storing vehicles on on-ramps may prevent or delay congestion from reaching the upstream off-ramps and therefore help to reduce the total travel time.

(ii) Capacity drop phenomenon: this refers to a drop in the bottleneck flow rate below its capacity after the onset of congestion at that bottleneck given that the bottleneck flow is not being affected by the traffic conditions downstream to it [17]. In that case, keeping the bottleneck uncongested through ramp metering and other control methods avoids the drop in the flow rate and therefore helps to reduce the travel delay.

While the queue spill back mechanism is well understood, the same is not true about the capacity drop phenomenon [18]. The first studies that described the phenomenon with empirical data date from 1990's [11, 46, 17] and there is still not clear consensus of its exact mechanism. An evidence of the lack of consensus is the existence of capacity drop models based on different driving behavior such as bounded acceleration [71], drivers heterogeneity [22], and lane changing [76]. As consequence, it is still unclear the impact of the capacity drop phenomenon on the performance and design of freeway control strategies.

Nonetheless, there has been increasing interest on the impacts of capacity drop into freeway control. In [18] it was empirically shown that it is possible to recover the capacity outflow of a merge bottleneck through ramp metering. In other studies, stochastic models have been proposed to predict the probability of capacity drop (flow breakdown) being triggered based on current traffic conditions ([32, 74]) and these models have been considered in developing

control strategies [34]. Deterministic models have also been proposed [65, 82, 83] and recently some of these models were integrated into model based control strategies [98, 92].

In this study we attempt to examine the impacts of capacity drop on the performance of a freeway corridor in asymptotic equilibrium (stationary) states. In particular, we are interested in properties such as uniqueness, stability, and reachability of these states. These properties are important because they can answer some key questions related to freeway control. Uniqueness can answer whether there is only one equilibrium state that achieves the best performance. Stability can answer whether small variations on inputs and system dynamics will steer the system back to, as opposed to drifting away, from a desired equilibrium state. Reachability answer the question whether it is possible to steer the system to the desirable equilibrium state for given boundary conditions. Understanding the system behavior regarding such properties can help on the design of new control strategies as well as on the analysis of control strategies that have already been proposed in the literature.

The study object is depicted in Figure 4.1. The freeway contain M blocks of alternating merge and diverge segments, indexed as $1, 1', 2, 2', ..., I', I$ where i refers to the diverging segment of $i-th$ block; similarly, i' refers to the merging segment of the same block. Vehicles join the freeway through the on-ramps, assumed to be metered, and from section 0, which is not metered. Vehicles exit the freeway through the off-ramps and on the very last section, I'. This work is an extension of [42] in which they also studied the equilibrium properties in a freeway corridor assuming cell transmission model dynamics [28]. The fundamental difference of this study is considering the capacity drop phenomenon on the analysis. To that end, we adapted a capacity drop model for lane drop bottlenecks [65] to merge bottlenecks and integrated that on the link transmission model [139] to perform the analysis. The system is analyzed for two cases, without and with capacity drop, enabling us to identify the distinct features imposed by the capacity drop.

The rest of the paper is organized as follows. In Section 2 we introduce fundamental concepts

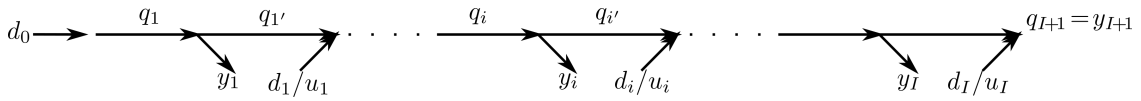


Figure 4.1: Schematic of freeway with alternating on- and off-ramps.

used in the following analysis showing how they impact the study object. In Section 3 we present the link transmission model [139] which is used for the analysis. In Section 4, the system properties are analyzed without considering the capacity drop phenomenon and in Section 5 a similar analysis is presented, but considering the capacity drop phenomenon. Finally, in Section 5 we state our conclusions and the implications for future work.

4.2 Equilibrium state definition and its Properties

While in the later analysis is based on the link transmission model, the concepts related to equilibrium states holds regardless of the model assumed with the same intuitive interpretation and practical relevance. The definitions are model-agnostic and therefore applicable also on different models. In this section we introduce the concept of equilibrium along with its properties on a freeway corridor.

The notation used throughout this document is summarized in Table 5.1 with the related description and unit. In addition to that, the same symbols with bold letters refers to a vector of the appropriate dimensions comprising all the individual elements.

The freeway corridor is split into I blocks containing a merge and a diverge segment in each block. Each segment has a set of state variables $\mathbf{x}_i$ and the vector $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_{1'}, \mathbf{x}_2, \mathbf{x}_{2'}, ..., \mathbf{x}_{I'}]$ contains all the state variables of the system. The demand is defined as the vector $\mathbf{d} = [d_0, d_1, d_2, ..., d_I]$ where d_0 is the upstream demand and not metered where all the following are the on-ramp demands assumed to be metered under constant metering rate $\mathbf{u} = [u_1, ..., u_I]$. Both, demands and metering rates, are assumed to be constant.

Symbol	Description	unit
$q_i(t)$, $gr_i(t)$	Flow from section i to $i+1$ at time t, flow on on-ramp i	veh/s
$f_i(t)$, $g_i(t)$	Inflow and outflow on section i	veh/s
$\lambda_i(t), \gamma_i(t)$,	Link queue and vacancy at section i	
D_i,D_{ri}, S_i	Demand, on-ramp demand, and supply of section i	veh/s
C_i, C_i^-,C_{ri},	Capacity, congested capacity, and on-ramp capacity	veh/s
$u_i(t)$, u_i^{min}	metering rate, minimum metering rate	veh/s
β_i	ratio of vehicles on link i that leaves the freeway at section i	-
Δ_i	Relative capacity drop ($\Delta_i = 1 - \frac{C_i^-}{C_i}$)	-

Table 4.1: Notation

.

The system dynamics is defined as:

$$\begin{aligned} &\frac{d}{dt}\mathbf{x}(t) = \Phi(\mathbf{x}(t), \mathbf{u}, \mathbf{d}, \psi) \\ &\quad s.t. \\ &\mathbf{x}(0) = \mathbf{x}_0 \end{aligned} \tag{4.1}$$

where the function Φ describes the system dynamics based on current state, metering rates, and demand. It is also considered the vector ψ with any further parameters that are model-specific such as fundamental diagram parameters, turning ratios, capacity drop ratio and so on. The system outputs is a scalar $\mathbf{y}(t)$ consisting in our case as the sum of outflows on off-ramps at sections 1 to I+1. There is an unique relationship between the state variables and the system output:

$$y(t) = C(\mathbf{x}(t), \mathbf{u}, \mathbf{d}, \psi). \tag{4.2}$$

The system (4.1) is in equilibrium when:

$$\frac{d}{dt}\mathbf{x}(t) = 0 \tag{4.3}$$

which is equivalent to $\Phi(\mathbf{x}(t), \mathbf{u}, \mathbf{d}, \psi) = 0$. It is assumed the function $\mathbf{X}$ and Y such that:

$$\begin{aligned} \mathbf{x}_{eq} &= \mathbf{X}(\mathbf{x}_0, \mathbf{u}, \mathbf{d}, \psi) \\ y_{eq} &= Y(\mathbf{x}_0, \mathbf{u}, \mathbf{d}, \psi) \end{aligned} \tag{4.4}$$

The reason we defined outflow, y, as the output is its direct relationship with total time spent. We can compute the cumulative arrival as:

$$A_i(t) = \sum_{i=0}^{I} \int_0^t d_i d\tau = t \sum_{i=0}^{I} d_i, \tag{4.5}$$

similarly the departure curve can be written as:

$$A_o(t) = A_0 + \int_{t_0}^{t} y_{eq} d\tau = A_0 + (t - t_0) y_{eq} \tag{4.6}$$

where t_0 is the instant in which it reaches the equilibrium and A_0 is the number of vehicles discharged up to time t_0. Thus, we can use basic queuing theory relationship to compute

the total time spent (TTS):

$$TTS = \int_0^t \big[A_i(\tau) - A_o(\tau)\big] d\tau \tag{4.7}$$

as is not possible to change the cumulative arrivals curve, the total time spent is decreased when the outflow is maximized [1]. Therefore, the optimal equilibrium is defined as:

$$y^\star = \max_{\mathbf{u},\mathbf{x}_0} Y(\mathbf{x}_0, \mathbf{u}, \mathbf{d}, \psi). \tag{4.8}$$

The classic formulation presented by [134] is consistent with this definition. On that formulation the goal was to maximize on-ramp flows - which in equilibrium is equivalent to maximizing outflow - while keeping mainline flows not exceeding the capacity at any section. The formulation for obtaining optimal flows without capacity drop presented on Section 4 is very similar to that work.

The maximum outflow $y^\star$ is reachable from initial state $\mathbf{x}_0$ if:

$$\max_{\mathbf{u}} Y(\mathbf{x}_0, \mathbf{u}, \mathbf{d}, \psi) = y^\star, \tag{4.9}$$

and $y^\star$ is defined as reachable if it is reachable from any $\mathbf{x}_0$

The system is defined to be stable with respect to initial states if for a small δ there is a ϵ

[1] Similar relationship and reasoning based on the cumulative arrivals and departures is presented in ([89], Equation 14)

such that:

$$||\mathbf{X}(\mathbf{x}_0, \mathbf{u}, \mathbf{d}, \psi) - \mathbf{X}(\mathbf{x}_0 + \epsilon, \mathbf{u}, \mathbf{d}, \psi)|| \leq \delta \tag{4.10}$$

which resembles Lyapunov Stability [6], but note this definition is only related to the equilibrium relationship. Lyapunov stability requires all trajectory $\mathbf{x}(t)$ to be close to equilibrium and this definition only requires as $t \to \infty$.

Similarly, the system output is stable with respect to initial states if for a small δ there is a ϵ such that:

$$||Y(\mathbf{x}_0, \mathbf{u}, \mathbf{d}, \psi) - Y(\mathbf{x}_0 + \epsilon, \mathbf{u}, \mathbf{d}, \psi)|| \leq \delta \tag{4.11}$$

It is defined similarly stability with respect to demands, $\mathbf{d}$, and metering rates, $\mathbf{u}$. This definition conceptually assesses the impact of small variations around an equilibrium. It is unstable when for small variations the system will drift away from the equilibrium leading to a completely different state and outflows. It is desirable that the equilibrium state of highest outflow to be stable because the system keeps discharging close to its maximum even in the presence of small disturbances. It does not happen with an unstable equilibrium and therefore it is unlikely to be observed [29].

Finally, another important property is uniqueness in which our case can be looked from different perspectives. The optimal output is unique with respect to metering rates if $y^\star$ can be achieved with only one $\mathbf{u}$. Similarly, y_{eq} is unique with respect to state if it can only be achieved by only one $\mathbf{x}_{eq}$.

4.3 Traffic Flow Model

On this study the dynamics on the mainline freeway is based on the link transmission model [139]. On the merge bottlenecks, the model is extended with a phenomenological capacity drop model from [65]. First, we detail the link dynamics on on-ramps and mainline freeway and later we introduce the node models on merges and diverges used to compute boundary flows.

4.3.1 Link Transmission Model

The state variables on the link transmission model is the upstream and downstream cumulative curves, similarly the Newell's model [100] which the link transmission model is derived from. On this study, however, we present an equivalent formulation [61, 63] based on queue and vacancy size in each link rather than upstream and downstream cumulative flows.

Each link has three parameters three parameters associated to the fundamental diagram [100]: the free-flow speed, V_i, the shock-wave traveling speed, W_i, and the jam density, K_i. These three parameters defines the critical density $K_{ci} = \frac{V_i W_i}{V_i + W_i}$ and the capacity $C_i = V_i K_{ci}$. Additionally, the link length is denoted as L_i.

Each link has two state variables, the queue (λ_i), and vacancy, (γ_i) which evolve as:

$$\begin{aligned} \frac{d}{dt}\lambda_i(t) &= f_i(t - \frac{L_i}{V_i}) - g_i(t) \\ \frac{d}{dt}\gamma_i(t) &= g_i(t - \frac{L_i}{W_i}) - f_i(t) \end{aligned} \tag{4.12}$$

where the boundary condition is determined by the initial queue, $\lambda_i(0)$, vacancy size, $\gamma_i(0)$,

the upstream flows, f_i, on the period $-\frac{L_i}{V_i} \leq t < 0$ and downstream flows, g_i, on the period $-\frac{L_i}{W_i} \leq t < 0$. If the link is initially empty with $f_i(t) = g_i(t) = 0$ for $t < 0$ corresponds to $\lambda_i = 0$ and $\gamma_i^{empty} = K_i L_i$, which we refer as holding capacity.

Supply and demand is computed based on the state variables:

$$\begin{aligned} D_i(t) &= \min\{f_i(t - \frac{L_i}{V_i}) + H(\lambda_i(t)), C_i\} \\ S_i(t) &= \min\{g_i(t - \frac{L_i}{W_i}) + H(\gamma_i(t)), C_i\} \end{aligned} \tag{4.13}$$

where $H(y)$, with $y > 0$, is the indicator function:

$$H(y) = \lim_{\Delta t \to 0^+} \frac{y}{\Delta t} = \begin{cases} 0, & y = 0, \\ \infty, & y > 0. \end{cases} \tag{4.14}$$

The on-ramp dynamics is not modeled explicitly and it is assumed a time dependent $d_i(t)$ at each on-ramp. The on-ramp flow can be metered by the metering rate $u_i(t)$ which is assumed to be on the interval

$$\min\{d_i(t), u_i^{min}\} \leq u_i(t) \leq \min\{d_i(t), Cr_i\}\}, \tag{4.15}$$

where u_i^{min} is the minimum metering rate at on-ramp i, Cr_i is the on-ramp capacity. With $u_i(t)$ on that range the on-ramp demand is determined by $u_i(t)$. The boundary flows are computed based on demand and supplies and are detailed on the following sub-sections.

4.3.2 Computation of Flows at Nodes

On the freeway under study, for each block there is a diverge node (on link i) and a merge node (on link i'). Flows are also needed to be computed at lane drop bottlenecks. A schematic with the related variables of a freeway block is depicted on Figure 4.2.

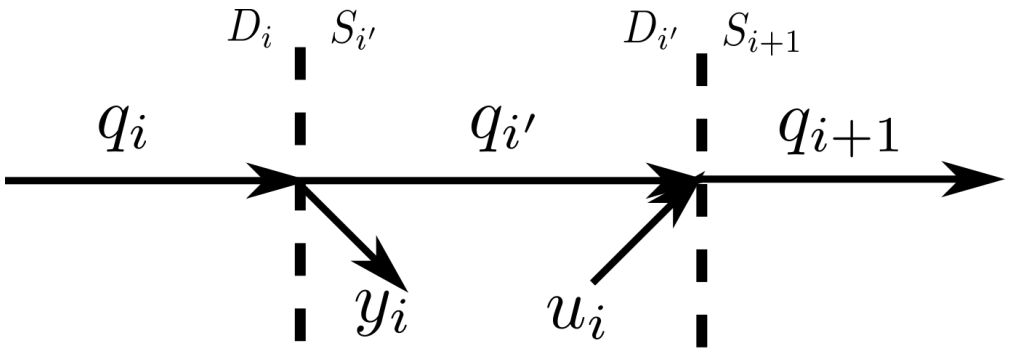


Figure 4.2: Schematic of a building block and variables associated to boundary flows computation

We refer q_i as the flow on link i that remains on the freeway, therefore the flows in each link become:

$$\begin{aligned} f_i(t) &= q_{(i-1)'}(t) + u_{i-1}(t) \\ g_i(t) &= q_i(t) + y_i(t) \\ f_{i'}(t) &= q_i(t) \\ g_{i'}(t) &= q_{i'}(t) \end{aligned} \tag{4.16}$$

and each demand and supply can be computed based on these flows, queues and vacancies from (4.13). We show in the next sub-sections how the flows are computed on the merge and diverge case.

On that case the downstream supply, S_{i+1} is serving both the upstream traffic demand,

$D_{i'}$ and the on-ramp demand, u_i. A distribution scheme is necessary when the sum of the demands exceeds the downstream supply. In this study, it is assumed absolute priority to the on-ramp [56] and the flows on the merge when the capacity drop is not considered as:

$$\begin{aligned} gr_i &= \min\{u_i, S_{i+1}\}, \quad \text{(on-ramp flow no capacity drop case)} \\ q_{i'} &= \min\{D_{i'}, S_{i+1} - gr_i\} \quad \text{(mainline flow no capacity drop case)}, \end{aligned} \tag{4.17}$$

where gr_i is the outflow of the on-ramp. Therefore, the on-ramp is first served and the remaining supply can serve the upstream flow.

When the capacity drop is considered the flow computation is slightly different. We integrate the capacity drop model from [61] in which the total flow is reduced when the total demand exceeds the downstream supply. Defining $\tilde{S}_{i+1}$ as effective supply computed as:

$$\tilde{S}_{i+1} = \min\{S_{i+1}, C_{i+1}(1 - \Delta_{i+1}\delta(D_{i'} + u_i - S_{i+1}))\} \tag{4.18}$$

where $\delta(x)$ is the step function:

$$\delta(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0, \end{cases} \tag{4.19}$$

The flows can be computed based on the effective supply, $\tilde{S}_{i+1}$ rather than the downstream

supply, S_{i+1}:

$$\begin{aligned} gr_i &= \min\{u_i, \tilde{S}_{i+1}\}, \quad \text{(on-ramp flow capacity drop case)} \\ q_{i'} &= \min\{D_{i'}, \tilde{S}_{i+1} - gr_i\} \quad \text{(mainline flow no capacity drop case)}, \end{aligned} \tag{4.20}$$

Note that the flow computation is neither modified demand or supply as $\tilde{S}_{i+1}$ depends on both upstream demand and downstream supply. Therefore, whenever the upstream demand is greater than the downstream supply the flow rate will be bounded by $C_{i+1}(1 - \Delta_{i+1})$, modeling the capacity drop phenomenon. When the downstream supply is enough to serve the upstream demand, the maximum flow is the downstream supply.

On the diverge node, we assume that the off-ramp has always enough supply to serve the incoming flow. So in this case the flow is determined by the upstream demand and downstream supply:

$$\begin{aligned} q_i &= \min\{D_i(1 - \beta_i), S_{(i+1)'}\} = \min\{D_i(1 - \hat{\beta}_i), S_{(i+1)'}\} \\ y_i &= q_i \beta_i \end{aligned} \tag{4.21}$$

where $\hat{\beta}_i = 1 - \beta_i$ is the proportion vehicle traveling at freeway at link i that remains at the freeway after on-ramp at link i.

4.3.3 Properties of Equilibrium States in a Link and in a block

Based on the link dynamics and the node models, we show some basic properties of the equilibrium on a general links and along a merge-diverge block. The properties shown in

this subsection will be used when presenting the equilibrium properties along the whole corridor.

Part of the following discussion is also present in [61] and [59] regarding the stationary states on a link. A link reaches an equilibrium on time t_0 if the following holds:

$$\begin{aligned} \frac{d}{dt}\lambda_i(t) &= 0 \quad t > t_0 \\ \frac{d}{dt}\gamma_i(t) &= 0 \quad t > t_0 \end{aligned} \tag{4.22}$$

Lemma 4.1. *An equilibrium is reached with* $f_i(t) = g_i(t) = z_i$.

Proof. From (4.12) and (4.22) follows that $f_i(t) = g_i(t) = z_i(t)$ for $t > t_0$. When, for a given $t_1 > t_0$, $z(t_1^-) \neq z(t_2^+)$, leads to $\lambda(t_1^-) \neq \lambda(t_1^+)$ and $\gamma(t_1^-) \neq \gamma(t_1^+)$ drifting away from the equilibrium. Therefore to remain at the equilibrium, $f_i(t) = g_i(t) = z_i(t) = z_i$ for $t > t_0$. □

This is equivalent to ([61], Theorem 4.2). On the link transmission model, the flow z_i can be achieved either by the uncongested or congested side of the triangular fundamental diagram, except when at capacity in which the relationship is unique. From a given stationary state, with flow, z_i, queue, λ_i, and vacancy γ_i we can obtain the density over space.

Let's Assume the link i initially empty at $t = 0$. In this situation $\gamma_i(0) = K_i L_i$, $\lambda_i = 0$, and the inflow for $t > 0$ $f_i(t) = D_i = z_i \leq C_i$. Assuming there is enough supply on the downstream link, the outflow will be $g_i(t) = z_i$ for $t \geq \frac{L_i}{V_i}$ and 0 for $t < \frac{L_i}{V_i}$. From mass

conservation, the number of vehicles $N_i(t)$ inside the link i is given by:

$$N_i(t) = N_i(0) + \int_0^t f_i(t)dt - \int_0^t g_i(t)dt, \tag{4.23}$$

The queue, from (4.12), is computed as:

$$\lambda_i(t) = \lambda_i(0) + \int_0^t f_i(t - \frac{L_i}{V_i}) - \int_0^t g_i(t)dt, \tag{4.24}$$

as it is assumed links initially empty, we can find a relationship between the number of vehicles inside the link and the queue:

$$N_i(t) = \lambda_i(t) + \int_{t-\frac{L_i}{V_i}}^t f(t)dt = \lambda_i(t) + z_i\frac{L_i}{V_i}. \tag{4.25}$$

In the case the link is uncongested and therefore $\lambda_i(t) = 0$, the density on the link,$k_i^u(z_i)$, can be computed as:

$$k_i^u(z_i) = \frac{N_i(t)}{L_i} = \frac{z_i}{V_i}, \tag{4.26}$$

corresponding to the uncongested branch of the triangular fundamental diagram.

Now let's assume similar situation with downstream supply as such that $g_i(t) = z_i$ for $t > \frac{L_i}{V_i}$, but with a higher upstream demand: $D_i = f_i = z_i + \epsilon$ with $\epsilon > 0$. In such situation, the

vacancy and queue for $t > \frac{L_i}{W_i}$ evolve as:

$$\begin{aligned}\lambda_i(t) &= \epsilon(t - \frac{L_i}{V_i}) \\ \gamma_i(t) &= K_i L_i + z_i(t - \frac{L_i}{V_i} - \frac{L_i}{W_i}) - (z_i + \epsilon)t\end{aligned} \tag{4.27}$$

which the queue will steadily grow and the vacancy will decrease. Considering it decreases until an arbitrary γ_i taking time t_1 such that:

$$t_1 = \frac{K_i L_i - \gamma_i}{\epsilon} - \frac{z_i}{\epsilon}(\frac{L_i}{V_i} + \frac{L_i}{W_i}), \tag{4.28}$$

and onwards $f_i = g_i = z_i$. The queue will settle after at time $t = t_1 + \frac{L_i}{V_i}$ with value:

$$\lambda_i(t_1 + \frac{L_i}{V_i}) = K_i L_i - \gamma_i - z_i(\frac{L_i}{V_i} + \frac{L_i}{W_i}). \tag{4.29}$$

In particular, the maximum queue in equilibrium for flow z_i is then:

$$\lambda_i^{max}(z_i) = K_i L_i - z_i(\frac{L_i}{V_i} + \frac{L_i}{W_i}) \tag{4.30}$$

From the maximum queue, we can plug into (4.25) to compute the congested density, $k_i^c(z_i)$:

$$k_i^c(z_i) = \frac{N_i}{L_i} = K_i - \frac{z_i}{W_i}, \tag{4.31}$$

which corresponds to the congested branch of the triangular fundamental diagram. Observe from (4.29) the unique relationship between queue and vacancy in equilibrium. Therefore we need just one of the state variables; having, for example, λ_i we can compute γ_i, D_i and S_i.

For any intermediary equilibrium state with $\gamma_i > 0$ and $\lambda_i > 0$, we define α_i such that:

$$\alpha_i = 1 - \frac{\lambda_i}{\lambda_i^{max}(z_i)} \tag{4.32}$$

as the uncongested fraction of the road. Considering the position v along the link, being $w = 0$ the upstream end and $w = L_i$, we can write the density in stationary state as:

$$k_i(w) = H(\alpha_i L_i - w)k_i^u(z_i) + (1 - H(\alpha_i L_i - w))k_i^c(z_i) \tag{4.33}$$

where $H(\cdot)$ is the Heaviside function. Assuming a link subject to total upstream demand D_i, no off-ramp flow (i.e., $\beta_i = 0$), upstream supply S_{i+1}, and with queue and vacancy defining α_i can reach 4 types of equilibrium states [61, 59]:

- Strictly under-critical (SUC): flow below capacity with the whole link uncongested which corresponds to $D_i < S_{i+1}$, $\alpha_i = 1$ ($\lambda_i = 0$);

- Strictly over-critical (SOC): flow below capacity with the whole link congested which corresponds to $D_i > S_{i+1} < C_i$, $\alpha_i = 0$ ($\gamma_i = 0$);

- Critical (C): flow, g_i, at capacity and the whole link is at critical density which corresponds to $g_i = C_i = D_i = S_{i+1}$, $\lambda_i = \gamma_i = 0$ (α_i is undetermined in this case);

- Zero-speed shock-wave (ZS): flow under capacity, the downstream end of the link congested while the upstream end of the link is uncongested. This corresponds to $g_i < C_i$, $D_i = S_{i+1}$ and $\alpha_i \in (0, 1)$.

Now we study the diverge node and its two related links where the link upstream of the off-ramp is i and downstream of the off-ramp the link (i'). We are assuming throughout this study the same road characteristics on both links.

Lemma 4.2. *At equilibrium,* $\lambda_i > 0 \implies \lambda_{i'} = \lambda_{i'}^{max}$

Proof. Let an initial state with $\lambda_i(0) > 0$ and $\lambda_{i'} = 0$. It leads to supply $S_{i'} = C_{i'} = C_i$ which will lead to $g_i = C_i$; if $f_i < C_i$ it will lead to $\lambda_i = 0$ eventually. If $f_i = C_i$, it is the critical case and the vacancy will be reduced until $\lambda_i = \gamma_i = 0$. With $\lambda_{i'} = \lambda_{i'}^{max}$, $S_{i'} = q_{i'}$ which can lead to $D_i > S_{i'}$ and is the only equilibrium case with $\lambda_i > 0$. □

In other words, with supply at capacity, there will be no queue on the link upstream of the off-ramp. To our discussion, it means that for computing the upstream supply, S_i, at equilibrium can be based on $\lambda^i = \lambda_i + \lambda_{i'}$ where λ^i is the queue associated with the $i - th$ block. We also define onwards $\beta^i = \beta_i$ ($\hat{\beta}^i = \hat{\beta}_i$), $gr^i = gr_i$, and $u^i = u_i$. We can compute $\lambda^{i,max}(q^i, fr_i)$ as:

$$\lambda^{i,max}(q^i, gr^i) = \lambda_{i'}^{max}(q_i) + \lambda_i^{max}(\frac{q_i}{\hat{\beta}^i}), \tag{4.34}$$

where the flow on the merge link is $q_{i'} = q^i$ and the flow on the diverge link is $q_i = \frac{q^i}{\hat{\beta}^i}$. The maximum queues is given by (4.30). We can recover the individual queues from λ^i as:

$$\begin{aligned} \lambda_{i'}(q^i) &= \min\{\lambda^i, \lambda_{i'}^{max}(q^i)\} \\ \lambda_i(q^i) &= \max\{0, \lambda^i - \lambda_{i'}^{max}(q^i)\} \end{aligned} \tag{4.35}$$

We can therefore compute the supply of the block based on λ^i:

$$S^i = \min\{\frac{q^i}{\hat{\beta}^i} + H(\lambda^i - \lambda^{i,max}), C^i\}. \tag{4.36}$$

Therefore given known on-ramp flows and turning ratios at the diverge nodes, we can deal with only the aggregated queue on a block merge-diverge. Onwards, we base our analysis assuming as state variables the bottleneck queue λ^i with associated outflow $q^i = q_{i'}$. Assuming the changes on the queues happened on a higher time scale compared to free-flow and shock-wave travel times, the queue evolve as:

$$\frac{d}{dt}\lambda^i(t) = q^{i-1}(t) + gr^{i-1}(t) - y_i(t) - q^i(t) = \hat{\beta}^i(q^{i-1}(t) + gr^{i-1}(t)) - q^i(t). \tag{4.37}$$

4.4 Equilibrium, Stability and Reachability Without Capacity Drop

In this section the properties of the equilibrium states of a freeway corridor. First, we analyzed the model and show the properties and later we confirm the results with numerical experiments.

4.4.1 Model Analysis

We now use the link transmission model presented on Section 4.3 to analyze the system without capacity drop. This scenario is very similar to [42] and the properties shown are equivalent. One assumption made is that the metering rate is in the range $\min\{u^{min,i}, S^{i+1}\} \leq u^i \leq \min\{d^i, S^{i+1}\}$ and in that case $gr^i = u^i$ on all on-ramps.

The equilibrium condition for a single link is given by (4.22). It is assumed known and constant metering rates $\mathbf{u} = [u^1, u^2, ..., u^I]$ and upstream demand, d_0. We are considering as state variables the bottlenecks queues, λ^i. It is considered equilibrium if all bottlenecks are in equilibrium. In this case, the equilibrium is reached at time t_0 if:

$$\frac{d}{dt}\lambda^i(t) = 0, t > t_0 \qquad i = 1, ..., I, \tag{4.38}$$

which plugging into (4.37) and with all variables not varying in time:

$$q^i = \hat{\beta}^i(q^{i-1} + u^{i-1}). \tag{4.39}$$

Applying recursively for $i-1$ as function of $i-2$ to the first section we can find q^i in terms of the on-ramp flows and upstream flow q^0:

$$q^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i-1} u^k \hat{\beta}^{ki} \quad i \geq 1 \tag{4.40}$$

where q^0 is the upstream flow into the first section and $\hat{\beta}^{ki} := \hat{\beta}^{k+1} ... \hat{\beta}^i$ is the share of the vehicles traveling or enter at on-ramp k that are still traveling on section i. We also define:

$$q^i(j, q^j) = \begin{cases} q^j \hat{\beta^{ji}} + \sum_{k=j}^{i-1} u^k \hat{\beta^{ki}} & i > j \\ \frac{q^j}{\hat{\beta^{ij}}} - \sum_{k=j}^{i-1} u^k \hat{\beta^{jk}} & i < j \end{cases} \tag{4.41}$$

as the resulting flow on section i given section j flow is q^j. A pair $(\mathbf{q}, \lambda)$ if the boundary flows based on states λ leads to flows $\mathbf{q}$. Applying that on the merges:

$$\begin{aligned} q^i &= \min\{D^i(\lambda^i), S^{i+1}(\lambda^{i+1}) - u^i\} \\ &= \min\{q^i + H(\lambda^i), C^i, \frac{1}{\hat{\beta}^{i+1}} q^{i+1} + H(\lambda^{i+1,max} - \lambda^{i+1}) - u^i, C^{i+1} - u^i\} \\ &= \min\{q^i + H(\lambda^i), C^i, q^i + H(\lambda^{i+1,max} - \lambda^{i+1}), C^{i+1} - u^i\} \end{aligned} \tag{4.42}$$

Observe from (4.40) that $\lambda^i > 0$ leads to $D^i = \min\{q^i + H(\lambda^i), C^i\} = C^i$. As for $q^i = C^i$ would be a critical equilibrium in a link and for that case $\lambda^i = 0$, necessarily $S^{i+1} - u^i < C^i$. The flow in that case will be $q^i = S^{i+1} - u^i = \min\{C^{i+1} - u^i, q^i + H(\lambda^{i+1,max} - \lambda^{i+1})\}$. If $C^{i+1} - u^i < q^i + H(\lambda^{i+1,max} - \lambda^{i+1})\}$ corresponds to the active bottleneck case with the congestion starting just at bottleneck i. Otherwise, it corresponds to the queue spill-back

mechanism and q^i is the remaining supply of a queue that has started on an active bottleneck downstream of section i.

In the case of the queue spill-back case, the Lemma 4.2 applies and the state can be grouped with its downstream active bottleneck as in Equation (4.35), but grouping queues of different blocks instead of grouping the two queues of the same block. This means the state of the freeway corridor is determined by the active bottleneck queues. It is consistent with [42] that having K active bottlenecks, the system state can be given by its related K queues $\mathbf{\Lambda} = [\Lambda^1, \Lambda^2, ..., \Lambda^K]$.

We define $\widehat{d^i}$ the reflected demand into bottleneck i computed as:

$$\widehat{d^i} = \hat{\beta}^{0i} d_0 + \sum_{k=1}^{i} u^k \hat{\beta}^{ki} \tag{4.43}$$

which is the flow induced by the demand into section $i+1$ by on-ramps and upstream demands.

Lemma 4.3. *$\widehat{d^i} < C^{i+1}$ there is an unique equilibrium with $q^i = d^i - u^i$ and $\lambda = 0$*

Proof. If all sections starts congested, the bottleneck I will have flow $\bar{q}^I = C^{I+1} - u^I$ leading eventually to supply $S^I = \frac{1}{\beta_i}\bar{q}^I$; the bottleneck $I-1$ flow will be $\bar{q}^{I-1} = q^{i-1}(I, q^I) = \frac{1}{\beta_i}\bar{q}^I - u^{I-1}$ and similarly to the very first section. At the first section $d_0 < \frac{1}{\beta_1}\bar{q}^1 = \frac{1}{\beta_1}q^1(I, q^I)$ and $\frac{d}{dt}\lambda_1(t) < 0$ until $\lambda_1 = 0$ and later $q_1 = \beta_1 d_0 = \widehat{d^1} - u^1$ which in turn leads to $\frac{d}{dt}\lambda_2(t) < 0$ until clear the queue and $q_2 = \beta_2\beta_1 d_0 + \beta_2 u^1 = \tilde{d^2} - u^2$ and then successively until section I. □

Lemma 4.3 shows that regardless of initial conditions, eventually at some upstream link (and it was shown to the worst case), the upstream demand will be strictly smaller than

downstream supply and the queue is dissipated followed by a smaller flow to the upstream link. This then lead to smaller demand downstream repeating this process until congestion at all sections is relieved. In case up to link i is initially congested, $q^i = \min\{\widehat{d^i} - u^i, C^{i+1}\} = \widehat{d^i} - u^i$ as $\tilde{d^i} < C^{i+1}$ and the link remains uncongested.

Lemma 4.4. *For $\widehat{d^i} < C^{i+1} \quad \forall i \neq j$ and $\widehat{d^j} > C^{j+1}$ there is an unique equilibrium with flows $q^j = C^{j+1} - u^j$ and $q^i(j, q^j) \quad for i \neq j$ and queues $\lambda^i = \lambda^{i,max}(q^i)$ for $i \leq j$ and $\lambda^i = 0$ for $i > j$.*

Proof. From section 1 to $j-1$ Lemma 1 applies and would have no queues if $\lambda^j < \lambda^{j,max}$. At section j, $\widehat{d^j} > C^{j+1}$ leading to $q^j = C^{j+1} - u^j$ and $\frac{d}{dt}\lambda^j(t) > 0$. Queue grows until their maximum and then reducing supply and $q^{j-1} = q^{j-1}(j, q^j)$ successively until the first section in which $q^0 < d_0$. Downstream to section j, from Lemma 4.3 $\lambda^i = 0$ as $\widehat{d^i} + u^i < C^{i+1}$ from Lemma 4.3. □

Therefore, when one bottleneck has reflected demand strictly higher than its downstream capacity, the congestion will start at that bottleneck and propagate until the first section eventually leading to $q^0 < d^0$.

Lemma 4.5. *For $\widehat{d^i} < C^{i+1} \quad \forall i \neq j$ and $\widehat{d^j} = C^{j+1}$ there is an unique equilibrium with flows $q^j = C^{j+1} - u^j$ and $q^i(j, q^j) = \tilde{d^i} - u^i \quad for i \neq j$ and queues λ^i depending on $\lambda(0)$ for $i \leq j$ and $\lambda^i = 0$ for $i > j$.*

Proof. It follows Lemma 4.4, but at section j, $q^j = C^{j+1} - u^j$. Case $\lambda_i(0) = 0$ for all $i < j$, it leads to $\frac{d}{dt}\lambda_j(t) = 0$ and therefore $\lambda_j(t) = \lambda_j(0)$ and $\lambda_i(0)$ for $i \neq j$. The dependency on $\lambda(0)$ we show with a counterexample. Let the initial state be the same $\lambda_j(0)$, but $\lambda^{j-1}(0) > 0$. It leads immediately to $q^j(t) = C^{j+1} - u^j$ and $q^{j-1}(t) = C^j - u^{j-1} > q^{j-1}(j, q^j)$ leading to $\frac{d}{dt}\lambda_{j-1}(t) < 0$ and $\frac{d}{dt}\lambda_j(t) > 0$ until either $\lambda^j(t) = \lambda^{j,max}(q^j)$ or $\lambda^{j-1}(t) = 0$ reaching equilibrium afterwards. □

Note that for any section $i < j$ the Lemma 4.2 applies as $\widehat{d^i} < C^{i+1}$ and there will be queues in equilibrium at section i only if the section $i+1$ is at its maximum. For $\lambda^{j,max}(q^j) > \lambda^j(0) > 0$ and $\lambda_i(0) = 0$ for $i \neq j$. In that case, $\lambda(t) = \lambda(0)$. The counterexample case, with $0 < \lambda^j(0) < \lambda^{j,max}$ and $0 < \lambda^{j-1}(0) < \lambda^{j-1,max}$ the queue on $j-1$ is transferred to section j until it clears queue $j-1$ or queue at section j is at maximum. What happens in this case is that the vehicles queued upstream of section j that remains at the corridor after section j will remain queued, but the queue starts from section j and Lemma 4.7 applies. Therefore, we can combine queues as Eq. (4.35) for queue $j-1$ and j and then the combined queue with $j-2$ until the first section resulting in:

$$\begin{aligned}
\Lambda^1(t) &= \sum_1^j \lambda^j(t) \\
\lambda^j(t) &= \min\{\Lambda^j(t), \lambda^{j,max}(q^j = C^{j+1} - u^j)\} \\
\lambda^{j-1} &= \min\{\Lambda^j(t) - \lambda^j(t), \lambda^{j-1,max}(q^{j-1}(j, q^j))\} \\
&\ldots \\
\lambda^1 &= \min\{\Lambda^j(t) - \sum_{i=2}^j \lambda^i(t), \lambda^{1,max}(q^1(j, q^j))\}
\end{aligned} \tag{4.44}$$

Lemma 4.6. *For $\lambda(0)$ respecting (4.44) and $\widehat{d^i} + < C^{i+1} \quad \forall i \neq j$ and $\widehat{d^j} + gr^j = C^{j+1}$, the equilibrium is reached with $\Lambda^1(t) = \Lambda^1(0)$.*

Proof. Immediately $q^i = \widehat{d^i} - u^i$ and $\frac{d}{dt}\lambda_i(t) = 0$ for all sections and therefore $\frac{d}{dt}\Lambda^1(t) = 0$ and $\Lambda^1(t) = \Lambda^1(0)$. □

In all previous cases it was assumed that at one specific bottleneck the reflected demand, $\widehat{d}$, was equal or exceeded the downstream capacity. We now extend to the case in which happens in multiple bottlenecks.

Lemma 4.7. *If $\widehat{d^i} \geq C^{i+1}$ for $i < j$, $\widehat{d^j} \geq C^{j+1}$ and for some $k > j$ $q^k(j, C^{j+1} - u^j) + u^k > C^{k+1}$, the active bottleneck k and the queue state is based on k.*

Proof. In that case, the flow at section k is $q^k = C^{k+1} - u^k$, the queues grows from section k and eventually $S_{j+1} = q^j(k, q^k) + u^j < C^{j+1}$ therefore the active bottleneck is k. □

In that case the queue starts from section k and will propagate until section 1 as Lemma 4.4 holds.

Lemma 4.8. *If $\widehat{d^i} \geq C^{i+1}$ for $i < j$, $\widehat{d^j} \geq C^{j+1}$ and for some k $q^k(j, C^{j+1} - u^j) + u^k = C^{k+1}$, with queues from $j+1$ to k merged as (4.44) as $\Lambda^2(t) = \lambda^{j+1}(t) + ...\lambda^k(t)$ the state is $\Lambda^2(t) = \Lambda^2(0)$.*

Proof. It follows from Lemma 4.7 by split the sections $j+1$ to k, numbered 1 to $k-j$, as independent corridor with upstream flow $d_0 = C^{j+1}$ □

From section k, it is possible that Lemma 4.8 holds successively into downstream sections. However, note that if Lemma 4.7 holds, all independent corridors are again merged to that bottleneck. With that the equilibrium states are qualitatively characterized. We identify the most downstream overcritical bottleneck, define as k_1, by checking from the first to the last section as follows:

$$\begin{aligned} &q_0 = d_0, k_0 = 0 \\ &q^j = \min\{\hat{\beta}^j(q^{j-1} + u^{j-1}), C^{j+1} - u^j\}, j = 1, ..., I \\ &k_0 = j \quad if \quad C^{j+1} - u^j > \hat{\beta}^j(q^{j-1} + u^{j-1}) \end{aligned} \tag{4.45}$$

The corridor will be completely congested from section k_1 until section 0. Note that if the

section if $k_1 = I$, the whole corridor congested is the unique equilibrium. After k_1, k_2, ..., k_m can be found from Lemma 4.8 with $j = k_0$ and $k_1, ..., k_m, ..., k_M$ are sections in which $q^m(k_0, q^{k_0}) + u^{k_m} = C^{k_m+1}$. Therefore, the queues from $k_0 + 1$ to k_1 can be grouped as Λ^1, from $k_1 + 1 to k_2$ as Λ^2 until Λ^M and in equilibrium the system state can be given by its related M active bottlenecks queues $\mathbf{\Lambda} = [\Lambda^1, \Lambda^2, ..., \Lambda^M]$. This transformation is equivalent of ([42], Theorem 4.1).

Note it is consistent with the trivial cases. If $\widehat{d^i} + u^i < C^{i+1}$, $k_0 = 0$ and $M = 0$ is the case with an unique uncongested equilibrium. If $k_0 = I$, $M = 0$, the whole corridor is congested. Figure 4.3 illustrate a case in with $k_0 > 0$ and $M = 2$.

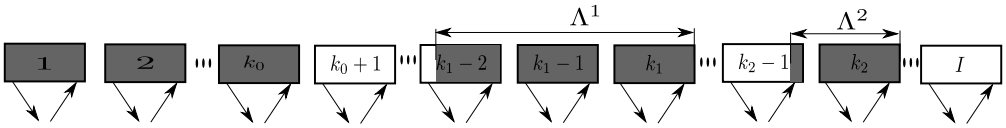


Figure 4.3: Representation of state based on the active bottlenecks with $M = 2$

We now can show the qualitative aspects of the equilibrium such as stability, reachability, and uniqueness. The total outflow can be computed, in equilibrium, as:

$$y = q_0 + \sum_{i=1}^{I} u^i \tag{4.46}$$

which is total inflow entering the corridor as in equilibrium the inflow and outflow are balanced. We are considering a given equilibrium as a pair $(\lambda_{eq}, \mathbf{u})$ with its associated y_{eq}. Similarly, it is referred as $\tilde{y}_{eq}$, $\tilde{\lambda}$ and $\tilde{\mathbf{q}}$ an equilibrium reached with a perturbation in the state or inputs. Also, a λ defines a $\mathbf{\Lambda}$ and vice-versa.

Theorem 4.9. *The system state and output is stable with respect to the initial states.*

Proof. Consider a pair $(\mathbf{\Lambda}_{eq})$. Case with $M = 0$ and $k = 0$ the equilibrium is no queues and the system returns to the uncongested equilibrium based on Lemma 4.3 and for $M = 0$ and

$k = I$ it returns to the complete congested equilibrium based on Lemma 4.4 and 4.7. For the cases with $k > 0$, for each bottleneck $\Lambda^k(0) = \Lambda^k_{eq} + \epsilon$ leads to $\Lambda^k(t) = \Lambda^k(0)$ from 4.6 with no change on inflows and outflows. Therefore for all cases exist a δ respecting inequalities (4.10) and (4.11). □

Basically, if it starts either completely congested or uncongested ($k = 0$) the system returns to the previous equilibrium state. For $k > 0$ the perturbation on the initial state remains unchanged. It follows intuition in each case, if the demand is smaller than capacity, any queue that just builds up (say, following a small incident) is cleared afterwards. In the case of demands higher than capacity, a small decrease on the queue (say, following some periods of higher bottleneck outflow) will be followed by a return to the maximum queue. If demand and capacity are almost the same, a perturbation on the initial state remains untouched. That is, if the queue grew after a period of smaller outflow, the system will remain close to this state of larger queue.

In the case of perturbation on the input flows the behavior is different with respect to the impact of perturbation on the system states, but similar with respect to the output. We define the on-ramp i as under critical on-ramp if $\widehat{d^j} < C^{j+1}$ for $i \geq j$, critical if $\widehat{d^j} = C^{j+1}$ for $i \geq j$ and overcritical if $\widehat{d^j} > C^{j+1}$ for $i \geq j$.

Lemma 4.10. *The system state and output is stable with respect to perturbation on an under critical on-ramp.*

Proof. An equilibrium in this case is $\widehat{d^j} < C^{j+1}$, $\lambda = 0$, and flows $q^j = \widehat{d^j} - u^j$ for $j > i$. If $\tilde{u^i} = u^i + \epsilon$ leads to $\tilde{d^j} + \epsilon\beta_{i,j} < C^{j+1}$. From Lemma 4.3, the equilibrium is $\lambda = 0$ ($\tilde{\lambda_{\mathbf{eq}}} - \lambda_{\mathbf{eq}} = 0$) and $\tilde{y_{eq}} - y_{eq} = \epsilon$. □

Lemma 4.11. *The system state is unstable for a perturbation in a critical on-ramp while the output is stable for the same perturbation.*

Proof. An equilibrium with states λ_{eq} with input i at on-ramp i and $\widehat{d^j} = C^{j+1}$ for $j > i$. With $\tilde{u}^i = u^i + \epsilon$ leads to $\tilde{d}^j + \epsilon\beta_{i,j} \neq C^{j+1}$. If $\epsilon > 0$ it leads to $\lambda^k = \lambda^{max}$ for $k < j$ by Lemma 4.4 and if $\epsilon < 0$ to $\lambda = 0$ for $i < k \leq j$ by Lemma 4.3. Therefore there is no ϵ such that $|\lambda^k_{eq} - \tilde{\lambda}^k_{eq}| < \delta$ for small δ. Regarding the output, for $\epsilon < 0$, $\tilde{y_{eq}} - y_{eq} = \epsilon$ and for $\epsilon > 0$, $\tilde{y}_{eq} = y_{eq} + \epsilon - (q^0(j, C^{j+1}) - \tilde{q}^0(j, C^{j+1})) = y_{eq} - \epsilon\frac{1-\beta^{0i}}{\beta^{0i}}$. □

The interpretation of Lemma 4.11 is that in a scenario leading to demand equal to capacity for some bottleneck, a small increase (decrease) in the demand will turn section fully congested (uncongested), but with rather small change in the total outflow. It means that a congested freeway does not necessarily means that it is leading to larger delays.

Lemma 4.12. *The system state and output is stable with respect to perturbation on an over critical on-ramp.*

Proof. It keeps $\lambda_{eq} = \lambda^i_{eq} = \tilde{\lambda}^i_{eq} = \lambda^{i,max}$ for $i < j$ from Lemma 4.4 as $\widehat{d^j} \geq C^{j+1}$ from where j is the active bottleneck. The change in the outflow is $\tilde{y}_{eq} = y_{eq} + \epsilon - (q^0(j, C^{j+1}) - \tilde{q}^0(j, C^{j+1})) = y_{eq} - \epsilon\frac{1-\beta^{0i}}{\beta^{0i}}$. □

From Lemmas 4.10,4.11 and 4.12 we can conclude the output is always stable. Again, the importance is that variations on the on-ramp flows will never lead to large changes in the total outflow of the corridor. On the case of a critical on-ramp, the freeway can turn from completely uncongested to congested with a small decrease in the on-ramp flows but the impact in the total outflow is marginal.

After establishing basic properties of equilibrium states without the presence of capacity drop phenomenon, we now turn the discussion to the properties of equilibrium states regarding performance. Questions to related to this is whether there is optimal equilibrium, whether it is unique and whether we can settle the system at that equilibrium based on given initial conditions.

In this problem, we want to maximize outflow y by changing the on-ramp flows $\mathbf{u}$. The flows in equilibrium state is given by (4.40) with u^k given by the control, but q^0 is not controlled. Also observe from (4.42) that flows respecting (4.40) that does not exceed downstream capacity (i.e., $q^i < C^{i+1} - u^i$) always has enough supply due to downstream queue as (as $S_{i+i} = \min\{q^i + H(\lambda^{i+1,max} - \lambda^{i+1}, C^{i+i}\})$. In that case, we can formulate the problem of optimal equilibrium as:

$$
\begin{aligned}
&P1 : \max_{\mathbf{u},q_0} y = q_0 + \sum_{i=1}^{I} u^i \qquad \text{s.t.} \\
&\min\{u^{i,min}, d^i\} \leq u_i \leq \min\{u^{i,max}, d^i\} \qquad i = 1, ..., I \\
&q^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i-1} u^k \hat{\beta}^{ki} \qquad i = 1, ..., I \\
&q^i + u^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i} u^k \hat{\beta}^{ki} \leq C^{i+1} \qquad i = 1, ..., I \\
&q_0 = \min\{\frac{q^1}{\beta^1}, d_0\}
\end{aligned}
\tag{4.47}
$$

The optimization problem (4.47) resembles the classic linear programming formulation in [134]. The only conceptual difference is the possibility of part of the upstream demand not being served. Observe that with min operator on Eq. (4.47) the optimization problem is not convex.

Theorem 4.13. *The* $x = \min\{a, b\}$ *operator on* (4.47) *can be exchanged by two linear inequalities* $x \leq a$ *and* $x \leq b$.

Proof. An initial solution to the problem is $\mathbf{u} = \mathbf{u}^{min}$. The remaining set of constraint is $\hat{\beta}^{0i} q^0 + \sum_{k=1}^{i} u^k \hat{\beta}^{ki} < C^{i+1}$. For any section i, the highest increase in the objective is to

add flow on the on-ramp j such βji is the smallest. As $\beta_{i-1,i} > \beta_{i-2,i} ... > \beta_{1,i}$ therefore the solution is obtained by increasing in order q_0,u_1, u_2..., u_I. Therefore, the min is always attained as the optimal solution is the higher value possible of q_0 in the feasible set. □

Therefore the problem (4.47) can be solved through a linear programming solver. The optimal equilibrium is achieved by the $\mathbf{u}$ that solves (4.47). Also see from the structure of the problem that whenever there are enough demand, the optimal solution will be binding with $q^i + u^i = C^{i+1}$ in one or more bottlenecks.

Another aspect worth pointing out is the lack of dependence of the optimal equilibrium with respect to initial initial state. The optimal outflow can be reached either from $\lambda_{eq} = 0$ or $\lambda = \lambda^{\mathbf{max}}$ or any intermediate state.

Lemma 4.14. *The optimal* $\mathbf{u}^\star$ *is unique if* $\beta_i > 0 \quad \forall i$

Proof. It follows from Theorem 4.13 that the optimal is always reached by increasing on-ramp flows from upstream to downstream whenever $\hat{\beta}_{i-1,j} < \hat{\beta}_{i,j}$ which holds for $\beta_i > 0$. □

Therefore, without capacity drop the ramp metering control problem has all desirable properties in equilibrium: it is stable, reachable and unique. We confirm all these properties with numerical experiments in the next sub-section.

4.4.2 Numerical Experiments

We verify the results using the link transmission model in a small network with $I = 2$ with $\beta^1 = 0$ and $\beta^2 = 0.1$. All the blocks have the same parameters with all the links with length $L = 600$m, free-flow speed $V = 30$m/s, shock-wave speed $\omega = \frac{1}{5}V$ and $k_j = 3/10$ veh/m ($\approx$ 3 lanes). With that parameters, the capacity is $C = 1.5veh/s$. Demands are $d_1 = d_2 = 0.3C$

and $d_0 = 0.8C$ and minimum metering rates are all and set to $u^{i,min} = u^{min} = 0.1C$. The total simulation time is 3 hours.

The optimal on this case is let the first bottleneck flow to capacity (i.e., $u^1 \geq C - d_0$) and $u^2 = C\beta^2$. Setting $u^1 > C - d_0$ will make the first bottleneck congested, but will not affect the outflow as there is no vehicle leaving the freeway upstream to that bottleneck ($\beta^1 = 0$). Increasing u^2 would be harmful because the total outflow would decrease. We show on Figure the total outflow and contour plot for 4 cases:

- case (a) - with $u^1 = C - d_0$ and $u^2 = C\beta^2$;
- case (b) - with $u^1 = 1.01(C - d_0)$ and $u^2 = C\beta^2$;
- case (c) - with $u^1 = C - d_0$ and $u^2 = 1.2C\beta^2$;
- case (d) - three different demand patterns. First third of the simulation with $u^1 = 1.01(C - d_0)$ and $u^2 = 1.2C\beta^2$, second third with $u^1 = 1.01(C - d_0)$ and $u^2 = C\beta^2$, and the last third with $u^1 = 0.95(C - d_0)$ and $u^2 = 0.95C\beta^2$.

The higher metering rate than optimal on case (c) throughout all the simulation should lead to the smaller outflow in all cases through the same principle in Eq. (4.20). On Figure 4.4 the queues (top and middle graphs) for each case and the total outflow and cumulative flow in oblique coordinates on the bottom graphs. The oblique coordinate (bottom right) helps us to see the cumulative effect of the flow compared to a baseline flow in which we picked $1.09C$.

Both cases (a) and (b) leads to exactly same outflow as the outflow and cumulative flow curves for each case exactly overlaps, but the first bottleneck is congested on case (b). However, observe the comparison between case (c) and (d) on the first third of the simulation. The case (c) the queue starts from section 3 and spills back until the very first section. Observe it

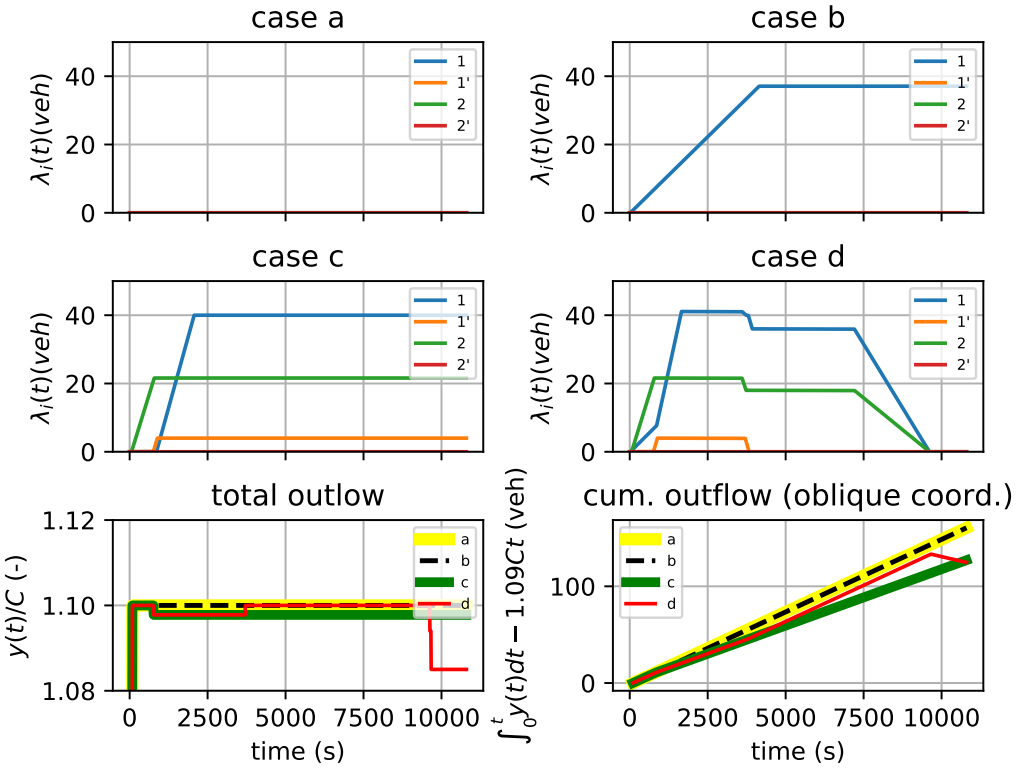


Figure 4.4: Cumulative outflow for each of experiment with no-capacity drop.

follows Lemma 4.2 and queues appears on upstream links once the queue on the downstream reaches its maximum. The outflow starts at the maximum and decreases around $t = 800s$ when the queue reaches the off-ramp. Case (d) leads it to start with two active bottlenecks with λ^1 and λ^2 increasing on the beginning, but the queue of the downstream hits the bottleneck and the both queues are combined (Lemma 4.7) and when the equilibrium in the fist half is reached (around $t = 2000s$) there is a single queue on the whole corridor. However, on case (d) the metering rates are set exactly as the case (b) and, like case (b), the maximum outflow was reached after the it, but with queues in both bottlenecks. This confirms it is always reachable regardless of the initial state.

The results also depicts the behavior with respect to small perturbation. From case (a) to (b) a small perturbation on u^1 turned the first section completely congested. Similar to case (d) in which towards the end of the simulation a small perturbation on both on-ramp flows

relieved the congestion. However, the variations in the total outflow were rather small for each case which confirms that the output is stable.

4.5 Equilibrium, Stability and Reachability With Capacity Drop

We now look to the equilibrium properties for the case in which capacity drop is considered. In the following sub-section we analyze the system model considering the capacity drop model and then later we confirm the results with numerical experiments.

4.5.1 Model Analysis

The relationships for equilibrium flows still holds for the capacity drop case and more specifically we can use (4.40) and (4.41). The fundamental difference regarding the capacity drop is the relationship between flows and queues on the merges in equilibrium is given by:

$$\begin{aligned} q^i &= \min\{D_i, \tilde{S}_{(i+1)'} - u^i\} \\ &= \min\{q^i + H(\lambda^i), C^i, \frac{1}{\hat{\beta}^{i+1}} q^{i+1} + H(\lambda^{i+1,max} - \lambda^{i+1}) - u^i, C^{i+1}(1 - \Delta_{i+1}\delta(x)) - u^i\} \\ &= \min\{q^i + H(\lambda^i), C^i, q^i + H(\lambda^{i+1,max} - \lambda^{i+1}), C^{i+1}(1 - \Delta_{i+1}\delta(x)) - u^i\}, \end{aligned} \tag{4.48}$$

where $x = D^i + u^i - S^{i+1}$. The demand, D^i, and supply, S^{i+1}, depend on the queues λ^i and λ^{i+1} respectively. Assuming $C^i = C^{i+1}$, the computation of x can be divided based on the queues:

1. $\lambda^i = 0$, $\lambda^{i+1} < \lambda^{i+1,max}$: $x = q^i + u^i - C^{i+1}$,

2. $\lambda^i = 0$, $\lambda^{i+1} = \lambda^{i+1,max}$: $x = q^i + u^i - q^i = u^i > 0$,

3. $\lambda^i > 0$, $\lambda^{i+1} < \lambda^{i+1,max}$: $x = C^i + u^i - C^{i+1} = u^i > 0$,

4. $\lambda^i > 0$, $\lambda^{i+1} = \lambda^{i+1,max}$: $x = C^i + u^i - q^i = u^i + (C^i - q^i) > u^i > 0$.

Note that only the first case can lead to $x < 0$ with $q^i + u^i < C^{i+1}$; all other cases the flow in equilibrium will be bounded by the dropped capacity. Nonetheless, the flow computation for $\lambda^i > 0$ follows the same pattern as the no capacity drop case. The queue spill-back case is when $\lambda^{i+1} = \lambda^{i,max}$ and the flow at segment i decreases as soon as the queue at bottleneck $i+1$ reaches its maximum. The active bottleneck case happens when the upstream demand exceeds the downstream supply and in that case $q^i = C^i(1-\Delta_{i+1}) - u^i$ and queues will grow.

We now highlight the differences between the case in which the capacity drop is not considered.

Lemma 4.15. *For $I = 1$ and $C^2(1-\Delta_2) < \hat{d}^1 < C^2$, there exist an uncongested equilibrium with $\lambda^1 = 0$ and $q^i = \hat{d}^1$ for $\lambda^1(0) = 0$ and a congested equilibrium with $q^i = C^2(1-\Delta_2) - u^1$ and $\lambda^1 = \lambda^{1,max}$ for $\lambda^1(0) > 0$.*

Proof. If $\lambda^1(0) = 0$, applying in (4.48) with $D_1 = \hat{d}^1$ leads to $q^1 = \hat{d}^1$ and $\lambda^1(t) = 0$. For $\lambda^1(0) > 0$, $D^1 = C^1$ and $q^1 = C^2(1-\Delta_2)$ and $\frac{d}{dt}\lambda^1(t) > 0$ growing until $\lambda^1 = \lambda^{1,max}$. □

Lemma 4.15, though restricted to a single bottleneck, shows that the flows are dependent on the initial state unlike the case without capacity drop. It also contradicts Lemma 4.3 and therefore an uncongested equilibrium is not necessarily reached by setting $\hat{d}^i < C^{i+1}$ at all bottlenecks.

Lemma 4.16. *$\widehat{d^i} < C^{i+1}(1-\Delta_{i+1})$ there is an unique equilibrium with $q^i = d^i - u^i$ and $\lambda = 0$*

Proof. Follows the same condition and proof as Lemma 4.3 □

Lemma 4.16 is the counterpart of 4.3, but an uncongested equilibrium is only guaranteed with reflected demand smaller than the dropped capacity rather than capacity. When that happens, any perturbation that causes a congestion eventually is relieved; otherwise, if $\widehat{d^i} > C^{i+1}(1-\Delta_{i+1})$ the queue would grow once the section i becomes congested.

Lemma 4.15 was limited to a single bottleneck because the propagation of the congestion upstream to that bottleneck is also different. Without capacity drop, whenever $\widehat{d^i} < C^{i+1}$, the section 1 to i will be congested when the equilibrium is reached by Lemma 4.4.

Lemma 4.17. *For an active bottleneck i with $0 < \lambda^i(t) < \lambda^{i,max}$ with $q^i = C^{i+1}(1-\Delta_{i+1}) - u^{i+1} > \hat{d}^i - u^i$, the active bottleneck becomes $i-1$ if $C^i(1-\Delta_i) < \frac{1}{\hat{\beta}^i}[C^{i+i}(1-\Delta_{i+1}) - u^i]$ leading to $\frac{d}{dt}\lambda^i(t) < 0$ and $\frac{d}{dt}\lambda^{i-1}(t) > 0$.*

Proof. Once $\lambda^i(t) = \lambda^{i,max}$ the flow $q^{i-1} + u^{i-1} = \min\{S^i, C^i(1-\Delta^i)\} = \min\{\frac{1}{\hat{\beta}^i}[C^{i+1}(1-\Delta_{i+1}) - u^i], C^i(1-\Delta_i)\}$ and for $C^i(1-\Delta_i) < S^i$ leads to $q^{i-1} + u^{i-1} < \frac{1}{\hat{\beta}^i}q^i$ and therefore $\frac{d}{dt}\lambda^i(t) < 0$ which decreases until $\lambda^i = 0$. □

The outcome given by Lemma 4.17 is counterintuitive. When a congestion that started at section i reaches the section $i-1$ can trigger the capacity drop at the upstream section $i-1$. If the flow fall belows the downstream supply, it will relieve the congestion at section i. It means that a congestion that has started in a given segment may reach an uncongested equilibrium. Note however that this uncongested equilibrium has smaller outflow compared to what would be observed should that segment remains an active bottleneck. The congestion is moved upstream and become more severe. Also, either way the upstream section $i-1$ becomes congested. After queue grows to its maximum on section $i-1$, the section $i-2$ will become congested all the way to the section 1. Therefore, similar to the analysis without

capacity drop, the congestion propagates until the very first section, but not necessarily all sections will remain congested.

Lemma 4.18. *An initially uncongested freeway with* $\widehat{d^i} > C^{i+1}$ *and* $\widehat{d^k} \leq C^{k+1}$ *for* $i \neq k$ *reaches an equilibrium with active bottleneck* j *(*$1 \leq j \leq i$*)) with* $\lambda^i = \lambda^{i,max}$ *for* $i \leq j$ *and* $\lambda^i = 0$ *for* $i > j$ *and* $q^i = q^i(j, C^{j+1}(1 - \Delta_{j+1}) - u^j)\ \forall i$.

Proof. Starting with $j = i$ and set $j = j - 1$ if Lemma 4.17 applies and section j becomes uncongested. Applying recursively to $j - 1$,$j - 2$..,1, j will be one in which congestion was triggered and $q^k(j, q^j) < C^{k+1}(1 - \Delta_{k+1}) - u^k$ for $k < j$. □

Therefore, an equilibrium will be reached, but it depends on the bottleneck interactions which sections will be eventually congested. Observe that all previous derivation was related to the triggering of the capacity drop and a reduction of the bottleneck flow from capacity to dropped capacity. The increase of the queues is much faster compared to without capacity drop case and if the inflow does not change, the congestion reaches the first section also faster.

In the case without capacity drop, the bottleneck with i subject $\widehat{d^i} = C^{i+1}$ can reach equilibrium with $q^i = C^{i+1} - u^i$ with queues that can vary from zero to all upstream sections congested. However, when capacity drop is present this equilibrium only exists for $\lambda^i = 0$. If a perturbation happens that leads to $\lambda^i > 0$, λ^i will start to grow at rate of $C^{i+1}\Delta_{i+1}$ and the outflow becomes different. Take into account this difference, we classify bottlenecks into:

Every bottleneck in equilibrium can be classified into one of them and different types can coexist in a corridor similar to the without capacity drop case. There can be only one over-critical and it leads the congestion propagates until the first section. Downstream to the congested over-critical there might be any combination of the other types. The structure as Figure 4.3 remains exactly the same.

Classification	Condition	St wrt demand	St wrt. state
U-UC	$\hat{d}^i < C^{i+1}$ and $\lambda^i(0) = 0$	stable	stable if $\hat{d}^i < C^{i+1,-}$
U-C	$\hat{d}^i = C^{i+1}$ and $\lambda^i(0) = 0$	unstable	unstable
C-OC	$\hat{d}^i > C^{i+1,-}$ and $\lambda^i(0) > 0$	stable	stable
C-C	$\hat{d}^i = C^{i+1,-}$ and $\lambda^i(0) > 0$	stable	stable

Table 4.2: Classification, condition and stability with respect to perturbation in demand and state for the different types of equilibrium. U refers to uncongested, C refer to congested, OC to over-critical and C to critical, $C^{i+1,-} = C^{i+1}(1 - \Delta_{i+1})$. Proofs are on appendix.

The optimal equilibrium can be found with P1 (4.47). The difference in the case of capacity drop is that whenever $q^i + u^i > C^{i+1}(1 - \Delta_{i+1})$ it can only be achieved with the bottleneck being uncongested. To turn the bottleneck uncongested it is necessary to have flows under the dropped capacity until the queues are dissipated. Let $\widehat{d}^{i,min}$ the demand induced on bottleneck i by applying the minimum metering rate at all on-ramps and $\widehat{d}^{i,\star}$ the demand induced on the same bottleneck by applying the optimal metering rate $u^\star$.

Lemma 4.19. *An optimal equilibrium $y^\star$ through metering rates $\mathbf{u}^\star$ is always reachable if $\widehat{d}^{i,min} < C^{i+1}(1 - \Delta_{i+1})$ for all i such that $\widehat{d}^{i,\star} > C^{i+1}(1 - \Delta_{i+1})$.*

Proof. By applying $u^{\min}$ will turn all the sections uncongested through Lemma 4.3. Once uncongested, $\mathbf{u}^\star$ can be applied. □

The counterpart of Lemma 4.19 is that the optimal equilibrium point may not be reachable if $\widehat{d}^{i,min} \geq C^{i+1}(1-\Delta_{i+1})$ for some i. In that case, it is not possible to relieve an already formed congestion. Therefore, the reachability property, unlike the without capacity drop case, is also dependent on the initial conditions. The useful question under this situation is "what is the optimal reachable equilibrium state?". To answer that question we can formulate the problem in order to consider the initial condition.

The process to find the optimal equilibrium in that case is more complex. We assume without loss of generality that there is either zero or one active bottleneck in the optimal solution; it

is possible an equilibrium state with more than one active bottleneck but they are essentially equivalent to a single bottleneck located at the farthest downstream bottleneck as the section flows are the same. In that case we want to find what is the active bottleneck, k_0, and the flows, $\mathbf{u}$, that leads to the optimal flow given demands, $\mathbf{d}$, minimum metering rates, $\mathbf{u}^{min}$, capacities, $\mathbf{C}$ and drop ratios, $\mathbf{\Delta}$. Further, the initial condition is represented as the current most downstream congested bottleneck, labeled as v. We also assume that all sections are initially uncongested and $d^{i.min} > C^{i+1}$ for some section i, this last assumption also imply that the solution given by P_1 is not feasible.

Therefore, our problem is to choose the section k_0 that will be the active bottleneck and what is the control input $\mathbf{u}$ for that case. The problem $P_2(k_1)$ (4.49), if feasible, return the control sequence in which the section k_1 is the only one over capacity and therefore the active bottleneck. It P_2 is not feasible, there is no control sequence that turns the section k_1 congested while not leading to some downstream section become congested at the same time.

Once $P_2(k_0)$ is found to be feasible, the total flow and the optimal sequence is given by $P_3(k_0)$. The optimal flow will be the highest $y(k_0)$ for all values of k_0 feasible. In order to drive the system to that state, the sequence given by $P_2(k_0)$ is applied and once in that state, the sequence given by $P_3(k_0)$ should be applied. We first assumed that all sections are initially uncongested; the congestion may be relieved through $\mathbf{u}^{min}$. If that is the case, the uncongested equilibrium exists and the optimal is given by P_1. If the congestion cannot be relieved, the problem is equivalent to consider only the downstream sections of v with $d_0 = C_{v+1}(1 - \Delta_{v+1})$ to obtain the optimal flows.

The optimal given by $P_3(k_0)$ follows similar properties to the original problem without capacity drop. The optimal is obtained by optimizing upstream flows which means that $u_i = u^{min}$ for $i < k_1$ and the same structure of the optimal of P_1 in downstream sections with metering

rates increasing from upstream to downstream so as to flow not exceed capacity.

$$
\begin{aligned}
&P2(k_0): \max_{\mathbf{u},q_0,e} \quad e \qquad \text{s.t.} \\
&\min\{u^{i,min}, d^i\} \le u_i \le \min\{u^{i,max}, d^i\} \qquad i = 1, ..., I \\
&q^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i-1} u^k \hat{\beta}^{ki} \qquad i = 1, ..., I \\
&q^i + u^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i} u^k \hat{\beta}^{ki} \le C^{i+1} \qquad i = 1, ., k_0 - 1, k_0 + 1, .., I \\
&q^{k_0} + u^{k_0} = C^{k_0+1} \\
&q_0 = \frac{q^1}{\beta^1} \\
&e < C^{i+1} - q^i - u^i \quad i \neq k_1 \\
&e > 0
\end{aligned}
\tag{4.49}
$$

$$
\begin{aligned}
&P3(k_1): \max_{\mathbf{u},q_0} y^\star(k_1) = q_0 + \sum_{i=1} I u_i \quad e \qquad \text{s.t.} \\
&\min\{u^{i,min}, d^i\} \le u_i \le \min\{u^{i,max}, d^i\} \qquad i = 1, ..., I \\
&q^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i-1} u^k \hat{\beta}^{ki} \qquad i = 1, ..., I \\
&q^i + u^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i} u^k \hat{\beta}^{ki} \le C^{i+1}(1 - \Delta_{i+1}) \qquad i = 1, ., k_0 \\
&q^i + u^i = \hat{\beta}^{0i} q^0 + \sum_{k=1}^{i} u^k \hat{\beta}^{ki} \le C^{i+1} \qquad k_0 + 1, .., I \\
&q^{k_0} + u^{k_0} = C^{k_0+1} \\
&q_0 = \frac{q^1}{\beta^1}
\end{aligned}
\tag{4.50}
$$

4.5.2 Numerical Experiments

We use the same two bottleneck scenarios to verify the analytical results with capacity drop under a similar scenario compared to without capacity drop case. The upstream demand is $d_0 = 0.8C$ while at on-ramps $d_1 = d_2 = 0.3C$. The turning ratios at off-ramps are also the same with $\beta^1 = 0$ and $\beta^2 = 0.1$. The baseline scenario is the optimal pattern for such case which is:

$$\begin{aligned} q_0^\star &= d_0 \\ u_1^\star &= 0.2C \\ u_2^\star &= 0.1C \end{aligned} \tag{4.51}$$

Compared to this case, we show the results for relatively small perturbations in some of the pattern. The total simulation time $T = 50kseconds$. For all cases $\tilde{q}_1(40K) = q_1(40K) - 0.02C$, that is, a small perturbation occurs at $t = 40000s$ leading to a slightly smaller outflow at that time step. This setting is tested in 4 different cases:

- case (a) - the optimal pattern;
- case (b) - $u_2(t) = u_2^\star + 0.01C$ for $T/3 < t < 2T/3$ and as the optimal pattern otherwise
- case (c) - $u_1(t) = u_1^\star + 0.01C$ for $t > 2T/3$ and as the optimal pattern otherwise;
- case (d) - $u_1(t) = u^\star$ and $u_2(t) = u^\star$ but on that case the upstream demand is set to $d_0 = 0.75C$

The case (a) is the optimal pattern and would lead to the highest outflow if no disturbances happen. Cases (b) and (c) lead to demand over capacity when its respective metering rates

change. Case (d) follow the optimal pattern as (a) but observe the upstream demand is smaller. The results are shown in Figure 4.5. The pattern (a) had no queues until the disturbance at segment 1 which had triggered the capacity drop and reduced the flow. As the demand was close to capacity at the bottleneck 1 it is an unstable case and there was a sudden drop on the total outflow. The same instability is present on cases (c) and (d) and the flow drops as soon as the disturbance in u_2 (case b) and u_1 (case c) is applied. Observe the difference between case (a) and (d); the case (d) is stable with respect to small disturbances as the flow is not at capacity on the first bottleneck. Therefore, when the disturbance occurred (a slightly smaller outflow), it has just slightly increased the outflow in the previous time steps.

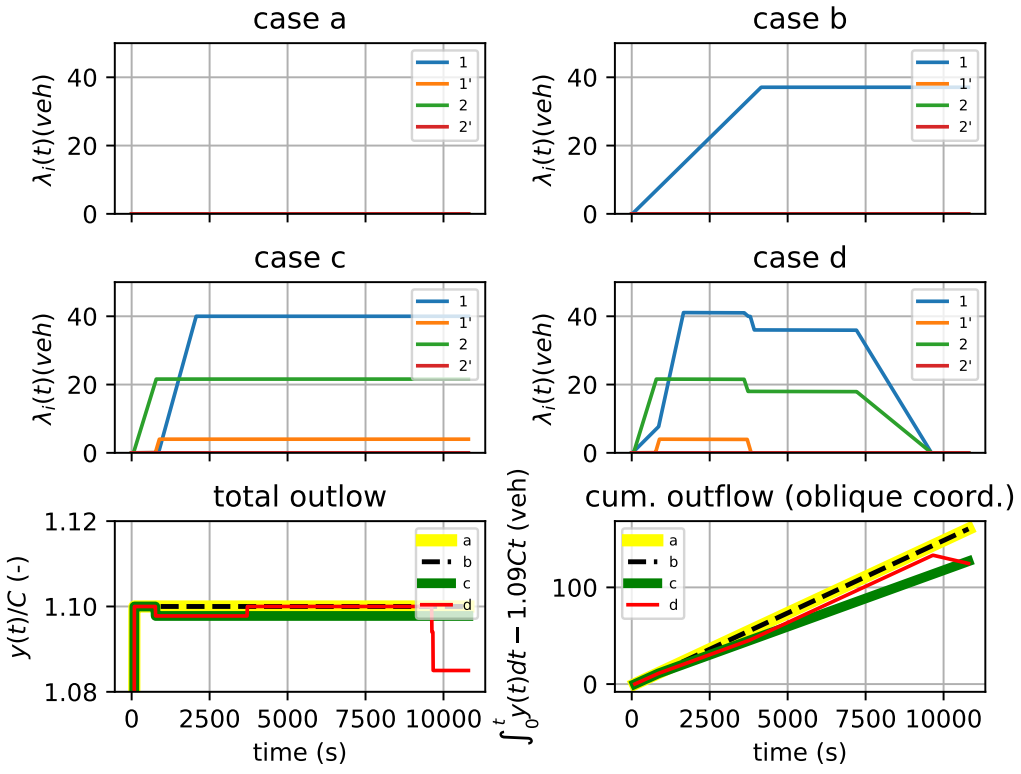


Figure 4.5: Cumulative outflow for each case considering the capacity drop.

Now we verify the reachability and stability results. All the numerical experiments so far the uncongested equilibrium existed, but we have shown that it may not be case. Also, the

reachability property is dependent on the initial state and the minimum metering rates. We show these properties by comparing a set of four similar cases. The demands are:

$$d_0 = \begin{cases} 0.85C & t < T/3 \\ 0.94C & T/3 \leq t \leq 2T/3 \\ 0.7C & t > 2T/3 \end{cases} \tag{4.52}$$

$$d_1 = 0.2C$$

$$d_2 = 0.2C$$

Similarly to the previous experiments, $\Delta_1 = \Delta_2 = \Delta = 0.1$, $\beta^1 = 0$, $\beta^2 = 0.04$. The cases differ on the following:

- case (a) - $u^{i,min} = 0.08C$ with all sections initially uncongested. Optimal policy applied considering the capacity drop phenomenon (P_2 and P_3)
- case (b) - $u^{i,min} = 0.08C$ with section 2 initially congested. Optimal policy applied considering the capacity drop phenomenon (P_2 and P_3);
- case (c) - $u^{i,min} = 0.04C$ with section 2 initially congested. Optimal policy applied considering the capacity drop phenomenon (P_2 and P_3);
- case (d) - $u^{i,min} = 0.08C$ with all sections initially uncongested. Optimal policy applied without considering the capacity drop phenomenon (P_1)

The results are shown in Figure 4.6. In case (a) the metering rates were set so as to yield capacity flow on sections 1 and 2 for the first upstream demand pattern. When the upstream demand slightly increases, it is not possible to keep both sections congested and the optimal

policy is to keep the first bottleneck congested while leaving the second bottleneck uncongested. Observe that by doing so it is possible to increase the metering rates of the second on-ramp. In case (b) the section 2 is initially congested and it is not possible to relieve the congestion with the minimum metering rates. In this experiment there was nothing the controller could do to avoid the congestion. However, observe the case (c) where it starts from similar situation, but the congestion is relieved as the minimum metering rates is similar. Following that, it follows similar pattern as case (a). Case (d) neglects the capacity drop. That was not harmful while it was possible to keep both sections uncongested, but it lead to capacity drop being triggered at section 2 and to a smaller outflow overall compared to cases (a) and (c).

One important aspect of this case, though simple, is that the capacity drop being forced at section 1 did not significantly change the outflow. In this specific case, the smaller outflow on the first section could be compensated by a higher metering rates at the following section. The impact in this case is a smaller outflow through the second off-ramp which is in the order of $\Delta_1 \beta^2 \approx 0.4\%$. This suggests that the impact of capacity drop can be significantly undermined even if its occurrence is unavoidable.

4.6 Conclusions

An analysis of equilibrium states on freeways based on the link transmission model combined with a capacity drop was presented. Properties of the equilibrium states was derived from the model and the impacts of the capacity drop phenomenon was highlighted by comparing with no capacity drop case. Also, we presented the optimization problems to obtain optimal equilibrium states.

Compared to a previous study [42], the capacity drop changes on the following. First, the

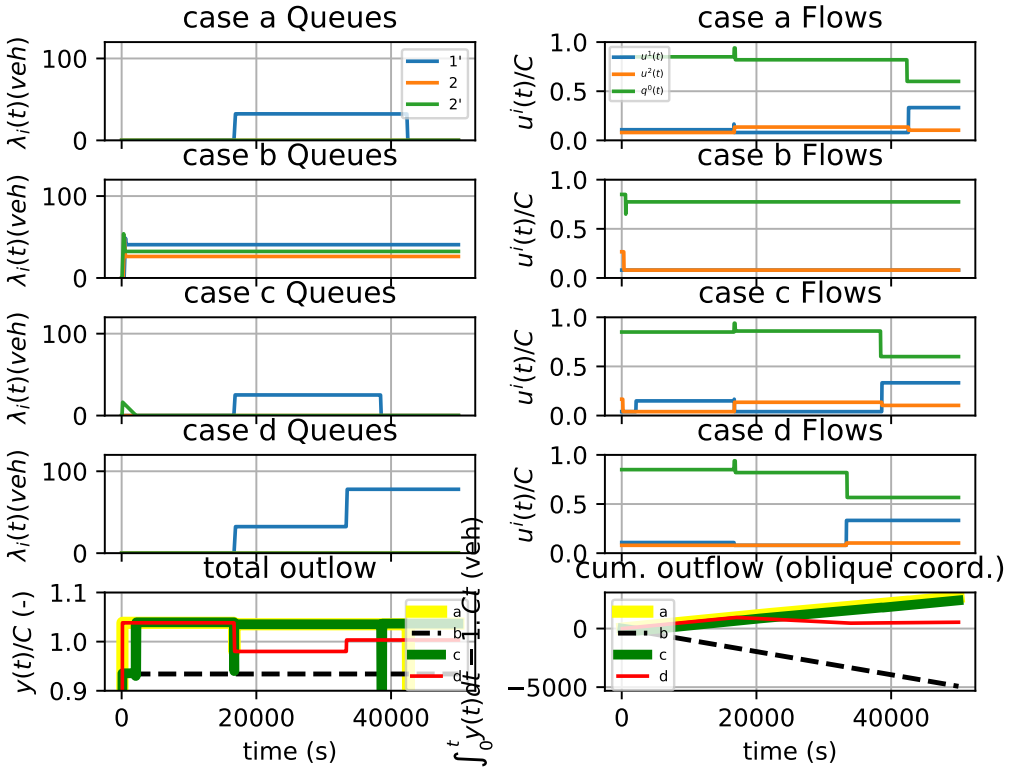


Figure 4.6: Queues and metering rates for each of the four cases. At bottom left the total outflow and bottom right the oblique cumulative curves.

active bottleneck flows are not unique. If there is no capacity drop, the active bottleneck flows are unique - at capacity - regardless of the queues. Second, the output is unstable with respect to disturbance in the states and control if the flow at some bottleneck is close to capacity. It is different compared to without capacity drop case in which it is always stable. Third, the optimal equilibrium state may not be reachable depending on the upstream demand and initial state.

We presented two optimization problems for obtaining optimal equilibrium states. As uncongested equilibrium states may not be reachable, the goal turns to find the best reachable state given the initial state. This is particular interesting because it is often the case in several freeways especially in peak hours. Even if possible to keep uncongested with small metering rates, the queues on on-ramps grows quickly and later the metering rates must be

increased due to the queue override rule. On that situation, it is possible to find the optimal given the current operational situation. On the two bottlenecks scenario, it was possible to achieve an outflow just 0.5% smaller than the uncongested equilibrium.

On future studies we will be interested on the equilibrium states based on origin-destination demands. With the emergence of connected vehicles it may not be a strong assumption the knowledge of origin-destination matrix. Also, the problem of reachability is due to the uncontrolled upstream demand. We would like to study the effect of variable speed limits on this scenario to confirm whether with both types of control an optimal equilibrium state is always reachable.

Chapter 5

Reachability and Stability of A Local Ramp Metering System

So they push me then from side to side.
They're pushing 'til there's nothing
more to hear

Helloween (I Want Out)

5.1 Introduction

Severe traffic congestion usually occurs during peak hours and causes delays, higher fuel consumption, and air pollution [117]. Ramp metering is one of the possible techniques to improve freeway performance. Significant reductions in total travel time have been observed in field deployments (e.g, [108],[84]) and simulation studies (e.g., [55], [124], [110]). Reducing congestion can also provide safety benefits [39]. There are two mechanisms that cause a reduction in total travel time.

The first is related to the queue spill back mechanism. When a queue starts from a bottleneck, propagates upstream and reaches an upstream off-ramp, vehicles leaving at that off-ramp are delayed due to congestion ahead [109]. These vehicles leaving on that off-ramp do not go through the bottleneck, but are nonetheless impacted. Holding vehicles on on-ramps might prevent, or at least postpone, the queue from reaching that off-ramp and consequently reducing the total travel time.

The second is the capacity drop phenomenon, a drop in the discharge flow rate of a merge bottleneck when its upstream section gets congested [17] while its downstream section is uncongested. The magnitude of the drop varies and a typical value is around 10% [24]. As the flow rate is lower, the bottleneck takes longer to discharge the same amount of vehicles, thereby increasing the total travel time. Again, storing the vehicles on on-ramps might prevent or postpone the onset of the congestion and therefore keep longer discharging at capacity.

Several field deployments have confirmed the benefits of ramp metering. An evocative case is that of the Twin Cities in Minnesota in which the meters were turned off for a period of two weeks in order to assess the impacts of ramp metering [84] on several metrics. On the four freeways studied, the travel time were smaller when the meters were operating, ranging from around 2% to as high as 36% ([84], Table 3). Significant benefits were also reported in [108, 31].

The benefits on these cases are a combination of avoiding queue spill back and capacity drop phenomenon; however, it is not an easy task to quantify the share of each mechanism on the overall improvement [107]. This difficult to breakdown the effects might explain why few empirical studies have related the relationship between ramp metering and capacity drop. An important study is reported in [18] that shows that through a more restrictive metering rate, it is possible to recover the discharge flow rate to capacity on a isolated merge bottleneck.

Simulation studies have considered the effect of the capacity drop implicitly or explicitly. Using second order models, in [119] it was shown that variations of ALINEA were able to sustain a higher outflow for local control. Similar models have been used for coordinated control as in [110, 75]. More recently, capacity drop was taken explicitly and integrated in model based controllers (see [49, 92]).

However, no study has combined a model in order to analytically study the effect of capacity drop on the system dynamics controlled by an on-ramp meter. Through these studies, system properties can be established in closed form solution helping us to understand important features of the system. This ultimately can be used to specify requirements to warranty a meter and parameter tuning of established algorithm. A better understanding may also give insights into the design of new algorithms.

Nonetheless, there are relevant and recent analytical studies on ramp metering. An interesting study on closed-loop ramp metering and its operating regime is reported in [40]. Using cell transmission models [28] with few cells, it was shown that different ramp metering algorithms can be analyzed from operation "modes" and its transitions. It was established controllability and observability with respect to detector placement. The analysis, the authors claim, suggests that ALINEA [91] is a superior strategy compared to %-Occ [1].

The set of equilibrium states and their characteristics in a single freeway was studied in [42]. It was shown that all equilibrium states leads to the same flow rate on the bottlenecks, but keeping those bottlenecks uncongested is beneficial as it diminishes the aforementioned queue spill back effect. They show that through ramp metering, it is possible to steer the system to an uncongested equilibrium state and therefore reducing delays.

Closed loop stability for ramp metering also has been the subject of recent research, in particular on ALINEA and its variations. Through linearization and Lyapunov theory, stability is established for PI-ALINEA in [130]. For PI-Controllers and a class of systems that local

ramp metering fits in, stability was also derived in [69]. The aforementioned study [40] also establishes stability range for ALINEA.

All of these studies have not considered the capacity drop. Its understanding is an essential step in order to analyze and design ramp metering algorithms especially for local ramp metering control. We attempt to fill some of the gaps by analytically studying essential open and closed loop (with PI-ALINEA control law) properties. This is enabled by using models that are simple and yet capable of reproducing essential traffic flow characteristics. A link queue model [58] is used for the traffic dynamics inside the merging segment. This model is an approximation of the LWR model and extends the cell demand and supply [28, 80] functions to a link. Second, a simple model is incorporated to replicate the outcome of the capacity drop phenomenon at a merge bottleneck [66], that is, a decrease in flow after following the onset of congestion. The combined model leads to a switched linear ordinary differential equation [85]. We were able to establish the following:

1. the system has hysterical nature with respect to the demand pattern. Demand higher than the capacity triggers the congestion; however, in order to clear a formed congestion, it is necessary a demand lower than the current discharge flow rate, which is lower than the capacity;

2. depending on the amount of the capacity drop, it might not be possible for a local ramp controller to eliminate the congestion. In this situation, while the meter can at some extent change the proportion of the delays on on-ramp and mainline freeway, it is not able to clear the congestion; and

3. if the controller is able to effectively eliminate the congestion, the stability region of the widely applied and studied (PI-)ALINEA algorithm is derived.

The rest of the paper is organized as follows. In Section 2 the system's model is presented. In Section 3 the impact of the capacity drop on performance is presented. In Section 4 we show

when ramp metering is effective by analyzing the system's equilibrium states, its transitions, and reachability. In Section 5, the closed-loop system are analyzed and it was shown the range of parameters in which conditions PI-ALINEA is stable and reach an optimal state, uncongested and discharging at capacity. In Section 6, the controller stability region based on Link Queue Model is compared to its counterpart in using cell transmission model. Finally, Section 7 has the conclusions and future work.

5.2 System Description and Model

The system under study is depicted in Figure 5.1 referred as a merge bottleneck containing three components: a merge of the on-ramp and freeway streams at $x = 0$, a bottleneck (lane drop) located downstream, at $x = L$, where the on-ramp acceleration lane ends, and the merging segment is the area between them. On the upstream boundary and on-ramp any unserved traffic is modeled as point queues [63].

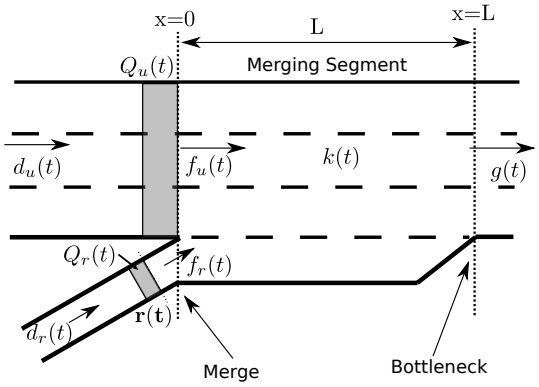


Figure 5.1: Schematic of merge bottleneck and the model variables.

The state variable is the average density on the merge bottleneck, $k(t)$. Longitudinal and lateral variations inside this zone are ignored. The inputs demands are the mainline demand, $d_u(t)$, and the on-ramp demand, $d_r(t)$. Total demand is denoted by $d(t) = d_u(t) + d_r(t)$.

The demand can be limited by the metering rate, $r(t)$. In addition, $f_u(t)$ is the mainline

Symbol	Description	units
$d_u(t)$, $d_r(t)$, $d(t)$, $\hat{d}(t)$	Upstream, ramp, total, controlled demand	veh/s
$f_u(t)$, $f_r(t)$, $f(t)$, $g(t)$	Upstream, ramp, total in-flux, out-flux	veh/s
D, S	Demand, supply	veh/s
Q_u, Q_r	Upstream, on-ramp queues	veh
$r(t)$, r_{min}	current, minimum metering rate	veh/s
$k(t)$, k_c,k_j, k_o	Current, critical, jam, target density	veh/m
C (C_l), k_{cd}	Downstream (per lane) capacity	veh/s
k_{cd}	Critical density downstream	veh/m
Δ	Relative drop amount	veh/s
k_1, $\bar{k}$	congested and uncongested for $g = C(1-\Delta)$	veh/m
v_f, ω	free flow, shock-wave speed	m/s
K_p	Proportional gain of PI controller	m/s
K_i	Integral gain of PI controller	m/s^2
$x(t)$	Excess density $(k(t) - k_{cd})$	veh/m
$v(t)$	Excess metering rate	veh/s
L	Merge bottleneck length	m

Table 5.1: Notation

.

in-flux, $f_r(t)$ the on-ramp in-flux, and $g(t)$ the out-flux. Table 5.1 summarizes the notation used.

The traffic dynamics inside the merging segment can be described by the Lighthill-Whitham-Richards (LWR) model [86, 113], which has been successfully applied to analyze the initialization, propagation, and dissipation of traffic congestion with spatial and temporal density waves (kinematic waves). However, the LWR model is a partial differential equation, more specifically a hyperbolic conservation law for which the control problem is not well studied. In this study, we resort to an approximation, the Link-Queue Model (LQM) [58], which only considers dynamical variations of spatially average densities and is therefore an ordinary differential equation. In [54], this model has been successfully applied to analyze and design the variable speed limit strategy and the results are validated in the LWR model through Cell Transmission Model simulation. Thus we follow the same approach by studying the

control of merge bottlenecks with the LQM:

$$\dot{k}(t) = \frac{1}{L}(f(t) - g(t)), \tag{5.1}$$

where $f(t)$ and $g(t)$ are the in and out-fluxes. Equation (5.1) can be viewed as a reservoir in which level increases or decreases based on the in- and out-fluxes difference. The fluxes are computed based on demand and supply concepts [28, 80]:

$$\begin{aligned} D(t) &= \min\{v_f k(t), v_f k_c\} \\ S(t) &= \min\{v_f k_c, \omega(k_j - k(t))\}, \end{aligned} \tag{5.2}$$

which are respectively the increasing and decreasing parts of the triangular fundamental diagram [100]:

$$q(k) = \min\{v_f k, \omega(k_j - k)\}, \tag{5.3}$$

where $q(k)$ is the flow-rate, v_f is the free flow speed, k_j the jam density, and ω the shock-wave speed. The density which yields maximum flow is $k_c = \frac{\omega k_j}{v_f + \omega}$; at this point the flow is the capacity, $C = v_f k_c$. Ramp, acceleration lane, and freeway lanes share the same characteristics, as v_f and ω. After the lane drop it is assumed the same per-lane fundamental diagram holds. The per-lane capacity is denoted by C_l. Also, hereafter C refers to downstream capacity unless stated otherwise. The density that yields capacity is $k = k_{cd}$. Figure 5.2 presents the fundamental diagram at each segment. In addition, the on-ramp lane has a capacity $C_r = C_l$.

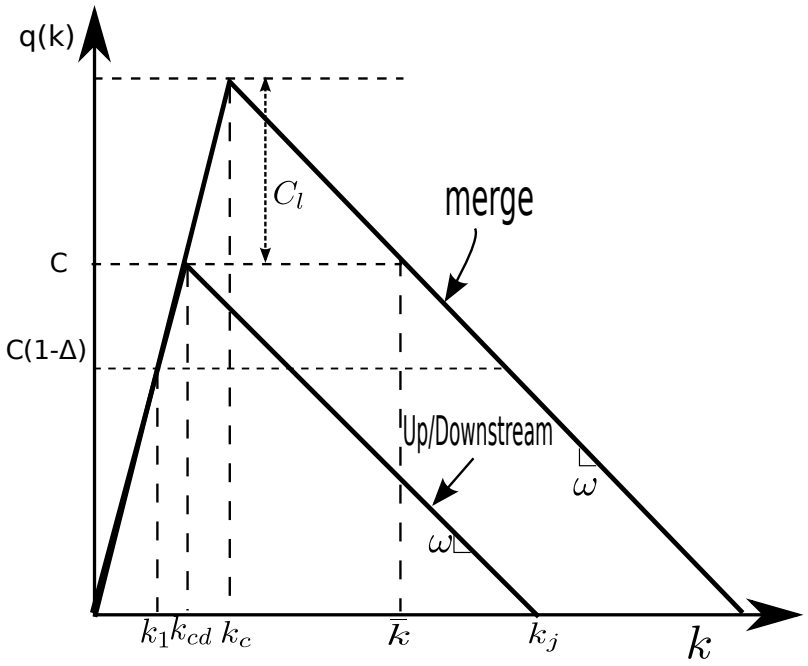


Figure 5.2: Fundamental diagram of upstream/downstream and merging segments.

On the upstream and on-ramp unserved vehicle are modeled as point queues:

$$\begin{aligned} \dot{Q}_u(t) &= d_u(t) - f_u(t) \\ \dot{Q}_r(t) &= d_r(t) - f_r(t), \end{aligned} \tag{5.4}$$

which demands are computed [63]:

$$\begin{aligned} D_u(t) &= \min\{Q_u(t) + d_u(t), v_f k_c\} \\ D_r(t) &= \min\{Q_r(t) + d_r(t), r(t), C_r\}, \end{aligned} \tag{5.5}$$

note that the upstream demand can be limited by the metering rate.

With upstream demands fluxes are computed. It is assumed absolute priority for the on-

ramp:

$$f_r(t) = \min\{D_r(t), S(t)\}, \tag{5.6}$$

and the remaining supply can serve the upstream demand:

$$f_u(t) = \min\{D_u(t), S(t) - f_r(t)\}, \tag{5.7}$$

and the total in-flux is $f(t) = f_u(t) + f_r(t)$. We also denote as $D_m(t) = D_u(t) + D_r(t)$ as the total demand on the merge. Note that $f(t) = \min\{D_m(t), S(t)\}$.

At the downstream boundary of the merging segment, the out-flux is determined by:

$$g(t) = \min\{D(t), C(1 - \Delta H(k(t) - k_{cd}))\}, \tag{5.8}$$

where $H(x)$ is the Heaviside function:

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{5.9}$$

and Δ is the capacity drop ratio.

It is assumed that there is no congestion on the downstream mainline freeway, i.e., the merge bottleneck is active. Note that the capacity drop model proposed in [66] is used here to replicate the capacity drop phenomenon: when there is no queue on the merging segment, the out-flux can reach the downstream capacity, but if a queue forms, the out-flux

is the dropped capacity $g(t) = C(1 - \Delta)$. An important aspect is that the drop ratio, Δ, is exogenous and should be determined for each case.

In the following two sections we use this model to show the impact of the capacity drop and the conditions in which ramp metering is effective.

5.3 Why Ramp Metering? The Impact of the Capacity Drop on The Delay

From the presented model, it is possible to assess the impact of capacity drop by comparing the drop ($\Delta > 0$) and no-drop case ($\Delta = 0$). It is assumed that any transitory period is small compared to the total time considered. There are two cases. One discharging at capacity, C, and another at dropped capacity, $C(1 - \Delta)$. The total demand, $d = d_u + d_r = \alpha C$ is assumed to be over capacity (i.e., $\alpha > 1$) by $t = T$ and zero thereafter.

In Figure 5.3 cumulative curves $N(t)$ are depicted. The continuous line is the cumulative arrival. The dashed lines are the departure rates for the case in which it discharges at capacity, C and the case discharging at dropped capacity, $C(1 - \Delta)$. The vertical difference between the arrival and departure curve is the instantaneous queue. This queue could be at on-ramp, mainline, or both depending on the upstream and ramp demand and the metering rates.

The area between the arrival and departure curve are the total delay:

$$\begin{aligned} D_{nd} &= \frac{1}{2} C T^2 \big[(\alpha - 1)\alpha\big] \\ D_d &= \frac{1}{2} C T^2 \left[\frac{(\alpha + \Delta - 1)\alpha}{1 - \Delta}\right] \end{aligned} \tag{5.10}$$

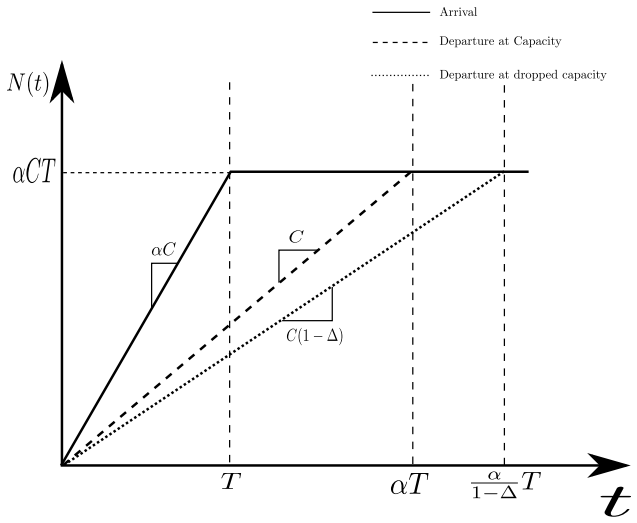


Figure 5.3: Arrival (continuous) and departure (dashed) cumulative curves. The outflow (departure) could be either at capacity, C, or at dropped capacity, $C(1-\Delta)$.

The relative improvement of avoiding the capacity drop is given by:

$$D(\%) = 1 - \frac{D_{nd}}{D_d} = \frac{\alpha\Delta}{\alpha + \Delta - 1} \tag{5.11}$$

For example, if $\alpha = 1.1$ (that is, demand 10 % higher than capacity) and $\Delta = 0.05$, the improvement is 36%; if the drop amount $\Delta = 0.1$, the difference goes to 55%. Therefore, a well designed ramp meter can drastically decrease the delay. The question turns to which conditions should be satisfied to achieve such reduction.

5.4 When Ramp Metering is effective? The Equilibrium States and Reachability Property

We consider the meter effective when it is able to reduce the total delay. As it is assumed that the metering rate cannot change the demand, the delay will be lower as the outflow

increases. Therefore, in equilibrium it is desired to discharge at capacity when there is enough demand. When the demand is below capacity, the system should be able to remain uncongested and discharge all vehicles with no delay.

First, we show the equilibrium states and its characteristics. It is shown that keeping at uncongested equilibrium states is beneficial as the outflow is always higher. Then, we show in which conditions it is possible to lead the system to the uncongested equilibrium state.

5.4.1 The equilibrium states and their Behavior

We analyze the equilibrium states of the systems subject to constant metering rate ($r = C_r$ for no control case) and ignoring the on-ramp and mainline queues. In this case the total demand is constant:

$$D_m = \hat{d} = d_u + \min(r, dr). \tag{5.12}$$

given an initial state and constant demands. The system reaches equilibrium states classified as follows:

State 1 If $\hat{d} < C(1-\Delta)$, the system reaches an uncongested equilibrium density $k_{eq} = \frac{\hat{d}}{v_f} < k_1$ from any initial state.

State 2 If $C(1-\Delta) \leq \hat{d} \leq C$, the system reaches an uncongested equilibrium density $k_{eq} = \frac{\hat{d}}{v_f} \in [k_1, k_{cd}]$ from an initial state $k(0) \leq k_{cd}$.

State 3 If $C(1-\Delta) \leq \hat{d} \leq C$, the system reaches a congested equilibrium density $k_{eq} = \bar{k}$ from an initial state $k(0) > k_{cd}$.

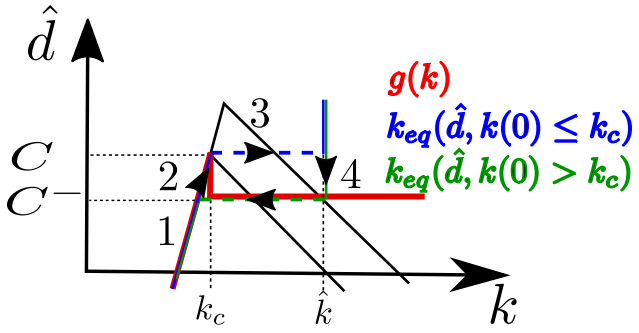


Figure 5.4: Transition in equilibrium states subject to the change in the demand level. Uncongested states are 1 and 2; congested states are 3 and 4.

State 4 If $\hat{d} > C$, the system reaches a congested equilibrium density $k_{eq} = \bar{k}$ from any initial state.

Even though the model is based on continuous variables, its essential operating regimes and its transitions can be characterized by a finite state machine, depicted in Figure 5.4. The system shifts to state 1 whenever $\hat{d} < C(1-\Delta)$ it if is initially either at state 2 or 3. Similarly, reaches state 4 when $\hat{d} > C$ from states 2 or 3. However, it shifts to state 2 when $C(1-\Delta) \leq \hat{d} \leq C$ and initially at state 1. Likewise, reaches state 3 for the same demand level, but initially on state 4.

A complete cycle can be done clockwise, but it is not possible on the other way around. This fact shows the inherent hysteresis: when the system is initially at state 2, it is necessary a demand greater than capacity ($\hat{d} > C$) to reach state 3; however, it is necessary demands lower than the dropped capacity (i.e., $\hat{d} < C(1-\Delta)$) to return to state 1 again and then a demand to $C(1-\Delta) \leq \hat{d} \leq C$ to settle at state 2. It is not possible to switch between state 3 and 2 without state 1 as intermediate.

It is clear that keeping in state 2 has advantages over state 3. In state 2 yields higher out-flux while keep the bottleneck uncongested. However, to shift state 3 to 2 is not straightforward. It needs a sharp reduction on the demand to shift to state 1 and then the demand can be increased to levels higher than $C(1-\Delta)$ again which then the out-flux will be higher and

the merging segment uncongested.

The hysteretical nature of transportation networks has been discussed with empirical evidence in [36]. However, the difference is that on that study the hysteresis is an outcome of the queue spill back mechanism while in this study it is an outcome of the capacity drop phenomenon. The queue spill back mechanism reduces the out-flux of upstream links when the congestion propagates until of an upstream boundary . The capacity drop reduces the flow to downstream links when its downstream boundary is congested.

Equilibrium States Classification

The equilibrium state can be characterized over different aspects regarding its equilibrium. We analyze for convergence and stability.

Under constant demand, the system is convergent [42]: given constant demand d, it always converges to one equilibrium state, either the $k_{eq,u} = \frac{\hat{d}}{v_f}$ or congested $k_{eq,c} = \bar{k}$. Any density in the interval $(k_{cd}, \bar{k})$ is an unstable equilibrium state for $\hat{d} = C(1 - \Delta)$.

For stability, we analyze based on Lyapunov stability [6] in which an equilibrium state is stable if the initial condition is close to an equilibrium, it will remain close to this equilibrium. This analysis could be with respect to density and the demand.

With respect to demand level, there are two cases in which it fails. It initially at critical density and $\hat{d} = C$, a demand $\hat{d} + \gamma$ where γ is small and greater than zero will lead the system to the congested equilibrium $\bar{k}$. Likewise, if $k(0) > k_{cd}$ and $\hat{d} = C(1 - \Delta)$; a demand $\hat{d} - \gamma$ will lead the system to $k = \frac{\hat{d}-\gamma}{v_f} << \bar{k}$. Both cases, with a small perturbation on demand, the system settle far away from its equilibrium. In this sense, it can be classified as bistable: the system has two distinct equilibrium points depending on the sign of the perturbation on the demand.

The same is true for a small perturbation on the density. If $C(1-\Delta) \leq \hat{d} \leq C$ and $k = k_{cd}$. A small perturbation positive on density leads to $\bar{k}$ while a negative leads to $k_{eq,u} = \frac{d}{v_f}$.

Figure 5.5 depicts the bifurcation diagram considering d as bifurcation parameter [136]. Continuous line represents stable equilibrium and dashed lines unstable equilibrium. Bifurcations have been discussed for traffic network in [29] and [60]. The existence of multiple stationary states in a network in [62] and multiple equilibrium states in a single freeway as in [42] implies bifurcations. However, all of them the underlying principle is a queue spill back effect which means that flows reduces on upstream links or sections due to a congestion downstream. In this case, the possibility of multiple equilibrium states arise in a single merge and affects also the downstream flow.

As mentioned and it also can be seen on the bifurcation diagram that $k(0) > k_{cd}$ and $C(1-\Delta)$ is an unstable equilibrium state. The fact it is unstable it does not change the operation regimes in 5.4 as the stable equilibrium states are the ones likelier to be observed in practice [29] and therefore more important to be studied.

5.4.2 Reachability with dynamic metering rates

As we have shown, for demands between the capacity and the dropped capacity, the system can be either congested or uncongested. In order to shift it is needed a demand lower than the dropped capacity for a sustained period. As the metering rate can limit the on-ramp demand, we ask the following question: in which condition the ramp meter is able to avoid the congestion? If initially congested, in which condition is it possible to dissipate the congestion?

We use the terminology of control theory in which reachability is the capability to reach an arbitrary state [1] through any function $r(t)$ [6] that satisfies the constraints (this case

[1] The term controllability and reachability are often exchangeable depending on the textbook [6], here we

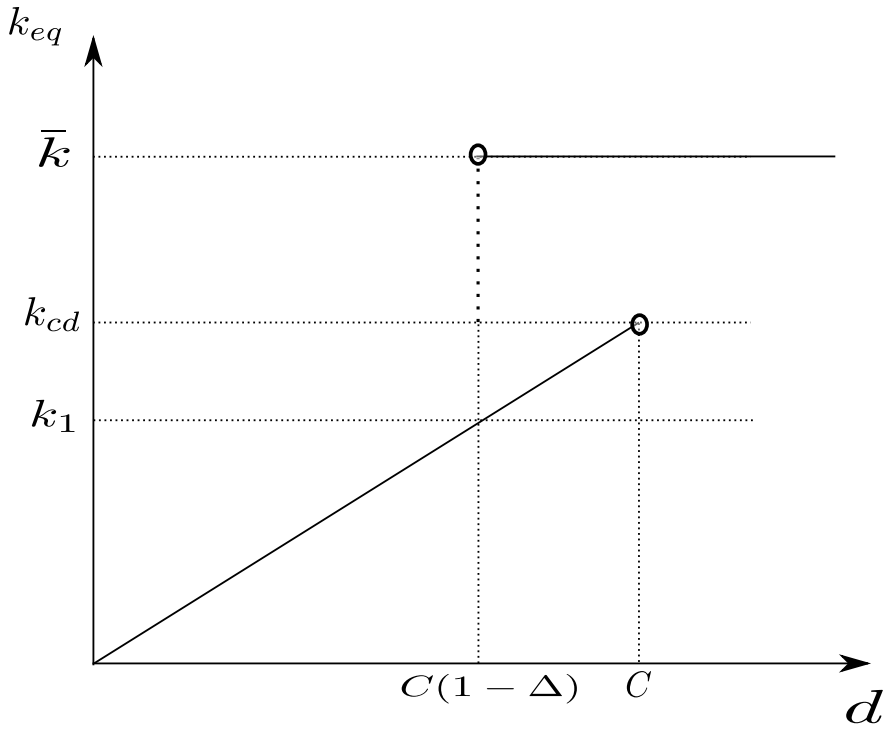


Figure 5.5: Bifurcation diagram: the set of equilibrium states for varying demands.

$r_{min} \leq r(t) \leq C_r$). When there exists at least one $r(t)$ that satisfies this condition, the state it is reachable. For unconstrained linear systems, a general test is often used [68]. However, the system under analysis is switched and $r(t)$ constrained.

In this case, the goal is to keep the system uncongested and therefore discharging at capacity. Then, the test consists in keeping $k(t) \leq k_{cd}$. Let Z the set of points such $k(t) \leq k_{cd}$, then Z is reachable if the controller is able to lead the system to at least one element of Z. We assume constant demands, either upstream, $d_u(t) = d_u$, or ramp, $d_r = (t) = d_r$.

Theorem 5.1. *Z is reachable for $d_u < C(1-\Delta) - r_{min}$ for any initial condition.*

Proof. If $k(t_0) > k_1$, $g(t) = C(1-\Delta)$. Setting $r(t) = r_{min}$, $f(t) = \min\{v_f k_c, \omega(k_j - k(t), d_u + r_{min}\}$. As long as $d_u < C(1-\Delta) - r_{min}$, $f(t) < C(1-\Delta)$ and $\dot{k(t)} < 0$ and eventually, at $t = t_1$, $k \leq k_{cd}$. Once it is uncongested, either initially or at $t = t_1$, it will remain so as

follow the definition as in [126]: controllability is related to reaching the origin and reachability is related to an arbitrary state.

long as $r(t) \leq C - d_u$ because $\dot{k(t)} = \frac{1}{L}(f(t) - \min(C, v_f k(t))$ and $g(t) \geq f(t)$ at boundary, $k = k_{cd}$, and $\dot{k(t)} \leq 0$ so $k(t) \leq k_{cd}$ for $t > t_1$. □

Theorem 5.2. *Z is reachable for $d_u < C - r_{min}$ if $k(t_0) < k_{cd}$*

Proof. With $r(t) = C - r_{min}$, follows the same condition of Theorem 5.1 for $t > t_1$. □

Outside this region, the controller is no longer effective. For example, if the system is in State 3 and $d_u > C - (1 - \Delta) - r_{min}$, even with $r(t) = r_{min}$ it does not switch to any of the uncongested states (1 or 2). In this case, a drop in the upstream demand, d_u, is necessary to relieve the congestion.

Though still able to control $k(t)$, there might exist unserved demand. In order to keep at full capacity, the metering rate can be set to levels lower than ramp demand forming a queue that may spillover to local streets.

Theorem 5.3. *For Z reachable, all demand is served for $d < C$.*

Proof. If Z is reachable implies that eventually $k(t) \leq k_{cd}$ and it can operate at capacity. After this instant, $r(t) = C - d_u(t)$ can be set. The maximum influx is $f(t) = d_u + \min(C - d_u, d_r)$ and therefore $f(t) \leq d_u(t) + C - d_u$, thus $f(t) \leq C$. If $d = d_u + d_r < C$, then $d_r < C - d_u$, so $r(t) \geq d_r$ and both upstream and ramp demand are served. □

When demand exceeds capacity, queues will grow either on on-ramp or upstream. In case $d_u + d_r > C$, setting the metering rate as $r = C - d_u$ the upstream and ramp flux would be $f_u = d_u$ and $f_r = r$ respectively. The on-ramp queue would evolve as:

$$\dot{Q}_r = d_r(t) - f_r(t) = d_r - r = d_u + d_r - C \geq 0, \tag{5.13}$$

The queue would steadily increase. In practice, this queue has a maximum length in order to avoid the congestion to spill over to local streets. Often, the meter has a queue override feature that forces a higher metering rate to avoid long queues on on-ramps (see Equation (5.11)). It is not considered explicitly in this study. However, at this point either delay will increase on local streets, due to queue spill back, or at mainline due to the capacity drop.

When reachability is not guaranteed, the control system is able, at some extent, to change the share of the delays on on-ramp or mainline freeway, but it will discharge at dropped capacity.

This result, while in this case for a single merge, differs from [42]. On that study, it was shown that there could be multiple equilibrium states for a bottleneck with demand larger than capacity. As the capacity drop phenomenon was not considered, the resulting flow rate at the bottleneck is unique and always at capacity. Considering the capacity drop phenomenon, the flow rate is lower when the bottleneck is congested.

Also on that study, it was proven that it is possible to *steer* the system towards the uncongested equilibrium state. It is a similar to what we defined on this study for reachability. One of the differences is a minimum metering rate introduced in this study. Nonetheless, even for $r = r_{min} = 0$ if $d_u > C(1 - \Delta)$ it is not possible to dissipate the congestion.

On the case of coordinated control, d_u is function of metering rates on upstream on-ramps on previous time. Clearly, if $r_{min} = 0$ on all on-ramps it is possible to induce a $d_u = 0 < C(1 - \Delta)$. However, if the flow induced by $r_{min} \geq 0$ on all upstream on-ramps lies in the interval $(C(1 - \Delta), C)$ it is possible to avoid the capacity drop, but not recover from it.

This also shows the impact of the minimum metering rate. A higher minimum metering rates can make Z not reachable. For this purpose, ideally $r_{min} = 0$; however, usually agencies might impose higher minimum metering rates due to other operational issues, such as Caltrans in California [123].

5.5 How lead the system to the desired state? Closed-Loop Analysis

The model equations are combined with PI-ALINEA [132] in order to analyze the response in closed-loop. First, the PI-ALINEA algorithm is briefly described. Second, the choice of set-point, k_o, and the equilibrium states are discussed. Then, we show for which parameters, K_p and K_i, the system in closed loop is stable. A Poincaré map analysis is presented for the case which the response is oscillatory.

5.5.1 PI-ALINEA

ALINEA [91] is a feedback control algorithm based on PID Controller family. The metering rate is updated based on the observed occupancy close to the lane-drop location. While the traditional ALINEA [91] is an I-controller, in this study we consider the extended PI-ALINEA [132], which it is used a PI-Controller rather than an I-Controller. Also, the ALINEA control law is considered in discrete time. In this study we consider the continuous PI-Controller, given by [6]:

$$r(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau, \tag{5.14}$$

where K_p and K_i are the proportional and integral coefficients respectively, and the error, $e(t)$, is the difference between the real-time density $k(t)$ and the target density $k_o(t)$:

$$e(t) = k_o(t) - k(t). \tag{5.15}$$

In addition, the control signal $r(t)$ is bounded:

$$r_{min} \leq r(t) \leq C_r. \tag{5.16}$$

Thus, it is necessary to determine the following parameters: the coefficients K_p and K_i, the target density $k_o(t)$, and the minimum metering rate r_{min}.

5.5.2 Set-Point Specification and Equilibrium States

From the analysis of equilibrium states, the fundamental diagram, and the PI-ALINEA control law (Eq. 6.8), we can find the optimal set point from the following observations:

1. if $d > C$ the maximum out-flux is when $k(t) = k_{cd}$ with out-flux $g(t) = C$;
2. if $d \leq C$ and $k(t) \geq k_{cd}$ the system could have been operating with the same out-flux, but at free-flow speed ($k \leq k_{cd}$). In this case any set point less or equal to k_{cd} will force the freeway to operate at free-flow speed; and
3. if $d \leq C$ and $k(t) \leq k_{cd}$ the system will not operate at capacity; however, any control action in the direction of a higher flux (i.e., increase of metering rate) is always desirable or at least does not affect the performance. The integral effect will push the metering rate to $r(t) = C_r$ as long as the set point, k_o, is such that $k(t) < k_o \leq k_{cd}$.

Therefore, the set point $k_o = k_{cd}$ always leads the system to its maximum throughput in steady state, considering no fluctuation in demand or modeling errors.

The equilibrium states depend on the PI-controller set-point and demand. From Equation (6.8), the PI-Controller holds constant metering rate $r(t)$ when $e(t) = 0$, as long as $K_i > 0$.

In other words, the PI-Controller assures that the only equilibrium point is $k(t) = k_o = k_{cd}$. However, this state may not be reachable depending on demand patterns.

If the demand is high enough to not match the reachability condition (see Section 5.4.2), it is not possible to prevent the capacity drop phenomenon. Indeed, when $k(t) > k_{cd}$, $r(t)$ will steadily decrease until the lower bound $r(t) = r_{min}$. Once $d_u(t) > C(1-\Delta) - r_{min}$, the influx will be higher than out-flux until $k(t) = \bar{k}$, which is the equilibrium state in this case. Once Z is reachable again (i.e., $d_u < C(1-\Delta) - r_{min}$), the density decreases and eventually $k(t) \leq k_{cd}$.

On the other hand, the set-point might not be reached for low demands. For $k(t) < k_{cd}$ the metering rate will steadily increase until the upper bound $r(t) = C_r$, but as long as $d(t) < C$ it is always possible to serve the demand and the equilibrium state is $k(t) = \frac{d(t)}{v_f}$.

5.5.3 Closed Loop Response and Stability

For stability analysis, we assume:

1. Z is reachable so that there is a $r(t)$ that can lead to the set point;
2. the ramp flux is determined by the metering rate (i.e., $f_r(t) = \min(r(t), d_r(t)) = r(t)$). The on-ramp queue evolve as $\dot{Q}_r(t) = d_r(t) - r(t)$ and will not be included in the state space modeling;
3. the upstream queue is ignored and assumed to be zero and $f_u(t) = d_u(t)$. In this case it is implicitly assumed that $d_u(t)$ reflect any queue or unserved vehicles up to time t.

Also, the upstream demand is split in a constant and a variable term: $d_u(t) = d_{u0} + \delta(t)$ and we define excess density as $x(t) = k(t) - k_{cd}$ and excess demand as $v(t) = r(t) + d_{u0} - C$.

Combining equations (5.1), (5.6), (5.8), and the control law (6.8), the system can be described as:

$$\begin{aligned} \dot{x(t)} &= \frac{1}{L}(r(t) + d_{u0} + \delta(t) - v_f x(t) - C) \\ \dot{r(t)} &= -\frac{K_p}{L}(r(t) + d_{u0} + \delta(t) - v_f x(t) - C) - K_i x(t), \end{aligned} \tag{5.17}$$

which is valid for $x(t) \leq 0$. Similarly, for $x(t) > 0$:

$$\begin{aligned} \dot{x(t)} &= \frac{1}{L}(r(t) + d_{u0} + \delta(t) - C(1 - \Delta)) \\ \dot{r(t)} &= -\frac{K_p}{L}(r(t) + d_{u0} + \delta(t) - C(1 - \Delta)) - K_i x(t). \end{aligned} \tag{5.18}$$

Setting excess density and demand as the state variables, $Y(t) = [x(t), v(t)]^T$, we have the following switched affine system [87]:

$$\dot{Y} = A_1 Y + B_1 + P\delta(t), x(t) \leq 0, \tag{5.19a}$$

$$\dot{Y} = A_2 Y + B_2 + P\delta(t), x(t) > 0, \tag{5.19b}$$

where

$$A_1 = \begin{bmatrix} -\frac{v_f}{L} & \frac{1}{L} \\ \frac{K_p}{L} v_f - K_i & -\frac{K_p}{L} \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, P = \begin{bmatrix} \frac{1}{L} \\ \frac{-K_p}{L} \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0 & \frac{1}{L} \\ -K_i & -\frac{K_p}{L} \end{bmatrix}, \text{ and } B_2 = \begin{bmatrix} \frac{C\Delta}{L} \\ \frac{-K_p C\Delta}{L} \end{bmatrix}.$$

The demand variation is treated as perturbation, $\delta(t)$, which is assumed zero throughout this analysis. The equilibrium points are $Y_1^\star(t) = [0, 0]^T$ for (5.19a) and $Y_2^\star(t) = [0, -C\Delta]^T$ for (5.19b). In both cases $x = 0$ is the equilibrium point; however, $Y_1^\star$ is ideal because the out-flux is greater.

Locally, the dynamic is determined by the eigenvalues of the respective matrix A_k ($k = 1, 2$):

$$\lambda(A_1) = \sigma_{1n} \pm j\omega_1 = \frac{-\frac{v_f + K_p}{L} \pm \sqrt{\frac{(v_f + K_p)^2}{L^2} - \frac{-4K_i}{L}}}{2}, \tag{5.20a}$$

$$\lambda(A_2) = \sigma_{2n} \pm j\omega_2 = \frac{-\frac{K_p}{L} \pm \sqrt{\frac{K_p^2}{L^2} - \frac{-4K_i}{L}}}{2}, \tag{5.20b}$$

where $\lambda(A_k)$ denotes the eigenvalues of matrix A_k.

Through the nature of the eigenvalues in respect to the sign of its real part and whether it is a complex number and the initial condition after switching between regimes, the stability is derived. The complete derivation is in Section 5.8.

The stability is guaranteed by two basic condition. After switching from congested to uncongested state, it remains uncongested and converges to $Y_1^\star$. Also, if initially congested it should be guaranteed that the system will eventually switch to the uncongested state.

The first condition is guaranteed by real and negative eigenvalues of A_1. While it is possible

a switch to the congested state depending on the initial conditions, an eventual transition back to the uncongested state always will be $Y(0) = [0, -C\Delta + \epsilon]$, $\epsilon > 0$ and for this initial condition real and negative eigenvalues will lead the system to the origin.

The second condition, is to assure that the system initially congested eventually switches to the uncongested regime. In this case, a real and negative eigenvalue can settle the system on the congested equilibrium state congested side. Complex eigenvalues assures that it does not happen. In the specific case where eigenvalues are real and positive, the system always switches back as long as $K_i > 0$ due to the saturation. As the real part is positive, the system initially diverges from $x = 0$ and reaches $r(t) = C_r$; at this point $K_p x(t)$ keeps constant while the integral term increases; when the integral term exceeds the proportional, the system is pushed back to the uncongested side.

Combining all these cases, the system is stable and converge to $Y_1^\star$ when:

$$\begin{aligned} K_p &> -v_f \\ H(K_p)\frac{K_p^2}{4L} \leq K_i &\leq \frac{(v_f + K_p)^2}{4L} \end{aligned} \tag{5.21}$$

An interesting fact is that the drop amount, Δ, does not influence the eigenvalues. So, these results would be the same as long as the drop amount is greater than zero.

Another addressed case is when eigenvalues of A_1 and A_2 are all complex numbers. In this case, whatever the initial conditions, it has $x(t) = a_n sin(\omega_n t)$ and it always cross the line $x = 0$ at $\frac{\pi}{\omega_n}$. In the new region, it changes ω_n and a_n, but not the functional form and it always switches back after half a period. We use Poincaré Map [136] analysis to study the behavior of the oscillations over multiple cycles.

Consider the Figure 5.6, we assume the initial condition at $Y(0) = [0, v_1]^T$. It follows a

sinusoidal trajectory and intercepts again $x = 0$ in the point $Y_2 = [0, v_2]$. This process repeats until point $Y_3 = [0, v_3]^T$ and so on. At each segment, $v_i = f_1(v_{i-1})$ and therefore $v_i = f_2(v_{i-2})$. After obtaining $f_2(v)$ it is possible to compute when it will cross the segment $x = 0$ coming from the same dynamic region after n cycles and what is the asymptotic behavior when n approaches infinity.

With the response given by Equation 5.32 and (5.33) and (5.34), v_2 and v_3 are obtained:

$$\begin{aligned} v_2 = Y(\frac{\pi}{\omega_1}) = \frac{v_1}{\omega_1 L} e^{\frac{\pi\sigma_1}{\omega_1}}, \\ v_3 = Y(\frac{\pi}{\omega_1} + \frac{\pi}{\omega_2}) = (\frac{v_2 + C\Delta}{\omega_1 L}) e^{\frac{\pi\sigma_2}{\omega_2}}. \end{aligned} \tag{5.22}$$

Combining both equations:

$$v_3 = v_1 e^{\frac{\pi\sigma_1}{\omega_1} + \frac{\pi\sigma_2}{\omega_2}} + C\Delta e^{\frac{\pi\sigma_2}{\omega_2}}. \tag{5.23}$$

Equation (5.23) has one fixed point ($v_1 = v_3 = v^\star$) at $v^\star = \frac{C\Delta e^{\frac{\pi\sigma_2}{\omega_2}}}{1 - e^{\frac{\pi\sigma_1}{\omega_1} + \frac{\pi\sigma_2}{\omega_2}}}$, configuring a stable limit cycle. Also note that if $\frac{\pi\sigma_1}{\omega_1} + \frac{\pi\sigma_2}{\omega_2}$ is positive, the trajectories will increase over time; when negative it asymptotically goes to $v^\star$. The period is $\frac{\pi}{\omega_1} + \frac{\pi}{\omega_2}$.

On the other hand, for $\frac{\pi\sigma_1}{\omega_1} + \frac{\pi\sigma_2}{\omega_2}$ is negative, the system will approach $v = v^\star$ as $t \to \infty$.

5.5.4 Numerical Examples

It is considered a section of freeway with four lanes, dropping to three as depicted in Figure 5.1. The length of merging segment is $L = 600m$, $v_f = 30m/s$, $\omega = 35/8m/s$, and $k_j =$

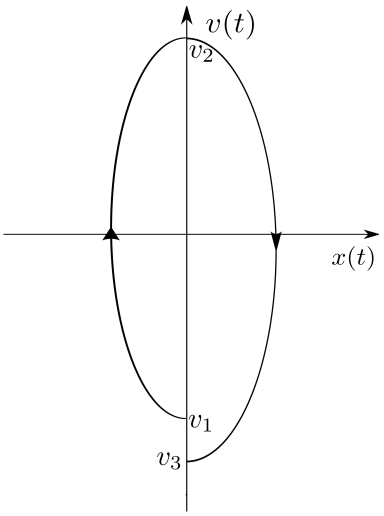


Figure 5.6: Phase diagram $x(t)$x$v(t)$ with initial condition in $Y(0) = [0, v_1]^T$.

$4/7veh/m$. In these conditions, $k_c = 4/55veh/m$ and the downstream capacity is $C = \frac{3}{4}v_f k_c$ and drop magnitude $\Delta = 10\%$. The controller is set with set-point $\bar{x} = 0$ (k_{cd}) and $r_{min} = 0$. Euler's method with 1 second time-step was used as discretization method as it is modeled in continuous time.

There are a lower and upper bound for K_i given a valid K_p in the region defined in (5.21). To analyze both, we do two sets of experiments, one for the lower and another for the upper bound. All cases, demands are $d_u = 0.8C$ and $d_r = 0.2C$ and the proportional gain is $K_p = 10$.

For the upper bound, we set the initial conditions close to $Y_2^\star$, $Y(0) = Y_2^\star + \epsilon$ where $\epsilon = [0.01k_{cd}, 0.01C]^T$. It expected to be stable for $K_i < \frac{16}{24}$. We consider two cases, $K_i = \frac{15}{24}$ and $K_i = \frac{17}{24}$. The simulation-time is 3000 seconds.

While for the lower bound, we set the initial conditions as $k(0) = 1.5k_{cd}$ and $r(t) = C(1 - \Delta) - d_u(t)$. For $K_i < \frac{1}{24}$ the eigenvalues are negative real numbers and we expect to not switch from one side of the half-plane to another; when greater, it is expected to switch from congested to uncongested region, but not the other way around. It is considered $K_i = \frac{1}{23}$ and $K_i = \frac{1}{25}$. The simulation time is 6000 seconds.

The results are as expected. For the upper-bound, both cases have similar trajectories until the vicinity of the origin, but for $K_i = \frac{17}{24}$ the system response is oscillatory and cross the segment $x = 0$, which it is not the case for $K_i = \frac{15}{24}$. While in the lower bound case, both converge to one of the equilibrium points at $x = 0$, but the out-flux are different, as it is possible to observe on the out-flux graph.

Note that the case (b) is a stable limit cycle. As σ_1 and σ_2 are negative, then $\frac{\pi\sigma_1}{\omega_1} + \frac{\pi\sigma_2}{\omega_2}$ is also negative. In that case, it will oscillate with increasing trajectories if $v(0) < v^\star$ and increasing when $v(0) > v^\star$ and eventually reach $v(t) = v^\star$. These two cases are depicted in Figure 5.10.

5.6 Validation of the Results in Cell Transmission Model

All the analyses in this study was based in the Link Queue Model with the capacity drop extension. It is approximation of the LWR model and capable of reproducing the fundamental features of kinematic wave model: demand and supply derived from fundamental diagram and flux functions at junctions [58].

The fact that it is based in ordinary differential equations instead of partial differential equations allowed us to thoroughly analyze the system, especially in closed loop as PI-ALINEA algorithm is itself an ordinary differential equation and the solution to the LWR model considering metering rates based on the downstream density is not known. On the other hand, the Link Queue Model any dynamic that might arise from spatially different densities inside the merging segment is ignored. In this section we validate the results obtained on a model closer to the LWR model, the cell transmission model [28] (CTM).

In the CTM method the discretization is both in time and space. A link of length L is split into n cells of $\Delta x = \frac{L}{n}$, time step size is Δt and the state variable is the average density

within a cell in a given time step, $k^i(t)$, density on cell i on time interval $[t, t + \Delta t]$. Cell length and step sizes should yield CFL number lower than or equal to 1 [25] (i.e., $v_f \frac{\Delta t}{\Delta x} \leq 1$). As $v_f = 30m/s$, we keep the same time step as the previous experiments, $\Delta t = 1s$, then $\Delta x = 30m$ and $n = 20$.

For each cell demand, D_i, and supply, S_i, are defined as:

$$\begin{aligned} D_i &= \min\left(v_f k^i, v_f k_c\right) \qquad i = 1, ..., n \\ S_i &= \min\left(v_f k_c, \omega(k_j - k^i)\right) \qquad i = 1, ..., n \end{aligned} \tag{5.24}$$

The downstream portion is assumed to be uncongested and its supply is the downstream capacity:

$$S_{n+1} = C. \tag{5.25}$$

For all cells, except at the downstream boundary, the flux is the inter cell flux is the minimum between the demand of the downstream cell and supply of the upstream cell:

$$q_{i,i+1} = \min(D_i, S_{i+1}), \qquad i = 1, ..., n. \tag{5.26}$$

For the last cell, the phenomenological capacity drop model is integrated to compute the

flux:

$$q_{n,n+1} = \min(D_n, S_{n+1}, C(1 - \Delta H(D_n - S_{n+1}))) \tag{5.27}$$

Therefore the out-flux is lower when demand, D_n is greater than supply, S_{n+1}.

To compute the in-flux at the first cell, it depends on upstream and ramp demand as well as the metering rate:

$$\begin{aligned} q_{r,1} &= \min(r, Q_r + d_r, S_1), \\ q_{0^-,1} &= \min\{d_u + Q_u, S_1 - q_{r,1}\}, \\ q_{.,1} &= q_{r,1} + q_{0^-,1}, \end{aligned} \tag{5.28}$$

where $q_{r,1}$, $q_{0^-,1}$ and $q_{.,1}$ are the ramp, upstream and total flux to the first cell respectively. With all fluxes computed, the density for each cell is updated:

$$k^i(t + \Delta t) = k^i(t) + \frac{\Delta t}{\Delta x}(q_{i-1,i}(t) - q_{i,i+1}(t)) \qquad i = 1, ..., n \tag{5.29}$$

5.6.1 Verification of Stability

To validate the stability in this case, the upstream demand is constant, $d_u(t) = 0.8C$, and ramp demand:

$$\begin{aligned} d_r(t) &= 0 \qquad 0 < t \leq 300s \\ d_r(t) &= 0.25C \qquad t > 300s, \end{aligned} \tag{5.30}$$

and all other parameters are the same as the previous cases. Thus, it is expected an increasing metering rate in the beginning and then a sudden increase on the ramp demand leads to the congested regime. At this point, the controller must push the system to the uncongested state again and eventually allowing the out-flux close to capacity. We are looking if the system remains sufficiently close and almost steady at $Y_1^\star$, for a long time. The simulation time is 5 hours (18000 seconds).

The criterion of stability is based on the last hour of simulation. It is computed the average out-flux, $\bar{g}$ and the normalized standard deviation of the density $\sigma(k^n)$ and it is considered stable if $\bar{g} \geq 0.97C$, ensuring discharging close to capacity, and $\sigma(k^n) < 10^{-4}$, ensuring that it is close to a steady state.

With and without capacity drop effect cases were simulated so is possible to analyze its impact. The stability region for no capacity drop effect is on the Appendix. Figure 5.11 shows the result obtained. Blue dots show CTM with, red dots CTM without capacity drop, the black straight line is the stability region for LQM without capacity drop, and blue straight lines is the same model but with capacity drop. In the case where of no capacity drop, it is stable for $K_i > 0$ and $k_p > -v_f/2$ and the regions coincide for lower values of K_p.

Especially for the case where capacity drop occurs, the stability region analytically obtained

Variable	[0s,450s)	[450s,900s)	[900s,1350s)	[1350s,1800s)
$d_r(t)$	0.1C	0.1C	0.1C	0.1C
$d_{u1}(t)$	0.9C	0.96C	0.83C	0.7C
$d_{u2}(t)$	0.9C	0.96C	0.86C	0.67C

Table 5.2: Demands

.

is a good approximation. Note that both K_i and K_p has a maximum value in the CTM case which is not the case for LQM in which K_p and K_i can grow as long as it respects Equation (5.21). There are two reasons for that.

First, in CTM there is an inherent dead-time between a change in the metering rate and the instant that the density changes in its response. It makes the control prone to oscillations and reduces the stability region [103]. It is a relatively small dead time compared with, for example, the control time step, but it makes harder to avoid oscillations which in the end triggers the capacity drop phenomenon.

Second, from matrix P in Equation (5.19), any not modeled dynamic implies $\delta(t) \neq 0$ and it is amplified by K_p/L, therefore as K_p increases the higher is the impact of any deviation between both models. Even respecting the stability region, it is expected that K_p cannot grow indefinitely in practice.

5.6.2 Reachability

The reachability conditions established in Section 5.4.2 was related to the upstream demand and the drop amount. Therefore in this case we keep the ramp demand constant and vary the upstream demand for two demand pattern such that the total demand over the period of the simulation is the same.

The minimum metering rate is $r_{min} = 0.05C$ and $d_r(t) = 0.1C$. The upstream demand and ramp demands are in Table 5.2. In the beginning, Z is reachable and it is possible to

Case	Mainline TTS (veh-s)	On-Ramp TTS (veh-s)	TTS (veh-s)
Reachable (1)	2673	7732	10405
Not Reachable (2)	23302	41586	64883
(%)	-88%	-81%	-84%

Table 5.3: Performance Metrics for each case

.

discharge at capacity; on the interval $[450, 900)$ it is not possible to avoid a congestion as $d_u + r_{min} > C$. On the interval $[900s, 1350s)$ the first demand pattern leads to Z reachable while the second does not. Note that both cases have the same total demand.

The results are shown in Figure 5.12. After $t = 900s$ the congestion starts to dissipate which eventually allow an increase on the metering rate and discharge more vehicles. On the other hand, for the second demand pattern, the first increase also leads to the congested regime; however, it decreases to a level lower than the dropped capacity and the controller is able to dissipate the congestion. After that the controller is able to clear the on-ramp queue.

The performance metrics for this case is presented at Table 5.3. A small perturbation on the demand leads to a large difference on the total time spent. In this specific case the total time spent increased around 6 times controlled exactly by the same algorithm.

While it might seem a simple example, it can be the underlying principle to fairly high day to day variation on congestion at the same place. A small changes on the upstream demand might impair the capability of the system to avoid or alleviate the congestion resulting in smaller out-flux.

5.7 Conclusion

In this study, by combining a simple link queue model to describe the traffic dynamics of a merge bottleneck, we were able to show analytically the hysteresis imposed by the capacity

drop phenomenon, the reduced reachability region, and the stability range when the merge is controlled by PI-ALINEA.

The reachability is a direct consequence of the hysteresis imposed by the capacity drop phenomenon. The maximum metering rate in which the capacity drop can be avoided is greater than the metering rate necessary to recover from the capacity drop. A quite possible scenario is being possible to avoid the capacity drop, but if a disturbance on the system leads to capacity drop, it might not be possible to recover from it unless the upstream demand ceases. This result is general and regardless of the control strategy.

This is a disadvantage of ramp metering compared to variable speed limit. A reduced speed lower the upstream flow while a lower metering rate reduces the ramp flow. In general, the ramp demand is a small share of the total demand and acting only on the ramp demand may not be enough. On the other hand, variable speed limit moves the congestion upstream which can hit upstream off-ramps first which in not the case for ramp metering in which the congestion is moved to the on-ramp.

We derived the stability range for the (PI-)ALINEA, one of the most studied ramp metering algorithms. Considering the capacity drop phenomenon, ALINEA can lead the system to the density in which yields maximum throughput if it is in the stability region, that is, theoretically, the target density can be the critical downstream density. In practice, a "slightly undercritical" [108] is set. From the proposed model and experiments a possible reason is the following is the asymmetrical effect of a small disturbance. A small decrease on the upstream demand leads to a small decrease on the out-flux. However, a small increase can trigger the capacity drop phenomenon and severely decrease the out-flux. Therefore, a slightly undercritical target occupancy avoids the capacity drop at expense of lower out-flux. The effect of target occupancy subject to random arrivals and varying parameters is subject of future studies.

We will also be considering the impacts of the on-ramp queue overriding rule as well as practical implementation issues such as detector placement and controller time steps.

5.8 Appendix - Stability Region Derivation

The stability is assessed by each possible combination of eigenvalues of A_1 and A_2. Real and negative, saddle, complex number, and real and positive. For two different matrices and 4 classification, it is possible 16 combinations; however, some of the combinations are not possible.

We are interested in given initial conditions, $Y(0) = [x(0), v(0)]^T$, eventually will reach the origin in steady state, $Y(t) = [0, 0]$ as $t \to \infty$ assuming reachability condition (Section 5.4.2) and $f_r(t) = r(t)$.

Recall that the model in closed loop is a switch affine system with two state variables. For $x(t) \leq 0$ the state transition matrix is A_1, and for $x(t) > 0$ it is A_2. The eigenvalues and eigenvector defines the response. The eigenvalues can be:

1. real numbers. The system response is described by equation (5.31). The constants c_1 and c_2 are obtained by the initial condition.

$$Y(t) = c_1 \begin{bmatrix} r_{11} \\ r_{12} \end{bmatrix} e^{\sigma_{k1} t} + c_2 \begin{bmatrix} r_{21} \\ r_{22} \end{bmatrix} e^{\sigma_{k2} t} + Y_n^\star, \tag{5.31}$$

   where r_{1j} denotes the elements of eigenvectors associated to σ_{k1}; similarly for r_{2j} and σ_{k2}.

   When the eigenvalues are negative ($\sigma_{kn} < 0$) it is stable node and reaches the equilibrium in both cases. In the uncongested region it corresponds to $K_p > -v_f$ and

$0 < K_i < \frac{(v_f+K_p)^2}{4L}$, while for congested $K_p > 0$ and $0 < K_i < \frac{K_p^2}{4L}$.

2. complex numbers. In this case the system response is oscillatory, given by:

$$\begin{aligned} x(t) &= a_1 e^{\sigma_n t} sin(\omega_n t) + a_2 e^{\sigma_n t} cos(\omega_n t) + y_{n1} \\ v(t) &= a_3 e^{\sigma_n t} sin(\omega_n t) + a_4 e^{\sigma_n t} cos(\omega_n t) + y_{n2}, \end{aligned} \tag{5.32}$$

where $Y_n^\star = [y_{n1}, y_{n2}]^T$.

Once the transition between states is at $x = 0$ and $y_{n1} = 0$, it switches with frequency ω_n. It will be oscillatory for uncongested if $K_i > \frac{(v_f+K_p)^2}{4L}$, while congested for $K_i > \frac{K_p^2}{4L}$. Note that there is a region when it is complex for congested and it is not for uncongested. In that case, the system initially congested eventually switches to uncongested and never switches back.

Assuming initial conditions after it has just switched, $Y_n(0) = [0, v_0]$, the constants for uncongested case are:

$$a_1 = \frac{v_0}{\omega_1 L} \qquad a_2 = 0 \qquad a_3 = -\frac{v_0}{\omega_1}[\frac{K_p}{L} + \sigma_1] \qquad a_4 = v_0. \tag{5.33}$$

while for congested:

$$a_1 = \frac{v_0 - C\Delta}{\omega_2 L} \qquad a_2 = 0 \qquad a_3 = -\frac{v_0}{\omega_2}[\frac{K_p}{L} + \sigma_2] \qquad a_4 = v_0 - C\Delta. \tag{5.34}$$

3. real and identical. The response is critically damped, given by:

$$\begin{aligned} x(t) &= a_1 e^{\sigma_n t} + a_2 t e^{\sigma_n t} + y_{n1} \\ v(t) &= a_3 e^{\sigma_n t} + a_4 t e^{\sigma_n t} + y_{n2}, \end{aligned} \tag{5.35}$$

where constants a_i are chosen so as to respect the initial conditions.

First, we show that $Y_1^\star$ is reached only from the uncongested region and then we split the analysis in how to ensure that the system initially uncongested converges to $Y_1^\star$ and how to push the system from congested to uncongested state.

Theorem 5.4. *The system cannot reach $Y_1^\star$ from the congested state $x(0) > 0$.*

Proof. Assuming an initial condition, $Y(0)$, close to $Y_1^\star$, to reach $Y_1^\star$, $x(0)$ must be decreasing. Let $Y(0) = [x(0), v(0)]^T$ such that $x(0) \to 0^+$ and $v(0) \to 0$. From Equation (5.19b), x ($\dot{x} < 0$) is decreasing for $v(0) \leq -C\Delta$, but it is a contradiction since $v(0) \to 0$, for $C > 0$ and $\Delta > 0$. □

Theorem 5.5. *For uncongested initial condition $x(0) < 0$, a transition to the congested state at time t, $x(0) > 0$, occurs only if $v(t) > 0$.*

Proof. Let $Y(0) = [x(0), v(0)]^T$ such that $x(0) \to 0^-$. From Equation (5.19a), x ($\dot{x} > 0$) is increasing for $v(t) > v_f x(t)$. As $x(t) \to 0$ then $v(t) > 0$.

□

Theorem 5.4 shows that the system reaches $Y_1^\star$ only from uncongested state and, if it switches from uncogested to congested at time t_{s1}, the initial condition will be $Y(t_{s1}) = [0, v(t_{s1})]^T$ and $v(t_{s1}) < -C\Delta$. From Theorem 5.5, a transition from uncongested to congested at time t_{s2} implies the initial conditions is $Y(t_{s2}) = [0, v(t_{s2})]^T$, $v(t_{s2})$ such that $v(t_{s2}) > 0$.

The nature of response is determined by the eigenvalues (Eq. (5.20a)). If they are complex numbers, the response is oscillatory with center at $Y_1^\star$ and eventually cross $x = 0$ for any initial condition different from $Y_1^\star$. Case they are distinct and one of them is greater than

zero, it diverges from $Y_1^\star$. If the eigenvalues are identical and negative the response is:

$$x(t) = c_1 e^{\sigma_{1*}} t + c_2 t e^{\sigma_{1*}}, \tag{5.36}$$

which solution is $c_1 = x(0)$ and $c_2 = \frac{-v_f x(0)}{L\sigma_{1*}} + \frac{v(0)}{L\sigma_{1*}} - x(0)$. Note that if c_1 and c_2 have opposite signs, $x(t)$ eventually changes the sign and switches to congested state, but note that for initial condition $Y(0) = [0, v(0)]^T$, $v(0) < -C\Delta$ it does not. Finally, two distinct eigenvalues $\sigma_{11} < \sigma_{12} \leq 0$ the response is given by Eq. (5.31), $r_1 = [1, \sigma_{11} + \frac{v_f}{L}]^T$, $r_2 = [1, \sigma_{12} + \frac{v_f}{L}]^T$,and for initial condition $Y(0) = [0, v(0)]^T$, $c_1 = -c_2 = \frac{v(0)}{\sigma_{11} - \sigma_{12}}$ and $x(t) \leq 0$ and also does not change signs for $v(0) < 0$.

Therefore, for these two cases, for a given initial condition there could be two possibilities. It can converge monotonically to $Y_1^\star$ or at some instant t_0 change the sign. In this case, if it switches back from uncongested to congested at time $t = t_1$, from Theorem 5.4, the initial condition will be $v(0) < 0$ which was shown that goes to $Y_1^\star$.

We also look for the eigenvalues to analyze when the system is congested. First we consider the case of two real non-positive eigenvalues of A_2 ($\sigma_{21} < \sigma_{22} < 0$) with initial conditions $Y(0) = [0, -C\Delta - \epsilon]^T$, with $\epsilon > 0$, which response is given by:

$$Y(t) = c_1 \begin{bmatrix} 1 \\ \sigma_{21} L \end{bmatrix} e^{\sigma_{21} t} + c_2 \begin{bmatrix} 1 \\ \sigma_{22} L \end{bmatrix} e^{\sigma_{22} t}, \tag{5.37}$$

where $c_1 = \frac{\epsilon}{\sigma_{22} - \sigma_{21}}$ which eventually switches, but it would not if $\epsilon < 0$. Therefore, real negative eigenvalues may not switch to uncongested state depending on the initial conditions. If one is positive, it diverges from $x = 0$ and also does not switch.

If eigenvalues of A_2 are complex, the response is given by Equation (5.32) and, applying the constants in (5.34), $x(t) = \frac{v_0 - C\Delta}{w_1 L} e^{\sigma_2 t} sin(\omega_2 t)$ and it eventually switches to uncongested state again at $t = \frac{\pi}{\omega_2}$.

In all cases presented the fact $r(t)$ is constrained was ignored, but it plays a role in one additional case. If $K_p < 0$ and $K_i > 0$, there is a positive and a negative real eigenvalue which is classified as saddle. This case, having a positive eigenvalue the trajectory goes farther away from $x = 0$ with $r(t)$ increasing as it becomes more congested until $r(t) = C_r$. At this point $r(t)$ stops changing, but as $K_i > 0$ the metering rate will decrease as $K_i(k_{cd} - k) < 0$ until the metering rate is able to make $f(t) < g(t)$ decreasing the density until it switches to the uncongested region. Then, two real and negative eigenvalues will make it converge to $Y_1^\star$.

A schematic for these combinations of initial conditions and eigenvalues type can be seen in Figure 5.13. For complex eigenvalues, blue lines, the system switch from one state to another regardless of the initial condition. When is real and negative, the system may or may not switch to another state depending on the initial conditions. The shaded area represents the initial conditions which it does switch. The red line represents the case which it has a positive and a negative eigenvalue in congested area ($K_p < 0$ and $K_i > 0$). Note that after switching it necessarily converges to respective equilibrium point if the eigenvalues are real and negative inside the area it just switched to.

Combining all these observations, the system will converge to $Y_1^\star$ if eigenvalues of A_1 are real and negative, which ensures that it goes to $Y_1^\star$ after switching from congested state, and eigenvalues of A_2 must be complex or saddle ($K_i > 0$ and $K_p < 0$) to ensure that it will eventually converge to $Y_1^\star$ regardless of the initial condition. It defines the following stability

region:

$$\begin{aligned} K_p &> -v_f \\ H(K_p)\frac{K_p^2}{4L} \leq K_i &\leq \frac{(v_f + K_p)^2}{4L} \end{aligned} \tag{5.38}$$

5.8.1 No Capacity Drop Effect

All derivations in this study was considering the capacity drop phenomenon. In this subsection the analysis is extended to the case where $\Delta = 0$ and there is no capacity drop effect.

The system in this case is still switched; however, the origin is the equilibrium point in both regimes. The difference is that $g(t) = C$ when congested and $g = v_f k_{cd}$; in the boundary both values are the same. The Poincaré map analysis is also valid for that case, but note that $\Delta = 0$ and $\sigma_k < 0$, $v^\star = 0$ as long as $\frac{\pi\sigma_1}{\omega_1} + \frac{\pi\sigma_2}{\omega_2} < 0$, therefore achieving the desired state.

Thus, for no capacity drop effect, it will be stable inside the region in Equation 5.21, but adding two combinations of eigenvalues: real and negative of both A_1 and A_2, and complex numbers with decreasing trajectories. For positive K_p and K_i both sides have real negative components and therefore is stable. For $K_p > -v_f/2$ and $\frac{(v_f+K_p)^2}{4L}$ both sides are complex with decreasing trajectories and therefore stable, which is not the case for lower values of $K_p < -v_f/2$ which case the same conditions as in Equation 5.21 holds. Combining all cases,

the stability region becomes:

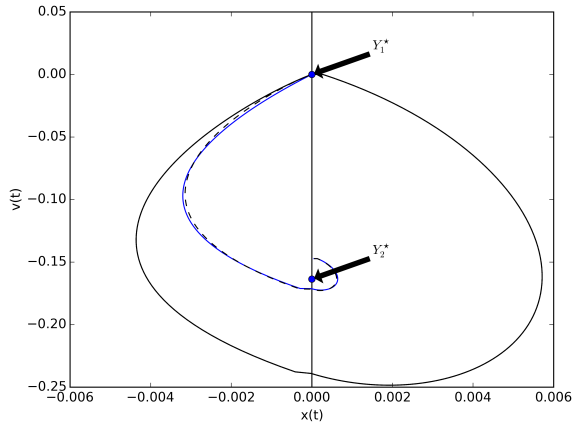
$$\begin{aligned} K_i > 0 \qquad & K_p > -\frac{v_f}{2} \\ 0 \leq K_i \leq \frac{(v_f + K_p)^2}{4L} \qquad & K_p \leq -\frac{v_f}{2}. \end{aligned} \tag{5.39}$$

Positive values of K_i and K_p always lead to stability in closed loop; K_p can be negative, but the region is stricter in this case. Two aspects are worth mentioning in this result. First, the discretization was not considered. It changes the closed loop dynamic if the output variable, $k(t)$, can present large variation between two successive sampling times which in this case it would happen as K_p become much greater than v_f.

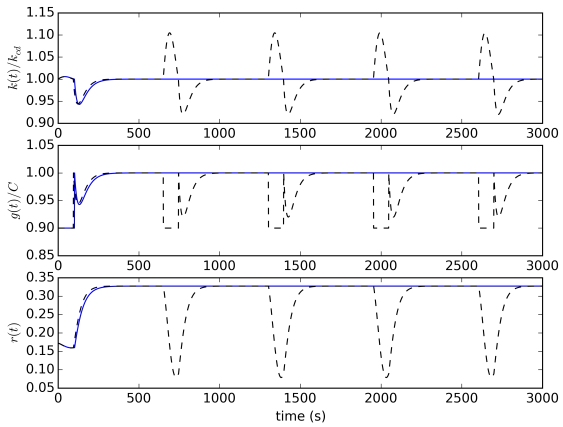
Second, this analysis assumes triangular fundamental diagram, but it can be extended to piece-wise linear if one changes v_f by $\frac{dq}{dk}$ close to maximum out-flux in the FD. For example, with trapezoidal FD it would be $K_i, K_p > 0$. Therefore, this result should be understood that with suitable controller time-step and gains that do not lead the time-response to something closer to the time-step, the system is stable [2].

The interesting point in this case is a smaller stability region when the capacity drop is considered. Even though this result was derived from some assumptions and in a simplified model, experiments using the Cell Transmission-Model (Section 5.6) also shows that both regions are in fact different.

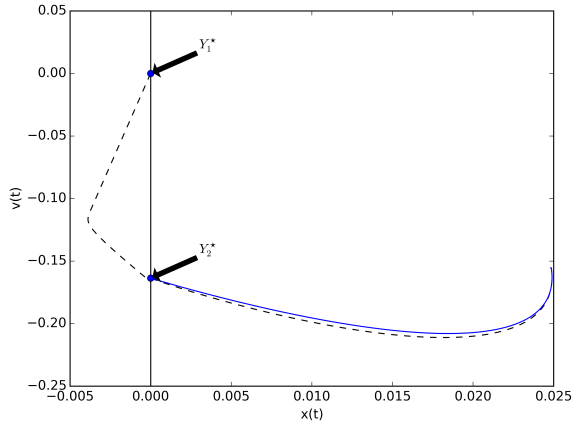
[2]It is still valid for traditional ALINEA which sample time is usually in the same magnitude of L/v_f which the time-constant (real part of the inverse of the largest eigenvalue) in closed loop should be, at most, higher than the sample time. In [130] a similar analysis for this case (i.e. without CD) is done in discrete-time which takes the discretization into account.



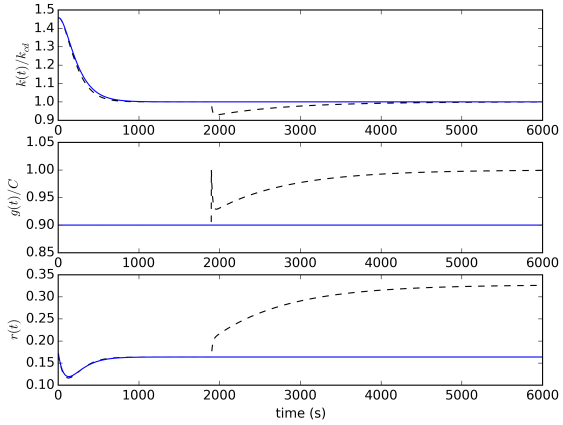
(a) Phase Plane for $K_i = \frac{15}{24}$ and $K_i = \frac{17}{24}$.



(b) Normalized density, out-flux, and metering rates for $K_i = \frac{15}{24}$ and $K_i = \frac{17}{24}$.



(c) Phase Plane for $K_i = \frac{1}{25}$ and $K_i = \frac{1}{23}$.



(d) Normalized density, out-flux, and metering rates for $K_i = \frac{1}{25}$ and $K_i = \frac{1}{23}$.

Figure 5.7: In (a) and (b) blue line is $K_i = \frac{15}{24}$ and dashed lines $K_i = \frac{17}{24}$. The stable case, blue line, it goes to $Y_1^\star$, while the unstable case it keeps oscillating. Note in the phase plane (a), it follows the same trajectory multiple times as it can be observed in (b) and the dashed lines became continuous. In (c) and (d) blue line corresponds to $K_i = \frac{1}{25}$, and dashed to $K_i = \frac{1}{23}$. Both follow similar trajectory, but when $K_i = \frac{1}{25}$, $Y_2^\star$ is a equilibrium point, whereas for $K_i = \frac{1}{23}$ it is not and goes to $Y_1^\star$ instead.

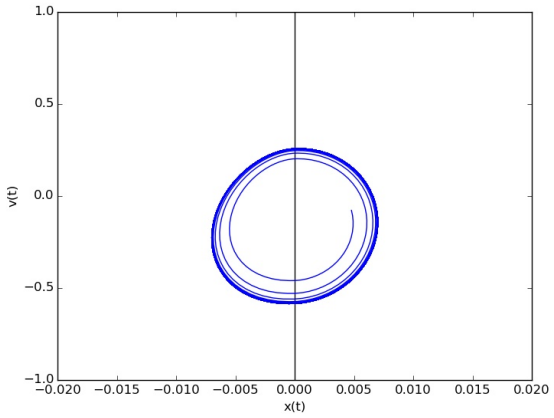


Figure 5.8: $v_0 < v^\star$

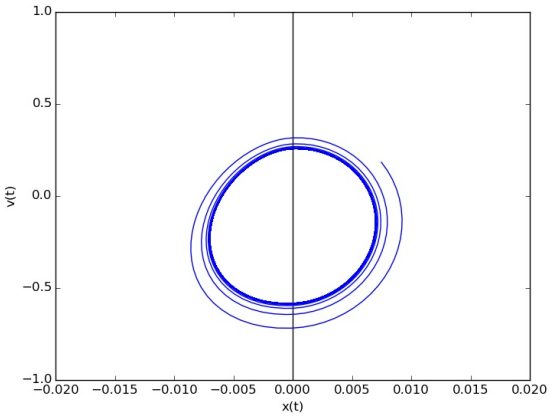


Figure 5.9: $v_0 > v^\star$

Figure 5.10: Oscillating Trajectories for (a) $v_0 < v^\star$ and (b) $v_0 > v^\star$.

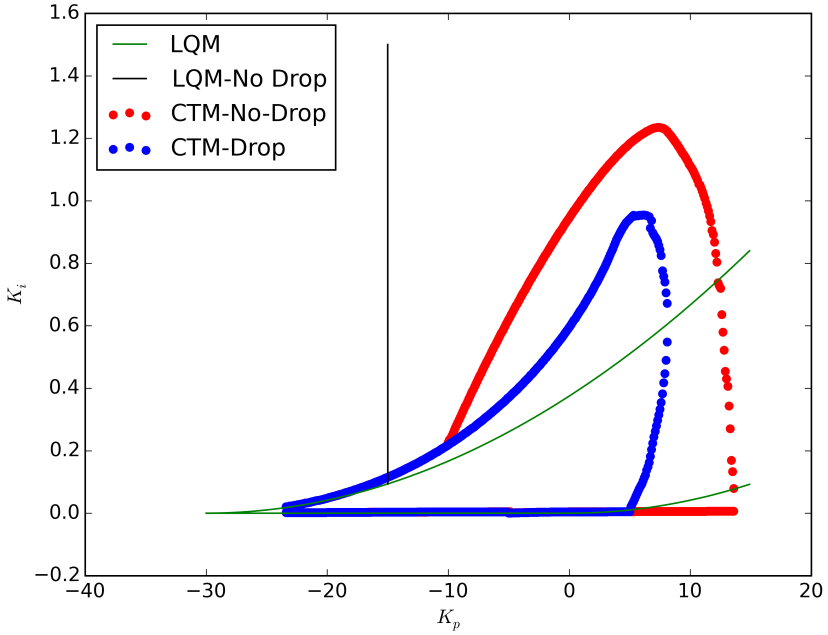


Figure 5.11: Analytical stability region using LQM (straight lines) and dots are the stability through CTM simulation, blue for $\Delta = 10\%$ and red for $\Delta = 0\%$.

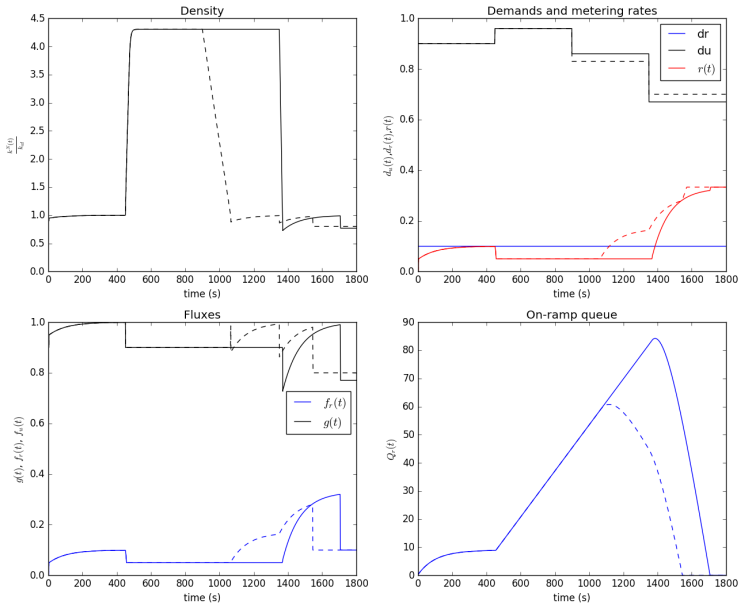


Figure 5.12: Density (top left), demands and metering rates (top right), ramp and out fluxes (bottom left), and on-ramp queue (bottom right).

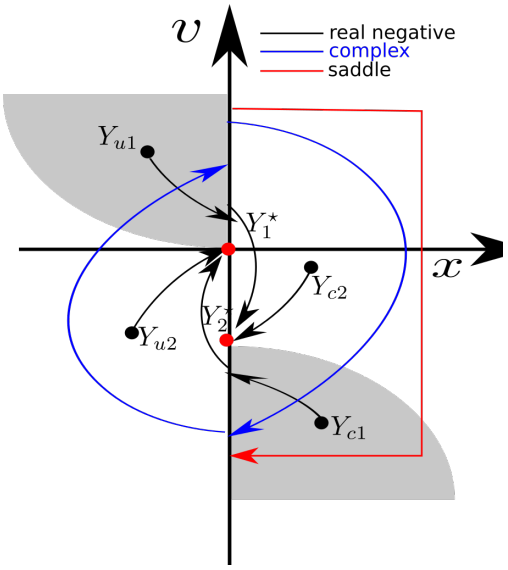


Figure 5.13: Path that the system can follow depending on eigenvalues nature and initial condition. The blue lines represent complex eigenvalues, black real and negative, and red saddle ($K_i > 0$ and $K_p < 0$). The shaded area represents the region where the system with real and negative eigenvalues switches to the another state.

Chapter 6

Integrating a Smith Predictor into Control of Freeways

I'm the master of the universe
And I have seen it all before
Before the war

Helloween (Before the War)

Ramp metering is a freeway management technique that aims to limit on-ramp demand in order to alleviate congestion in the mainline freeway. This action might seem to simply shift of congestion from the mainline freeway to the on-ramp, but its benefits are twofold. First, it has been empirically observed that the maximum mainline freeway discharge rate (capacity) drops with the onset of congestion [24]. Therefore, when the mainline is uncongested its discharge rate is higher. Second, if the congestion grows backwards until an upstream off-ramp, vehicles exiting at that location, which would not be impacted if there was no congestion ahead, take longer to leave the freeway and therefore increasing the total time spent [109].

Figure 6.1 depicts a typical ramp metering scenario, a merge bottleneck consisting of a

mainline freeway merging with an on-ramp, whose lane eventually drops downstream at $x = L$. The controller sets the metering rate, $r(t)$, which determines the maximum flow rate on the on-ramp. Many ramp metering algorithms have been proposed and they vary by the extent of the application and complexity, control objectives, feedback information and control logic. The simplest is to implement a fixed metering rate based on the predicted demand for different periods of time. The most sophisticated method is metering rates updated dynamically for either only a single location or multiple on-ramps when the goal is to achieve an optimal at the system level [55]. Here we are interested in the feedback control by adjusting metering rates based on the measurements from detectors placed at specific locations.

ALINEA is one of the most widely studied ramp metering algorithms [91]. It is based on the feedback control logic, in which the observed occupancy is fed back to a PI controller so that the system can settle down at a desired state. Ideally, two goals can be achieved with the ramp metering algorithm: keeping the freeway uncongested (at the desired level) and with inflow not exceeding the maximum outflow (capacity). In this scheme, the upstream demand is not directly measured, and its variation is treated as disturbance. ALINEA is originally based on an I-controller; recently, ALINEA was extended to a PI-Controller in order to achieve better performance when the lane drop is far from the on-ramp [132], which is the case we address here.

ALINEA requires a detector placed close to lane drop location, at $x = L$ on Figure 6.1, and any change in the metering rate is sensed after the time vehicles take to cross the whole section. This travel time between the on-ramp and the detector is the cause of the dead time. A dead time is a the time elapsed between a control action being performed and the system present any variation in response to this change. In that case, it is riskier an overreaction of the controller. It may keep prescribing corrective measures because the system output is not at the desired state, but the past control action might have been sufficient and the

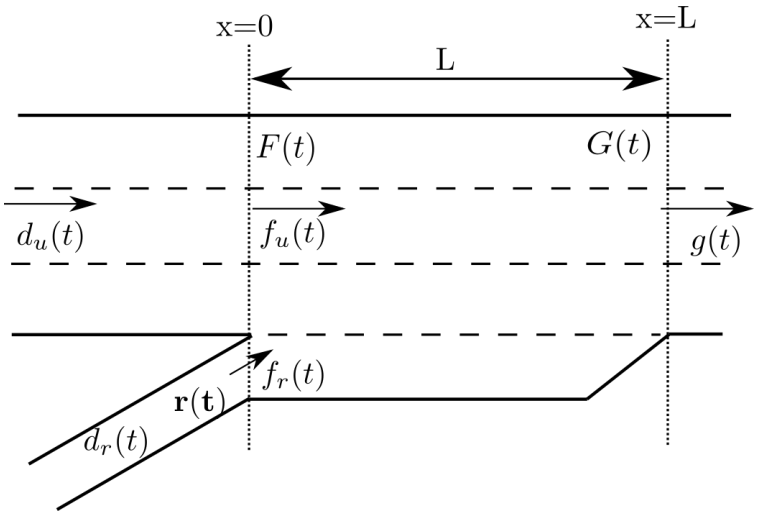


Figure 6.1: The merge bottleneck: A freeway merging with an on-ramp with a downstream lane drop bottleneck.

additional effort leads to an undesired oscillation and even to instability if the overreaction is high enough. In order to avoid these effects, the controller response should be slower which ultimately can have effect on performance.

Our goal is to undermine the effects of the dead time by integrating a Smith Predictor into the controller. The Smith Predictor and its modifications has been used in different areas and especially in industrial processes [102] to overcome the effects of the dead time, but to the best our knowledge has not been applied to freeway control. It explicitly uses a model of the system to compute future outputs in order to compensate the dead time. If there are no modeling errors, the effect of dead time is completely eliminated and the performance of the PI-Controller is recovered as though the system did not have any dead time [103].

In this study, we describe the dynamics of the traffic system with the Link Transmission Model [139], and capture the capacity drop phenomenon with a discontinuous boundary flux function [66]. We show that incorporating a Smith Predictor into ALINEA can provide two major benefits. First, the response to disturbance, which is the primary goal of ALINEA, can be improved. Second, the stability region, which in the literature was obtained through a first order differential equation approximation (e.g.[130, 30]), can be found analytically

considering a more complex and realistic model. Numerical experiments confirm the analytical results and the Smith Predictor is able to achieve better performance even with model uncertainties.

6.1 The System Model

The most established dynamic traffic model is the Lighthill-Witham [86], and Richards [113] (LWR) model which is inspired in hydrodynamic theory. It is a macroscopic model which describes the evolution of macroscopic variables (average density, flow, and speed) in time and space. The two basic principles are vehicle conservation and an unique relationship between density flow referred as fundamental diagram. The solution of the model provides flows and density in space and time.

We consider in this study the link transmission model [139] (LTM). This model is based on the Newell's formulation of the LWR model [100]. The main assumption is a triangular fundamental diagram. We stick with this model for two reasons. First, the dead time is clearer on this formulation although the results would still apply for other fundamental diagram relationships. Second, the control strategy needs a prediction model and the triangular shape leads to a simpler and yet efficient prediction model.

The triangular fundamental diagram [100] is given by:

$$Q(\rho) = \min\{v_f\rho, \omega(\rho_j - \rho)\}, \tag{6.1}$$

where Q is flow, ρ density, v_f free-flow speed, ρ_j jam density (a density in which vehicles are so close that do not move), and ω the shock-wave speed, the density which yields maximum flow is $k_c = \frac{\rho_j\omega}{v_f+\omega}$ and maximum flow is $C = v_f\rho_c$. Figure 6.2 depicts this relationship.

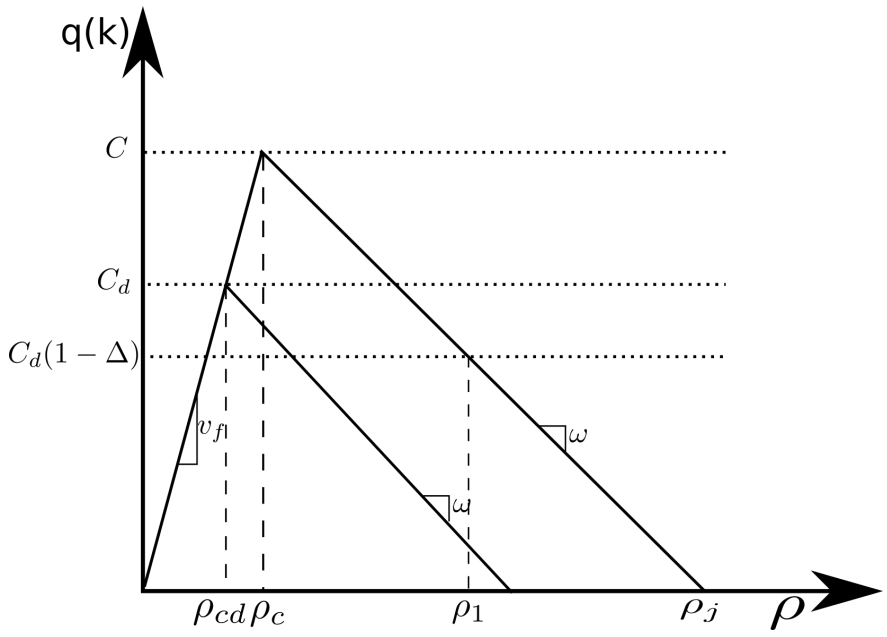


Figure 6.2: The triangular and a general traffic fundamental diagram, the relationship between flow and density.

On the link transmission model the state variables are the upstream, $F(k)$, and downstream, $G(k)$, cumulative flows computed at discrete steps k such that time t is $t = k\Delta t$ where Δt is the time step. The state variables are updated based on the flows:

$$\begin{aligned} F(k+1) &= F(k) + f(k) \\ G(k+1) &= G(k) + g(k) \end{aligned} \tag{6.2}$$

where $f(k)$ and $g(k)$ are the upstream and downstream flows respectively.

Similarly to the cell transmission [28] model, flows are computed through demand and supply

[28, 80]. For each link at every time step, demand, D, and supply, S, are computed:

$$\begin{aligned} D(k) &= \min\{F(k - T_1 + 1) - G(k), C\Delta t\} \\ S(k) &= \min\{G(k - T_2 + 1) + \rho_j L - F(k), C\Delta t\} \end{aligned} \tag{6.3}$$

where $T_1 = \lfloor \frac{v_f}{L\Delta t} \rfloor$ and $T_2 = \lfloor \frac{\omega}{L\Delta t} \rfloor$ are the free-flow and shock-wave traveling speeds respectively. Flows at each boundary are computed based on the upstream demand and downstream supply. In the case of ramp metering, there are the on-ramp demand, $d_r(k)$, that may be limited by the metering rate, $r(k)$, and upstream demand, $d_u(k)$.The upstream demand is computed as:

$$D_u(k) = d_u(k) + \min\{r(k), Q_r(k) + d_r(k)\}, \tag{6.4}$$

where $d_u(k)$ and $d_r(k)$ can be given as boundary conditions or computed as (6.3) if modeled as links. At the downstream end, it is assumed that is uncongested and the supply is determined by the downstream capacity:

$$S_d(k) = S_d = C_d \Delta t. \tag{6.5}$$

On the upstream boundary the flow is computed as:

$$f(k) = \min\{D_u(k), S(k)\} \tag{6.6}$$

where it is assumed absolute priority to the on-ramp so $f_r(k) = \min\{r(k), d_r(k), f(k)\}$ and the remaining flow the upstream section, $f_u(t) = f(t) - f_r(t)$. On the downstream end it is similar, though we integrate the capacity drop model from [66]:

$$g(k) = \begin{cases} D(k), & D(k) \leq S_d \\ C_d(1-\Delta) & D(k) > S_d. \end{cases} \tag{6.7}$$

If the demand is lower than downstream supply (capacity), the flow is the demand and there is no unserved demand. When demand is greater than capacity, there will be unserved demand on the downstream end and congestion arises. The decreased flow is modeled in order to replicate the capacity drop phenomenon [17], a drop in the downstream flow when its upstream section becomes congested. Figure 6.2 depicts these two possible fluxes.

6.2 Closed Loop Analysis

In this section we present the baseline control strategy, the ALINEA control algorithm. Following, we show that the distance between the on-ramp and the bottleneck is large, there might be significant dead time on the control loop that can undermine the controller's performance. After, we show that by integrating a Smith Predictor into the control strategy the effect of dead-time can be mitigated allowing a faster response.

6.2.1 Control Strategy

Our base control strategy is, the (PI-)ALINEA [91, 132]. Probably the most studied ramp meteriing algorithm. It is simple and yet powerful. Several field deployments have reported good performance of PI-ALINEA for local ramp metering control. Traditionally ALINEA is based on a I-Controller of the control theory [6]; more recently it has extended to a PI-Control [132].

We consider the PI-ALINEA on this study. The metering rate is updated based on the downstream occupancy measured by loop detectors placed closed to the lane-drop ($x = L$ at Figure 6.1). The reason to feedback occupancy as it is directly related to density. Here we assume that the downstream density,$\rho^D(t)$, is directly measured and the metering rate is updated as:

$$r(i) = r(i-1) + T_s K_i(\rho_o - \rho^D(i)) + K_p(\rho^D(i-1) - \rho^D(i)) \\ , r_{min} \leq r(k) \leq C_r, \tag{6.8}$$

where the minimum metering rate, r_{min}, is a value defined due to operational constraints (if too low it is less likely that all drivers entering the freeway would respect the signal), and C_r is the capacity of ramp lane. The target density, ρ_o, is a value close to the downstream critical density, which yields flow close to capacity. In practice, a value slightly below ρ_{cd} is picked [108]. The metering rates are updated with period T_s and time is $t = iT_s$.

6.2.2 The Dead-Time in the Control Loop

The formulation presented, based on cumulative curves as state variables, density is not an explicit. The model is nonetheless consistent with the LWR model and therefore any observed flow respects the flow-density relationship. Let's assume the density, $\rho^D(k)$ is exactly measured at $x = L$ where the flow $g(k)$ given by Equation (6.7).

If the demand is lower than supply, all the demand is served and $g(k) = D(k)$ and the traffic is uncongested. Therefore, the density at $x = L$ correspond to that flow on the uncongested side of the fundamental diagram; in this case $\rho^D(k) = g(k)/v_f$. If the demand is greater than supply, $g(k) = C(1-\Delta)$ and the density correspond to that flow on the congested side of the fundamental diagram; $\rho^D(k) = \rho_j - \frac{C(1-\Delta)}{\omega}$.

The unserved vehicles on the link transmission model is given by $Q(k) = G(k) - F(k - T_1)$ [63] that can be seen as a vertical queue at the link's downstream end. With that definition, the density can be computed as function of the queue:

$$\rho^D(k) = \begin{cases} g(k)/v_f, & Q(k) = 0 \\ C(1-\Delta), & Q(k) > 0. \end{cases} \tag{6.9}$$

The queue $Q(k)$ can be computed recursively:

$$Q(k+1) = Q(k) + f_u(k - T_1) - f_r(k - T_1) - g(k), \tag{6.10}$$

and $D(k) = Q(k) + f_u(k-T_1) - f_r(k-T_1)$ and the set-point ρ_o is achieved when $D(k) = v_f \rho_o$.

Assuming the on-ramp flow determined by the metering rate,i.e., $f_r(t) = r(t)$:

$$Q(k+1) = Q(k) + f_u(k - T_1) + r(k - T_1) - g(k). \tag{6.11}$$

As $\rho^D(k)$ is a function of $Q(k)$, the control action, $r(k)$, takes effect after T_1 steps. Note that if the downstream end is congested and constant at $\rho^D = k_j - \frac{C(1-\Delta)}{\omega}$ as long $Q(k) + f_u(k - T_1) + r(k - T_1) > S_d$. Nonetheless it is necessary $Q(k)$ to decrease. Even though the flow is not changed, the dead time of T_1 steps is always present.

This reasoning does not limit to a point measurement at $x = L$. If density is measured on a defined section with a given length, the correspondence between density would not be exactly as (6.9) when congested. However, still $Q(k) = 0$ signs no congestion throughout the section and higher values indicates the extent of a congested region (indeed an one-to-one function as the outflow at $x = L$ is constant when congested). In that case, the density would be a function of $Q(k)$ and the past upstream flows.

The dead-time computed as $t_d = L/v_f$ is present in any setting. However, its effect is significant when t_d is significantly larger than the control time step, otherwise it is not even observed due to the controller sample time. The control time step itself cannot be too small due to some practical aspects. One of them is that the metering rate computed as Eq 6.8 should be discretized to number of vehicles allowed to enter the freeway per time-step.

We target the cases in which t_d is larger than one minute which would be cases where the segment lengths are higher than one kilometer.

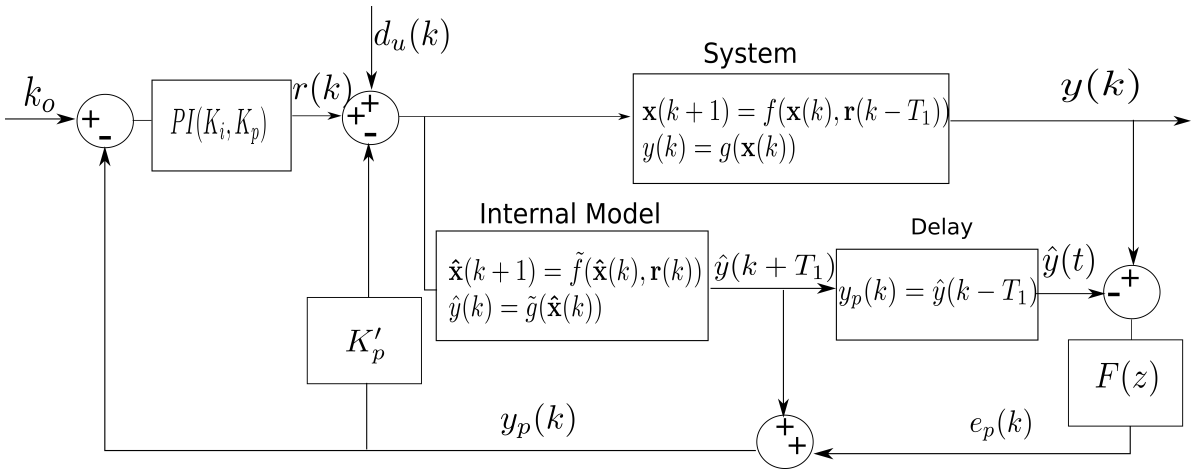


Figure 6.3: Proposed control structure with Smith Predictor and output filter to improve robustness as in [101]. It is assumed all blocks in discrete time or discretized.

6.2.3 The Smith Predictor

To overcome the effect of the dead time some techniques can be used such as Model Predictive Control or dead time compensators [102]. Model Predictive Control takes a model of the system explicitly, and therefore considers the dead time, to compute the control action. The control action is usually the result of an optimization problem [14]. Dead time compensators also consider a model of the system, but it is integrated into a control loop to improve the performance of, for example, a PI-Controller. Most of dead time compensators usually are variations of the Smith Predictor [120]. It is simpler and inherits most of the PI-Controller properties. When suitably designed and there are not modeling errors, it completely diminishes the effect of the dead time.

We extend the ALINEA with a Smith Predictor based structure (SP-ALINEA) in order to compensate the inherent dead time. The block diagram is depicted in Figure 6.3. Considering all elements, it is the robust smith predictor as in [101], with output of the model equal to 0 and $K'_p = 0$ is the original PI-ALINEA, with $K'_p = 0$, and $F(z) = 1$ is the classic Smith Predictor [120].

The assumption of the Smith Predictor is that the control action affects the system after T_1

time steps. The output, $y(k)$, is a function of states at the same time step. If the system model, $f(\mathbf{x}, \mathbf{r})$ and delay, T_1, are known it is possible to estimate the system output after the delay. This predicted value, $\hat{y}(k + T_1)$ is fed back to the controller and therefore the controller "sees" the effect of its past actions. Otherwise, the controller would respond to errors that it had already provided corrective measures.

The internal model consider a system model $\tilde{f}(.)$ while the actual model is represented by $f(.)$. The designer goal is to have both function as close as possible. Assuming $f(.) = \tilde{f}(.)$, the value of $y(k) = \hat{y}(k)$ and $e_p(k) = 0$. In that case, the internal model, which does not have the dead time, T_1, is begin controlled. In other words, the effect of the delay is diminished. However, in reality there will be modeling errors and the internal system and model will have different responses. In that case, terms proportional to the difference between the model and the system appears on the closed loop response. It still presents advantage as long as the errors are small. We do not detail it on this study, see [102] for more details.

The robust structure is considered as the capacity drop imposes a discontinuity on the outflow, $g(t)$, and, hence, density. It becomes important when the free flow speed, v_f is underestimated. The output of the internal model predicts a higher downstream density before then it would actually happen in the system causing successive switchings preventing the system to reach a steady state. The filter $F(z)$ attenuates this oscillation and is able to stabilize the system. Also note, as pointed out in [101], if the system and the model are exactly equal, the filter does not change the dynamic of the system as $e_p = 0$ in this case.

Stability Analysis

The system can be described in two distinct regimes. We can disregard the delay it is compensated by the Smith Predictor. We model as the state variables the metering rates and the downstream demand.

It is assumed that (i) on-ramp flow is determined by the metering rate, $r(k)$, and (ii) the upstream demand is served and split in a constant and disturbance term, $d_u(k) = d_o + \delta(k)$, and (iii) the model and control time step is the same, that is, $T_s = \Delta t$.

When uncongested, there is no unserved vehicles and the system evolve as:

$$D(k+1) = d_o + \delta(k) + r(k) \tag{6.12a}$$

$$\begin{aligned} r(k+1) = r(k) + K_i \Delta t(\rho_{cd} - \rho(k+1)) + K_p(v_f \rho(k) - v_f \rho(k+1) = \\ r(k) + \frac{K_i \Delta t}{v_f}(C_d - D(k+1)) + \frac{K_p}{v_f}(D(k) - D(k+1)) \end{aligned} \tag{6.12b}$$

$$Q(k+1) = 0 \tag{6.12c}$$

It holds as long as $Q(k) = 0$. When congested, the density observed is always $\rho(k) = \rho_1$, the outflow is always $g(t) = C_d(1 - \Delta)$ until the queue is cleared:

$$D(k+1) = C \tag{6.13a}$$

$$r(k+1) = r(k) + K_i \Delta t(\rho_{cd} - \rho_1) \tag{6.13b}$$

$$Q(k+1) = Q(k) - C_d(1-\Delta) + d_o + \delta(k) + r(k) \tag{6.13c}$$

It holds as long as $Q(k) > 0$. As the density is always around ρ_1 the proportional term does not interfere as $\rho(k+1) - \rho(k) = 0$. On the congested regime $r(k)$ decreases with rate $K_i(\rho_1 - \rho_{cd})$, the demand is always at capacity, the flow always at dropped capacity and the queue $Q(k)$ will decrease as long as $r(k) < C_d(1-\Delta) + d_o + \delta(k)$.

On this side there is no equilibrium state. It is necessary $K_i > 0$ to obtain decreasing metering rates; to assure that $Q(k)$ will decrease is necessary that $r_{min} < C_d(1-\Delta) + d_o + \delta(k)$. If the upstream demand is high enough, even with the minimum metering rate it is not possible to alleviate the congestion.

On the uncongested side, however, it needs to settle at the set-point. It can be modeled as a linear discrete system:

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{P}\delta(k) + \mathbf{B} \tag{6.14}$$

where:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ \frac{K_p}{v_f} & 1 - \frac{K_i}{v_f}\Delta t - \frac{K_p}{v_f} \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} d_o \\ -\frac{d_o}{v_f}(K_p + \Delta t K_i) + \frac{K_i}{v_f}\Delta t C_d \end{bmatrix},$$

$$\mathbf{P} = \begin{bmatrix} 1 \\ -\frac{1}{v_f}(K_p + \Delta t K_i) \end{bmatrix}$$

and the state variables $\mathbf{x}(k) = [D(k), r(k)]^T$.

Disregarding disturbances, the system has an unique equilibrium at $r(k) = Cd - d_o$, $D(k) = C_d$. From linear systems theory, the response depends on the eigenvalues of the matrix A:

$$\begin{aligned} z_1 &= \frac{1}{2}\left[(1 - \frac{K_i\Delta t}{v_f} - \frac{K_p}{v_f}) + \sqrt{(1 - \frac{K_i\Delta t}{v_f} - \frac{K_p}{v_f})^2 + 4\frac{K_p}{v_f}}\right] \\ z_2 &= \frac{1}{2}\left[(1 - \frac{K_i\Delta t}{v_f} - \frac{K_p}{v_f}) - \sqrt{(1 - \frac{K_i\Delta t}{v_f} - \frac{K_p}{v_f})^2 + 4\frac{K_p}{v_f}}\right] \end{aligned} \tag{6.15}$$

The response is given by:

$$r(k) = (C - d_o) - a_1 z_1^k - a_2 z^k, \tag{6.16}$$

Without loss of generality we can assume $C - d_o = 1$, the results for other values will scale the constants a_1 and a_2. The constant a_1 and a_2 are defined by the initial condition $r(0)$ and $r(-1)$. In order to be initially at the uncongested dynamics, $r(0) < 1$ and note that the system needs $r(0) < (C - d_o)(1 - \Delta)$ to have switched to the uncongested state.

In order to settle $r(k) = 1$, the function $r(k)$ must be monotonically increasing and have a stable equilibrium at $r = 1$. A stable equilibrium is guaranteed with z_1 and z_2 within the unit circle. Considering $K_i, K_p > 0$ both z_1 and z_2. As both are not complex:

$$\begin{aligned} (\frac{K_i\Delta t}{v_f} + \frac{K_p}{v_f}) + \sqrt{(\frac{K_i\Delta t}{v_f} + \frac{K_p}{v_f} - 1)^2 + 4\frac{K_p}{v_f}} \leq 3 \\ (\frac{K_i\Delta t}{v_f} + \frac{K_p}{v_f}) - \sqrt{(\frac{K_i\Delta t}{v_f} + \frac{K_p}{v_f} - 1)^2 + 4\frac{K_p}{v_f}} > -1 \end{aligned} \tag{6.17}$$

Considering the positive quadrant, $K_i > 0$ and $K_p > 0$, the eigenvalues are such: (i) $z_1 > 0$; (ii) $z_2 < 0$; (iii) $|z_1| > |z_2|$. From the initial conditions, we have $a_1 + a_2 = 1$ and the condition reduces to:

$$\begin{aligned} r(k) = 1 - a_1 z_1^k - (1 - a_1) z_2^k > 1 \\ a_1 z_1^k - (1 - a_1) z_2^k > 0 \\ a_1 (z_1^k - z_2^k) - z_2^k > 0 \end{aligned} \tag{6.18}$$

Note that $z_1^k - z_2^k > 0$ and decreasing and z_2^k also decreasing. The second root, z_2, however, can change sign at every step. Therefore, if the inequality holds for $k = -1$ and $k = 0$ it holds for all $k > 0$. In this case if $r(-1) \leq 1$ $r(z)$ is monotone. From the analysis the stability range become:

$$\begin{aligned} K_i, K_p > 0 \\ K_i \Delta t + K_p < v_f \\ K_i \Delta t + K_p < -2 v_f \end{aligned} \tag{6.19}$$

The stability range cannot be derived in a similar way if the dead-time is not compensated. In that case, $r(k)$ changes the demand on time $k + t_d = k + \frac{L}{v_f \Delta t}$. It can be represented in state space, by adding $\lfloor t_d / T_s \rfloor$ states. There would not be a closed form for the eigenvalues in that case.

Nevertheless, this stability range is larger than it would be without the Smith Predictor and it does not depend on the segment length, L. We show in the next section that it is larger

even if the internal model contains modeling error.

6.3 Simulation Experiments

Numerical experiments were conducted to evaluate the control structure and the analysis of the system in closed loop. First stability is assessed through multiple simulations to confirm both the stability region obtained and also for model variations. Then a particular case where the effect $F(z)$ is discussed and the benefits of the larger stability range are shown when the controller responds to variation in the upstream demand.

The parameters used for the following simulations are $v_f = 30m/s$, $\omega = 35/8m/s$, $\rho_j = \frac{4}{7}veh/m$ at $x = L$ the freeway drops from 4 to 3 lanes, then $C_d = \frac{3}{4}v_f k_c$, $\rho_{cd} = \rho_o = \frac{3}{4}k_c$, $\Delta = 0.1$, time-step of the controller and simulation are both 1 second.

6.3.1 Stability Region

We check the stability region in Equation (6.19) through successive simulations for different controller parameters K_i and K_p. For each simulation, the downstream density of the last hour is used to assess stability. It is considered stable if the system keeps with average downstream density at $x = L$ on the interval $0.97 \leq \bar{k}/k_{cd} \leq 1$ and the standard deviation lower than 10^{-3}. The sum upstream and ramp demand exceeds the downstream capacity and one disturbance (increase in the upstream demand) is applied. The total simulation time is 5 hours. It is also assumed that the system states are directly measured.

For this experiment, the upstream demand is $d_u(t) = 0.8C_d$ until $t = 1000s$ and increasing to $d_u(t) = 0.89C_d$ afterwards. Ramp demand is $d_r = 0.3C_d$. With this setting there will be unserved demand on on-ramp, in the mainline, or both. It is expected that the controller

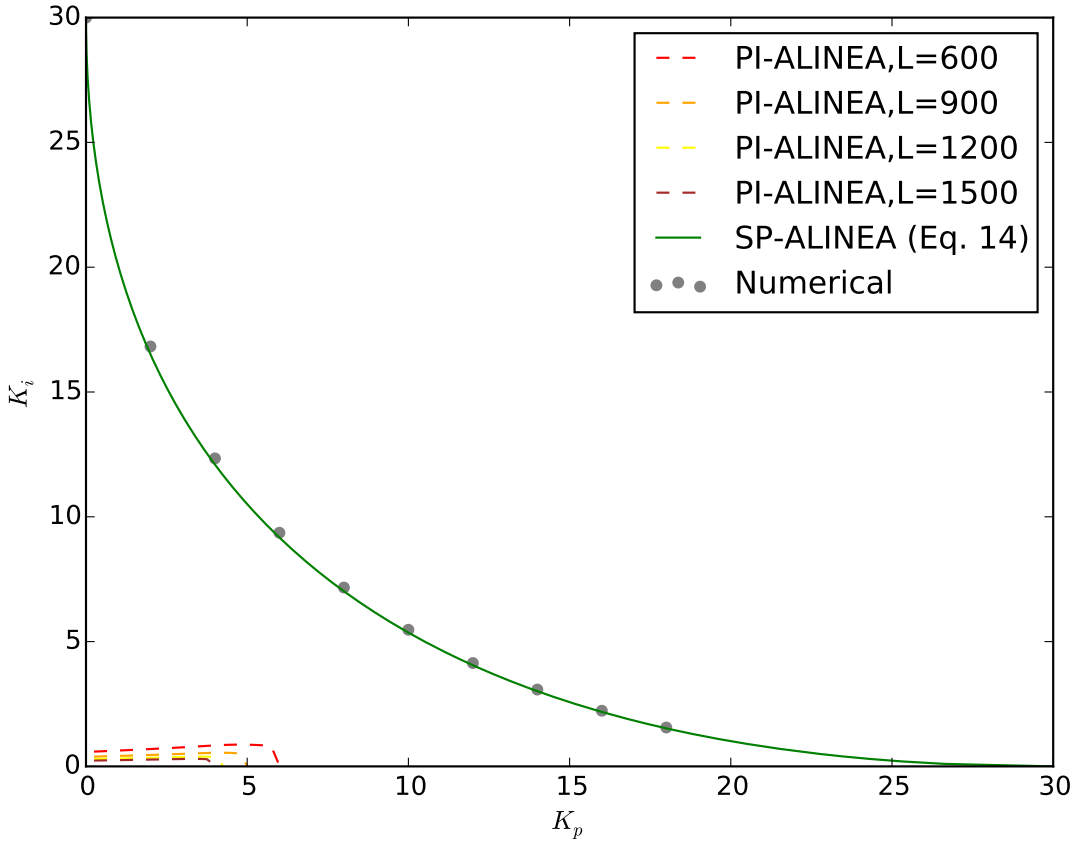


Figure 6.4: Comparison between the stability region of SP-ALINEA with different α and PI-ALINEA for different segment lengths, L.

will keep the freeway uncongested.

Figure 6.4 shows the obtained results. The green line is the Equation (6.19); the dots around the line is the stability boundary found numerically through simulations. For a given K_p, if the K_i is above the dot, the controller is unstable; it is stable otherwise. The dashed lines at the bottom left is the stability region when the Smith Predictor is not used (PI-ALINEA). It is dependent on the segment length, L and values between $600m$ and $1500m$ were plotted.

The Smith Predictor substantially increases the stability range. The practical implication is a faster dynamic response as the PI-Controller coefficients can be larger.

6.3.2 Robustness

The Smith Predictor, in opposition to a PI-Controller, needs a model to compute future responses. However, the performance improvement by the application of Smith Predictor can be undermined if the model is not accurate.

The delay in this case is influenced by the free flow speed, v_f. The controller also needs $k_o = k_c = \frac{\omega}{v_f+\omega}k_j$ which is dependent on both v_f and ω, but this one affects both the cases with and without Smith Predictor. It also cannot be overestimated. If $k_o > k_{cd}$, the system will eventually become congested yielding $C(1-\Delta)$ which will lead the downstream density to $k = k_2 = k_j - C_d(1-\Delta)/\omega$ and remains as long as the congestion is not dissipated.

Then, all the experiments are with $k_o = k_{cd}$. With a target density lower (underestimated) would be easier to achieve stability as a small overshoot around this point does not entail a switch to the congested region. The parameters change is v_f and ω such as the critical density, k_c, remains unchanged.

The results are depicted on Figure 6.5. The cases labeled from 0.92 to 1.0 is the ratio of errors on parameters, α. It is computed $\tilde{v_f} = \alpha v_f$ and $\tilde{\omega} = (1/\alpha)\omega$ where $\tilde{v_f}$ and $\tilde{\omega}$ are the values considered in the internal model and v_f and ω are the actual values.

The stability of PI-ALINEA is much smaller than the SP-ALINEA even with modeling errors and the difference is higher as the section length, L, grows.

The case in which the free flow speed is underestimated ($\alpha > 1$) is more challenging. The internal model predicts that capacity will drop before the actual time leading to a sudden drop in the metering rate. The capacity drop happens afterwards, but the anticipated actions prevents the system to converge to the set-point. This oscillation is mitigated with the filter $F(z)$ and with the internal model not imposing a capacity drop $\Delta_{model} = 0$. In Figure 6.6 the case where $\alpha = 1.03$ is tested with $F(z) = 1$, until $t = 1000s$ and $F(z) = \frac{1}{(0.9z+0.1)^2}$

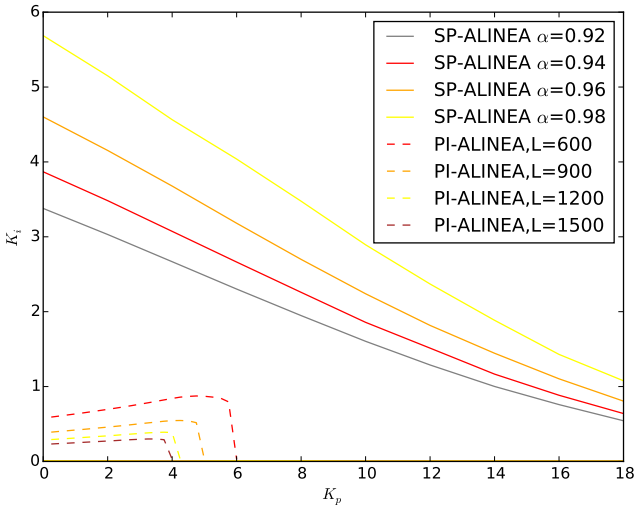


Figure 6.5: Stability of the system for different case of controller and model uncertainties. The dashed lines is the PI-ALINEA for L ranging from $600m$ to $1500m$. The values of α from 0.92 to 0.98 is the SP-ALINEA with internal model with different free flow speeds.

afterwards. In this case $K_i = 2.5$ and $K_p = 3$.

Therefore the filter can indeed improve robustness, but it still presents a high-frequency oscillation in steady state. It is more damaging to underestimate the free-flow speed parameter rather than the opposite. Also, it is an additional step in the controller design and adds another dimension to deal with in terms of performance and stability. The stability region through simulations are not shown as in the previous case as it depends on the filter $F(z)$ as well.

6.3.3 Performance

As the downstream density depends on the upstream demand and it is not taken into account in the control law. Any variation on it is treated as a disturbance ($\delta(t) \neq 0$). An increase (decrease) in the upstream should be followed by a decrease (increase) in the metering rate. In this experiment, a piecewise constant demands is applied, whose profile is shown in the bottom graph of Figure 6.7. For this comparison, the gains of the PI-ALINEA are $K_i = 0.4$

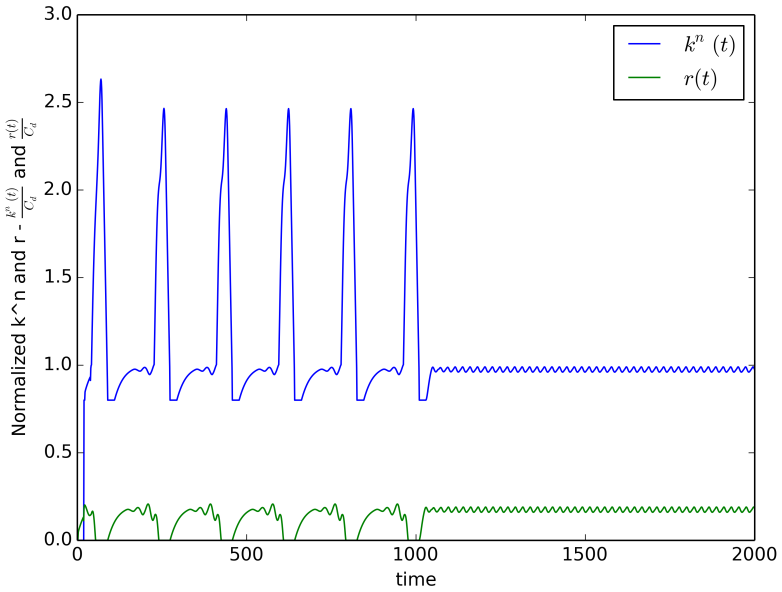


Figure 6.6: System response and metering rate when speed is overestimated. With $F(z) = 1$ for $t < 1000s$ and a second order filter $F(z) = \frac{1}{(0.9z+0.1)^2}$ for $t \geq 1000$.

and $K_p = 1.5$ which is close to the largest possible but still inside the stability region; while for the SP-ALINEA, $\alpha = 0.96$, i.e., with the internal model different from the actual system, $F(z) = 1$, $K_i = 1.5$ and $K_p = 3$. The evolution of downstream density is in the top graph of Figure 6.7 and the metering rate in the bottom. Both PI-ALINEA and SP-ALINEA are able to reject the disturbances, but the Smith Predictor allows a faster response as the gains are higher. The difference would be even higher as L increases because SP-ALINEA would need more conservative gains.

Table 6.1 presents the time spent on the mainline freeway, on the on-ramp and the total time spent for the PI-ALINEA, SP-ALINEA and No-Control case. As expected, both feedback control cases, PI-ALINEA and SP-ALINEA, perform better than the no-control case by avoiding the capacity drop phenomenon. The SP-ALINEA yields better performance than PI-ALINEA due to a faster response.

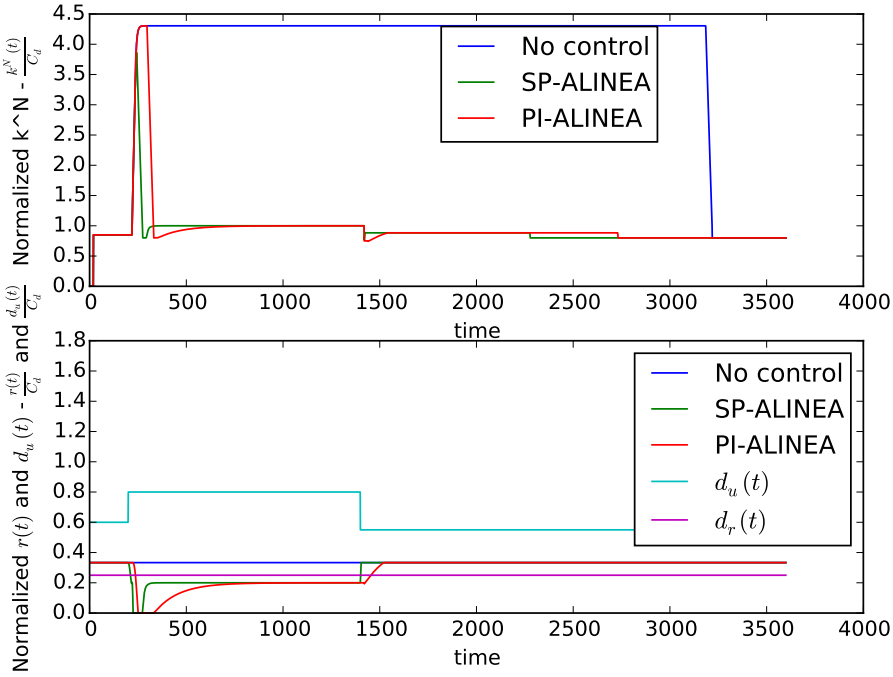


Figure 6.7: System response to disturbance without and with dead time compensation ($\alpha = 0.96$). Even with modeling errors, the response of the SP-ALINEA is faster than PI-ALINEA.

Case	Mainline (veh-s)	On-Ramp (veh-s)	Total (veh-s)	(%)
No-Control	545.9×10^3	0	545.9×10^3	-
PI-ALINEA	104.7×10^3	241.7×10^3	346.4×10^3	-36.5%
SP-ALINEA	104.2×10^3	129.9×10^3	234.2×10^3	-57.0%

Table 6.1: Total time spent for SP-ALINEA, PI-ALINEA and No-Control.

6.4 Conclusion

This study shows that incorporating a Smith Predictor to ALINEA strategy can improve its performance especially when the lane-drop (point of measurement) is distant from the on-ramp. The proposed control structure is able to stabilize in the presence of model uncertainties, even though it was shown that an overestimated free-flow speed, v_f, is harder to deal with compared to an underestimation of that parameter.

Both controllers, PI-ALINEA and SP-ALINEA, are able to reject disturbances and lead the system to the desired equilibrium state, which is no congestion on the merging area and discharging at capacity. However, the Smith Predictor diminishes the dead time effect and allows a faster response. The behavior of the response change in uncongested regime. When congested, both quickly decreases the metering rate, the difference is that the Smith Predictor detects and respond to the transition in advance. On the other hand, when uncongested and able to discharge at capacity, the response of PI-ALINEA should be slower in order to avoid oscillations and therefore a transition to the congested state. With the Smith Predictor, the response in free flow speed can be faster and still not oscillatory.

This work also shows the possible benefits of having additional detectors or other sources of data (e.g. on-line vehicle trajectories from GPS sent by cell-phones). The Smith Predictor is, probably, the simplest way to use this information: as all states and the model are known, it is an easy task to compute future outputs. Another predictive structure using a single point of measurement could effectively deal with the inherent dead time, but any variation on the upstream demand, treated as disturbance in the ALINEA scheme, would only be sensed when it reaches the lane drop. The proposed structure naturally takes advantage of that information and still keeps the essential features of the controller.

Two aspects were not investigated in this study and are objects of future research. First, it needs measurements of all states of the system or, at least, a well designed state estimator.

It was shown in [40] that the system is observable when uncongested with a detector on the downstream end; and with a detector on the upstream end when congested. Therefore, it will be generally observable with one detector in each end of the merging segment. However, due to the two distinct dynamics (one for congested and another for uncongested) it is not trivial task to design a state estimator. However, some work had already been proposed as [125, 131] even though aiming the coordinate (several on- and off-ramps) case. Second, the control sample time was one second, but in practice it is usually around the dead time, L/v_f, mainly because the metering rate is actually discrete, as the output of the controller is converted into number of vehicles allowed during the following time step.

Chapter 7

Practical Aspects and Validation of Results in Microsimulation Models

Is that the real life?
Is this just fantasy?
Caught in a landslide
No escape from reality

Queen (Bohemian Rhapsody)

Microsimulation has been widely used for transportation analysis with applications ranging from transportation planning, traffic management, dynamic routing and driver-infrastructure collaboration [21]. Such models require calibration to ensure that the model portrays as accurately and consistently as possible critical features of traffic [23, 33].

Travel times and delays on freeways are strongly dependent on the capacity of their bottlenecks [33], hence, an accurate representation of the capacity is often the first step in calibration procedures. Capacity is defined as the flow observed in a facility in prevailing conditions, according to the Highway Capacity Manual [94]. A common assumption is that

this flow is observed whenever the car arrival rate is higher than the capacity or there are queued vehicles. However, it has been empirically observed that the discharge rate of a freeway bottleneck drops when queues are formed upstream of the bottleneck (e.g., when "prevailing" condition is achieved) [17]. This is the so-called capacity drop phenomenon. A typical value of the drop in the flow rate is around 10%. There is still much debate in the traffic flow theory literature on the exact mechanisms behind the capacity drop, including differences in behaviors or characteristics such as bounded acceleration [71], drivers heterogeneity [22], or lane-changing maneuvers [76].

Several authors have used macroscopic models to understand and replicate the capacity drop (e.g., see [64, 128]), yet few studies have investigated this phenomenon from a microscopic perspective. This includes the work of [72] who analyzed the effect of drivers' relaxation on bottleneck capacity, and [20] who replicated successfully the capacity drop with the INTEGRATION software as a result of acceleration, lane-changing behavior, and fleet composition [20]. Microscopic simulation models have important advantages over macroscopic models in that they may provide detailed insights into the mechanisms that explain the capacity drop. This is of utmost importance to evaluate traffic management strategies since the capacity drop is one of the main causes of delay on freeways [109].

Despite this progress made, no published studies exist in the literature that have used real traffic data to understand and analyze the capacity drop with microscopic simulation models. Here, we use a combined car-following and a lane-changing model to simulate traffic flow in a merge bottleneck in the I-405 near Irvine, CA, where the capacity drop is consistently observed. The parameters of this model are calibrated against the observed data, and used to study the capacity drop.

The paper is organized as follows. In Section 7.1, we describe the simulation model and calibration procedure. Then, in Section 7.2, we present the calibration results, followed in Section 7.3 with a discussion of our main findings.

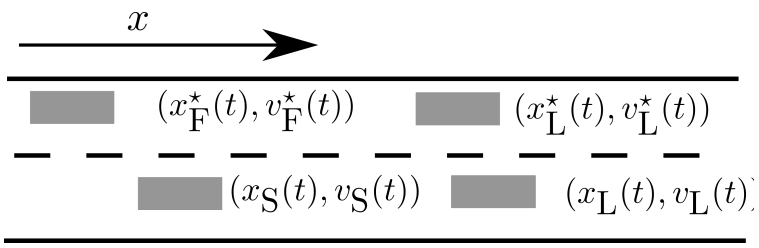


Figure 7.1: Notation of a subject vehicle driving in vicinity of its neighbor vehicles.

7.1 Simulation Model

The simulation model combines a car-following model with a lane-changing model. Route choice is not necessary in our application. A schematic notation is shown in Figure 7.1. State variables of the subject vehicle (S), are speed $v_{\mathrm{S}}(t)$ and position $x_{\mathrm{S}}(t)$. The car immediately in front of the subject vehicle in the same lane is referred to as leader with state variables $(x_{\mathrm{L}}(t), v_{\mathrm{L}}(t))$. A given vehicle aiming to perform a lane-changing defines the potential leader and follower in the target lane with state variables $(x^\star_{\mathrm{L}}(t), v^\star_{\mathrm{L}}(t))$ and $(x^\star_{\mathrm{F}}(t), v^\star_{\mathrm{F}}(t))$, respectively.

We use the Gipps framework [37, 38] for our car-following and lane-changing model with obligatory lane-changing rules based on [53]. For car-following, the subject vehicle updates its speed as:

$$v_{\mathrm{S}}(t) = \min\{v^{\mathrm{u}}_{\mathrm{S}}(t), v^{\mathrm{c}}_{\mathrm{S}}(t)\} \tag{7.1a}$$

$$v^{\mathrm{u}}_{\mathrm{S}}(t) = v_{\mathrm{S}}(t-\tau) + 2.5A\tau\left(1 - \frac{v_{\mathrm{S}}(t-\tau)}{V}\right)\sqrt{0.025 + \frac{v_{\mathrm{S}}(t-\tau)}{V}}, \tag{7.1b}$$

$$
\begin{aligned}
v_{\mathrm{S}}^{\mathrm{c}}(t) &= \max\{v_a, v_b\} \\
v_a &= -B(\frac{\tau}{2}+\theta)+ \\
&\quad \sqrt{B^2(\frac{\tau}{2}+\theta)^2 + B\{2(x_{\mathrm{L}}(t-\tau) - x_{\mathrm{S}}(t-\tau) - S_{\mathrm{L}}) - \tau v_{\mathrm{S}}(t-\tau) + \frac{v_{\mathrm{L}}(t-\tau)^2}{\widehat{B}_{\mathrm{L}}}\}}, \\
v_b &= v_{\mathrm{S}}(t-\Delta t) - B\Delta t
\end{aligned}
\tag{7.1c}
$$

where, $v_{\mathrm{S}}^{\mathrm{c}}$ (t) signifies the speed constrained by the leader vehicle, $v_{\mathrm{S}}^{\mathrm{u}}$ (t) denotes the speed when movement is not constrained by the leader car, A (m/s^2) and B (m/s^2) are the maximum acceleration and deceleration rates which the subject wishes to undertake, respectively, $\widehat{B}$ (m/s^2) is the maximum deceleration rate of the leader as *estimated* by vehicle S, V (m/s) is the free-flow speed; S_L (m) characterizes the effective length of the leader, τ (s) is the reaction time, θ (s) signifies the safety margin, and Δt (s) is the integration time step. All drivers are assumed to have the same parameters.

To discourage deceleration rates higher than B_s, the second term on the max operator of Eq. (1c) is added to the Gipps car-following model, assuming simulation time steps to be smaller than τ. Gipps' updates the state variables via a so-called *iterated coupled map* [70] using $\Delta t = \tau$. For smaller time steps, this term was added to maintain the same behavioral principle which is a maximum deceleration rate of B. To compute the vehicles' speed at $t-\tau$ when τ is not a multiple of the time step, Δt, we use an interpolation method [70] with variable z at time $t-\tau$ computed using:

$$
z(t-\tau) = \beta z(t-(n+1)\Delta t) + (1-\beta) z(t-n\Delta t) \tag{7.2}
$$

where n constitutes the integer part of $\tau/\Delta t$ and $\beta = \tau/\Delta t - n$.

To simulate lane-changing, we use a gap acceptance rule with speed and position of the follower (x_1, v_1) and leader (x_2, v_2) that must respect:

$$B^{\mathrm{L}} = \frac{(v_2 - v_1)^2}{x_1 - x_2 + \tau(v_2 - v_1) - S_1} < B^{\max}, \qquad X_1 - X_2 > g, \tag{7.3}$$

where g (m) signifies the minimum gap. The first equation is based on the assumption that the leading vehicle travels at a constant speed, v_2, and the following vehicle travels at speed v_1 after τ seconds and decelerates right after. For a lane-changing maneuver, the subject vehicle acts as follower with respect to its potential leader and $v_1 = v_{\mathrm{S}}$ (t), $x_1 = x_{\mathrm{S}}$ (t), $v_2 = v_{\mathrm{L}}^{\star}$ (t) and $x_2 = x_{\mathrm{L}}^{\star}$ (t), and as leader with respect to its potential follower in the target lane using $v_1 = v_{\mathrm{F}}^{\star}$ (t), $x_1 = x_{\mathrm{F}}^{\star}$ (t), $v_2 = v_{\mathrm{S}}$ (t) and $x_2 = x_{\mathrm{S}}$ (t).

For discretionary lane-changing, the subject vehicle considers a lane change desirable either for speed advantage or an anticipated mandatory lane change when its distance (based on the current speed) from the mandatory lane-changing location is between 8 and 50 seconds. Once lane-changing is deemed desirable, the driver selects the target lane. In the case of speed advantage, both adjacent lanes, if existent, are considered. If the vehicle needs to undertake a mandatory lane-changing maneuver in less than 50 seconds, the selected lane will be the one which would ensure that the driver can follow the planned route. Regardless of which lane is chosen, the lane-changing maneuver can be performed if the required deceleration rate, from Eq. (7.3), is lower than or equal to the maximum deceleration rate for discretionary lane changes, denoted as B^{D} (i.e., $B^{\max} = B^{\mathrm{D}}$). The required maximum deceleration rate and minimum gap are the only conditions to execute the maneuver in the case of an anticipated mandatory lane-changing. In the case of speed advantage, the leader in the target lane

should have a speed that sufficiently larger than that of the current leader:

$$\frac{v_{\mathrm{L}}^{\star}(t)}{v_{\mathrm{L}}(t)} > \alpha \frac{x_{\mathrm{L}}^{\star}(t) - x_{\mathrm{S}}^{\star}(t)}{x_{\mathrm{L}}^{\star}(t) - x_{\mathrm{S}}(t)}, \tag{7.4}$$

where $\alpha > 1$ (dimensionless) denotes a threshold that weights speed advantage and available spacing in the target lane.

Mandatory lane-changing is performed when the vehicle is within 8 seconds from the point where a lane change is required. Lane-changing is considered required when the vehicle cannot continue its preferred route. In the single merge bottleneck of this study, the end of the acceleration lane of the on-ramp is where a lane change is mandatory.

We implement the lane-changing model in [53] which was especially designed for merging sections based on recorded lane-changing maneuvers. The lane-changing on the merge is especially challenging because in congested periods an acceptable gap in the target lane appears rarely according to discretionary rules. In this case, according to [53], the merging vehicle performs a forced lane-change or the follower in the target lane allows the vehicle to enter, which is called cooperative lane-changing, or both. There are some implementation changes in this study compared to [53] as we consider another car-following model and consider different set for parameters, as for example "driver aggressiveness" does not exist in our implementation.

If a lane-changing maneuver is deemed necessary by a driver, then at every time step, the subject vehicle scans the gaps on the intended lane up to the visible distance. To check whether there is enough gap to fit, we used a distance of 60 meters backward and 30 meters ahead with deceleration of $B^{\mathrm{L}} < B^{\mathrm{f}}$, and minimum gap (7.3), where B^{f} signifies the maximum deceleration to a forced lane changing maneuver. It is assumed that the driver is more aggressive for this kind of lane change and therefore $B^{\mathrm{f}} > B^{\mathrm{d}}$. Still, during periods of congestion it is often not possible to find such a gap. In this case, when the vehicle is

at distance L from the end of the acceleration lane, the subject vehicle sets the current follower and leader in the target lane as the intended gap, regardless of the size of the gap between them. This distance L was found to be critical in other studies (see [64]) and here we consider this to be a calibration parameter.

When assessing a given gap, the driver also checks if it is possible to reach the gap before the end of the acceleration lane, based on the actual speeds and positions of the subject, potential leader and follower, either by accelerating if the potential follower is ahead or braking if the potential follower is behind. During this maneuver, the lane-changing model overrides the car-following model and the vehicle adjusts its speed to reach the gap. We do not explain the details here.

Similarly to the original model [53], once the gap is defined both the new follower and the merging vehicle undertake actions to perform the lane-changing. Both vehicles cannot exceed the constrained speed of Gipps' model, v^{c}, but both vehicles sets acceleration rates in order to the follower vehicle to allow necessary gap to the maneuver and the merging vehicle to be exactly on the proper position as the gap is enough:

$$\begin{aligned} \frac{\mathrm{d}v_{\mathrm{F}}(t)}{dt} &= K_1(x_{\mathrm{F}}^{\mathrm{ref}}(t) - x_{\mathrm{F}}^{\star}(t)) + (v_{\mathrm{S}}^{\star}(t) - v_{\mathrm{F}}^{\star}(t)), \\ \frac{\mathrm{d}v_{\mathrm{S}}(t)}{dt} &= K_2(x_{\mathrm{S}}^{\mathrm{ref}}(t) - x_{\mathrm{S}}(t)) + (v_{\mathrm{L}}^{\star}(t) - v_{\mathrm{S}}(t)), \end{aligned} \tag{7.5}$$

where $x_{\mathrm{F}}^{\mathrm{ref}}(t) = \min\{x_{\mathrm{L}}^{\star}(t) - 2g, x_{\mathrm{S}}(t) - g\}$ and $x_{\mathrm{S}}^{\mathrm{ref}}(t) = (x_{\mathrm{F}^{\star}}(t) + x_{\mathrm{L}^{\star}}(t))/2$ where K_1 ($1/\mathrm{s}^2$) and K_2 ($1/\mathrm{s}^2$) are driving parameters weighting how fast drivers wish to perform the maneuver. Thus, both vehicles follow a reference position in which eventually leads to a successful maneuver where the follower stays at least two minimum gaps from the leader and one gap to the merging vehicle and the merging vehicle stays in the middle of the gap. If the acceleration from Eq. (7.5) is higher than the maximum acceleration (deceleration), the

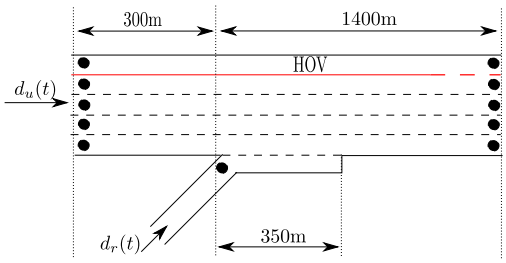


Figure 7.2: Study site with position of the loop detectors

maximum acceleration (deceleration) is taken instead.

7.2 Calibration Procedure and Results

A schematic of the study site is depicted in Figure 7.2, the stretch of I-405N merging with Jeffrey Road in Irvine-CA, United States. Loop detector data are obtained from the California Department of Transportation Performance Measurement System (PeMS). The detectors' positions are as shown in Figure 7.2. The occupancy and vehicle counts at each detector are recorded every 30 seconds. On the PeMS system, the upstream station is labeled as 1201211, the downstream station as 1201222 and the ramp detector as 1201203. Data from the High Occupancy Vehicles (HOV) lane are disregarded as its access is closed for most part of the study site including the most critical section, the area between the on-ramp and the lane drop.

Capacity drop happened most of the morning peaks throughout the first semester of 2012. We chose a day which was representative and the total counts at the upstream and downstream stations were consistent (April-19-2012) [1]. The observed data is depicted in Figure 7.3. The left graphs (a-b) show the occupancies upstream (top) and downstream (bottom). The observed counts are on the middle graphs. The upstream counts (c) include mainline and

[1] In a scenario like the study site, the total upstream and downstream count should be very similar. Due to miscounts, it is common to have a difference between both as reported in [105] when they studied various bottlenecks also using PeMS data

on-ramp. The downstream counts are according to the middle bottom graph (d). Note that the downstream is always uncongested as the downstream occupancy is almost constant throughout the period; however, high occupancies were observed at the upstream detector between 8:00 - 9:00 AM and decreased around 9:00 AM. It means congestion started at the bottleneck, reached the upstream detector, and dissipated when the upstream demand ceased with the congestion at the upstream detector being eliminated just after. The dashed line on the upstream and downstream flows (c-d) refers to the average flow at the downstream between 8:03 - 9:08 AM (q_0).

It is not very clear, though possible, to identify the decrease in the outflow on the downstream flow graph (d) when the section is congested. However, it becomes clear on the top right graph (e) in which the T-curve is plotted. The T-curve is the area between the outflow and a baseline outflow, q_0. Note that $q_0 = 0$ is the cumulative flow. The baseline outflow is arbitrary and we chose it as the average flow during the congested period. This curve has positive slope whenever the flow is greater than q_0 and negative otherwise. Clearly, the outflow was smaller during the congested period compared to the 20 minutes before the drop in the downstream flow, a decrease from 2350 vphpl to 2130 vphpl (9.5%), as annotated on the graph. The vertical difference between the blue and red curves is approximately the number of vehicles between the two stations (apart from initial conditions). The distance between the two curves increases between 7AM and 8AM on the congestion build up and decreases around 9AM when the congestion dissipates.

For the input demands, we considered the demands varying in steps of 10 minutes. For the on-ramp, the arrival rate is the average flow reported by the detector in the 10 minute periods. For the upstream demand the counts were also considered, but with a slight modification. Between 8:00 -9:00 AM, the occupancy reached the upstream detector, so queues were formed before the upstream detector. It means that the flow would have been higher if the bottleneck discharge flow was higher. To address this, we considered a higher arrival pattern during

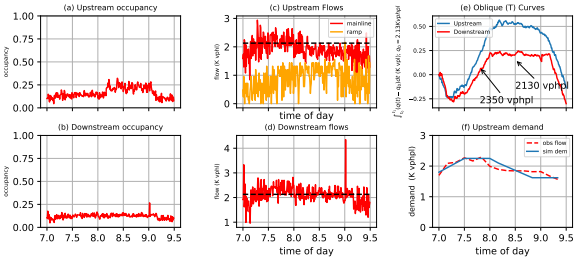


Figure 7.3: Observed data at location. Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depict upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130$K vphpl and graph (f) depicts the upstream demand used in the simulation.

this period, but keeping the same total flow over the simulation period. The input demand can be seen in Figure 7.3 (f). Uniform arrivals were considered for both cases.

The parameters part of the calibration with their respective bounds and units are: V [28-36] (m/s), A [0.8-1.5] (m/s^2), $B = \widehat{B}$ [3-5] (m/s^2), τ [0.5-0.9] (s), S [6-9] (m), L [50-150] (m) and α [0.2-1.1] (-). All the lane-changing parameters except α are fixed. The fixed parameters are: $B^{\mathrm{D}} = B^{\mathrm{F}} = B$ (m/s^2),$K_1 = K_2 = 0.5$ (1/s^2) and $g = 2S$. The time step is set to $\Delta t = 0.4$ (s). We consider three residuals:

$$
\begin{aligned}
QE^2 &= \frac{1}{T}\sum_{i=0}^{T}\left(\widehat{q}(i) - q(i)\right)^2 \\
OE^2 &= \frac{1}{T}\sum_{i=0}^{T}\left(\widehat{O}(i) - O(i)\right)^2 \\
TE^2 &= \frac{1}{T}\sum_{i=0}^{T}\left(\widehat{T}(i) - T(i)\right)^2
\end{aligned}
\tag{7.6}
$$

where $\widehat{q}$ and q signify the observed and simulated downstream flow, respectively. Similarly, O refers to the upstream occupancy and $T(i) = \sum_{j=0}^{i} q(j)$ denotes the cumulative flow, QE stands for downstream flow error, OE occupancy error and TE is the cumulative (T-curve) error. Two minutes moving average is considered for flow and occupancy.

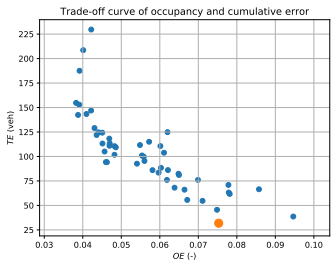


Figure 7.4: Trade-off curve (Pareto Frontier) between occupancy (OE) and cumulative flow error

Most of the calibration procedures consider occupancy and downstream flow error but not cumulative error. As it is easier to observe the capacity drop on T-curve, we consider the cumulative curve a reasonable objective function as well.

The calibration procedure is to minimize the three residuals by adjusting the parameters within their respective bounds. As we consider three residuals, the multi-objective differential evolution algorithm is used based on [114], with population size of 50 and "rand1bin" selection rule. The multi-objective procedure is suitable when the objectives are conflicting, that is, a change in some parameters may decrease one objective to detriment of the others.

This conflict happened between the occupancy and cumulative flow error as depicted in Figure 7.4. Each dot represents one candidate solution with its position with respect to the x-axis representing the occupancy error (OE) and the distance with respect to y-axis representing the cumulative flow error (TE).

Figure 7.5 shows the calibration result of the capacity drop if one considers separately each of the objectives. The left graph shows the upstream occupancy, the middle graph the downstream flow and the right graph the T-Curve for $q_0 = 2130$ vphpl. The yellow, orange, and blue are the results when upstream occupancy, downstream flow and cumulative outflow error are minimized, respectively. Observe that the results are similar for downstream and cumulative flow error, but not exactly the same. Therefore, it is important to consider the cumulative flow as an objective.

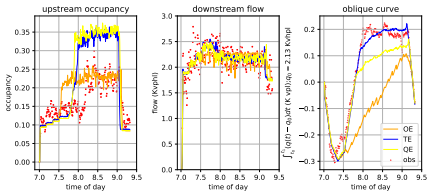


Figure 7.5: Results for calibration minimizing only one of the three objectives considered.

As we are focusing on the capacity drop, we pick the solution with minimum value of cumulative error. The values for which we show the results refer to the orange dot in Figure 7.4. The correspondent parameter values for that case are $V = 33.46$ (m/s), $A = 1.27$ (m/s)^2, $B = \widehat{B} = 3.06$ (m/s^2), $S = 6.33$ (m), $\tau = 0.82$ (s), $\alpha = 0.87$, and $L = 81$ (m).

The comparison between the simulation and actual data is depicted in Figure 7.6. The upstream occupancy and flow are depicted on the top graphs. The downstream occupancy and flow are the middle graphs. The T-curve for $q_0 = 2130$ vphpl is depicted at the bottom left. In all graphs the red dots are the data observed and continuous blue line is the simulated data. Notice that the simulated T-curve almost overlaps with the observed data. It means that the model captures both the capacity (uncongested capacity) and capacity drop (congested capacity) with high accuracy. The shape of the upstream occupancy is similar and the congestion reaches the upstream detector slightly earlier than its observed time. The bottom right graph depicts the scatter plot of flow and density for both upstream and downstream location. The continuous line is the corresponding fundamental diagram based on Gipps' car-following parameters assuming stationary traffic. The downstream area is always uncongested while at the upstream there are transitions to and from the congestion side of the fundamental diagram.

We validate our results by applying the calibrated model in a different period. In Feb-1-2012 the capacity drop also happened during the morning peak and we compared the simulation outcomes with the observed data of data period. The results is depicted in Figure 7.7. Though the duration of congestion was smaller in the simulation, the results are in general

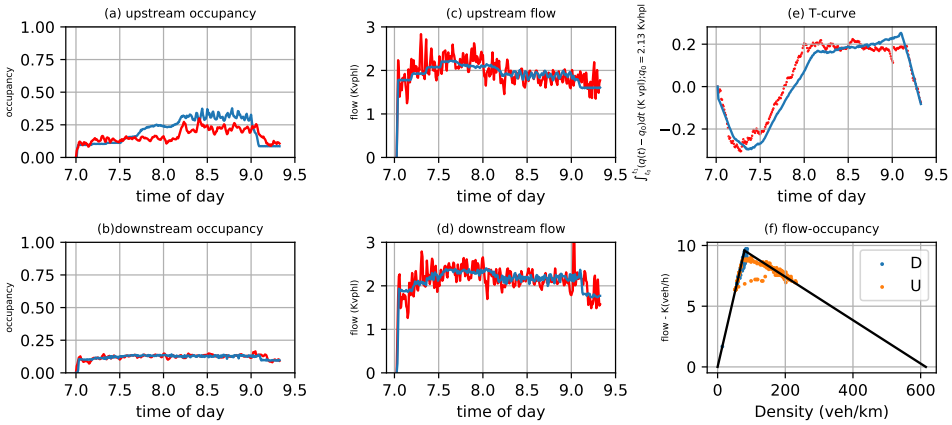


Figure 7.6: Calibration results vs. actual data. Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depicts upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130K$ vphpl and graph (f) depicts the density flow relationship with orange and blue denoting the upstream and downstream location respectively.

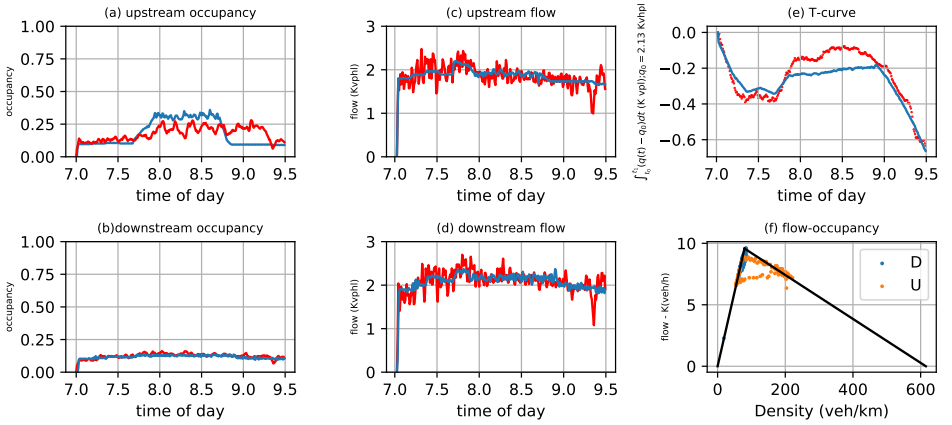


Figure 7.7: Simulation results vs. actual data for validation (Feb-1-2012). Graphs (a) and (b) depict upstream and downstream occupancies, respectively. Graphs (c) and (d) depicts upstream and downstream flow in vehicles per hour per lane. Graph (e) depicts the T curve for $q_0 = 2130K$ vphpl and graph (f) depicts the density flow relationship with orange and blue denoting the upstream and downstream location respectively.

consistent. The capacity drop is triggered around the same time and the outflow during the congested period was similar.

7.3 Conclusion

A microscopic simulation model for a merge bottleneck was calibrated using loop detector data with an observed capacity drop. The model was able to reproduce the capacity drop phenomenon yielding an accurate estimate of the uncongested and congested capacities as well as their trigger time. These preliminary results are encouraging, nevertheless, much further analysis is required to draw conclusions. Nevertheless, we demonstrate evidence that microscopic models can reproduce the capacity drop through a proper calibration procedure. We applied the model in a different day and the results were consistent showing the method is consistent.

The calibrated model can be used to evaluate traffic management strategies such as ramp metering, variable speed limits and cooperative merging at specific locations. In particular, local ramp metering control (i.e., single bottleneck) has received little attention with microscopic traffic flow models. We will report the findings of our studies in due course.

Chapter 8

Concluding Remarks

> The closer you get to the meaning
> The sooner you'll know that you're dreaming
> So it's on and on and on, oh it's on and on and on
> It goes on and on and on, Heaven and Hell
>
> Black Sabath (Heaven and Hell)

In this research we could highlight various aspects relevant to freeway control that the capacity drop phenomenon imposes using different methodologies. In the following section this research is briefly summarized. It is followed by the conclusions achieved. Finally, further questions for future research are discussed

8.1 Summary

In Chapter 4 the coordinated ramp metering problem was analyzed from its equilibrium states assuming the continuous link transmission model dynamics combined with a capacity drop model. Several key aspects were derived: (i) the system performance may depend on the initial state; (ii) an optimal outflow may not be achievable due to operational constraints such as minimum metering rates; and (iii) the system has unstable equilibrium points with capacity drop which is not the case when the capacity drop is inexistent. It has practical implications especially when evaluating the performance of controllers in scenario in which the controller is not capable to effectively interfere in the system.

We also showed that an uncongested equilibrium point always leads to the maximum outflow. This is not the case for congested equilibrium points. This conclusion corroborates with the intuition that it is desirable to keep the freeway uncongested. It was formulated the optimization problem to find the optimal reachable equilibrium state for the cases in which it is not possible to reach an uncongested equilibrium state.

In Chapter 5 we analyzed further the local ramp metering problem considering an ordinary differential equation approximation of the LWR model. We could show the hysteresis cycle that occurs in local ramp metering and derived direct relationship for reachability. The closed loop stability was derived assuming the system is controlled by PI-ALINEA controller for the cases with and without capacity drop. The stability range is shorter when the capacity drop is considered. The results were confirmed with numerical experiments.

In Chapter 6 the local ramp metering control was studied for the particular situation in which the distance between the on-ramp and the lane drop is significantly large. This distance leads to a dead time in the control system which can undermine the system performance. We showed that by incorporating a Smith Predictor in the loop the effect of the dead-time is mitigated and the system responds as if there was no dead-time. In addition, it was possible

to derive the stability range considering LWR model dynamics. The results were confirmed with numerical experiments.

In Chapter 7 the capacity drop phenomenon was studied based on microscopic models as oppose to the previous studies that were based on macroscopic models. We showed that microscopic models can replicate the capacity drop phenomenon with a proper calibration procedure. A multi-objective approach is proposed in order to obtain the car-following and lane-changing parameters. The calibrated model was validated and the results results were satisfactory.

8.2 Conclusions

All results using different methodologies points to the direction that the existence of capacity drop phenomenon imposes additional challenges in the system control. Two important properties of the system was analytically derived from equilibrium state analysis.

First, an uncongested and optimal equilibrium state may not be reachable depending on the arrival demand, minimum metering rates on the on-ramps and initial conditions. This property has practical implications as in specific conditions the controller cannot effectively interfere in the system. Especially for local control, there are some combinations of initial conditions and demand pattern that no ramp metering strategy can be effective. Though that is an intuitive statement, in this research we analytical derived this conclusion from meaningful models and it was provided the theoretical foundation for that intuition.

Second, the capacity drop phenomenon leads to instability of the equilibrium states in a freeway corridor. One of the practical implication of it is that the system becomes more unpredictable as small disturbances or incidents can lead to large variations in the traffic state which turns more difficult the control design. A previous study had investigated the

stability of the freeway equilibrium states without the presence of the capacity drop and had reached opposite conclusion - that the equilibrium states are stable. Therefore, this is a contribution of this work.

A further study into local ramp metering control was performed in which the conclusions are in line with the previous study. In that particular study, it was also investigated the closed loop stability when controlled with the ALINEA algorithm. Following previous conclusions, the stability range is smaller when the capacity drop phenomenon is considered. It was also highlighted the importance of the reachability concept. With numerical experiments we showed that the same control strategy - ALINEA - can lead to a completely different outcomes with a small disturbance in the arrival demand. It imposes further challenges when evaluating control strategies as the very same strategy may distinct outcomes for similar demand patterns.

Nonetheless, it was shown that ALINEA strategy is able to control a single merge bottleneck effectively as long as the input parameters are chosen within the stability range. This can be effective in most cases; however, the stability range becomes smaller when the distance between the on-ramp and the lane drop is large. We showed that by incorporating a Smith Predictor into ALINEA control this effect can be significantly mitigated. The main limitation of the study was the assumption that all traffic state between the on-ramp and the lane drop is known which may not be straightforward to obtain in practice.

This work also touch upon the issue of ramp metering effectiveness. Though a large body of research had confirmed that statement, in recent years it was put in doubt with the shutdown Minnesota shutdown experiment [84]. We could show that ramp metering is in fact effective with few nuances that follows from the reachability property. In general ramp metering can provide overall benefits, but there are cases in which it is not achievable regardless of the control strategy.

From a different perspective, another contribution of this research was the analysis of local ramp metering through microsimulation models. For local ramp metering control it is crucial that the microsimulation model replicates the capacity drop phenomenon accurately and this is one achievement of the dissertation. The ramp metering control analysis conducted based on the microsimulation model corroborated the previous conclusions. Several further aspects can be studied with microscopic models such as taking the input data from loop detectors and the conversion of the metering rate into a discrete number of vehicles per sample time.

8.3 Open Questions for Future Research

Throughout this research some of the impacts of capacity drop phenomenon were described and some strategies to improve freeway control were proposed. Nonetheless, this research also shed some light into new questions that adds to several unanswered question in freeway control.

Most of the analysis and simulation experiments conducted in this research was focused in ramp metering control. There are further ways to perform freeway control that may lead to similar effects such as variable speed limit and congestion pricing. Similar to ramp metering, these strategies may limit the flow in specific locations in order to increase the overall performance. The question for that applications is whether all the conclusions that hold for ramp metering also do in different applications. The reachability results show that ramp metering may not be effective mainly because it cannot control the upstream demand; those applications may be able to control the upstream traffic and therefore not being subject of this drawback. Preliminary analysis as well as recent research points in the direction that ramp metering and variable speed limit has complementary features.

An aspect of utmost importance in traffic flow was not studied into details in this research

is the inherent stochasticity present in traffic flow. Arrival demand is a factor on that, but its relationship with the capacity drop phenomenon was not further investigated in this research. Also, the traffic dynamics is stochastic in nature due to drivers heterogeneity. A possible approach is to use the similar setting of the microsimulation study in Chapter 7 into a Bayesian framework in order to quantify uncertainties. These type of analysis can be a basis for more robust control strategies.

Though the results with ALINEA were encouraging there is still room for improvements. In this research it was kept the approach to feedback density for local ramp metering control, but it may be possible to feedback different variables. Back in 1960s it was reported a successful control strategy by measuring speed at a critical location instead of occupancy. Also, assuming models derived from Newells formulation of the LWR model it may be possible to feedback other variables such as the link queue of the link transmission model.

The methodology and approach used in this research can also be extended to assess potential impacts of automated and connected vehicles. Most of this research it was assumed point measurements; however, recently there has been availability of real time information with moving observers such as vehicle trajectory information. More detailed information can make strategies more robust as the measurements are more accurate and also open opportunities for other strategies that takes as input more detailed data. Regarding automated vehicle, ramp metering strategy can be further improved by smoother merging process. This application is referred to as longitudinal control. The traffic at merge bottleneck may be improved if automated vehicles perform lane changes so as to improve overall performance.

A aspect not explored in this research was the capacity reduction due to higher on-ramp flows. As the flow through the on-ramp is linked with more lane-changing, additional lane-changes may reduce the total outflow. Previous research have consistently reached this conclusion and the microsimulation study also pointed in that direction. This fact is especially important in model-based (rolling horizon) strategies as most of the model does not consider this effect.

Potentially, integrating that aspect into the model may improve performance.

The microsimulation setting that was proposed in this research can be further extended. The calibration procedure can be further refined and tested in different models and locations as well. A specific question that arises is whether it is possible to obtain a set of unique model parameters that reproduce the capacity drop accurately in different cases, namely for different number of freeway lanes. Adding stochasticity to the model, the same setting also can be used to study travel time reliability of a bottleneck.

Bibliography

[1] *Guidelines for design and operating of ramp control systems*. Stanford Research Institute, 1975.

[2] K. Aboudolas, M. Papageorgiou, and E. Kosmatopoulos. Store-and-forward based methods for the signal control problem in large-scale congested urban road networks. *Transportation Research Part C: Emerging Technologies*, 17(2):163–174, 2009.

[3] K. Ahmed, M. Ben-Akiva, H. Koutsopoulos, and R. Mishalani. Models of freeway lane changing and gap acceptance behavior. *Transportation and traffic theory*, 13:501–515, 1996.

[4] P. Allaby, B. Hellinga, and M. Bullock. Variable speed limits: Safety and operational impacts of a candidate control strategy for freeway applications. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):671–680, 2007.

[5] R. Ansorge. What does the entropy condition mean in traffic flow theory? *Transportation Research Part B: Methodological*, 24(2):133–143, 1990.

[6] K. J. Aström and R. M. Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2010.

[7] K. J. Åström and B. Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[8] A. Aw and M. Rascle. Resurrection of" second order" models of traffic flow. *SIAM journal on applied mathematics*, 60(3):916–938, 2000.

[9] J. Bank. Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering. *Transportation Research Record*, 1320:64–69, 1991.

[10] J. H. Banks. Flow processes at a freeway bottleneck. *Transportation Research Record*, (1287), 1990.

[11] J. H. Banks. The two-capacity phenomenon: some theoretical issues. *Transportation Research Record*, (1320), 1991.

[12] M. Brackstone and M. McDonald. Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2(4):181–196, 1999.

[13] C. Cai, C. K. Wong, and B. G. Heydecker. Adaptive traffic signal control using approximate dynamic programming. *Transportation Research Part C: Emerging Technologies*, 17(5):456–474, 2009.

[14] E. F. Camacho and C. B. Alba. *Model predictive control*. Springer Science & Business Media, 2013.

[15] R. C. Carlson, I. Papamichail, M. Papageorgiou, and A. Messmer. Optimal mainstream traffic flow control of large-scale motorway networks. *Transportation Research Part C: Emerging Technologies*, 18(2):193–212, 2010.

[16] R. C. Carlson, I. Papamichail, M. Papageorgiou, and A. Messmer. Optimal motorway traffic flow control involving variable speed limits and ramp metering. *Transportation Science*, 44(2):238–253, 2010.

[17] M. J. Cassidy and R. L. Bertini. Some traffic features at freeway bottlenecks. *Transportation Research Part B: Methodological*, 33(1):25–42, 1999.

[18] M. J. Cassidy and J. Rudjanakanoknad. Increasing the capacity of an isolated merge by metering its on-ramp. *Transportation Research Part B: Methodological*, 39(10):896–913, 2005.

[19] M. J. Cassidy and J. R. Windover. *Methodology for assessing dynamics of freeway traffic flow*. Number 1484. 1995.

[20] E. Chamberlayne, H. Rakha, and D. Bish. Modeling the capacity drop phenomenon at freeway bottlenecks using the integration software. *Transportation Letters*, 4(4):227–242, 2012.

[21] B. Chen and H. H. Cheng. A review of the applications of agent technology in traffic and transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 11(2):485–497, 2010.

[22] D. Chen, S. Ahn, J. Laval, and Z. Zheng. On the periodicity of traffic oscillations and capacity drop: the role of driver characteristics. *Transportation research part B: methodological*, 59:117–136, 2014.

[23] L. Chu, H. X. Liu, J.-S. Oh, and W. Recker. A calibration procedure for microscopic traffic simulation. In *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, volume 2, pages 1574–1579. IEEE, 2003.

[24] K. Chung, J. Rudjanakanoknad, and M. J. Cassidy. Relation between traffic density and capacity drop at three freeway bottlenecks. *Transportation Research Part B: Methodological*, 41(1):82–95, 2007.

[25] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische annalen*, 100(1):32–74, 1928.

[26] R. Courant, K. Friedrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM journal of Research and Development*, 11(2):215–234, 1967.

[27] C. F. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, 28(4):269–287, 1994.

[28] C. F. Daganzo. The cell transmission model, part ii: network traffic. *Transportation Research Part B: Methodological*, 29(2):79–93, 1995.

[29] C. F. Daganzo, V. V. Gayah, and E. J. Gonzales. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. *Transportation Research Part B: Methodological*, 45(1):278–288, 2011.

[30] F. de Souza and W. Jin. System performance and controller design of the pi-alinea ramp metering scheme. In *Transportation Research Board 95th Annual Meeting*, number 16-6183, 2016.

[31] C. Diakaki, M. Papageorgiou, and T. McLean. Integrated traffic-responsive urban corridor control strategy in glasgow, scotland: Application and evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, (1727):101–111, 2000.

[32] J. Dong and H. S. Mahmassani. Stochastic modeling of traffic flow breakdown phenomenon: Application to predicting travel time reliability. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1803–1809, 2012.

[33] R. Dowling, A. Skabardonis, J. Halkias, G. McHale, and G. Zammit. Guidelines for calibration of microsimulation models: framework and applications. *Transportation Research Record: Journal of the Transportation Research Board*, (1876):1–9, 2004.

[34] L. Elefteriadou, A. Kondyli, W. Brilon, F. L. Hall, B. Persaud, and S. Washburn. Enhancing ramp metering algorithms with the use of probability of breakdown models. *Journal of Transportation Engineering*, 140(4):04014003, 2014.

[35] N. H. Gartner. *OPAC: A demand-responsive strategy for traffic signal control*. Number 906. 1983.

[36] N. Geroliminis and J. Sun. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. *Transportation Research Part A: Policy and Practice*, 45(9):966–979, 2011.

[37] P. G. Gipps. A behavioural car-following model for computer simulation. *Transportation Research Part B: Methodological*, 15(2):105–111, 1981.

[38] P. G. Gipps. A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological*, 20(5):403–414, 1986.

[39] T. F. Golob and W. W. Recker. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of transportation engineering*, 129(4):342–353, 2003.

[40] G. Gomes and R. Horowitz. A study of two onramp metering schemes for congested freeways. In *American Control Conference, 2003. Proceedings of the 2003*, volume 5, pages 3756–3761. IEEE, 2003.

[41] G. Gomes and R. Horowitz. Optimal freeway ramp metering using the asymmetric cell transmission model. *Transportation Research Part C: Emerging Technologies*, 14(4):244–262, 2006.

[42] G. Gomes, R. Horowitz, A. A. Kurzhanskiy, P. Varaiya, and J. Kwon. Behavior of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Emerging Technologies*, 16(4):485–513, 2008.

[43] B. Greenshields, W. Channing, H. Miller, et al. A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board, 1935.

[44] H. Hadj-Salem, J. Blosseville, and M. Papageorgiou. Alinea: a local feedback control law for on-ramp metering; a real-life study. In *Road Traffic Control, 1990., Third International Conference on*, pages 194–198. IET, 1990.

[45] F. L. Hall. Traffic stream characteristics. *Traffic Flow Theory. US Federal Highway Administration*, 1996.

[46] F. L. Hall and K. Agyemang-Duah. Freeway capacity drop and the definition of capacity. *Transportation research record*, (1320), 1991.

[47] F. L. Hall and D. Barrow. *Effect of weather on the relationship between flow and occupancy on freeways*. Number 1194. 1988.

[48] K. Han, B. Piccoli, and W. Szeto. Continuous-time link-based kinematic wave model: formulation, solution existence, and well-posedness. *Transportmetrica B: Transport Dynamics*, 4(3):187–222, 2016.

[49] Y. Han, Y. Yuan, A. Hegyi, and S. P. Hoogendoorn. Linear quadratic mpc for integrated route guidance and ramp metering. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 1150–1155. IEEE, 2015.

[50] A. Hegyi, B. De Schutter, and H. Hellendoorn. Model predictive control for optimal coordination of ramp metering and variable speed limits. *Transportation Research Part C: Emerging Technologies*, 13(3):185–209, 2005.

[51] A. Hegyi, B. De Schutter, and J. Hellendoorn. Optimal coordination of variable speed limits to suppress shock waves. *Intelligent Transportation Systems, IEEE Transactions on*, 6(1):102–112, 2005.

[52] P. Hidas. Modelling lane changing and merging in microscopic traffic simulation. *Transportation Research Part C: Emerging Technologies*, 10(5):351–371, 2002.

[53] P. Hidas. Modelling vehicle interactions in microscopic simulation of merging and weaving. *Transportation Research Part C: Emerging Technologies*, 13(1):37–62, 2005.

[54] H.-Y. Jin and W.-L. Jin. Control of a lane-drop bottleneck through variable speed limits. *Transportation Research Part C: Emerging Technologies*, 2014.

[55] W. Jin and M. Zhang. Evaluation of on-ramp control algorithms. *California Partners for Advanced Transit and Highways (PATH)*, 2001.

[56] W.-L. Jin. Continuous kinematic wave models of merging traffic flow. *Transportation research part B: methodological*, 44(8):1084–1103, 2010.

[57] W.-L. Jin. A kinematic wave theory of lane-changing traffic flow. *Transportation research part B: methodological*, 44(8-9):1001–1021, 2010.

[58] W.-L. Jin. A link queue model of network traffic flow. *arXiv preprint arXiv:1209.2361*, 2012.

[59] W.-L. Jin. The traffic statics problem in a road network. *Transportation research part B: methodological*, 46(10):1360–1373, 2012.

[60] W.-L. Jin. Stability and bifurcation in network traffic flow: A poincaré map approach. *Transportation Research Part B: Methodological*, 57:191–208, 2013.

[61] W.-L. Jin. Continuous formulations and analytical properties of the link transmission model. *Transportation Research Part B: Methodological*, 74:88–103, 2015.

[62] W.-L. Jin. On the existence of stationary states in general road networks. *Transportation Research Part B: Methodological*, 81:917–929, 2015.

[63] W.-L. Jin. Point queue models: A unified approach. *Transportation Research Part B: Methodological*, 77:1–16, 2015.

[64] W.-L. Jin. A first-order behavioral model of capacity drop. *Transportation Research Part B: Methodological*, 105:438–457, 2017.

[65] W.-L. Jin, Q.-J. Gan, and J.-P. Lebacque. A kinematic wave theory of capacity drop. *Transportation Research Part B: Methodological*, 81:316–329, 2015.

[66] W.-L. Jin, Q.-J. Gan, and J.-P. Lebacque. A kinematic wave theory of capacity drop. *Transportation Research Part B: Methodological, Accepted Paper*, 2015.

[67] R. E. Kalman. When is a linear control system optimal? *Journal of Basic Engineering*, 86(1):51–60, 1964.

[68] R. E. Kalman et al. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexicana*, 5(2):102–119, 1960.

[69] I. Karafyllis and M. Papageorgiou. Stability results for simple traffic models under pi-regulator control. *IMA Journal of Mathematical Control and Information*, page dnu040, 2014.

[70] A. Kesting and M. Treiber. How reaction time, update time, and adaptation time influence the stability of traffic flow. *Computer-Aided Civil and Infrastructure Engineering*, 23(2):125–137, 2008.

[71] M. M. Khoshyaran and J. P. Lebacque. Capacity drop and traffic hysteresis as a consequence of bounded acceleration. *IFAC-PapersOnLine*, 48(1):766–771, 2015.

[72] S. Kim and B. Coifman. Driver relaxation impacts on bottleneck activation, capacity, and the fundamental relationship. *Transportation Research Part C: Emerging Technologies*, 36:564–580, 2013.

[73] V. Knoop. Introduction to traffic flow theory: An introduction with exercises. 2017.

[74] A. Kondyli, L. Elefteriadou, W. Brilon, F. L. Hall, B. Persaud, and S. Washburn. Development and evaluation of methods for constructing breakdown probability models. *Journal of Transportation Engineering*, 139(9):931–940, 2013.

[75] A. Kotsialos, M. Papageorgiou, M. Mangeas, and H. Haj-Salem. Coordinated and integrated control of motorway networks via non-linear optimal control. *Transportation Research Part C: Emerging Technologies*, 10(1):65–84, 2002.

[76] J. A. Laval and C. F. Daganzo. Lane-changing in traffic streams. *Transportation Research Part B: Methodological*, 40(3):251–264, 2006.

[77] T. Le, H. L. Vu, Y. Nazarathy, Q. B. Vo, and S. Hoogendoorn. Linear-quadratic model predictive control for urban traffic networks. *Transportation Research Part C: Emerging Technologies*, 36:498–512, 2013.

[78] J. Lebacque. The godunov scheme and what it means for rst order tra c ow models. In *Proceedings of the 13th International Symposium on Transportation and Trafc Theory, Lyon, France, July*, volume 2426. Citeseer, 1995.

[79] J. Lebacque. Two-phase bounded-acceleration traffic flow model: analytical solutions and applications. *Transportation Research Record: Journal of the Transportation Research Board*, (1852):220–230, 2003.

[80] J.-P. Lebacque. The godunov scheme and what it means for first order traffic flow models. In *Internaional symposium on transportation and traffic theory*, pages 647–677, 1996.

[81] L. Leclercq. Bounded acceleration close to fixed and moving bottlenecks. *Transportation Research Part B: Methodological*, 41(3):309–319, 2007.

[82] L. Leclercq, J. A. Laval, and N. Chiabaut. Capacity drops at merges: An endogenous model. *Procedia-Social and Behavioral Sciences*, 17:12–26, 2011.

[83] L. Leclercq, F. Marczak, V. L. Knoop, and S. P. Hoogendoorn. Capacity drops at merges: analytical expressions for multilane freeways. *Transportation Research Record: Journal of the Transportation Research Board*, (2560):1–9, 2016.

[84] D. Levinson and L. Zhang. Ramp meters on trial: evidence from the twin cities metering holiday. *Transportation Research Part A: Policy and Practice*, 40(10):810–828, 2006.

[85] D. Liberzon. *Switching in systems and control*. Springer Science & Business Media, 2012.

[86] M. J. Lighthill and G. B. Whitham. On kinematic waves. ii. a theory of traffic flow on long crowded roads. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 229, pages 317–345. The Royal Society, 1955.

[87] H. Lin and P. J. Antsaklis. Stability and stabilizability of switched linear systems: a survey of recent results. *Automatic control, IEEE Transactions on*, 54(2):308–322, 2009.

[88] S. Lin, B. De Schutter, Y. Xi, and H. Hellendoorn. Efficient network-wide model-based predictive control for urban traffic networks. *Transportation Research Part C: Emerging Technologies*, 24:122–140, 2012.

[89] D. J. Lovell and C. F. Daganzo. Access control on networks with unique origin–destination paths. *Transportation Research Part B: Methodological*, 34(3):185–202, 2000.

[90] X.-Y. Lu, P. Varaiya, R. Horowitz, D. Su, and S. Shladover. Novel freeway traffic control with variable speed limit and coordinated ramp metering. *Transportation Research Record: Journal of the Transportation Research Board*, (2229):55–65, 2011.

[91] J.-M. B. M. Papageorgiou, H. Hadj-Salem. Alinea: A local feedback control law for on-ramp metering. *Transportation Research*, pages 58–64, 1991.

[92] L. Maggi, S. Sacone, and S. Siri. Freeway traffic control considering capacity drop phenomena: comparison of different mpc schemes. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 457–462. IEEE, 2015.

[93] H. C. Manual. Hcm2010. *Transportation Research Board, National Research Council, Washington, DC*, 2010.

[94] H. C. Manual et al. Transportation research board. *National Research Council, Washington, DC*, 113, 2000.

[95] A. Messner and M. Papageorgiou. Metanet: A macroscopic simulation program for motorway networks. *Traffic Engineering & Control*, 31(8-9):466–470, 1990.

[96] V. Milanés, J. Godoy, J. Villagrá, and J. Pérez. Automated on-ramp merging system for congested traffic situations. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):500–508, 2011.

[97] P. Mirchandani and L. Head. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies*, 9(6):415–432, 2001.

[98] A. Muralidharan and R. Horowitz. Computationally efficient model predictive control of freeway networks. *Transportation Research Part C: Emerging Technologies*, 58:532–553, 2015.

[99] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part ia discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

[100] G. F. Newell. A simplified theory of kinematic waves in highway traffic, part i: General theory. *Transportation Research Part B: Methodological*, 27(4):281–287, 1993.

[101] J. Normey-Rico, C. Bordons, and E. Camacho. Improving the robustness of dead-time compensating pi controllers. *Control Engineering Practice*, 5(6):801–810, 1997.

[102] J. E. Normey-Rico. *Control of dead-time processes*. Springer Science & Business Media, 2007.

[103] J. E. Normey-Rico and E. F. Camacho. Dead-time compensators: A survey. *Control engineering practice*, 16(4):407–428, 2008.

[104] K. Ogata. *Modern control engineering*. Prentice Hall PTR, 2001.

[105] S. Oh and H. Yeo. Estimation of capacity drop in highway merging sections. *Transportation Research Record: Journal of the Transportation Research Board*, (2286):111–121, 2012.

[106] G. Paesani, J. Kerr, P. Perovich, and F. Khosravi. System wide adaptive ramp metering (swarm). In *Merging the Transportation and Communications Revolutions. Abstracts for ITS America Seventh Annual Meeting and Exposition*, 1997.

[107] M. Papageorgiou. Some remarks on macroscopic traffic flow modelling. *Transportation Research Part A: Policy and Practice*, 32(5):323–329, 1998.

[108] M. Papageorgiou, H. Hadj-Salem, and F. Middelham. Alinea local ramp metering: Summary of field results. *Transportation Research Record: Journal of the Transportation Research Board*, (1603):90–98, 1997.

[109] M. Papageorgiou and A. Kotsialos. Freeway ramp metering: An overview. In *Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE*, pages 228–239. IEEE, 2000.

[110] I. Papamichail, A. Kotsialos, I. Margonis, and M. Papageorgiou. Coordinated ramp metering for freeway networks–a model-predictive hierarchical control approach. *Transportation Research Part C: Emerging Technologies*, 18(3):311–331, 2010.

[111] H. J. Payne. Discontinuity in equilibrium freeway traffic flow. *Transp. Res. Rec*, 1977.

[112] L. A. Pipes. An operational analysis of traffic dynamics. *Journal of applied physics*, 24(3):274–281, 1953.

[113] P. I. Richards. Shock waves on the highway. *Operations research*, 4(1):42–51, 1956.

[114] T. Robič and B. Filipič. Differential evolution for multiobjective optimization. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 520–533. Springer, 2005.

[115] C. Roncoli, M. Papageorgiou, and I. Papamichail. Traffic flow optimisation in presence of vehicle automation and communication systems–part i: A first-order multi-lane model for motorway traffic. *Transportation Research Part C: Emerging Technologies*, 57:241–259, 2015.

[116] R. Scattolini. Architectures for distributed and hierarchical model predictive control–a review. *Journal of Process Control*, 19(5):723–731, 2009.

[117] D. Schrank, B. Eisele, and T. Lomax. Ttis 2012 urban mobility report. *Texas A&M Transportation Institute. The Texas A&M University System*, 2012.

[118] A. Skabardonis, P. Varaiya, and K. Petty. Measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record: Journal of the Transportation Research Board*, (1856):118–124, 2003.

[119] E. Smaragdis, M. Papageorgiou, and E. Kosmatopoulos. A flow-maximizing adaptive local ramp metering strategy. *Transportation Research Part B: Methodological*, 38(3):251–270, 2004.

[120] O. Smith. Closer control of loops with dead time. 1957.

[121] A. Srivastava and W.-L. Jin. A lane changing cell transmission model for modeling capacity drop at lane drop bottlenecks. In *Transportation Research Board 95rd Annual Meeting*, number 16-5452, 2016.

[122] A. Srivastava and W.-L. Jin. Framework for deriving macroscopic demand functions from microscopic acceleration models. *Transportation Research Record: Journal of the Transportation Research Board*, (2623):40–48, 2017.

[123] X. Sun and R. Horowitz. A localized switching ramp-metering controller with a queue length regulator for congested freeways. In *American Control Conference, 2005. Proceedings of the 2005*, pages 2141–2146. IEEE, 2005.

[124] X. Sun and R. Horowitz. Set of new traffic-responsive ramp-metering algorithms and microscopic simulation results. *Transportation Research Record: Journal of the Transportation Research Board*, (1959):9–18, 2006.

[125] X. Sun, L. Munoz, and R. Horowitz. Highway traffic state estimation using improved mixture kalman filters for effective ramp metering control. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 6, pages 6333–6338. IEEE, 2003.

[126] Z. Sun, S. S. Ge, and T. H. Lee. Controllability and reachability criteria for switched linear systems. *Automatica*, 38(5):775–786, 2002.

[127] T. Toledo, H. Koutsopoulos, and M. Ben-Akiva. Modeling integrated lane-changing behavior. *Transportation Research Record: Journal of the Transportation Research Board*, (1857):30–38, 2003.

[128] J. P. van der Gun, A. J. Pel, and B. Van Arem. Extending the link transmission model with non-triangular fundamental diagrams and capacity drops. *Transportation Research Part B: Methodological*, 98:154–178, 2017.

[129] Y. Wang and S. Boyd. Fast model predictive control using online optimization. *Control Systems Technology, IEEE Transactions on*, 18(2):267–278, 2010.

[130] Y. Wang, E. B. Kosmatopoulos, M. Papageorgiou, and I. Papamichail. Local ramp metering in the presence of a distant downstream bottleneck: Theoretical analysis and simulation study. *Intelligent Transportation Systems, IEEE Transactions on*, 15(5):2024–2039, 2014.

[131] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended kalman filter: a general approach. *Transportation Research Part B: Methodological*, 39(2):141–167, 2005.

[132] Y. Wang, M. Papageorgiou, J. Gaffney, I. Papamichail, G. Rose, and W. Young. Local ramp metering in random-location bottlenecks downstream of metered on-ramp. *Transportation Research Record: Journal of the Transportation Research Board*, (2178):90–100, 2010.

[133] J. Wattleworth. Some aspects of macroscopic freeway traffic flow theory. *Traffic Engineering*, 34(2):15–20, 1963.

[134] J. A. Wattleworth. Peak period analysis and control of a freeway system/with discussion. *Highway Research Record*, (157), 1967.

[135] G. Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proc. R. Soc. Lond. A*, 229(1178):317–345, 1955.

[136] S. Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*, volume 2. Springer Science & Business Media, 2003.

[137] W. Xin, P. Michalopoulos, J. Hourdakis, and D. Lau. Minnesota's new ramp control strategy: design overview and preliminary assessment. *Transportation Research Record: Journal of the Transportation Research Board*, (1867):69–79, 2004.

[138] Q. Yang and H. N. Koutsopoulos. A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies*, 4(3):113–129, 1996.

[139] I. Yperman. The link transmission model for dynamic network loading. 2007.

[140] H. Zhang, S. Ritchie, and W. Recker. Some general results on the optimal ramp control problem. *Transportation Research Part C: Emerging Technologies*, 4(2):51–69, 1996.

[141] Z. Zheng. Recent developments and research needs in modeling lane changing. *Transportation research part B: methodological*, 60:16–32, 2014.