## Stochastic models for unsignalized road traffic intersections

Abhishek

**Publication date**
2019
**Document Version**
Final published version
**License**
Other

Stochastic Models for Unsignalized Road Traffic Intersections

Abhishek

# Stochastic Models for Unsignalized Road Traffic Intersections

Abhishek

# Stochastic Models for Unsignalized Road Traffic Intersections

**Abhishek**

**Stochastic Models for Unsignalized Road Traffic Intersections**

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K. I. J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 22 januari 2019, te 10:00 uur

door

**Abhishek**

geboren te Said Alipur, Haryana

**Promotiecommissie**

**Promotores:**

| | |
|---|---|
| Prof. dr. R. Núñez-Queija | Universiteit van Amsterdam |
| Prof. dr. M. R. H. Mandjes | Universiteit van Amsterdam |

**Copromotores:**

| | |
|---|---|
| Dr. ir. M. A. A. Boon | Technische Universiteit Eindhoven |
| Prof. dr. ir. O. J. Boxma | Technische Universiteit Eindhoven |

**Overige leden:**

| | |
|---|---|
| Prof. dr. ir. B. van Arem | Technische Universiteit Delft |
| Prof. dr. R. J. Boucherie | Universiteit Twente |
| Dr. J. L. Dorsman | Universiteit van Amsterdam |
| Prof. dr. N. M. van Dijk | Universiteit van Amsterdam |
| Prof. dr. R. D. van der Mei | Vrije Universiteit Amsterdam |
| Dr. ir. E. M. M. Winands | Universiteit van Amsterdam |

Faculteit der Natuurwetenschappen, Wiskunde en Informatica



This research has been carried out at the Korteweg-de-Vries Institute for Mathematics.

*This thesis is dedicated to my loving daughter*
*Abhidi Jangid.*

# Acknowledgments

First of all, I wish to express my deepest gratitude to my supervisors, Marko Boon, Onno Boxma, Michel Mandjes, and Rudesindo Núñez-Queija for the valuable supervision throughout this project. I think without your constant guidance, constructive comments and suggestions, the project would not have been successful. Sindo, I appreciate your immediate availability for me at times whenever I needed. I still remember my first day in Amsterdam when you were waiting to pick me up at the Schiphol airport and helped me with the essential arrangements. I also remember how our short meetings ended up in two to three hours long discussions next to your whiteboard, even though these meetings were usually started at the end of the day. Your guidance and support have been instrumental in helping me toward making my professional decisions, and for that, I am forever grateful to you. I would like to thank Michel for teaching me the art of presenting the longer proofs in a compact and comprehensible fashion. A special thanks to Onno for his guidance specifically in writing Chapter 1 of this thesis. It was truly a great pleasure to learn from you on the academics writing and the scientific community in general. Marko, you always help me in improving my presentation and numerical skills. I learned a lot from you specifically while working on numerical results. I am truly grateful also to the members of my defense committee, Bart van Arem, Richard Boucherie, Jan-Pieter Dorsman, Nicolaas van Dijk, Rob van der Mei and Erik Winands for taking their time to read this thesis.

I am grateful to Evelien Wallet, Monique Onderwater and Marieke Kranenburg for taking care of organizational and administrative issues. I would like to express a special gratitude to Marieke for arranging the financial support from the Networks during the very first month of my PhD.

I would like to thank my colleagues that I met or worked with them during the PhD: Alex, Brendan, Danilo, Daphne, David, Jacob, Jacobien, Jarno, Julia, Kayed, Liron, Masha, Madelon, Mariska, Masoumeh, Mayank, Murtuza, Nicos,

Peter, Raymond, Reinier, Sara, Wendy, Wessel, and Zijian. I am very thankful to my officemates Jacobien, Liron, Masoumeh and Raymond for the delightful conversations we had. I would like to express a special gratitude to Danilo, Julia and Nicos for being a part of my wedding ceremony in India and making the moment more memorable with their blessings. Many thanks to Alex and Mariska for helping me during shifting to the new house, and for enjoying the Indian delicacies at my place. A special thanks to Mariska for giving the initial feedback on the Dutch summary of the thesis.

I would like to extend my thanks to my friends for their continuous motivation, encouragement and technical support. Murtuza, it is an immense pleasure to have a friend like you. Our late night calls discussions were always helpful and fruitful for me. You were always a support system whether it was personal or professional. I am particularly grateful to Vivek and Anuja for being a family. It was always fun being with you guys whether it was having food together, photo sessions or birthday celebrations. And of course, your support at the hospital during my daughter's birth (Abhidi) was truly grateful. Thanks to Anuj, Nitin and Pushkar for being such good friends.

Finally, I would like to thank my parents (Beg Raj & Shakuntla), parents-in-law (Rajender Parshad & Sharda), brothers-in-law (Amit & Hitkul). Without their moral support and continuous encouragement, I would not accomplish this work. A special thanks to my sister and niece (Bharti & Sumaira) for supporting us during Abhidi's birth. I would like to dedicate this thesis to my loving daughter Abhidi. You make me so happy just being around. Thank you for being such a wonderful and lovely child. Last but not least, I am truly grateful to my wife, Divya, for her unconditional love, care, and support. You have no idea how much my life has changed since we met. You made me smile in a special kind of way whenever I was tensed during the PhD. Apart from being a soulmate, you are a very good friend. You even helped me with a figure in the thesis and read the thesis even though, being a management student, it was not so easy for you to understand it technically.

# Contents


**Page**

*Acknowledgments* vii

**1 Introduction** 1
- 1.1 General introduction and motivation . . . 1
  - 1.1.1 Signalized intersections . . . 3
  - 1.1.2 Unsignalized intersections . . . 3
- 1.2 Modeling and performance analysis of unsignalized priority-controlled intersections . . . 5
- 1.3 The single server queue . . . 11
- 1.4 Literature overview . . . 18
  - 1.4.1 Gap acceptance models . . . 18
  - 1.4.2 Models based on additive conflict flows (ACF) methods, empirical regression methods, and simulation methods 20
- 1.5 Main results of the thesis . . . 22
- 1.6 Organization of the thesis . . . 24

**2 Congestion analysis of unsignalized intersections** 27
- 2.1 Introduction . . . 28
- 2.2 Model description . . . 29
- 2.3 Analysis of queue lengths and delays . . . 30
- 2.4 Impact of driver behavior on capacity . . . 34
- 2.5 Numerical results and practical examples . . . 37
  - 2.5.1 Example 1: ordering of the capacities . . . 37
  - 2.5.2 Example 2: paradoxical behavior . . . 37
  - 2.5.3 Example 3: the impact of resampling . . . 39
  - 2.5.4 Example 4: impatience . . . 41
- 2.6 Discussion and conclusion . . . 43

**3 Markov platooning** 45
3.1 Introduction . . . 46
3.2 Model description . . . 47
3.3 Capacity . . . 47
3.4 Numerical results . . . 54
3.4.1 Example 1: the impact of Markov platooning . . . 54
3.4.2 Example 2: platoon lengths . . . 56
3.5 Discussion and conclusion . . . 58

**4 Generalized $M^X$/semi-Markov/1 queue** 59
4.1 Introduction . . . 60
4.2 Model description . . . 62
4.2.1 The $M^X$/semi-Markov/1 queue . . . 63
4.2.2 General model . . . 64
4.3 The queue length distribution at departure epochs . . . 66
4.3.1 Steady-state analysis . . . 67
4.3.2 Transient analysis . . . 72
4.4 Poisson batch arrivals: stationary queue length at arrival and arbitrary epochs . . . 76
4.5 The queueing model with two customer types: departure epochs 77
4.6 Numerical results . . . 82
4.6.1 Example 1 . . . 83
4.6.2 Example 2 . . . 84
4.6.3 Example 3 . . . 86
4.6.4 Example 4: Transient-state analysis . . . 86
4.7 Discussion and conclusion . . . 87

**5 Extension with exceptional first service** 89
5.1 Introduction . . . 90
5.2 Model description . . . 90
5.3 Stationary queue length analysis . . . 94
5.3.1 Stationary queue length analysis: departure epochs . 94
5.3.2 Special cases . . . 101
5.3.3 Stationary queue length analysis: arrival and arbitrary epochs . . . 103
5.4 Waiting time and sojourn time . . . 103
5.5 Applications to road traffic . . . 107

5.6 Numerical results . . . . . 115
5.6.1 Example 1: the impact of batch arrivals . . . . . 115
5.6.2 Example 2: the impact of Markov platooning . . . . 116
5.7 Discussion and conclusion . . . . . 117

**6 Heavy-traffic analysis of the $M^X/SM/1$ queue** 119
6.1 Introduction . . . . . 119
6.2 Heavy-traffic analysis . . . . . 120
6.3 Numerical example . . . . . 129
6.4 Discussion and conclusion . . . . . 131

**7 Generalized gap acceptance models** 133
7.1 Introduction . . . . . 134
7.2 Model description . . . . . 137
7.3 Queue length analysis . . . . . 139
7.3.1 Preliminaries . . . . . 139
7.3.2 A queueing model with semi-Markovian service times and exceptional first service . . . . . 141
7.4 Capacity . . . . . 144
7.5 Numerical results . . . . . 154
7.5.1 Example 1 . . . . . 154
7.5.2 Example 2 . . . . . 157
7.5.3 Example 3 . . . . . 158
7.6 Discussion and conclusion . . . . . 160

**8 Generalized gap acceptance models with Markov platooning** 161
8.1 Model description . . . . . 161
8.2 Capacity . . . . . 163
8.3 Discussion and conclusion . . . . . 167

*Summary* 169

*Samenvatting* 173

*Publications of the author* 177

*About the author* 179

*References* 181

Chapter 1

# Introduction

## 1.1 General introduction and motivation

This PhD thesis is devoted to the mathematical analysis of some specific forms of road traffic congestion. In urban areas across the world, road traffic congestion is one of the main transportation issues. One can hardly find a day, in which one can travel by car from one place to another without experiencing traffic congestion. Road traffic congestion can be either recurring or non-recurring. Recurring congestion is based on daily events in fixed time periods such as commuting by car to and from work in peak hours. Non-recurring congestion occurs due to incidents or bad weather conditions such as accidents, roadworks, ice storms, heavy rain, heavy snow, strong wind, etc. Therefore it is highly unpredictable, making it more difficult to let the system run on full capacity.

Road traffic congestion has a huge societal, environmental and economic impact. It increases the travel time of each individual vehicle, which leads to aggravation and additional energy (fuel) consumption. Fuel combustion is one of the main reasons for increasing $CO_2$ emissions in the atmosphere [9], which has a negative impact on the environment and health [95, 116]. In 2009, the estimated cost of road traffic congestion in the UK was about €24 billion, or 1.6% of GDP [28]. And in the same year, the total cost of congestion in the European Union member states was about €111 billion, which was 1% of the European GDP. In 2017, the total cost of congestion for road transport in the Netherlands exceeded €1 billion. Clearly, the cost of congestion is huge and has an adverse effect on the world economy. See Fleuren [45, Chapter 1] for an extensive list of references regarding the environmental and economic impact of road traffic congestion.

Traffic congestion can be mitigated by either building new infrastructures, or improving (adding capacity to) the existing infrastructures, or reducing the demands on roads. The existing infrastructures can be improved by adding new lanes or traffic lights, or placing a roundabout. The demands on roads can be reduced by providing enough information about accidents, traffic jams and road congestions so that drivers can choose alternative routes, or by promoting alternative modes of transportation, e.g., ride-sharing, walking, cycling, public transport, etc. Building new infrastructures like new roads or transportation networks, is generally more expensive in terms of urban space and financial resources, than modifying the existing infrastructures or reducing the demands on roads.

One of the earliest known attempts to mitigate road traffic congestion was at the old London bridge at the end of the $18^{th}$ century. The London municipality then constructed a novel control system by installing traffic lights as well as dividing the road into two different lanes: one for vehicles and another for pedestrians. In the beginning of the $19^{th}$ century, traffic signs such as stop or yield signs were introduced to mitigate traffic congestion at intersections — places where two or more roads cross or merge at angles. Thus, it became clear that traffic congestion can be reduced with the help of a proper control system.

Using traffic control technology, intersections can be classified into two

main categories: signalized and unsignalized. In the subsequent sections, we give a brief overview of these two types of intersections.

### 1.1.1 Signalized intersections

In signalized intersections, vehicles that approach intersections are controlled by traffic lights to provide safe and efficient movement. Signalized intersections are commonly used when traffic volumes of vehicles on roads connected to intersections are more than moderate. At modern signalized intersections, there are three types of traffic lights: green, yellow, and red. Green light allows the traffic to move ahead in the direction shown by the green signal. Yellow light alerts drivers to slow down and to stop at the intersection. If the light turns red, then vehicles have to stop immediately and wait for the light to turn green before traveling through the intersection. Therefore, these lights work in a fixed cyclic order through green, yellow and red at regular and synchronized intervals to reduce congestion for a consistent traffic flow.

Traffic signals operate in either fixed time or dynamic time or some combination of the two. Fixed time signals are programmed according to a fixed cycle length to clear off the traffic, and thus these are useful when the traffic volume of vehicles on roads is high. On the other hand, dynamic time signals regulate the phase and cycle according to fluctuating traffic demand using detectors installed at the approaches. Hence they are better able to respond to the actual traffic situation.

### 1.1.2 Unsignalized intersections

Unsignalized intersections are not regulated by means of traffic lights, and are generally used when traffic volumes of vehicles on roads that are connected to intersections range from low to moderate. Unsignalized intersections can be either priority-controlled, uncontrolled, or roundabouts.

#### Priority-controlled intersections

Priority-controlled intersections are regulated with the help of priority (traffic) signs. The priority signs on roads can be either ‘Stop’ signs or ‘Yield’ signs, which are described below.

- **Stop sign-controlled**: When drivers of vehicles face stop signs on roads, the drivers must stop completely before approaching intersections, and can only proceed when no other traffic is present at the intersections. When stop signs are placed at each and every road that is connected to an intersection, each driver must first stop completely, and then yield the right of way to the driver of the vehicle on its right.

- **Yield sign-controlled**: Drivers facing yield signs must first slow down, and stop only if necessary to let vehicles from other directions proceed first. If there are safe gaps available on major roads, drivers immediately proceed through intersections.

We can conclude that the stop sign requires drivers to stop completely, but the yield sign allows drivers to proceed immediately, without stopping, through the intersection if a safe gap is available on the major road.

Unsignalized priority-controlled intersections generally consist of major and minor roads, in which vehicles on the major roads cross the intersections without observing the streams of vehicles on the minor roads, and vehicles on the minor roads are controlled by stop signs or yields signs. Therefore, vehicles arriving on the major roads theoretically do not experience any delay while crossing the intersections. Due to traffic signs, vehicles arriving on the minor roads can be required to wait before entering the intersections, which leads to queues of low-priority vehicles on the minor roads. As a consequence, the performance of the intersections is fully determined by the characteristics of these queues of vehicles on the minor roads.

### Uncontrolled intersections

Uncontrolled intersections have no traffic signs or lights. Therefore, vehicles approaching intersections follow the basic 'right-of-way' rules of the road in countries with driving on the right hand side of the road. More specifically, whenever two or more vehicles approach the intersection at the same time, each vehicle must yield the right of way to the driver of the vehicle on their right. These uncontrolled intersections are commonly found in rural or residential areas, where traffic volumes of vehicles on roads are very low.

### Roundabouts

A roundabout is a type of circular intersection, in which drivers travel almost continuously in one direction around a central island. Traffic approaching the roundabout is not required to stop, but must yield the right of way to the traffic on the roundabout. In roundabouts, there is no need to specify roads connected to intersections as major or minor roads. When most vehicles that approach intersections from each direction are not heavy (like trucks, private buses, etc.), and like to take a right turn (in right-hand driving countries), then roundabouts can be used to control traffic at intersections. However, modern roundabouts are designed in such a way that vehicles of all sizes, including trucks and private buses, can comfortably use them.

In this dissertation, our main focus is on the mathematical analysis of stochastic models for unsignalized priority-controlled intersections. Therefore, we shall give a global introduction to modeling and performance analysis of these intersections in the next section.

## 1.2 Modeling and performance analysis of unsignalized priority-controlled intersections

In this section, we consider an unsignalized priority-controlled intersection, where vehicles arriving on the major road have priority over the vehicles on the minor road (see Figure 1.1). More specifically, the major road vehicles cross the intersection without 'observing' the low-priority stream, whereas vehicles on the minor road enter the major road without interrupting the stream on the major road.

Models for unsignalized priority-controlled intersections typically contain a stochastic element to capture the inherent uncertainty about future arrivals of vehicles/drivers and their characteristics such as inter-vehicle-type heterogeneity (due to the vehicle's features, in particular the size and acceleration speed of the vehicle), inter-driver heterogeneity (due to the driver characteristics such as age, gender, years of driving experience), intra-driver heterogeneity (due to the driver's behavior in different situations), etc. In the existing literature, various quantitative methods have been used to study stochastic models

for unsignalized priority-controlled intersections. The two most common procedures are simulation methods and methods based on mathematical models. Simulation methods are based on simulation models that measure the impact of uncertain inputs of a system through the use of computer software; see [87] for an overview of simulation of traffic models. The simulation models are useful in dealing with uncertainty, and with complex relations between the behaviors of drivers that seem difficult to analyze by conventional mathematical methods (see [17, 93]). Several types of methods based on mathematical models have been developed to study the capacity of the minor road. The most commonly used methods are so-called gap acceptance procedures in the framework of queueing theory (see [60, 101]), additive conflict flows techniques (see [6, 15, 16, 121, 122]), and empirical procedures in the framework of regression theory (see [14, 18, 65]). We discuss these models in more details later in this chapter.

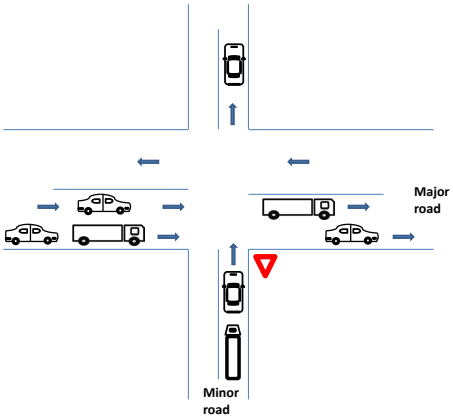



**Figure 1.1.** An unsignalized intersection.

Based on the level of detail, stochastic models are usually categorized into three main types: microscopic, mesoscopic, and macroscopic. A microscopic model takes into account behavior of individual vehicles, in particular, interactions between two subsequent vehicles approaching the intersection and entering the intersection. On the other hand, a macroscopic model formulates

relationships among traffic flow characteristics like density, flow, mean speed of a traffic stream, etc., without considering the behavior of the individual vehicles. In particular, we can say that microscopic models deal with vehicles, whereas macroscopic models deal with traffic flows. Mesoscopic models strike a balance between microscopic and macroscopic models. Mesoscopic models do not distinguish or trace individual vehicles, but consider an aggregated behavior of different groups of vehicles. We refer to Chapter 2 of Baër [8] for discussions of microscopic, mesoscopic, and macroscopic models of highway traffic flows.

At the end of this section, we briefly explain two other applications such as freeways [37, 38, 78] and pedestrian crossings [103, 113], where all these stochastic models are applicable.

In this dissertation, unsignalized intersections are modeled by microscopic mathematical stochastic models, namely *gap acceptance models in the framework of queueing theory*. In gap acceptance models, the low-priority drivers have to either accept (i.e., use) or reject the available gaps on the major road, at the intersection. A minimum gap that a low-priority driver accepts, is referred to as a *critical* gap (or headway). Therefore, when the time gap between two subsequent vehicles on the major road is at least the critical gap, say $T$, then one vehicle, from the minor road, enters the major road.

There are three main components in stochastic models of unsignalized intersections, using the gap acceptance method: (i) arrival processes, (ii) gap acceptance behavior, (iii) merging behavior. We now successively discuss each of the three components.

**(i) Arrival processes.** In general, the arrival processes of vehicles on the major and the minor roads are represented by stochastic processes, i.e., collections of random variables. In this dissertation, the arrival processes on both these roads will be either a Poisson process, a Markov modulated Poisson process, or a batch Poisson process.

- **Poisson process**: Let $N(t)$ denote the total number of 'events' that occur by time t. $\{N(t), t \geq 0\}$ is said to be a Poisson process having rate $\lambda$, $\lambda > 0$, if (see Ross [89, p. 313])

  (i) $N(0) = 0$.

(ii) The process has independent increments, i.e., the number of arrivals in any given time interval is independent of the number of arrivals in any other non-overlapping time interval.
(iii) The number of events in any interval of length $t$ is Poisson distributed with mean $\lambda t$, i.e., for all $s, t \geq 0$

$$\mathbb{P}(N(t+s) - N(s) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad \text{for } n = 0, 1, 2, \ldots \quad (1.1)$$

Now we mention two important and mathematically attractive features of the Poisson process: 1) interarrival times are independent and exponentially distributed with parameter $\lambda$, and 2) if $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson processes with rates $\lambda_1$ and $\lambda_2$ respectively, then $\{N(t) = N_1(t) + N_2(t), t \geq 0\}$ is a Poisson process with rate $\lambda = \lambda_1 + \lambda_2$. Poisson processes occur naturally when there are many candidates to arrive, but they all have a very small probability of actually arriving in the next small time unit, and they operate independently of each other. In such a situation, the time until the next event does not depend on the length of time since the occurrence of the last event. This is often referred to as the *memoryless property*: with $I$ an interevent time, $\mathbb{P}(I > t + s | I > t) = e^{-\lambda s} = \mathbb{P}(I > t), \forall s, t \geq 0$ (see Ross [90, p. 35]). The exponential distribution is known to be the only continuous-time memoryless distribution.

- **Markov modulated Poisson process (MMPP)**: MMPP is a generalization of the Poisson process. The MMPP can be constructed by varying the arrival rate of a Poisson process according to a $d$-state irreducible continuous-time Markov chain which is independent of the arrival process. When the Markov chain is in state $i$, arrivals occur according to a Poisson process of rate $q_i$. The MMPP is parameterized by the $d$-state continuous-time Markov chain with infinitesimal generator $M$, $M = (\mu_{ij})_{i,j=1}^{d}$, and the $d$ Poisson arrival rates $q_1, q_2, \ldots, q_d$ (see [44]). The MMPP is a well-studied arrival process which is generally used to model dependencies between interarrival times. The flexibility of the MMPP allows us to vary the interarrival times in such a way, that we can create platoons, single arrivals, or combinations thereof.

- **Batch Poisson process**: Like the MMPP, the batch Poisson process is also a generalization of the Poisson process. In the batch Poisson

process, vehicles (or customers) arrive in batches according to a Poisson process with rate $\lambda$. Vehicles within a batch are assumed to be ordered arbitrarily, and the batch sizes form a sequence of (positive, integer-valued) independent and identically distributed (i.i.d.) random variables. A batch could be formed due to vehicles clustering behind one slow driver, or due to platoons of vehicles arriving from an upstream intersection. Of course, one might argue about the assumption of exponential interarrival times of batches, but we refrain from that discussion in this thesis.

**(ii) Gap acceptance behavior.** The gap acceptance behavior of drivers has a strong impact on the performance as well as safety of unsignalized intersections. Due to the driver characteristics such as gender, age, years of driving experience, as well as the vehicle's features, in particular the size and acceleration speed of the vehicle, critical gaps differ from vehicle to vehicle, and even from one driver to another driver. A driver may behave differently on different occasions under similar circumstances. Furthermore, the driver may at some stage accept a gap shorter than a gap that he had earlier rejected. We distinguish three types of *behavior* when making the gap acceptance decision.

- $\mathbf{B}_1$: The critical headway $T$ is deterministic, and uniform across all low-priority vehicle drivers.
- $\mathbf{B}_2$: The vehicle driver at the front of the queue *resampling* $T$ (from a given distribution) at any new attempt (where an 'attempt' amounts to comparing this sampled $T$ to the gap between the two subsequent vehicles that he is currently observing). It is also known as inconsistent gap acceptance behavior.
- $\mathbf{B}_3$: Each driver selects a random value of $T$, but then sticks to that same value for all attempts, rather than resampling these. Therefore, it is known as consistent gap acceptance behavior.

The first model $\mathrm{B}_1$ is the most simplistic. In $\mathrm{B}_2$, we take into account that there is heterogeneity in the driver behavior: one could expect a broad range of 'preferences', ranging from very defensive to very reckless drivers. $\mathrm{B}_3$ reflects *persistent* differences between drivers.

For each of the aforementioned behavior types, we also consider a variant that includes *impatience*. Due to the higher traffic rate on the major road, or long delay on the minor road, the low-priority driver, at the head of the queue, may get more and more impatient after each rejection. With impatience, the critical headway decreases after each failed attempt, reflecting the impatience of drivers, resulting in the willingness to accept smaller and smaller gaps. In more detail, we define a critical headway $T_j$ for the $j$-th attempt to enter the major road ($j = 1, 2, \dots$). Note that, depending on the distributions of $T_1, T_2, \dots$, situations might occur in model $B_2$ where $T_{i+1} > T_i$, despite $T_{i+1}$ being stochastically smaller than $T_i$. This is a typical feature of the model with resampling.

**(iii) Merging/crossing behavior.** In reality, drivers always demand, for safety purposes, a larger critical gap to enter the major road, than the time actually needed to cross (or merge on) the major road. As a consequence, they use only a part of the critical gap to cross the major road, and the remainder can be utilized by subsequent drivers on the minor road. In the literature, that first part of the critical gap is usually referred to as the *merging* time of the driver. It can be easily seen in practical situations that the merging time depends on the behavior of the driver as well as the size of the vehicle that he is driving. For example, trucks require larger merging times than motor bikes.

Now that we have discussed the three main components of stochastic models of unsignalized intersections, we can turn to the performance analysis of such models.

## Performance analysis

As the high-priority vehicles on the major road do not experience any interference from the low-priority vehicles, the performance of the vehicles on the major road is not affected by the vehicles on the minor road, and thus the system's performance is fully determined by the characteristics of the queue of low-priority vehicles on the minor road. One of the most important performance measures on the minor road is the capacity of the minor road [13], which is defined as the maximum possible number of vehicles per time unit that can pass in the long run through an intersection from the minor road. Other relevant performance measures are the queue length and the delay on

the minor road.

The performance analysis of such a probabilistic model lies in the scope of queueing theory (see, e.g., the survey [117]). In such a queueing setting, vehicles are viewed as customers, and a specified segment of road as a server providing service to the customers. Then a careful study of the resulting queueing model should lead to mathematical expressions for key performance measures like the capacity of the minor road, the queue length and the delay on the minor road.

### Applications to other contexts of road traffic

We conclude this section by noting that stochastic models for unsignalized intersections can be applied in other contexts as well, e.g. when analyzing freeways [37, 38, 78] and pedestrian crossings [103, 113]. In a freeway merging, we consider the situation of a ramp (or side-road street) and a highway (or freeway), where the drivers on the highway have priority over the drivers on the ramp, while vehicles on the ramp wish to merge into the stream of traffic on the highway. This situation is similar to that of unsignalized intersections, and can be analyzed using the same (or similar) mathematical models. In a pedestrian crossing, pedestrians arrive randomly at a curb of a road. Vehicles arriving on the road have priority over the pedestrians at the curb. The main difference with the aforementioned models is that pedestrians make a crossing decision immediately after their arrival at the curb, and cross the road either individually or in groups (which is usually referred to as batch (bulk) services in the corresponding queueing models).

In the next section, we shall give a global introduction to queueing theory.

## 1.3 The single server queue

Queues are a part of everyday life, i.e., they are everywhere: in call centers, in supermarkets, in hospitals, in road traffic networks, in communication networks, at airports, at computer systems, etc. Congestion typically occurs when there is a demand for a service, and there is a service facility that can offer that service but that has limited capacity. In general, people don't like to wait and delays can be very costly. Hence, from both a social and an economic perspective, queues are undesirable. So, models and techniques are required

to model, analyze, and alleviate such queueing problems. Since the pioneering work of the Danish engineer A.K. Erlang [39], who developed a queueing model for a telephone system, queueing theory has developed into a large and mature subdiscipline of applied probability and stochastic operations research.

A queueing model is a mathematical description of a service facility. It involves *servers* and *customers* arriving for service, waiting for service if it is not immediately available, and leaving the system after being served. The customers arriving at a queue may be vehicles waiting in lines to enter an intersection, telephone calls waiting to be answered, patients waiting to see a doctor, messages waiting to be transmitted, etc. And the servers that provide service to the arriving customers could be checkout counters in supermarkets, agents (operators) at call centers, doctors and/or nurses in hospitals, conflict areas (or specified road segments) at road traffic intersections, etc.

In this section, we give a brief overview of the fundamental concepts of queueing theory, because they play an essential role in this dissertation. Several books and survey papers summarize the main contributions made in the area of queueing theory. A summary of the evolution of queueing theory until the mid-1980's can be found in Cohen and Boxma [30], while more recent developments are discussed in Harchol-Balter [53]. We refer the reader to Kleinrock [67], Cohen [29], Takagi [100], Nelson [79] for extensive expositions of queueing systems. The ultimate goal of the performance analyses of queues is to come to a better design, fine-tuning and operation of a service facility. Kleinrock [67] also pays much attention to these aspects.

A standard and basic framework of a queueing model, consisting of a single server, is described in Kendall's notation $G/G/1$, where the first $G$ in the notation stands for a general distribution of the independent sequence of interarrival times, the second $G$ stands for a general distribution of the independent sequence of service times, and the number 1 refers to a single server (see Figure 1.2). Customers who cannot immediately be taken into service join the queue or waiting room. The waiting room is often assumed to have infinite capacity. If the system can only hold $N$ customers, it is referred to as $G/G/1/N$.
An example of the $G/G/1$ queue is the $M/G/1$ queue, where $M$ stands for *memoryless*, which means that the interarrival times are exponentially dis-

tributed. The $M/G/1$ model is a natural model and mathematically more accessible than the $G/G/1$ queue. Another special case of the $G/G/1$ queue is the $M/M/1$ queue, where both $M$ stand for *memoryless*, which means that the interarrival and the service times are exponentially distributed.

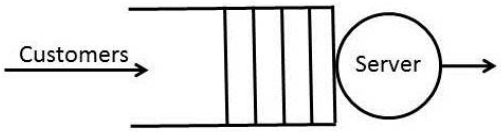



**Figure 1.2.** The $G/G/1$ queue.

A service discipline must be specified in order to describe the behavior of the queueing system. There are many service disciplines studied in queueing theory, namely FCFS (First Come First Served), LCFS (Last Come First Served), SIRO (Service in Random Order), PS (Processor Sharing), etc. Furthermore, one also often distinguishes several customer classes, these classes having different priority levels. For example, high-priority customers always have priority of service above low-priority customers, and then one can distinguish, e.g., between preemptive and non-preemptive priorities. In the case of preemptive priority, the service of a low-priority customer is interrupted when a high-priority customer arrives; it is only resumed (or perhaps repeated) when there is no longer any high-priority customer in the system. In the case of non-preemptive priority, a low-priority customer is allowed to finish its service when a high-priority customer arrives. Furthermore, one can distinguish between static and dynamic priorities. In the case of static priorities, some customers have a fixed priority over other customers. However, dynamic priorities may alternate over time, as in so-called polling systems, or the largest job may have the highest priority, etc. For further details, we refer the reader to Harchol-Balter [53] and Cohen [29].

The models that we analyze in this thesis show strong resemblance with priority queueing models. Due to the complex interaction of different traffic flows (i.e., vehicles on the major road and vehicles on the minor road), our

models can not be cast into one of the standard priority models, but in our analysis we are inspired by the analysis methods of priority models.

Key performance measures in most queueing systems are the number of customers in the queue or in the system, and the waiting time of a customer or the time spent by a customer in the system (sojourn time). In view of the fact that the key input variables (interarrival and service times) are random variables, the key performance measures are also random variables and we need to study their probability distributions.

Now the important question arises whether the number of customers in the system doesn't grow indefinitely? Put differently, what is the stability condition of a queueing model? Let $\rho = \frac{\lambda}{\mu}$, where $1/\lambda$ and $1/\mu$ are, respectively, the mean interarrival and the mean service times of customers in the $G/G/1$ queueing model. It has been proved that when $\rho < 1$, the number of customers in the system stays finite in the $G/G/1$ queue (see Cohen [29, Section II.5.7]). And, when $\rho \uparrow 1$, the queueing system is in the heavy-traffic regime (see [66]). Working in the heavy-traffic regime and using appropriate heavy-traffic asymptotics often allows one to get more explicit and insightful results. In more detail, heavy-traffic analysis of a queueing system provides exact insight in how to rescale the queue so that it converges to a non-degenerate distribution.

Next, we briefly discuss two important results of queueing theory that are often used to determine the above mentioned key performance measures such as waiting and sojourn times of customers, and the number of customers in the system.

- **Little's law**: It relates the mean number of customers in a stable system and the mean sojourn time. More specifically, it states that the mean number of customers in a stable system is equal to the average number of customers arriving in the system per time unit, multiplied by the mean sojourn time of a customer (see [73]). In fact, Keilson and Servi [64] show that under some additional assumptions, one can even relate the *distribution* of the number of customers in the system and the sojourn time.
- **Poisson Arrivals See Time Averages (PASTA) property**: It states

that with Poisson arrivals, the fraction of arrivals finding a queue in some state equals the fraction of time the queue is in that state (see [118]). The PASTA property is very useful to compute the distribution of the queue length in queueing models in which customers arrive according to a Poisson process, e.g., $M/G/1$, $M/M/1$.

Firstly, we will determine the distribution of the number of customers in the $M/M/1$ queueing system, in which customers arrive according to a Poisson process with rate $\lambda$, and the service times of customers are exponentially distributed with parameter $\mu$. Let $X(t)$ denote the number of customers in the system at time $t$ such that

$$\pi_n(t) = \mathbb{P}(X(t) = n) \quad \text{for } n = 0, 1, 2, \dots. \tag{1.2}$$

Notice that $X(t)$ jumps up by amount 1 at an arrival time and jumps down by amount 1 at a departure time. And the arrival and the departure times are exponentially distributed with parameters $\lambda$ and $\mu$ respectively. Therefore, $X(t)$ is a continuous time Markov chain with state space $\{0, 1, 2, \dots\}$ and transition rates

$$q_{ij} = \begin{cases} \lambda, & j = i + 1 \\ \mu, & j = i - 1 \\ -(\lambda + \mu), & j = i \\ 0, & \text{otherwise.} \end{cases}$$

The equilibrium (balance) equations of the Markov chain are given by

$$\begin{aligned} \lambda\pi_0 &= \mu\pi_1, \\ (\lambda + \mu)\pi_n &= \lambda\pi_{n-1} + \mu\pi_{n+1}, \quad n = 1, 2, \dots, \end{aligned}$$

where $\pi_n = \lim_{t\to\infty} \pi_n(t)$ is the limiting (invariant) distribution.

After solving the equilibrium equations and using the normalization equation, $\sum_{n=0}^{\infty} \pi_n = 1$, the limiting probability, $\pi_n$, is shown to be

$$\pi_n = (1 - \rho)\rho^n, \tag{1.3}$$

provided $\rho = \frac{\lambda}{\mu} < 1$.

Furthermore, the probability generation function of the number of customers in the system is derived as

$$\begin{aligned}\tilde{X}(z) &= \sum_{n=0}^{\infty} \pi_n z^n = \sum_{n=0}^{\infty} (1-\rho)\rho^n z^n \\ &= \frac{1-\rho}{1-\rho z}, \quad \text{for } |z| \leq 1. \end{aligned} \tag{1.4}$$

Next, we will derive the distribution of the number of customers in the $M/G/1$ queueing system, in which the service times are i.i.d. random variables. Due to the general service-time distribution, $X(t)$ does not form a continuous time Markov chain. To find the distribution of the number of customers, we study the system immediately after departures. To do so, we define the recurrence relation:

$$X_n = \begin{cases} X_{n-1} - 1 + A_n & \text{if } X_{n-1} \geq 1 \\ A_n & \text{if } X_{n-1} = 0 \end{cases}, \quad n = 1, 2, 3, \ldots, \tag{1.5}$$

where $X_n$ is the number of customers immediately after the departure epoch of the $n$-th customer, and $A_n$ is the number of arrivals during the service time of the $n$-th customer. Here the sequence $\{A_n\}_{n\geq 1}$ is independent of the sequence $\{X_n\}_{n\geq 1}$. Due to the Poisson arrivals, it follows that $A_1, A_2, \ldots$ are i.i.d. random variables and, as a consequence, $\{X_n\}_{n\geq 0}$ forms a Markov chain. Since we look at the system specific moments, namely immediately after departures, this Markov chain is referred to as an embedded Markov chain.

From the recurrence relation (1.5), we find for the probability generating functions:

$$\begin{aligned}\mathbb{E}[z^{X_n}] =& \mathbb{E}[z^{X_{n-1}-1+A_n} | X_{n-1} \geq 1]\mathbb{P}(X_{n-1} \geq 1) \\ &+ \mathbb{E}[z^{A_n} | X_{n-1} = 0]\mathbb{P}(X_{n-1} = 0) \\ =& \frac{1}{z}\mathbb{E}[z^{A_n}]\mathbb{E}[z^{X_{n-1}} | X_{n-1} \geq 1]\mathbb{P}(X_{n-1} \geq 1) + \mathbb{E}[z^{A_n}]\mathbb{P}(X_{n-1} = 0) \\ =& \frac{1}{z}\mathbb{E}[z^{A_n}](\mathbb{E}[z^{X_{n-1}}] - 1) + \mathbb{E}[z^{A_n}]\mathbb{P}(X_{n-1} = 0). \end{aligned} \tag{1.6}$$

It can be proven (see Cohen [29, Chapter II.4]) that $\{X_n\}_{n\geq 0}$ has a steady-state distribution iff $\rho < 1$. Let $X$ be a random variable with that steady-state

distribution. We introduce the following notations to study the system in steady state:

$$F(z) = \mathbb{E}[z^X], \tag{1.7}$$

$$A(z) = \mathbb{E}[z^{A_n}], \quad |z| \leq 1, \tag{1.8}$$

with $F(0) = \mathbb{P}(X_n = 0)$.

In steady state, Equation (1.6) leads to the following equation

$$F(z) = \frac{1}{z} A(z)(F(z) - F(0)) + A(z)F(0),$$

which further implies that

$$F(z) = \frac{F(0)(z-1)A(z)}{z - A(z)}.$$

It is noted that $F(1) = 1$. As a consequence, using L'Hôpital's rule, $F(0) = 1 - \mathbb{E}[A]$, where $\mathbb{E}[A] = A'(1)$.

As the customers arrive according to a Poisson process with parameter $\lambda$, $A(z)$ is easily seen to be

$$A(z) = \tilde{G}(\lambda(1-z)), \tag{1.9}$$

where $\tilde{G}(.)$ is the LST of the service-time distribution. Finally, the probability generating function of the number of customers immediately after a departure is given by

$$F(z) = \frac{(1-\rho)(z-1)\tilde{G}(\lambda(1-z))}{z - \tilde{G}(\lambda(1-z))}, \tag{1.10}$$

where $\rho = \frac{\lambda}{\mu} = \mathbb{E}[A]$ and $1/\mu$ is the mean service time of a customer.

Note that, in steady state, the number of transitions per time unit from state $n$ to $n+1$ (that correspond to arrivals) is equal to the number of transitions per time unit from state $n+1$ to $n$ (that correspond to departures). Therefore, the distribution of the number of customers already in the system just before

an arrival must coincide with the distribution of the number of customers in the system immediately after a departure. And, from the PASTA property, the number of customers in the system at an arbitrary time equals in distribution the number of customers at an arrival epoch.

Now, we observe that the system satisfies the assumptions required for applying the distributional form of Little's law, and obtain the sojourn time of an arbitrary customer. The customers in the system at the departure epoch of a customer are exactly those that arrive during the sojourn time of that customer. As a consequence, we can write the following relation

$$F(z) = \tilde{S}(\lambda(1-z)), \tag{1.11}$$

where $\tilde{S}(.)$ is the LST of the sojourn time distribution. Hence $\tilde{S}(s)$ immediately follows from Equations (1.10) and (1.11).

Note that the sojourn time of a customer equals the sum of the waiting time and the service time of that customer. Therefore, the LST of the waiting time distribution is given by

$$\tilde{W}(s) = \frac{\tilde{S}(s)}{\tilde{G}(s)}. \tag{1.12}$$

## 1.4 Literature overview

In this section, we focus on the literature on unsignalized priority-controlled intersections. In that literature, the most commonly used procedures are methods based on gap acceptance models, which have the longest history so far, but other methods, e.g., simulation methods, empirical regression techniques and additive conflict flows techniques, have been employed as well. We review gap acceptance models in Subsection 1.4.1, and in Subsection 1.4.2 we briefly discuss other models.

### 1.4.1 Gap acceptance models

One of the first studies, using gap acceptance models, for unsignalized intersections was by Tanner [101], who studied an unsignalized intersection consisting of major and minor roads, where the traffic on the major road, which consists of a single stream of vehicles, has absolute priority over vehicles on the minor

road. Tanner first determined the mean delay of the low-priority vehicles in the case of constant critical gap and move-up time, and characterized the capacity of the minor road as the arrival rate of the low-priority vehicles, at which the mean delay grows without bound. Tanner's model has been generalized in various ways [22, 27, 58, 59, 111, 114, 126] by allowing random critical gaps and move-up times, also analyzing performance measures such as the queue length and the waiting time on the minor road. In [126], Yeo and Weesakul assumed random critical gaps and move-up times in such a way that a fixed critical gap is assigned to each low-priority driver, which varies from one driver to another. This type of behavior is referred to as consistent gap acceptance behavior in the literature. In that model, they also incorporated driver impatience behavior. Moreover, they were the ones to notice that Tanner's model is a special case of a queueing model called $M/G2/1$ (a term probably introduced by Daganzo [33]), i.e., a generalization of the $M/G/1$ queueing model in which the first customer of each busy period has an exceptional service time, and using such a queueing model, the transforms of the delay and the queue length on the minor road were obtained. In [57], Hawkes considered inconsistent gap acceptance behavior, in which the driver samples independently a new critical gap from a given gap acceptance distribution function at each new attempt, to enter the major road. In addition, he introduced a hesitation time for non-queuers, which is equivalent to the move-up time for queuers, and determined the distribution of delay of the minor road vehicles. In [31], Cowan extended Tanner's results [101] by considering a generalized stochastic structure for the arrival streams, and developed a delay formula for the minor road vehicles.

Tanner's model [101] has been further extended to include multiple lanes on the major road. In [56], Hawkes considered two opposite streams of vehicles on the major road, and a single lane of vehicles on the minor road, which wish to take either a left or a right turn on the major road (or move straight ahead). In this model, Hawkes assumed that all right turning vehicles require a larger critical gap than the left turning vehicles in countries where cars drive on the left, and determined the distribution of the delay of the minor road vehicles. In [102], Tanner considered $n$ lanes of traffic on the major road, and determined the capacity of the minor road. For further studies on models describing multi-lane situations, we refer the reader to Troutbeck [109], Wegmann [112], and Wu [123].

In [22], Catchpole and Plank briefly described four types of gap acceptance models with their relevant existing literature: homogeneous and consistent, homogeneous and inconsistent, nonhomogeneous and consistent, nonhomogeneous and inconsistent. In this model, they first included both nonhomogeneous (mixed) vehicle types and inconsistent gap acceptance driver behavior in the stream on the minor road, and then determined the capacity of the minor road for a general distribution of the critical gap and headways on the major road, and a fixed move-up time for all low-priority vehicles. Furthermore, they proved that the capacity equals a weighted harmonic mean of the capacities for each individual vehicle type, rather than a weighted arithmetic mean of the individual capacities as claimed by Evans et al. [40]. In [111], Wegmann further generalized the model studied by Catchpole and Plank [22] by including the gap-block process on the major road and impedance effects, and determined a capacity formula for unsignalized intersections.

To make the model more realistic, Heidemann and Wegmann [60], relying on results by Tanner [101], proposed a general framework based on gap-block models. In such models, vehicles form platoons which arrive according to a Poisson process. The lengths of these platoons are i.i.d. random variables with a general distribution, which can be suitably chosen such that it matches real-life clustering behavior. Moreover, Heidemann and Wegmann added a stochastic dependence between the critical gap and the merging (move-up) time, and studied the minor road as an M/G2/1 queue, to determine the queue length, the delay and the capacity on the minor road. A further generalization, dividing the time scale of the major stream into four regimes (namely free space (no vehicles), free flow (single vehicles), bunched traffic (platoons of vehicles), and queueing) was investigated by Wu [119]. By conditioning on the current regime, he applied the framework of [60] to set up a heuristic argument that provides a more general capacity formula that is valid under all four regimes. In [27], Cheng and Allam reviewed the stochastic modeling of delay and capacity at unsignalized priority intersections.

### 1.4.2 Models based on additive conflict flows (ACF) methods, empirical regression methods, and simulation methods

In the last subsection, we discussed the existing literature on gap acceptance models. In this subsection, we briefly discuss models based on additive conflict

flows (ACF) methods, empirical regression methods, and simulation methods.

**ACF methods**: In [16], Brilon and Wu stated a few drawbacks of gap acceptance methods. One of them was that gap acceptance methods fail when pedestrians or cyclists share the use of the intersection. Therefore, ACF methods [6, 15] that are based on graph theory are used to deal with such situations. The ACF method was first developed by Gleue [51] for signalized intersections. Later, Wu [122, 121] modified it for All-Way-Stop-Controlled (AWSC) intersections, where traffic on each road connected to intersections is controlled by a stop sign, and thus recommended a practical procedure for the determination of capacity at AWSC intersections. In [16], Brilon and Wu used the same procedure to study Two-Way Stop-Controlled (TWSC) intersections that consist of major and minor roads, in which vehicles on the major road have priority over vehicles on the minor road, and a stop sign is used to control the traffic on the minor road. In [120], Wu further modified and extended the existing ACF procedure for TWSC in such a way that microscopic parameters such as critical gaps and follow-up times from gap acceptance methods can directly be used in the extended ACF method. For a further overview of the existing literature on ACF methods, we refer to Asaithambi and Anuroop [6], and Prasetijo and Ahmad [85].

**Empirical regression methods**: Empirical regression methods, which are particularly used in the United Kingdom, are based on an entirely different approach. Instead of building a queueing model that captures the essential features of the gap acceptance process, these methods rely on large amounts of collected data that measure the relevant performance measures, such as delay and service time, plus additional characteristics such as minor and major street flow rate, sizes of the accepted gaps, and directional movement (through-vehicle or turning vehicle). Using a standard regression analysis, it is determined which of these factors influence the capacity and a linear model is developed to describe this relation. There is considerable literature on the empirical regression methods for unsignalized intersections. In [65], Kimber and Coombe computed the capacity of major/minor priority junctions using empirical methods. Kyte et al. [70] developed a set of preliminary empirical models to estimate delay and capacity at TWSC intersections. In [69], Kyte et al. studied the empirical capacity and delay models for AWSC intersections. For a further overview of the existing literature on models based on empirical

regression methods, we refer to Brilon et al. [14] and Brilon [17, 18].

**Simulation models**: When mathematical models cannot provide satisfactory practical solutions, or when certain features of road traffic, such as time-varying traffic demands, multi-lane traffic situations, different vehicle types, and aggressive and impatience behaviors of drivers, are difficult to describe in analytical ways, simulation models can be considered to be effective and useful tools to determine the performance measures of unsignalized intersections. In the existing literature, several different types of simulation models such as vehicle/car-following models, cellular automata models, and multi-agent models have been studied. Vehicle-following models are based on the idea that vehicles follow each other, interact with each other, and thus take into account individual vehicle data, e.g., size, headway, velocity, acceleration, together with interactions in the vehicle-following process (see [11, 50, 63, 83, 92]). Cellular automata models divide a space into an uniform grid with cells that can either be occupied by a vehicle or be empty, where variables at each cell are updated simultaneously in discrete time steps according to a set of local interaction rules (see [68, 72, 91, 124]). Multi-agent models are based on the interaction and coordination of agents in order to achieve their goals, where agents may represent vehicles, drivers, pedestrians, intersection controllers, or other traffic participants, and goals can be destinations and crossing/merging intersections/roads (see [10, 26, 35, 36]). In simulation models, the required input parameters are estimated according to video recordings of the actual traffic flow at several intersections. Relationships between delay, capacity, queue length, major and minor road flow rate, are developed using regression or other empirical evaluation techniques. For more details, we refer the reader to Caliendo [21], Chan and Teply [24], Grossmann [52], Sayed et al. [93], Tian et al. [105], Tracz and Gondek [107], and Zhang [127].

## 1.5 Main results of the thesis

The goal of this thesis is to analyze an unsignalized priority-controlled intersection where drivers on the major road have priority over the drivers on the minor road. The main contributions of the thesis are the following:

- We first reveal some interesting results which were not observed in the existing literature. In particular, we show a strict ordering among

the capacities for the existing gap acceptance models for unsignalized intersections: the models with constant critical gaps ($B_1$), inconsistent gap acceptance behavior ($B_2$), and consistent behavior ($B_3$). More specifically, it is shown that $B_2$ has a larger capacity than $B_1$, and the capacity of $B_3$ is the smallest (with the mean critical headway of models $B_2$ and $B_3$ chosen equal to the deterministic critical headway of model $B_1$). This implies that randomness within the critical gaps in model $B_2$ has a positive impact on the capacity.

- To make the existing models more realistic, we include two additional features into these models: *impatience* of the drivers that are waiting to cross the major road (the longer the driver has to wait, the lower the critical headway), and Markov *platooning* on the major road (modeling the fluctuations in the traffic density on that road). In the existing literature, platooning on the major road was analyzed using a so-called gap-block process [60]. In such a process, vehicles form platoons which arrive according to a Poisson process, and the lengths of these platoons are i.i.d. random variables with a general distribution. We introduce a novel way to model different traffic-flow regimes on the major road, using a well-established method to model dependence between successive interarrival times. We assume that the arrival process on the major road is modeled by a Markov modulated Poisson process (MMPP). In an MMPP, at time $t$, the time till the next arrival is exponentially distributed with mean $1/q_i$, where $i$ is the state of the background process at time $t$. It is noticed that the strict ordering among the capacities of the gap acceptance models that was observed in the case without impatience, is no longer preserved after including the impatience of drivers. Moreover, we numerically show that when drivers rapidly decrease the critical headway after each rejection of the gap between two subsequent vehicles on the major road, the capacity first decreases up to a certain threshold, as a function of the flow rate on the major road, but it starts increasing after that threshold. Based on the numerical results, it can be observed that platooning has a positive effect on the capacity of the minor road for given mean rate, but only for models $B_1$ and $B_3$. In a model with inconsistent behavior, it depends on the model parameters whether platooning increases or decreases the capacity.

- When there is Markov platooning on the major road, there are fluctuations in the traffic density, which lead to a dependency between two subsequent entering vehicles, from the minor road, on the major road. To capture these kinds of dependencies, we investigate a single server queueing model with batch arrivals, in which the sequence of service times is governed by a modulating Markovian process. One of the findings of this queueing model is that when the variance of the number of arrivals during a service time increases, due to the dependence between service times, the mean number of customers may decrease.

- We generalize the queueing model with batch arrivals and correlated service times into a model in which the first customer of each busy period has an exceptional service time, and we compute the stationary queue length distribution and the steady-state waiting time and sojourn time distributions of an arbitrary customer. Furthermore, we show that the heavy-traffic distribution of the scaled stationary queue length is exponential.

- We further extend all three gap acceptance models $\mathrm{B}_1$, $\mathrm{B}_2$, and $\mathrm{B}_3$ into a more realistic model that incorporates multiple classes of gap acceptance behavior (including the impatience of the drivers) as well as merging behaviors of drivers, where the remaining (unused) part of a critical headway can be used by subsequent drivers. An important feature of this model is that it allows us to distinguish between different driver types and/or different vehicle types that play a crucial role in determining the capacity of the minor road.

## 1.6 Organization of the thesis

The remainder of the thesis is structured as follows. Chapter 2 introduces *impatience* of the drivers in the existing gap acceptance models $B_1$ up to $B_3$ for unsignalized intersections, where vehicles arrive according to a Poisson process on the minor road as well as on the major road. We then determine the Laplace-Stieltjes transformation (LST) of the service time, which is used to derive the transforms of the distributions of the queue length and the delay on the minor road. Furthermore, the capacity of the minor road is also obtained via the LST of the service time.

In Chapter 3, Markov platooning on the major road is modeled by a Markov modulated Poisson process. We then develop methods, using the renewal reward theorem, to investigate the capacity of the minor road for the models $B_1$, $B_2$, and $B_3$. Lastly, using several practical examples, we present numerical results for the capacity of the minor road.

In Chapter 4, we describe a single server queueing model with batch arrivals and semi-Markov service times. Although set up with general applications in mind, this framework turns out to be particularly useful in road traffic models, where the semi-Markov (correlated) service times can be used to model clustering of vehicles on the major road, to differentiate between multiple types of driver behavior, or to account for heterogeneous traffic. We first discuss how the transient and the stationary probability generating functions of the number of customers in the system immediately after a departure can be obtained for this general framework. Next, for the case of batch Poisson arrivals, we derive the generating functions of the stationary number of customers at an arbitrary instant, at batch arrival instants, and at customer arrival instants. In the end, we numerically explore the impact of the dependencies between two subsequent service times on the mean number of customers in the system.

The queueing framework studied in Chapter 4 is generalized in Chapter 5 by allowing that the first customer in a busy period has a different service-time distribution than regular customers served in the busy period. Firstly, we provide the analysis to compute the stationary queue length distribution immediately after a departure. Subsequently, we use that result to derive the steady-state waiting time and sojourn time distributions of an arbitrary customer in the case of batch Poisson arrivals, which depend on its position in the batch, as well as on the type of service of the first customer in the batch. In the end, we apply this extension of the queueing model, to road traffic situations involving multiple streams of conflicting traffic. In particular, we use it in the context of gap acceptance models, studied in Chapter 3, where low-priority traffic needs to cross (or, merge with) other traffic of higher priority at an unsignalized intersection.

In Chapter 6, we investigate the queueing system studied in Chapter 5, in the heavy-traffic regime, i.e., the expected number of arrivals during the service

time (of a regular customer) is approaching 1. We show that the heavy-traffic distribution of the scaled stationary queue length is exponential.

In Chapter 7, we introduce a generalized gap acceptance model for unsignalized intersections, where vehicles arrive according to a Poisson process on the major road and according to a *batch* Poisson process on the minor road. The generalized model consists of multiple classes of gap acceptance behavior (including impatience of the drivers) as well as merging behavior of drivers. Due to the merging behavior of drivers, the merging time depends on the profile of the driver, and is less than the critical headway. As a consequence, the remaining part of a critical headway can be used by subsequent drivers. Therefore, we can use the generalized queueing framework introduced in Chapter 5, to obtain the queue length distribution on the minor road. Subsequently, the capacity of the minor road is derived from the expectation of the service time for waiting vehicles. Finally, we present some numerical examples in order to get more insight into the model.

To see the impact of the Markov platooning on the capacity, we further generalize the gap acceptance model in Chapter 8. We here assume that vehicles arrive according to a Markov modulated Poisson process on the major road. We determine the LST of the service time of an arbitrary low-priority driver, which is further used to derive the capacity of the minor road.

Chapter 2

# Congestion analysis of unsignalized intersections

In this chapter, we consider an unsignalized intersection, described in Section 1.2, where a stream of cars arriving on the primary (major) road has priority over the stream of cars on the secondary (minor) road; see Figure 1.1. Cars belonging to the latter stream cross the primary road if the gap between two subsequent cars on the primary road is larger than their critical headways, which are minimum gaps that low-priority drivers accept. Questions that naturally arise are: given the arrival pattern of the cars on the primary road, what is the maximum arrival rate of low-priority cars such that the number of such cars remains stable? In the second place, what can be said about the delay experienced by a typical car at the secondary road? In this chapter, we address such issues by considering a model that sheds light on the dynamics of the considered unsignalized intersection. The model, which is of a queueing-theoretic nature, reveals interesting insights into the impact of the

user behavior on the above stability and delay issues. First, we obtain new results for the aforementioned model with driver impatience. Secondly, we reveal some surprising aspects that have remained unobserved in the existing literature so far, many of which are caused by the fact that the capacity of the minor road cannot be expressed in terms of the *mean* gap size; instead more detailed characteristics of the critical headway distribution play a role.

## 2.1 Introduction

A common situation in any road traffic network is that of an unsignalized intersection that is used by two traffic streams which have different priorities. The priority class consists of cars on the main road which arrive at the intersection, according to some inherently random process; the fact that they have priority essentially means that they cross the intersection without observing the low-priority stream on the secondary road. Cars of the low priority stream, however, only cross when the duration (in time) of a gap between two subsequent cars passing by is sufficiently large, i.e., larger than a, possibly car-specific, threshold $T$.

As the high-priority cars do not experience any interference from the low-priority cars, the system's performance is fully determined by the characteristics of the queue of low-priority cars on the secondary road. A first issue concerns the stability of this queue: for what arrival rate of low-priority cars can it be guaranteed that the queue does not explode? Formulated differently: what is the capacity of the minor road? The answer to this question evidently depends on the distribution of the gaps between subsequent cars on the primary road. In addition, however, the specific 'preferences' of the low-priority car drivers play a crucial role: how does the individual car driver choose the threshold $T$ which determines the minimal gap needed. In the existing literature, various models have been studied, the simplest variant being the situation in which all low-priority drivers use the same deterministic $T$ (referred to as model $B_1$ in this thesis, cf. Section 1.2). A second, more realistic, model ($B_2$) is the one in which the driver draws a random $T$ from a distribution, where the $T$ is resampled for any new attempt; this randomness models the heterogeneity in the preferences of the low-priority car drivers. A further refinement is a model ($B_3$) in which the driver sticks to the same (random) $T$ for all his attempts.

Gap acceptance models are mainly applied to unsignalized intersections (cf. [22, 60, 101]), pedestrian crossings (cf. [77, 103]), and freeways (cf. [37, 38]). Although the gap acceptance process in these three application areas exhibits similar features, the queueing aspects are fundamentally different. In this chapter, we focus on motorized vehicles, but all results regarding the capacity of the minor road can easily be applied to pedestrian crossings or freeway merging. Heidemann and Wegmann [60] give an excellent overview of the existing results in gap acceptance theory, including the three types of user behavior ($B_1$ up to $B_3$) that were discussed. In this chapter, we first include *impatience* of the drivers, waiting to cross the major road, in the model. This phenomenon, which is indeed encountered in practice [4], has been studied before (cf. [37, 38, 114, 126]), but (to the best of our knowledge) not yet in the context of model $B_2$, where randomness is encountered in the critical headway $T$. Secondly, we reveal some surprising aspects that have remained unobserved in the existing literature so far. In particular we show there is a strict ordering in the capacities, resulting from the different types of driver behavior. This is far from obvious, since it is known that the capacity is *not* in terms of the mean quantity $\mathbb{E}[T]$, but more precise distributional information of the random variable $T$ is needed. Perhaps counterintuitively, when comparing two gap time distributions $T_1$ and $T_2$ one could for instance encounter situations in which $\mathbb{E}[T_1] < \mathbb{E}]T_2]$, but in which still the capacity under $T_1$ is smaller than the one under $T_2$.

The remainder of this chapter is structured as follows. In the next section, we describe in more detail the aforementioned types of driver behavior, when waiting for a gap on the main road, and build a model which is used in Section 2.3 to analyze queue lengths and delays. In Section 2.4, we study the traffic congestion on the minor road. In Section 2.5 we present numerical results for several practical examples, focusing on some surprising, paradoxical features that one might encounter. Finally, we give several concluding remarks.

## 2.2 Model description

We consider an intersection used by two traffic streams, both of which wishing to cross the intersection. There are two priorities: the car drivers on the major road have priority over the car drivers on the minor road. The high-priority car drivers arrive at the intersection according to a Poisson process of intensity

$q$, meaning that the inter arrival times between any pair of subsequent cars are exponentially distributed with mean $1/q$. The low-priority car drivers, on the minor road, cross the intersection as soon as they come across a gap with duration larger than $T$ between two subsequent high-priority cars, commonly referred to as the *critical headway*. On the minor road cars arrive according to a Poisson process with rate $\lambda$.

Above we were intentionally imprecise regarding the exact definition of the criterion based on which the low-priority cars decide to cross. Importantly, in this chapter, we focus on the three types of 'behavior' ($B_1$, $B_2$ and $B_3$) when making this decision, which are discussed in Section 1.2. For each of these behavior types, we also consider a variant that includes impatience. With impatience, the critical headway decreases after each failed attempt, reflecting the impatience of drivers, resulting in the willingness to accept smaller and smaller gaps. In more detail, we define a critical headway $T_j$ for the $j$-th attempt to enter the main road ($j = 1, 2, \dots$). Note that, depending on the distributions of $T_1, T_2, \dots$, in model $B_2$ situations might occur where $T_{i+1} > T_i$, despite $T_{i+1}$ being stochastically smaller than $T_i$. This is a typical feature of the model with resampling. Exact details regarding the manner in which impatience is incorporated will be discussed in more detail in the next section.

We conclude this section by noting that in case the primary road actually consists of two lanes that have to be crossed (without a central reservation), with cars arriving (potentially in opposite directions) at Poisson rates (say) $q^{\leftarrow}$ and $q^{\rightarrow}$, our model applies as well, as an immediate consequence of the fact that the superposition of two Poisson processes is once again a Poisson process with the parameter $q := q^{\leftarrow} + q^{\rightarrow}$; see also the discussion in [119, Section 5].

## 2.3 Analysis of queue lengths and delays

The three models $B_1$, $B_2$ and $B_3$ can be analyzed using queueing-theoretic techniques. Since results for the variants without impatience have been known in the existing literature (see, for example, Heidemann and Wegmann [60] for an overview), we will carefully study the formulas for the capacity of the minor road under different circumstances, which turns out to lead to a few interesting new insights. In addition, we will consider the situation with impatient drivers.

The underlying idea is that the model can be cast in terms of an $M/G/1$ queue, i.e., a queue with Poisson arrivals, general service times (which will have their specific form for each of the models $B_1$ up to $B_3$) and a single server. We start our exposition by introducing some notation. In the first place, we let $X_n$ denote the number of cars in the queue on the minor road when (*right after*, that is) the $n$-th low-priority car crosses the primary road; in addition, $T_n^{\#}$ is the time that this happens.

In our analysis, we need to distinguish between times that the queue on the secondary road is empty or not. When $X_{n-1} \geqslant 1$ we define the inter departure time between the $(n-1)$-st and $n$-th car from the secondary road as $G^{(n)} := T_n^{\#} - T_{n-1}^{\#}$. It is clear that for all rules $B_1$ up to $B_3$, we can write $X_n$ as

$$X_n = \begin{cases} X_{n-1} - 1 + A_n, & \text{if } X_{n-1} \geqslant 1 \\ A_n, & \text{if } X_{n-1} = 0 \end{cases}$$

where $A_n$ is the number of arrivals, on the minor road, during the departure time $G^{(n)}$.

It is directly verified that, due to the exponentiality assumptions imposed, $\{X_n, n = 1, 2, \ldots\}$ is a Markov chain, the random variables $\{G^{(n)}\}$ are independent identically distributed (and have a definition for each of the models $B_1$ up to $B_3$). In fact, the process $\{X_n, n = 1, 2, ...\}$ has the dynamics of an M/G/1 queue length process at departures where $G^{(n)}$ corresponds to the $n$-th service time. As a consequence, we have that $X_n$ has a stationary distribution which is uniquely characterized through its probability generating function (directly following from the celebrated Pollaczek–Khinchine formula (1.10))

$$F(z) = \mathbb{E}[z^X] = (1-\rho)\frac{(1-z)\tilde{G}(\lambda(1-z))}{\tilde{G}(\lambda(1-z)) - z} \tag{2.1}$$

where $\tilde{G}(s) := \mathbb{E}[e^{-sG}]$ and $\rho := \lambda\mathbb{E}[G]$. The queue is stable when $\rho < 1$.

Furthermore, the mean queue length is given by

$$\mathbb{E}[X] = \frac{\lambda^2\mathbb{E}[G^2]}{2(1-\rho)} + \rho. \tag{2.2}$$

Let $W$ and $S$, respectively, be the stationary waiting time and sojourn time corresponding to an arbitrary arriving car on the secondary road. Since

$\{X_n, n = 1, 2, \ldots\}$ is an $M/G/1$ queue length process at departures, the Laplace-Stieltjes transformations (LSTs) of $W$ and $S$ are obtained respectively from Equations (1.12) and (1.11) as

$$\tilde{W}(s) = \frac{s(1-\rho)}{\lambda\tilde{G}(s) + s - \lambda}, \quad \tilde{S}(s) = \frac{s(1-\rho)\tilde{G}(s)}{\lambda\tilde{G}(s) + s - \lambda}.$$

Now we study the impact of the three types of the driver's behavior on stability and delay.

**$B_1$ (constant gap):** Every driver on the minor road needs the same constant critical headway $T_j$ for the $j$-th attempt to enter the main road ($j = 1, 2, \ldots$). We assume that $T_1 \geq T_2 \geq \cdots \geq T_{\min}$. In this case, each user tries a number of attempts (use the memoryless property!), with success probability $\mathbb{P}(\tau_q > T_j) = e^{-qT_j}$ for the $j$-th attempt, where $\tau_q$ is an exponential random variable with mean $1/q$. The Laplace transform of the 'service time', $\mathbb{E}[e^{-sG}]$, hence follows from

$$\sum_{k=0}^{\infty} \left( \prod_{j=1}^{k} \mathbb{E}[e^{-s\tau_q} 1_{\{\tau_q < T_j\}}] \right) \mathbb{E}[e^{-sT_{k+1}} 1_{\{\tau_q \geqslant T_{k+1}\}}], \tag{2.3}$$

where $1_{\{.\}}$ denotes the indicator function.
Elementary computations yield

$$\begin{aligned} \mathbb{E}[e^{-s\tau_q} 1_{\{\tau_q < T_j\}}] &= \frac{q}{q+s} \left(1 - e^{-(q+s)T_j}\right), \\ \mathbb{E}[e^{-sT_{k+1}} 1_{\{\tau_q \geqslant T_{k+1}\}}] &= e^{-(q+s)T_{k+1}}, \end{aligned}$$

which leads to

$$\mathbb{E}[e^{-sG}] = \sum_{k=0}^{\infty} \left(\frac{q}{s+q}\right)^k e^{-(s+q)T_{k+1}} \prod_{j=1}^{k} (1 - e^{-(s+q)T_j}). \tag{2.4}$$

It is also verified that

$$\mathbb{E}[G] = \sum_{k=0}^{\infty} e^{-qT_{k+1}} \left[ \frac{k}{q} + T_{k+1} - \sum_{i=1}^{k} \frac{T_i e^{-qT_i}}{1 - e^{-qT_i}} \right] \prod_{j=1}^{k} (1 - e^{-qT_j}).$$

These expressions simplify considerably in the case without driver impatience, i.e. $T_1 = T_2 = \cdots =: T$, namely

$$\mathbb{E}[e^{-sG}] = \frac{(s+q)e^{-(s+q)T}}{s+qe^{-(s+q)T}},$$
$$\mathbb{E}[G] = \frac{e^{qT}-1}{q}.$$

**$B_2$ (sampling per attempt):** With this behavior type, which is also sometimes referred to as "inconsistent behavior", every car driver samples a random $T_j$ for its $j$-th 'attempt' (where 'attempt' corresponds to comparing the resulting $T_j$ with the gap between two subsequent cars on the major road). Although things are slightly more subtle than for $B_1$, with the $T_j$ being random, we can still use the memoryless property of the gaps between successive cars on the major road in combination with the independence of the $T_j$, to argue that expression (2.3) is also valid for this model, but now the $T_j$ are random, leading to

$$\mathbb{E}[e^{-s\tau_q}1_{\{\tau_q<T_j\}}] = \frac{q}{q+s}\left(1-\mathbb{E}[e^{-(q+s)T_j}]\right),$$
$$\mathbb{E}[e^{-sT_{k+1}}1_{\{\tau_q\geqslant T_{k+1}\}}] = \mathbb{E}[e^{-(q+s)T_{k+1}}],$$

which, after substitution in (2.3), leads to

$$\mathbb{E}[e^{-sG}] = \sum_{k=0}^{\infty}\left(\frac{q}{s+q}\right)^k \mathbb{E}[e^{-(s+q)T_{k+1}}]\prod_{j=1}^{k}(1-\mathbb{E}[e^{-(s+q)T_j}]).$$

The mean service time $\mathbb{E}[G]$ readily follows:

$$\mathbb{E}[G] = \sum_{k=0}^{\infty}\left[\mathbb{E}[T_{k+1}e^{-qT_{k+1}}] + \mathbb{E}[e^{-qT_{k+1}}]\left(\frac{k}{q}-\sum_{i=1}^{k}\frac{\mathbb{E}[T_ie^{-qT_i}]}{1-\mathbb{E}[e^{-qT_i}]}\right)\right] \times \prod_{j=1}^{k}(1-\mathbb{E}[e^{-qT_j}]).$$

Again, these expressions simplify considerably in the case without driver impatience:

$$\mathbb{E}[e^{-sG}] = \frac{(s+q)\mathbb{E}[e^{-(s+q)T}]}{s+q\,\mathbb{E}[e^{-(s+q)T}]}, \qquad \mathbb{E}[G] = \frac{1-\mathbb{E}[e^{-qT}]}{q\mathbb{E}[e^{-qT}]}.$$

Observe that $B_1$ (with deterministic $T_j$) is a special case of $B_2$.

**$B_3$ (sampling per driver):** In this variant, sometimes referred to as "consistent behavior", every car driver samples a random $T_1$ at his first attempt. This (random) value determines the complete sequence of $T_2, T_3, \ldots$. To model this kind of behavior, we introduce a sequence of functions $h_j(\cdots)$, for $j = 1, 2, \ldots$, such that the critical headway $T_j$ is defined as $T_j := h_j(T_1)$, where $h_1$ is the identity function. This model has the advantages of differentiating between various types of drivers, while still modeling a form of impatience.

Since only the first critical headway is random, we obtain the LST of $G$ by conditioning on the value of $T_1$, and using (2.4). We obtain

$$\mathbb{E}[e^{-sG}] = \mathbb{E}\left[\sum_{k=0}^{\infty}\left(\frac{q}{s+q}\right)^k e^{-(s+q)T_{k+1}}\prod_{j=1}^{k}(1-e^{-(s+q)T_j})\right],$$

where $T_j := h_j(T_1)$. The mean service time is

$$\mathbb{E}[G] = \mathbb{E}\left[\sum_{k=0}^{\infty} e^{-qT_{k+1}}\left(\frac{k}{q} + T_{k+1} - \sum_{i=1}^{k}\frac{T_i e^{-qT_i}}{1-e^{-qT_i}}\right)\prod_{j=1}^{k}(1-e^{-qT_j})\right].$$

For completeness, we also give expressions for the variant without driver impatience, which simply result from taking the expectations of the results from model $B_1$:

$$\mathbb{E}[e^{-sG}] = \mathbb{E}\left[\frac{(s+q)e^{-(s+q)T}}{s+qe^{-(s+q)T}}\right], \qquad \mathbb{E}[G] = \frac{\mathbb{E}[e^{qT}]-1}{q}.$$

## 2.4 Impact of driver behavior on capacity

In the previous section, we have introduced three behavior types, each with and without driver impatience. For these models we determined the distributions of the stationary number of low-priority cars in the queue, being characterized by the corresponding Laplace transform $\tilde{G}(s) = \mathbb{E}[e^{-sG}]$, and hence also the Laplace transform $\tilde{W}(s)$ of the waiting time. In this section we consider

the impact of the model of choice ($B_1$ up to $B_3$, that is) on the 'capacity' of the secondary road; here 'capacity' is defined as the maximum arrival rate $\lambda$ such that the corresponding queue does not explode. A standard result from queueing theory is that for the M/G/1 queue the stability condition is $\rho := \lambda\,\mathbb{E}[G] < 1$. As a consequence, the capacity of the minor road, denoted by $\bar{\lambda}$, can be determined for each of the models, with or without impatience:

$$\bar{\lambda} = \frac{1}{\mathbb{E}[G]},$$

where $\mathbb{E}[G]$ depends on the driver behavior, as explained before. In this section we focus on drivers *without* impatience, denoting the capacity for model $B_i$ by $\bar{\lambda}_i$, for $i = 1, 2, 3$. Although the expressions below can also be found in, for example, Heidemann and Wegmann [60], we add some new observations regarding the capacities.

**$B_1$ (constant gap):** In this case, we find the stability condition $\lambda < \bar{\lambda}_1$ where

$$\bar{\lambda}_1 := \frac{q}{e^{qT} - 1}.$$

**$B_2$ (sampling per attempt):** Using the expression for $\mathbb{E}[G]$, we now find the condition $\lambda < \bar{\lambda}_2$ where

$$\bar{\lambda}_2 = \frac{q}{(\mathbb{E}[e^{-qT}])^{-1} - 1}.$$

**$B_3$ (sampling per driver):** Here, the stability condition is in the form of $\lambda < \bar{\lambda}_3$ where

$$\bar{\lambda}_3 = \frac{q}{\mathbb{E}[e^{qT}] - 1}.$$

Importantly, it is here tacitly assumed that the moment generating function $\mathbb{E}[e^{qT}]$ of $T$ exists. A consequence that has not received much attention in the existing literature, is that it also means that in case $T$ has a polynomially decaying tail distribution (i.e., $\mathbb{P}(T > t) \approx C\,t^{-\beta}$ for some $C, \beta > 0$ and $t$ large) *the queue at the secondary road is never stable.* The reason is that for this type of distributions it is relatively likely that an extremely large $T$ is drawn, such that it takes very long before the car can cross the intersection (such that in the meantime the low-priority queue has built up significantly).

In fact, also for certain light-tailed distributions we find that $B_3$ has an undesirable impact on the capacity. Take, for example, $T$ exponentially distributed with parameter $\alpha$. In this case, we have

$$\mathbb{E}[G] = \begin{cases} 1/(\alpha - q), & q < \alpha, \\ \infty & q \geq \alpha, \end{cases}$$

implying that the capacity of the minor street drops to zero when $q \geq 1/\mathbb{E}[T]$. Actually, the situation might be even worse than it seems, because it can be shown that $\mathbb{E}[G^k] = \infty$ if $q \geq \alpha/k$, for $k = 1, 2, \ldots$. As a consequence, when $\alpha > q \geq \alpha/2$ the capacity is positive, but it follows from (2.2) and Little's law that the mean queue length and the mean delay at the minor road grow beyond any bound.

Another interesting observation, is that the arrival rates $\bar{\lambda}_1, \bar{\lambda}_2$ and $\bar{\lambda}_3$ obey the ordering

$$\bar{\lambda}_2 \geqslant \bar{\lambda}_1 \geqslant \bar{\lambda}_3$$

(where in $B_1$ we have chosen $T$ equal to the mean $\mathbb{E}[T]$ used in the other variants). This is an immediate consequence of Jensen's inequality, as we show now. To compare $\bar{\lambda}_1$ and $\bar{\lambda}_3$ realize that Jensen's inequality implies

$$\frac{1}{q}(\mathbb{E}[e^{qT}] - 1) \geqslant \frac{1}{q}(e^{q\mathbb{E}[T]} - 1),$$

which directly entails $\bar{\lambda}_3 \leq \bar{\lambda}_1$. Along the same lines, again appealing to Jensen's inequality, $\mathbb{E}[e^{-qT}] \geqslant e^{-q\mathbb{E}[T]}$, and hence $\bar{\lambda}_2 \geq \bar{\lambda}_1$.

We conclude this section by stating a number of general observations. In the first place, the above closed-form expressions show that *the stability conditions depend on the full distribution of* $T$, as opposed to just the mean values of the random quantities involved.

Secondly, in many dynamic systems introducing variability degrades the performance of the system. The fact that $\bar{\lambda}_2 \geqslant \bar{\lambda}_1$ indicates that in this case this 'folk theorem' does not apply: the fact that one resamples $T$ often (every driver selects a new value for each new attempt) actually *increases* the capacity of the low-priority road.

## 2.5 Numerical results and practical examples

In this section, we perform numerical experiments to investigate the impact of driver behavior with or without impatience, on the capacity of the minor road as well as the queue length on the minor road.

### 2.5.1 Example 1: ordering of the capacities

In this example, in which we try to take realistic parameter settings, we illustrate the impact of driver behavior on the capacity of the system and on the queue lengths. In particular, we compare the following three scenarios (corresponding to the three behavior types):

(1) All drivers search for a gap between consecutive cars on the major road, that is at least 7 seconds long.

(2) A driver on the minor street, waiting for a suitable gap on the major street, will sample a new (random) critical headway every time a car passes on the major street. With probability $9/10$ this critical headway is $6.22$ seconds, and with probability $1/10$ it is exactly $14$ seconds. Note that the expected critical headway is $0.9 \times 6.22 + 0.1 \times 14 = 7$ seconds, ensuring a fair comparison between this scenario and the previous scenario.

(3) In this scenario we distinguish between slow and fast traffic. We assume that 90% of all drivers on the minor road need a gap of (at least) $6.22$ seconds. The other 10% need at least $14$ seconds.

We do not yet incorporate impatience in this example. Fig. 2.1 depicts the capacity (veh/h) of the minor street as a function of $q$, the flow rate on the main road (veh/h). The relation $\bar{\lambda}_2 \geq \bar{\lambda}_1 \geq \bar{\lambda}_3$ is clearly visible. In the next example we will include driver impatience.

### 2.5.2 Example 2: paradoxical behavior

In this example we illustrate that the gap acceptance model can exhibit unexpected behavior under specific circumstances, caused by the randomness in the behavior of the drivers. To this end, we compare the following two distributions for $T$:

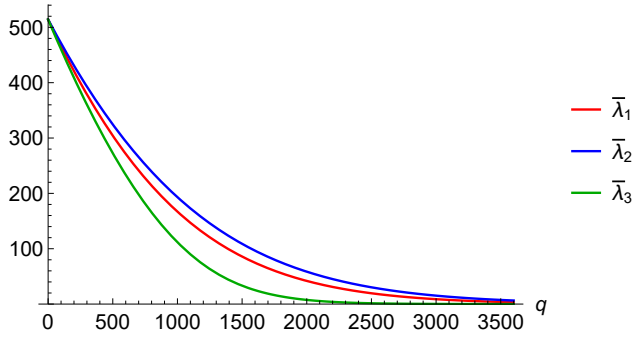
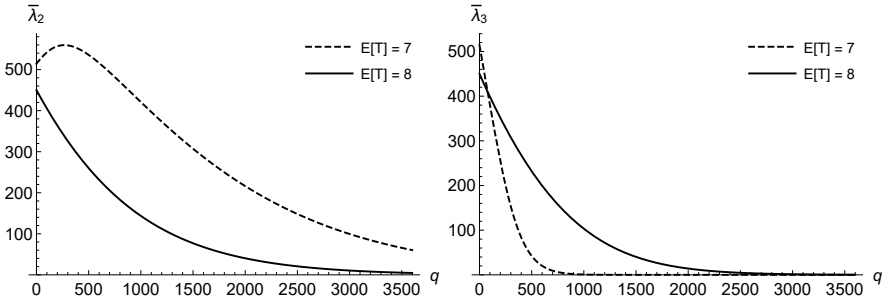



**Figure 2.1.** Capacity of the minor street (veh/h) as a function of the flow rate on the main road (veh/h) in Example 1.

1. $T_A$ is equal to 4 seconds with probability $9/10$, and 34 seconds with probability $1/10$. The expected critical headway is $\mathbb{E}[T] = 7$ seconds.

2. $T_B$ is equal to respectively 6 or 10 seconds, each with probability $1/2$. Now $\mathbb{E}[T] = 8$ seconds.

It is apparent that $\mathbb{E}[T_A] < \mathbb{E}[T_B]$, whereas the standard deviation of the first model (9 seconds) is much *greater* than in the second model (2 seconds). In Fig. 2.2 we plot the capacities $\bar{\lambda}_2$ and $\bar{\lambda}_3$, when $T$ has the same distributions as $T_A$ and, subsequently, when $T$ has the same distributions as $T_B$. We notice two things. In Fig. 2.2(a), corresponding to model $\mathrm{B}_2$, we see that the capacity when $T \stackrel{d}{=} T_A$ first *increases* when $q$ grows larger; only for larger values of $q$ it starts to decrease. This counterintuitive behavior, which is not uncommon in model $\mathrm{B}_2$, is explained in more detail in the next numerical example. Now, we mainly focus on the second striking result: in model $\mathrm{B}_3$, for $q > 78$ veh/h, the capacity when $T \stackrel{d}{=} T_A$ is *less* than the capacity when $T \stackrel{d}{=} T_B$, even though $\mathbb{E}[T_A] < \mathbb{E}[T_B]$. This is a consequence of the fact that the stability condition does not depend on the *mean* quantities only, as is common in most queueing models, but it depends on the entire distribution. In this case, the large variance of $T_A$ has a negative impact on the capacity of the minor street.




(a) Capacities for model $B_2$.

(b) Capacities for model $B_3$.

**Figure 2.2.** Capacities $\bar{\lambda}_2$ and $\bar{\lambda}_3$ as a function of $q$ (veh/h) in Example 2.

The last surprising result of this example is encountered when comparing the mean queue lengths resulting from the different types of driver's behavior. Although the capacities for $B_1$, $B_2$, and $B_3$ are strictly ordered, $\lambda_2 \geq \lambda_1 \geq \lambda_3$, we show in this example that this is not necessarily true for the mean queue lengths on the minor street. To this end, we only consider the situation where the critical headway is distributed as $T_A$. We fix the traffic flow on the major road at $q = 60$ vehicles per hour. The mean queue length on the minor road, as a function of $\lambda$, is depicted in Fig. 2.3, where we have taken $T = \mathbb{E}[T_A] = 7$ seconds in model $B_1$. The figure displays a paradoxical situation, where the mean queue length corresponding to $B_2$ (resampling) is *higher* than with $B_1$ (constant) for $71.2 < \lambda < 445.1$. Indeed, for $\lambda < 71.2$, which is hardly visible, $B_2$ also has a smaller mean queue length than $B_1$. The reason why this behavior was not encountered in the previous numerical example, is that it only occurs for small values of $q$, the traffic flow on the major road. In fact, it can be shown that, for the given distribution of $T$, this paradox only takes place when $q < 124.6$ vehicles per hour. Note that, due to Little's law, the mean delays exhibit the same behavior.

### 2.5.3 Example 3: the impact of resampling

The previous examples have illustrated that resampling, as described in $B_2$, has a positive impact on the capacity of the minor street. In this example we show that, under specific circumstances, this positive impact may be even bigger than expected. We show this by varying the probability distribution of the critical headway $T$, taking the following five distributions (all with $\mathbb{E}[T] = 7$

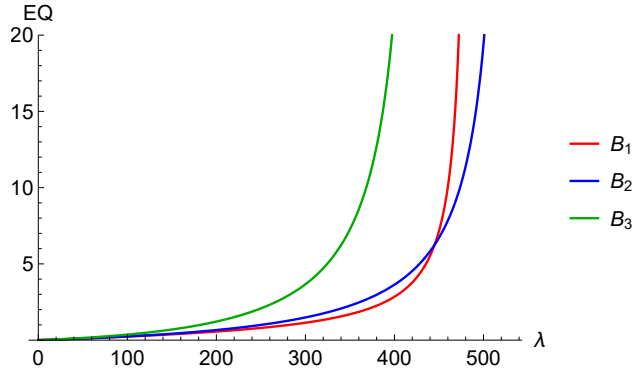



**Figure 2.3.** Mean queue lengths at the minor street (veh) as a function of the flow rate $\lambda$ (veh/h) in Example 2.

seconds):

1. $T$ is equal to 14 seconds with probability $1/10$, and 6.22 seconds with probability $9/10$. This distribution, referred to as High/Low $(14, 6.22)$ in Fig. 2.4, is the same as in Example 1.

2. $T$ is equal to 28 with probability $1/10$, and 4.67 with probability $9/10$. This is a similar distribution as the previous, but with more extreme values.

3. $T$ is equal to 42 with probability $1/10$, and 3.11 with probability $9/10$. Even more extreme values, as in Example 1.

4. $T$ is exponentially distributed with parameter $1/7$.

5. $T$ has a gamma distribution with shape parameter $1/2$ and rate $1/14$.

In Fig. 2.4 we have plotted the capacity of the minor road as a function of $q$, the traffic intensity of the main road. Several conclusions can be drawn. First, we notice that in all High/Low distributions, the capacity drops to zero as $q$ increases towards infinity. However, when zooming in at $q$ close to zero, it turns out that model 1 has an immediate capacity drop, whereas model 3 actually *increases* in capacity up to $q \approx 437$. A possible explanation for this striking phenomenon, that we also encountered in the previous numerical

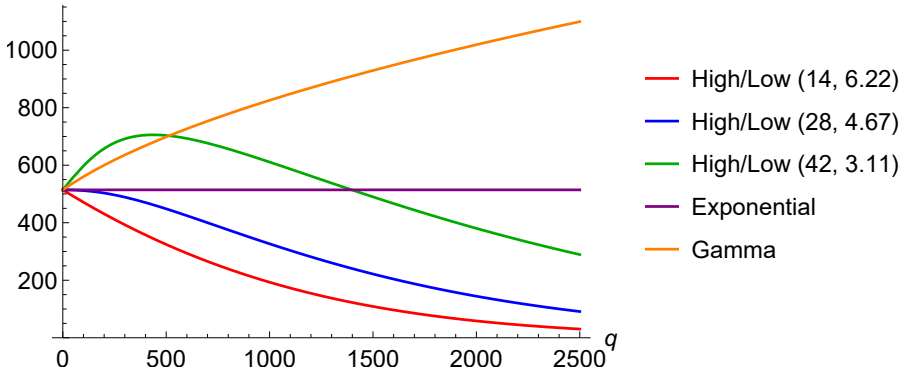



**Figure 2.4.** Capacity of the minor road (veh/h) for several distributions of $T$.

example, is that in the third model the critical headway is either extremely low, or extremely high. When a driver has a low critical capacity (which is quite likely to happen), he experiences no delay before leaving the minor street anyway. In contrast to this, when a driver has a high critical capacity, he will have to resample, meaning that he has to wait for the next car to pass on the major street. If $q$ increases, the frequency of resampling increases, meaning that he has to wait less before getting a new chance to obtain a small critical headway.

Another extreme case is the exponential distribution, which, due to its memoryless property, has a *constant* capacity, not depending on the traffic flow on the major street.

Clearly, the most paradoxical case is the Gamma distribution with a shape parameter less than one. If $T$ has this particular distribution, it can be shown that the capacity of the minor road keeps on *increasing* as $q$ increases. Although this case, admittedly, may not be a realistic one, it certainly provides valuable insight in the system behavior.

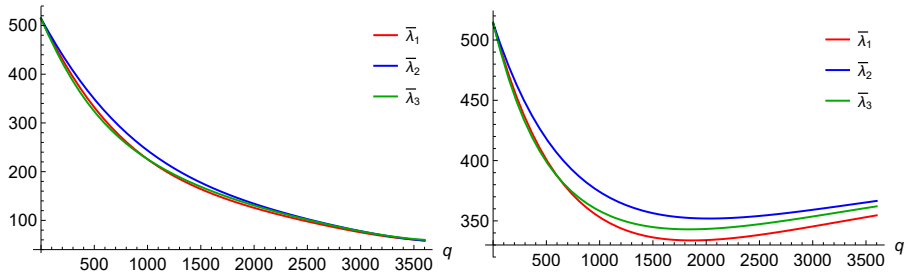
### 2.5.4 Example 4: impatience

We now revisit Example 1, but introducing driver impatience into the model, in the following specific form:

$$T_{k+1} = \alpha(T_k - \Delta) + \Delta, \qquad k = 1, 2, \ldots; 0 < \alpha < 1, \tag{2.5}$$

which means that the critical headway decreases in every next attempt, approaching the limiting value of $\Delta$. The parameter $\alpha$ determines the speed at which the patience decreases. In scenario 1 all $T_k$ are fixed, with $T_1 = 7$ seconds. In scenario 2, each of the $T_k$ is a random variable, with $T_1$ equal to $6.22$ or $14$ seconds, with probability $9/10$ and $1/10$ respectively. The distribution of $T_k$ for $k > 1$ can be determined from (2.5). Note that the impatience is a new random sample at each attempt, independent of the value of $T_{k-1}$. Scenario 3, as before, is similar to scenario 2, but each driver samples a random impatience $T_1$ exactly once. The value of $T_1$ (which is again either $6.22$ or $14$ seconds) determines the whole sequence of critical gap times at the subsequent attempts according to (2.5).

Fig. 2.5(a) shows the capacity as a function of $q$, when $\alpha = 9/10$ and $\Delta = 4$ seconds. It is noteworthy that the strict ordering that was observed in the case without impatience, is no longer preserved, even though $\mathbb{E}[T_k]$ is the same in all scenarios, for $k$ fixed. Another interesting phenomenon, depicted in Fig. 2.5(b), occurs when we decrease the parameter values to $\alpha = 8/10$ and $\Delta = 1$ second. Now, the capacity actually *increases* when $q$ exceeds a certain threshold. Due to the increase in $q$, gaps between cars on the major road will be smaller in general, but apparently the benefit of having a (much) lower critical headway at each attempt outweighs the disadvantage of having smaller gaps.




(a) $\alpha = 9/10, \Delta = 4$ sec. (b) $\alpha = 8/10, \Delta = 1$ sec.

**Figure 2.5.** Capacity of the minor street (veh/h) as a function of the flow rate on the main road (veh/h) in Example 4.

## 2.6 Discussion and conclusion

Our main target in this chapter has been to investigate the impact of randomness in the critical headway on the capacity for traffic flows of low priority at a road intersection. For that, we have analyzed three versions of a queueing-theoretic model; each with its own dynamics. Special attention was paid to drivers' impatience under congested circumstances: the value of the critical headway decreases with subsequent attempts to cross the main road.

In our first model ($B_1$) we have assumed that the sequence of critical headways is a deterministically decreasing sequence, and that all cars use the same sequence. In the second model ($B_2$), we let each car sample new values for the critical headway, according to a *stochastically decreasing* sequence. In the third model ($B_3$) we sample the first value for the critical headway for each car, but then use a deterministic decreasing sequence throughout the attempts of a particular car.

Several instances of our models have previously been studied in the literature, but not with the focus on assessing the impact of the drivers' behaviors on the capacity for the low priority flow. Our main observation is that randomness has a strong impact on the capacity. More specifically, the capacity region depends on the entire distribution(s) of the critical headway durations, and not only on the mean value(s). We also observe that resampling of the critical headway values has a benign impact on the capacity of the minor road.

The models studied in this chapter will be extended in the next chapter by including Markov platooning on the major road to model the fluctuations in the traffic density on that road.

Chapter 3

# Markov platooning

The main contribution of this chapter is the introduction of a new model for vehicle clustering on the main (major) road of an unsignalized intersection (see Figure 1.1), which we will refer to as *Markov platooning* throughout this chapter. The tractability of this model allows one to study the impact of various platoon formations on the main road, on the capacity of the minor road.
Based on the numerical results, we observe that platooning has a positive effect on the capacity of the minor road for given mean rate, but only for models with constant critical gaps and consistent gap acceptance behavior ($B_1$ and $B_3$). In a model with inconsistent behavior ($B_2$), it depends on the model parameters whether platooning increases or decreases the capacity.

## 3.1 Introduction

Assuming a Poisson arrival process on the major road is realistic in periods of free traffic flow, where it is assumed that any vehicle does not affect vehicles behind it. To make the model more realistic, platoon forming has been studied in the existing literature on gap acceptance models. Tanner [101] considers a model where platoon lengths are distributed as the busy period of a single-server queue. Heidemann and Wegmann [60], relying on results by Tanner [101], propose a general framework based on gap-block models. In such models, vehicles form platoons which arrive according to a Poisson process. The lengths of these platoons are i.i.d. random variables with a general distribution, which can be suitably chosen such that it matches real-life clustering behavior. Wu [119] observed that, in practice, traffic flow in the major stream can have up to four different regimes: free space (no vehicles), free flow (single vehicles), bunched traffic (platoons of vehicles), and queueing. By conditioning on the current regime, he applies the framework of [60] to set up a heuristic argument that provides a more general capacity formula that is valid under all four regimes. However, all of these models assume no (or a very weak form of) dependence between successive block sizes and gap sizes.

In this chapter, we introduce a new form of bunching on the main road, called *Markov platooning*. The tractability of this model allows us to study the impact of various platoon formations [47, 62, 71, 111] on the main road on the capacity of the minor road. By introducing Markov platooning, an arrival process based on Markov modulation, we allow for a more refined way of bunching on the major road that includes dependence between successive gap sizes.

The remainder of this chapter is structured as follows. Section 3.2 gives the model description. The capacity of the minor road is determined in Section 3.3. In Section 3.4, we present numerical examples to demonstrate the impact of Markov platooning, on the capacity of the minor road. Finally, in Section 3.5, we make several concluding remarks.

## 3.2 Model description

In this section, we introduce a novel way to model different traffic-flow regimes on the major road of an unsignalized intersection, using a well-established method to model dependence between successive interarrival times. We assume that the arrival process on the major road is modeled by a *Markov modulated Poisson process* (MMPP), described in Section 1.2. In an MMPP arrivals are generated at a Poisson rate $q_i$ when an exogeneous, autonomously evolving continuous-time Markov process (commonly referred to as the *background process*) is in state $i$. We denote by $d \in \{1, 2, \dots\}$ the number of states of the background process (where $d = 1$ corresponds to a non-modulated, ordinary Poisson process). We assume the background process to be irreducible; the corresponding stationary distribution is given by the vector $\pi$. In the sequel we denote by $M = (\mu_{ij})_{i,j=1}^{d}$ the transition rate matrix of the background process, and define $\mu_i := -\mu_{ii}$. Therefore, an MMPP allows different traffic-flow regimes on the major road. For example, in Figure 3.1, we show the arrival patterns of two MMPP's, each with two background states. The red squares mark arrivals during the high traffic intensity ($q_1$), while the green squares mark arrivals during the low intensity ($q_2$). It can be seen that platoons are generally longer when the background process is in state 1, corresponding to a high arrival rate. Additionally, we observe in Figure 3.1(a) that the background process stays longer in state 2 ($\mu_2 = 1/40$) than in state 1 ($\mu_1 = 1/20$). Another difference between the two sub-figures is that we choose $q_2 = 1/15$ in Figure 3.1(a) and $q_2 = 1/5$ in Figure 3.1(b). This explains why, in state 2, we see no platooning at all in Figure 3.1(a), but Figure 3.1(b) still shows some mild platoon forming.

## 3.3 Capacity

The main objective of this section is to develop methods that determine the capacity of the minor road under MMPP arrivals on the major road, for the models $B_1$ up to $B_3$ introduced in Section 1.2. Because of this focus on the capacity, we can simplify the model by taking away the queueing aspect on the minor road, assuming that this road is *saturated*: there are *always* low-priority cars waiting for gaps. The reason underlying this reduction is that capacity is a quantity that corresponds to stability of the associated queue, and stability

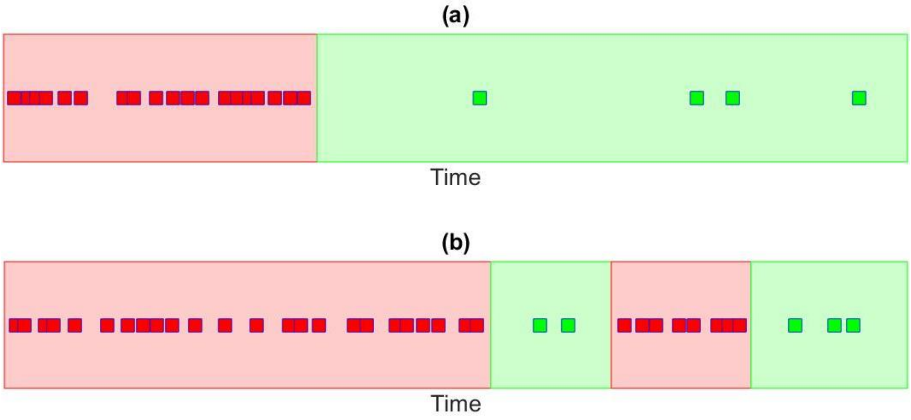



**Figure 3.1.** Simulated examples of two MMPP's with two background states. On the horizontal axis we depict the time, while the squares (red or green) mark the arrivals. In (a), we have chosen $\mu_1 = 1/20, \mu_2 = 1/40$ and arrival rates $q_1 = 1, q_2 = 1/15$ vehicles per time unit. In (b) we use $\mu_1 = 1/20, \mu_2 = 1/20$ and arrival rates $q_1 = 1, q_2 = 1/5$ vehicles per time unit. The red areas indicate that the background process is in state 1 (more platooning) and the green areas correspond to state 2 (less platooning).

essentially amounts to the queue being able to process all input in the long run.

The capacity, to be denoted by $\bar{\lambda}$, is the ratio of the mean number of arrived cars in a cycle (which we define below) to the mean duration of a cycle, which equals (due to renewal theory) the number of cars that can be served per unit time. The system is stable when $\lambda$, the arrival intensity on the minor road, is less than $\bar{\lambda}$. Again, we distinguish between the three behavior types $B_1$, $B_2$, and $B_3$, each with its own capacity $\bar{\lambda}_i$, for $i = 1, 2, 3$.

Our objective is to assess the impact of the three types of the driver's behavior on stability. As we have seen before, the capacity can be interpreted as the reciprocal of the time it takes for an arbitrary car to cross the major road. At first sight, the following procedure seems to provide us with $\bar{\lambda}$. Define $S_i$ as the time it takes for an arbitrary car to cross the major road, given the background process is in state $i$ when the car (which has reached the head of the queue) starts his attempt. Recalling that $\pi_i$ represents the long-run

fraction of time that the background process resides in state $i$, it is tempting to conclude that the capacity would equal

$$\sum_{i=1}^{d} \frac{\pi_i}{\mathbb{E}[S_i]}. \tag{3.1}$$

Alternatively, one might try to first take a weighted average of the mean service times, and then take the reciprocal to find the capacity,

$$\frac{1}{\sum_{i=1}^{d} \pi_i \mathbb{E}[S_i]}. \tag{3.2}$$

There is, however, a conceptual mistake in these (naïve) approaches. It is true that $\pi$ is indeed the distribution of the background process that is seen by cars that arrive at the queue, due to the well-known PASTA property. The distribution seen by the car that has reached the head of the queue, however, differs from $\pi$. To see this, think of the extreme case in which $q_1 = q > 0$ has some moderate value, and $q_2 = M$ is large. For large values of $M$, only cars who find the queue empty, may start their attempt while the background process is in state 2.

This reasoning illustrates how careful one should be when weighing capacities that belong to different regimes by the fractions of time in which those regimes apply. A very similar decomposition approach was followed by Wu [119]; he distinguishes four different regimes, as described above, each with an own capacity, and those are combined into a single capacity. The formulas obtained by Wu [119] likely provide a reasonable indication of the capacity across a wide range of parameters, but there are also many cases in which the approach fails to do so. Later on, we provide an example which illustrates what errors may result when following the naïve approaches.

**$B_1$ (constant gap):** In this model, every driver on the minor road needs the same constant critical headway $T$ to enter the major road. In our analysis we use the renewal reward theorem, which entails that the capacity can be written as the mean number of cars arriving in a regenerative cycle divided by the mean duration of that cycle. For our purposes, an appropriate definition of a cycle is: the time elapsed between two consecutive epochs such that (i) the background process is in a reference state (say state 1), and (ii) a service is completed (i.e., a low-priority car is served).

To make our model Markovian, we approximate this deterministic $T$ by an Erlang random variable with $k$ phases of average length $T/k$. It is well known that a deterministic $T$ can be approximated by the sum of $k$ independent exponential random variables, each with parameter $\kappa := k/T$, with $k$ large; to see this, observe that this Erlang random variable has mean $T$ (as desired), and variance $k/\kappa^2 = T^2/k$, which goes to 0 as $k$ grows large. In the sequel we write $\varrho_i := \mu_i + q_i + \kappa$. Define $h_{ij}$ as the mean number of cars that is served till the cycle ends, given that the current state of the background process is $i \in \{1, \ldots, d\}$ and the car in service has finished $j \in \{0, \ldots, k-1\}$ phases of the Erlang distribution. To find the mean number of arrived cars in a cycle, we need to find $h_{10}$. This can be done as follows.

Relying on 'standard Markovian reasoning', by conditioning on the first jump,

$$h_{1j} = \sum_{\ell \neq 1} \frac{\mu_{1\ell}}{\varrho_1} h_{\ell j} + \frac{q_1}{\varrho_1} h_{10} + \frac{\kappa}{\varrho_1} \left( h_{1,j+1} 1_{\{j<k-1\}} + 1_{\{j=k-1\}} \right),$$

where $1_{\{.\}}$ denotes the indicator function.
In addition, for $i \neq 1$,

$$h_{ij} = \sum_{\ell \neq i} \frac{\mu_{i\ell}}{\varrho_i} h_{\ell j} + \frac{q_i}{\varrho_i} h_{i0} + \frac{\kappa}{\varrho_i} \left( h_{i,j+1} 1_{\{j<k-1\}} + (1 + h_{i0}) 1_{\{j=k-1\}} \right).$$

This can be written as a linear system of $dk$ equations with $dk$ unknowns of the form $A\vec{h} = \vec{b}$, where entries of the matrix $A = [a_{mn}]$, $\vec{h}$ and $\vec{b} = [b_m]$ are given as follows, with $i = \lceil m/k \rceil$,

$$a_{mn} = \begin{cases} -\dfrac{\kappa}{\varrho_i}, & \text{if } n = m+1;\ m \neq k, 2k, ..., dk; \\ -\dfrac{q_i}{\varrho_i}, & \text{if } n = (i-1)k+1;\ m \neq 1, k+1, 2k, 2k+1, 3k, ..., \\ & \quad (d-1)k+1, dk; \\ -\dfrac{\kappa + q_i}{\varrho_i}, & \text{if } n = (i-1)k+1;\ m = 2k, ..., dk; \\ -\dfrac{\mu_{i,\ell+1}}{\varrho_i}, & \text{if } n = (\ell - i + 1)k + m;\ \ell \in \{0, 1, ..., d-1\} \setminus \{i-1\}; \\ 1 - \dfrac{q_i}{\varrho_i}, & \text{if } n = m;\ m = 1, k+1, ..., (d-1)k+1; \\ 1, & \text{if } n = m;\ m \neq 1, k+1, ..., (d-1)k+1; \\ 0, & \text{else,} \end{cases}$$

$\vec{h} = [h_{10}, h_{11}, ..., h_{1,k-1}, h_{20}, h_{21}, ..., h_{2,k-1}, ..., h_{d0}, h_{d1}, ..., h_{d,k-1}]^{\mathrm{T}}$, and

$$b_m = \begin{cases} \dfrac{\kappa}{\varrho_\ell}, & \text{if } m = \ell k, \ell = 1, 2, ..., d \\ 0, & \text{else.} \end{cases}$$

It is noted that $|a_{mm}| = \sum_{n \neq m} |a_{mn}|$ for all $m \neq k$ and for $m = k$, $|a_{kk}| > \sum_{n \neq k} |a_{kn}|$. Therefore, the matrix $A$ is weak diagonally dominant where one row is strictly dominant and $A$ is also an irreducible matrix, and hence $A$ is invertible [61]. Therefore, the solution of the system of equations $A\vec{h} = \vec{b}$ is $\vec{h} = A^{-1}\vec{b}$. We thus find the desired quantity $h_{10}$.

To determine the capacity we need, in addition to the mean number of arrived cars in a cycle, also the mean duration of a cycle. To this end we define $\tau_{ij}$ as the mean time till the end of the current cycle, given that the current state of the background process is $i \in \{1, \ldots, d\}$ and the car in service has finished $j \in \{0, \ldots, k-1\}$ phases of the Erlang distribution. The objective is now to find the mean duration of a cycle, which is given by $\tau_{10}$.

Similarly to the procedure we set up above,

$$\tau_{1j} = \frac{1}{\varrho_1} + \sum_{\ell \neq 1} \frac{\mu_{1\ell}}{\varrho_1} \tau_{\ell j} + \frac{q_1}{\varrho_1} \tau_{10} + \frac{\kappa}{\varrho_1} \tau_{1,j+1} 1_{\{j<k-1\}}.$$

In addition, for $i \neq 1$,

$$\tau_{ij} = \frac{1}{\varrho_i} + \sum_{\ell \neq i} \frac{\mu_{i\ell}}{\varrho_i} \tau_{\ell j} + \frac{q_i}{\varrho_i} \tau_{i0} + \frac{\kappa}{\varrho_i} \left( \tau_{i,j+1} 1_{\{j<k-1\}} + \tau_{i0} 1_{\{j=k-1\}} \right).$$

Also this system can be written as $dk$ linear equations with $dk$ unknowns. More precisely, we have $A\vec{\tau} = \vec{c}$ with $A$ as defined before and

$$\vec{\tau} = [\tau_{10}, \tau_{11}, ..., \tau_{1,k-1}, \tau_{20}, \tau_{21}, ..., \tau_{2,k-1}, ..., \tau_{d0}, \tau_{d1}, ..., \tau_{d,k-1}]^{\mathrm{T}},$$
$$c_m = \frac{1}{\varrho_\ell} \quad \text{for } (\ell - 1)k + 1 \leq m \leq \ell k, \text{ and } \ell = 1, 2, ..., d.$$

We already proved that $A$ is invertible, and therefore the unique solution of the system of equations $A\vec{\tau} = \vec{c}$ is $\vec{\tau} = A^{-1}\vec{c}$. We thus find $\tau_{10}$.

The capacity of this system can now be evaluated as $\bar{\lambda}_1 := h_{10}/\tau_{10}$, meaning that the stability condition of the low-priority queue is $\lambda < \bar{\lambda}_1$. In the

numerical procedure, the value of $k$ should be chosen large, to ensure that the Erlang distribution is sufficiently 'close-to-deterministic'.

**$B_2$ (sampling per attempt):** As pointed out before, in this behavior type every driver samples a 'fresh' random $T$ for every attempt to enter the major road. Let us assume that the gap size $T$ equals some deterministic $T_n$ with probability $p_n$; below we present the computational procedure for $n \in \{1, 2\}$, but it can be extended in an evident manner to the situation in which $T$ can attain more than 2 possible values. Analogously to what we did in the procedure to evaluate the capacity for $B_1$, we approximate $T_n$ by an Erlang random variable with $k_n$ phases; each of the phases is exponentially distributed with parameter $\kappa_n = k_n/T_n$.

We write $\varrho_i^{(n)} := \mu_i + q_i + \kappa_n$. Let $h_{ij}^{(n)}$ be the mean number of cars that is served till the cycle ends, given that the current state of the background process is $i \in \{1, \ldots, d\}$, the car in service has gap size $T_n$ and the car in service has finished $j \in \{0, \ldots, k_n - 1\}$ phases. We wish to find $h_{10}$ where

$$h_{i0} = p_1 h_{i0}^{(1)} + p_2 h_{i0}^{(2)} \text{ for } i = 1, 2. \tag{3.3}$$

Then

$$h_{1j}^{(n)} = \sum_{\ell \neq 1} \frac{\mu_{1\ell}}{\varrho_1^{(n)}} h_{\ell j}^{(n)} + \frac{q_1}{\varrho_1^{(n)}} h_{10} + \frac{\kappa_n}{\varrho_1^{(n)}} \left( h_{1,j+1}^{(n)} 1_{\{j<k_n-1\}} + 1_{\{j=k_n-1\}} \right).$$

Observe how the resampling is incorporated in this system: when an attempt has failed a 'fresh' new gap size is sampled, explaining the $h_{10}$ (rather than $h_{10}^{(n)}$) in the right hand side. In addition, for $i \neq 1$,

$$h_{ij}^{(n)} = \sum_{\ell \neq i} \frac{\mu_{i\ell}}{\varrho_i^{(n)}} h_{\ell j}^{(n)} + \frac{q_i}{\varrho_i^{(n)}} h_{i0} + \frac{\kappa_n}{\varrho_i^{(n)}} \left( h_{i,j+1}^{(n)} 1_{\{j<k_n-1\}} + (1 + h_{i0}) 1_{\{j=k_n-1\}} \right).$$

The $h_{i0}$ in the right hand side of the previous display corresponds to the event that an attempt has succeeded, after which a new gap size is sampled.

The above equations can be written as a linear system of the type $A\vec{h} = \vec{b}$ for a matrix $A$ and vector $\vec{b}$ (which evidently differ from the matrix $A$ and vector $\vec{b}$ that were used in the model $B_1$) consisting of $d(k_1 + k_2)$ equations with $d(k_1 + k_2)$ unknowns. With the same argument as we have used for $B_1$, it follows that the coefficient matrix $A$ is invertible. Using (3.3), this

facilitates the computation of $\vec{h}$ and in particular the desired quantity $h_{10}$ (from $h_{10} = p_1 h_{10}^{(1)} + p_2 h_{10}^{(2)}$).

We then define $\tau_{ij}^{(n)}$ as the mean time till the current cycle ends, given that the current state of the background process is $i \in \{1, \ldots, d\}$, the car in the service has gap size $T_n$ and the car in service has finished $j \in \{0, \ldots, k_n - 1\}$ phases. The objective is to set up a numerical procedure to evaluate $\tau_{10}$ where $\tau_{i0} = p_1 \tau_{i0}^{(1)} + p_2 \tau_{i0}^{(2)}$ for $i = 1, 2$. Using the same argumentation as above,

$$\tau_{1j}^{(n)} = \frac{1}{\varrho_1^{(n)}} + \sum_{\ell \neq 1} \frac{\mu_{1\ell}}{\varrho_1^{(n)}} \tau_{\ell j}^{(n)} + \frac{q_1}{\varrho_1^{(n)}} \tau_{10} + \frac{\kappa_n}{\varrho_1^{(n)}} \tau_{1,j+1}^{(n)} 1_{\{j < k_n - 1\}}.$$

In addition, for $i \neq 1$,

$$\tau_{ij}^{(n)} = \frac{1}{\varrho_i^{(n)}} + \sum_{\ell \neq i} \frac{\mu_{i\ell}}{\varrho_i^{(n)}} \tau_{\ell j}^{(n)} + \frac{q_i}{\varrho_i^{(n)}} \tau_{i0} + \frac{\kappa_n}{\varrho_i^{(n)}} \left( \tau_{i,j+1}^{(n)} 1_{\{j < k_n - 1\}} + \tau_{i0} 1_{\{j = k_n - 1\}} \right),$$

with $\tau_{i0} = p_1 \tau_{i0}^{(1)} + p_2 \tau_{i0}^{(2)}$ for $i = 1, 2$. Again, this system can be written as a linear system of $d(k_1 + k_2)$ equations with $d(k_1 + k_2)$ unknowns, say $A\vec{\tau} = \vec{c}$, with $A$ as above (and hence invertible). Therefore, the solution of the system of equations $A\vec{\tau} = \vec{c}$ is $\vec{\tau} = A^{-1}\vec{c}$, and we can compute $\tau_{10} = p_1 \tau_{10}^{(1)} + p_2 \tau_{10}^{(2)}$. The capacity of the low-priority queue under $B_2$ is therefore $\bar{\lambda}_2 = h_{10}/\tau_{10}$.

**$B_3$ (sampling per driver):** We finally consider the model with consistent behavior, i.e., each driver sticks to the gap size he or she initially sampled. The procedure is similar to the ones we developed for $B_1$ and $B_2$, and therefore we restrict ourselves to the main steps.

Define, as before, $h_{i0} = p_1 h_{i0}^{(1)} + p_2 h_{i0}^{(2)}$ for $i = 1, 2$. The mean number of cars served during the cycle follows from

$$h_{1j}^{(n)} = \sum_{\ell \neq 1} \frac{\mu_{1\ell}}{\varrho_1^{(n)}} h_{\ell j}^{(n)} + \frac{q_1}{\varrho_1^{(n)}} h_{10}^{(n)} + \frac{\kappa_n}{\varrho_1^{(n)}} \left( h_{1,j+1}^{(n)} 1_{\{j < k_n - 1\}} + 1_{\{j = k_n - 1\}} \right);$$

it is instructive to compare this equation with the corresponding one for $B_2$: when the attempt has failed the gap size is *not* resampled. Also, for $i \neq 1$, along the same lines,

$$h_{ij}^{(n)} = \sum_{\ell \neq i} \frac{\mu_{i\ell}}{\varrho_i^{(n)}} h_{\ell j}^{(n)} + \frac{q_i}{\varrho_i^{(n)}} h_{i0}^{(n)} + \frac{\kappa_n}{\varrho_i^{(n)}} \left( h_{i,j+1}^{(n)} 1_{\{j < k_n - 1\}} + (1 + h_{i0}) 1_{\{j = k_n - 1\}} \right);$$

resampling is only done when an attempt has been successfully completed.

Similarly, the system of equations for the mean cycle length is

$$\tau_{1j}^{(n)} = \frac{1}{\varrho_1^{(n)}} + \sum_{\ell \neq 1} \frac{\mu_{1\ell}}{\varrho_1^{(n)}} \tau_{\ell j}^{(n)} + \frac{q_1}{\varrho_1^{(n)}} \tau_{10}^{(n)} + \frac{\kappa_n}{\varrho_1^{(n)}} \tau_{1,j+1}^{(n)} 1_{\{j<k_n-1\}},$$

and, for $i \neq 1$,

$$\tau_{ij}^{(n)} = \frac{1}{\varrho_i^{(n)}} + \sum_{\ell \neq i} \frac{\mu_{i\ell}}{\varrho_i^{(n)}} \tau_{\ell j}^{(n)} + \frac{q_i}{\varrho_i^{(n)}} \tau_{i0}^{(n)} + \frac{\kappa_n}{\varrho_i^{(n)}} \left( \tau_{i,j+1}^{(n)} 1_{\{j<k_n-1\}} + \tau_{i0} 1_{\{j=k_n-1\}} \right),$$

with $\tau_{i0} = p_1 \tau_{i0}^{(1)} + p_2 \tau_{i0}^{(2)}$ for $i = 1, 2$. The linear system can solved as before, yielding $h_{10}$ and $\tau_{10}$. Therefore, the capacity of the system can be evaluated as $\bar{\lambda}_3 = h_{10}/\tau_{10}$.

## 3.4 Numerical results

The purpose of this collection of numerical examples is to exhibit specific, interesting features of gap acceptance models that relate to the impact of Markov platooning, on the capacity of the minor road. In the literature, it has already been observed that platoon forming on the major road may have a positive impact on the capacity of the minor road.

### 3.4.1 Example 1: the impact of Markov platooning

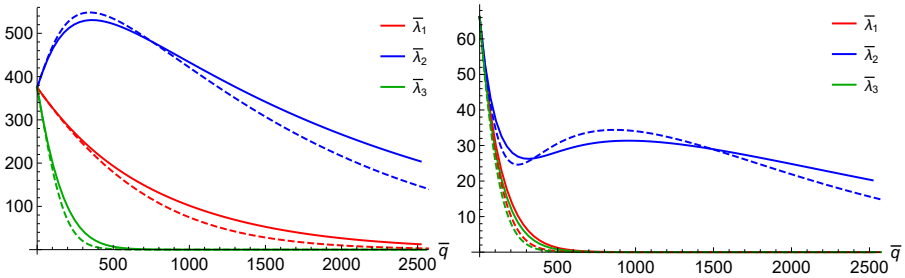
In this example, we compare the capacity of the minor road for the three behavior types $B_1$, $B_2$, and $B_3$. For the last two behavior types, we assume that a driver requires either a short gap of $T_1 = 4$ seconds, or an extremely long gap of $T_2 = 60$ seconds. Obviously these values are not chosen with the intention to mimic realistic behavior, but to point out extreme situations that might occur. For behavior type $B_1$, we take $T = p_1 T_1 + p_2 T_2$ seconds long, where $p_2 := 1 - p_1$.

For these settings, we compare the model with and without Markov platooning. With platooning, we take $\mu_1 = 1/60$ and $\mu_2 = 1/240$, resulting in exponential periods of, on average, one minute where the arrival rate on the major road is $q_1$, followed by exponential periods of, on average, four minutes,

with arrival rate $q_2$. We assume a fixed ratio of $q_1$ and $q_2$, namely $q_1 = 3q_2$. The long-term average arrival rate equals

$$\bar{q} := \frac{q_1/\mu_1 + q_2/\mu_2}{1/\mu_1 + 1/\mu_2} = \frac{q_1\mu_2 + q_2\mu_1}{\mu_1 + \mu_2}.$$

We compare the capacities with those obtained from the model without platooning, where we assume Poisson arrivals with rate $\bar{q}$.




(a) $p_1 = 0.9$ (b) $p_1 = 0.1$.

**Figure 3.2.** Capacity of the minor street (veh/h) as a function of the average flow rate on the main road (veh/h) in Example 1. The solid lines correspond to the model with Markov platooning; the dashed lines correspond to the model without platooning.

Figure 3.2 depicts the capacity (veh/h) of the minor street as a function of $\bar{q}$, the average flow rate on the main road (veh/h), for $p_1 = 0.9$ and $p_1 = 0.1$ respectively. As in the non-modulated case, we observed the relation $\bar{\lambda}_2 \geqslant \bar{\lambda}_1 \geqslant \bar{\lambda}_3$ in Section 2.4. Due to the lack of explicit expressions for $\bar{\lambda}_1$, $\bar{\lambda}_2$, and $\bar{\lambda}_3$, we cannot prove the strict ordering now. We did, however, observe it in all numerical examples that we conducted, and conjecture the ordering to hold true in general.

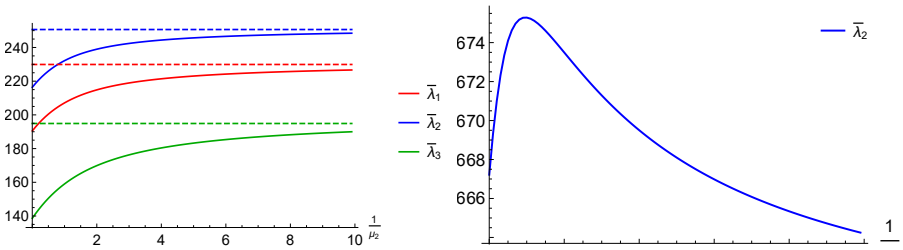
Based on the results of this example (and many other examples that are not discussed in the present chapter) we are inclined to believe that platooning has a positive effect on the capacity of the minor road, but *only* for models $B_1$ and $B_3$. In a model with inconsistent behavior, it really depends on the model parameters whether platooning increases or decreases the capacity. This is nicely illustrated in Figure 3.2(a) and even better in Figure 3.2(b).

### 3.4.2 Example 2: platoon lengths

In this example we fix the overall arrival rate on the major road, but we vary the platoon sizes. In more detail, we assume that $q_1 = 600$ veh/h and $q_2 = 2400$ veh/h. This means that phase 1 can be considered as a situation of moderate traffic (every 6 seconds a car passes), whereas phase 2 can be considered as one big platoon (on average every 1.5 seconds a car passes). The overall arrival rate $\bar{q}$ is fixed at 900 vehicles per hour, which implies that $\mu_1/\mu_2 = 1/5$. By varying the mean platoon length $1/\mu_2$ (in seconds) between 0 and 10, we will get better insight in the relation between platoon lengths and the capacity. Wegmann [111, Section 5] conducted a very similar experiment, varying the mean number of vehicles per bunch. He observed that the capacity increases with increasing variance of gaps.

We consider two different distributions for the critical headways. First, we consider the situation with $T_1 = 6.22$, $T_2 = 14$, and $p_1 = 0.9$, which can be considered as a quite realistic situation that we have used in Chapter 2. In Figure 3.3(a) we show the results for behavior types $B_1$, $B_2$, and $B_3$. The relation between the capacity and the mean platoon length is in line with [111, Figure 3]. Our numerical experiments confirm that this is indeed typical behavior for $B_1$, $B_2$, and $B_3$. Nevertheless, we want to show that it is possible to create a situation where model $B_2$ exhibits completely different behavior. When changing the distribution of the critical headway such that $T_1 = 3$ and $T_2 = 60$, we no longer see a monotonous relation between the capacity and the mean platoon length; see Figure 3.3(b). Bearing in mind that this inconsistent behavior type in combination with the extreme values for $T_1$ and $T_2$ might not be all too realistic, we do not find it likely that this type of behavior occurs in practical situations, but the model shows that it is not entirely impossible. For completeness, we want to mention that under extreme circumstances such as mean platoon lengths of 1000 seconds, the capacity with consistent behavior $B_3$ will also exhibit a drop, but not as drastically as in Figure 3.3(b).

The final conclusion that can be drawn from this example, is that one should be cautious when developing capacity estimates based on Equations (3.1) and (3.2). This type of reasoning may create a substantial bias, due to the fact that the vehicle at the head of the queue typically does not see the background process in equilibrium. It is noted that such argumentation underlies the capacity formulae in e.g. [119], where the capacity is calculated by conditioning on the state of the background process, i.e., the state of the




(a) $T_1 = 6.22, T_2 = 14, p_1 = 0.9$ (b) $T_1 = 3, T_2 = 60, p_1 = 0.9$

**Figure 3.3.** Capacity of the minor street (veh/h) as a function of the mean platoon length (sec) in Example 2. The dashed lines in (a) indicate the limiting capacities for $\mu_2 \downarrow 0$ while keeping the ratio $\mu_1/\mu_2$ fixed.

traffic on the major road (free space, free flow, bunching, or queueing). This example, and also Wegmann's example, show that there is a dependency between the mean platoon size and the capacity. The parameters in these examples are carefully chosen, such that the steady-state distribution of the background process (the vector $\pi$) remains unchanged. In our case, the major road is in state "free flow" for a fraction $\pi_1 = 5/6$ of the time, and in state "bunched" for a fraction $\pi_2 = 1/6$ of the time. If one would use the naïve approach and determine $\mathbb{E}[S_1]$ and $\mathbb{E}[S_2]$ by considering two separate models with regular Poisson arrivals, with intensities respectively $q_1$ and $q_2$, and use Equation (3.1), the capacities for models $B_1$, $B_2$, and $B_3$, respectively, would be

$$\bar{\lambda}_1 = 229.91, \bar{\lambda}_2 = 250.65, \bar{\lambda}_3 = 194.89,$$

independent of $\mu_1$ and $\mu_2$. From Figure 3.3(a) and Figure 3 in [111], it is clearly visible that these values (indicated by the dashed lines in Figure 3.3(a)) may differ substantially from the actual capacities. In fact, the capacities calculated from (3.1) can be interpreted as the limiting capacities from our MMPP model when $\mu_2 \downarrow 0$ while keeping the ratio $\mu_1/\mu_2$ fixed. When using (3.2) to compute the capacities, one would obtain

$$\bar{\lambda}_1 = 96.28, \bar{\lambda}_2 = 130.74, \bar{\lambda}_3 = 11.63,$$

leading to even greater errors.

## 3.5 Discussion and conclusion

In this chapter, we have provided a framework that allows the systematic evaluation of the effect of platooning on the primary road (modeling the fluctuations in the traffic density on that road). Unlike studies that appeared before, we focus on assessing the impact of the drivers' behaviors ($B_1$ up to $B_3$) together with platooning on the primary road, on the capacity for the low priority flow. Based on the numerical results, we observed that platooning has a positive effect on the capacity of the minor road for given mean rate, but only for models $B_1$ and $B_3$. In a model with inconsistent behavior, it really depends on the model parameters whether platooning increases or decreases the capacity. It is possible to include impatience as discussed in the previous chapter in these models $B_1$ up to $B_3$, with only minor adaptations.

One of the possible extensions of the model studied in this chapter, is to include the fact that cars may not need the entire duration of their critical headway to cross the main road (the remainder of that duration is the driver's *safety margin* to cross the road). Naturally, letting subsequent cars use the remaining critical headway leads to an increase of the capacity of the minor road. This extension will be discussed in more details in Chapter 7.

Chapter 4

# Generalized $M^X$/semi-Markov/1 queue

In this chapter, we investigate the transient and stationary queue-length distributions of a class of service systems with correlated service times. The classical $M^X/G/1$ queue with semi-Markov service times ($M^X$/semi-Markov/1) is the most prominent example in this class and serves as a vehicle to display our results. The sequence of service times is governed by a modulating process $J(t)$. The state of $J(\cdot)$ at a service initiation time determines the joint distribution of the subsequent service duration and the state of $J(\cdot)$ at the next service initiation.

Several earlier works have imposed technical conditions, on the zeros of a matrix determinant arising in the analysis, that are required in the computation of the stationary queue length probabilities. The imposed conditions in several of these articles are difficult or impossible to verify. Without such assumptions, we determine both the transient and the steady-state joint distribution of the

number of customers immediately after a departure and the state of the process $J(t)$ at the start of the next service.

We numerically investigate how the mean queue length is affected by variability in the number of customers that arrive during a single service time. Our main observations here are that increasing variability may *reduce* the mean queue length, and that the Markovian dependence of service times can lead to large queue lengths, even if the system is not in heavy traffic.

## 4.1 Introduction

Service systems with correlated service durations have a long tradition in the queueing literature. Such systems enjoy a large variety of application domains, including logistics, production management and telecommunications [5, 48, 82, 99]. Our main motivation stems from road traffic analysis, where traffic flows may interact at junctions or crossings [2, 98]. Focus, for illustration, on a traffic flow that merges into a main flow (very similar considerations are valid for road intersections). If the traffic density on the main flow is high, vehicles in the secondary flow may queue up before merging into the main flow. The merging times required for two subsequent vehicles will be strongly correlated as they experience similar traffic conditions on the main flow. In this chapter, we will capture this dependence in a queueing model in which the sequence of service times is governed by a modulating Markovian process. Although our analysis allows for a slightly larger class of models, we will use the classical $M/G/1$ queue with semi-Markov service times [82], and more specifically its extension to batch arrivals [80] to compare our results with existing literature.

The first to have investigated this class of queueing models was Gaver [48], who derived the waiting time in a single-server queue with two types of customers arriving according to independent Poisson processes. In that model, service times are class-specific and when service switches from one type to the other, an additional switch-over time is required. This framework was generalized by Neuts [82], allowing for more than two customer types and the sequence of service times forming a semi-Markov process. Under technical assumptions (these will be discussed later in detail) Neuts obtained the transient and stationary distributions of queue lengths, waiting times and busy periods. Subsequently, Çinlar [23] obtained the transient and stationary

queue length distributions under less restrictive assumptions, and Purdue [86] showed that the assumptions imposed by Neuts and Çinlar are not necessary for the analysis of the busy period, presenting an alternative approach. The literature on extensions of this model steadily expanded in the next two decades. In [81], Neuts studied the multi-type $M/G/1$ queue with change-over times when switching service from one type of customer to another. A further generalization allowing for Poisson arrivals of groups (batches) of customers of arbitrary random size was investigated by Neuts in [80], obtaining the busy period, queue length and waiting time distributions.
The departure process of a related model with single Poisson arrivals and exponential service times was determined by Magalhães and Disney [76]. In that model, the rate of the exponential service times depends on the type of the customer being served as well as that of its predecessor.
Models with single arrivals, but with both the arrivals and the services depending on a common semi-Markov process have been investigated by De Smit [97] and Adan and Kulkarni [5]. Using the Wiener-Hopf factorization technique, De Smit [97] obtained the waiting time and queue length distributions. Adan and Kulkarni [5] considered a similar setting, but with the customer type being determined at arrival instants (independent of the service durations).

In this chapter, we investigate the transient and stationary queue length distributions in a single server model with semi-Markov service times and with batch arrivals (our framework includes Poisson arrivals of batches as the most prominent example). In order to explain the technical contribution of our work, it is best to compare with the expositions of Neuts [82] and Çinlar [23]. In those papers only single Poisson arrivals were allowed, but the subsequent analysis is very similar. The earlier mentioned technical assumptions made by Neuts entail that the zeros of a particular matrix determinant appearing in the transient analysis are either strictly separated or completely coincide. This ensures that the zeros are analytic functions of the entries of the matrix and, consequently, that the stationary distribution can be obtained from the transient distribution. The assumptions were relaxed by Çinlar [23] while maintaining the analyticity of the zeros. Unfortunately, it remains hard, if not impossible, to verify the required conditions in practice, as they must hold for the zeros as *functions* of the matrix entries. As noted earlier, Purdue [86] showed that the assumptions imposed by Neuts and Çinlar are not necessary for the analysis of the busy period. Our work shows that these assumptions

are also not needed for the analysis of the queue length distribution. This comes at the expense of a separate analysis for the stationary distribution, which is more involved than that of the transient distribution. Specifically, we determine the generating function of the number of customers immediately after the departure of an arbitrary customer, considering both transient and steady-state behavior. For Poisson batch arrivals, in steady state we further obtain the queue length distribution at batch-arrival instants and at arbitrary times, which are identical due to PASTA. Note that this distribution is in general *not* the same as that at departure times (for single arrivals they would coincide).

A further contribution is an extensive numerical investigation of the mean queue length in steady state. We show that due to the dependence between service times, the mean number of customers may be very large, even if the load on the system is not large. A noteworthy observation is that *increasing* the variability in the number of customers arriving during a service time may in fact *decrease* the mean queue length.

The remainder of this chapter is organized as follows. In Section 4.2, we describe (a slight generalization of) the $M^X/G/1$ queue with semi-Markov services ($M^X$/semi-Markov/1). In Section 4.3, we derive the transient and stationary probability generating functions of the number of customers in the system immediately after a departure. In Section 4.4, we derive the generating functions of the stationary number of customers at an arbitrary epoch, at batch arrival epochs and at customer arrivals. The special case with only two customer types is specified in Section 4.5. In Section 4.6, we present numerical examples to demonstrate the impact of the correlated arrivals, and of the variability of the number of customers arriving during a service time, on the expected number of customers in the system. We summarize our results in Section 4.7.

## 4.2 Model description

We start by describing the $M^X$/semi-Markov/1 queueing model, which is the most natural example in our framework. Our analysis extends directly to any model that satisfies the dynamics described in the recurrence relation (4.9) below.

### 4.2.1 The $M^X$/semi-Markov/1 queue

Customers arrive in batches at a single server queue according to a Poisson process with rate $\lambda$; the batch size is denoted by the random variable $B$ with generating function $B(z)$, for $|z| \leq 1$. Customers are served in order of arrival, with speed 1. Customers within a batch are assumed to be ordered arbitrarily. The service times are governed by a Markov process $J_n$, $n = 0, 1, \ldots$, that can take values in $\{1, 2, \ldots, N\}$, for some integer $N$. It will be convenient to refer to $J_n$ as the *type* of the $n$-th customer; thus there are $N$ customer types. The service time of the $n$-th customer is denoted with $G^{(n)}$. An essential feature of our model is that the type of the $(n+1)$-th customer depends both on the type of the $n$-th customer *and* on the service duration of the $n$-th customer. This exactly matches the framework of semi-Markov (SM) service times introduced by Neuts [82], and thus the queueing system is referred to as the $M^X$/SM/1. We define

$$G_{ij}(x) = \mathbb{P}(G^{(n)} \leq x, J_{n+1} = j | J_n = i), \quad x \geq 0, \quad i, j = 1, 2, \ldots, N. \tag{4.1}$$

For future use we introduce the Laplace-Stieltjes transform (LST)

$$\tilde{G}_{ij}(s) = \mathbb{E}[\mathrm{e}^{-sG^{(n)}} 1_{\{J_{n+1}=j\}} | J_n = i], \quad \mathrm{Re}\, s \geq 0, \quad i, j = 1, 2, \ldots, N, \tag{4.2}$$

where $1_{\{.\}}$ denotes the indicator function. In particular,

$$P_{ij} = G_{ij}(\infty) = \mathbb{P}(J_{n+1} = j | J_n = i), \quad i, j = 1, 2, \ldots, N. \tag{4.3}$$

The type of a customer, and its service time, do not depend on the arrival process.

It should be observed that $\{J_n,\ n = 1, 2, \ldots\}$ forms a finite-state Markov chain. We shall restrict ourselves to irreducible Markov chains. The stationary distribution $\mathbb{P}(J = j)$ of the Markov chain $J_n$ is given by the unique solution of the set of equations

$$\mathbb{P}(J = j) = \sum_{i=1}^{N} \mathbb{P}(J = i) P_{ij}, \quad j = 1, 2, \ldots, N, \tag{4.4}$$

with normalizing condition $\sum_{j=1}^{N} \mathbb{P}(J = j) = 1$.
The mean service time of an arbitrary customer is given by

$$\mathbb{E}[G] := \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{P}(J_n = i) \mathbb{E}[G^{(n)} 1_{\{J_{n+1}=j\}} | J_n = i]. \tag{4.5}$$

The stability condition for this model is given by

$$\rho := \lambda \mathbb{E}[B]\mathbb{E}[G] < 1. \tag{4.6}$$

This can be formalized using Theorem 3 from Loynes [75], by describing the workload process in terms of "super customers" whose service times are the aggregate service times of customers in a single batch. Let $\mathcal{G}^{(m)}$ be the service time of the super customer corresponding to the $m$-th arriving batch, and $\mathcal{J}_m$ the type of the first customer in the $m$-th batch. Starting from a stationary version of the sequence $(G^{(n)}, J_{n+1})$, one can readily construct a stationary sequence $(\mathcal{G}^{(m)}, \mathcal{J}_{m+1})$ for the super customers. Note that by construction $\mathcal{G}^{(m)}$ is also stationary and, together with the arrival epochs of batches (which form an independent Poisson process), this sequence completely determines the workload process. This description of the workload process satisfies the criteria to use the characterization for stability in Loynes [75].

We will investigate the queue length process at departure times of customers. For that it will be convenient to define $A_n$ as the number of customers arriving during the service time of the $n$-th customer and $B_n$ as the size of the batch in which the $n$-th customer arrived. Note that for $i, j = 1, 2, \ldots, N$, $|z| \leq 1$,

$$A_{ij}(z) := \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i] = \tilde{G}_{ij}(\lambda(1 - B(z))). \tag{4.7}$$

The queue length distribution at customer departure times is fully determined by the sequences $A_n$ and $B_n$. For the analysis, it is not needed that the arrivals during service times occur in batches at Poisson instants. For that reason we will now formulate our general model in terms of the $A_n$ and $B_n$ only; to specify our later results for the $M^X/G/1$ queue with semi-Markov services, we will simply substitute the relation given in (4.7).

### 4.2.2 General model

The inputs to our general model are probability generating functions of non-negative discrete random variables $A_{ij}(z)$, $i, j \in \{1, 2, \ldots, N\}$, and $B(z)$. From the $A_{ij}(z)$, we construct a Markov process $(A_n, J_{n+1})$, $n = 1, 2, \ldots$, satisfying

$$\mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i] = A_{ij}(z). \tag{4.8}$$

In this construction it is implicit that $(A_n, J_{n+1})$ conditional on $J_n$ is independent of $A_{n-1}$. The sequence $B_n$ is i.i.d. with generating function $B(z)$ and independent of the sequence $A_n$.

Next we define the recurrence relation

$$X_n = \begin{cases} X_{n-1} - 1 + A_n & \text{if } X_{n-1} \geq 1 \\ A_n + B_n - 1 & \text{if } X_{n-1} = 0 \end{cases}, \quad n = 1, 2, 3, \ldots. \tag{4.9}$$

*Note:* If the $A_{ij}(z)$ are set equal to (4.7), then the sequence $X_n$ follows the same law as the number of customers at departure times in the $M^X/G/1$ queue with semi-Markov services. The role of the $B_n$ is subtle in this representation: $B_n$ is only included if the $(n-1)$-th customer leaves the system empty upon departure. The $n$-th customer is therefore the first to be served of a batch that arrives into an empty system. Only for that reason, the sequence $B_n$ can be taken independent of the $A_n$ in the $M^X/G/1$ queue with semi-Markov services.

In the sequel we will study the transient and stationary distributions of $X_n$ defined by (4.9). Again using Theorem 3 of Loynes [75], we may conclude that the stability condition in this case is

$$\rho := \mathbb{E}[A] < 1. \tag{4.10}$$

Here $\mathbb{E}[A]$ denotes the expectation of the $A_n$ in stationarity:

$$\mathbb{E}[A] = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{P}(J = i)\alpha_{ij},$$

with

$$\alpha_{ij} = \mathbb{E}[A_n 1_{\{J_{n+1}=j\}} | J_n = i] = A'_{ij}(1). \tag{4.11}$$

Note that at first sight (4.9) does not seem to fit the framework in Loynes [75], because of the special condition when the system is empty. However, the behavior when the system is empty (or more generally, the behavior of the system in any finite set of states) is irrelevant for stability.

## 4.3 The queue length distribution at departure epochs

We shall determine the transient and steady-state joint distribution of the number of customers immediately after a departure, and the type of the next customer to be served. From the recurrence relation (4.9) we find for the probability generating functions:

$$\begin{aligned}
&\mathbb{E}[z^{X_n}1_{\{J_{n+1}=j\}}] = \mathbb{E}[z^{X_{n-1}-1+A_n}1_{\{J_{n+1}=j\}}1_{\{X_{n-1}\geq 1\}}] \\
&\quad + \mathbb{E}[z^{A_n+B_n-1}1_{\{J_{n+1}=j\}}1_{\{X_{n-1}=0\}}] \\
=&\mathbb{E}[z^{X_{n-1}-1+A_n}1_{\{J_{n+1}=j\}}] - \frac{1}{z}\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}1_{\{X_{n-1}=0\}}] \\
&\quad + \mathbb{E}[z^{A_n+B_n-1}1_{\{J_{n+1}=j\}}1_{\{X_{n-1}=0\}}] \\
=&\sum_{i=1}^{N}\mathbb{E}[z^{X_{n-1}-1+A_n}1_{\{J_{n+1}=j\}}|J_n=i]\mathbb{P}(J_n=i) \\
&\quad - \frac{1}{z}\sum_{i=1}^{N}\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}1_{\{X_{n-1}=0\}}|J_n=i]\mathbb{P}(J_n=i) \\
&\quad + \sum_{i=1}^{N}\mathbb{E}[z^{A_n+B_n-1}1_{\{J_{n+1}=j\}}1_{\{X_{n-1}=0\}}|J_n=i]\mathbb{P}(J_n=i), \\
&\quad \text{for } n=1,2,3,\ldots, \quad j=1,2,\ldots,N.
\end{aligned}$$

Now we exploit the fact that $X_{n-1}$ and $(A_n, J_{n+1})$ are conditionally independent given $J_n$, and the $B_n$ are also independent of all other random variables:

$$\begin{aligned}
&\mathbb{E}[z^{X_n}1_{\{J_{n+1}=j\}}] \\
=&\sum_{i=1}^{N}\mathbb{E}[z^{X_{n-1}-1}|J_n=i]\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}|J_n=i]\mathbb{P}(J_n=i) \\
&\quad + \frac{B(z)-1}{z}\sum_{i=1}^{N}\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}|J_n=i]\mathbb{P}(X_{n-1}=0|J_n=i)\mathbb{P}(J_n=i) \\
=&\frac{1}{z}\sum_{i=1}^{N}\mathbb{E}[z^{X_{n-1}}1_{\{J_n=i\}}]\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}|J_n=i] \\
&\quad + \frac{B(z)-1}{z}\sum_{i=1}^{N}\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}|J_n=i]\mathbb{P}(X_{n-1}=0|J_n=i)\mathbb{P}(J_n=i),
\end{aligned}$$

$$\text{for } n = 1, 2, 3, \ldots, \quad j = 1, 2, \ldots, N. \tag{4.12}$$

### 4.3.1 Steady-state analysis

In this subsection, we restrict ourselves to the steady-state queue length distribution, assuming that the stability condition (4.10) holds. In the next subsection, we will analyze the transient behavior of the queue length.
It will be useful to introduce some further notation: for $i = 1, 2, \ldots, N$,

$$A_i(z) = \sum_{j=1}^{N} A_{ij}(z), \tag{4.13}$$

and

$$\alpha_i = \sum_{j=1}^{N} \alpha_{ij}, \tag{4.14}$$

where the $\alpha_{ij}$ are defined in (4.11). Furthermore, for $j = 1, 2, \ldots, N$, $|z| \leq 1$:

$$f_j(z) = \lim_{n\to\infty} \mathbb{E}[z^{X_n} 1_{\{J_{n+1}=j\}}], \tag{4.15}$$

$$f_j(0) = \lim_{n\to\infty} \mathbb{P}(X_n = 0, J_{n+1} = j), \tag{4.16}$$

and note that

$$f_j(1) = \lim_{n\to\infty} \mathbb{P}(J_{n+1} = j) = \mathbb{P}(J = j). \tag{4.17}$$

The probability generating function of the steady-state queue length distribution immediately after a departure is denoted by

$$F(z) = \sum_{j=1}^{N} f_j(z). \tag{4.18}$$

In steady state, Equation (4.12) leads to the following $N$ equations, for $j = 1, 2, \ldots, N$:

$$(z - A_{jj}(z))f_j(z) - \sum_{i=1, i\neq j}^{N} A_{ij}(z) f_i(z) = (B(z) - 1) \sum_{i=1}^{N} A_{ij}(z) f_i(0). \tag{4.19}$$

We can also write these $N$ linear equations in matrix form as

$$M(z)^T f(z) = b(z),$$

where

$$M(z) = \begin{bmatrix} z - A_{11}(z) & -A_{12}(z) & \dots & -A_{1N}(z) \\ -A_{21}(z) & z - A_{22}(z) & \dots & -A_{2N}(z) \\ \dots & \dots & \dots & \dots \\ -A_{N1}(z) & -A_{N2}(z) & \dots & z - A_{NN}(z) \end{bmatrix}, \tag{4.20}$$

$$f(z) = \begin{bmatrix} f_1(z) \\ f_2(z) \\ \dots \\ f_N(z) \end{bmatrix}, \quad b(z) = (B(z) - 1) \begin{bmatrix} \sum_{i=1}^N A_{i1}(z) f_i(0) \\ \sum_{i=1}^N A_{i2}(z) f_i(0) \\ \dots \\ \sum_{i=1}^N A_{iN}(z) f_i(0) \end{bmatrix}.$$

Therefore, using Cramer's rule, solutions of the non-homogeneous linear system $M(z)^T f(z) = b(z)$ are in the form:

$$f(z) = \frac{1}{\det M(z)^T} (\text{cof } M(z)^T)^T b(z), \text{ provided } \det M(z) \neq 0. \tag{4.21}$$

Here cof $M(z)^T$ is the cofactor matrix of $M(z)^T$. It remains to find the values of $f_1(0), f_2(0), \dots, f_N(0)$. We shall derive $N$ linear equations for $f_1(0), f_2(0), \dots, f_N(0)$.

**First equation:**

Note that $M(z)^T f(z) = b(z)$, which implies that

$$\lim_{z \uparrow 1} \frac{1}{z-1} \hat{e} M(z)^T f(z) = \lim_{z \uparrow 1} \frac{1}{z-1} \hat{e} b(z),$$

where $\hat{e}$ is a row vector with all entries one.

After simplification, we can write this as

$$\lim_{z \uparrow 1} \frac{\sum_{i=1}^N \left( z - \sum_{j=1}^N A_{ij}(z) \right) f_i(z)}{z-1} = \lim_{z \uparrow 1} \frac{B(z) - 1}{z-1} \sum_{j=1}^N \sum_{i=1}^N A_{ij}(z) f_i(0).$$

Using $\sum_{i=1}^{N} f_i(1) = 1$ and $\sum_{i=1}^{N} f_i(1)\alpha_i = \rho$, and after simplification, we get,

$$\sum_{i=1}^{N} f_i(0) = \frac{1-\rho}{\mathbb{E}[B]}. \tag{4.22}$$

**$(N-1)$ remaining equations:**
To find the remaining $N-1$ equations, we first prove that $\det M(z)$ has exactly $N-1$ zeros in $|z|<1$ and the zero $z=1$ on $|z|=1$. Since $f_i(z)$ is an analytic function in $|z|<1$, the numerator of $f_i(z)$ also has $N-1$ zeros in the unit disc $|z|<1$. As a consequence, these $N-1$ zeros provide $N-1$ linear equations for $f_1(0), f_2(0), \ldots, f_N(0)$.

To find the $N-1$ zeros, we use a method that has also been applied in [5, 46, 96]. It is based on the concept of (strict) diagonal dominance in a matrix. The proof consists of 4 steps:
Step 1: Prove that each element on the diagonal of $M(z)$ has exactly one zero in $|z|<1$.
Step 2: Introduce a matrix $M(t,z)$, $0 \le t \le 1$, with $M(1,z) = M(z)$, and prove strict diagonal dominance of $M(t,z)$, i.e., each diagonal element of $M(t,z)$ is in absolute value larger than the sum of the absolute values of the non-diagonal terms in the same row of the matrix.
Step 3: Prove that $\det M(t,z)$ has exactly $N$ zeros in $|z|<1$ and none on $|z|=1$ for $0 \le t < 1$.
Step 4: Use continuity of $\det M(t,z)$ in $t$ for $0 \le t < 1$ to prove that, indeed, $\det M(z)$ has $N-1$ zeros in $|z|<1$ and one zero $z=1$ on $|z|=1$.

**Step 1:** Prove that each element on the diagonal of $M(z)$ has exactly one zero in $|z|<1$.
It follows from (4.20) that $M(z) = D(z) + O(z)$, where $D(z)$ is the diagonal matrix

$$D(z) = \begin{bmatrix} z - A_{11}(z) & 0 & \ldots & 0 \\ 0 & z - A_{22}(z) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & z - A_{NN}(z) \end{bmatrix}, \tag{4.23}$$

and $O(z)$ is the off-diagonal matrix which corresponds to $M(z)$.

**Proposition 4.1.** $\det D(z)$ *has exactly* $N$ *zeros (counting multiplicities) in* $|z| < 1$ *and none satisfying* $|z| = 1$.

*Proof.* First observe that $\det D(z) = \prod_{i=1}^{N}(z - A_{ii}(z))$. Because $|\frac{A_{ii}(z)}{z}| \leq P_{ii} < 1$ on $|z| = 1$, Rouché's theorem implies that the numbers of zeros of $z$ and $z - A_{ii}(z)$ are the same in $|z| < 1$. $z$ has exactly one zero in $|z| < 1$, and hence $z - A_{ii}(z)$ also has exactly one zero in $|z| < 1$, for $i = 1, 2, \ldots, N$.
On $|z| = 1$, $|z - A_{ii}(z)|$ has no zeros, because $|z - A_{ii}(z)| \geq |z| - |A_{ii}(z)| \geq 1 - P_{11} > 0$.
Hence $\det D(z)$ has $N$ zeros in $|z| < 1$ and none on $|z| = 1$.

Now we define the matrix $M(t, z) := D(z) + tO(z)$, where $0 \leq t \leq 1$ is a real parameter. Note that $M(0, z) = D(z)$ and $M(1, z) = M(z)$.

**Step 2:** Prove diagonal dominance for matrix $M(t, z)$.

**Proposition 4.2.** $\det M(t, z) \neq 0$ *for* $0 \leq t < 1, |z| = 1$ *and for* $t = 1, |z| = 1, z \neq 1$.

*Proof.* Consider an arbitrary $i \in \{1, 2, \ldots, N\}$.

$$\begin{aligned} |z - A_{ii}(z)| &\geq |z| - |A_{ii}(z)| \\ &\geq 1 - P_{ii} = \sum_{j \neq i} P_{ij} > t \sum_{j \neq i} P_{ij} \qquad \text{for } 0 \leq t < 1, |z| = 1. \end{aligned} \tag{4.24}$$

On the other hand, $\sum_{j \neq i} |tA_{ij}(z)| \leq t \sum_{j \neq i} P_{ij}$ for $0 \leq t < 1, |z| = 1$.
Therefore, $|z - A_{ii}(z)| > |t \sum_{j \neq i} A_{ij}(z)|$ for $0 \leq t < 1, |z| = 1$. This holds for $i = 1, 2, \ldots, N$.
Thus, $M(t, z)$ is strictly diagonally dominant. This implies that $M(t, z)$ is a non-singular matrix, i.e., $\det M(t, z) \neq 0$, for $0 \leq t < 1, |z| = 1$. This concludes the proof for the case $0 \leq t < 1$, with $|z| = 1$.
We next turn to the case $t = 1, |z| = 1, z \neq 1$, again considering an arbitrary $i \in \{1, 2, \ldots, N\}$. Now (4.24) is replaced by $|z - A_{ii}(z)| > \sum_{j \neq i} P_{ij}$ for $|z| = 1, z \neq 1$. On the other hand, $\sum_{j \neq i} |A_{ij}(z)| < \sum_{j \neq i} P_{ij}$. Therefore, $|z - A_{ii}(z)| > |\sum_{j \neq i} A_{ij}(z)|$ for $|z| = 1, z \neq 1$. This holds for $i = 1, 2, \ldots, N$.
In this way we have proven the strict diagonal dominance, and hence the non-singularity, also for $t = 1, |z| = 1, z \neq 1$.

**Step 3:** Prove that $\det M(t,z)$ has exactly $N$ zeros in $|z|<1$ and none on $|z|=1$ for $0 \le t < 1$.

**Proposition 4.3.** *The function* $\det M(t,z)$ *has exactly* $N$ *zeros in* $|z|<1$ *and none on* $|z|=1$ *for* $0 \le t < 1$.

*Proof.* Let $n(t)$ be the number of zeros of $\det M(t,z)$ in $|z|<1$. By the argument principle, see Evgrafov [41, p. 97],

$$n(t) = \frac{1}{2\pi i} \int_{|z|=1} \frac{\frac{\partial}{\partial z} \det M(t,z)}{\det M(t,z)} \mathrm{d}z, \tag{4.25}$$

where it should be noticed that $\det M(t,z) \neq 0$ on $|z|=1$ for $0 \le t < 1$ according to Proposition 4.2. Here, $n(t)$ is a continuous integer-valued function of $t$ for $0 \le t < 1$ and $n(0) = N$ according to Proposition 4.1. So $n(t) = n(0) = N$.

From the above we may conclude that $\det M(1,z) = M(z)$ has *at least* $N$ zeros in the closed unit disc, because the zeros of $\det M(t,z)$ are continuous functions for $0 \le t \le 1$. Finally we need to prove that there are *exactly* $N$ zeros in $|z| \le 1$, one of which ($z=1$) lies on $|z|=1$.

**Step 4:** Use continuity of $\det M(t,z)$ in $t$ for $0 \le t \le 1$ to prove that $\det M(z)$ has $N-1$ zeros in $|z|<1$ and one zero $z=1$ on $|z|=1$.

**Proposition 4.4.** $\frac{d}{dz}\{\det M(z)\}|_{z=1} > 0$ *and* $z=1$ *is a simple zero of* $\det M(z)$.

*Proof.* Firstly, $z=1$ is a zero of $\det M(z)$. Now we show that it is a simple zero. Use that $\lim_{z\uparrow 1} \frac{\det\ M(z)}{z-1} = \frac{d}{dz}\{\det\ M(z)\}|_{z=1} > 0$, where the inequality is a consequence of the stability condition. Hence, $z=1$ is a simple zero of $\det M(z)$.

**Proposition 4.5.** $\det M(t,1) > 0$ *for* $0 \le t < 1$.

*Proof.* We shall exploit the fact that $\det M(t,1)$ is the product of all eigenvalues of $M(t,1)$. So we need to prove that the product of these eigenvalues is positive.

Consider the matrix $I - M(t,1)$, where $I$ is the identity matrix:

$$I - M(t,1) = \begin{bmatrix} P_{11} & tP_{12} & tP_{13} & \cdots & tP_{1N} \\ tP_{21} & P_{22} & tP_{23} & \cdots & tP_{2N} \\ tP_{31} & tP_{32} & P_{33} & \cdots & tP_{3N} \\ \vdots & \vdots & \vdots & & \vdots \\ tP_{N1} & tP_{N2} & tP_{N3} & & P_{NN} \end{bmatrix}.$$

Note that $I - M(t,1)$ is a substochastic matrix, so if $z$ is an eigenvalue of the matrix $I - M(t,1)$, it lies in the unit disc $|z| < 1$. Hence every eigenvalue $z$ of the matrix $M(t,1)$ lies in $|z-1| < 1$. $M(t,1)$ is a real matrix, so if $M(t,1)$ has a complex eigenvalue, then the conjugate of this complex eigenvalue is also one of the eigenvalues of $M(t,1)$. This implies that if $M(t,1)$ has complex eigenvalues, then the product of these complex eigenvalues is positive. The product of the real eigenvalues is also positive because every eigenvalue $z$ of the matrix $M(t,1)$ lies in $|z-1| < 1$. This concludes the proof.

**Proposition 4.6.** *The function* $\det M(z)$ *has exactly* $N-1$ *zeros in* $|z| < 1$ *and one zero on* $|z| = 1$ *(at* $z = 1$*).*

*Proof.* We follow the argument of Gail et al. [46, p. 372]. By letting $t \to 1$ in Proposition 4.3, it follows that $\det M(z)$ has at least $N$ zeros in $|z| \leq 1$. By Proposition 4.4, given $\epsilon > 0$, there is a real $z'$, $1 - \epsilon < z' < 1$, such that $\det M(z')$ is negative. By continuity, there is a real $t'$, $1 - \epsilon < t' < 1$, such that $\det M(t', z')$ is negative. Since $\det M(t', 1)$ is positive according to Proposition 4.5, there is a real $z''$, $z' < z'' < 1$ with $\det M(t', z'') = 0$. As $t \to 1$, one of the $N$ zeros of $\det M(t,z)$ from inside the unit disc must be approaching the zero of $\det M(z)$ at $z = 1$. In summary, as $t \to 1$, the limiting positions of the $N$ zeros of $\det M(t,z)$ are: one at $z = 1$ and the other $N-1$ in $|z| < 1$.

### 4.3.2 Transient analysis

In this subsection, we shall determine the transient behavior of the probability generating function of the number of customers. The analysis proceeds largely

analogous to the stationary case. In fact, for the transient analysis, it turns out to be less involved to demonstrate the location of the roots. We define

$$f_j(r,z) = \sum_{n=0}^{\infty} r^n \mathbb{E}[z^{X_n} 1_{\{J_{n+1}=j\}}] \quad \text{for} \quad |r| < 1, j = 1, 2, ..., N, \tag{4.26}$$

so that,

$$f_j(r,0) = \sum_{n=0}^{\infty} r^n \mathbb{P}(X_n = 0, J_{n+1} = j). \tag{4.27}$$

Using (4.12) with $\mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i] = A_{ij}(z)$ in (4.26), we get

$$\begin{aligned} f_j(r,z) =& \mathbb{E}[z^{X_0} 1_{\{J_1=j\}}] + \frac{1}{z} \sum_{i=1}^{N} A_{ij}(z) \sum_{n=1}^{\infty} r^n \mathbb{E}[z^{X_{n-1}} 1_{\{J_n=i\}}] \\ &+ \left( \frac{B(z)-1}{z} \right) \sum_{i=1}^{N} A_{ij}(z) \sum_{n=1}^{\infty} r^n \mathbb{P}(X_{n-1} = 0, J_n = i) \\ =& z^{x_0} \mathbb{P}(J_1 = j) + \frac{1}{z} \sum_{i=1}^{N} A_{ij}(z) \sum_{n=0}^{\infty} r^{n+1} \mathbb{E}[z^{X_n} 1_{\{J_{n+1}=i\}}] \\ &+ r \left( \frac{B(z)-1}{z} \right) \sum_{i=1}^{N} A_{ij}(z) f_i(r,0), \end{aligned}$$

provided the initial number of customers in the system is deterministic and equal to $x_0$.

Using (4.26) and after simplification, we get the following $N$ equations:

$$\begin{aligned} (z - rA_{jj}(z)) f_j(r,z) - r \sum_{i=1, i\neq j}^{N} A_{ij}(z) f_i(r,z) &= z^{X_0+1} \mathbb{P}(J_1 = j) \\ + r\,(B(z)-1) \sum_{i=1}^{N} A_{ij}(z) f_i(r,0), \qquad & j = 1, 2, \ldots, N. \end{aligned} \tag{4.28}$$

We can also write these $N$ linear equations in matrix form as

$$M(r,z)^T f(r,z) = b(r,z),$$

where

$$M(r,z) = \begin{bmatrix} z - rA_{11}(z) & -rA_{12}(z) & \dots & -rA_{1N}(z) \\ -rA_{21}(z) & z - rA_{22}(z) & \dots & -rA_{2N}(z) \\ \dots & \dots & \dots & \dots \\ -rA_{N1}(z) & -rA_{N2}(z) & \dots & z - rA_{NN}(z) \end{bmatrix},$$

$$f(r,z) = \begin{bmatrix} f_1(r,z) \\ f_2(r,z) \\ \dots \\ f_N(r,z) \end{bmatrix},$$

$$b(r,z) = z^{X_0+1} \begin{bmatrix} \mathbb{P}(J_1 = 1) \\ \mathbb{P}(J_1 = 2) \\ \dots \\ \mathbb{P}(J_1 = N) \end{bmatrix} + r(B(z) - 1) \begin{bmatrix} \sum_{i=1}^N A_{i1}(z) f_i(r,0) \\ \sum_{i=1}^N A_{i2}(z) f_i(r,0) \\ \dots \\ \sum_{i=1}^N A_{iN}(z) f_i(r,0) \end{bmatrix}.$$

Therefore, using Cramer's rule, solutions of the non-homogeneous linear system $M(r,z)^T f(r,z) = b(r,z)$ are in the form:

$$f(r,z) = \frac{1}{\det M(r,z)^T} (\text{cof M(r,z)}^{\text{T}})^{\text{T}} \text{b(r,z)}, \text{ provided } \det \text{M(r,z)} \neq 0. \tag{4.29}$$

It remains to find the values of $f_1(r,0), f_2(r,0), \dots, f_N(r,0)$. We shall derive $N$ linear equations for $f_1(r,0), f_2(r,0), \dots, f_N(r,0)$.

To find $N$ linear equations for $f_1(r,0), f_2(r,0), \dots, f_N(r,0)$, we fist prove that $\det M(r,z)$ has exactly $N$ zeros for fixed $r$ in $|z| < 1$. Since $M(r,z) = zI - rA(z)$, $\det M(r,z)$ is a continuous function in $r$ for $0 \leq r \leq 1$, and therefore the zeros are continuous in $0 \leq r \leq 1$.

**Remark 4.1.** *It is worth emphasizing that it is at this point that our approach is different from the analysis by Neuts [82] and Çinlar [23]. We do not require for each pair of elementary roots that they either be strictly different for all values of* $0 \leq r \leq 1$ *or coincide for all* $0 \leq r \leq 1$. *The main price to pay is that we can not use that the roots are analytic in* $r$ *and we can therefore not obtain the stationary distribution from the transient distribution as* $r \to 1$.

Compared to the steady-state analysis, the proof is simpler and only consists of two steps:

**Step 1:** Prove diagonal dominance of the matrix $M(r,z)$.

**Proposition 4.7.** $\det M(r,z) \neq 0$ *for* $0 \leq r < 1, |z| = 1$.

*Proof.* Consider an arbitrary $i \in \{1, 2, \ldots, N\}$.

$$\begin{aligned} |z - rA_{ii}(z)| &\geq |z| - r|A_{ii}(z)| \\ &> 1 - P_{ii} = \sum_{j \neq i} P_{ij} > r \sum_{j \neq i} P_{ij} \qquad \text{for } 0 \leq r < 1, |z| = 1. \end{aligned} \tag{4.30}$$

On the other hand, $\sum_{j \neq i} |rA_{ij}(z)| \leq r \sum_{j \neq i} P_{ij}$ for $0 \leq r < 1, |z| = 1$.
Therefore, $|z - rA_{ii}(z)| > |r \sum_{j \neq i} A_{ij}(z)|$ for $0 \leq r < 1, |z| = 1$. This holds for $i = 1, 2, \ldots, N$.
Thus, $M(r,z)$ is strictly diagonally dominant. This implies that $M(r,z)$ is a non-singular matrix, i.e., $\det M(r,z) \neq 0$, for $0 \leq r < 1, |z| = 1$. This completes the proof.

**Step 2:** Prove that $\det M(r,z)$ has exactly $N$ zeros in $|z| < 1$ for $0 \leq r < 1$.

**Proposition 4.8.** *The function* $\det M(r,z)$ *has exactly* $N$ *zeros in* $|z| < 1$ *for* $0 \leq r < 1$.

*Proof.* Let $n(r)$ be the number of zeros of $\det M(r,z)$ in $|z| < 1$. As before, by the argument principle [41, p. 97],

$$n(r) = \frac{1}{2\pi i} \int_{|z|=1} \frac{\frac{\partial}{\partial z} \det M(r,z)}{\det M(r,z)} \mathrm{d}z, \tag{4.31}$$

where it should be noticed that $\det M(r,z) \neq 0$ on $|z| = 1$ for $0 \leq r < 1$ according to Proposition 4.7. Here, $n(r)$ is a continuous integer-valued function of $r$ for $0 \leq r < 1$ and $n(0) = N$ because $\det M(0,z) = z^n$ . So $n(r) = n(0) = N$.

## 4.4 Poisson batch arrivals: stationary queue length at arrival and arbitrary epochs

In the previous section, we determined the stationary and the transient queue length distributions at departure times of customers. In the general framework, the exact arrival process of customers is not specified, but for the model with Poisson batch arrivals, we can obtain the stationary queue length distribution at *arbitrary time*, at *batch arrival instants* and at *customer arrival instants*. Because of PASTA, the distribution of the number of customers already in system just before a new batch arrives (let us denote this by a generic random variable $X^{ba}$) coincides with the distribution of the number of customers in the system at an arbitrary time ($X^{arb}$). The number of customers at customer arrival instants (denoted with $X^{ca}$) needs to be further specified, because with batch arrivals all customers in the same batch have the same arrival time. As noted previously, customers within one batch are assumed to be (randomly) ordered. Although they arrive at the same time, they see different numbers of customers in front of them. In particular, the last customer in a batch sees all the customers that were already in the system *plus* all other customers (excluding him/her) arriving in the same batch. In the *customer average* distribution at arrival times, this must be taken into account. In Figure 4.1 we depict three batch arrivals, two of which contain multiple customers and thus coincide with more than one customer arrival. Applying a simple level crossing argument

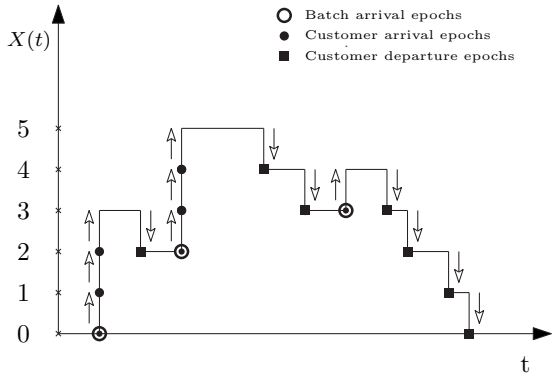



**Figure 4.1.** Up and down-crossing.

with the aid of Figure 4.1, it is readily seen that the distributions of $X$ (at

departure times) and $X^{ca}$ must coincide: indeed, for each level $k = 1, 2, \ldots$, customer departures that decrease the queue length from $k$ to $k-1$ must be matched by customer arrivals increasing the level from $k-1$ to $k$ (since the arrival of each customer within a batch is counted separately, the difference can be at most 1 which is negligible in the long run).
We can also link the distributions of $X^{ba}$ and $X^{ca}$: A customer in an arriving batch sees in front of him the number of customers already in system ($X^{ba}$) and the number of customers in front of him in the same batch. For an arbitrary customer in the batch the number of customers in front of him in the same batch has the forward recurrence distribution of $B$. Summarizing,

$$\mathbb{E}[z^X] = \mathbb{E}[z^{X^{ca}}] = \mathbb{E}[z^{X^{ba}}] \frac{1 - B(z)}{\mathbb{E}[B](1-z)}, \tag{4.32}$$

where we use independence of the batch size and the number of customers already in system, and

$$\mathbb{E}[z^{X^{arb}}] = \mathbb{E}[z^{X^{ba}}]. \tag{4.33}$$

From these relations we can obtain all the required distributions. It can be verified that these distributions agree with the results from Chaudhry [25] for the model without dependencies between successive service times.

## 4.5 The queueing model with two customer types: departure epochs

In this section, we restrict ourselves to the case of two customer types, i.e., $N = 2$. In this case, we are able to give an explicit and compact expression for the probability generating function of the number of customers in the system immediately after a departure. For large values of $N$, the algebraic becomes tedious and gives no additional insight. For the steady-state behavior it follows from (4.19) that:

$$\begin{aligned} f_1(z) =& \frac{f_1(0)(B(z)-1)\Big(zA_{11}(z) + A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)} \\ &+ \frac{z f_2(0) A_{21}(z)(B(z)-1)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)}, \end{aligned} \tag{4.34}$$

$$f_2(z) = \frac{f_2(0)(B(z)-1)\Big(zA_{22}(z) + A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)}{(z-A_{11}(z))(z-A_{22}(z)) - A_{12}(z)A_{21}(z)} + \frac{zf_1(0)A_{12}(z)(B(z)-1)}{(z-A_{11}(z))(z-A_{22}(z)) - A_{12}(z)A_{21}(z)}, \tag{4.35}$$

where

$$f_1(0) = \frac{1-\rho}{\mathbb{E}[B]}\frac{A_{11}(\hat{z}) - \hat{z}}{A_{11}(\hat{z}) + A_{12}(\hat{z}) - \hat{z}}, \quad f_2(0) = \frac{1-\rho}{\mathbb{E}[B]}\frac{A_{22}(\hat{z}) - \hat{z}}{A_{21}(\hat{z}) + A_{22}(\hat{z}) - \hat{z}}, \tag{4.36}$$

such that $f_1(0) + f_2(0) = \frac{1-\rho}{\mathbb{E}[B]}$ and $z = \hat{z}$ is the zero of $(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)$ with $|\hat{z}| < 1$.

It is noted that the probability generating function of $X_n$ in steady state is

$$F(z) = \lim_{n\to\infty}\mathbb{E}[z^{X_n}].$$

From Equation (4.18), for $N = 2$, we can write $F(z)$ as the sum of $f_1(z)$ and $f_2(z)$, i.e.,

$$F(z) = f_1(z) + f_2(z).$$

After substituting the values of $f_1(z)$ and $f_2(z)$ from Equations (4.34) and (4.35) respectively, we obtain $F(z)$ as

$$F(z) = \frac{z(B(z)-1)\Big(f_1(0)(A_{11}(z) + A_{12}(z)) + f_2(0)(A_{21}(z) + A_{22}(z))\Big)}{(z-A_{11}(z))(z-A_{22}(z)) - A_{12}(z)A_{21}(z)} + \frac{(B(z)-1)(f_1(0)+f_2(0))\Big(A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)}{(z-A_{11}(z))(z-A_{22}(z)) - A_{12}(z)A_{21}(z)}.$$

Equation (4.13) states that $A_i(z) = A_{i1}(z) + A_{i2}(z)$ for $i = 1, 2$. After substituting the values of $f_i(0)$ and $A_i(z)$ for $i = 1, 2$, $F(z)$ becomes

$$F(z) = \frac{z(B(z)-1)(1-\rho)\Big(c_1A_1(z) + c_2A_2(z)\Big)}{\mathbb{E}[B]\Big((z-A_{11}(z))(z-A_{22}(z)) - A_{12}(z)A_{21}(z)\Big)} + \frac{(B(z)-1)(1-\rho)\Big(A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)}{\mathbb{E}[B]\Big((z-A_{11}(z))(z-A_{22}(z)) - A_{12}(z)A_{21}(z)\Big)}, \tag{4.37}$$

where $c_1 = \frac{A_{11}(\hat{z})-\hat{z}}{A_{11}(\hat{z})+A_{12}(\hat{z})-\hat{z}}, c_2 = \frac{A_{22}(\hat{z})-\hat{z}}{A_{21}(\hat{z})+A_{22}(\hat{z})-\hat{z}}$.

Let us now determine the expected number of customers $\mathbb{E}[X] = F'(1)$.

After differentiating $F(z)$ w.r.t. $z$ and taking the limit $z \uparrow 1$, we get

$$\mathbb{E}[X] = \frac{\rho}{2} + \frac{\mathrm{Var}(A)}{2(1-\rho)} + \frac{\mathbb{E}[B(B-1)]}{2\mathbb{E}[B]} + \frac{-\rho + \mathbb{E}[B](f_1(0)\alpha_1 + f_2(0)\alpha_2) + \rho(\alpha_{11}+\alpha_{22}) + \alpha_{12}\alpha_{21} - \alpha_{11}\alpha_{22}}{(P_{12}+P_{21})(1-\rho)}. \tag{4.38}$$

For the transient distribution it follows from (4.28) that

$$f_1(r,z) = \frac{z^{X_0+1}\Big(z\mathbb{P}(J_1=1) + r(A_{21}(z)\mathbb{P}(J_1=2) - A_{22}(z)\mathbb{P}(J_1=1))\Big)}{\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)} + \frac{rz(B(z)-1)\sum_{i=1}^{2} A_{i1}(z)f_i(r,0)}{\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)} + \frac{r^2(B(z)-1)\Big(A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)f_1(r,0)}{\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)}, \tag{4.39}$$

$$f_2(r,z) = \frac{z^{X_0+1}\Big(z\mathbb{P}(J_1=2) + r(A_{12}(z)\mathbb{P}(J_1=1) - A_{11}(z)\mathbb{P}(J_1=2))\Big)}{\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)} + \frac{rz(B(z)-1)\sum_{i=1}^{2} A_{i2}(z)f_i(r,0)}{\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)} + \frac{r^2(B(z)-1)\Big(A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)f_2(r,0)}{\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)}, \tag{4.40}$$

where

$$f_1(r,0) = \frac{-\hat{z}_1^{X_0}(\hat{B}^{(2)}-1)\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)})\mathbb{P}(J_1=1)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}$$
$$+ \frac{\hat{z}_2^{X_0}(\hat{B}^{(1)}-1)\hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\mathbb{P}(J_1=1)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}$$
$$+ \frac{r\left(\hat{z}_2^{X_0}(\hat{B}^{(1)}-1) - \hat{z}_1^{X_0}(\hat{B}^{(2)}-1)\right)\hat{A}_{21}^{(1)}\hat{A}_{21}^{(2)}\mathbb{P}(J_1=2)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}, \tag{4.41}$$

$$f_2(r,0) = \frac{1}{r}\frac{\hat{z}_1^{X_0}(\hat{B}^{(2)}-1)\left(\hat{z}_1 - r\hat{A}_{22}^{(1)}\right)\left(\hat{z}_2 - r\hat{A}_{22}^{(2)}\right)\mathbb{P}(J_1=1)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}$$
$$+ \frac{1}{r}\frac{-\hat{z}_2^{X_0}(\hat{B}^{(1)}-1)\left(\hat{z}_1 - r\hat{A}_{22}^{(1)}\right)\left(\hat{z}_2 - r\hat{A}_{22}^{(2)}\right)\mathbb{P}(J_1=1)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}$$
$$+ \frac{-\hat{z}_2^{X_0}(\hat{B}^{(1)}-1)\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)})\mathbb{P}(J_1=2)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}$$
$$+ \frac{\hat{z}_1^{X_0}(\hat{B}^{(2)}-1)\hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\mathbb{P}(J_1=2)}{(\hat{B}^{(1)}-1)(\hat{B}^{(2)}-1)\left(\hat{A}_{21}^{(2)}(\hat{z}_1 - r\hat{A}_{22}^{(1)}) - \hat{A}_{21}^{(1)}(\hat{z}_2 - r\hat{A}_{22}^{(2)})\right)}, \tag{4.42}$$

$z = \hat{z}_1$ and $z = \hat{z}_2$ are the zeros in the unit disc $|z| < 1$ of $\Big(z - rA_{11}(z)\Big)\Big(z - rA_{22}(z)\Big) - r^2A_{12}(z)A_{21}(z)$ and $\hat{A}_{ij}^{(1)} := A_{ij}(\hat{z}_1)$, $\hat{A}_{ij}^{(2)} := A_{ij}(\hat{z}_2)$, $\hat{B}^{(i)} := B(\hat{z}_i)$ for $i,j = 1,2$.

**Remark 4.2.** *It can be observed that the first three terms in the right-hand-side of Equation* (4.38) *are exactly equal to the mean queue length at departure epochs of the standard $M^X/G/1$ queue without dependencies, cf. Gaver [49] and Cohen [29, Section III.2.3], and the remaining term appears due to the dependent service times.*

**Corollary 4.1.** *The queue length in the system considered in the present chapter has the same distribution as the number of customers in an $M^X/G/1$ queuing model if $A_1(z) = A_2(z) = A(z)$, cf. Gaver [49] and Cohen [29, Section III.2.3].*

*Proof.* Let $A_1(z) = A_2(z) = A(z)$, and let $X^{M^X/G/1}$ denote the queue length of an $M^X/G/1$ queue with arrival rate $\lambda$ and service-time LST $\tilde{G}^{M^X/G/1}(\lambda(1-B(z))) = A(z)$ where $B$ is the group size. After substituting $A_1(z) = A_2(z) = A(z)$ with $c_1 + c_2 = 1$ in Equation (4.37), and after simplification, we get

$$F(z) = \frac{(1-\rho)(B(z)-1)\Big(zA(z) + A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z)\Big)}{\mathbb{E}[B]\Big(z^2 - (A_{11}(z)+A_{22}(z))z + A_{11}(z)A_{22}(z) - A_{12}(z)A_{21}(z)\Big)}.$$

Note that $A_{11}(z) + A_{12}(z) = A(z)$ and $A_{21}(z) + A_{22}(z) = A(z)$, which implies that $A_{12}(z)A_{21}(z) - A_{11}(z)A_{22}(z) = A(z)(A_{12}(z) - A_{22}(z))$.

Therefore, we can write $F(z)$ as

$$\begin{aligned}
F(z) &= \frac{(1-\rho)(B(z)-1)(z + A_{12}(z) - A_{22}(z))A(z)}{\mathbb{E}[B](z^2 - (A_{11}(z)+A_{22}(z))z + A(z)(A_{22}(z) - A_{12}(z)))} \\
&= \frac{(1-\rho)(B(z)-1)A(z)(z + A_{12}(z) - A_{22}(z))}{\mathbb{E}[B](z^2 - (A(z)+A_{22}(z) - A_{12}(z))z + A(z)(A_{22}(z) - A_{12}(z)))} \\
&= \frac{(1-\rho)(B(z)-1)A(z)(z + A_{12}(z) - A_{22}(z))}{\mathbb{E}[B](z - A(z))(z + A_{12}(z) - A_{22}(z))} \\
&= \frac{(1-\rho)(1-B(z))A(z)}{\mathbb{E}[B](A(z) - z)} \\
&= \mathbb{E}[z^{X^{M^X/G/1}}].
\end{aligned}$$

□

**Corollary 4.2.** *The* expected *number of customers in the system considered in the present chapter is equal to the* expected *number of customers in the corresponding $M^X/G/1$ queueing model if $\alpha_1 = \alpha_2 = \mathbb{E}[A]$, again cf. Gaver [49] and Cohen [29, Section III.2.3].*

*Proof.* Let $\alpha_1 = \alpha_2 = \mathbb{E}[A]$, and let $X^{M^X/G/1}$ denote the queue length of an $M^X/G/1$ queue with arrival rate $\lambda$ and service-time distribution $G^{M^X/G/1}$, satisfying the following two properties:

$$\begin{aligned}\lambda\mathbb{E}[B]\mathbb{E}[G^{M^X/G/1}] &= \mathbb{E}[A] = \rho,\\ (\lambda\mathbb{E}[B])^2\mathrm{Var}(G^{M^X/G/1}) + \lambda\mathbb{E}[B^2]\mathbb{E}[G^{M^X/G/1}] &= \mathrm{Var}(A),\end{aligned}$$

where $B$ is the group size. After substituting $\alpha_1 = \alpha_2 = \mathbb{E}[A] = \rho$ with $f_1(0) + f_2(0) = \frac{1-\rho}{\mathbb{E}[B]}$ in Equation (4.38), and after simplification, we get

$$\begin{aligned}\mathbb{E}[X] =& \frac{\rho}{2} + \frac{\mathrm{Var}(A)}{2(1-\rho)} + \frac{\mathbb{E}[B(B-1)]}{2\mathbb{E}[B]}\\ &+ \frac{-\rho^2 + \rho(\alpha_{11}+\alpha_{22}) + \alpha_{12}\alpha_{21} - \alpha_{11}\alpha_{22}}{(P_{12}+P_{21})(1-\rho)}.\end{aligned}$$

Note that $\alpha_{11} + \alpha_{12} = \alpha_{21} + \alpha_{22} = \rho$, which implies that

$$\begin{aligned}&\alpha_{12}(\alpha_{21}+\alpha_{22}) = \rho\alpha_{12}\\ \Longrightarrow\ &\alpha_{12}\alpha_{21} + (\rho-\alpha_{11})\alpha_{22} = \rho(\rho-\alpha_{11})\\ \Longrightarrow\ &-\rho^2 + \rho(\alpha_{11}+\alpha_{22}) + \alpha_{12}\alpha_{21} - \alpha_{11}\alpha_{22} = 0.\end{aligned}$$

Therefore, we can write the expected number of customers $\mathbb{E}[X]$ as

$$\mathbb{E}[X] = \frac{\rho}{2} + \frac{\mathrm{Var}(A)}{2(1-\rho)} + \frac{\mathbb{E}[B(B-1)]}{2\mathbb{E}[B]} = \mathbb{E}[X^{M^X/G/1}]. \qquad \square$$

## 4.6 Numerical results

In this section, we present four numerical examples in order to get more insight in the consequences of introducing dependencies between the service times of consecutive customers. For simplicity, we restrict ourselves to two customer types ($N = 2$). In all four examples we assume that the overall batch arrival process is a Poisson process with rate $\lambda$ and the load $\rho$ equals $\frac{3}{4}$.

### 4.6.1 Example 1

In this example we consider an almost symmetric system, with $\mathbb{P}(J = 1) = \mathbb{P}(J = 2) = \frac{1}{2}$ and $\alpha_{ij} = \frac{3}{8}$ for $\forall i, j = 1, 2$. It follows that $\mathbb{E}[A] = \frac{3}{4}$, $P_{11} = P_{22}$ and we shall vary $P_{11}$. The batch sizes are geometrically distributed with

$$\mathbb{P}(B = k) = p^{k-1}(1 - p), \qquad k = 1, 2, \ldots$$

We take $p = 3/4$, resulting in a mean batch size of $\mathbb{E}[B] = 4$. The conditional service times are respectively exponential and Erlang distributed random variables, with

$$G_{ij}(x) = \left(1 - \sum_{m=0}^{k_j-1} \frac{(\mu_{ij}x)^m}{m!} e^{-\mu_{ij}x}\right) P_{ij},$$

for $\mu_{ij} > 0$, $i, j = 1, 2$. In this example we will take an Erlang distribution with four phases. If we define

$$k_j = \begin{cases} 1 & \text{if } j = 1, \\ 4 & \text{if } j = 2, \end{cases}$$

we can use Equation (4.8) to obtain

$$A_{ij}(z) = P_{ij} \left(\frac{\mu_{ij}}{\lambda(1 - B(z)) + \mu_{ij}}\right)^{k_j},$$

for $i = 1, 2$ and $j = 1, 2$.

The variance of the number of arrivals during one arbitrary service time, written as a function of $P_{11}$, directly follows. For $0 < P_{11} < 1$,

$$\mathrm{Var}(A) = \frac{75}{16} + \frac{117}{512(1 - P_{11})P_{11}}.$$

We observe that $\alpha_1 = \alpha_2$, but $A_1(z) \neq A_2(z)$. From Corollary 4.2, we know that the mean queue length in our model is equal to the mean queue length of a standard $M^X/G/1$ queue, but for higher moments of the queue length, this equality is not true unless we can construct a case with $A_1(z) = A_2(z)$. This is confirmed by Table 4.1, which depicts numerical values for the means

and variances of the queue lengths in our model and in the corresponding $M^X/G/1$ queue. Indeed, the mean queue lengths of both systems are equal, whereas the variances of the queue lengths are only equal in the case $P_{11} = \frac{1}{2}$, where $A_1(z) = A_2(z)$. Since $\alpha_1 = \alpha_2$, we immediately conclude that the mean queue length and the variance of $A$ are minimal when $P_{11} = 1/2$ (see Corollary 4.2).

| $P_{11}$ | $\mathbb{E}[X] = \mathbb{E}[X^{M^X/G/1}]$ | $\mathrm{Var}(X)$ | $\mathrm{Var}(X^{M^X/G/1})$ |
|---|---|---|---|
| 0.1 | 17.8281 | 374.4642 | 374.4631 |
| 0.3 | 14.9263 | 237.6202 | 237.6198 |
| 0.5 | 14.5781 | 223.8303 | 223.8303 |
| 0.7 | 14.9263 | 237.6184 | 237.6198 |
| 0.9 | 17.8281 | 374.4185 | 374.4631 |

**Table 4.1.** Means and variances of $X$ and $X^{M^X/G/1}$ for various values of $P_{11}$ in Example 1.

### 4.6.2 Example 2

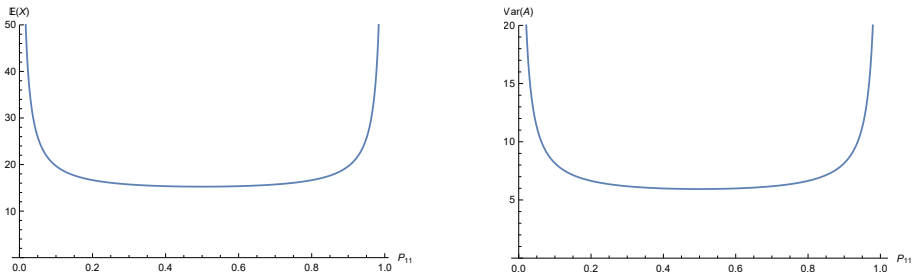
In this example we take a similar setting as in the previous example, but we make two adjustments. First, for even more simplicity, we assume that all conditional service times are exponentially distributed, i.e.,

$$G_{ij}(x) = (1 - e^{-\mu_{ij}x})P_{ij}, \qquad \forall i, j = 1, 2.$$

Secondly, we take $\alpha_{11} = \alpha_{12} = \frac{1}{2}$ and $\alpha_{21} = \alpha_{22} = \frac{1}{4}$. As in the previous example, we let $\mathbb{P}(J = 1) = \mathbb{P}(J = 2) = \frac{1}{2}$. We observe that the difference with Example 1 is that all conditional service-time distributions are exponential now, but with different parameters. Moreover, in this model $\alpha_1 \neq \alpha_2$.

An interesting question is, how the mean queue length and the variance of the number of arrivals during an arbitrary service time are related. Since $\alpha_1 \neq \alpha_2$, the setting of Corollary 4.2 does not apply. In Figure 4.2 we show $\mathbb{E}[X]$ and $\mathrm{Var}(A)$ plotted versus $P_{11}$. When studying the two plots carefully, one can see that the plots are not completely symmetric, which is obviously caused by the asymmetric service times. However, another observation that is

not visible to the human eye, is that the minima of both plots are *not* attained at the same value of $P_{11}$. It can be shown analytically, that the variance of $A$ is minimal at exactly $P_{11} = 1/2$, and numerically, that $\mathbb{E}[X]$ is minimal for $P_{11} \approx 0.500411$. Although this is a small difference, it means that this system exhibits an interesting, perhaps counterintuitive feature: it is possible to obtain a *smaller* mean queue length by having a *greater* variance in the number of arrivals during one service time. In Example 3 we will create a setting in which this effect is more pronounced.




(a) Mean queue length (b) Variance of $A$

**Figure 4.2.** The mean queue length $\mathbb{E}[X]$ and the variance of $A$ in Example 2.

From Figures 4.2(a) and (b), we can observe that, except for the small region where $0.5 < P_{11} < 0.500411$, the expected number of customers is increasing when the variance of the number of arrivals during a customer service time is increasing and conversely. This means that bigger variance of the number of arrivals implies a larger expected number of customers. This also implies that the expected number of customers can grow beyond any bound in a stable system due to the very large variance of the number of arrivals during one service time. This scenario occurs when $P_{11}$ tends to 0 or 1 in Figure 4.2. Therefore, we can observe dependencies when $P_{11}$ or $(1 - P_{11})$ is small. Otherwise, $\mathbb{E}[X]$ and $\mathrm{Var}(A)$ appear to be rather insensitive to the value of $P_{11}$.

Of course, the reason for the large variance in the number of arrivals during a customer service time lies in the dependence. When, e.g., $P_{11} = P_{22}$ is very small, services alternate for a long time between $\exp(\mu_{12})$ and $\exp(\mu_{21})$ services with small mean; rarely is there an $\exp(\mu_{11})$ or $\exp(\mu_{22})$ service which has a huge mean.

### 4.6.3 Example 3

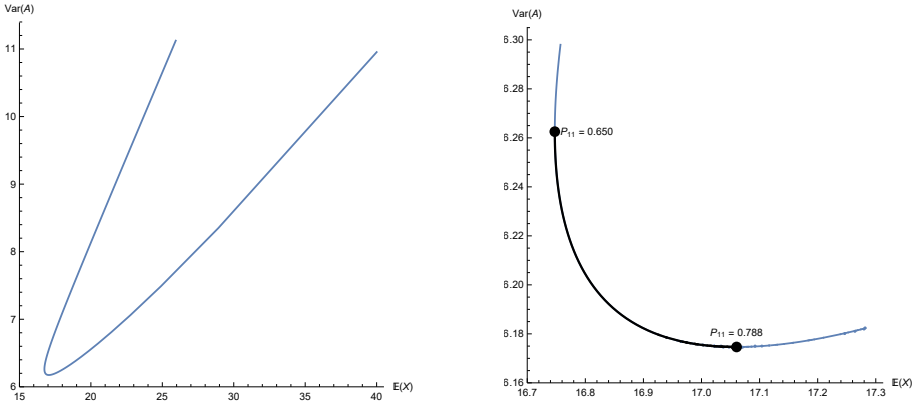
Once again, we assume that the conditional service times are exponentially distributed, but in this example we choose less symmetric settings. Let $\mathbb{P}(J = 1) = \frac{7}{16}, \mathbb{P}(J = 2) = \frac{9}{16}, \alpha_{11} = \alpha_{12} = \alpha_{21} = \frac{3}{20}$ and $\alpha_{22} = \frac{19}{20}$. From these settings we obtain $P_{21} = \frac{7}{9}P_{12}$, $\alpha_1 = 0.3$, and $\alpha_2 = 1.1$. The interesting phenomenon observed in Example 2, is also taking place here. In fact, in this example there is a bigger difference between the value of $P_{11}$ for which the mean queue length is minimal ($P_{11} \approx 0.65$), and the value resulting in a minimum variance of the number of arrivals during an arbitrary service time ($P_{11} \approx 0.788$). More details can be found in Table 4.2. The interesting region is obviously $0.650 < P_{11} < 0.788$, because in this region we know that an increase in $\mathrm{Var}(A)$ results in a decrease in $\mathbb{E}[X]$. This is illustrated even better in Figure 4.3, where $\mathrm{Var}(A)$ and $\mathbb{E}[X]$ are plotted against each other, for varying values of $P_{11}$.

| $P_{11}$ | $\mathbb{E}[X]$ | $\mathrm{Var}(A)$ |
|---|---|---|
| 0.100 | 20.377 | 8.327 |
| 0.300 | 17.931 | 7.056 |
| 0.500 | 16.969 | 6.493 |
| **0.650** | **16.747** | 6.263 |
| 0.700 | 16.780 | 6.214 |
| **0.788** | 17.060 | **6.175** |
| 0.900 | 18.587 | 6.333 |

**Table 4.2.** Mean queue length and variance of the number of arrivals during an arbitrary service time, for various values of $P_{11}$ in Example 3.

### 4.6.4 Example 4: Transient-state analysis

We return to the system in Example 2, but now we study the transient analysis. In this example we start with an empty system, $\mathbb{E}[z^{X_0}] = 1$, and set $P_{11} = 1/10$. Next, we repeatedly apply Equation (4.12) to express $\mathbb{E}[z^{X_n}]$ in terms of $\mathbb{E}[z^{X_{n-1}}]$. We have taken four different distributions for the conditional service times, namely exponential, gamma with shape parameter $1/2$, gamma




(a) $0.01 < P_{11} < 0.99$ (b) $0.62 < P_{11} < 0.82$

**Figure 4.3.** The variance of the number of arrivals versus the expected number of customers during an arbitrary customer service time. This implicit plot is obtained by varying $P_{11}$. Figure (b) is a zoomed in version of Figure (a).

with shape parameter 5, and deterministic. The results are shown in Figure 4.4, where we depict the mean queue length after the departure of the $n$-th customer, for $n = 0, 1, 2, \ldots, 200$. In this example, it can clearly be seen that service-time distributions with higher coefficients of variation result in longer queues. Also, it seems to take longer to reach steady state. For completeness, we give the steady-state mean queue lengths for the four systems below:

| Distribution | Deterministic | Gamma 5 | Exponential | Gamma 1/2 |
|---|---|---|---|---|
| $\mathbb{E}[X]$ | 16.224 | 16.918 | 19.696 | 23.168 |

## 4.7 Discussion and conclusion

In this chapter, we have studied a generalized $M^X$/semi-Markov/1 queuing model. An essential feature of our model is that the type of customer $n+1$ not only depends on the type of customer $n$, but also on the length of the service of customer $n$. We have determined the transient and stationary probability generating functions of the number of customers in the system immediately after a departure. We used that result to derive the generating functions of the stationary number of customers at an arbitrary epoch, at batch arrival epochs

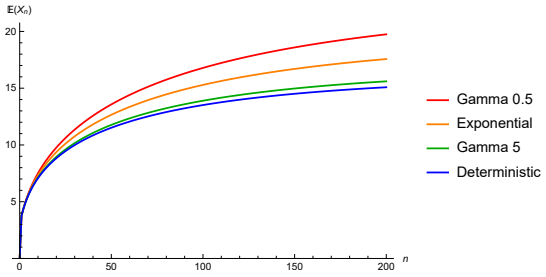



**Figure 4.4.** Numerical example 4: Transient mean queue-length analysis.

and at customer arrivals. Using these results, we explored the impact of the dependencies on mean number of customers. In particular, it was shown that the mean number of customers may become very large, even when the system is not in heavy traffic, due to a very high variance of the number of arrivals during a service time.

Chapter 5

# Extension with exceptional first service

In this chapter, we extend the queueing model, $M^X$/semi-Markov/1, studied in Chapter 4, with exceptional first service, i.e., the first customer in a busy period has a different service-time distribution than regular customers served in the busy period. Based on the results from the previous chapter on the steady-state distribution of the queue length, we derive the waiting time and sojourn time distributions of an arbitrary customer, showing that these depend on the position of the customer inside the batch, as well as on the type of the first customer in the batch.

We apply this extension of the queueing model to road traffic situations involving multiple conflicting traffic streams. In particular, we use it in the context of gap acceptance models for unsignalized intersections, studied in Chapter 3, where high-priority vehicles arrive according to an MMPP, and low-priority vehicles arrive in batches according to a Poisson process.

## 5.1 Introduction

The current model is a slight extension of that in Chapter 4. An overview of the earlier existing literature can be found in Chapter 4, in which we investigated the transient and stationary queue length distributions in a single server model with batch arrivals and semi-Markov service times [1]. In this chapter, we allow the service duration of a customer arriving into an empty system to have a distribution that differs from the service-time distributions of other customers. For the stationary analysis of the model, this requires minor adaptations of that in Chapter 4. In addition, we investigate the stationary waiting time and sojourn time distributions of an arbitrary customer, and apply this extension to road traffic situations involving multiple conflicting traffic streams.

The remainder of this chapter is organized as follows. In Section 5.2, we present the description of the queueing model. In Section 5.3, we first determine the stationary probability generating function of the queue length of the system at departure times. Subsequently, we use that result to derive the stationary generating functions of the queue length at an arbitrary time, at batch arrival instants, and at customer arrival instants. Using these results, we obtain the LST (Laplace-Stieltjes transform) of the steady-state waiting time and sojourn time distributions of customers as well as batches in Section 5.4. In Section 5.5, we first give several applications in which the extended queueing model arises, and then study the application to road traffic situations involving multiple conflicting traffic streams. In Section 5.6, we present numerical examples to demonstrate the impact of the three types of the driver's behavior ($B_1$, $B_2$, and $B_3$), described in Section 3.3, on the delay on the minor road. We present our conclusions in Section 5.7.

## 5.2 Model description

We consider a single-server queuing system. Customers arrive in batches at the system according to a Poisson process with intensity parameter $\lambda$. The arriving batch size is denoted by the random variable $B$, with generating function $B(z)$, for $|z| \leq 1$ (we do not allow zero-size batches without loss of generality, i.e. $B \geq 1$). Customers are served individually, and the first

customer in a busy period has a different service-time distribution than regular customers served in the busy period. There are $N$ types of customers, which we number $1, 2, \ldots, N$. Denote by $J_n$ the type of the $n$-th customer and $G^{(n)}$ its service time, $n = 1, 2, \ldots$. The type of a customer is only determined at the moment its service begins. More specifically, the type of the $n$-th customer depends on the type, and on the service duration of the $(n-1)$-th customer, as well as on whether the queue is empty at the departure time of the $(n-1)$th customer. We introduce, for $i = 1, 2, \ldots, N$,

$$\tilde{G}_{ij}(s) = \mathbb{E}[e^{-sG^{(n)}} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1], \tag{5.1}$$

$$\tilde{G}^*_{ij}(s) = \mathbb{E}[e^{-sG^{(n)}} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} = 0], \tag{5.2}$$

where $X_{n-1}$ is the number of customers in the system at the departure of the $(n-1)$-th customer.

In particular, for $i, j = 1, 2, \ldots, N$, we define

$$P_{ij} = \tilde{G}_{ij}(0) = \mathbb{P}(J_{n+1} = j | J_n = i, X_{n-1} \geq 1), \tag{5.3}$$

$$P^*_{ij} = \tilde{G}^*_{ij}(0) = \mathbb{P}(J_{n+1} = j | J_n = i, X_{n-1} = 0). \tag{5.4}$$

We assume that $P = [P_{ij}]_{i,j\in\{1,2,\ldots,N\}}$ is the transition probability matrix of an irreducible discrete time Markov chain, with stationary distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ such that

$$\pi P = \pi. \tag{5.5}$$

For intuition we may think of $\pi$ as the conditional equilibrium distribution of $J_n$ in case the queue would never empty.

Using Cramer's rule with the normalizing equation $\sum_{i=1}^N \pi_i = 1$, the solution of the system of equations (5.5) is given by

$$\pi_i = \frac{d_i}{d}, \qquad i = 1, 2, \ldots, N, \tag{5.6}$$

where $d = \sum_{i=1}^N d_i$, and $d_i$ is the cofactor of the entry in the $i$-th row and the

first column of the matrix $(I - P)$, which is given by

$$d_1 = \begin{vmatrix} 1-P_{22} & -P_{23} & \dots & -P_{2N} \\ -P_{32} & 1-P_{33} & \dots & -P_{3N} \\ \vdots & \vdots & \ddots & \vdots \\ -P_{N2} & -P_{N3} & \dots & 1-P_{NN} \end{vmatrix}, \tag{5.7}$$

$$d_i = (-1)^{i+1} \begin{vmatrix} -P_{12} & -P_{13} & \dots & -P_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ -P_{i-12} & -P_{i-13} & \dots & -P_{i-1N} \\ -P_{i+12} & -P_{i+13} & \dots & -P_{i+1N} \\ \vdots & \vdots & \ddots & \vdots \\ -P_{N2} & -P_{N3} & \dots & 1-P_{NN} \end{vmatrix}, \quad i = 2, 3, \dots, N-1, \tag{5.8}$$

$$d_N = (-1)^{N+1} \begin{vmatrix} -P_{12} & -P_{13} & \dots & -P_{1N} \\ 1-P_{22} & -P_{23} & \dots & -P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -P_{N-12} & -P_{N-13} & \dots & -P_{N-1N} \end{vmatrix}. \tag{5.9}$$

In the next section, to study the queue length distribution at departure times of customers, we denote by $A_n$ the number of arrivals during the service time of the $n$-th customer (including the customers coming in the batches). We introduce, for $i = 1, 2, \dots, N$,

$$A_i(z) = \sum_{j=1}^{N} A_{ij}(z), \tag{5.10}$$

$$A_i^*(z) = \sum_{j=1}^{N} A_{ij}^*(z), \tag{5.11}$$

with

$$A_{ij}(z) = \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1], \tag{5.12}$$

$$A_{ij}^*(z) = \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} = 0]. \tag{5.13}$$

Let us define

$$\rho = \sum_{i=1}^{N} \pi_i \alpha_i, \tag{5.14}$$

where

$$\alpha_i = \sum_{j=1}^{N} \alpha_{ij}, \tag{5.15}$$

with

$$\alpha_{ij} = \mathbb{E}[A_n 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1]. \tag{5.16}$$

Intuitively, we can think of $\rho$ as being the expected number of arrivals during a service time if the process $(J_n, X_{n-1})$ would never hit the level $X_{n-1} = 0$.

Introduce some further notations:

$$\alpha_i^* = \sum_{j=1}^{N} \alpha_{ij}^*, \tag{5.17}$$

with

$$\alpha_{ij}^* = \mathbb{E}[A_n^* 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} = 0]. \tag{5.18}$$

Note that the number of arrivals during the service time of a customer is a batch Poisson process. Therefore, we can write the following relations:

$$A_{ij}(z) = \tilde{G}_{ij}(\lambda(1 - B(z))), \tag{5.19}$$

$$A_{ij}^*(z) = \tilde{G}_{ij}^*(\lambda(1 - B(z))), \quad \text{for } i, j = 1, 2, \ldots, N. \tag{5.20}$$

To derive the stability condition for our model, we use the results from Chapter 4. Note that the dynamics in the current model only differs from that in Chapter 4 when the queue length is zero. More specifically, the two processes have identical transition rates, except in a finite number of states. This implies that the two processes are either both positive recurrent, both null recurrent or both transient. The condition for stability reads $\rho < 1$, in accordance with Chapter 4, and similarly, both processes are null recurrent if $\rho = 1$. Hence, if we modify the parameters such that $\rho \uparrow 1$, the processes move from positive recurrence to null recurrence. In particular, $\mathbb{P}[X = 0] > 0$ if $\rho < 1$ and $\mathbb{P}[X = 0] \to 0$ as $\rho \uparrow 1$.

**Remark 5.1.** *It is worth to emphasize that the definition of $\tilde{G}_{ij}(s)$ in this chapter (and the subsequent chapters) is not the same as in Chapter 4. In this chapter, $\tilde{G}_{ij}(s)$ equals $\mathbb{E}[e^{-sG^{(n)}}1_{\{J_{n+1}=j\}}|J_n = i, X_{n-1} \geq 1]$, whereas this equals $\mathbb{E}[e^{-sG^{(n)}}1_{\{J_{n+1}=j\}}|J_n = i]$ in Chapter 4. The same also applies to $A_{ij}(z)$.*

## 5.3 Stationary queue length analysis

In this section, we shall first determine the steady-state joint distribution of the number of customers in the system immediately after a departure, and the type of the next customer to be served. Subsequently, we will use this result to derive the generating functions of the stationary number of customers at an arbitrary time, at batch arrival instants, and at customer arrival instants.

### 5.3.1 Stationary queue length analysis: departure epochs

Starting-point of the analysis is the following recurrence relation:

$$X_n = \begin{cases} X_{n-1} - 1 + A_n & \text{if } X_{n-1} \geq 1 \\ A_n + B_n - 1 & \text{if } X_{n-1} = 0 \end{cases}, \quad n = 1, 2, 3, \ldots, \tag{5.21}$$

where $X_n$ is the number of customers at the departure times of the $n$-th customer and $B_n$ is the size of the batch in which the $n$-th customer arrived, with generating function $B(z)$, for $|z| \leq 1$. Due to dependent successive service times, $\{X_n\}_{n\geq 0}$ here is not a Markov chain. In order to obtain a Markovian model, it is required to keep track of the type of the departing customer together with the number of customers in the system immediately after the departure of that customer. As a consequence, $(X_n, J_{n+1})$ forms a Markov chain.

Taking generating functions and exploiting the fact that $X_{n-1}$ and $(A_n, J_{n+1})$ are conditionally independent, given $J_n$ and $X_{n-1} \geq 1$, we find:

$$\begin{aligned} &\mathbb{E}[z^{X_n}1_{\{J_{n+1}=j\}}] \\ &= \sum_{i=1}^{N} \mathbb{E}[z^{X_{n-1}-1}|J_n = i, X_{n-1} \geq 1]\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}|J_n = i, X_{n-1} \geq 1] \end{aligned}$$

$$\times \mathbb{P}(J_n = i, X_{n-1} \geq 1) + \frac{B(z)}{z} \sum_{i=1}^{N} \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} = 0]$$
$$\times \mathbb{P}(J_n = i, X_{n-1} = 0)$$
$$= \frac{1}{z} \sum_{i=1}^{N} \mathbb{E}[z^{X_{n-1}} 1_{\{J_n=i\}}] \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1]$$
$$+ \frac{1}{z} \sum_{i=1}^{N} \Big( B(z) \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} = 0]$$
$$- \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1] \Big) \mathbb{P}(J_n = i, X_{n-1} = 0),$$
$$\text{for } n = 1, 2, 3, \ldots, \quad j = 1, 2, \ldots, N. \tag{5.22}$$

Now, we restrict ourselves to the stationary situation, assuming that the stability condition holds.
Introduce, for $i, j = 1, 2, \ldots, N$ and $|z| \leq 1$:

$$f_i(z) = \lim_{n\to\infty} \mathbb{E}[z^{X_n} 1_{\{J_{n+1}=i\}}], \tag{5.23}$$

with, for $i = 1, 2, \ldots, N$,

$$f_i(0) = \lim_{n\to\infty} \mathbb{P}(X_n = 0, J_{n+1} = i) \tag{5.24}$$

such that

$$F(z) = \sum_{i=1}^{N} f_i(z). \tag{5.25}$$

In stationarity, Equation (5.22) leads to the following $N$ equations:

$$(z - A_{jj}(z)) f_j(z) - \sum_{i=1, i\neq j}^{N} A_{ij}(z) f_i(z) = \sum_{i=1}^{N} (B(z) A^*_{ij}(z) - A_{ij}(z)) f_i(0),$$
$$\text{for } j = 1, 2, \ldots, N. \tag{5.26}$$

We can also write these $N$ linear equations in matrix form as

$$M(z)^T f(z) = b(z),$$

where

$$M(z) = \begin{bmatrix} z - A_{11}(z) & -A_{12}(z) & \dots & -A_{1N}(z) \\ -A_{21}(z) & z - A_{22}(z) & \dots & -A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ -A_{N1}(z) & -A_{N2}(z) & \dots & z - A_{NN}(z) \end{bmatrix}, \tag{5.27}$$

$$f(z) = \begin{bmatrix} f_1(z) \\ f_2(z) \\ \vdots \\ f_N(z) \end{bmatrix}, b(z) = \begin{bmatrix} b_1(z) \\ b_2(z) \\ \vdots \\ b_N(z) \end{bmatrix},$$

$$\text{with } b_j(z) = \sum_{i=1}^{N} (B(z)A_{ij}^*(z) - A_{ij}(z)) f_i(0). \tag{5.28}$$

Therefore, by Cramer's rule, solutions of the non-homogeneous linear system $M(z)^T f(z) = b(z)$ are in the form: for $i = 1, 2, \dots, N$,

$$f_i(z) = \frac{\det L_i(z)}{\det M(z)^T}, \quad \det M(z)^T \neq 0, \tag{5.29}$$

where $L_i(z)$ is the matrix formed by replacing the $i$-th column of $M(z)^T$ by the column vector $b(z)$.

It remains to find the values of $f_1(0), f_2(0), \dots, f_N(0)$. We shall derive $N$ linear equations for $f_1(0), f_2(0), \dots, f_N(0)$.

**First equation:**

Note that $\det M(z)^T = \det M(z)$. After replacing the first column by the sum of all $N$ columns in (5.27), and using (5.10), we get

$$\det M(z)^T = \begin{vmatrix} z - A_1(z) & -A_{12}(z) & \dots & -A_{1N}(z) \\ z - A_2(z) & z - A_{22}(z) & \dots & -A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z - A_N(z) & -A_{N2}(z) & \dots & z - A_{NN}(z) \end{vmatrix}. \tag{5.30}$$

This implies that

$$\det M(z)^T = \sum_{i=1}^{N} (z - A_i(z)) u_{i1}(z), \tag{5.31}$$

where $u_{i1}(z)$ is the cofactor of the entry in the $i$-th row and the first column of the matrix

$$\begin{bmatrix} z - A_1(z) & -A_{12}(z) & \dots & -A_{1N}(z) \\ z - A_2(z) & z - A_{22}(z) & \dots & -A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z - A_N(z) & -A_{N2}(z) & \dots & z - A_{NN}(z) \end{bmatrix}.$$

Note that $\{z - A_i(z)\}|_{z=1} = 0$, $\frac{d}{dz}\{z - A_i(z)\}|_{z=1} = 1 - \alpha_i$, and $u_{i1}(1) = d_i$, where $d_i$ are given by Equations (5.7),(5.8),(5.9), and $\alpha_i$ are defined in (5.15), for $i = 1, 2, \dots, N$. Therefore, we obtain

$$\begin{aligned} \frac{d}{dz}\{\det M(z)^T\}|_{z=1} &= \sum_{i=1}^{N}(1 - \alpha_i)d_i \\ &= d - \sum_{i=1}^{N} \alpha_i d_i \\ &= d(1 - \sum_{i=1}^{N} \alpha_i \pi_i) \\ &= d(1 - \rho), \end{aligned} \tag{5.32}$$

where $d = \sum_{i=1}^{N} d_i$.

In the determinant form, $\det L_i(z)$ is given by

$$\det L_1(z) = \begin{vmatrix} b_1(z) & -A_{21}(z) & \dots & -A_{N1}(z) \\ b_2(z) & z - A_{22}(z) & \dots & -A_{N2}(z) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(z) & -A_{2N}(z) & \dots & z - A_{NN}(z) \end{vmatrix}, \tag{5.33}$$

$$\begin{aligned} &\det L_i(z) \\ &= \begin{vmatrix} z - A_{11}(z) & \dots & -A_{i-11}(z) & b_1(z) & -A_{i+11}(z) & \dots & -A_{N1}(z) \\ -A_{12}(z) & \dots & -A_{i-12}(z) & b_2(z) & -A_{i+12}(z) & \dots & -A_{N2}(z) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -A_{1N}(z) & \dots & -A_{i-1N}(z) & b_N(z) & -A_{i+1N}(z) & \dots & z - A_{NN}(z) \end{vmatrix}, \\ &\qquad \text{for } i = 2, 3, \dots, N. \end{aligned} \tag{5.34}$$

This implies that

$$\det L_i(z) = \sum_{j=1}^{N} b_j(z) r_{ji}(z), \quad i = 1, 2, \ldots, N, \tag{5.35}$$

where $b_j(z)$ is given by (5.28), and $r_{ji}(z)$ is the cofactor of the entry in the $j$-th row and $i$-th column of the matrix $L_i(z)$, which is given by

$$r_{11}(z) = \begin{vmatrix} z - A_{22}(z) & -A_{32}(z) & \ldots & -A_{N2}(z) \\ -A_{23}(z) & z - A_{33}(z) & \ldots & -A_{N3}(z) \\ \vdots & \vdots & \ddots & \vdots \\ -A_{2N}(z) & -A_{3N}(z) & \ldots & z - A_{NN}(z) \end{vmatrix}, \tag{5.36}$$

$$r_{1i}(z) = (-1)^{i+1} \times \begin{vmatrix} -A_{12}(z) & \ldots & -A_{i-12}(z) & -A_{i+12}(z) & \ldots & -A_{N2}(z) \\ -A_{13}(z) & \ldots & -A_{i-13}(z) & -A_{i+13}(z) & \ldots & -A_{N3}(z) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -A_{1N}(z) & \ldots & -A_{i-1N}(z) & -A_{i+1N}(z) & \ldots & z - A_{NN}(z) \end{vmatrix}, \quad \text{for } i = 2, 3, \ldots, N, \tag{5.37}$$

$$r_{j1}(z) = (-1)^{j+1} \times \begin{vmatrix} -A_{21}(z) & -A_{31}(z) & -\ldots & -A_{N1}(z) \\ \vdots & \vdots & \ddots & \vdots \\ -A_{2j-1}(z) & -A_{3j-1}(z) & \ldots & -A_{Nj-1}(z) \\ -A_{2j+1}(z) & -A_{3j+1}(z) & \ldots & -A_{Nj+1}(z) \\ \vdots & \vdots & \ddots & \vdots \\ -A_{2N}(z) & -A_{3N}(z) & \ldots & z - A_{NN}(z) \end{vmatrix}, \quad \text{for } j = 2, 3, \ldots, N, \tag{5.38}$$

$$r_{ji}(z) = (-1)^{i+j} \times \begin{vmatrix} z - A_{11}(z) & \dots & -A_{i-11}(z) & -A_{i+11}(z) & \dots & -A_{N1}(z) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -A_{1j-1}(z) & \dots & -A_{i-1j-1}(z) & -A_{i+1j-1}(z) & \dots & -A_{Nj-1}(z) \\ -A_{1j+1}(z) & \dots & -A_{i-1j+1}(z) & -A_{i+1j+1}(z) & \dots & -A_{Nj+1}(z) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -A_{1N}(z) & \dots & -A_{i-1N}(z) & -A_{i+1N}(z) & \dots & z - A_{NN}(z) \end{vmatrix},$$

$$\text{for } i, j = 2, 3, \dots, N. \tag{5.39}$$

Subsequently,

$$\begin{aligned} \frac{d}{dz}\{\det L_i(z)\}|_{z=1} &= \sum_{j=1}^{N} (b_j(1) r'_{ji}(1) + b'_j(1) r_{ji}(1)) \\ &= \sum_{j=1}^{N} \sum_{k=1}^{N} \Big( r'_{ji}(1)(P^*_{kj} - P_{kj}) + r_{ji}(1)(\mathbb{E}[B] P^*_{kj} + \alpha^*_{kj} - \alpha_{kj}) \Big) f_k(0). \end{aligned} \tag{5.40}$$

After replacing the first row by the sum of all $N$ rows of $\det L_i(z)$ in (5.34), we obtain $\det L_i(z)$, $i = 2, 3, \dots, N$, as

$$\det L_i(z) = \begin{vmatrix} z - A_1(z) & \dots & z - A_{i-1}(z) & \sum_{j=1}^{N} b_j(z) & z - A_{i+1}(z) & \dots & z - A_N(z) \\ -A_{12}(z) & \dots & -A_{i-12}(z) & b_2(z) & -A_{i+12}(z) & \dots & -A_{N2}(z) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -A_{1N}(z) & \dots & -A_{i-1N}(z) & b_N(z) & -A_{i+1N}(z) & \dots & z - A_{NN}(z) \end{vmatrix}. \tag{5.41}$$

In particular, for $i = 2, 3, \dots, N$,

$$\det L_i(1) = \begin{vmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ -P_{12} & \dots & -P_{i-12} & b_2(1) & -P_{i+12} & \dots & -P_{N2} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -P_{1N} & \dots & -P_{i-1N} & b_N(1) & -P_{i+1N} & \dots & 1 - P_{NN} \end{vmatrix} = 0.$$

Similarly, $\det L_1(1) = 0$ and $\det M(1)^T = 0$.

Therefore, for $i = 1, 2, \ldots, N$, we obtain

$$\begin{aligned} f_i(1) &= \lim_{z\uparrow 1} \frac{\det L_i(z)}{\det M(z)^T} \\ &= \frac{\frac{d}{dz}\{\det L_i(z)\}|_{z=1}}{\frac{d}{dz}\{\det M(z)^T\}|_{z=1}}. \end{aligned} \tag{5.42}$$

Note that $F(1) = 1$, which implies that $\sum_{i=1}^N f_i(1) = 1$. And, as a consequence, we obtain

$$\frac{\sum_{i=1}^N \frac{d}{dz}\{\det L_i(z)\}|_{z=1}}{\frac{d}{dz}\{\det M(z)^T\}|_{z=1}} = 1.$$

This implies that

$$\begin{aligned} \sum_{k=1}^N \left( \sum_{i=1}^N \sum_{j=1}^N \Big( r'_{ji}(1)(P^*_{kj} - P_{kj}) + r_{ji}(1)(\mathbb{E}[B]P^*_{kj} + \alpha^*_{kj} - \alpha_{kj}) \Big) \right) f_k(0) \\ = d(1-\rho). \end{aligned} \tag{5.43}$$

$(N-1)$ **equations:**
Note that under the stability condition, $\det M(z)^T$ has exactly $N-1$ zeros: $\hat{z}_l$, $l = 1, 2, \ldots, N-1$, in $|z| < 1$ (see in Chapter 4), and $F(z)$ is an analytical function in $|z| < 1$. Therefore, the numerator of $F(z)$ also has $(N-1)$ zeros in $|z| < 1$. And, as a consequence, these $(N-1)$ zeros provide $(N-1)$ linear equations for $f_1(0), f_2(0), \ldots, f_N(0)$:

$$\begin{aligned} \sum_{i=1}^N \det L_i(\hat{z}_l) = 0, \\ \Longrightarrow \sum_{i=1}^N \sum_{j=1}^N b_j(\hat{z}_l) r_{ji}(\hat{z}_l) = 0, \\ \Longrightarrow \sum_{k=1}^N \left( \sum_{i=1}^N \sum_{j=1}^N r_{ji}(\hat{z}_l) \Big( B(\hat{z}_l) A^*_{kj}(\hat{z}_l) - A_{kj}(\hat{z}_l) \Big) \right) f_k(0) = 0, \\ \text{for } |\hat{z}_l| < 1 \text{ and } l = 1, 2, \ldots, N-1. \end{aligned} \tag{5.44}$$

### 5.3.2 Special cases

We are not able to find explicit and compact expressions for general $N$. However, for $N = 2$, we can explicitly determine the probability generating function of the number of customers as follows.
Steady state behaviour:

$$\begin{aligned} f_1(z) =& \frac{(z - A_{22}(z)) \sum_{i=1}^{2} (B(z)A_{i1}^*(z) - A_{i1}(z)) f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)} \\ &+ \frac{A_{21}(z) \sum_{i=1}^{2} (B(z)A_{i2}^*(z) - A_{i2}(z)) f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)}, \end{aligned} \tag{5.45}$$

$$\begin{aligned} f_2(z) =& \frac{(z - A_{11}(z)) \sum_{i=1}^{2} (B(z)A_{i2}^*(z) - A_{i2}(z)) f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)} \\ &+ \frac{A_{12}(z) \sum_{i=1}^{2} (B(z)A_{i1}^*(z) - A_{i1}(z)) f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)}. \end{aligned} \tag{5.46}$$

In particular,

$$\begin{aligned} f_1(1) =& \lim_{z \uparrow 1} \frac{(z - A_{22}(z)) \sum_{i=1}^{2} (B(z)A_{i1}^*(z) - A_{i1}(z)) f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)} \\ &+ \lim_{z \uparrow 1} \frac{A_{21}(z) \sum_{i=1}^{2} (B(z)A_{i2}^*(z) - A_{i2}(z)) f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)} \\ =& \frac{\sum_{i=1}^{2} \Big( P_{21}(P_{i1}^* \mathbb{E}[B] + \alpha_{i1}^* - \alpha_{i1}) + (1 - \alpha_{22})(P_{i1}^* - P_{i1}) \Big) f_i(0)}{(1 - P_{11})(1 - \alpha_{22}) + (1 - P_{22})(1 - \alpha_{11}) - P_{12}\alpha_{21} - P_{21}\alpha_{12}} \\ &+ \frac{\sum_{i=1}^{2} \Big( P_{21}(P_{i2}^* \mathbb{E}[B] + \alpha_{i2}^* - \alpha_{i2}) + \alpha_{21}(P_{i2}^* - P_{i2}) \Big) f_i(0)}{(1 - P_{11})(1 - \alpha_{22}) + (1 - P_{22})(1 - \alpha_{11}) - P_{12}\alpha_{21} - P_{21}\alpha_{12}} \\ =& \frac{\sum_{i=1}^{2} \Big( P_{21}(\mathbb{E}[B] + \alpha_i^* - \alpha_i) + (1 - \alpha_2)(P_{i1}^* - P_{i1}) \Big) f_i(0)}{(P_{12} + P_{21}) \left( 1 - \frac{P_{21}}{P_{12}+P_{21}} \alpha_1 - \frac{P_{12}}{P_{12}+P_{21}} \alpha_2 \right)} \\ =& \frac{\sum_{i=1}^{2} \Big( P_{21}(\mathbb{E}[B] + \alpha_i^* - \alpha_i) + (1 - \alpha_2)(P_{i1}^* - P_{i1}) \Big) f_i(0)}{(P_{12} + P_{21})(1 - \rho)}. \end{aligned} \tag{5.47}$$

Similarly,

$$f_2(1) = \frac{\sum_{i=1}^{2} \Big( P_{12}(\mathbb{E}[B] + \alpha_i^* - \alpha_i) + (1 - \alpha_1)(P_{i2}^* - P_{i2}) \Big) f_i(0)}{(P_{12} + P_{21})(1 - \rho)}. \tag{5.48}$$

As a consequence of $f_1(1) + f_2(1) = 1$, we obtain

$$\sum_{i=1}^{2} \Big( (P_{12} + P_{21})(\mathbb{E}[B] + \alpha_i^* - \alpha_i) + (\alpha_1 - \alpha_2)(P_{i1}^* - P_{i1}) \Big) f_i(0) = (P_{12} + P_{21})\,(1 - \rho)\,. \tag{5.49}$$

We can write $F(z)$ as the sum of $f_1(z)$ and $f_2(z)$, i.e.,

$$F(z) = f_1(z) + f_2(z).$$

After substituting the values of $f_1(z)$ and $f_2(z)$ from Equations (5.45) and (5.46) respectively, we obtain $F(z)$ as

$$F(z) = \frac{\sum_{i=1}^{2}(z + A_{12}(z) - A_{22}(z))(B(z)A_{i1}^*(z) - A_{i1}(z))f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)} + \frac{\sum_{i=1}^{2}(z + A_{21}(z) - A_{11}(z))(B(z)A_{i2}^*(z) - A_{i2}(z))f_i(0)}{(z - A_{11}(z))(z - A_{22}(z)) - A_{12}(z)A_{21}(z)}. \tag{5.50}$$

Let $z = \hat{z}$ be the zero of the denominator of $F(z)$ such that $|\hat{z}| < 1$. Therefore, $z = \hat{z}$ is also zero of the numerator of $F(z)$. As a consequence, we obtain the following equation in terms of $f_1(0)$ and $f_2(0)$:

$$\sum_{i=1}^{2} \Big( (\hat{z} + A_{12}(\hat{z}) - A_{22}(\hat{z}))(B(\hat{z})A_{i1}^*(\hat{z}) - A_{i1}(\hat{z})) + (\hat{z} + A_{21}(\hat{z}) - A_{11}(\hat{z}))(B(\hat{z})A_{i2}^*(\hat{z}) - A_{i2}(\hat{z})) \Big) f_i(0) = 0. \tag{5.51}$$

After solving Equations (5.49) and (5.51), we obtain

$$f_1(0) = \frac{-(P_{12} + P_{21})\,(1 - \rho)\,R_{12}}{\det R}, \tag{5.52}$$

$$f_2(0) = \frac{(P_{12} + P_{21})\,(1 - \rho)\,R_{11}}{\det R}, \tag{5.53}$$

where $\det R$ is the determinant of the matrix $R = [R_{ij}]$, whose elements are given by

$$\begin{aligned} R_{1j} =& (\hat{z} + A_{12}(\hat{z}) - A_{22}(\hat{z}))(B(\hat{z})A_{j1}^*(\hat{z}) - A_{j1}(\hat{z})) \\ &+ (\hat{z} + A_{21}(\hat{z}) - A_{11}(\hat{z}))(B(\hat{z})A_{j2}^*(\hat{z}) - A_{j2}(\hat{z})), \\ R_{2j} =& (P_{12} + P_{21})(\mathbb{E}[B] + \alpha_j^* - \alpha_j) + (\alpha_1 - \alpha_2)(P_{j1}^* - P_{j1}), \quad j = 1, 2. \end{aligned}$$

### 5.3.3 Stationary queue length analysis: arrival and arbitrary epochs

In the previous subsection, we determined the probability generating function of the stationary queue length distribution at the departure epoch of an arbitrary customer for general batch arrivals. As customers arrive at the system according to a batch Poisson process with rate $\lambda$, from the PASTA property, the distribution of the number of customers in the system at the arrival time of a batch is the same as the distribution of the number of customers at an arbitrary time. After using PASTA and level-crossing arguments (see Section 4.4 for more details), we obtain the following relations:

$$\mathbb{E}[z^X] = \mathbb{E}[z^{X^{ca}}] = \mathbb{E}[z^{X^{ba}}]\frac{1-B(z)}{\mathbb{E}[B](1-z)}, \tag{5.54}$$

with,

$$\mathbb{E}[z^{X^{arb}}] = \mathbb{E}[z^{X^{ba}}]. \tag{5.55}$$

where $X$ and $X^{ca}$ are the number of customers at the departure and the arrival epoch of the customer respectively; $X^{arb}$ and $X^{ba}$ are the number of customers at the arbitrary time and the arrival time of a batch respectively.

Hence, from these relations, we can obtain all the required distributions.

## 5.4 Waiting time and sojourn time

In this section, we shall determine the waiting time and sojourn time distributions of an arbitrary batch as well as an arbitrary customer, noticing that the waiting time and sojourn time of a customer depend on its position in the batch, as well as on the type of service of the first customer in its batch.

Let $G^{(n)}$ be the service time of the $n$-th customer such that

$$\tilde{G}_{ij}(s) = \mathbb{E}[e^{-sG^{(n)}}1_{\{J_{n+1}=j\}}|J_n = i, X_{n-1} \geq 1], \tag{5.56}$$

$$\tilde{G}^*_{ij}(s) = \mathbb{E}[e^{-sG^{(n)}}1_{\{J_{n+1}=j\}}|J_n = i, X_{n-1} = 0], \tag{5.57}$$

where $X_{n-1}$ is the number of customers in the system at the departure of the $(n-1)$-th customer.

To determine the waiting times and sojourn times of customers, firstly, we modify our model in such a way that all customers in the same batch are served together as a *super customer*. Let $\mathcal{G}^{(n)}$ and $\mathcal{J}_n$ be the service time and the service type of the $n$-th super customer respectively. Then, the LST of the conditional service time of a super customer is defined as, for $\text{Re}(s) \geq 0, i, j = 1, 2, \ldots, N$,

$$\tilde{\mathcal{G}}_{ij}(s) = \mathbb{E}[e^{-s\mathcal{G}^{(n)}} 1_{\{\mathcal{J}_{n+1}=j\}} | \mathcal{J}_n = i, X_{n-1} \geq 1], \tag{5.58}$$

$$\tilde{\mathcal{G}}^*_{ij}(s) = \mathbb{E}[e^{-s\mathcal{G}^{(n)}} 1_{\{\mathcal{J}_{n+1}=j\}} | \mathcal{J}_n = i, X_{n-1} = 0]. \tag{5.59}$$

Now, we can obtain the LST of the conditional service time of a super customer in terms of the LST of the conditional service time of an individual customer as

$$\tilde{\mathcal{G}}_{ij}(s) = \mathbb{E}\left[[\tilde{\mathbf{G}}(s)^B]_{ij}\right], \tag{5.60}$$

$$\tilde{\mathcal{G}}^*_{ij}(s) = \sum_{k=1}^{N} \tilde{G}^*_{ik}(s) \mathbb{E}\left[[\tilde{\mathbf{G}}(s)^{(B-1)}]_{kj}\right], \quad i, j = 1, 2, \ldots, N, \tag{5.61}$$

where $\tilde{\mathbf{G}}(s) = [\tilde{G}_{ij}(s)]$ is a matrix of order $N \times N$, and $[\tilde{\mathbf{G}}(s)^B]_{ij}$ is the $(i, j)$th element of matrix $\tilde{\mathbf{G}}(s)^B$, for $i, j = 1, 2, \ldots, N$.

Let $\mathcal{X}^d_n, \mathcal{X}^{bs}_n$ be the number of super customers in the queue at the departure of, and the beginning of service of the $n$-th super customer respectively. We can derive the probability generating function of the number of super customers in the queue, in steady state, at the departure of a super customer by letting $A_{ij}(z) = \tilde{\mathcal{G}}_{ij}(\lambda(1-z))$, $A^*_{ij}(z) = \tilde{\mathcal{G}}^*_{ij}(\lambda(1-z))$ and $B(z) = z$ in Equation (5.29).

Therefore, now, we know the distribution of the number of super customers at the departure of the super customer. But, to determine the waiting time of a super customer, using the distributional form of Little's law (see Section 1.3), we need to find the distribution of the number of super customers at the beginning of the service of a super customer.

We can write

$$\mathcal{X}^{bs}_n = \begin{cases} \mathcal{X}^d_{n-1} - 1, & \text{if } \mathcal{X}^d_{n-1} \geq 1, \\ 0, & \text{if } \mathcal{X}^d_{n-1} = 0. \end{cases}$$

This implies that

$$\mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}}] = \mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}] + \mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d\geq 1\}}], \tag{5.62}$$

where

$$\mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}] = \mathbb{P}(\mathcal{X}_{n-1}^d = 0, \mathcal{J}_n = i), \tag{5.63}$$

and

$$\begin{aligned}
&\mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d\geq 1\}}] \\
&= \mathbb{E}[z^{\mathcal{X}_{n-1}^d - 1} 1_{\{\mathcal{J}_n=i\}}] - \mathbb{E}[z^{\mathcal{X}_{n-1}^d - 1} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}] \\
&= \frac{1}{z}\left(\mathbb{E}[z^{\mathcal{X}_{n-1}^d} 1_{\{\mathcal{J}_n=i\}}] - \mathbb{P}(\mathcal{X}_{n-1}^d = 0, \mathcal{J}_n = i)\right).
\end{aligned} \tag{5.64}$$

Let $\mathcal{W}_n^{sc}$ and $\mathcal{S}_n^{sc}$ be the waiting time and sojourn time of the $n$-th super customer respectively. By the distributional form of Little's law, we obtain

$$\begin{aligned}
\mathbb{E}[z^{\mathcal{X}_n^d} 1_{\{\mathcal{J}_{n+1}=i\}}] &= \mathbb{E}[e^{-\lambda(1-z)\mathcal{S}_n^{sc}} 1_{\{\mathcal{J}_{n+1}=i\}}], \\
\mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}] &= \mathbb{E}[e^{-\lambda(1-z)\mathcal{W}_n^{sc}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}], \\
\mathbb{E}[z^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d\geq 1\}}] &= \mathbb{E}[e^{-\lambda(1-z)\mathcal{W}_n^{sc}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d\geq 1\}}].
\end{aligned}$$

Letting $s = \lambda(1-z)$, then yields

$$\mathbb{E}[e^{-s\mathcal{S}_n^{sc}} 1_{\{\mathcal{J}_{n+1}=i\}}] = \mathbb{E}\left[\left(1-\frac{s}{\lambda}\right)^{\mathcal{X}_n^d} 1_{\{\mathcal{J}_{n+1}=i\}}\right], \tag{5.65}$$

$$\mathbb{E}[e^{-s\mathcal{W}_n^{sc}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}] = \mathbb{E}\left[\left(1-\frac{s}{\lambda}\right)^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d=0\}}\right], \tag{5.66}$$

$$\mathbb{E}[e^{-s\mathcal{W}_n^{sc}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d\geq 1\}}] = \mathbb{E}\left[\left(1-\frac{s}{\lambda}\right)^{\mathcal{X}_n^{bs}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^d\geq 1\}}\right]. \tag{5.67}$$

Subsequently, we obtain

$$\mathbb{E}[e^{-s\mathcal{S}^{sc}}] = \mathbb{E}\left[\left(1-\frac{s}{\lambda}\right)^{\mathcal{X}^d}\right], \tag{5.68}$$

$$\mathbb{E}[e^{-s\mathcal{W}^{sc}}] = \mathbb{E}\left[\left(1-\frac{s}{\lambda}\right)^{\mathcal{X}^{bs}}\right], \tag{5.69}$$

where $\mathcal{S}^{sc} = \lim_{n\to\infty} \mathcal{S}_n^{sc}, \mathcal{W}^{sc} = \lim_{n\to\infty} \mathcal{W}_n^{sc}, \mathcal{X}^{d} = \lim_{n\to\infty} \mathcal{X}_n^{d}, \mathcal{X}^{bs} = \lim_{n\to\infty} \mathcal{X}_n^{bs}$.

We conclude the following relations between the waiting times and sojourn times of customers in the batches, with the waiting times and sojourn times of super customers:

- the waiting time of the first customer in the batch is equal to the waiting time of the super customer,
- the sojourn time of the first customer in the batch is equal to the waiting time of the super customer, plus the service time of the first customer,
- the sojourn time of the last customer in the batch is equal to the sojourn time of the super customer.

Let $W^{(m)}$ and $S^{(m)}$ be the steady-state waiting time and sojourn time of the $m$-th customer served in his batch, respectively. Therefore, we obtain

$$\mathbb{E}[e^{-sW^{(1)}}] = \mathbb{E}[e^{-s\mathcal{W}^{sc}}], \tag{5.70}$$

$$\begin{aligned}\mathbb{E}[e^{-sW^{(m)}}] = &\sum_{k=1}^{N}\sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{E}[e^{-s\mathcal{W}_n^{sc}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^{d}=0\}}]\tilde{G}_{ik}^{*}(s)[\tilde{\mathbf{G}}(s)^{m-2}]_{kj} \\ &+ \sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{E}[e^{-s\mathcal{W}_n^{sc}} 1_{\{\mathcal{J}_n=i\}} 1_{\{\mathcal{X}_{n-1}^{d}\geq 1\}}][\tilde{\mathbf{G}}(s)^{m-1}]_{ij}, \quad m \geq 2,\end{aligned} \tag{5.71}$$

$$\mathbb{E}[e^{-sS^{(m)}}] = \mathbb{E}[e^{-sW^{(m+1)}}], \quad m \geq 1. \tag{5.72}$$

Now, we are interested in the probability of being the $m$-th customer served in a batch. For that, we define the probability of the arriving batch-size as $b_k = \mathbb{P}(B = k)$ for $k \geq 1$. Therefore, the probability that an arbitrary customer arrives in a batch of size $k$, is equal to $\frac{kb_k}{\mathbb{E}[B]}$ (see Burke [19]). And hence, the probability of being the $m$th customer served in a batch is given by

$$r_m = \sum_{k=m}^{\infty} \frac{kb_k}{\mathbb{E}[B]}\frac{1}{k} = \frac{1}{\mathbb{E}[B]}\sum_{k=m}^{\infty} b_k. \tag{5.73}$$

Hence, the steady-state waiting and sojourn time LST of an arbitrary customer are given by

$$\mathbb{E}[e^{-sW}] = \sum_{m=1}^{\infty} r_m \mathbb{E}[e^{-sW^{(m)}}], \tag{5.74}$$

$$\mathbb{E}[e^{-sS}] = \sum_{m=1}^{\infty} r_m \mathbb{E}[e^{-sS^{(m)}}]. \tag{5.75}$$

**Remark 5.2.** *In case that batches have a maximum size of, say, $M$, we can still use Equations* (5.71) *and* (5.72)*. However, we note that although we define $\mathbb{E}[e^{-sW^{(m)}}]$ for $m = 1, 2, ..., M + 1$, there is in fact no $(M + 1)$-th customer in the batch. Still, we need to define $\mathbb{E}[e^{-sW^{(M+1)}}]$ to determine $\mathbb{E}[e^{-sS^{(M)}}]$.*

## 5.5 Applications to road traffic

The queueing model considered in this chapter arises in several applications including logistics, production/inventory systems, computer and telecommunication networks. In this section, we focus on the application to road traffic situations involving multiple conflicting traffic streams. More specifically, we consider an unsignalized priority-controlled intersection, described in Section 1.2. The low-priority car drivers, on the minor road, cross the intersection as soon as they come across a gap with duration larger than $T$ between two subsequent high-priority cars, commonly referred to as the *critical headway*. On the major road, we consider Markov platooning, described in Chapter 3, which can be used to model the fluctuations in the traffic density with a dependency between successive gap sizes.

On the minor road, cars arrive in batches according to a Poisson process with rate $\lambda$, where the batch size is denoted by the random variable $B$ with generating function $B(z)$, for $|z| \leq 1$. The arrival process on the major road is an MMPP such that, for $i = 1, 2, \ldots, N$, $q_i$ is the Poisson rate when the continuous time Markov process, $J(t)$, is in phase $i$.

Note that the transition probabilities of the background process of the MMPP are given by

$$\mathbb{P}(J(T) = j | J(0) = i) = [e^{TQ}]_{ij}, \quad \text{for } i, j = 1, 2, \ldots, N,$$

with the transition rate matrix

$$Q = \begin{bmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1N} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2N} \\ \vdots & \vdots & \dots & \vdots \\ \mu_{N1} & \mu_{N2} & \dots & \mu_{NN} \end{bmatrix},$$

where $-\mu_{ii} = \mu_i = \sum_{j \neq i} \mu_{ij}$.
Let $J_n$ and $\bar{J}_n$ be the phase, on the major road, seen by the $n$-th low priority car at the beginning of its service when the $(n-1)$-th car left the system non empty and empty respectively. In other words, we can say that $J_n$ is the phase on the major road when the $(n-1)$-th car enters the major road.
We can write

$$A^*_{ij}(z) = \sum_k \bar{P}_{ik} A_{kj}(z),$$

where $\bar{P}_{ik} = \mathbb{P}(\bar{J}_n = k | J_n = i, X_{n-1} = 0)$ which is given by

$$\bar{P}_{ik} = \frac{\lambda}{\lambda + \mu_i} 1_{\{k=i\}} + \frac{\mu_i}{\lambda + \mu_i} \sum_{l \neq i} \frac{\mu_{il}}{\mu_i} \bar{P}_{lk}.$$

This implies that

$$\lambda \bar{P}_{ik} - \sum_l \mu_{il} \bar{P}_{lk} = \lambda 1_{\{k=i\}}.$$

We can write this in matrix form as

$$\lambda \bar{P} - Q\bar{P} = \text{diag}(\lambda),$$

and hence we obtain

$$\bar{P} = (I - \frac{1}{\lambda} Q)^{-1}.$$

Now we study the impact of the three types of the driver's behavior, described in Section 3.3, on the delay on the minor road.

**$\mathbf{B_1}$ (constant gap):** Every driver on the minor road needs the same constant critical headway T to enter the major road. Denote by $G^{(n)}$ the service time of the $n$-th minor road car and $J(t)$ the phase seen by the low priority car driver on the major road at time $t$ such that, for $i, j = 1, 2, \ldots, N$,

$$\begin{aligned} G_{ij}(x) &= \mathbb{P}(G^{(n)} \leq x, J_{n+1} = j | J_n = i, X_{n-1} \geq 1) \\ &= \mathbb{P}(G^{(n)} \leq x, J(G^{(n)}) = j | J(0) = i), \end{aligned} \tag{5.76}$$

with

$$\tilde{G}_{ij}(s) = \mathbb{E}[e^{-sG^{(n)}} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1]. \tag{5.77}$$

Now, firstly, we determine the probability that there is no car on the major road in [0,T] and $J(T) = j$, given that $J(0) = i$. For that we define $u_i(t) = \int_{u=0}^{t} 1_{\{J(u)=i\}} \mathrm{d}u$, with $\sum_{i=1}^{N} u_i(t) = t$. Therefore, we obtain,

$$\mathbb{P}(\text{No car on the major road in [0,T] and } J(T) = j | J(0) = i) = \phi_{ij}(T), \tag{5.78}$$

where

$$\begin{aligned} \phi_{ij}(T) &= e^{-q_i T} e^{-\mu_i T} 1_{\{i=j\}} + \int_{u=0}^{T} \mu_i e^{-\mu_i u} e^{-q_i u} \sum_{k \neq i} \frac{\mu_{ik}}{\mu_i} \phi_{kj}(T-u) \mathrm{d}u \\ &= e^{-(q_i+\mu_i)T} 1_{\{i=j\}} + \int_{u=0}^{T} e^{-(\mu_i+q_i)u} \sum_{k \neq i} \mu_{ik} \phi_{kj}(T-u) \mathrm{d}u. \end{aligned}$$

The Laplace-Stieltjes transform of $\phi_{ij}(t)$ is given by

$$\begin{aligned} &\tilde{\phi}_{ij}(\omega) = \int_{t=0}^{\infty} e^{-\omega t} \phi_{ij}(t) \mathrm{d}t \\ &= \frac{1}{\omega + \mu_i + q_i} 1_{\{i=j\}} + \frac{1}{\omega + \mu_i + q_i} \sum_{k \neq i} \mu_{ik} \tilde{\phi}_{kj}(\omega) \\ &= \frac{1}{\omega + \mu_i + q_i} 1_{\{i=j\}} + \frac{1}{\omega + \mu_i + q_i} \sum_{k=1}^{N} \mu_{ik} \tilde{\phi}_{kj}(\omega) + \frac{\mu_i}{\omega + \mu_i + q_i} \tilde{\phi}_{ij(\omega)}. \end{aligned}$$

This implies that, for $i, j = 1, \dots N$,

$$\frac{\omega + q_i}{\omega + \mu_i + q_i} \tilde{\phi}_{ij(\omega)} = \frac{1}{\omega + \mu_i + q_i} \left( 1_{\{i=j\}} + \sum_{k=1}^{N} \mu_{ik} \tilde{\phi}_{kj}(\omega) \right). \tag{5.79}$$

We can write the above system of equations in matrix form as

$$\text{diag}\left( \frac{\omega + q_i}{\omega + \mu_i + q_i} \right) \tilde{\phi}(\omega) = \text{diag}\left( \frac{1}{\omega + \mu_i + q_i} \right) \Big( I + Q\tilde{\phi}(\omega) \Big),$$

where $\text{diag}(d_i) = \text{diag}(d_1, d_2, \dots, d_N)$ is a diagonal matrix and $\tilde{\phi}(\omega) = [\tilde{\phi}_{ij}(\omega)]_{N \times N}$.

After simplification, we obtain $\tilde{\phi}(\omega)$ as

$$\tilde{\phi}(\omega) = \left( I - \text{diag}\left( \frac{1}{\omega + q_i} \right) Q \right)^{-1} \text{diag}\left( \frac{1}{\omega + q_i} \right). \tag{5.80}$$

Now, we need to determine the probability that at least one car arrives on the major road before time $T$. Let $T_{\text{next car}}$ be the time when the next car passes on the major road and $\Psi_{ij}(t) = \mathbb{P}(T_{\text{next car}} \leq t, J(T_{\text{next car}}) = j | J(0) = i)$. The LST is given by

$$\begin{aligned} \tilde{\psi}_{ij}(\omega) &= \mathbb{E}[e^{-\omega T_{\text{next car}}} 1_{\{J(T_{\text{next car}})=j\}} | J(0) = i] \\ &= \frac{\mu_i + q_i}{\omega + \mu_i + q_i} \Big( \frac{\mu_i}{\mu_i + q_i} \sum_{k \neq i} \frac{\mu_{ik}}{\mu_i} \tilde{\psi}_{kj}(\omega) + \frac{q_i}{\mu_i + q_i} 1_{\{i=j\}} \Big). \end{aligned}$$

After simplification, we can write this as

$$(\omega + q_i)\tilde{\psi}_{ij}(\omega) = \sum_{k=1}^{N} \mu_{ik} \tilde{\psi}_{kj}(\omega) + q_i 1_{\{i=j\}} \text{ for } i, j = 1, \dots, N,$$

and hence, in matrix form as

$$\text{diag}(\omega + q_i)\tilde{\psi}(\omega) = Q\tilde{\psi}(\omega) + \text{diag}(q_i),$$

where $\tilde{\psi}(\omega) = [\tilde{\psi}_{ij}(\omega)]_{N \times N}$.

Therefore, we obtain $\tilde{\psi}(\omega)$ as

$$\tilde{\psi}(\omega) = \left(I - \text{diag}\left(\frac{1}{\omega + q_i}\right) Q\right)^{-1} \text{diag}\left(\frac{q_i}{\omega + q_i}\right). \tag{5.81}$$

From Equations (5.80) and (5.81), we conclude the following relation

$$\tilde{\psi}(\omega) = \tilde{\phi}(\omega)\text{diag}(q_i). \tag{5.82}$$

Let $\phi_{ij}(t) = \mathcal{L}^{-1}(\tilde{\phi}_{ij}(\omega)), \psi_{ij}(t) = \mathcal{L}^{-1}(\tilde{\psi}_{ij}(\omega))$ for $i, j = 1, 2, \ldots, N$ such that $\phi(t) = [\phi_{ij}(t)]_{N\times N}, \psi(t) = [\psi_{ij}(t)]_{N\times N}$, where $\mathcal{L}^{-1}$ is the inverse Laplace–Stieltjes transform operator. From Equation (5.82), we obtain

$$\psi(t) = \phi(t)\text{diag}(q_i). \tag{5.83}$$

Hence, for $i, j = 1, 2, \ldots, N$, the LST of the conditional service time is given by

$$\begin{aligned}\tilde{G}_{ij}(s) = e^{-sT}\mathbb{P}(&\text{No car on the major road in [0,T] and } J(T) = j | J(0) = i) \\ &+ \int_{t=0}^{T} \sum_{k=1}^{N} \psi_{ik}(t) e^{-st} \tilde{G}_{kj}(s) \mathrm{d}t, \\ = e^{-sT}\phi_{ij}(T) &+ \sum_{k=1}^{N} \tilde{G}_{kj}(s) \int_{t=0}^{T} e^{-st}\psi_{ik}(t)\mathrm{d}t. \end{aligned} \tag{5.84}$$

We can solve this system of equations to obtain $\tilde{G}_{ij}(s)$.

**Special case:** Let $N = 2$, i.e., the MMPP is having two phases on the major road. In this case, we obtain $\tilde{\phi}(\omega)$ from (5.80) as

$$\begin{aligned}\tilde{\phi}(\omega) &= \begin{bmatrix} 1 + \frac{\mu_1}{\omega+q_1} & -\frac{\mu_1}{\omega+q_1} \\ -\frac{\mu_2}{\omega+q_2} & 1 + \frac{\mu_2}{\omega+q_2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\omega+q_1} & 0 \\ 0 & \frac{1}{\omega+q_2} \end{bmatrix} \\ &= \left(\frac{(\omega + q_1)(\omega + q_2)}{\omega^2 + (q_1 + \mu_1 + q_2 + \mu_2)\omega + \mu_1 q_2 + \mu_2 q_1 + q_1 q_2}\right) \\ &\times \begin{bmatrix} 1 + \frac{\mu_2}{\omega+q_2} & \frac{\mu_1}{\omega+q_1} \\ \frac{\mu_2}{\omega+q_2} & 1 + \frac{\mu_1}{\omega+q_1} \end{bmatrix} \begin{bmatrix} \frac{1}{\omega+q_1} & 0 \\ 0 & \frac{1}{\omega+q_2} \end{bmatrix} \\ &= \left(\frac{1}{\omega^2 + (q_1 + \mu_1 + q_2 + \mu_2)\omega + \mu_1 q_2 + \mu_2 q_1 + q_1 q_2}\right) \end{aligned}$$

$$\times \begin{bmatrix} \omega + q_2 + \mu_2 & \mu_1 \\ \mu_2 & \omega + q_1 + \mu_1 \end{bmatrix}. \tag{5.85}$$

Now, firstly, we determine the zeros (say $\omega_1, \omega_2$) of the polynomial $\omega^2 + (q_1 + \mu_1 + q_2 + \mu_2)\omega + \mu_1 q_2 + \mu_2 q_1 + q_1 q_2$ which are given by

$$\omega = \frac{-(q_1 + \mu_1 + q_2 + \mu_2)}{2} \pm \frac{\sqrt{q_1^2 + q_2^2 + \mu_1^2 + \mu_2^2 + 2q_1\mu_1 + 2\mu_1\mu_2 + 2q_2\mu_2 - 2\mu_1 q_2 - 2q_1\mu_2 - 2q_1 q_2}}{2}. \tag{5.86}$$

From Equation (5.86), we observe that the zeros $\omega_1$ and $\omega_2$ are real, distinct and non-positive. Moreover, without loss of generality, we assume that $\omega_1 > \omega_2$.

Therefore, we can write Equation (5.85) as

$$\tilde{\phi}(\omega) = \begin{bmatrix} \frac{\omega+q_2+\mu_2}{(\omega-\omega_1)(\omega-\omega_2)} & \frac{\mu_1}{(\omega-\omega_1)(\omega-\omega_2)} \\ \frac{\mu_2}{(\omega-\omega_1)(\omega-\omega_2)} & \frac{\omega+q_1+\mu_1}{(\omega-\omega_1)(\omega-\omega_2)} \end{bmatrix}.$$

After partial fractions, we obtain

$$\tilde{\phi}(\omega) = \frac{1}{\omega_1 - \omega_2} \begin{bmatrix} \frac{\omega_1+q_2+\mu_2}{\omega-\omega_1} - \frac{\omega_2+q_2+\mu_2}{\omega-\omega_2} & \frac{\mu_1}{\omega-\omega_1} - \frac{\mu_1}{\omega-\omega_2} \\ \frac{\mu_2}{\omega-\omega_1} - \frac{\mu_2}{\omega-\omega_2} & \frac{\omega_1+q_1+\mu_1}{\omega-\omega_1} - \frac{\omega_2+q_1+\mu_1}{\omega-\omega_2} \end{bmatrix}.$$

After taking the inverse Laplace transformation, the elements $\phi_{ij}(t)$ of the matrix $\phi(t)$ are given by

$$\phi_{ij}(t) = \begin{cases} \frac{\mu_i}{\omega_1-\omega_2}(e^{\omega_1 t} - e^{\omega_2 t}), i \neq j \\ \frac{1}{\omega_1-\omega_2}\Big((\omega_1 + q_{3-i} + \mu_{3-i})e^{\omega_1 t} - (\omega_2 + q_{3-i} + \mu_{3-i})e^{\omega_2 t}\Big), i = j \end{cases} \tag{5.87}$$

From Equation (5.83), we obtain the following relations

$$\psi_{ij}(t) = q_j \phi_{ij}(t), \qquad \text{for } i, j = 1, 2. \tag{5.88}$$

Now, we know the expressions for $\psi_{ij}(t)$ and $\phi_{ij}(t)$ which we need to determine the LST of the conditional service time. For $N = 2$, Equation (5.84) becomes

$$\tilde{G}_{ij}(s) = e^{-sT}\phi_{ij}(T) + \tilde{G}_{1j}(s)\int_{t=0}^{T} e^{-st}\psi_{i1}(t)\mathrm{d}t + \tilde{G}_{2j}(s)\int_{t=0}^{T} e^{-st}\psi_{i2}(t)\mathrm{d}t.$$

For $i = 1$, after substituting the values of $\psi_{ij}$, we obtain the above expression as

$$\tilde{G}_{1j}(s) = e^{-sT}\phi_{1j}(T) + \tilde{G}_{1j}(s)\int_{t=0}^{T} e^{-st}\frac{q_1}{\omega_1-\omega_2}\Big((\omega_1+q_2+\mu_2)e^{\omega_1 t} - (\omega_2+q_2+\mu_2)e^{\omega_2 t}\Big)\mathrm{d}t + \tilde{G}_{2j}(s)\int_{t=0}^{T} e^{-st}\frac{\mu_1 q_2}{\omega_1-\omega_2}(e^{\omega_1 t}-e^{\omega_2 t})\mathrm{d}t.$$

After simplification, we can write this as

$$\left(1 - \frac{q_1}{\omega_1-\omega_2}\Big(\frac{(\omega_1-\omega_2)(s+q_2+\mu_2)}{(s-\omega_1)(s-\omega_2)} - \frac{\omega_1+q_2+\mu_2}{s-\omega_1}e^{-(s-\omega_1)T} + \frac{\omega_2+q_2+\mu_2}{s-\omega_2}e^{-(s-\omega_2)T}\Big)\right)\tilde{G}_{1j}(s) - \frac{\mu_1 q_2}{\omega_1-\omega_2}\left(\frac{\omega_1-\omega_2}{(s-\omega_1)(s-\omega_2)} - \frac{1}{s-\omega_1}e^{-(s-\omega_1)T} + \frac{1}{s-\omega_2}e^{-(s-\omega_2)T}\right)\tilde{G}_{2j}(s) = e^{-sT}\phi_{1j}(T), \quad (5.89)$$

where $\phi_{1j}(T)$ is given by Equation (5.87) with $i = 1, t = T$.

Similarly, for $i = 2$, we obtain the following equation in $\tilde{G}_{1j}(s)$ and $\tilde{G}_{2j}(s)$

$$-\frac{\mu_2 q_1}{\omega_1-\omega_2}\left(\frac{\omega_1-\omega_2}{(s-\omega_1)(s-\omega_2)} - \frac{1}{s-\omega_1}e^{-(s-\omega_1)T} + \frac{1}{s-\omega_2}e^{-(s-\omega_2)T}\right) \times \tilde{G}_{1j}(s) + \left(1 - \frac{q_2}{\omega_1-\omega_2}\Big(\frac{(\omega_1-\omega_2)(s+q_1+\mu_1)}{(s-\omega_1)(s-\omega_2)} - \frac{\omega_1+q_1+\mu_1}{s-\omega_1} \times e^{-(s-\omega_1)T} + \frac{\omega_2+q_1+\mu_1}{s-\omega_2}e^{-(s-\omega_2)T}\Big)\right)\tilde{G}_{2j}(s) = e^{-sT}\phi_{2j}(T), \quad (5.90)$$

where $\phi_{2j}(T)$ is given by Equation (5.87) with $i = 2, t = T$.
Hence, we have two linear equations with two unknowns $\tilde{G}_{1j}(s)$ and $\tilde{G}_{2j}(s)$ for fixed $j$, which we can solve to obtain $\tilde{G}_{1j}(s)$ and $\tilde{G}_{2j}(s)$ for fixed $j$.

**$B_2$ (sampling per attempt):** Every car driver samples a random $T$ for each new 'attempt'. Therefore, using (5.78), for $i, j = 1, 2, \ldots, N$, the LST of the conditional service time is given by

$$\begin{aligned}\tilde{G}_{ij}(s) &= \mathbb{E}[e^{-sG^{(n)}} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1] \\ &= \mathbb{E}\Big[e^{-sT}\mathbb{P}(\text{No car on the major in [0,T] and } J(T) = j | J(0) = i) \\ &\qquad + \int_{t=0}^{T} \sum_{k=1}^{N} \psi_{ik}(t) e^{-st} \tilde{G}_{kj}(s) \mathrm{d}t\Big] \\ &= \mathbb{E}[e^{-sT}\phi_{ij}(T)] + \sum_{k=1}^{N} \tilde{G}_{kj}(s) \mathbb{E}\Big[\int_{t=0}^{T} \psi_{ik}(t) e^{-st} \mathrm{d}t\Big]. \end{aligned} \tag{5.91}$$

Now we have $N^2$ linear equations for $\tilde{G}_{ij}(s)$. The solution of this system of equations provides $\tilde{G}_{ij}(s)$.

**$B_3$ (sampling per driver):** Every car driver samples a random $T$ at his first attempt and this (random) value will be used consistently for each new attempt by this driver. Let $G^{(T,n)}$ be the service time of the $n$-th low-priority vehicle corresponding to the deterministic $T$ in the model $B_1$, then for $i, j = 1, 2, \ldots, N$, the LST of the conditional service time is given by

$$\tilde{G}_{ij}(s) = \mathbb{E}[e^{-sG^{(n)}} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1] = \mathbb{E}\Big[\tilde{G}_{ij}(T, s)\Big], \tag{5.92}$$

where $\tilde{G}_{ij}(T, s) = \mathbb{E}[e^{-sG^{(T,n)}} 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1]$ which we obtain from (5.84) as

$$\tilde{G}_{ij}(T, s) = e^{-sT}\phi_{ij}(T) + \sum_{k=1}^{N} \tilde{G}_{kj}(T, s) \int_{t=0}^{T} e^{-st} \psi_{ik}(t) \mathrm{d}t. \tag{5.93}$$

In this model, firstly, $\tilde{G}_{ij}(T, s)$ is determined from the system of equations (5.93). We then obtain the LST of the conditional service time, $\tilde{G}_{ij}(s)$, from Equation (5.92).

## 5.6 Numerical results

In this section, we present two numerical examples to demonstrate the impact of batch arrivals and Markov platooning, on the delay on the minor road. For simplicity, we restrict ourselves to two phases of the background process of the MMPP on the major road (N = 2).

### 5.6.1 Example 1: the impact of batch arrivals

In this example, we study similar settings as in Example 3.4.1 with batch arrivals and more realistic critical gaps, and compare the expected waiting times on the minor road, for the three behavior types $B_1$, $B_2$, and $B_3$. In the behavior type $B_2$, we assume that a minor road driver samples the critical gap (headway) of $6.22$ seconds with probability $0.9$ and $14$ seconds with probability $0.1$, at each new attempt. In $B_3$, $90\%$ of the minor road drivers need a gap of at least $6.22$ seconds; the other $10\%$ need at least $14$ seconds. For the behavior type $B_1$, we take the critical gap $T = 6.22 \times 0.9 + 14 \times 0.1 = 7$ seconds. On the major road, vehicles arrive according to rate $q_i$ (veh/h) in phase $i$, for $i = 1, 2$, with the fixed ratio $q_1 = 3q_2$, where the background process of the MMPP stays exponentially distributed times of, on average, $60$ seconds in phase $1$, and $240$ seconds in phase $2$, i.e., $\mu_1 = 1/60$ and $\mu_2 = 1/240$. Therefore, the long-term average arrival rate on the major road is given by

$$\bar{q} := \frac{q_1/\mu_1 + q_2/\mu_2}{1/\mu_1 + 1/\mu_2} = \frac{q_1\mu_2 + q_2\mu_1}{\mu_1 + \mu_2}. \tag{5.94}$$

We assume that the batch (platoon) arrival rate on the minor road is $\lambda = 50$ (batches per hour). We consider the following two distributions with the same mean $\mathbb{E}[B] = 4$ for the batch sizes on the minor road:

- Uniform distribution: $\mathbb{P}(B = k) = \begin{cases} 1/7, & \text{if } k = 1, 2, \ldots, 7 \\ 0, & \text{otherwise.} \end{cases}$
- Low and high distribution: $\mathbb{P}(B = k) = \begin{cases} 1/2, & \text{if } k = 1 \text{ and } k = 7 \\ 0, & \text{otherwise.} \end{cases}$

For the model without batch arrivals on the minor road (i.e., mean batch size 1), we take the arrival rate as $50 \times 4 = 200$ (veh/hour) to make a fair comparison

with the model with batch arrivals of the mean batch size $4$ and the batch arrival rate $50$ (per hour). From Figure 5.1, it is first noticed that the expected waiting times for the all three behavior types (denoted by $\mathbb{E}[W_1], \mathbb{E}[W_2]$ and $\mathbb{E}[W_3]$) depend not only on the mean batch size, but also on the full distribution of the batch sizes. Second, batch arrivals on the minor road, have a negative effect (compared to the individual arrivals on that road), as a function of the average flow rate on the major road, on the expected waiting times.

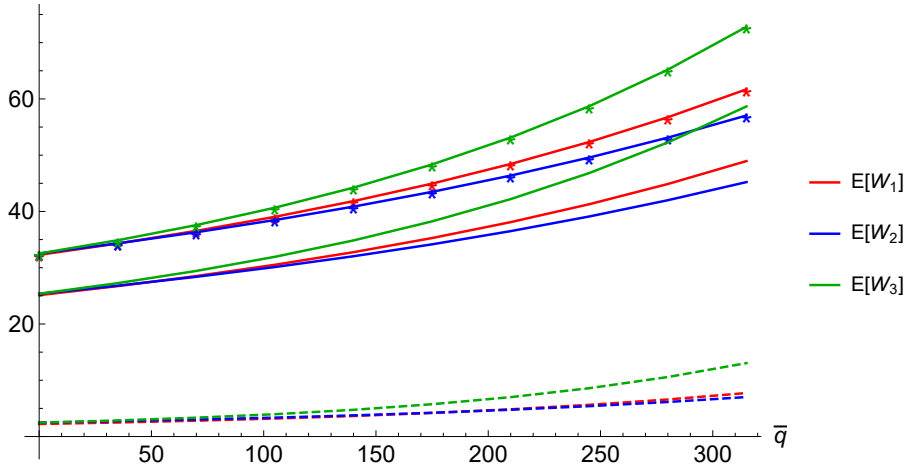
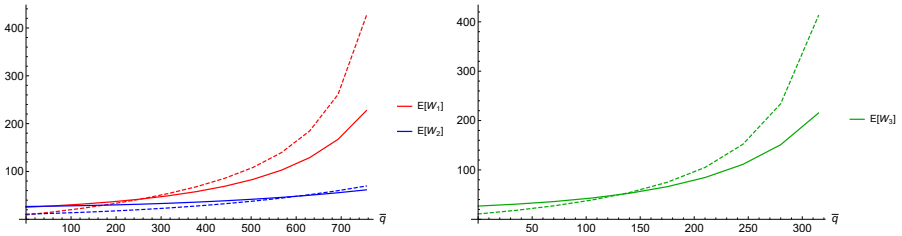



**Figure 5.1.** Expected waiting times of an arbitrary vehicle (seconds) on the minor road, as a function of the average flow rate on the major road (veh/h) in Example 1. The solid lines correspond to the model with batch arrivals where the solid lines with stars are for the low and high distribution of the batch sizes, and the solid lines without stars are for the uniformly distributed batch sizes; the dashed lines correspond to the model without batches.

### 5.6.2 Example 2: the impact of Markov platooning

In this example, we take the same settings as in Example 1, but we make two adjustments. First, we change the critical gaps from $6.22$ to $5$ seconds and from $14$ to $25$ seconds, for the behavior types $\mathrm{B}_2$ and $\mathrm{B}_3$, such that the expected critical gap equals $5 \times 0.9 + 25 \times 0.1 = 7$. For $\mathrm{B}_1$ we again take $T = 7$. Second, we fix the uniform distribution for the batch sizes on the

minor road, as considered in Example 1, where the batch arrival rate is taken as $\lambda = 50$ (batches per hour). For these settings, we compare the expected waiting times of the model with and without Markov platooning on the major road, where in the case without platooning, we assume Poisson arrivals on this road, with rate $\bar{q}$, which can be obtained from Equation (5.94). From Figure 5.2, one can observe that the impact of the Markov platooning depends on the parameters chosen. In particular, we see that the expected waiting times without Markov platooning, are first smaller than those with platooning, up to certain thresholds of $\bar{q}$, as a function of the flow rate on the major road, and these then become greater after those thresholds.




(a) Expected waiting times for models $B_1$ and $B_2$.

(b) Expected waiting times for model $B_3$.

**Figure 5.2.** Expected waiting times of an arbitrary vehicle (seconds) on the minor road, as a function of the average flow rate on the major road (veh/h) in Example 2. The solid lines correspond to the model with Markov platooning; the dashed lines correspond to the model without platooning.

## 5.7 Discussion and conclusion

In this chapter, we have first studied waiting and sojourn time distributions in the $M^X/\mathrm{SM}/1$ queue with exceptional first service. Later, we presented several applications in which this queueing model arises, and then studied the application to road traffic situations involving multiple conflicting traffic streams. In two numerical examples, we demonstrated the impact of the three types of driver's behavior ($B_1$, $B_2$, and $B_3$), on the delay on the minor road. In the next chapter, we will present a heavy-traffic analysis of this extended $M^X/\mathrm{SM}/1$ queueing model.

Chapter 6

# Heavy-traffic analysis of the $M^X$/SM/1 queue

In the previous chapter, we studied waiting and sojourn times in the $M^X/\mathrm{SM}/1$ queue with exceptional first service. This chapter presents a heavy-traffic analysis of that queueing model. In particular, we show that the distribution of the scaled stationary queue length in heavy traffic is exponential. A noteworthy observation is that the exceptional first service does not influence the final result.

## 6.1 Introduction

As stated in Section 5.1, a global overview of the earlier existing literature on the $M^X/\mathrm{SM}/1$ queue can be found in Chapter 4. In this section, we give a brief overview of the literature related in particular to the heavy-traffic

analysis of that queueing model.

One of the first studies for Markov-modulated single-server queueing systems in heavy traffic was by Burman and Smith [20], who study the mean delay and the mean number in queue in a single-server system in both light-traffic and heavy-traffic regimes, where customers arrive according to a nonhomogeneous Poisson process with rate equal to a function of the state of an independent Markov process. In their model, service times are independent and identically distributed. Later, G. Falin and A. Falin [42] suggest another approach to analyze the same queueing model, which is based on certain 'semi-explicit' formulas for the stationary distribution of the virtual waiting time and its mean value under heavy traffic. Dimitrov [34] applies the same approach of G. Falin and A. Falin to a single-server queueing system with arrival rate and service time depending on the state of Markov chain at the arrival epoch, and shows that the distribution of the scaled stationary virtual waiting time is exponential under a heavy-traffic scaling. Several other authors [7, 104] also study Markov-modulated $M/G/1$-type queueing systems in heavy traffic. However, we are not aware of any prior work analyzing the $M^X/\mathrm{SM}/1$ queue with exceptional first service under a heavy-traffic scaling.

The remainder of this chapter is organized as follows. In Section 6.2, we analyze the $M^X/\mathrm{SM}/1$ queue with exceptional first service under a heavy-traffic scaling. In Section 6.3, we present a numerical example in order to get more insight into the heavy-traffic behavior of the $M^X/\mathrm{SM}/1$ queue. We conclude with a summary in Section 6.4.

## 6.2 Heavy-traffic analysis

Throughout this chapter, we use the notations and assumptions of the extended $M^X/\mathrm{SM}/1$ queueing model with exceptional first service, described in Section 5.2. In this section, we shall determine the heavy-traffic limit of the scaled queue length at the departure epochs. In particular, we will show that under some conditions the distribution of the scaled stationary queue length in heavy traffic is exponential. This will be formally stated in Theorem 6.1.

As discussed before, denote by $J_n$ the type of the $n$-th customer, by $A_n$ the number of arrivals during its service time, and by $X_n$ the queue length at its departure time. In Section 5.2, it was noted that the system is stable

when $\rho < 1$, and $\rho$ is defined in Equation (5.14) as $\rho = \sum_{i=1}^{N} \pi_i \alpha_i$ where $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ is the stationary distribution of the irreducible discrete time Markov chain, with transition probabilities $P_{ij} = \mathbb{P}(J_{n+1} = j | J_n = i, X_{n-1} \geq 1)$, and $\alpha_i = \sum_{j=1}^{N} \alpha_{ij}$, with $\alpha_{ij} = \mathbb{E}[A_n 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1]$. Thus, the system is in heavy traffic when $\rho \uparrow 1$.

Let us define the heavy-traffic limit for the LST of the scaled queue length, $(1-\rho)X$, for $i = 1, 2, \ldots, N$:

$$\bar{f}_i(s) = \lim_{\rho \uparrow 1} \mathbb{E}[e^{-s(1-\rho)X_n} 1_{\{J_{n+1}=i\}}],$$

with

$$\bar{F}(s) = \sum_{i=1}^{N} \bar{f}_i(s).$$

Firstly, we introduce the following notations: for $i = 1, 2, \ldots, N$, $|z| \leq 1$,

$$\mathcal{A}_i(z) = \mathbb{E}[z^{A_n} | J_{n+1} = i, X_{n-1} \geq 1], \tag{6.1}$$

and

$$\gamma_i = \mathbb{E}[A_n | J_{n+1} = i, X_{n-1} \geq 1]. \tag{6.2}$$

We assume $z = e^{-s(1-\rho)}$, which is given by

$$z = 1 - s(1-\rho) + \frac{1}{2} s^2 (1-\rho)^2 + O((1-\rho)^3), \quad \text{as } \rho \uparrow 1. \tag{6.3}$$

Therefore, after using $z = e^{-s(1-\rho)}$, we can write the generating function of the batch size, $B(z)$, as

$$B(e^{-s(1-\rho)}) = 1 - s(1-\rho)\mathbb{E}[B] + \frac{1}{2} s^2 (1-\rho)^2 \mathbb{E}[B^2] + O((1-\rho)^3), \tag{6.4}$$

where $\mathbb{E}[B]$ is the mean batch size.
Similarly, for $i, j = 1, 2, \ldots, N$, $A_{ij}(z)$, $A^*_{ij}(z)$ and $\mathcal{A}_i(z)$ are obtained, with $z = e^{-s(1-\rho)}$, from Equations (5.12), (5.13) and (6.1) respectively as, for $\rho \uparrow 1$,

$$A_{ij}(e^{-s(1-\rho)}) = P_{ij} - s(1-\rho)\alpha_{ij} + \frac{1}{2} s^2 (1-\rho)^2 \hat{\alpha}_{ij} + O((1-\rho)^3), \tag{6.5}$$

$$A^*_{ij}(e^{-s(1-\rho)}) = P^*_{ij} - s(1-\rho)\alpha^*_{ij} + \frac{1}{2}s^2(1-\rho)^2\hat{\alpha}^*_{ij} + O((1-\rho)^3), \quad (6.6)$$

$$\mathcal{A}_i(e^{-s(1-\rho)}) = 1 - s(1-\rho)\gamma_i + \frac{1}{2}s^2(1-\rho)^2\hat{\gamma}_i + O((1-\rho)^3), \text{ as } \rho \uparrow 1, \quad (6.7)$$

where $\alpha_{ij}, \alpha^*_{ij}$, and $\gamma_i$ are respectively defined in Equations (5.16), (5.18) and (6.2), and

$$\hat{\alpha}_{ij} = \mathbb{E}[(A_n)^2 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} \geq 1], \quad (6.8)$$

$$\hat{\alpha}^*_{ij} = \mathbb{E}[(A^*_n)^2 1_{\{J_{n+1}=j\}} | J_n = i, X_{n-1} = 0], \quad (6.9)$$

$$\hat{\gamma}_i = \mathbb{E}[(A_n)^2 | J_{n+1} = i, X_{n-1} \geq 1]. \quad (6.10)$$

As a consequence of $A_i(z) = \sum_{j=1}^N A_{ij}(z)$ and $A^*_i(z) = \sum_{j=1}^N A^*_{ij}(z)$, we obtain,

$$A_i(e^{-s(1-\rho)}) = 1 - s(1-\rho)\alpha_i + \frac{1}{2}s^2(1-\rho)^2\hat{\alpha}_i + O((1-\rho)^3), \quad (6.11)$$

$$A^*_i(e^{-s(1-\rho)}) = 1 - s(1-\rho)\alpha^*_i + \frac{1}{2}s^2(1-\rho)^2\hat{\alpha}^*_i + O((1-\rho)^3), \text{ as } \rho \uparrow 1, \quad (6.12)$$

where $\alpha_i$ and $\alpha^*_i$ are respectively defined in Equations (5.15) and (5.17), and

$$\hat{\alpha}_i = \sum_{j=1}^N \hat{\alpha}_{ij}, \quad (6.13)$$

$$\hat{\alpha}^*_i = \sum_{j=1}^N \hat{\alpha}^*_{ij}. \quad (6.14)$$

After substituting the values of $B(z), A_{ij}(z)$ and $A^*_{ij}(z)$ from Equations (6.4), (6.5) and (6.6) respectively, we obtain $b_j(z)$, with $z = e^{-s(1-\rho)}$, from (5.28) as

$$b_j(e^{-s(1-\rho)}) = \sum_{i=1}^N \Big( (P^*_{ij} - P_{ij}) - s(1-\rho)(P^*_{ij}\mathbb{E}[B] + \alpha^*_{ij} - \alpha_{ij}) + \frac{s^2(1-\rho)^2}{2}(\mathbb{E}[B^2]P^*_{ij} + 2\mathbb{E}[B]\alpha^*_{ij} + \hat{\alpha}^*_{ij} - \hat{\alpha}_{ij}) + O((1-\rho)^3)\Big) f_i(0),$$

$$\text{as } \rho \uparrow 1. \quad (6.15)$$

It is noted that $\sum_{j=1}^{N} P_{ij} = 1, \sum_{j=1}^{N} P^*_{ij} = 1$. This implies that

$$\sum_{j=1}^{N} b_j(e^{-s(1-\rho)}) = -s(1-\rho)\sum_{i=1}^{N}\Big((\mathbb{E}[B] + \alpha_i^* - \alpha_i) - \frac{s(1-\rho)}{2}(\mathbb{E}[B^2] + 2\mathbb{E}[B]\alpha_i^* + \hat{\alpha}_i^* - \hat{\alpha}_i) + O((1-\rho)^2)\Big) f_i(0), \text{ as } \rho \uparrow 1. \quad (6.16)$$

Substituting the values of $z, A_{ij}(z), A_i(z), b_j(z)$, and $\sum_{j=1}^{N} b_j(z)$ from Equations (6.3), (6.5), (6.11), (6.15), and (6.16), respectively, in Equation (5.41), and after simplification, we obtain $\det L_i(z)$, with $z = e^{-s(1-\rho)}$, as

$$\det L_i(e^{-s(1-\rho)}) = -s(1-\rho)\times \begin{vmatrix} 1-\alpha_1 & \dots & 1-\alpha_{i-1} & \sum_{k=1}^{N}(\mathbb{E}[B] + \alpha_k^* - \alpha_k) f_k(0) & 1-\alpha_{i+1} & \dots & 1-\alpha_N \\ -P_{12} & \dots & -P_{i-12} & \sum_{k=1}^{N}(P^*_{k2} - P_{k2}) f_k(0) & -P_{i+12} & \dots & -P_{N2} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -P_{1N} & \dots & -P_{i-1N} & \sum_{k=1}^{N}(P^*_{kN} - P_{kN}) f_k(0) & -P_{i+1N} & \dots & 1-P_{NN} \end{vmatrix} + c_i s^2(1-\rho)^2 + O((1-\rho)^3), \quad i = 2, 3, \dots, N, \quad (6.17)$$

where $c_i$ is the coefficient of $s^2(1-\rho)^2$ term such that

$$\lim_{\rho \uparrow 1} c_i = 0, \quad (6.18)$$

because of $\lim_{\rho\uparrow 1} f_k(0) = 0$ for all $1 \le k \le N$, which can be concluded by $\lim_{\rho\uparrow 1} \mathbb{P}(X = 0) = 0$ with $\mathbb{P}(X = 0) = \sum_{k=1}^{N} f_k(0)$.

Now, differentiating Equation (5.41) w.r.t. $z$, and substituting $z = 1$, we get,

$$\frac{d}{dz}\{\det L_i(z)\}|_{z=1} = \begin{vmatrix} 1-\alpha_1 & \dots & 1-\alpha_{i-1} & \sum_{k=1}^{N}(\mathbb{E}[B] + \alpha_k^* - \alpha_k) f_k(0) & 1-\alpha_{i+1} & \dots & 1-\alpha_N \\ -P_{12} & \dots & -P_{i-12} & \sum_{k=1}^{N}(P^*_{k2} - P_{k2}) f_k(0) & -P_{i+12} & \dots & -P_{N2} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -P_{1N} & \dots & -P_{i-1N} & \sum_{k=1}^{N}(P^*_{kN} - P_{kN}) f_k(0) & -P_{i+1N} & \dots & 1-P_{NN} \end{vmatrix}. \quad (6.19)$$

After using Equations (5.32), (5.42) and (6.19) in Equation (6.17), we can write

$$\begin{aligned}
&\det L_i(e^{-s(1-\rho)}) \\
&= -s(1-\rho)\frac{d}{dz}\{\det L_i(z)\}|_{z=1} + c_i s^2(1-\rho)^2 + O((1-\rho)^3) \\
&= -s(1-\rho)\frac{d}{dz}\{\det M(z)^T\}|_{z=1} f_i(1) + c_i s^2(1-\rho)^2 + O((1-\rho)^3) \\
&= -sd(1-\rho)^2(f_i(1) - \frac{c_i s}{d}) + O((1-\rho)^3), \quad i = 2, 3, \ldots, N.
\end{aligned}$$

Similarly,

$$\det L_1(e^{-s(1-\rho)}) = -sd(1-\rho)^2(f_1(1) - \frac{c_1 s}{d}) + O((1-\rho)^3).$$

Hence, we can write, for $i = 1, 2, \ldots, N$,

$$\det L_i(e^{-s(1-\rho)}) = -sd(1-\rho)^2(f_i(1) - \frac{c_i s}{d}) + O((1-\rho)^3). \tag{6.20}$$

From Equation (5.30), $\det M(z)^T$ is given by

$$\begin{aligned}
\det M(z)^T &= \begin{vmatrix} z - A_1(z) & -A_{12}(z) & \ldots & -A_{1N}(z) \\ z - A_2(z) & z - A_{22}(z) & \ldots & -A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z - A_N(z) & -A_{N2}(z) & \ldots & z - A_{NN}(z) \end{vmatrix} \\
&= \frac{1}{\prod_{i=1}^N \pi_i} \begin{vmatrix} \pi_1(z - A_1(z)) & -\pi_1 A_{12}(z) & \ldots & -\pi_1 A_{1N}(z) \\ \pi_2(z - A_2(z)) & \pi_2(z - A_{22}(z)) & \ldots & -\pi_2 A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \pi_N(z - A_N(z)) & -\pi_N A_{N2}(z) & \ldots & \pi_N(z - A_{NN}(z)) \end{vmatrix}, \\
&\qquad \text{since } \pi_i \neq 0, i = 1, 2, \ldots, N.
\end{aligned} \tag{6.21}$$

It is noted that $\lim_{\rho\uparrow 1} \mathbb{P}(X = 0) = 0$ with $\mathbb{P}(X = 0) = \sum_{k=1}^N f_k(0)$. This implies that $\lim_{\rho\uparrow 1} f_k(0) = 0$ for all $1 \leq k \leq N$. Using $\lim_{\rho\uparrow 1} f_k(0) = 0$, we will first show that $\lim_{\rho\uparrow 1} f_j(1) = \pi_j$ for all $1 \leq j \leq N$. To do so, we first write $\lim_{\rho\uparrow 1} f_j(1)$ as

$$\lim_{\rho\uparrow 1} f_j(1) = \lim_{\rho\uparrow 1} \mathbb{P}(J_{n+1} = j)$$

$$
\begin{aligned}
&= \lim_{\rho\uparrow 1}(\mathbb{P}(J_{n+1}=j, X_{n-1}=0) + \mathbb{P}(J_{n+1}=j, X_{n-1}\geq 1))\\
&= \lim_{\rho\uparrow 1}\mathbb{P}(J_{n+1}=j|X_{n-1}\geq 1)\mathbb{P}(X_{n-1}\geq 1)\\
&= \lim_{\rho\uparrow 1}\sum_{i=1}^{N}\mathbb{P}(J_{n+1}=j|J_n=i, X_{n-1}\geq 1)\mathbb{P}(J_n=i|X_{n-1}\geq 1)\mathbb{P}(X_{n-1}\geq 1)\\
&= \lim_{\rho\uparrow 1}\sum_{i=1}^{N}P_{ij}P(J_n=i, X_{n-1}\geq 1)\\
&= \lim_{\rho\uparrow 1}\sum_{i=1}^{N}P_{ij}(\mathbb{P}(J_n=i) - \mathbb{P}(J_n=i, X_{n-1}=0))\\
&= \lim_{\rho\uparrow 1}\sum_{i=1}^{N}P_{ij}(f_i(1) - f_i(0))\\
&= \sum_{i=1}^{N}P_{ij}\lim_{\rho\uparrow 1}f_i(1), \quad \text{for } j=1,2,\ldots,N.
\end{aligned}
$$

As $P=[P_{ij}]_{i,j\in\{1,2,\ldots,N\}}$ is the transition probability matrix of an irreducible discrete time Markov chain, with stationary distribution $\pi=(\pi_1,\pi_2,\ldots,\pi_N)$, $\pi$ is the unique solution of the system of equations $\pi(I-P)=0$, and, hence, $\lim_{\rho\uparrow 1}f_j(1)=\pi_j$ for all $1\leq j\leq N$. As a consequence, we obtain $\lim_{\rho\uparrow 1}\mathbb{P}(J_n=j|X_{n-1}\geq 1)=\pi_j$.

Furthermore,

$$
\begin{aligned}
&\lim_{\rho\uparrow 1}\mathbb{P}(J_{n+1}=j|X_{n-1}\geq 1)\\
&= \lim_{\rho\uparrow 1}\sum_{i=1}^{N}\mathbb{P}(J_{n+1}=j|J_n=i, X_{n-1}\geq 1)\mathbb{P}(J_n=i|X_{n-1}\geq 1)\\
&= \sum_{i=1}^{N}P_{ij}\pi_i\\
&= \pi_j. \qquad (6.22)
\end{aligned}
$$

As a consequence, $\lim_{\rho\uparrow 1}\mathcal{A}_j(z)$ is given by

$$
\lim_{\rho\uparrow 1}\mathcal{A}_j(z) = \lim_{\rho\uparrow 1}\frac{\mathbb{E}[z^{A_n}1_{\{J_{n+1}=j\}}|X_{n-1}\geq 1]}{\mathbb{P}(J_{n+1}=j|X_{n-1}\geq 1)}
$$

$$= \lim_{\rho\uparrow 1} \frac{\sum_{i=1}^{N} \mathbb{P}(J_n = i|X_{n-1} \geq 1)\mathbb{E}[z^{A_n} 1_{\{J_{n+1}=j\}}|J_n = i, X_{n-1} \geq 1]}{\mathbb{P}(J_{n+1} = j|X_{n-1} \geq 1)}$$
$$= \frac{\sum_{i=1}^{N} \pi_i A_{ij}(z)}{\pi_j}. \tag{6.23}$$

Subsequently, we obtain,

$$\gamma_j = \frac{\sum_{i=1}^{N} \pi_i \alpha_{ij}}{\pi_j}, \quad \text{as } \rho \uparrow 1. \tag{6.24}$$

Replacing the first row by the sum of all $N$ rows in Equation (6.21), and using $\mathcal{A}_j(z) = \frac{1}{\pi_j} \sum_{i=1}^{N} \pi_i A_{ij}(z)$ as $\rho \uparrow 1$ and $\sum_{i=1}^{N} \pi_i = 1$ , we obtain $\det M(z)^T$ as, for $\rho \uparrow 1$,

$$\det M(z)^T$$
$$= \frac{1}{\prod_{i=1}^{N} \pi_i} \begin{vmatrix} z - \sum_{i=1}^{N} \pi_i A_i(z) & \pi_2(z - \mathcal{A}_2(z)) & \dots & \pi_N(z - \mathcal{A}_N(z)) \\ \pi_2(z - A_2(z)) & \pi_2(z - A_{22}(z)) & \dots & -\pi_2 A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \pi_N(z - A_N(z)) & -\pi_N A_{N2}(z) & \dots & \pi_N(z - A_{NN}(z)) \end{vmatrix}$$
$$= \frac{1}{\pi_1} \begin{vmatrix} z - \sum_{i=1}^{N} \pi_i A_i(z) & \pi_2(z - \mathcal{A}_2(z)) & \dots & \pi_N(z - \mathcal{A}_N(z)) \\ z - A_2(z) & z - A_{22}(z) & \dots & -A_{2N}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z - A_N(z) & -A_{N2}(z) & \dots & z - A_{NN}(z) \end{vmatrix}. \tag{6.25}$$

Substituting the values of $z, A_{ij}(z), A_i(z)$, and $\mathcal{A}_i(z)$ from Equations (6.3), (6.5), (6.11), and (6.7), respectively, in Equation (6.25), and after simplification, with $z = e^{-s(1-\rho)}, \rho = \sum_{i=1}^{N} \pi_i \alpha_i, \hat{\alpha} = \sum_{i=1}^{N} \pi_i \hat{\alpha}_i, \pi_1 = \frac{d_1}{d}$, we obtain

$$\det M(e^{-s(1-\rho)})^T = \frac{d}{d_1} \times$$
$$\begin{vmatrix} -s(1-\rho)^2(1 - \frac{s}{2}(1-\hat{\alpha})) & -\pi_2 s(1-\rho)(1-\gamma_2) & \dots & -\pi_N s(1-\rho)(1-\gamma_N) \\ -s(1-\rho)(1-\alpha_2) & 1 - P_{22} & \ddots & -P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -s(1-\rho)(1-\alpha_N) & -P_{N2} & \ddots & 1 - P_{NN} \end{vmatrix}$$

$$+ O((1-\rho)^3)$$

$$= \frac{-sd(1-\rho)^2}{d_1} \times$$

$$\begin{vmatrix} 1-\frac{s}{2}(1-\hat{\alpha}) & -\pi_2 s(1-\gamma_2) & \dots & -\pi_N s(1-\gamma_N) \\ 1-\alpha_2 & 1-P_{22} & \ddots & -P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 1-\alpha_N & -P_{N2} & \ddots & 1-P_{NN} \end{vmatrix} + O((1-\rho)^3)$$

$$= \frac{-sd(1-\rho)^2}{d_1}\Big((1-\frac{s}{2}(1-\hat{\alpha}))d_1 - s\sum_{k=2}^{N}\pi_k(1-\gamma_k)q_k\Big) + O((1-\rho)^3)$$

$$= -sd(1-\rho)^2\left(1+s\left(\frac{\hat{\alpha}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\gamma_k)q_k\right)\right) + O((1-\rho)^3), \tag{6.26}$$

where $d_1$ is defined in Equation (5.7), and $q_k$, $k = 2, 3, \dots, N$, is the cofactor of the entry in the first row and the $k$-th column of the matrix

$$\begin{bmatrix} 1-\frac{s}{2}(1-\hat{\alpha}) & -\pi_2 s(1-\gamma_2) & \dots & -\pi_N s(1-\gamma_N) \\ 1-\alpha_2 & 1-P_{22} & \ddots & -P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 1-\alpha_N & -P_{N2} & \ddots & 1-P_{NN} \end{bmatrix},$$

which is given by

$$q_2 = -\begin{vmatrix} 1-\alpha_2 & -P_{23} & -P_{24} & \dots & -P_{2N} \\ 1-\alpha_3 & 1-P_{33} & -P_{34} & \dots & 1-P_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1-\alpha_N & -P_{N3} & -P_{N4} & \dots & 1-P_{NN} \end{vmatrix}, \tag{6.27}$$

$$q_k = (-1)^{k+1}\begin{vmatrix} 1-\alpha_2 & 1-P_{22} & \dots & -P_{2k-1} & -P_{2K+1} & \dots & -P_{2N} \\ 1-\alpha_3 & -P_{32} & \dots & -P_{3k-1} & -P_{3K+1} & \dots & 1-P_{3N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1-\alpha_N & -P_{N2} & \dots & -P_{Nk-1} & -P_{NK+1} & \dots & 1-P_{NN} \end{vmatrix},$$

$$\text{for } k = 3, 4, \dots, N. \tag{6.28}$$

Therefore, using Equation (6.18) and $\lim_{\rho\uparrow 1} f_i(1) = \pi_i$, the heavy-traffic limit

of the scaled queue length, $\bar{f}_i(s) = \mathbb{E}[e^{-s(1-\rho)X_n} 1_{\{J_{n+1}=i\}}]$, $i = 1, 2, \ldots, N$, is obtained from Equation (5.29) as

$$\begin{aligned}\bar{f}_i(s) &= \frac{\det L_i(e^{-s(1-\rho)})}{\det M(e^{-s(1-\rho)})^T} \\ &= \frac{-sd(1-\rho)^2(f_i(1) - \frac{c_i s}{d}) + O((1-\rho)^3)}{-sd(1-\rho)^2\left(1 + s\left(\frac{\hat{\alpha}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\gamma_k)q_k\right)\right) + O((1-\rho)^3)} \\ &\approx \frac{\pi_i}{1 + s\left(\frac{\bar{\hat{\alpha}}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k\right)} \quad \text{as } \rho \uparrow 1, \end{aligned} \tag{6.29}$$

where $\lim_{\rho\uparrow 1}\hat{\alpha} = \bar{\hat{\alpha}}$, $\lim_{\rho\uparrow 1}\gamma_k = \bar{\gamma}_k$ and $\lim_{\rho\uparrow 1} q_k = \bar{q}_k$.

Hence $\bar{F}(s) = \sum_{i=1}^{N}\bar{f}_i(s)$ tends to

$$\frac{1}{1 + s\left(\frac{\bar{\hat{\alpha}}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k\right)} \quad \text{as } \rho \uparrow 1, \tag{6.30}$$

provided $\left(\frac{\bar{\hat{\alpha}}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k\right) > 0$. We have now proved the following theorem:

**Theorem 6.1.** *If $E[B^2]$ and $\hat{\alpha}_{ij}$ are finite for $i, j = 1, 2, \ldots, N$, then*

$$\lim_{\rho\uparrow 1}\bar{F}(s) = \lim_{\rho\uparrow 1}\mathbb{E}[e^{-s(1-\rho)X}] = \frac{1}{1 + s\left(\frac{\bar{\hat{\alpha}}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k\right)},$$

*provided $\left(\frac{\bar{\hat{\alpha}}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k\right) > 0$, which is the LST of the exponentially distributed random variable with the rate parameter*

$$\frac{1}{\frac{\bar{\hat{\alpha}}-1}{2} - \frac{1}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k}.$$

**Remark 6.1.** *If $\alpha_i = \alpha$ for all $i = 1, 2, \ldots, N$, then Equation* (5.14) *implies that $\rho = \alpha$. Therefore, the system then is in heavy traffic when $\alpha \uparrow 1$. And as a consequence, $\alpha_i \uparrow 1$ for all $i = 1, 2, \ldots, N$.*

*Note that each element of the first column of $q_k$, $k = 2, 3, \ldots, N$, tends to zero, as $\alpha_i \uparrow 1$ for all $i = 1, 2, \ldots, N$. And as a consequence, $q_k = 0$, and hence*

$\bar{q}_k = 0$ *for all* $k = 2, 3, \ldots, N$. *Therefore,* $\bar{F}(s)$ *tends to*

$$\frac{1}{1 + s\left(\frac{\bar{\hat{\alpha}}-1}{2}\right)}, \quad \text{as } \alpha = \rho \uparrow 1,$$

*which is the heavy-traffic limit of the scaled queue length of the standard* $M^X/G/1$ *without dependencies at the departure epochs. Furthermore, we can conclude that the term* $-\frac{s}{d_1}\sum_{k=2}^{N}\pi_k(1-\bar{\gamma}_k)\bar{q}_k$ *in Equation* (6.30) *appears due to the dependent service times.*

**Remark 6.2.** *For* $N = 2$,

$$\lim_{\rho\uparrow 1}\bar{F}(s) = \frac{1}{1 + s\left(\frac{\bar{\hat{\alpha}}-1}{2} + \frac{(1-\bar{\alpha}_2)}{P_{12}+P_{21}}\left(\frac{P_{12}}{P_{21}}(1-\bar{\alpha}_{22}) - \bar{\alpha}_{12}\right)\right)}, \quad \text{as } \rho \uparrow 1,$$

*where* $\lim_{\rho\uparrow 1}\hat{\alpha} = \bar{\hat{\alpha}}$, $\lim_{\rho\uparrow 1}\alpha_i = \bar{\alpha}_i$ *and* $\lim_{\rho\uparrow 1}\alpha_{ij} = \bar{\alpha}_{ij}$ *for* $i, j = 1, 2$.

*Furthermore, when* $\frac{(1-\bar{\alpha}_2)}{P_{12}+P_{21}}\left(\frac{P_{12}}{P_{21}}(1-\bar{\alpha}_{22}) - \bar{\alpha}_{12}\right) = 0$, *then* $\bar{F}(s)$ *becomes* $\left(1 + s\left(\frac{\bar{\hat{\alpha}}-1}{2}\right)\right)^{-1}$, *which is the heavy-traffic limit of the scaled queue length of the standard* $M^X/G/1$ *without dependencies at the departure epochs.*

**Remark 6.3.** *After using Equation* (6.30) *in* (5.54) *and* (5.55), *it can be shown by substituting* $z$ *with* $e^{-s(1-\rho)}$ *and taking* $\rho \uparrow 1$ *that the heavy-traffic distribution of the scaled stationary queue length at an arbitrary epoch is the same as the heavy-traffic distribution of the scaled stationary queue length at a departure epoch. Furthermore, this heavy-traffic limit can be used in Equations* (5.74) *and* (5.75) *to derive the scaled stationary waiting and sojourn time distributions in heavy traffic.*

## 6.3 Numerical example

We present this numerical example in order to get more insight into the heavy-traffic behavior of the $M^X/SM/1$ queue. In particular, we show that when $\rho$ tends to 0 or 1 (i.e. light-traffic or heavy-traffic regime), the mean scaled queue length of the $M^X/SM/1$ model is equal to the mean scaled queue length of the corresponding $M^X/G/1$ model; however, for $0 < \rho < 1$ (excluding the scenario where $\rho$ tends to 0 or 1), this equality is not necessarily true.

In this example, for simplicity, we assume $N = 2$, $B(z) = z$ and $\tilde{G}^*_{ij}(s) = \tilde{G}_{ij}(s)$ for all $i, j = 1, 2$, i.e., there are two customer types, the batch size is one, and the customers arriving in the empty system have the same service-time distributions as regular customers. The conditional service times are Erlang distributed random variables, with

$$G_{ij}(x) = \left(1 - \sum_{m=0}^{k_{ij}-1} \frac{(\mu_{ij}x)^m}{m!} e^{-\mu_{ij}x}\right) P_{ij},$$

where $k_{ij} = i + j, \mu_{ij} > 0, i, j = 1, 2$.

We can use Equation (5.19) to obtain

$$A_{ij}(z) = P_{ij} \left(\frac{\mu_{ij}}{\lambda(1 - B(z)) + \mu_{ij}}\right)^{k_{ij}}, \quad \text{for } i, j = 1, 2.$$

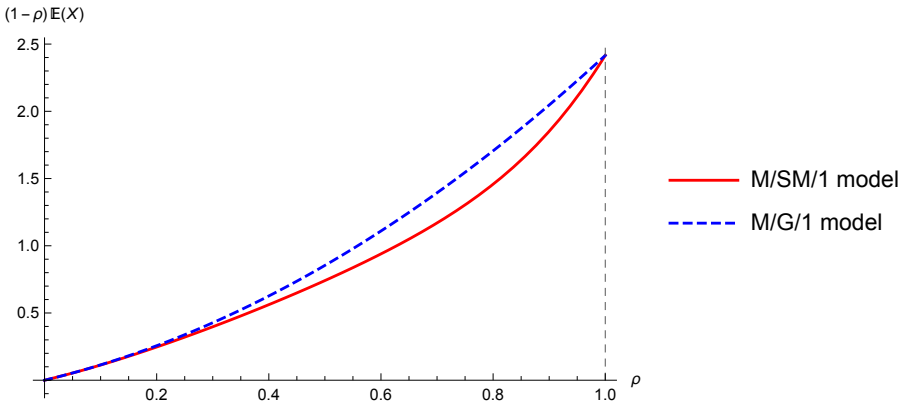



**Figure 6.1.** The mean scaled queue length versus the number of arrivals per time unit.

We choose model parameters $P_{11} = 0.9, P_{22} = 0.951138, \alpha_{11} = \lambda, \alpha_{12} = 3\lambda, \alpha_{21} = 10\lambda$, and $\alpha_{22} = 20\lambda$ in such a way that $\frac{P_{12}}{P_{21}}(1 - \bar{\alpha}_{22}) - \bar{\alpha}_{12} = 0$ as $\rho \uparrow 1$ where $\lim_{\rho\uparrow 1} \alpha_{ij} = \bar{\alpha}_{ij}$ for $i, j = 1, 2$. From Remark 6.2, we know that the heavy-traffic limit of the scaled queue length in the $M/SM/1$ model is equal to the heavy-traffic limit of the scaled queue length of the corresponding $M/G/1$ model. In Figure 6.1, it can be observed that, except for the scenario

where $\rho$ tends to 0 or 1, the mean scaled queue length of the $M/SM/1$ model is smaller than the mean scaled queue length of the corresponding $M/G/1$ model. This difference occurs due to the dependence between the subsequent service times of customers. However, when $\rho$ tends to 0 or to 1 (i.e. light-traffic or heavy-traffic regime), the dependence between the subsequent service times no longer influences the scaled queue length, and thus the system can be analyzed as an $M/G/1$ queueing system in such types of situations. Furthermore, in Figure 6.2, it can be seen that the density of the scaled queue length converges to the limiting density when the traffic intensity $\rho$ approaches 1.

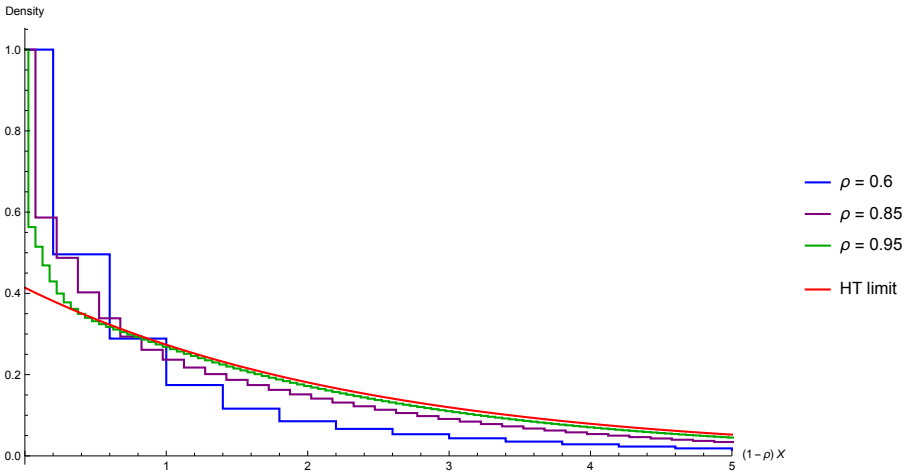



**Figure 6.2.** The density of the scaled queue length.

## 6.4 Discussion and conclusion

In this chapter, we have investigated the $M^X/SM/1$ queue with exceptional first service under a heavy-traffic scaling. In particular, it was shown that the distribution of the scaled stationary queue length is exponential in the heavy-traffic limit, which does not depend on the service time of the first customer of each busy period. This heavy-traffic limit can be further used in Equations (5.74) and (5.75) to derive the scaled stationary waiting and sojourn time distributions under the heavy-traffic regime.

Chapter 7

# Generalized gap acceptance models

The main contribution of this chapter is to develop a series of generalizations of the gap acceptance models, studied in Chapters 2 and 3, thus increasing the model's practical applicability significantly. First, we incorporate driver impatience behavior while allowing for a realistic merging behavior; we do so by distinguishing between the critical gap and the merging time, thus allowing *multiple* vehicles to use a sufficiently large gap. Incorporating this feature is particularly challenging in models with driver impatience. Secondly, we allow for multiple classes of gap acceptance behavior, enabling us to distinguish between different driver types and/or different vehicle types. Thirdly, we use the novel $M^X/\mathrm{SM}/1$ queueing model with exceptional first service ($M^X/\mathrm{SM2}/1$), studied in Chapter 5, which has batch arrivals, dependent service times, and a different service-time distribution for vehicles arriving in an empty queue on the minor road. This setup facilitates the analysis of the service-time distribution of an arbitrary vehicle on the minor road and of the queue length on the minor road. In particular, we can compute the *mean*

service time, thus enabling the evaluation of the capacity of the minor road.

## 7.1 Introduction

In this chapter, we contribute to the modeling and analysis of unsignalized intersections. In classical gap acceptance models vehicles on the minor road accept any gap greater than the *critical* gap (or critical headway), and reject gaps below this threshold, where the gap is defined as the time between two subsequent vehicles on the major road. However, in reality, drivers typically do not need the full critical gap; the remaining part can be used by the next vehicle on the minor road. In the literature, the part of the critical gap that is really used by the vehicle is referred to as the *merging* time.

Although gap acceptance models have improved greatly since their introduction more than fifty years ago, there are still some fundamental limitations to their practical applicability. As pointed out recently by Liu *et al.* [74], two of its main limitations are the inability to incorporate (1) different driver behaviors and (2) heterogeneous traffic; see also [84, 85]. The main reason why these features could not be included is that the currently used analysis techniques are based on a single-server queueing model with exceptional first service (see also [115, 125, 126]). In the gap acceptance literature, this model is commonly referred to as the $M/G2/1$ queue, a term seemingly introduced by Daganzo [33]. Importantly, in this queue, the 'service times' (corresponding to the time required to search for a sufficiently large gap and crossing the intersection or, depending on the application, merging with the high-priority traffic flow) are assumed independent. Incorporating more realistic features, such as driver impatience or different types of driver behavior, creates dependencies that make the model significantly more difficult to analyze.

An important development, that helps us overcome this obstacle, is the recently developed framework for the analysis of queueing models that allows a very general correlation structure between successive service times (the $M^X/\mathrm{SM}2/1$ queue, studied in Chapter 5). Although set up with general applications in mind, this framework turns out to be particularly useful in road traffic models, where the dependencies can be used to model clustering of vehicles on the major road, to differentiate between multiple types of driver behavior, or to account for heterogeneous traffic. The present chapter focuses on the latter two features.

In the existing literature, various quantitative methods have been used to study unsignalized intersections. The two most common procedures are empirical regression techniques and gap acceptance models [14], but other methods have been employed as well (for example the additive conflict flows technique by Brilon and Wu [16]). A topic of particular interest concerns the capacity of the minor road [13], which is defined as the maximum possible number of vehicles per time unit that can pass through an intersection from the minor road. Other relevant performance measures are the queue length and the delay on the minor road.

In Chapters 2 and 3, three variations of gap acceptance models ($B_1$ up to $B_3$) are studied. The present chapter generalizes all three gap acceptance models into a more realistic model that covers various realistic driving behavior features.

One of the first studies using queueing models for unsignalized intersections was by Tanner [101], who assumes constant critical gap and move-up time. Tanner first determines the mean delay of the low priority vehicles, and characterizes the capacity of the minor road as the arrival rate of the low-priority vehicles, at which the mean delay grows beyond any bound. Initially one worked with a constant critical gap (for all users), resulting in a nice, compact, closed-form expression for the capacity (cf. [115, 125, 126]). Then one empirically observed that the statistical variations of the drivers' critical gaps have impact; Siegloch [94] and Harders [54, 55] developed some theoretically based variations and corrections. One of the adaptations was applied to the capacity estimate in which the old number is multiplied with $f := 1 - 10^{-7} \cdot q^2$, where $q$ is the arrival rate on the major road, to account for various sources of randomness. Later studies [12] mention that the approach with a universal factor $f$ does not always work. In addition, it is empirically observed that drivers tend to finally accept gaps of a value they have rejected before [4, 43, 88, 106, 110]. This resulted in models that also included driver impatience (cf. [2, 37, 38, 114]), which complicates the analysis drastically (cf. [60]), in particular when trying to maintain a realistic merging behavior. Tanner's model has been generalized in various ways [22, 27, 58, 59, 111, 114] by allowing random critical gaps and move-up times, also analyzing performance measures such as the queue length and the waiting time on the minor road. Heidemann and Wegmann [60] give an excellent overview of the earlier existing literature. Moreover, they add a stochastic dependence between the

critical gap and the merging (move-up) time, and study the minor road as an $M/G2/1$ queue, to determine the queue length, the delay and the capacity on the minor road. A further generalization, dividing the time scale of the major stream into four regimes (viz. free space, single vehicle, bunching, and queueing) was investigated by Wu [119].

Although the literature on gap acceptance models is relatively mature and the existing gap acceptance models have proven their value, they are sometimes criticized for *looking like* rigorous mathematics, but in reality being based on pragmatic simplifications. As a consequence, the produced results might be of a correct magnitude, but are, only of a approximative nature [16]. The main goal of this chapter is to increase the significance of gap acceptance models by taking away some of these concerns via inclusion of essential new features that are typically encountered in practice. In more detail, the main novelties of this chapter are the following.

- First, we incorporate driver impatience behavior while allowing for a realistic merging behavior. Put more precisely, we distinguish between the critical gap and the merging time, thus allowing *multiple* vehicles to use a sufficiently large gap. Incorporating this feature is particularly challenging in models with driver impatience.

- Secondly, we allow for multiple classes of gap acceptance behavior, enabling us to distinguish between different driver types and/or different vehicle types.

- Thirdly, we use a queueing model in which vehicles arriving in an empty queue on the minor road have different service-time distributions than the queueing vehicles (where 'service time' is meant in the sense introduced above). This setup facilitates the analysis of the queue length on the minor road as well as the service-time distribution of an arbitrary low-priority vehicle. The capacity for the minor road vehicles is then derived from the expectation of the service time for queuers.

The remainder of this chapter is organized as follows. In Section 7.2 we introduce our model. Then the full queue-length analysis is presented in Section 7.3, whereas the capacity of the minor road is determined in Section 7.4. In Section 7.5 numerical examples demonstrate the impact of driver behavior

on the capacity of the minor road. Finally, we present our conclusions in Section 7.6.

## 7.2 Model description

In this section we provide a detailed mathematical description of our new model. We study an unsignalized, priority controlled intersection where drivers on the major road have priority over the drivers on the minor road. For notational convenience, we will focus on the situation with one traffic stream on each road, but several extensions also fall in our framework (see, for example, Figure 1.1). Vehicles on the major road arrive according to a Poisson process with intensity $q$. On the minor road, we have a *batch* Poisson arrival process with $\lambda$ denoting the arrival intensity of the batches (platoons) and $B$ denoting the (random) platoon size. We assume that traffic on the major road is not hindered by traffic on the minor road, which is a reasonable assumption in most countries (but not all, see [74]).

The drivers on the minor road cross the intersection as soon as they come across a sufficiently large gap between two subsequent vehicles on the major road. Any gap that is too small for a driver on the minor road is considered a *failed attempt* to cross the intersection and adds to the impatience of the driver.

In the existing literature, *three* model variants have been introduced: (1) the standard model with constant critical gaps (referred to as model $B_1$ in this thesis, cf. Section 1.2), (2) inconsistent gap acceptance behavior where a new random critical gap is sampled at each attempt ($B_2$), and (3) consistent behavior where drivers sample a random critical gap which they use for all attempts ($B_3$). In this chapter, we apply a framework that allows us to create a *single* model that generalizes all three aforementioned model variations, with a combination of consistent *and* inconsistent behavior, allowing for heterogeneous traffic and driver impatience. This is a major enhancement of the existing models, as such a general model has not been successfully analyzed so far. Unfortunately, this requires a significant adaptation of the underlying $M/G2/1$ queueing model that has been the basis of all gap acceptance models so far.

We will now describe the full model dynamics in greater detail. We dis-

tinguish between $R \in \{1, 2, \ldots\}$ driver/vehicle profiles. The profile of any arriving vehicle on the minor road is modeled as a random variable that equals $r$ with probability $p_r$ (with $\sum_{r=1}^{R} p_r = 1$). The profile represents the *consistent part* of the gap selecting behavior. Each driver profile has its own critical gap acceptance behavior, allowing us to distinguish between drivers that are willing to accept small gaps and (more cautious) drivers that require longer gaps, but also to allow for heterogeneous traffic with, for example, trucks requiring larger gaps than cars or motor cycles.

We introduce $T_{(i,r)}$ to denote the critical gap of a driver with profile $r$ during his $i$-th attempt. The *impatient behavior* is incorporated in our framework by letting the critical gap $T_{(i,r)}$ depend on the attempt number $i = 1, 2, 3, \ldots$. This generalizes classical gap acceptance models where the critical gap is the same for all driver types and throughout all attempts.

Finally, we introduce the *inconsistent* component of our model by allowing the critical gap of a driver of type $r$ in attempt $i$ ($T_{(i,r)}$, that is) to be a discrete random variable which can take any of the following $M$ values:

$$T_{(i,r)} = \begin{cases} u_{(i,1,r)} & \text{with probability } p_{(i,1,r)}, \\ u_{(i,2,r)} & \text{with probability } p_{(i,2,r)}, \\ \vdots & \\ u_{(i,M,r)} & \text{with probability } p_{(i,M,r)}, \end{cases}$$

with $\sum_{k=1}^{M} p_{(i,k,r)} = 1$ for all $(i, r)$. Inconsistent behavior is used in the existing literature when the driver samples a new random critical gap after each failed attempt. It is generally being criticized for not being very realistic, because it is unlikely that a driver's gap acceptance behavior fluctuates significantly throughout multiple attempts. In our model, however, it is an excellent way to include some randomness in the gap selection process (which is due to the fact that not every driver from the same profile will have *exactly* the same critical gap at his $i$-th attempt), because we can limit the variability by suitably choosing the values $u_{(i,k,r)}$.

Realistic values for the critical gaps $u_{(i,k,r)}$ can be empirically obtained by measuring the lengths of all rejected and accepted gaps for each vehicle type. One needs to distinguish between systematic fluctuations due to different driver/vehicle types and (smaller) random fluctuations within the same driver profile. Since the index $i$ represents impatience, it makes sense that $u_{(i,k,r)}$

is decreasing in $i$. Moreover, since $r$ denotes the driver profile, $u_{(i,k,r)}$ is relatively large for driver profiles $r$ that represent slow vehicles/drivers and vice versa. The variability in $u_{(i,1,r)}, \ldots, u_{(i,M,r)}$ can be limited to avoid unrealistic fluctuations. For example: if driver profile $r = 1$ corresponds to trucks and $r = 2$ corresponds to motor cycles, then it may be quite realistic that

$$\min_{1 \leq k \leq M} u_{(i,k,1)} > \max_{1 \leq k \leq M} u_{(i,k,2)}.$$

Paraphrasing, the slowest motor cycle will always be faster than the fastest truck.

To make the model even more realistic, we assume that the actual vehicle *merging time*, denoted by $\Delta_r$ for profile $r$, is less than the critical gap $T_{(i,r)}$. As a consequence, the remaining part of the critical headway can be used by following drivers.

At first sight, our model has a few limitations. In the first place, we assume the critical gap has a discrete distribution with $M$ possible values. In the second place, as will be discussed in great detail in the next section, the analysis technique requires that there is a maximum number of possible attempts: $i \in \{1, 2, \ldots, N\}$. It is important to note, however, that both issues do not have major practical consequences, since one can choose $M$ and $N$ quite large. This does come at the expense of increased computation times, obviously.

## 7.3 Queue length analysis

The main objective of this section is to determine the stationary queue length on the minor road at departure epochs of low-priority vehicles. In the next section we will use these results to determine the service-time distribution of the low-priority vehicles, which enables us to evaluate the capacity of the minor road.

### 7.3.1 Preliminaries

We start by describing the queueing process on the minor road. All arriving vehicles experience three stages before leaving the system: queueing, scanning and merging. The queueing phase is defined as the time between joining the queue and reaching the front of the queue (for queuers this is the moment

that the preceding driver accepts a gap and departs). The scanning phase commences when the driver starts scanning for gaps and ends when a sufficiently large gap is found. The scanning phase is followed by the merging phase, which ends at the moment that the vehicle merges into the major stream. Note that phases 1 and 2 may have zero length. The length of phase 1 is called the *delay*. The length of phase 3 for a vehicle of profile $r$ is denoted by the profile-specific constant $\Delta_r$. We define the *service time* as the total time spent in phases 2 and 3 (scanning and merging).

**Remark 7.1.** *In some papers, an additional move-up phase is distinguished but, depending on the application (freeway or unsignalized intersection), this phase can be incorporated in one of the other phases. See Heidemann and Wegmann [60, Section 3] for more details.*

The queue-length analysis is based on observing the system at *departure epochs*. In contrast to the classical $M/G2/1$ queueing models, in our current model we now have *dependent* service times. This dependence is caused by the fact that the length of the lag (i.e., the remaining gap) left by the previous driver, due to the impatience, now depends on the attempt number *and* the profile of the previous driver. Using the notation introduced in the previous section: we need to know the critical gap $T_{(i,r)}$ of the previous driver of profile $r$ that led to a successful merge/crossing of the intersection. A queueing model with this type of dependencies is called a queueing model with *semi-Markovian* service times, sometimes referred to as the $M/\mathrm{SM}/1$ queue. Although several papers have studied such a queueing model (cf. [1, 5, 23, 97, 48, 82, 81, 80, 86]), we still need to make several adjustments in order to make it applicable to our situation. The analysis below is based on the framework that was introduced in Chapter 5.

Denote by $X_n$ the queue length at the departure epoch of the $n$-th vehicle, i.e. the number of vehicles *behind* the $n$-th driver at the moment that he merges into the major stream. We use $G^{(n)}$ to denote the service time of the $n$-th vehicle and $A_n$ to denote the number of arrivals (on the minor road) during this service time. Since vehicles arrive in batches (platoons), we denote by $B_n$ the size of the batch in which the $n$-th vehicle arrived. Let $F(\cdot)$, $A(\cdot)$, and $B(\cdot)$ denote, respectively, the probability generating functions (PGFs) of the limiting distributions of these random variables,

$$F(z) = \lim_{n\to\infty} \mathbb{E}[z^{X_n}], A(z) = \lim_{n\to\infty} \mathbb{E}[z^{A_n}], B(z) = \lim_{n\to\infty} \mathbb{E}[z^{B_n}], \quad |z| \leq 1,$$

whereas $\tilde{G}(\cdot)$ is the limiting Laplace-Stieltjes transform (LST) of $G^{(n)}$:

$$\tilde{G}(s) = \lim_{n\to\infty} \mathbb{E}[e^{-sG^{(n)}}], \qquad s \geq 0.$$

Note that the arrivals constitute a batch Poisson process. Therefore, $A(z)$ can be expressed in terms of $B(z)$ and $\tilde{G}(s)$,

$$A(z) = \tilde{G}(\lambda(1 - B(z))), \tag{7.1}$$

but finding an expression for $\tilde{G}(s)$ is quite tedious. For this reason, we split the analysis in two parts. This section discusses how to find $F(z)$ in terms of $A(z)$, while Section 7.4 shows how to find $G(z)$ for the gap acceptance model considered in this chapter.

The starting-point of our analysis is the following recurrence relation:

$$X_n = \begin{cases} X_{n-1} - 1 + A_n & \text{if } X_{n-1} \geq 1 \\ A_n + B_n - 1 & \text{if } X_{n-1} = 0 \end{cases}, \quad n = 1, 2, 3, \ldots. \tag{7.2}$$

As argued before, $\{X_n\}_{n\geq 1}$ is not a Markov chain. In order to obtain a Markovian model, we need to keep track of all the characteristics of the $(n-1)$-th driver. Let $J_n = (i, k, r)$ contain these characteristics of the $(n-1)$-th driver, where $i$ is the 'succeeded attempt number', $k$ indicates which random gap the driver sampled, and $r$ denotes the profile of the driver.

In Chapter 5 it is shown that the ergodicity condition of this system is given by

$$\lim_{n\to\infty} \mathbb{E}[A_n | X_{n-1} \geq 1] < 1. \tag{7.3}$$

Throughout this chapter we assume that this condition holds.

### 7.3.2 A queueing model with semi-Markovian service times and exceptional first service

To find the probability generating function (PGF) of the queue-length distribution at departure epochs, we use the framework introduced in Chapter 5. There an analysis is presented of the $M^X$/SM2/1 queue, which is a very general type of single-server queueing model with batch arrivals and semi-Markovian service times with exceptional service when the queue is empty, extending

earlier results (cf. also [1, 23, 48, 82]) to make it applicable to situations as the one discussed in the present chapter. To improve the readability of this chapter, we briefly summarize the most important results from Chapter 5 that are also valid for our model.

In Chapter 5, the type of the $n$-th customer is denoted by $J_n \in \{1, 2, \ldots, N\}$. In order to apply the results from Chapter 5 correctly to our model, we need to make one small adjustment. In order to determine the service time of the $n$-th customer, we need to know exactly how much of the critical gap of the $(n-1)$-th customer remains for this $n$-th customer. As it turns out in the next section, this requires the following knowledge about the $(n-1)$-th customer:

(A) we need to know the sampled critical gap $u_{(i,k,r)}$ of the $(n-1)$-th customer;

(B) we need to know whether the $(n-1)$-th customer emptied the queue at the minor road and the $n$-th customer was the first in a new batch arriving some time after the departure of his predecessor.

To start with the latter: a consequence of (B) is that we need a queueing model with so-called *exceptional first service*. As a consequence of (A), we use the triplet $(i, k, r)$, extensively discussed in Section 7.2, to denote the 'customer type', which should be interpreted as a vehicle of profile $r$ that accepted the $i$-th gap, where it sampled $u_{(i,k,r)}$ as its critical headway. It means that if the $n$-th customer is of type $(i, k, r)$, then his *predecessor* (which is the $(n-1)$-th customer) succeeded in his $i$-th attempt, while being of profile $r$ and having critical gap $u_{(i,k,r)}$.

A translation from our model to the model in Chapter 5 can easily be made by mapping our $\bar{N} := NMR$ customer types onto their $N$ customer types (for instance, if $j = (i-1)MR + (k-1)R + r$, then $j \in \{1, 2, \ldots, \bar{N}\}$ can serve as the customer type in Chapter 5).

Mimicking the steps in Section 5.3.1, it is immediate that

$$\begin{aligned}
\mathbb{E}[z^{X_n} 1_{\{J_{n+1}=(j,l,r_1)\}}] &= \frac{1}{z} \sum_{r_0=1}^{R} \sum_{k=1}^{M} \sum_{i=1}^{N} \mathbb{E}[z^{X_{n-1}} 1_{\{J_n=(i,k,r_0)\}}] \\
&\times \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=(j,l,r_1)\}} | J_n = (i,k,r_0), X_{n-1} \geq 1] \\
&+ \frac{1}{z} \sum_{r_0=1}^{R} \sum_{k=1}^{M} \sum_{i=1}^{N} \Big( B(z) \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=(j,l,r_1)\}} | J_n = (i,k,r_0), X_{n-1} = 0]
\end{aligned}$$

$$- \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=(j,l,r_1)\}} | J_n = (i,k,r_0), X_{n-1} \geq 1]\Big)$$
$$\times \mathbb{P}(X_{n-1} = 0, J_n = (i,k,r_0)), \tag{7.4}$$

for $n \in \mathbb{N}$, $j \in \{1,2,\ldots,N\}$, $l \in \{1,2,\ldots,M\}$, and $r_1 \in \{1,2,\ldots,R\}$.

To further evaluate these expressions, we need to introduce some additional notation. Define, for $i,j \in \{1,2,\ldots,N\}$ $k,l \in \{1,2,\ldots,M\}$, $r_0, r_1 \in \{1,2,\ldots,R\}$, and $|z| \leq 1$,

$$A^{(j,l,r_1)}_{(i,k,r_0)}(z) = \lim_{n\to\infty} \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=(j,l,r_1)\}} | J_n = (i,k,r_0), X_{n-1} \geq 1],$$
$$A^{*(j,l,r_1)}_{(i,k,r_0)}(z) = \lim_{n\to\infty} \mathbb{E}[z^{A_n} 1_{\{J_{n+1}=(j,l,r_1)\}} | J_n = (i,k,r_0), X_{n-1} = 0],$$

where $A^*$ corresponds to the number of arrivals during the (exceptional) service of a vehicle arriving when there is no queue in front of him. In addition,

$$f_{(j,l,r_1)}(z) = \lim_{n\to\infty} \mathbb{E}[z^{X_n} 1_{\{J_{n+1}=(j,l,r_1)\}}], \tag{7.5}$$

such that

$$f_{(j,l,r_1)}(0) = \lim_{n\to\infty} \mathbb{P}(X_n = 0, J_{n+1} = (j,l,r_1)). \tag{7.6}$$

The above definitions entail that

$$F(z) = \sum_{r_1=1}^{R} \sum_{l=1}^{M} \sum_{j=1}^{N} f_{(j,l,r_1)}(z). \tag{7.7}$$

In steady state, Equation (7.4) thus leads to the following system of $\bar{N}$ linear equations: for any $j$, $l$, and $r_1$,

$$\left(z - A^{(j,l,r_1)}_{(j,l,r_1)}(z)\right) f_{(j,l,r_1)}(z) - \sum_{\substack{r_0=1,\\ r_0\neq r_1}}^{R} A^{(j,l,r_1)}_{(j,l,r_0)}(z) f_{(j,l,r_0)}(z) \tag{7.8}$$
$$- \sum_{r_0=1}^{R} \sum_{\substack{k=1,\\ k\neq l}}^{M} A^{(j,l,r_1)}_{(j,k,r_0)}(z) f_{(j,k,r_0)}(z) - \sum_{r_0=1}^{R} \sum_{k=1}^{M} \sum_{\substack{i=1,\\ i\neq j}}^{N} A^{(j,l,r_1)}_{(i,k,r_0)}(z) f_{(i,k,r_0)}(z)$$

$$= \sum_{r_0=1}^{R} \sum_{k=1}^{M} \sum_{i=1}^{N} (B(z) A^{*(j,l,r_1)}_{(i,k,r_0)}(z) - A^{(j,l,r_1)}_{(i,k,r_0)}(z)) f_{(i,k,r_0)}(0).$$

From the above we conclude that when we know $A^{(j,l,r_1)}_{(i,k,r_0)}(z)$ and $A^{*(j,l,r_1)}_{(i,k,r_0)}(z)$, the PGFs of our interest can be evaluated. In the next section we show how to find the Laplace-Stieltjes transforms of the conditional service-time distributions, which lead to $A^{(j,l,r_1)}_{(i,k,r_0)}(z)$ and $A^{*(j,l,r_1)}_{(i,k,r_0)}(z)$.

We refer to Section 5.3.1 for more details on how to write the linear system of equations (7.8) in a convenient matrix form and solve it numerically. In Section 5.3.3, and in more detail in Section 4.4, it is also discussed how the PGF of the queue length at *arbitrary epochs* (denoted by $X^{\text{arb}}$) can be found; in particular the clean relation between $X^{\text{arb}}$ and $X$:

$$\mathbb{E}[z^{X^{\text{arb}}}] = \mathbb{E}[z^X] \frac{\mathbb{E}[B](1-z)}{1-B(z)} \tag{7.9}$$

is useful in this context. The interpretation is that the number of customers at an *arbitrary epoch* is equal to the number of customers left behind by an arbitrary *departing* customer, excluding those that arrived in the same batch as this departing customer. We refer to Chapter 4 for more details and a proof of (7.9).

## 7.4 Capacity

In this section we derive the Laplace-Stieltjes transform of the service-time distribution of our gap acceptance model. This service-time analysis serves two purposes. Firstly, it is used to complete the queue-length analysis of the previous section. Secondly, it is an important result by itself, because it is directly linked to the capacity of the intersection, which is defined as the maximum possible number of vehicles per time unit that can pass from the minor road. An alternative but equivalent definition of capacity, also employed by Heidemann and Wegmann [60], is the maximum arrival rate for which the corresponding queue remains stable. Combining (7.1) and (7.3), and denoting

$$g := \lim_{n\to\infty} \mathbb{E}[G^{(n)} \mid X_{n-1} \geq 1],$$

we can rewrite the stability condition as

$$\lambda \mathbb{E}[B]\, g < 1, \tag{7.10}$$

where $\lambda$ is the arrival rate of batches, implying that the arrival rate of individual vehicles is $\lambda\mathbb{E}[B]$. As a consequence, the capacity of the minor road is given by

$$C = \frac{1}{g}. \tag{7.11}$$

Every driver (of profile $r$) samples a random $T_{(i,r)}$ at his $i$-th attempt, irrespective of his previous attempts. However, he uses only a constant $\Delta_r$ of the gap to cross the main road. As a consequence, when accepting the gap, the remaining part $(T_{(i,r)} - \Delta_r)$ can be used by subsequent drivers. In order to keep the analysis tractable, we need the following assumption.

**Assumption 7.1 (Limited gap reusability assumption).** *We assume that* at most one *driver can reuse the remaining part of a critical gap accepted by his predecessor. The means that* $(T_{(i,r)} - \Delta_r)$ *should* not *be large enough for more than one succeeding vehicle to use it for his own critical headway.*

To avoid confusion, we stress that this assumption does *not* mean that a large gap between successive vehicles on the major road cannot be used by more than one vehicle from the minor road. It *does* mean, however, that the difference between *critical* gaps of two successive vehicles cannot be so large that the second vehicle's critical gap is completely contained in its predecessor's critical gap (excluding the merging time).

**Remark 7.2.** *This assumption seems to be realistic in most, but not all, cases, because one can envision for instance situations in which (parts of) a critical gap accepted by, say, a slow truck might be reused by two fast accelerating vehicles following this truck. It is noted, however, that if Assumption 7.1 is violated, despite our method no longer being exact, the computed capacity is still very close to the real, simulated capacity; numerical evidence of this property is provided in Section 7.5. It implies that our method can still be used as a very accurate approximation in those (rare) cases where Assumption 7.1 is violated.*

We recall that there is a complicated dependence structure between service times of two successive vehicles, due to our assumption that a vehicle can use part of the gap left behind by its predecessor *combined* with the assumption that drivers have different profiles and tend to become impatient. This obstacle

can be overcome by analyzing the service time of a vehicle of type $(j, l, r_1)$ in the situation that its predecessor was of type $(i, k, r_0)$.

We proceed by introducing some notation. Let $E_y$ be the event that there is a predecessor gap available of length $y$. Denote by $\pi_{(i,k,r_0)}$ the probability that an arbitrary driver with profile $r_0$ succeeds in his $i$-th attempt with critical headway $u_{(i,k,r_0)}$. In addition, let

$$\tilde{G}_{(j,l,r_1)}(s, y) = \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r_1)\}} \mid E_y\right], \tag{7.12}$$

for $0 \leq y \leq u_{(i,k,r_0)} - \Delta_{r_0}$, $j = 1, 2, \ldots, N$, and $l = 1, 2, \ldots, M$. The LST of an arbitrary service time can be found by conditioning on the type of the current vehicle and its predecessor:

$$\tilde{G}(s) = \mathbb{E}[e^{-sG^{(n)}}] = \sum_{r_0=1}^{R} \sum_{r_1=1}^{R} \sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{l=1}^{M} \sum_{k=1}^{M} \tilde{G}_{(i,k,r_0)}^{(j,l,r_1)}(s) \mathbb{P}(J_n = (i, k, r_0)), \tag{7.13}$$

where

$$\tilde{G}_{(i,k,r_0)}^{(j,l,r_1)}(s) = \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r_1)\}} | J_n = (i, k, r_0)\right], \tag{7.14}$$

and

$$\mathbb{P}(J_n = (i, k, r_0)) = \pi_{(i,k,r_0)} p_{(i,k,r_0)} p_{r_0}. \tag{7.15}$$

We are now ready to state the main result of this chapter. Define

$$\bar{f}_{(i,k,r)} := \frac{f_{(i,k,r)}(0)}{\pi_{(i,k,r)} p_{(i,k,r)} p_r}, \quad \bar{u}_{(i,k,r)} := u_{(i,k,r)} - \Delta_r.$$

**Theorem 7.1.** *The partial service-time LST of a customer, jointly with the event that he is of type $(j, l, r_1)$, given that his predecessor was of type $(i, k, r_0)$, is given by*

$$\begin{aligned} \tilde{G}_{(i,k,r_0)}^{(j,l,r_1)}(s) \; &= \tilde{G}_{(j,l,r_1)}(s, \bar{u}_{(i,k,r_0)}) \left(1 - \bar{f}_{(i,k,r_0)}\right) \\ &\quad + \left( \int_{x=0}^{\bar{u}_{(i,k,r_0)}} \lambda e^{-\lambda x} \tilde{G}_{(j,l,r_1)}(s, \bar{u}_{(i,k,r_0)} - x) \mathrm{d}x \right. \end{aligned}$$

$$+ \tilde{G}_{(j,l,r_1)}(s,0)e^{-\lambda \bar{u}_{(i,k,r_0)}}\Bigg)\bar{f}_{(i,k,r_0)}, \tag{7.16}$$

*where*

- $f_{(i,k,r_0)}(0)$ *is obtained by solving the system of equations* (7.8)*, and*
- *expressions for* $\tilde{G}_{(j,l,r_1)}(s,y)$ *and* $\pi_{(i,k,r_0)}$ *are provided in Lemmas 7.1 and 7.2 below.*

*Proof.* This partial LST $\tilde{G}^{(j,l,r_1)}_{(i,k,r_0)}(s)$ can be expressed in terms of $\tilde{G}_{(j,l,r_1)}(s,y)$ by first distinguishing between the case where there was no queue upon arrival of the $n$-th customer ($X_{n-1} = 0$), and the case where there was a queue ($X_{n-1} \geq 1$). If the customer arrives in an empty system at time $t$ and the previous arrival took place (merged on the major road) at time $t-x$, then we need to check whether the remaining gap $u_{(i,k,r_0)} - \Delta_{r_0} - x = \bar{u}_{(i,k,r_0)} - x$ is greater than zero, or not. If the system was not empty at arrival time, the lag is simply $\bar{u}_{(i,k,r_0)}$ due to Assumption 7.1 (we provide more details on this in Remark 7.3). This, combined with (7.6) and (7.15), yields

$$\begin{aligned}
\tilde{G}^{(j,l,r_1)}_{(i,k,r_0)}(s) &= \lim_{n\to\infty} \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r_1)\}} | X_{n-1} \geq 1, J_n = (i,k,r_0)\right] \\
&\qquad \times \mathbb{P}(X_{n-1} \geq 1 | J_n = (i,k,r_0)) \\
&+ \lim_{n\to\infty} \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r_1)\}} | X_{n-1} = 0, J_n = (i,k,r_0)\right] \\
&\qquad \times \mathbb{P}(X_{n-1} = 0 | J_n = (i,k,r_0)) \\
&= \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r_1)\}} | E_{(u_{(i,k,r_0)}-\Delta_{r_0})}\right]\left(1 - \bar{f}_{(i,k,r_0)}\right) + \\
&\left(\int_{x=0}^{\infty} \lambda e^{-\lambda x} \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r_1)\}} | E_{\max(u_{(i,k,r_0)}-\Delta_{r_0}-x,0)}\right] \mathrm{d}x\right)\bar{f}_{(i,k,r_0)},
\end{aligned}$$

which can easily be rewritten to (7.16). □

It thus remains to find expressions for $\tilde{G}_{(j,l,r_1)}(s,y)$ and $\pi_{(i,k,r_0)}$, which we do in the following two lemmas. We denote by $E_r^{(n)}$ the event that the $n$-th driver does not succeed in his first attempt, given that he has profile $r$.

**Lemma 7.1.** *The conditional service-time LST of a vehicle, given the event $E_y$, is given by*

$$\tilde{G}_{(1,l,r)}(s,y) = p_r p_{(1,l,r)} e^{-q\max\{u_{(1,l,r)}-y,0\}} e^{-s\Delta_r},$$

$$\tilde{G}_{(j,l,r)}(s,y) = p_r p_{(j,l,r)} \left(1 - \mathbb{E}\left[e^{-(s+q)\max\{T_{(1,r)}-y,0\}}\right]\right) \times e^{-(s(y+\Delta_r)+qu_{(j,l,r)})} \left(\frac{q}{s+q}\right)^{j-1} \prod_{m=2}^{j-1} \left(1 - \mathbb{E}\left[e^{-(s+q)T_{(m,r)}}\right]\right),$$

*for $j = 2, 3, \ldots, N$.*

*Proof.* For $j = 1$, the expression $\tilde{G}_{(j,l,r)}(s,y)$ simply follows from computing the probability that a customer is of type $r$ *and* merges successfully during its first attempt with critical gap $u_{(1,l,r)}$, for $l = 1, 2, \ldots, M$, *and* faces a gap that is greater than $u_{(1,l,r)} - y$. In this case, the customer's service time is $\Delta_r$. For the case $j \geq 2$ we find

$$\begin{aligned}
&\tilde{G}_{(j,l,r)}(s,y) \\
=&p_r \sum_{l_1=1}^{M} p_{(1,l_1,r)} 1_{\{u_{(1,l_1,r)}>y\}} \int_0^{u_{(1,l_1,r)}-y} qe^{-qt} e^{-s(y+t)} \\
&\quad \times \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r)\}} | E_r^{(n)}\right] \mathrm{d}t \\
=&p_r \sum_{l_1=1}^{M} p_{(1,l_1,r)} 1_{\{u_{(1,l_1,r)}>y\}} \frac{qe^{-sy}}{s+q} (1 - e^{-(s+q)(u_{(1,l_1,r)}-y)}) \\
&\quad \times \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r)\}} | E_r^{(n)}\right] \\
=&p_r \sum_{l_1=1}^{M} p_{(1,l_1,r)} \frac{qe^{-sy}}{s+q} (1 - e^{-(s+q)\max\{u_{(1,l_1,r)}-y,0\}}) \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r)\}} | E_r^{(n)}\right] \\
=&p_r \frac{qe^{-sy}}{s+q} \left(1 - \mathbb{E}\left[e^{-(s+q)\max\{T_{(1,r)}-y,0\}}\right]\right) \mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r)\}} | E_r^{(n)}\right];
\end{aligned}$$

here it should be kept in mind that

$$\mathbb{E}\left[e^{-sG^{(n)}} 1_{\{J_{n+1}=(j,l,r)\}} | E_r^{(n)}\right]$$

does not depend on $n$ and can be rewritten as

$$\left(\prod_{m=2}^{j-1} \sum_{k=1}^{M} p_{(m,k,r)} \int_0^{u_{(m,k,r)}} qe^{-qt} e^{-st} \mathrm{d}t\right) p_{(j,l,r)} \int_{u_{(j,l,r)}}^{\infty} qe^{-qt} e^{-s\Delta_r} \mathrm{d}t$$

$$= p_{(j,l,r)} e^{-(s\Delta_r + q u_{(j,l,r)})} \left(\frac{q}{s+q}\right)^{j-2} \prod_{m=2}^{j-1} \sum_{k=1}^{M} p_{(m,k,r)} (1 - e^{-(s+q)u_{(m,k,r)}})$$

$$= p_{(j,l,r)} e^{-(s\Delta_r + q u_{(j,l,r)})} \left(\frac{q}{s+q}\right)^{j-2} \prod_{m=2}^{j-1} \left(1 - \sum_{k=1}^{M} p_{(m,k,r)} e^{-(s+q)u_{(m,k,r)}}\right)$$

$$= p_{(j,l,r)} e^{-(s\Delta_r + q u_{(j,l,r)})} \left(\frac{q}{s+q}\right)^{j-2} \prod_{m=2}^{j-1} \left(1 - \mathbb{E}\left[e^{-(s+q)T_{(m,r)}}\right]\right).$$

Slightly rewriting this last expression completes the proof of Lemma 7.1. □

The next lemma shows how to compute the probabilities $\pi_{(i,k,r)}$. We denote by $\tau_q$ the generic interarrival time between two vehicles on the *major* road, which is exponentially distributed with parameter $q$. Let $\hat{v}$ be defined as $\max\{v, 0\}$, and $\check{v}$ as $\max\{-v, 0\}$. In addition, $v^{(1,k,r)}_{(i,l,r_0)} = u_{(1,k,r)} - u_{(i,l,r_0)} + \Delta_{r_0}$.

**Lemma 7.2.** *The probability that a driver of profile $r$ is served in his $i$-th attempt with critical headway $u_{(i,k,r)}$ is given by*

$$\pi_{(i,k,r)} = P_{(i,k,r)} \prod_{m=1}^{i-1} \left(1 - \sum_{k_m=1}^{M} p_{(m,k_m,r)} P_{(m,k_m,r)}\right), \tag{7.17}$$

*where, for $i \in \{2, 3, \ldots, N\}$ and $k \in \{1, 2, \ldots, M\}$,*

$$P_{(i,k,r)} = \mathbb{P}(\tau_q \geq u_{(i,k,r)}) = e^{-q u_{(i,k,r)}}. \tag{7.18}$$

*The remaining $P_{(1,k,r)}$ can be found by solving the system of linear equations*

$$\left(1 + p_{(1,k,r)} \left(-p_r e^{-q\hat{v}^{(1,k,r)}_{(1,k,r)}} + c_{(k,r,r)}\right)\right) P_{(1,k,r)} \tag{7.19}$$

$$+ \sum_{\substack{r_0=1,\\ r_0 \neq r}}^{R} \left(-p_{r_0} e^{-q\hat{v}^{(1,k,r)}_{(1,k,r_0)}} + c_{(k,r_0,r)}\right) p_{(1,k,r_0)} P_{(1,k,r_0)}$$

$$+ \sum_{r_0=1}^{R} \sum_{\substack{l_1=1,\\ l_1 \neq k}}^{M} \left(-p_{r_0} e^{-q\hat{v}^{(1,k,r)}_{(1,l_1,r_0)}} + c_{(k,r_0,r)}\right) p_{(1,l_1,r_0)} P_{(1,l_1,r_0)}$$

$$= \sum_{r_0=1}^{R} \sum_{l=1}^{M} \sum_{i=1}^{N} \left( 1 - e^{-\lambda \check{v}_{(i,l,r_0)}^{(1,k,r)}} - e^{-q\hat{v}_{(i,l,r_0)}^{(1,k,r)}} + \frac{\lambda}{\lambda+q} e^{-(\lambda+q)\check{v}_{(i,l,r_0)}^{(1,k,r)} - qv_{(i,l,r_0)}^{(1,k,r)}} \right.$$
$$\left. + \frac{q}{\lambda+q} e^{-\lambda \bar{u}_{(i,l,r_0)} - qu_{(1,k,r)}} \right) f_{(i,l,r_0)}(0) + \sum_{r_0=1}^{R} c_{(k,r_0,r)},$$

*for* $k \in \{1, 2, \ldots, M\}$ *and* $r \in \{1, 2, \ldots, R\}$*, where*

$$c_{(k,r_0,r)} = \sum_{l=1}^{M} \sum_{i=2}^{N} p_{r_0} p_{(i,l,r_0)} e^{-q(u_{(i,l,r_0)} + \hat{v}_{(i,l,r_0)}^{(1,k,r)})} \prod_{m=2}^{i-1} \left( 1 - \mathbb{E}[e^{-qT_{(m,r_0)}}] \right).$$

*Proof.* Expression (7.17) simply follows from the definition of $\pi_{(i,k,r)}$, by multiplying the probabilities that the driver does *not* succeed in attempts $1, 2, \ldots, i-1$ (each time distinguishing between all $M$ possible random critical gap values) and finally multiplying with the probability that he succeeds in the $i$-th attempt with critical headway $u_{(i,k,r)}$.

For $i \in \{2, 3, \ldots, N\}$ it is easy to determine $P_{(i,k,r)}$ because we do not have to take into account any gap left by the predecessor; we simply compute the probability that the critical gap $u_{(i,k,r)}$ is smaller than the remaining interarrival time $\tau_q$ (which, due to the memoryless property, is again exponentially distributed with rate $q$).

The case $i = 1$ is considerably more complicated. Using (7.15) and noting that $\pi_{(1,k,r)} = P_{(1,k,r)}$, we can find a system of equations for $P_{(1,k,r)}$ by conditioning on the type of the *predecessor* of the vehicle under consideration. Define

$$\chi_{(i,l,r_0)}^{(j,k,r)} := \mathbb{P}(J_n = (j,k,r) | J_{n-1} = (i,l,r_0)),$$

and

$$\phi_{(i,l,r_0)}^{(j,k,r)} := \mathbb{P}(J_n = (j,k,r) | J_{n-1} = (i,l,r_0), X_{n-2} = 0), \tag{7.20}$$

$$\psi_{(i,l,r_0)}^{(j,k,r)} := \mathbb{P}(J_n = (j,k,r) | J_{n-1} = (i,l,r_0), X_{n-2} \geq 1). \tag{7.21}$$

Evidently,

$$P_{(1,k,r)} = \frac{1}{p_{(1,k,r)} p_r} \mathbb{P}(J_n = (1,k,r)) \tag{7.22}$$
$$= \frac{1}{p_{(1,k,r)} p_r} \sum_{r_0=1}^{R} \sum_{l=1}^{M} \sum_{i=1}^{N} \chi_{(i,l,r_0)}^{(1,k,r)} \mathbb{P}(J_{n-1} = (i,l,r_0))$$

$$
\begin{aligned}
&= \frac{1}{p_{(1,k,r)}p_r} \sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N} \chi^{(1,k,r)}_{(i,l,r_0)}\, \pi_{(i,l,r_0)} p_{(i,l,r_0)} p_{r_0} \\
&= \frac{1}{p_{(1,k,r)}p_r} \sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N} \Big( \phi^{(1,k,r)}_{(i,l,r_0)}\, \mathbb{P}(X_{n-2}=0 | J_{n-1}=(i,l,r_0)) \\
&\quad + \psi^{(1,k,r)}_{(i,l,r_0)}\, \mathbb{P}(X_{n-2}\geq 1 | J_{n-1}=(i,l,r_0)) \Big) \pi_{(i,l,r_0)} p_{(i,l,r_0)} p_{r_0},
\end{aligned}
$$

which can be further evaluated, using notation introduced before, as

$$
\frac{1}{p_{(1,k,r)}p_r} \sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N} \left( \phi^{(1,k,r)}_{(i,l,r_0)}\, \bar{f}_{(i,l,r_0)} + \psi^{(1,k,r)}_{(i,l,r_0)} \left(1-\bar{f}_{(i,l,r_0)}\right) \right) \bar{\pi}_{(i,l,r_0)},
$$

with $\bar{\pi}_{(i,l,r_0)} := \pi_{(i,l,r_0)} p_{(i,l,r_0)} p_{r_0}$. We now consider the individual terms separately. Observe that, by conditioning on the value of $\tau_q$,

$$
\begin{aligned}
&\frac{1}{p_{(1,k,r)}p_r} \phi^{(1,k,r)}_{(i,l,r_0)}\, \bar{f}_{(i,l,r_0)} \bar{\pi}_{(i,l,r_0)} \\
&= \Bigg( \int_0^{\max\{\bar{u}_{(i,l,r_0)}-u_{(1,k,r)},0\}} \lambda e^{-\lambda x} \mathrm{d}x \\
&+ \int_{\max\{\bar{u}_{(i,l,r_0)}-u_{(1,k,r)},0\}}^{\bar{u}_{(i,l,r_0)}} \mathbb{P}(\tau_q \geq u_{(1,k,r)} - \bar{u}_{(i,l,r_0)} + x) \lambda e^{-\lambda x} \mathrm{d}x \\
&+ \int_{\bar{u}_{(i,l,r_0)}}^{\infty} \mathbb{P}(\tau_q \geq u_{(1,k,r)}) \lambda e^{-\lambda x} \mathrm{d}x \Bigg) f_{(i,l,r_0)}(0).
\end{aligned}
$$

Also,

$$
\begin{aligned}
&\frac{1}{p_{(1,k,r)}p_r} \psi^{(1,k,r)}_{(i,l,r_0)} \left(1-\bar{f}_{(i,l,r_0)}\right) \bar{\pi}_{(i,l,r_0)} \\
&= \mathbb{P}\left(\tau_q \geq \max\{u_{(1,k,r)} - \bar{u}_{(i,l,r_0)}, 0\}\right) \left( \pi_{(i,l,r_0)} p_{(i,l,r_0)} p_{r_0} - f_{(i,l,r_0)}(0) \right).
\end{aligned}
$$

Combining the above, it follows that $P_{(1,k,r)}$ equals, with $v^{(1,k,r)}_{(i,l,r_0)}$ as defined above,

$$
\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N} \Bigg( 1 - e^{-\lambda \check{v}^{(1,k,r)}_{(i,l,r_0)}} - e^{-q\hat{v}^{(1,k,r)}_{(i,l,r_0)}} + \frac{\lambda}{\lambda+q} e^{-(\lambda+q)\check{v}^{(1,k,r)}_{(i,l,r_0)} - q v^{(1,k,r)}_{(i,l,r_0)}}
$$

$$+\frac{q}{\lambda+q}e^{-\lambda(u_{(i,l,r_0)}-\Delta_{r_0})-qu_{(1,k,r)}}\Bigg)f_{(i,l,r_0)}(0)$$
$$+\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N}e^{-q\hat{v}_{(i,l,r_0)}^{(1,k,r)}}\pi_{(i,l,r_0)}p_{(i,l,r_0)}p_{r_0}.$$

These expressions can be written in a more convenient form. To this end, it is noted that

$$\begin{aligned}
&\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N}e^{-q\hat{v}_{(i,l,r_0)}^{(1,k,r)}}\pi_{(i,l,r_0)}p_{(i,l,r_0)}p_{r_0}\\
&=\sum_{r_0=1}^{R}\sum_{l=1}^{M}e^{-q\hat{v}_{(1,l,r_0)}^{(1,k,r)}}P_{(1,l,r_0)}p_{(1,l,r_0)}p_{r_0}+\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=2}^{N}e^{-q\hat{v}_{(i,l,r_0)}^{(1,k,r)}}\pi_{(i,l,r_0)}p_{(i,l,r_0)}p_{r_0}\\
&=\sum_{r_0=1}^{R}\sum_{l=1}^{M}e^{-q\hat{v}_{(1,l,r_0)}^{(1,k,r)}}P_{(1,l,r_0)}p_{(1,l,r_0)}p_{r_0}\\
&+\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=2}^{N}e^{-q\hat{v}_{(i,l,r_0)}^{(1,k,r)}}p_{(i,l,r_0)}p_{r_0}P_{(i,l,r_0)}\prod_{m=1}^{i-1}\left(1-\sum_{l_m=1}^{M}p_{(m,l_m,r_0)}P_{(m,l_m,r_0)}\right).
\end{aligned}$$

After some basic algebraic manipulations, and using (7.18) we obtain

$$\begin{aligned}
&\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=1}^{N}e^{-q\hat{v}_{(i,l,r_0)}^{(1,k,r)}}\pi_{(i,l,r_0)}p_{(i,l,r_0)}p_{r_0}=\sum_{r_0=1}^{R}\sum_{l_1=1}^{M}\Bigg(e^{-q\hat{v}_{(1,l_1,r_0)}^{(1,k,r)}} \qquad (7.23)\\
&-\sum_{l=1}^{M}\sum_{i=2}^{N}p_{(i,l,r_0)}e^{-q(u_{(i,l,r_0)}+\hat{v}_{(i,l,r_0)}^{(1,k,r)})}\prod_{m=2}^{i-1}\left(1-\sum_{l_m=1}^{M}p_{(m,l_m,r_0)}e^{-qu_{(m,l_m,r_0)}}\right)\Bigg)\\
&\times P_{(1,l_1,r_0)}p_{(1,l_1,r_0)}p_{r_0}+\sum_{r_0=1}^{R}\sum_{l=1}^{M}\sum_{i=2}^{N}p_{r_0}p_{(i,l,r_0)}e^{-q(u_{(i,l,r_0)}+\hat{v}_{(i,l,r_0)}^{(1,k,r)})}\\
&\times\prod_{m=2}^{i-1}\left(1-\sum_{l_m=1}^{M}p_{(m,l_m,r_0)}e^{-qu_{(m,l_m,r_0)}}\right).
\end{aligned}$$

Using Equation (7.23) in (7.22), and after simplification, we obtain the linear system of equations (7.19), which proves the lemma. □

The expressions for $A^{(j,l,r_1)}_{(i,k,r_0)}(z)$ and $A^{*(j,l,r_1)}_{(i,k,r_0)}(z)$, required for the queue-length analysis in Section 7.3, follow from substituting $s = \lambda(1 - B(z))$ in the two main terms in Equation (7.16):

$$A^{(j,l,r_1)}_{(i,k,r_0)}(z) = \tilde{G}_{(j,l,r_1)}(\lambda(1 - B(z)), \bar{u}_{(i,k,r_0)}), \tag{7.24}$$

$$\begin{aligned} A^{*(j,l,r_1)}_{(i,k,r_0)}(z) = &\int_{x=0}^{\bar{u}_{(i,k,r_0)}} \lambda e^{-\lambda x} \tilde{G}_{(j,l,r_1)}(\lambda(1 - B(z)), \bar{u}_{(i,k,r_0)} - x)\mathrm{d}x \\ &+ \tilde{G}_{(j,l,r_1)}(\lambda(1 - B(z)), 0)e^{-\lambda \bar{u}_{(i,k,r_0)}}. \end{aligned} \tag{7.25}$$

In order to determine the capacity from (7.11), we need to obtain $g$, which is the mean service time for *queuers*. This can be established by substituting $\bar{f}_{(i,k,r)}(0) = 0$ and $f_{(i,k,r)}(0) = 0$, respectively, in Theorem 7.1 and Lemma 7.2, and then use these expressions in Equation (7.13) to obtain the LST of the conditional service time distribution. This will be further used in Equation (7.11) to derive the capacity by differentiating at $s = 0$.

**Remark 7.3.** *We now discuss in more detail Assumption 7.1, and how it plays a role in our result. The assumption entails that the gap $T_{(i,r_0)} - \Delta_{r_0}$ left by a driver of profile $r_0$, while being successful in the $i$-th attempt, can be used by the subsequent driver only. If we would allow more than one following vehicle to use a gap left behind by a merging vehicle, the analysis becomes much more complicated due to the fact that we need to distinguish between all possible cases where multiple drivers fit into this lag.*

*For example, in the proof of Theorem 7.1 we use explicitly that, if the system was not empty at arrival time, the gap left behind by a predecessor of type $(j, l, r_1)$ is simply $\bar{u}_{(j,l,r_1)} = u_{(j,l,r_1)} - \Delta_{r_1}$. However, a situation where this is not necessarily true, occurs when $j = 1$ (meaning that the predecessor succeeded at his first attempt) and required a critical gap of $u_{(1,l,r_1)}$ that is* smaller *than the remaining gap of the* predecessor's predecessor*, say $\bar{u}_{(i,k,r_0)}$. In this case, the gap left behind by the predecessor is equal to*

$$u_{(i,k,r_0)} - \Delta_{r_0} - \Delta_{r_1} > u_{(1,l,r_1)} - \Delta_{r_1},$$

*meaning that the current vehicle and its predecessor both used the same gap $u_{(i,k,r_0)}$. In this sense, Assumption 7.1 ensures the tractability of the model.*

*Assumption 7.1 has not been stated in terms of the model input parameters, and is therefore not straightforward to check. To remedy this, we give an equivalent definition in terms of the critical headway and the merging time. Assumption*

*7.1 is satisfied if and only if*

$$u_{(1,l,r_1)} \geq \bar{u}_{(i,k,r_0)} = u_{(i,k,r_0)} - \Delta_{r_0}, \tag{7.26}$$

*for all* $i \in \{1, 2, \ldots, N\}$, $k, l \in \{1, 2, \ldots, M\}$, *and* $r_0, r_1 \in \{1, 2, \ldots, R\}$. *In the next section, we will numerically show that in cases that Assumption 7.1 is not fulfilled, the estimated capacity of the minor road is slightly smaller, but still highly accurate. The reason why the estimated capacity is a* lower *bound for the true capacity, is the fact that we waste some capacity by not allowing more than one vehicle to use the remaining part of a critical gap. The astute reader might have noticed that the maximum operators in Lemma 7.1 are not really needed due to condition* (7.26)*. It turns out, however, that by preventing the corresponding terms from becoming negative, the capacity is approximated much better even when the condition is violated.*

## 7.5 Numerical results

The analysis from the previous sections facilitates the evaluation of the performance of the system, including the assessment of the sensitivity of the capacity when varying model parameters. In this section, we present numerical examples to demonstrate the impact of these model parameters and of different driver behavior.

### 7.5.1 Example 1

In this illustrative example, we distinguish between two driver profiles. Profile 1 represents 'standard' traffic, whereas profile 2 can be considered as 'slower' traffic (for example large, heavily loaded vehicles). We assume that the ratio between profile 1 and profile 2 vehicles is $90\%/10\%$. The fact that profile 1 vehicles are faster than profile 2 vehicles is captured in their merging times (respectively $\Delta_1$ and $\Delta_2$) and in the length of their critical gaps. From the profile 1 drivers, we assume that 40% need a gap of at least 5 seconds, upon arrival at the intersection, while the remaining 60% need a gap of at least 6 seconds. If, however, the drivers do not find an acceptable gap right away, they will grow impatient and be more and more prepared to accept slightly smaller gaps. We introduce an impatience rate $\alpha \in (0, 1)$ that determines how fast critical gaps decrease. Profile 2 vehicles are assumed to be slower, meaning

that they need critical gaps of respectively 8 seconds (50%) or 9 seconds (50%) at their first attempt. Summarizing, we have the following model parameters in this example:

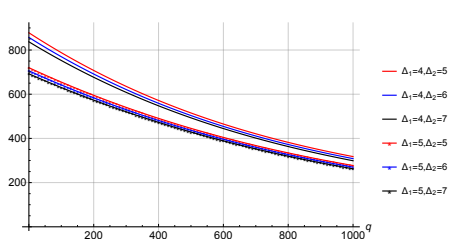
| Profile 1 | Profile 2 |
|---|---|
| $p_1 = 0.9$ | $p_2 = 0.1$ |
| $p_{i,1,1} = 0.4$ | $p_{i,1,2} = 0.5$ |
| $p_{i,2,1} = 0.6$ | $p_{i,2,2} = 0.5$ |
| $u_{1,1,1} = 5.0$ | $u_{1,1,2} = 8.0$ |
| $u_{1,2,1} = 6.0$ | $u_{1,2,2} = 9.0$ |

and, due to the impatience, we have for both profiles:

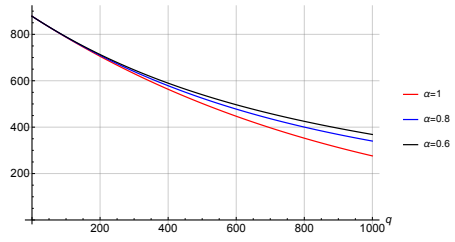
$$u_{(i+1,k,r)} = \alpha(u_{(i,k,r)} - \Delta_r) + \Delta_r, \quad i = 1, 2, \ldots, N-1,\ k = 1, 2,\ r = 1, 2. \tag{7.27}$$

We vary $\alpha$, $\Delta_1$ and $\Delta_2$ to gain insight into the impact of these parameters on the capacity of the minor road, while also varying $q$, the arrival rate on the major road. We take $\alpha \in \{0.6, 0.8, 0.9, 1.0\}$, $\Delta_1 \in \{4, 5\}$ and $\Delta_2 \in \{5, 6, 7\}$ and vary $q$ between 0 and 1000 vehicles per hour. Note that $\alpha = 1$ corresponds to a model *without* impatience. The results are depicted in Figures 7.1(a) and 7.1(b). In Figure 7.1(a), we can observe how the capacity of the minor road increases when merging time corresponding to at least one of the driver profiles decreases. From Figure 7.1(b) we conclude by how much the capacity of the minor road increases when drivers become more impatient. The biggest capacity gain is caused by a decrease in $\Delta_1$ from 5 to 4, obviously because profile 1 constitutes the vast majority of all vehicles.

So far we have only studied the impact of the model parameters on the capacity of the minor road. In practice, queue length and delay distributions are other important performance measures. In this numerical example we compute the distribution of the minor road queue length and show how it depends on the batch size distribution. In our current example we fix $q = 200$, $\alpha = 0.7$, $N = 10$, and we take an arrival rate of 300 vehicles per hour, arriving in batches of two vehicles on average. We compare the case where every batch consists of exactly two vehicles with random batch sizes where one, two, or three vehicles arrive simultaneously, each with probability $1/3$. The results are shown in Table 7.1. We observe that the mean queue length depends on the




(a) Impact of merging times of drivers for $\alpha = 0.9$.




(b) Impact of impatient behaviors of drivers for $\Delta_1 = 4$ seconds and $\Delta_2 = 5$ seconds

**Figure 7.1.** Capacity of the minor road (veh/h) as a function of the flow rate on the major road (veh/h) in Example 1.

*distribution* of the batch sizes, whereas the capacity only depends on the *mean* batch size. In this experiment, the differences are relatively small, but still the probability of having a queue length of more than five vehicles is almost twice as large (0.029) for random batch sizes than for fixed batch sizes (0.017). For other instances the difference between the the performance metrics in the two columns can be substantially more pronounced.

| | Fixed batch size | Random batch size |
|---|---|---|
| $g$ (in seconds) | 5.061 | 5.061 |
| $C$ (veh/h) | 711.331 | 711.331 |
| $\mathbb{E}[X^{\text{arb}}]$ | 0.977 | 1.094 |
| $\text{Var}[X^{\text{arb}}]$ | 2.188 | 2.921 |
| $\mathbb{E}[X]$ | 1.478 | 1.763 |
| $\text{Var}[X]$ | 2.443 | 3.508 |
| $\mathbb{P}(X^{\text{arb}} = 0)$ | 0.576 | 0.577 |
| $\mathbb{P}(X^{\text{arb}} > 5)$ | 0.017 | 0.029 |

**Table 7.1.** Numerical results for Example 1.

**Remark 7.4.** *Obtaining the numerical results for the queue lengths turned out to be significantly harder than computing only the service-time related performance measures (such as the capacity). We had to apply several numerical enhancements to the standard procedures. The most efficient way (that we found)*

*to solve the system of equations* (7.8) *was taking Taylor series approximations of the* $A^{(j,l,r_1)}_{(i,k,r_0)}(z)$ *and* $A^{*(j,l,r_1)}_{(i,k,r_0)}(z)$*, and using a mix of numerical procedures and symbolic computations.*

### 7.5.2 Example 2

The model parameters in Example 1 are such that condition (7.26) is met. In practice, however, situations might arise where this is not the case. The purpose of this example is to quantify errors made when computing the capacity using the methods proposed in Section 7.4, since these methods rely on condition (7.26). We take the same settings as in Example 1, but we adjust the critical gaps for profile 2 drivers from 8 and 9 to, respectively, 10 and 12 seconds. We keep the original merging times for both profile drivers of $\Delta_1 = 4$ seconds and $\Delta_2 = 5$ seconds. In order to quantify the errors, we use both the analysis from Section 7.4 and a computer program that simulates the model and gives accurate results. Table 7.2 compares the capacities obtained with both methods, for various values of $q$ and $\alpha$. It immediately becomes apparent that the relative error is below $0.5$ percent in all cases. We conducted similar experiments for different settings and in all cases we obtained relative errors below one percent, which clearly justifies using the analysis of this chapter in all practical situations, including those violating (7.26). As expected (see Remark 7.3), the approximated capacities are (slightly) underestimating the actual capacities.

| | | $q = 250$ | $q = 500$ | $q = 750$ | $q = 1000$ |
|---|---|---|---|---|---|
| $\alpha = 1.0$ | Approximation | 646.2 | 466.4 | 328.9 | 225.8 |
| | Simulation | 647.2 | 467.7 | 330.0 | 226.5 |
| $\alpha = 0.9$ | Approximation | 652.8 | 491.0 | 377.8 | 298.9 |
| | Simulation | 653.7 | 491.5 | 378.0 | 299.0 |

**Table 7.2.** Capacity of the minor road (veh/h) for various values of the flow rate, $q$, on the major road (veh/h) in Example 2.

### 7.5.3 Example 3

We demonstrate our methodology in an example with realistic parameters based on the empirical data study by Tupper [110], who recorded data on more than ten thousand critical gaps in Massachusetts and Oregon. In this study, it is argued that the standard correction factors suggested in the Highway Capacity Manual (HCM [108]) do not always adequately describe the true impact of various factors on the critical gap. In [110] the factors that influence this critical gap are studied; in this example we focus on vehicle types (car, van, SUV, or Truck) and the age of the driver (teen, adult, or elderly). We base the critical gaps in this example on [110, Chapter 6]; for ease we consider the case of no interaction between the driver type and vehicle type (but evidently any type of interaction could be included). In our experiment we consider a situation in which 10% of the drivers are teens, 70% are adults, and 20% are elderly people; in addition, 70% of the vehicles are cars, while vans, SUVs and trucks each constitute 10% of the total traffic. As a result, we have $R = 3 \times 4 = 12$ driver profiles. We matched the averaged values as much as possible with data from [110].

| | Vehicle type | | | | |
|---|---|---|---|---|---|
| | Car | Van | SUV | Truck | Average |
| Teen | 5.125 | 5.875 | 4.375 | 3.875 | 5.00 |
| Adult | 6.625 | 7.375 | 5.875 | 5.375 | 6.50 |
| Elderly | 7.375 | 8.125 | 6.625 | 6.125 | 7.25 |
| Average | 6.625 | 7.375 | 5.875 | 5.375 | 6.5 |

**Table 7.3.** Input data for Example 3: mean critical headways (in seconds) for the first attempt per combination of vehicle type and gender and their (weighted) averages.

We now explain how we incorporate impatience and random variations in the model. Denote the mean critical headways for the first attempt (see Table 7.3) by $u_r$, $r = 1, 2, \ldots, R$. Due to impatience, the mean critical headway should decrease from 6.5 seconds (first attempt) to 5.5 seconds (second attempt), 5.25 seconds (third attempt), and 5 seconds (all subsequent attempts). For this purpose, we define an impatience effect $\beta_1 = 0$, $\beta_2 = 6.5 - 5.5 = 1$, $\beta_2 = 6.5 - 5.25 = 1.25$, and $\beta_n = 6.5 - 5 = 1.5$ for $n = 3, 4, \ldots, N$. We take

the maximum number of attempts $N = 100$, which is more than sufficient to obtain accurate values for the capacities. Finally, we incorporate randomness by adding $-1$, $0$, or $1$, each with probability $1/3$, unless it would result in a critical headway less than 2.5 seconds, in which case we take 2.5. These effects combined result in the following random critical gaps:

$$T_{(i,r)} = \max(2.5, u_r - \beta_i + \epsilon),$$

with

$$\epsilon = \begin{cases} -1 & \text{w.p. } 1/3, \\ 0 & \text{w.p. } 1/3, \\ 1 & \text{w.p. } 1/3, \end{cases}$$

for $i = 1, \ldots, N$ and $r = 1, \ldots, R$. The values for $\Delta_r$ are chosen to be equal to the lowest possible critical headway that can be realized per driver profile:

$$\Delta_r = \min_{i,k} u_{(i,k,r)} = \max(2.5, u_r - \beta_N - 1) = \max(2.5, u_r - 2.5),$$

which varies from 2.5 to 5.625 seconds. The capacities for this example are given in Table 7.4. We have used two methods to compute these capacities:

(1) The full model discussed in this paper, with profiles, impatience and randomness;

(2) An aggregated model where critical headways are averaged over all profiles.

The reason to compare these two models is to check whether one could also determine the capacity by just averaging over all driver profiles and use simpler models; a procedure that was shown in the HCM to perform well in certain circumstances. For a fair comparison, we have taken $\Delta = \Delta_1 = 4.0175$ seconds in model 2, which is equal to the value found in the full model. The results in Table 7.4 indicate that the aggregated model overestimate the capacity. For small values of $q$ the approximation error decreases and even disappears in the limit $q \downarrow 0$, where the capacity approaches $3600/\Delta = 896.1$ vehicles per hour, as we have witnessed in Example 1. We conclude that for these parameter values, indeed, the simpler model is a reasonable approximation for the full model. This may be true for these parameters, but not in general, and therefore (if the increased computational complexity is no issue) we always recommend using the full model.

| | $q$ | | | |
|---|---|---|---|---|
| | 0 | 500 | 1000 | 1500 |
| Original model | 896.1 | 508.6 | 318.1 | 204.6 |
| No profiles | 896.1 | 514.0 | 326.6 | 215.2 |

**Table 7.4.** Capacities (veh/h) for Example 3.

## 7.6 Discussion and conclusion

In this chapter, we have presented a gap acceptance model for unsignalized intersections that considerably generalizes the existing models. The model proposed incorporates various realistic aspects that were not taken into consideration in previous studies: driver impatience, heterogeneous driving behavior, and the service time being dependent on the vehicle arriving at an empty queue or not. Despite the rather intricate system dynamics, we succeed in providing explicit expressions for the stationary queue-length distribution of vehicles on the minor road, which facilitate the evaluation of the corresponding capacity (of the minor road, that is).

We have concluded the chapter by presenting a series of numerical results, which are representative of the extensive experiments that we performed. Our techniques facilitate the quantitative evaluation of the system at hand, including the assessment of the sensitivity of the capacity when varying the model parameters (which is typically considerably harder when relying on simulation).

Chapter 8

# Generalized gap acceptance models with Markov platooning

The model in the present chapter is an extension of the model considered in Chapter 7. We now allow Markov platooning on the main road, described in Chapter 3 and Section 5.5, to study the impact of various platoon formations (on the main road), on the capacity of the minor road. As discussed in Section 7.4, to find the capacity of the minor road, it is required to determine the LST of the service time of an arbitrary low-priority driver.

## 8.1 Model description

We consider an unsignalized priority-controlled intersection, described in Section 1.2, where the vehicle drivers on the main road have priority over the vehicle drivers on the minor road. The low-priority vehicle drivers, on

the minor road, cross the intersection as soon as they come across a gap with duration larger than $T$ between two subsequent high-priority vehicles, which is referred to as the *critical headway* or *critical gap*. As discussed in Section 7.2 of the previous chapter, this critical gap depends on the profile of the driver as well as the attempt in which he makes an effort to enter the major road. Assume that the profile of an arriving low-priority driver is $r$ with probability $p_r$ for $r = 1, 2, \ldots, R$. Every driver of profile $r$ needs a *constant* critical gap $T_{(i,r)}$ at its $i$-th attempt, but he uses only a constant $\Delta_r$ of that gap to enter (or merge on) the major road. As a consequence, he leaves $(T_{(i,r)} - \Delta_r)$, while entering the $i$-th attempt on the major road, which can be used by the subsequent low-priority drivers for their very first attempts. In order to keep the analysis tractable, we assume that at most one driver can reuse the remaining part of a gap accepted by its predecessor. This means that $(T_{(i,r)} - \Delta_r)$ should not be large enough for more than one succeeding vehicle to use it for its own critical headway (see also Assumption 7.1).

On the minor road, vehicles arrive in batches according to a Poisson process with rate $\lambda$ and the batch size is denoted by the random variable $B$ with generating function $B(z)$, for $|z| \leq 1$. As considered in Section 5.5, the arrival process on the major road is a Markov modulated Poisson process (MMPP) such that, for $i = 1, 2, \ldots, M$, $q_i$ is the Poisson rate when the continuous time Markov process (so-called background process), $J(t)$, is in phase $i$. Therefore, the transition probabilities of the background process of the MMPP are given by

$$P_{ij}(T) = \mathbb{P}(J(T) = j | J(0) = i) = [e^{TQ}]_{ij}, \quad \text{for } i, j = 1, 2, \ldots, M,$$

with the transition rate matrix

$$Q = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1M} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{M1} & \mu_{M2} & \cdots & \mu_{MM} \end{bmatrix},$$

where $-\mu_{ii} = \mu_i = \sum_{j \neq i} \mu_{ij}$.

As discussed in Section 5.5, in order to determine the capacity of the minor road in the next section, this requires to find the following two expressions:

$$\phi_{ij}(t) = \mathbb{P}(\text{No car on the major road in [0,t] and } J(t) = j | J(0) = i), \quad (8.1)$$

$$\psi_{ij}(t) = \frac{d}{dt}\mathbb{P}(T_{\text{next car}} \leq t, J(T_{\text{next car}}) = j | J(0) = i), \tag{8.2}$$

where $T_{\text{next car}}$ is the time when next car passes on the major road. Both these expressions can be obtained from Equations (5.80) and (5.83) respectively.

## 8.2 Capacity

The main objective of this section is to determine the LST of the service-time distribution of the gap acceptance model considered in the previous section, which can be further used to derive the capacity of the minor road.

Let $G^{(n)}$ be the service time of the $n$-th low-priority vehicle. And let $X_n$ be the number of vehicles on the minor road immediately after the departure of an arbitrary low-priority vehicle. As discussed in Chapter 7, the capacity of the minor road is given from Equation (7.11) as

$$C = \frac{1}{g}, \tag{8.3}$$

where $g = \lim_{n\to\infty} \mathbb{E}[G^{(n)} \mid X_{n-1} \geq 1]$.

In order to determine the service time of the $n$-th low-priority driver, we need to know exactly how much of the critical gap of the $(n-1)$-th driver remains for this $n$-th driver. Let $J_n$ be an ordered triple $(i, k, r)$, associated with the $(n-1)$-th low priority driver, where $i$ is the 'succeeded attempt number', $k$ indicates the phase on the major road after accepting the critical gap at the $i$-th attempt, and $r$ denotes the 'profile' of the driver, for $i = 1, 2, \ldots, N, k = 1, 2, \ldots, M$ and $r = 1, 2, \ldots, R$. It can be noted that $(n-1)$-th driver of profile $r$ left $(T_{(i,r)} - \Delta_r)$ gap for the $n$-th driver while entering the $i$-th attempt on the major road. And the $n$-th driver (of profile $r_1$), who arrived in the nonempty system, can start scanning the remaining gap $(T_{(1,r_1)} - T_{(i,r)} + \Delta_r)$ in the phase $i$ on the major road for its very first attempt. It turns out that this requires determining $f_{(i,k,r)}(0) = \lim_{n\to\infty} \mathbb{P}(X_n = 0, J_{n+1} = (i, k, r))$, which can be obtained by solving the system of equations (7.8), with the conditional service-time distributions of the gap acceptance model considered in this chapter.

The LST of an arbitrary service time is obtained by conditioning on the type of a current low-priority vehicle and its predecessor:

$$\tilde{G}(s) = \mathbb{E}[e^{-sG^{(n)}}] = \sum_{i_1=1}^{N}\sum_{k_1=1}^{M}\sum_{r_1=1}^{R}\sum_{i_0=1}^{N}\sum_{k_0=1}^{M}\sum_{r_0=1}^{R}\tilde{G}_{(i_0,k_0,r_0)}^{(i_1,k_1,r_1)}(s)\pi_{(i_0,k_0,r_0)}, \tag{8.4}$$

where

$$\tilde{G}_{(i_0,k_0,r_0)}^{(i_1,k_1,r_1)}(s) = \mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|J_n=(i_0,k_0,r_0)\right], \tag{8.5}$$

and

$$\pi_{(i_0,k_0,r_0)} = \mathbb{P}(J_n=(i_0,k_0,r_0)). \tag{8.6}$$

We denote by $E_{(l,r)}^{(n)}$ the event that the $n$-th driver does not succeed in his first attempt, and the phase on the major road is $l$ at the end of the first attempt, given that he has profile $r$. We further denote $\hat{E}_{(k,y)}$ as the event that the phase on the major road is $k$ at the start of the current low-priority driver's first attempt, and there is a gap available of length $y$ after the departure of his predecessor. Furthermore, let

$$\tilde{G}_{(i_1,k_1,r_1)}(s,y,k_0) = \mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|\hat{E}_{(k_0,y)}\right], \tag{8.7}$$

for $0 \le y \le T_{(i_0,r_0)} - \Delta_{r_0}$, $i_0, i_1 \in \{1,2,\ldots,N\}$, $k_0, k_1 \in \{1,2,\ldots,M\}$, and $r_0, r_1 \in \{1,2,\ldots,R\}$.

Now, we can obtain $\tilde{G}_{(i_0,k_0,r_0)}^{(i_1,k_1,r_1)}(s)$ as

$$\begin{aligned}
&\tilde{G}_{(i_0,k_0,r_0)}^{(i_1,k_1,r_1)}(s) = \lim_{n\to\infty}\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|J_n=(i_0,k_0,r_0), X_{n-1}\ge 1\right]\\
&\times \mathbb{P}(X_{n-1}\ge 1|J_n=(i_0,k_0,r_0))\\
&+ \lim_{n\to\infty}\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|J_n=(i_0,k_0,r_0), X_{n-1}=0\right]\\
&\times \mathbb{P}(X_{n-1}=0|J_n=(i_0,k_0,r_0))\\
=&\tilde{G}_{(i_1,k_1,r_1)}(s,T_{(i_0,r_0)}-\Delta_{r_0},k_0)\left(1-\frac{f_{(i_0,k_0,r_0)}}{\pi_{(i_0,k_0,r_0)}}\right)\\
&+\left(\int_{t=0}^{T_{(i_0,r_0)}-\Delta_{r_0}}\lambda e^{-\lambda t}\tilde{G}_{(i_1,k_1,r_1)}(s,T_{(i_0,r_0)}-\Delta_{r_0}-t,k_0)\mathrm{d}t\right.
\end{aligned}$$

$$+\int_{t=T_{(i_0,r_0)}-\Delta_{r_0}}^{\infty}\lambda e^{-\lambda t}\sum_{k=1}^{M}P_{k_0k}(t-T_{(i_0,r_0)}+\Delta_{r_0})$$
$$\times\tilde{G}_{(i_1,k_1,r_1)}(s,0,k)\mathrm{d}t\Bigg)\frac{f_{(i_0,k_0,r_0)}}{\pi_{(i_0,k_0,r_0)}}, \tag{8.8}$$

where $\tilde{G}_{(i_1,k_1,r_1)}(s,y,k_0)$ is given by

$$\tilde{G}_{(i_1,k_1,r_1)}(s,y,k_0)=p_{r_1}\Bigg(1_{\{i_1=1\}}e^{-s\Delta_{r_1}}\phi_{k_0k_1}(T_{(1,r_1)}-y)$$
$$+1_{\{i_1\geq 2\}}\int_{t=0}^{T_{(1,r_1)}-y}\sum_{l_2=1}^{M}\psi_{k_0l_2}(t)e^{-s(t+y)}\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|E^{(n)}_{(l_2,r_1)}\right]\mathrm{d}t\Bigg)$$
$$=p_{r_1}\Bigg(1_{\{i_1=1\}}e^{-s\Delta_{r_1}}\phi_{k_0k_1}(T_{(1,r_1)}-y)+1_{\{i_1\geq 2\}}e^{-sy}$$
$$\times\sum_{l_2=1}^{M}\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|E^{(n)}_{(l_2,r_1)}\right]\left(\int_{t=0}^{T_{(1,r_1)}-y}\psi_{k_0l_2}(t)e^{-st}\mathrm{d}t\right)\Bigg). \tag{8.9}$$

Here $\phi_{ij}(t)$ and $\psi_{ij}(t)$ can be obtained from Equations (5.80) and (5.83) respectively, and for $i_1\geq 2$, $\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|E^{(n)}_{(l_2,r_1)}\right]$ is given by

$$\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|E^{(n)}_{(l_2,r_1)}\right]=1_{\{i_1=2\}}e^{-s\Delta_{r_1}}\phi_{l_2k_1}(T_{(2,r_1)})$$
$$+1_{\{i_1\geq 3\}}\sum_{l_3=1}^{M}\mathbb{E}\left[e^{-sG^{(n)}}1_{\{J_{n+1}=(i_1,k_1,r_1)\}}|E^{(n)}_{(l_3,r_1)}\right]\left(\int_{t=0}^{T_{(2,r_1)}}\psi_{l_2l_3}(t)e^{-st}\mathrm{d}t\right)$$
$$=\begin{cases}e^{-s\Delta_{r_1}}\phi_{l_2k_1}(T_{(2,r_1)}), & \text{if } i_2=2\\ e^{-s\Delta_{r_1}}\left(\prod_{j=3}^{i_1-1}\sum_{l_j=1}^{M}\left(\int_{t=0}^{T_{(j-1,r_1)}}\psi_{l_{j-1}l_j}(t)e^{-st}\mathrm{d}t\right)\right)\\ \times\sum_{l_{i_1}=1}^{M}\left(\int_{t=0}^{T_{(i_1-1,r_1)}}\psi_{l_{i_1-1}l_{i_1}}(t)e^{-st}\mathrm{d}t\right)\phi_{l_{i_1}k_1}(T_{(i_1,r_1)}), & \text{if } i_2\geq 3\end{cases} \tag{8.10}$$

It remains to find the expression for $\pi_{(i_1,k_1,r_1)}$ which can be written as

$$\pi_{(i_1,k_1,r_1)}=\mathbb{P}(J_{n+1}=(i_1,k_1,r_1))$$

$$
\begin{aligned}
&= \sum_{i_0=1}^{N}\sum_{k_0=1}^{M}\sum_{r_0=1}^{R}\Bigg(\mathbb{P}(J_{n+1}=(i_1,k_1,r_1)|J_n=(i_0,k_0,r_0),X_{n-1}\geq 1)\\
&\times \mathbb{P}(J_n=(i_0,k_0,r_0),X_{n-1}\geq 1)\\
&+ \mathbb{P}(J_{n+1}=(i_1,k_1,r_1)|J_n=(i_0,k_0,r_0),X_{n-1}=0)\\
&\times \mathbb{P}(J_n=(i_0,k_0,r_0),X_{n-1}=0)\Bigg)\\
&= \sum_{i_0=1}^{N}\sum_{k_0=1}^{M}\sum_{r_0=1}^{R}\Bigg(P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(T_{(i_0,r_0)}-\Delta_{r_0})\left(\pi_{(i_0,k_0,r_0)}-f_{(i_0,k_0,r_0)}\right)\\
&+\Bigg(\int_{t=0}^{T_{(i_0,r_0)}-\Delta_{r_0}}\lambda e^{-\lambda t}P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(T_{(i_0,r_0)}-\Delta_{r_0}-t)\mathrm{d}t\\
&+\int_{t=T_{(i_0,r_0)}-\Delta_{r_0}}^{\infty}\lambda e^{-\lambda t}\sum_{k=1}^{M}P_{k_0k}(t-T_{(i_0,r_0)}+\Delta_{r_0})P^{(i_1,k_1,r_1)}_{(i_0,k,r_0)}(0)\mathrm{d}t\Bigg)f_{(i_0,k_0,r_0)}\Bigg),
\end{aligned}
\tag{8.11}
$$

where $P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(y)=\tilde{G}^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(0,y)$.

After rearranging the terms corresponding to $\pi_{(i_0,k_0,r_0)}$ in Equation (8.11), we obtain the following linear system of $MNR$ equations for $\pi_{(i,k,r)}$, $i=1,2,\ldots,N, k=1,2,\ldots,M.r=1,2,\ldots,R$:

$$
\begin{aligned}
&(1-P^{(i_1,k_1,r_1)}_{(i_1,k_1,r_1)}(T_{(i_1,r_1)}-\Delta_{r_1}))\pi_{(i_1,k_1,r_1)}-\sum_{\substack{i_0=1,\\ i_0\neq i_1}}^{N}P^{(i_1,k_1,r_1)}_{(i_0,k_1,r_1)}(T_{(i_0,r_1)}-\Delta_{r_1})\\
&\times\pi_{(i_0,k_1,r_1)}-\sum_{i_0=1}^{N}\sum_{\substack{k_0=1,\\ k_0\neq k_1}}^{M}P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_1)}(T_{(i_0,r_1)}-\Delta_{r_1})\pi_{(i_0,k_0,r_1)}\\
&-\sum_{i_0=1}^{N}\sum_{k_0=1}^{M}\sum_{\substack{r_0=1,\\ r_0\neq r_1}}^{R}P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(T_{(i_0,r_0)}-\Delta_{r_0})\pi_{(i_0,k_0,r_0)}\\
&=\sum_{i_0=1}^{N}\sum_{k_0=1}^{M}\sum_{r_0=1}^{R}\Bigg(-P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(T_{(i_0,r_0)}-\Delta_{r_0})+\int_{t=0}^{T_{(i_0,r_0)}-\Delta_{r_0}}\lambda e^{-\lambda t}
\end{aligned}
$$

$$\times P^{(i_1,k_1,r_1)}_{(i_0,k_0,r_0)}(T_{(i_0,r_0)} - \Delta_{r_0} - t)\mathrm{d}t + \int_{t=T_{(i_0,r_0)}-\Delta_{r_0}}^{\infty} \lambda e^{-\lambda t}$$
$$\times \sum_{k=1}^{M} P_{k_0 k}(t - T_{(i_0,r_0)} + \Delta_{r_0}) P^{(i_1,k_1,r_1)}_{(i_0,k,r_0)}(0)\mathrm{d}t \Bigg) f_{(i_0,k_0,r_0)}. \tag{8.12}$$

Hence, $\pi_{(i_1,k_1,r_1)}$ is obtained as the solution of the above linear system of equations.

Similar to Chapter 7, the capacity can be derived from the LST of the service time distribution, $\tilde{G}(s)$.

**Remark 8.1.** *As the framework is essentially the same as in Chapter 7, we can use the same $M^X/SM2/1$ queueing model to obtain the stationary queue length distribution on the minor road.*

## 8.3 Discussion and conclusion

In this chapter, we have presented an extension of the generalized gap acceptance model for unsignalized priority-controlled intersections, studied in Chapter 7, that allows Markov platooning on the main road. We determined the LST of the service time of an arbitrary low-priority driver, which was further used to derive the capacity of the minor road. It is possible to include the randomness in the critical gaps at each attempt, within the same profile of the low-priority driver, as discussed in Chapter 7 in this model, with only minor adaptations.

An interesting direction for further research, would be to allow a more general arrival process on the main road. In the present chapter, we assume that the arrival process on the main road is a Markov modulated Poisson process that allows Markov platooning on that road. However, in practice, there might be platoon forming on this road that may not be Markovian, and several papers have shown that this clustering of vehicles will influence the capacity of the *minor* road (cf. [3, 101, 111, 119]). It would certainly be interesting to combine the framework in this chapter with, for example, the gap-block arrival process in Cowan [32].

# Summary

In this era of rapidly emerging new technologies for urban traffic control, the vast majority of urban traffic intersections are still priority-based unsignalized intersections, in which major roads have priority over minor roads. Due to the increasing level of traffic congestion, the need for state-of-the-art quantitative analysis methods is greater than ever. Performance analysis of unsignalized traffic intersections is traditionally based on so-called *gap acceptance models*, with roots that can be traced back to classical queueing models. The basis of these models is the assumption that the crossing decision of a driver on the minor road is based on the gap between two successive vehicles on the major road. In Chapter 1, we show that this model can be applied in other contexts as well, e.g. when analyzing freeways and pedestrian crossings.

In earlier existing works, three variations of gap acceptance models can be distinguished. The first is the basic model in which all low-priority drivers are assumed to use the same fixed critical gap (referred to as model $B_1$ in this thesis). In the second model ($B_2$), critical gaps are random, with drivers sampling a new critical gap (headway) at each new attempt. This is typically referred to as *inconsistent* driver behavior. The third model ($B_3$) is also known as the *consistent* model, in which a random critical gap is sampled by each driver for his first attempt only, and the driver then uses that same value at his subsequent attempts. A topic of particular interest concerns the capacity of the minor road, which is defined as the maximum possible number of vehicles per time unit that can pass through an intersection from the minor road. Other relevant performance measures are the queue length and the delay distributions on the minor road.

In Chapter 2, we study *impatience* of the drivers that are waiting to cross the major road (during the waiting time of the driver at the head of the queue, the critical headway becomes lower). We do this for the existing gap acceptance models $B_1$ up to $B_3$ for unsignalized intersections, where vehicles arrive according to a Poisson process on the minor road as well as on the major road. We reveal some interesting results which were not observed in

the existing literature. In particular, we show a strict ordering among the capacities for the models $B_1$, $B_2$ and $B_3$ without impatience, which is not necessarily true for the mean queue lengths on the minor road nor for the case of impatience. More specifically, it is shown that $B_2$ has a larger capacity than $B_1$, and the capacity of $B_3$ is the smallest (with the mean critical headway of models $B_2$ and $B_3$ chosen equal to the deterministic critical headway of model $B_1$). Hence, it can be observed that randomness within the critical gaps has a strong impact on the capacity and in particular, larger variability can improve the capacity.

In Chapter 3, we introduce Markov platooning on the major road to model fluctuations in the traffic density on that road. The tractability of this model allows one to study the capacity of the minor road for various platoon formations on the major road. We assume that the arrival process on the major road is modeled by a Markov modulated Poisson process (MMPP). Based on the numerical results, we observe for models $B_1$ and $B_3$ that platooning has a positive effect on the capacity of the minor road for given mean rate. In a model with inconsistent behavior, it depends on the model parameters whether platooning increases or decreases the capacity.

In Chapter 4, we investigate a single server queueing model with batch arrivals and semi-Markov service times. An essential feature of this queueing model is that the type of customer $n+1$ not only depends on the type of customer $n$, but also on the length of the service of customer $n$. We determine the transient and stationary probability generating functions of the number of customers in the system. We numerically investigate how the mean queue length is affected by variability in the number of customers that arrive during a single service time. Our main observations here are that increasing variability may *reduce* the mean queue length (note the similar observation in Chapter 2), and that the Markovian dependence of service times can lead to large queue lengths, even if the system is not heavily loaded.

In Chapter 5, we extend the queueing model, $M^X$/semi-Markov/1, studied in Chapter 4, with exceptional first services, i.e., the first customer of each busy period has a different service-time distribution than regular customers served in the busy period. Based on the results from the previous chapter, we derive the waiting time and sojourn time distributions of an arbitrary customer, showing that these depend on the position of the customer inside the batch, as well as on the type of the first customer in the batch. We present several applications in which this queueing model arises, and then study

the application to road traffic situations involving multiple conflicting traffic streams. In particular, we use it in the context of gap acceptance models for unsignalized intersections, studied in Chapter 3. We numerically demonstrate the impact of the three types of the driver's behavior ($B_1$, $B_2$, and $B_3$), on the delay on the minor road.

In Chapter 6, we present a heavy-traffic analysis of the queueing model studied in Chapter 5. We show that the distribution of the scaled stationary queue length in heavy traffic is exponential. This distribution does not depend on the service-time distribution of the first customer of each busy period that is different from service-time distributions of customers served in a busy period.

In Chapter 7, we introduce a generalized gap acceptance model for unsignalized intersections, where vehicles arrive according to a Poisson process on the major road and according to a *batch* Poisson process on the minor road. The generalized model consists of multiple classes of gap acceptance behavior (including impatience of the drivers) as well as merging behavior of drivers. In particular, we allow the subsequent drivers to use the unused part of a critical headway. Using the framework of Chapter 5, we obtain the queue length distribution and the capacity of the minor road.

In Chapter 8, we present an extension of the gap acceptance model studied in Chapter 7, where vehicles arrive according to an MMPP on the major road. We determine the LST of the service time of an arbitrary low-priority driver, which is further used to derive the capacity of the minor road.

# Samenvatting

In dit tijdperk van snel opkomende nieuwe technologieën voor grootstedelijke verkeersregulering zijn de meeste kruispunten in de stad nog steeds voorrangskruispunten (zonder stoplichten), waarbij de hoofdwegen voorrang krijgen op de secundaire wegen. Door de toenemende verkeersdrukte is de behoefte aan kwantitatieve analysemethoden groter dan ooit. Prestatieanalyse van voorrangskruispunten is traditioneel gebaseerd op zogenaamde gap-acceptatiemodellen, welke terug te voeren zijn op klassieke wachtrijmodellen. De basis van deze modellen is de aanname dat de beslissing van een bestuurder om de secundaire weg over te steken, afhangt van de afstand tussen twee opeenvolgende voertuigen op de hoofdweg. In Hoofdstuk 1 laten we zien dat dit model ook in andere contexten kan worden toegepast, bijvoorbeeld bij het analyseren van snelwegen en voetgangersoversteekplaatsen.

In bestaand werk kan men drie variaties van gap-acceptatie modellen onderscheiden. De eerste is het basismodel waarin van alle bestuurders die voorrang moeten verlenen wordt verondersteld dat ze dezelfde vaste kritieke afstand (volgafstand) gebruiken (in dit proefschrift aangeduid als model $B_1$). In het tweede model ($B_2$) is de kritieke afstand stochastisch, waarbij bestuurders bij elke nieuwe poging een nieuw kritieke afstand trekken uit een bepaalde verdeling. Dit wordt aangeduid als inconsistent rijgedrag. Het derde model ($B_3$) noemen we het consistente model, waarbij voor de eerste poging om over te steken (of in te voegen) een kritieke volgafstand getrokken wordt uit een kansverdeling en de bestuurder vervolgens dezelfde waarde gebruikt bij zijn volgende pogingen. Een bijzonder interessant onderwerp betreft ook de capaciteit van de secundaire weg (de voorrang verlenende weg), die wordt gedefinieerd als het maximaal mogelijke aantal voertuigen van de secundaire weg die het kruispunt per tijdseenheid kunnen passeren. Andere relevante prestatiematen zijn de lengte van de wachtrij en de vertraging op de secundaire weg.

In Hoofdstuk 2 bestuderen we ‘ongeduldige bestuurders’, die tijdens hun wachttijd om de hoofdweg over te kunnen steken een steeds kleinere kritieke

volgafstand hanteren. We behandelen de bestaande gap-acceptatiemodellen $B_1$, $B_2$ en $B_3$ voor voorrangskruispunten, waarbij voertuigen volgens onafhankelijke Poisson processen aankomen op zowel de secundaire weg als op de hoofdweg. We tonen enkele interessante resultaten aan die niet vermeld zijn in bestaande literatuur. In het bijzonder laten we een strikte ordening zien tussen de capaciteiten voor de modellen $B_1$, $B_2$ en $B_3$ zonder ongeduldige bestuurders, wat niet noodzakelijk geldt voor de *gemiddelde* wachtrijlengten op de secundaire weg. Specifiek wordt aangetoond dat $B_2$ een grotere capaciteit heeft dan $B_1$, en de capaciteit van $B_3$ de kleinste is (met de gemiddelde kritieke volgafstand in elk van de modellen $B_2$ en $B_3$ gelijk aan de deterministische kritieke volgafstand van model $B_1$). We concluderen daaruit dat stochasticiteit voor de kritieke volgafstand een sterke invloed heeft op de capaciteit; in het bijzonder kan grotere variabiliteit de capaciteit verbeteren. Onze resultaten blijken niet algemeen te gelden wanneer bestuurders ongeduldig zijn.

In Hoofdstuk 3 introduceren we Markov platooning op de hoofdweg om fluctuaties in de verkeersdichtheid op die weg te modelleren. Dit model stelt ons in staat om de capaciteit van de secundaire weg te bestuderen voor verschillende manieren van 'platoonvorming' de hoofdweg. We nemen aan dat het aankomstproces op de hoofdweg wordt gemodelleerd door een Markov-gemoduleerd Poisson-proces (MMPP). In op de numerieke resultaten observeren we voor modellen $B_1$ en $B_3$ dat platooning een positief effect heeft op de capaciteit van de secundaire weg voor een gegeven gemiddelde snelheid. In een model met inconsistent gedrag, hangt het van de modelparameters af of platooning de capaciteit verhoogt of verlaagt.

In Hoofdstuk 4 onderzoeken we een wachtrijmodel met groepsaankomsten en semi-Markov bedieningstijden. Een essentieel kenmerk van dit wachtrijmodel is dat het type van een klant niet alleen afhankelijk is van het type van zijn voorganger, maar ook van diens bedieningsduur. We bepalen de tijdsafhankelijke en stationaire kansgenererende functies van het aantal klanten in het systeem. We onderzoeken via numerieke analyse hoe de gemiddelde lengte van de wachtrij wordt beïnvloed door de variabiliteit in het aantal klanten dat binnenkomt tijdens een enkele bediening. Onze belangrijkste observaties hier zijn dat toenemende variabiliteit de gemiddelde wachtrijlengte kan verminderen (let op de vergelijkbare observatie in Hoofdstuk 2), en dat de Markovafhankelijkheid van bedieningstijden kan leiden tot grote wachtrijlengtes, zelfs als het systeem niet zwaar beladen is.

In Hoofdstuk 5 breiden we het wachtrijmodel, $M^X$/semi-Markov/1, uit

Hoofdstuk 4, uit met afwijkende eerste bedieningen, dat wil zeggen dat klanten die in een leeg systeem aankomen een andere bedieningstijdverdeling hebben dan reguliere klanten die bij aankomst aansluiten achter reeds aanwezige klanten. Op basis van de resultaten van het vorige hoofdstuk over de stationaire verdeling van de lengte van de wachtrij, leiden we de wachttijdverdeling en verblijftijdverdeling af van een willekeurige klant, waaruit blijkt dat deze afhankelijk zijn van de positie van de klant binnen de groep, evenals van het type van de eerste klant in de groep. We presenteren verschillende toepassingen van dit wachtrijmodel ontstaat en bestuderen vervolgens de toepassing bij wegverkeerssituaties met meerdere kruisende verkeersstromen. We gebruiken het in het bijzonder in de context van gap-acceptatie modellen voor voorrangskruispunten, bestudeerd in Hoofdstuk 3. We demonstreren via numerieke analyse de impact van de drie typen rijgedrag ($B_1$, $B_2$ en $B_3$) op de vertraging op de secundaire weg.

In Hoofdstuk 6 presenteren we een *heavy traffic* analyse van het wachtrijmodel dat is bestudeerd in Hoofdstuk 5. We laten zien dat de heavy traffic verdeling van de geschaalde stationaire wachtrijlengte exponentieel is. Deze verdeling is niet afhankelijk van de bedieningstijdverdeling van klanten die in een leeg systeem aankomen (wanneer deze verschillend is van die van klanten die in een wachtrij aansluiten).

In Hoofdstuk 7 introduceren wij een algemeen gap-acceptatie model voor voorrangskruispunten, waarin de voertuigen volgens een Poisson proces op de hoofdweg en volgens een Poisson proces met *groepsaankomsten* op de secundaire weg aankomen. Het algemene model bestaat uit meerdere klassen van gap-acceptatie gedrag (we nemen hierin ook het gedrag van ongeduldige bestuurders mee). In het bijzonder staan we toe dat de andere bestuurders het ongebruikte deel van de kritieke volgafstand gebruiken. Met behulp van Hoofdstuk 5 verkrijgen we de verdeling van de rijlengte en de capaciteit van de secundaire weg.

In Hoofdstuk 8 stellen we een uitbreiding voor van het gap-acceptatie model zoals bestudeerd in Hoofdstuk 7, waarbij we toestaan dat voertuigen volgens een MMPP op de hoofdweg aankomen. We bepalen de Laplace transformatie van de bedieningsduur van een willekeurige bestuurder die voorrang moet verlenen, waarmee we vervolgens de capaciteit van de secundaire weg berekenen.

# Publications of the author

The work in this thesis is based on the articles mentioned below. The content of Chapter 2 has been taken from both [1] and [2]. Chapters 3 and 4 contain the material of [2] and [3], respectively. Chapter 5 is based on both [4] and [5]. Chapters 6, 7 and 8 correspond to [5], [6] and [7], respectively. Only small changes have been made to improve readability. Abhishek carried out the research and wrote the articles, under the supervision of Marko Boon, Onno Boxma, Michel Mandjes, and Rudesindo Núñez-Queija. All co-authors gave feedback on the drafts of each corresponding article before it was submitted.


[1] Abhishek, M.A.A. Boon, M.R.H. Mandjes, and R. Núñez-Queija (2016). Congestion analysis of unsignalized intersections. *COMSNETS 2016: Intelligent Transportation Systems Workshop*.

[2] Abhishek, M.A.A. Boon, M.R.H. Mandjes, and R. Núñez-Queija. Congestion analysis of unsignalized intersections: The impact of impatience and Markov platooning. To appear in *European Journal of Operational Research*. ArXiv:1802.06732.

[3] Abhishek, M.A.A. Boon, O.J. Boxma, and R. Núñez-Queija (2017). A single server queue with batch arrivals and semi-Markov services. *Queueing Systems* 86(3-4), 217-–240.

[4] Abhishek, M.A.A. Boon, and R. Núñez-Queija. Applications of the $M^X$/semi-Markov/1 queue to road traffic. To be submitted for publication.

[5] Abhishek, M.A.A. Boon, and R. Núñez-Queija. Heavy-traffic analysis of the $M^X$/semi-Markov/1 queue. To be submitted for publication.

[6] Abhishek, M.A.A. Boon, and M.R.H. Mandjes. Generalized gap acceptance models for unsignalized intersections. Submitted for publication. ArXiv:1802.04192.

[7] Abhishek and R. Núñez-Queija. Gap acceptance models with Markov platooning (in preparation).

# About the author

Abhishek was born on 1st February 1989 in Said Alipur, Haryana (India). He received an M.Sc. in Mathematics from the University of Rajasthan, Jaipur (India) in 2011. He qualified GATE with All India Rank 46 and CSIR–JRF June with All India Rank 32 in Mathematics in 2012. He also received an M.Tech. in Industrial Mathematics and Scientific Computing (IMSC) from the Indian Institute of Technology (IIT) Madras, Chennai (India) in 2014 and his master thesis entitled 'Monte Carlo Simulation of Stochastic Volatility Models' was about forecasting the direction of future stock prices. He was awarded the Prof Helmut Neunzert Endowment Prize for the best academic record in the IMSC of the M.Tech. degree programme for the period 2012-2104. In October 2014, he started as a PhD researcher at the University of Amsterdam in collaboration with the Eindhoven University of Technology (The Netherlands), under the supervision of Dr. Marko Boon, Prof. dr. Onno Boxma, Prof. dr. Michel Mandjes, and Prof. dr. Rudesindo Núñez-Queija, and his project is about the congestion analysis of unsignalized road traffic intersections. In January 2018, he received the LNMB (in Dutch: Landelijk Netwerk Mathematische Besliskunde) diploma for PhD courses in the Mathematics of Operations Research. His main research interests are broadly in the field of applied and computational mathematics, in particular, applied probability, queueing theory, mathematical optimization, stochastic modeling and simulation.

# References


[1] Abhishek, M. A. A. Boon, O. J. Boxma and R. Núñez Queija. 'A single server queue with batch arrivals and semi-Markov services'. In: *Queueing Systems* 86.3–4 (2017), pp. 217–240.

[2] Abhishek, M. A. A. Boon, M. Mandjes and R. Núñez Queija. 'Congestion analysis of unsignalized intersections'. In: *COMSNETS 2016: Intelligent Transportation Systems Workshop*. 2016, pp. 1–6.

[3] Abhishek, M. A. A. Boon, M. Mandjes and R. Núñez Queija. *Congestion analysis of unsignalized intersections: The impact of impatience and Markov platooning*. ArXiv:1802.06732. University of Amsterdam, 2017.

[4] M. Abou-Henaidy, S. Teply and J. H. Hund. 'Gap acceptance investigations in Canada'. In: *Proceedings of the Second Int. Symp. on Highway Capacity*. Ed. by R. Akçelik. Vol. 1. 1994, pp. 1–19.

[5] I. J. B. F. Adan and V. G. Kulkarni. 'Single-server queue with Markov-dependent inter-arrival and service times'. In: *Queueing Systems* 45 (2003), pp. 113–134.

[6] G. Asaithambi and C. Anuroop. 'Analysis of occupation time of vehicles at urban unsignalized intersections in non-lane-based mixed traffic conditions'. In: *Journal of Modern Transportation* 24.4 (2016), pp. 304–313.

[7] S. Asmussen. 'The heavy traffic limit of a class of Markovian queueing models'. In: *Operations Research Letters* 6.6 (1987), pp. 301–306.

[8] N. Baër. 'Queueing and traffic'. PhD Thesis. Centre for Telematics and Information Technology, University of Twente, 2015.

[9] M. Barth and K. Boriboonsomsin. 'Traffic congestion and greenhouse gases'. In: *ACCESS Magazine* 1.35 (2009), pp. 2–9.

[10] A. L. C. Bazzan and F. Klügl. 'A review on agent-based technology for traffic and transportation'. In: *The Knowledge Engineering Review* 29.3 (2014), pp. 375–403.

[11] M. Brackstone and M. McDonald. 'Car-following: a historical review'. In: *Transportation Research Part F* 2 (1999), pp. 181–196.

[12] W. Brilon. 'Recent developments in calculation methods for unsignalized intersections in West Germany'. In: *Intersections without Traffic Signals*. Ed. by W. Brilon. Springer, Berlin, Heidelberg, 1988, pp. 111–153.

[13] W. Brilon and T. Miltner. 'Capacity at intersections without traffic signals'. In: *Transportation Research Record* 1920 (2005), pp. 32–40.

[14] W. Brilon, R. Troutbeck and M. Tracz. 'Review of international practices used to evaluate unsignalized intersections'. In: *Transportation Research Circular* 468 (1997). Transportation Research Board, Washington, DC..

[15] W. Brilon and N. Wu. 'Capacity at unsignalized intersections derived by conflict technique'. In: *Transportation Research Record* 1776 (2001), pp. 82–90.

[16] W. Brilon and N. Wu. 'Unsignalized Intersections - A Third Method for Analysis'. In: *Transportation and Traffic Theory in the 21st Century*. 2002. Chap. 9, pp. 157–178.

[17] W. Brilon (ed.) *Intersection without Traffic Signals*. Springer, Berlin, Heidelberg, 1988.

[18] W. Brilon (ed.) *Intersection without Traffic Signals II*. Springer, Berlin, Heidelberg, 1991.

[19] P. J. Burke. 'Delays in single-server queues with batch input'. In: *Operations Research* 23 (1975), pp. 830–833.

[20] D. Y. Burman and Smith D. R. 'An asymptotic analysis of a queueing system with Markov-modulated arrivals'. In: *Operations Research* 34.1 (1986), pp. 105–119.

[21] C. Caliendo. 'Delay time model at unsignalized intersections'. In: *Journal of Transportation Engineering* 140.9 (2014), pp. 1–13.

[22] E. A. Catchpole and A. W. Plank. 'The capacity of a priority intersection'. In: *Transportation Research-B* 20.6 (1986), pp. 441–456.

[23] E. Çinlar. 'Time dependence of queues with semi-Markovian services'. In: *J. Appl. Probab.* 4 (1967), pp. 356–364.

[24] P. L. Chan and S. Teply. 'Simulation of multilane stop-controlled T-intersections by Knosimo in Canada'. In: *Intersections without Traffic Signals II*. ed. by W. Brilon. Springer, Berlin, Heidelberg, 1991, pp. 308–319.

[25] M. L. Chaudhry. 'The queueing system $M^X/G/1$ and its ramifications'. In: *Naval Res. Logist. Quart.* 26 (1979), pp. 667–674.

[26] L. Chen and C. Englund. 'Cooperative intersection management: a survey'. In: *IEEE Transactions on Intelligent Transportation Systems* 17.2 (2016), pp. 570–586.

[27] T. E. C. Cheng and S. Allam. 'A review of stochastic modelling of delay and capacity at unsignalized priority intersections'. In: *EJOR* 60.3 (1992), pp. 247–259.

[28] P. Christidis and J. N. I. Rivas. *Measuring road congestion*. JRC Technical Notes. European Commission, Joint Research Centre, Institute for Prospective Technological Studies, Luxembourg, 2012.

[29] J. W. Cohen. *The Single Server Queue.* Amsterdam: North-Holland, 1969.

[30] J. W. Cohen and O. J. Boxma. 'A survey of the evolution of queueing theory'. In: *Statistica Neerlandica* 39 (1985), pp. 143–158.

[31] R. Cowan. 'An extension of Tanner's results on uncontrolled intersections'. In: *Queueing Systems* 1.3 (1987), pp. 249–263.

[32] R. J. Cowan. 'Useful headway models'. In: *Transportation Research* 9.6 (1975), pp. 371–375.

[33] C. F. Daganzo. 'Traffic delay at unsignalized intersections: clarification of some issues'. In: *Transportation Science* 11.2 (1977), pp. 180–189.

[34] M. Dimitrov. 'Single-server queueing system with Markov-modulated arrivals and service times'. In: *Pliska Stud. Math. Bulgar.* 20 (2011), pp. 53–62.

[35] A. Doniec, R. Mandiau, S. Piechowiak and S. Espié. 'A behavioral multi-agent model for road traffic simulation'. In: *Engineering Applications of Artificial Intelligence* 21.8 (2008), pp. 1443–1454.

[36] A. Doniec, S. Espié, R. Mandiau and S. Piechowiak. 'Multi-agent coordination and anticipation model to design a road traffic simulation tool'. In: *EUMAS'06, Proceedings of the Fourth European Workshop on Multi-Agent Systems.* Lisbon, 2006.

[37] D. R. Drew, J. H. Buhr and R. H. Whitson. *The determination of merging capacity and its applications to freeway design and control.* Report 430-4. Texas Transportation Institute, 1967.

[38] D. R. Drew, L. R. LaMotte, J. H. Buhr and J. A. Wattleworth. *Gap acceptance in the freeway merging process.* Report 430-2. Texas Transportation Institute, 1967.

[39] A. K. Erlang. 'Sandsynlighedsregning og telefonsamtaler'. In: *Nyt Tidsskrift for Matematik* 20 (1909), pp. 33–39.

[40] D. H. Evans, R. Herman and G. H. Weiss. 'The highway merging and queueing problem'. In: *Operations Research* 12 (1964), pp. 832–857.

[41] M. A. Evgrafov. *Analytic Functions.* New York: Dover, 1978.

[42] G. Falin and A. Falin. 'Heavy traffic analysis of M/G/1 type queueing systems with Markov-modulated arrivals'. In: *Sociedad de Estadistica e Investigacion Operativa Top* 7.2 (1999), pp. 279–291.

[43] H. G. Findeisen. 'Das Verhalten verkehrsrechtlich untergeordneter Fahrzeuge an nicht lichtsignalgesteuerten Knotenpunkten (The behaviour of subordinate vehicles at unsignalized intersections)'. In: *Strassenbau, Verkehrstechnik und Verkehrssicherheit* 15 (1971).

[44] W. Fischer and K. Meier-Hellstern. 'The Markov-modulated Poisson process (MMPP) cookbook'. In: *Performance Evaluation* 18 (1992), pp. 149–171.

[45] S. T. G. Fleuren. 'Optimizing pre-timed control at isolated intersections'. PhD Thesis. Beta Research School for Operations Management and Logistics, Eindhoven University of Technology, 2017.

[46] H. R. Gail, S. L. Hantler and B. A. Taylor. 'On a preemptive Markovian queue with multiple servers and two priority classes'. In: *Mathematics of Operations Research* 17 (1992), pp. 365–391.

[47] A. Gaur and P. Mirchandani. 'Method for real-time recognition of vehicle platoons'. In: *Transportation Research Record* 1748 (2001), pp. 8–17.

[48] D. P. Gaver. 'A comparison of queue disciplines when service orientation times occur'. In: *Naval Res. Logist. Quart.* 10 (1963), pp. 219–235.

[49] D. P. Gaver. 'Imbedded Markov chain analysis of a waiting-line process in continuous time'. In: *Ann. Math. Statist.* 30 (1959), pp. 698–720.

[50] P. G. Gipps. 'A behavioural car-following model for computer simulation'. In: *Transportation Research Part B* 15 (1981), pp. 105–111.

[51] A. W. Gleue. 'Vereinfachtes Verfahren zur Berechnung Signalgeregelter Knotenpunkte'. In: *Forschung Strassenbau und Strassenverkehrstechnik, No. 136, Bonn* (1972).

[52] M. Grossmann. 'KNOSIMO - A practicable simulation model for unsignalized intersections'. In: *Intersections without Traffic Signals.* Ed. by W. Brilon. Springer, Berlin, Heidelberg, 1988, pp. 263–273.

[53] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* New York, NY, USA: Cambridge University Press, 2013.

[54] J. Harders. 'Die Leistungsfähigkeit nicht signalgeregelter städtischer Verkehrsknoten (The capacity urban intersections)'. In: *Schriftenreihe Strassenbau und Strassenverkehrstechnik* 76 (1968).

[55] J. Harders. 'Grenz- und Folgezeit1ücken als Grundlage für die Leistungsfähigkeit von Landstrassen (Critical gaps and move-up times as the basis of capacity calculations for rural roads)'. In: *Schriftenreihe Strassenbau und Strassenverkehrstechnik* 216 (1976).

[56] A. G. Hawkes. 'Delay at traffic intersections'. In: *Journal of the Royal Statistical Society Series B* 28.1 (1966), pp. 202–212.

[57] A. G. Hawkes. 'Gap-acceptance in road traffic'. In: *J. Appl. Prob.* 5 (1968), pp. 84–92.

[58] A. G. Hawkes. 'Queueing for gaps in traffic'. In: *Biometrika* 52.1/2 (1965), pp. 79–85.

[59] D. Heidemann. 'Queue length and delays distributions at traffic signals'. In: *Transportation Research-B* 28.5 (1994), pp. 377–389.

[60] D. Heidemann and H. Wegmann. 'Queueing at unsignalized intersections'. In: *Transportation Research-B* 31.3 (1997), pp. 239–263.

[61] R. A. Horn and C. R. Johnson. *Matrix Analysis*. New York, NY, USA: Cambridge University Press, 1986. ISBN: 0-521-30586-1.

[62] D. Jia, K. Lu, J. Wang, X. Zhang and X. Shen. 'A survey on platoon-based vehicular cyber-physical systems'. In: *IEEE Commun. Surveys Tut.* 18.1 (2016), pp. 263–284.

[63] V. Kanagaraj, G. Asaithambi, C. H. N. Kumar, K. K. Srinivasan and R. Sivanandan. 'Evaluation of different vehicle following models under mixed traffic conditions'. In: *Procedia - Social and Behavioral Sciences* 104 (2013), pp. 390–401.

[64] J. Keilson and L. D. Servi. 'A distributional form of Little's law'. In: *Operations Research Letters* 7.5 (1988), pp. 223–227.

[65] R. M. Kimber and R. D. Coombe. *The traffic capacity of major/minor priority junctions*. Supplementary Report 582. Transport and Road Research Laboratory, 1980.

[66] J. F. C. Kingman. 'The single server queue in heavy traffic'. In: *Proc. Camb. Philos. Soc.* 57.4 (1961), pp. 902–904.

[67] L. Kleinrock. *Queueing Systems, Vol. I: Theory*. New York: Wiley, 1975.

[68] W. Knospe, L. Santen, A. Schadschneider and M. Schreckenberg. 'Towards a realistic microscopic description of highway traffic'. In: *Journal of Physics A: Mathematical and General* 33.48 (2000), pp. 1–7.

[69] M. Kyte, J. Zegeer and B. K. Lall. 'Empirical models for estimating capacity and delay at stop-controlled intersections in the United States'. In: *Intersections without Traffic Signals II*. ed. by W. Brilon. Springer, Berlin, Heidelberg, 1991, pp. 335–361.

[70] M. Kyte, C. Clemow, N. Mahfood, B. K. Lall and C. J. Khisty. 'Capacity and delay characteristics of two-way stop-controlled intersections'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1320 (1991), pp. 160–167.

[71] B. Li. 'Stochastic modeling for vehicle platoons (I): Dynamic grouping behavior and online platoon recognition'. In: *Transportation Research Part B* 95 (2017), 364––377.

[72] X. G. Li, Z. Y. Gao, B. Jia and X. M. Zhao. 'Cellular automata model for unsignalized T-shaped intersection'. In: *International Journal of Modern Physics C* 20.4 (2009), pp. 501–512.

[73] J. D. C. Little. 'A proof for the queuing Formula: L $= \lambda$ W'. in: *Operations Research* 9.3 (1961), pp. 383–387.

[74] M. Liu, G. Lu, Y. Wang and Z. Zhang. 'Analyzing drivers' crossing decisions at unsignalized intersections in China'. In: *Transportation Research Part F: Traffic Psychology and Behaviour* 24 (2014), pp. 244–255.

[75] R. M. Loynes. 'The stability of a queue with non-independent inter-arrival and service times'. In: *Proc. Camb. Phil. Soc.* 58 (1962), pp. 497–520.

[76] M. N. Magalhães and R. L. Disney. 'Departures from queues with changeover times'. In: *Queueing Systems* 5 (1989), pp. 295–312.

[77] A. J. Mayne. 'Some further results in the theory of pedestrians and road traffic'. In: *Biometrika* 41.3/4 (1954), pp. 375–389.

[78] P. K. Munjal and L. A. Pipes. 'Propagation of on-ramp density perturbations on unidirectional two- and three-lane freeways'. In: *Transportation Research* 5.4 (1971), pp. 241–255.

[79] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory — The Mathematics of Computer Performance Modeling*. New York: Springer, 1995.

[80] M. F. Neuts. 'Some explicit formulas for the steady-state behavior of the queue with semi-Markovian service times'. In: *Adv. in Appl. Probab.* 9 (1977), pp. 141–157.

[81] M. F. Neuts. 'The $M/G/1$ queue with several types of customers and changeover times'. In: *Adv. in Appl. Probab.* 9 (1977), pp. 604–644.

[82] M. F. Neuts. 'The single server queue with Poisson input and semi-Markov service times'. In: *J. Appl. Probab.* 3 (1966), pp. 202–230.

[83] S. Panwai and H. Dia. 'Neural agent car-following models'. In: *IEEE Transactions on Intelligent Transportation Systems* 8.1 (2007), pp. 60–70.

[84] J. Prasetijo. 'Capacity and traffic performance of unsignalized intersections under mixed traffic conditions'. PhD Thesis. Ruhr-University Bochum, 2007.

[85] J. Prasetijo and H. Ahmad. 'Capacity analysis of unsignalized intersection under mixed traffic conditions'. In: *Procedia - Social and Behavioral Sciences* 43 (2012). 8th International Conference on Traffic and Transportation Studies (ICTTS 2012), pp. 135–147.

[86] P. Purdue. 'A queue with Poisson input and semi-Markov service times: busy period analysis'. In: *J. Appl. Probab.* 12 (1975), pp. 353–357.

[87] M. Pursula. 'Simulation of traffic systems - an overview'. In: *Journal of Geographic Information and Decision Analysis* 3.1 (1999), pp. 1–8.

[88] H. G. Retzko. 'Vergleichende Bewertung verschiedener Arten der Verkehrsregelung an städtischen Strassenverkehrsknotenpunkten (Comparative assessment of different kinds of traffic control devices at urban intersections)'. In: *Schriftenreihe Strassenbau und Strassenverkehrstechnik* 12 (1961).

[89] S. M. Ross. *Introduction to Probability Models - 10th Edition*. Academic Press, Inc. Orlando, FL, USA, 2010.

[90] S. M. Ross. *Stochastic Processes - 2nd Edition*. Wiley, 1996.

[91] H. J. Ruskin and R. Wang. 'Modeling traffic flow at an urban unsignalized intersection'. In: *Computational Science — ICCS 2002*. Ed. by P. M. A. Sloot, A. G. Hoekstra, C. J. K. Tan and J. J. Dongarra. Springer, Berlin, Heidelberg, 2002, pp. 381–390.

[92] M. Saifuzzaman and Z. Zheng. 'Incorporating human-factors in car-following models: A review of recent developments and research needs'. In: *Transportation Research Part C* 48 (2014), pp. 379–403.

[93] T. Sayed, G. Brown and F. Navin. 'Simulation of traffic conflicts at unsignalized intersections with TSC-Sim'. In: *Accid. Anal. and Prev.* 26.5 (1994), pp. 593–607.

[94] W. Siegloch. 'Die Leistungsermittlung an Knotenpunkten ohne Lichtsignalsteuerung'. In: *Schriftenreihe Strassenbau und Strassenverkehrstechnik* 154 (1973).

[95] A. Singh and M. Agrawal. 'Acid rain and its ecological consequences'. In: *Journal of Environmental Biology* 29.1 (2008), pp. 15–24.

[96] J. H. A. de Smit. 'The queue $GI/M/s$ with customers of different types or the queue $GI/H_m/s$'. In: *Adv. Appl. Probab.* 15 (1983), pp. 392–419.

[97] J. H. A. de Smit. 'The single server semi-Markov queue'. In: *Stochastic Processes and their Applications* 22 (1986), pp. 37–50.

[98] R. Tachet, P. Santi, S. Sobolevsky, L. I. Reyes-Castro, E. Frazzoli and D. Helbing. 'Revisiting street intersections using slot-based systems'. In: (2016). *PLoS ONE* 11(3): e0149607. doi:10.1371/journal.pone.0149607.

[99] L. Takács. 'The transient behavior of a single server queueing process with a Poisson input'. In: *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 2. 1961, pp. 535–567.

[100] H. Takagi. *Queueing Analysis — A Foundation of Performance Evaluation*. Vol. 1-3. Amsterdam: Elsevier, 1991-1993.

[101] J. C. Tanner. 'A theoretical analysis of delays at an uncontrolled intersection'. In: *Biometrika* 49.1/2 (1962), pp. 163–170.

[102] J. C. Tanner. 'The capacity of an uncontrolled intersection'. In: *Biometrika* 54.3/4 (1967), pp. 657–658.

[103] J. C. Tanner. 'The delay to pedestrians crossing a road'. In: *Biometrika* 38.3/4 (1951), pp. 383–392.

[104] H. Thorsdottir and I. M. Verloop. 'Markov-modulated M/G/1-type queue in heavy traffic and its application to time-sharing disciplines'. In: *Queueing Systems* 83 (2016), pp. 29–55.

[105] Z. Z. Tian, M. Kyte, M. Vandehey, W. Kittelson and B. Robinson. 'Simulation-based study of traffic operational characteristics at all-way-stop-controlled intersections'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1776 (2001), pp. 75–81.

[106] F. Tonke. 'Wartezeiten bei instationarem Verkehr an Knotenpunkten ohne Lichtsignalanlagen (Delays with non-stationary traffic at unsignalized intersections)'. In: *Schriftenreihe Strassenbau und Strassenverkehrstechnik* 401 (1983).

[107] M. Tracz and S. Gondek. 'Use of simulation to analysis of impedance impact at unsignalised intersections'. In: *Transportation Research Circular E-C108: 4th International Symposium on Highway Capacity.* Ed. by W. Brilon. 2000, pp. 471–483.

[108] Transportation Research Board. *Highway Capacity Manual 2010.* 2010.

[109] R. J. Troutbeck. 'Average delay at an unsignalized intersection with two major streams each having a dichotomised headway distribution'. In: *Transportation Science* 20.4 (1986), pp. 272–286.

[110] S. M. Tupper. 'Safety and operational assessment of gap acceptance through large-scale field evaluation'. MSc. Thesis. University of Massachusetts Amherst, 2011.

[111] H. Wegmann. 'A general capacity formula for unsignalized intersections'. In: *Intersections without Traffic Signals II.* ed. by W. Brilon. Springer, Berlin, Heidelberg, 1991, pp. 177–191.

[112] H. Wegmann. 'Impedance-effects in capacity estimation'. In: *ZOR- Methods and Models of Operations Research* 36.1 (1992), pp. 73–91.

[113] D. Wei, W. Kumfer, D. Wu and H. Liu. 'Traffic queuing at unsignalized crosswalks with probabilistic priority'. In: *Transportation Letters* 10.3 (2018).

[114] G. H. Weiss and A. A. Maradudin. 'Some problems in traffic delay'. In: *Operations Research* 10.1 (1962), pp. 74–104.

[115] P. D. Welch. 'On a generalized M/G/1 queuing process in which the first customer of each busy period receives exceptional service'. In: *Operations Research* 12.5 (1964), pp. 736–752.

[116] C. M. Wickens and D. L. Wiesenthal. 'State driver stress as a function of occupational stress, traffic congestion, and trait stress susceptibility'. In: *Journal of Applied Biobehavioral Research* 10 (2005), pp. 83–97.

[117] T. van Woensel and N. Vandaele. 'Modeling traffic flows with queueing models: a review'. In: *Asia Pac. J. Oper. Res.* 24.4 (2007), pp. 435–461.

[118] R. Wolff. 'Poisson arrivals see time averages'. In: *Oper. Res.* 30.2 (1982), pp. 223–231.

[119] N. Wu. 'A universal procedure for capacity determination at unsignalized (priority-controlled) intersections'. In: *Transportation Research Part B* 35 (2001), pp. 593–623.

[120] N. Wu. 'ACF procedure for TWSC intersections - extensions and modifications'. In: *ICCTP 2010: Integrated Transportation Systems*. 2010.

[121] N. Wu. 'Capacity at all-way stop-controlled and first-in-first-out intersections'. In: *Proceedings of the 4th International Symposium on Highway Capacity, Hawaii, Transportation Research Circular E-C018, TRB, Washington, D.C.*. 2000 b.

[122] N. Wu. 'Determination of capacity at all-way stop-controlled intersections'. In: *Transportation Research Record* 1710 (2000 a), pp. 205–214.

[123] N. Wu. 'Impedance effects for streams of higher ranks at unsignalized intersections'. In: *Proceedings of the 3rd International Symposium on Highway Capacity*. 1998.

[124] Q. S. Wu, X. B. Li, M. B. Hu and R. Jiang. 'Study of traffic flow at an unsignalized T-shaped intersection by cellular automata model'. In: *The European Physical Journal B* 48 (2005), pp. 265–269.

[125] G. F. Yeo. 'Single server queues with modified service mechanisms'. In: *Journal of the Australian Mathematical Society* 2.4 (1962), pp. 499–507.

[126] G. F. Yeo and B. Weesakul. 'Delays to road traffic at an intersection'. In: *Journal of Applied Probability* 1 (1964), pp. 297–310.

[127] X. Zhang. 'The influence of partial constraint on delay at priority junctions'. In: *Intersections without Traffic Signals*. Ed. by W. Brilon. Springer, Berlin, Heidelberg, 1988, pp. 180–196.