THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Uncertain demand prediction for guaranteed automated vehicle fleet performance

Sten Elling Tingstad Jacobsen

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2023

Uncertain demand prediction for guaranteed automated vehicle fleet performance

Sten Elling Tingstad Jacobsen

Copyright © 2023 Sten Elling Tingstad Jacobsen
All rights reserved.

This thesis has been prepared using LaTeX.

Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Printed by Chalmers Digiyaltryck
Gothenburg, Sweden, February2023

To my family and friends

Abstract

Mobility-on-demand (MoD) services offer a convenient and efficient transportation option, using technology to replace traditional modes. However, the flexibility of MoD services also presents challenges in controlling the system. One of the major issues is supply-demand imbalance, caused by uneven stochastic travel demand. To address this, it is crucial to predict the network behavior and proactively adapt to future travel demand.

In this thesis, we present a stochastic model predictive controller (SMPC) that accounts for uncertainties in travel demand predictions. Our method make use of Gaussian Process Regression (GPR) to estimate passenger travel demand and predict time patterns with uncertainty bounds. The SMPC integrates these demand predictions into a receding horizon MoD optimization and uses a probabilistic constraining method with a user-defined confidence interval to guarantee constraint satisfaction. This result in a Chance Constrained Model Predictive Control (CCMPC) solution. Our approach has two benefits: incorporating travel demand uncertainty into the MoD optimization and the ability to relax the solution into a simpler Mixed-Integer Linear Program (MILP). Our simulation results demonstrate that this method reduces median customer wait time by 4% compared to using only the mean prediction from GPR. By adjusting the confidence bound, near-optimal performance can be achieved.

Keywords: Mobility-on-Demand, Travel Demand Uncertainty, Fleet Optimization, Gaussian Process Regression, Stochastic Model Predictive Control, Chance Constraint Optimization, Energy Efficiency

List of Publications

This thesis is based on the following publications:

[A] **Sten Elling Tingstad Jacobsen**, Anders Lindman, Balázs Kulcsár, "A Predictive Chance Constraint Rebalancing Approach to Mobility-on-Demand Services". Submitted to *Elsevier Communication in Transportation Research*, in Jan. 2023.

.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Balázs Kulcsár and Dr. Anders Lindman, for their deep knowledge, unwavering support, guidance, and mentorship throughout the first half of my PhD journey. Their insightful discussions and encouragement have greatly impacted my growth as a researcher. I am also grateful to my manager, Jonas Henningsson, for his mentorship and support. Finally I would like to thank my colleagues at Volvo Cars and Chalmers for creating a positive, fun and intellectually stimulating work environment.

Acronyms

MoD:	Mobility-on-Demand
AMoD:	Autonomous Mobility-on-Demand
CCMPC:	Chance Constrained Model Predictive Control
GPR:	Gaussian Process Regression
SMPC:	Stochastic Model Predictive Control
MILP:	Mixed-Integer Linear Program
LP:	Linear Program

Contents

Abstract i

List of Papers iii

Acknowledgements v

Acronyms vi

I Overview 1

1 Introduction 3
1.1 Motivation 3
1.2 Contributions 6
1.3 Thesis outline 6

2 Mobility-on-Demand simulation 9
2.1 Mobility-on-Demand 9
Operational Policies 11
2.2 Performance Metrics 15
2.3 Mathematical models for transportation networks 15
Graph model 16

Closed queueing network models . . . 16
Continuum model . . . 17
2.4 Travel demand prediction methods . . . 17
Parametric models . . . 17
Non-Parametric models . . . 18
2.5 Fleet Control Strategies . . . 19
2.6 Transportation simulators . . . 20

3 Modelling, prediction, and control **23**
3.1 Graph Network Model . . . 23
Vehicle Conservation and Imbalance Dynamics . . . 24
3.2 Gaussian Process regression . . . 26
3.3 Chance Constrained Optimization . . . 29
Chance Constraint Model Predictive Control . . . 30
Mixed Integer Linear Program . . . 31
Totally Unimodular Problems . . . 32

4 Summary of included papers **35**
4.1 Paper A . . . 35

5 Concluding Remarks and Future Work **37**
5.1 Concluding Remarks . . . 37
5.2 Future Work . . . 38

References **41**

II Papers **49**

A CCMPC **A1**
1 Introduction . . . A3
2 AMoD Modeling . . . A7
2.1 Model . . . A7
3 Model predictive control of AMoD with probabilistic guarantees A10
3.1 Model Predictive control of AMoD . . . A10
3.2 Chance Constrained MPC . . . A12
3.3 Separable Model . . . A12

3.4 Gaussian Processes Regression (GPR) for Time-Series Modelling A13
3.5 Chance Constraint MPC (CCMPC) with GPR A16
3.6 Minimal fleet size A17
3.7 Algorithm A18
4 Case Study A18
4.1 Simulation Environment A19
4.2 Travel Demand Prediction A19
4.3 Minimal fleet size for different confidence level A21
4.4 Confidence bound in CCMPC A21
4.5 Comparative evaluation of AMoDs A23
4.6 Computational Complexity A25
5 Conclusion and Future Work A27
1 Total Unimodular A31
References A31

Part I

Overview

CHAPTER 1

Introduction

1.1 Motivation

Transportation is an essential aspect of society that significantly impacts the quality of life. Over the past century, advancements in transportation have significantly improved connectivity between different societies and facilitated the growth of cities. Today, access to reliable transportation is essential for many individuals to lead fulfilling lives. However, many challenges still need to be addressed to reduce the negative effect of transportation and meet the needs of growing societies and cities.

One of the key challenges in transportation today is the reduction of greenhouse gas emissions. Road transport was responsible for approximately one-fifth of EU greenhouse gas emissions, with passenger cars being the most significant contributor at 61% [1]. To decrease the environmental impact of vehicles, it is necessary to consider the vehicle's entire lifecycle, from production to recycling. The question is, *how can we design, manufacture, and operate vehicles in a way that is the most energy and resource efficient?* Volvo cars is taking the lead as a car manufacturer by setting a goal to be a climate-neutral company by 2040 and that every new Volvo model will be pure electric

by 2030.

Another challenge is the rapid growth of population and urbanization. This has led to longer commuting times and increased congestion, resulting in increased pollution and strain on transportation infrastructure [2]–[4]. In response, many cities have implemented policies aimed at reducing the use of private cars, such as limiting parking spots, increasing parking fees, and implementing congestion charges [5], [6]. To address these challenges, there is a need for continued technological advancements and innovation in the transportation industry. This includes the development of more efficient and sustainable vehicles, integrating new technologies such as autonomous vehicles, and exploring new modes of transportation such as shared mobility services.

In recent years, the development of connectivity has led to the emergence of new forms of transportation, particularly in mobility-on-demand (MoD) services. MoD is a service in which shared vehicles are used for passenger travel. According to the US Department of Transportation, MoD is defined as a commodity that enables users to access transportation services as needed rather than owning a vehicle [7]. MoD services are adaptable and convenient and are not fixed to a schedule and route like traditional public transportation services such as trains, buses, trams, and metros, which make MoD popular to use. In 2019 the shared mobility market accounted for approximately 150 billion dollars in global consumer spending and more than 40 million daily users [8]. Today, MoD mainly consists of car-sharing (Volvo-on-Demand, Miles, Kinto) and ride-sharing (Uber, Didi). The increasing use of ride-sharing has caused a negative impact on congestion. Erhardt et al. showed in a study that ride-sharing increased the weekday vehicle hours of delay with 66% between 2010 and 2016 in San Francisco compared to a 22% increase without ride-sharing [9].

Another negative effect is the percentage of milages without customers, called empty distance. A study of ride-hailing data from Austin, Texas, from December 31, 2016, to March 31, 2017, showed that between 51.4% and 66.3% of the total mileages were without customers. This is related to one of the significant challenges with mobility-on-demand systems, which is that the pickup and dropoff locations are unevenly spread causing a supply-demand imbalance if left unattended. Vehicles hence accumulate in one part of the city where few ride-requests are causing long pickup times and a higher percentage of empty mileage. It is therefore important to have a fleet management that

considers these imbalances and proactively sends vehicles to areas with high demand-supply imbalance, called rebalancing, to decrease ride-sharing's negative impacts.

The development of self-driving vehicles is expected to greatly impact MoD. The benefits of autonomous cars are that they can rebalance themselves, are cost-effective and can be centrally controlled without uncontrollable behaviour from drivers. Hence, it is possible to use new centralized control algorithms that minimize both demand-supply imbalance and operational cost. So-called autonomous MoD (AMoD) can be a part of public transportation in cities and rural areas. Unlike traditional public transportation, the operation of AMoD is much more adaptable: it is not restricted to a timetable or specific line. It hence can improve the public transportation network where it is less effective. As an example, the automation and electrification of MoD is predicted to reduce the operational cost by as much as 84% in Berlin and 70% in Austin compared to the operational cost of taxis today [10]. Robotaxis has gained much interest in the US, with several actors operating test fleets of automated vehicles in California, Nevada and Arizona. For example, Waymo is now operating robotaxis in San Francisco and Phoenix, with over 20 million miles driven in Phoenix from 2019 to 2020 [11].

Many questions still need to be answered related to how AMoD should be operated and what impact these services will have on different cities. How can we use new technologies to reduce the negative aspects of MoD services while keeping and improving their positive effects? It is necessary to compare different mobility alternatives, what features are essential for operating vehicles and what infrastructure is needed. Furthermore, it is important to consider factors such as sustainability, accessibility, affordability, and the potential to reduce congestion and emissions. To be able to access future technology requires the development of new methodologies, models and simulation environments. To test these scenarios in real cities is neither cost-effective nor practical. One important aspect to consider when evaluating these systems is that they are stochastic dynamical systems. In this thesis, the focus has been on developing models and methods for operational aspects of ride-hailing services that consider uncertainty in stochastic travel demand predictions. The thesis investigates the following research questions:

1. What kind of models can we use to describe ride-hailing systems?

2. Shall we optimize AMoD for passenger wait time or overall service cost minimization?

3. How to handle travel demand prediction and embed them into the model?

4. What results can we have with high fidelity traffic simulator (sensitivity, robustness)?

1.2 Contributions

In order to answer the research questions above we focus the thesis to:

- Gaussian process regression is utilized to predict complex, non-linear travel demand.
- A stochastic model predictive controller is formulated using chance constraint optimization. The proposed chance-constrained model predictive controller (CCMPC) incorporates data-driven travel demand prediction using Gaussian process regression (GPR) and accounts for uncertainty in travel demand.
- The chance constraint is relaxed using a separable model, resulting in a Totally Unimodular MILP. This guarantees that the optimal solution of the integer-relaxed LP is always an integer and the MILP can be efficiently solved in polynomial time.
- A realistic case study is tested with the high-fidelity transportation simulator AMoDeus ([12]) to quantify the impact of the proposed model and controller on ride-hailing services in San Francisco (SF).
- A method to evaluate the minimal fleet size for different probabilistic guarantees on the service level and empirical results for the SF case study.

1.3 Thesis outline

The thesis starts with an introduction to the concept of Mobility-on-Demand (MoD) and an overview of control models and methods in Chapter 2. The

study's model, which includes the application of Gaussian Process Regression, is presented in Chapter 3, followed by the introduction of the CCMPC control method. Chapter 4 provides a summary of the paper's key points. Lastly, Chapter 5 concludes the thesis and highlights potential areas for future research.

CHAPTER 2

Mobility-on-Demand simulation

This chapter covers the concept of Mobility-on-Demand services and key factors to consider in simulating these services. It provides an overview of transportation network models, a literature review of fleet control and travel demand prediction methods. The chapter concludes with a brief review of various transportation simulation tools.

2.1 Mobility-on-Demand

Mobility-on-demand services (Mod) offer a wide range of options for individuals to access transportation. Autonomous Mobility on Demand (AMoD) is a mobility service utilizing autonomous vehicles. The operation of a traditional MoD service and AMoD are equivalent if it is assumed that drivers in MoD services follow instructions. There exist many different definitions of MoD but in this thesis, we consider that MoD can be divided into three subcategories: ride-hailing, ride-pooling, and car-sharing.

Ride-hailing is a service where an individual can request a vehicle to pick them up and take them to their desired destination. Uber, Bolt and Didi are example of ride-hailing services that exist today.

Ride-pooling is similar to ride-hailing, but instead of having a vehicle exclusively for an individual, multiple individuals share the same vehicle. Ride-pooling may have a lower service level with longer wait and travel times than traditional ride-hailing, but it is often more affordable. One study on ride-pooling, where at most two passengers can share the same ride, showed that the pickup time is similar to ride-hailing but with an 18% shorter total distance driven and 27% longer drive time [13].

Car-sharing is a service where an individual can book a vehicle to drive, often with the requirement that the vehicle is collected and returned to the same location. Car-sharing services, such as Volvo-on-Demand, allow customers to rent cars for short periods of time, by the hour or by the day.

It's important to notice that this text is limited to on-demand ride-hailing services, which have become the most popular among the subcategories.

Simulating mobility scenarios involves considering six key aspects to achieve an accurate representation, as shown in Fig. 2.1, which are:

- **Fleet Control Strategies**: Mobility-on-Demand services require real-time control of the vehicle fleet. This includes both rebalancing and dispatching of vehicles, charging strategies for battery electric vehicles (BEVs), and how vehicles should be routed.
- **Infrastructure**: Information about the service critical infrastructure, including where charging stations should be built and what power should they be able to deliver and possible expansion of power grid network due to increased electricity demand.
- **Fleet Specifications**: Determining the necessary fleet size to offer a good service and keep operating costs low is essential. This includes considering properties such as battery size, charging capacity, vehicle size, and vehicle-to-grid capability when selecting the types of vehicles to use in the fleet.
- **Road Network Properties**: The properties of the street network, such as travel times (deterministic or stochastic), travel distance, and static or dynamic traffic, should be taken into account.
- **Travel Demand Model**: A model that considers spatio-temporal stochasticity, mode choice, and decision-making processes.

- **Service Attributes**: Important service attributes, such as pricing and pick-up time constraints, should also be considered.

The focus of this thesis is on fleet control strategies, fleet specifications, and travel demand models, while acknowledging the importance of incorporating all relevant factors for a complete simulation of mobility scenarios. To achieve this, high-fidelity transportation simulators are utilized to account for various aspects."

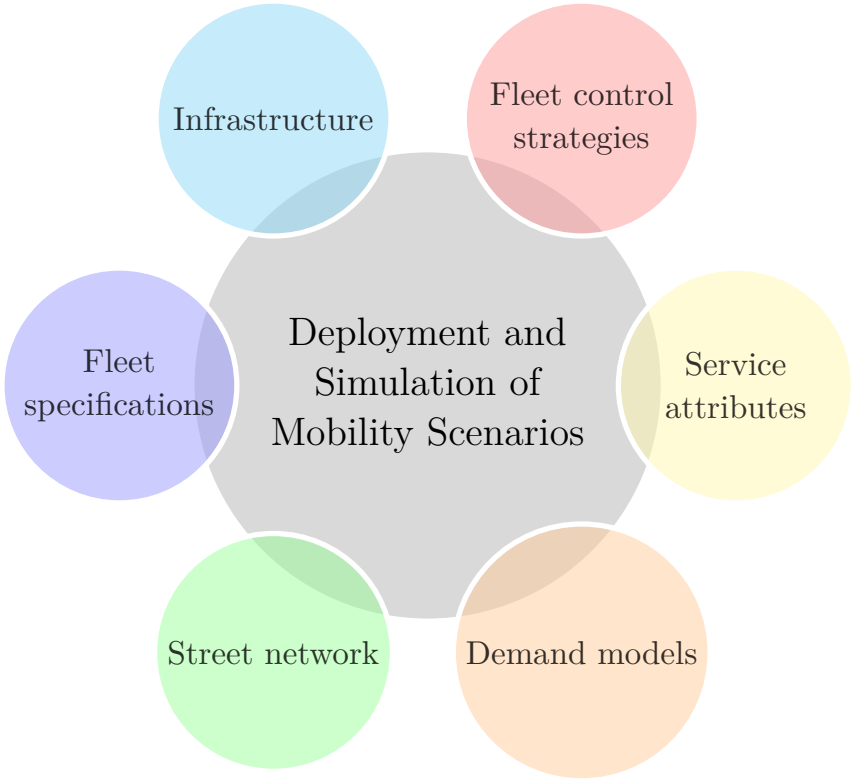

Figure 2.1: Six aspects that are important to consider when simulating mobility scenarios [14].

Operational Policies

When developing operational policies for MoD there are four different aspects that need to be considered: routing, dispatching, rebalancing, and constraints such as charging and limited driving range. While dispatching and rebalancing determine the destinations for the vehicles, routing involves finding the optimal route to reach those destinations.

Routing is the process of determining the most efficient and optimized route for a vehicle to travel from its current position to its destination. It is a necessary step after deciding on dispatching and rebalancing, as it ensures that the vehicle arrives at its destination in a timely and efficient manner. There are various methods and algorithms that can be used to optimize the routing of vehicles, depending on the specific needs of the transportation service. These methods take into account various metrics such as distance, travel time, and fuel consumption. Some routing algorithms also consider more complex factors such as traffic conditions, weather, and road closures to provide the most efficient routes. Additionally, many routing algorithms also incorporate real-time data and machine learning techniques to adapt to changing traffic conditions and provide dynamic routing. This helps to reduce travel time, increase the availability of vehicles, and improve the overall customer experience.

Dispatching of vehicles is a vital process in the transportation industry. It involves assigning vehicles to customers in real time. The goal of dispatching is to match the right vehicle with the right customer at the right time while also considering cost, distance and time factors. Dispatching can be viewed as an assignment problem, where the supply is the vehicles and the demand is the customer. Various methods can be used to find an optimal dispatching solution, given a set of vehicles and requests. One such method is the Hungarian method [15]. This method is known as a reactive method, meaning that it can only consider the current state of the system and not future predictions or trends.

Charging scheduling refers to the process of planning and coordinating the charging of electric vehicles in a fleet. It is an essential aspect of fleet management for electric vehicles since it ensures that the vehicles have enough energy to meet the transportation needs of customers. The charging scheduling can either be incorporated as a constraint in the rebalancing optimization or as a separate optimization problem. When charging scheduling is incorporated as a constraint in the rebalancing optimization, it ensures that the vehicles are charged and ready for their next trip. On the other hand, when charging scheduling is treated as a separate optimization problem, it involves determining the most efficient and cost-effective way to charge the vehicles, taking into account factors such as the availability of charging stations, electricity prices, and the energy needs of the vehicles.

Rebalancing is the proactive process of repositioning vehicles to areas of predicted high demand, to ensure that vehicles are always available to meet customer needs. It complements the dispatching process by being proactive rather than reactive, see Fig. 2.2. Typically, the rebalancing process consists of two steps:

1. Prediction of future demand: This step involves using data analysis, machine learning, and other techniques to predict the areas of high demand for vehicles in the future.

2. Optimization of rebalancing: Based on the demand prediction, this step involves determining the most efficient and cost-effective way to reposition vehicles to meet that predicted demand. This may involve optimizing routes, identifying the most suitable vehicles for a particular area, and scheduling the movement of vehicles in advance. Rebalancing can help to reduce empty mileage, wait times for customers, and overall operational costs. It also helps to improve the overall customer experience and the efficiency of the transportation service.

An example of why proactive rebalancing is beneficial can be seen in Fig. 2.2. The figure compares a reactive controller (left figures) and a predictive controller (right figures) using a simple example. The city or rural area has been divided into 9 stations, labeled A through I. At the initial state of the system, there is one customer at station A and two vehicles at stations C and D, respectively. The reactive controller finds the optimal static solution, which is to send the vehicle at station D to pick up the customer at station A. However, in the next time step, a customer at station G appears and the only available taxi is at station C. The predictive controller, on the other hand, first predicts future travel demand and then makes an optimal decision based on that prediction. As a result, the vehicle at station D is sent to pick up the customer at station G, while the vehicle at station C is sent to the customer at station A.

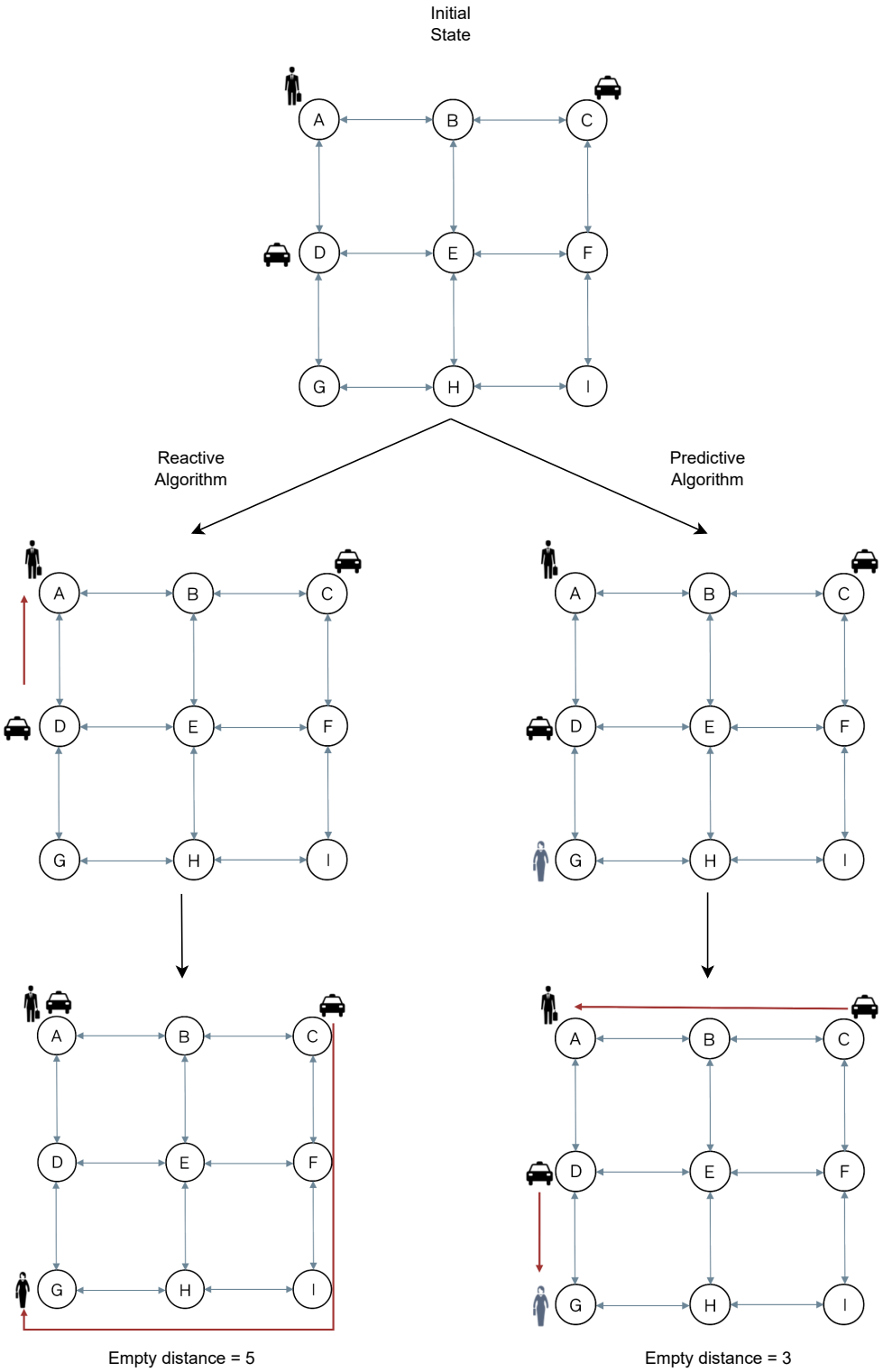

Figure 2.2: A comparison between a reactive (left figure) and a predictive controller (right figure) provides a simple illustration of the benefits of the latter. The predictive controller considers future states, enabling it to effectively minimize empty distance travel and overall pickup time.

2.2 Performance Metrics

Improving the efficiency of MoD fleets necessitates evaluating multiple crucial metrics, including waiting and travel time, costs and revenues, travel distance, demand satisfaction, fleet size, and societal factors. From an operator's perspective, it's essential to minimize the total cost of the fleet by optimizing utilization and reducing empty distance travel. From the customer's standpoint, wait time is a critical factor, but travel time, traffic congestion, and pollution levels should also be taken into account. These metrics are interdependent, and striking the right balance can be challenging. For example, reducing wait time might require a larger fleet or increased empty distance travel, while minimizing travel time and reducing pollution might result in higher costs. The optimal is to consider all aspects and tune the weighting of the different aspects to reach a equilibrium, see Fig. 2.3.

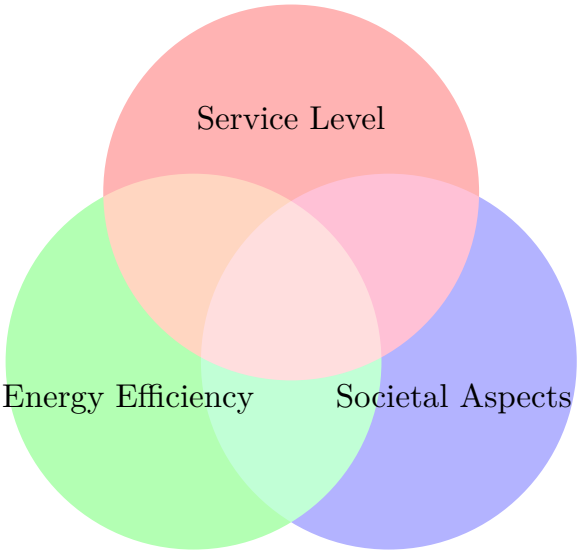

Figure 2.3: Performance metrics that are important to consider when operating MoD services.

2.3 Mathematical models for transportation networks

To plan the service ahead we need a model that models the dynamics of the transportation network. Three main models have been used to model trans-

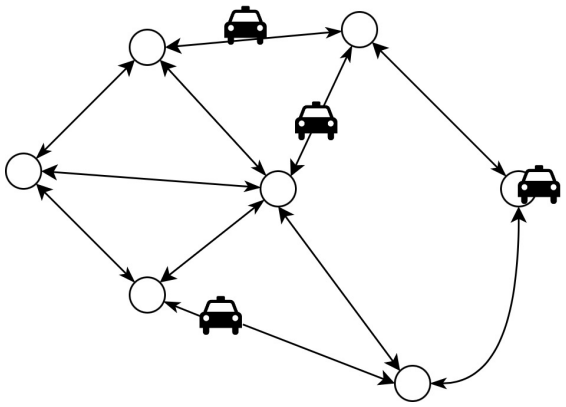

Figure 2.4: Graph model of MoD services.

portation networks for fleet control: graph models, queueing-based models and continuous models.

Graph model

The road network can be represented by a directed graph $\mathcal{G}\langle\mathcal{V},\mathcal{E}\rangle$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of edges, see Fig. 2.4. The granularity of the graph model can vary depending on the level of detail required.

At the most detailed level, each node $v_i \in \mathcal{V}$ represents a specific road section, and edges $e_{ij} \in \mathcal{E}$ represent the connections between these sections, represented by $e_{ij} = \langle v_i, v_j \rangle$. However, in some cases, nodes may represent an area of the city and edges represent the connections between these areas.

Additionally, each edge e_{ij} has an associated travel time and distance, which can be used to model the dynamics of transportation on the network. This information can be used to optimize the transportation network, such as finding the shortest route between two locations or determining how to allocate resources to improve network performance. An advantage with graph models is that they can be used in MPC formulations, [16]. One disadvantage with graph models is that often individual vehicles are not modeled.

Closed queueing network models

Queuing theory is a way of modeling operational systems with waiting lines. For AMoD systems, the operational area is partitioned into N areas. All of the areas are assumed to be indirectly connected, and the network is assumed to

be closed. In a closed queueing network, the number of vehicles in the system remains constant as no vehicles can enter or exit, creating a closed system. In each area, there is a queue of vehicles waiting to serve customers. Customers arrive at each station according to a stochastic process, such as a Poisson process, and their destinations are determined by a probability distribution. If there are available vehicles in the queue, they are assigned to customers. If not, unserved customers can either disappear (passenger loss model) or accumulate in a queue (passenger queue model). So far, only passenger loss models have been implemented ([17]–[22]), which is a disadvantage compared to graph models. The travel time between areas is also modeled as a stochastic variable, often exponentially distributed. To improve performance, vehicles can be proactively sent between different queues to prevent passenger loss. The key performance metric for queuing models is the availability of at least one vehicle in each queue. An advantage of closed queueing network models is that individual vehicles can easily be modeled.

Continuum model

In a continuum model the vehicles move freely in a bounded operational area of the real plane, $\Omega \in \mathbb{R}^2$. Where roads are located and how they are connected are neglected, which simplifies the control with such systems. The request in such systems is often modeled as a stochastic process. A continuum model was used in [23] where a solution to the Stacker Crane Problem, which is the problem of routing vehicles to a set of one-to-one travel requests, was proposed.

2.4 Travel demand prediction methods

An important aspect when controlling MoD is the travel demand predictions and several different prediction methods have been proposed in the literature. The different methods can be split into parametric and non-parametric models.

Parametric models

Parametric models assume that the underlying probability distribution is known [24]. One of the most widely used parametric travel demand mod-

els is the Poisson Process model. It has been used in several papers which use closed queueing network models [17]–[22]. This model is based on the Poisson distribution, a statistical tool commonly used to describe the probability of a certain number of events occurring within a given time period. In the context of travel demand modeling, the Poisson Process model represents the number of trips individuals make within a given time period, such as a day or an hour. Parametric models are effective for long time windows where the travel demand is averaged out but less effective for short time windows due to increased randomness in travel patterns.

Non-Parametric models

The travel demand often follows an unknown spatio-temporal probability distribution that can be very complex. Therefore, the parametric models might not model the travel demand correctly. This is one of the reasons why the research has shifted towards non-parametric models. Non-parametric models, unlike parametric models, do not rely on any specific assumptions about the underlying probability distribution of the data. Instead, non-parametric models use techniques such as machine learning and data mining to learn patterns and relationships in the data. This makes non-parametric models more flexible and better suited for data that may not follow a specific probability distribution.

One example of a non-parametric travel demand model that has been used is Long Short-Term Memory (LSTM) neural network [25], [26]. An LSTM neural network is a type of recurrent neural network that is able to process and predict time series data [27]. Other neural networks, such as Multi-Graph Convolution Networks, have also been proposed for travel demand prediction [28]. An advantaged of neural networks is that factors such as weather and other events can be included in the training.
Another example is Gaussian Process Regression (GPR) which is a type of Bayesian non-parametric model [24]. GPR has been shown to be an accurate model for time-series prediction [29]. One of the key advantages of GPR is that it can provide not only a point estimate of the function but also a measure of uncertainty in the form of a probability distribution over the function. This can be useful in applications where it is important to have a good understanding of the model's uncertainty, such as in control systems.

One challenge with non-parametric models is that their interpretation can

be difficult, and their performance relies heavily on careful hyperparameter tuning.

2.5 Fleet Control Strategies

There has been extensive research on various algorithms for modelling and controlling Mobility-on-Demand (MoD) and Automated Mobility-on-Demand (AMoD) systems. Several review papers have been written on topics related to AMoD, ranging from routing ([30]–[32],) to control of AMoD ([33]–[37]). This section will summarise what has been done in the field of AMoD control and identify research gaps.

Early research focused on reactive control methods, such as the Hungarian method [15], which aimed to solve static assignment problems. Later, research shifted to solving dynamic assignment problems using Markov Decision Processes, and adaptive learning algorithms [38]. Model Predictive Control (MPC) was subsequently proposed for dispatching and rebalancing vehicles [16], but initial implementations assumed known travel demand and were not scalable for large fleet sizes. To deal with the scalability issues queuing models were proposed, [17]–[22]. Queueing models necessitate simplification of the operation map through partitioning into smaller neighborhoods, as employed in many subsequent papers. The queueing models are suitable for real-time control, but these models relied on parametric demand models, and only passenger loss models were considered. At the same time as the first queue model was proposed a model-free adaptive controller was suggested [39]. The idea behind the controller is to send one extra vehicle to recent service request locations. The controller performs well in terms of service quality but is expensive to operate because of high percentages of rebalancing. More recent research has focused on MPC and developing better travel demand prediction methods, using techniques such as Long Short-Term Memory (LSTM) neural networks and Sample Average Approximation (SAA) [25], [26]. The importance of accounting for uncertainty in demand predictions has also been recognized, with methods such as robust optimization [40], [41] and distributionally robust optimization [42] proposed. However, these methods require large datasets and neglect demand predictions. In recent years, several different reinforcement learning (RL) methods have been proposed [43]. Each taxi is considered an agent, and the reward function can be trip fare ([44]), or

total operational costs ([45]). There have been different ways of representing the transportation network in RL. The most common representation is using a grid system of the network, using hexagonal or squares ([46], [47]), but also graph models of the entire transportation network have been used ([48]). All of the mentioned RL articles use value-based learning algorithms, which use a value function to estimate the expected future reward of a given action in a specific state. The benefit of the RL algorithms is that they can operate in real-time, but one drawback is that they require large datasets for training. Missing in the existing literature are algorithms that include uncertainty in travel demand predictions that do not require large datasets.

2.6 Transportation simulators

Transportation simulators are tools used to model and analyze the behavior of transportation systems. They can be classified into three main categories: microscopic, macroscopic, and mesoscopic simulators.

Microscopic simulators focus on the individual behavior of vehicles and pedestrians in a transportation system. These simulators model the interactions between vehicles, including lane changes, merging, and overtaking, as well as the interactions between vehicles and pedestrians. Microscopic simulators are typically used to study traffic flow at intersections, on-ramps, and other areas of congestion, [12], [49]–[51].

Macroscopic simulators, on the other hand, focus on the overall behavior of a transportation system. These simulators model the flow of traffic on a large scale, rather than individual vehicles. They use aggregate measures such as traffic volume and density to represent the transportation system. Macroscopic simulators are typically used to study the impacts of new infrastructure or policy changes on traffic flow.

Mesoscopic simulators fall between microscopic and macroscopic simulators. They model the behavior of groups of vehicles, rather than individual vehicles. Mesoscopic simulators use a combination of aggregate measures and individual vehicle interactions to represent the transportation system. They are commonly used to study the impacts of intelligent transportation systems (ITS) such as traffic signal control and variable message signs.

Mesoscopic transportation simulators can be used for a wide range of transportation planning and research applications, including traffic engineering,

transportation policy analysis, and the evaluation of new transportation technologies.

In this thesis we have chosen to work with AMoDeus (Autonomous Mobility on Demand Simulator) which is an open-source simulation platform that is specifically designed to model and analyze AMoD systems, [12]. One of the key strengths of AMoDeus is its ability to simulate large-scale, realistic AMoD systems in a highly accurate and detailed manner. It can be used to model various components of an AMoD system, such as the vehicles, charging infrastructure, and the transportation network. Additionally, it allows to simulate different scenarios, such as changes in land use, population growth, and technological advancements on AMoD systems.

AMoDeus can be used for a wide range of transportation planning and research applications, including evaluating the impacts of transportation policies, testing the performance of different AMoD systems, and assessing the feasibility of new autonomous technologies. It can also be used to compare different mobility alternatives, what features are important for operating vehicles and what infrastructure is needed.

CHAPTER 3

Modelling, prediction, and control

In this chapter, we first introduce the modelling framework for control of MoD. The model represents the dynamics of movement of vehicles and passengers in the transportation network. Second, we present Gaussian process regression (GPR), which is used for travel demand prediction. Lastly, we formulate a Chance Constrained Model Predictive Control (CCMPC) which gives probabilistic guarantees on the service level.

3.1 Graph Network Model

A graph model of a city consist of hundred of thousands of edges and vertices. This makes the model very complex and earlier studies have shown that considering individual vehicles in these types of models are not scalable [16]. Therefore, simplification of the transportation graph model have been proposed where the urban/rural area is split into larger areas, which we will call stations. The stations are represented by vertices v_i and the connection between the different stations are represented by edges e_{ij}. The graph is complete, meaning that all stations are connected to each other, see Fig. 3.1. Instead of modeling the position of every vehicle and passengers we model

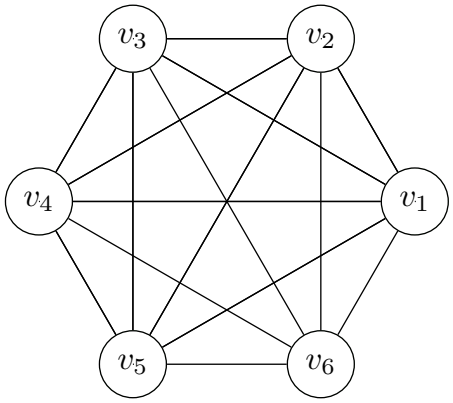

Figure 3.1: A city partitioned into $\mathcal{N} = 6$ smaller areas, where the vertices are stations and edges are paths between stations.

the number of vehicles and passengers in each station. This is an important property because the number of states doesn't depend on the fleet size nor the number of passenger which makes it suitable for control of large fleets. The driving time and the energy consumption in-between stations i and j are denoted by κ_{ij} and σ_{ij}, respectively. The model is discrete-time model with time intervals Δt. $\mathcal{T}$ denotes the set of periods, $\mathcal{T} = [1, .., T]$, where T is the number of periods.

Vehicle Conservation and Imbalance Dynamics

We first introduce the state and decision variables in the MoD model and then the MoD system dynamics. All states and decision variables are non-negative integer values and are presented in the list below:

- $\lambda_{ij}(t)$ - the number of passengers that wants to go from origin i to destination j.
- $x^c_{ijb}(t)$ - the number of vehicles that are driving passengers from station i to station j at time t.
- $x^r_{ij}(t)$ - the number of rebalancing vehicles in-between station i and station j at time t.
- $s_{ij}(t)$ - A decision variable for the imbalance, which describes how many customers to not pick-up at time t that wants to go from station i to station j.

Vehicle conservation

The total number of vehicles in the model should be conserved at every time instance. This is enforced with a vehicle conservation constraint for each station,

$$\sum_{j \in N} x^{\mathrm{c}}_{ij}(t) + x^{\mathrm{r}}_{ij}(t) = \phi_i(t) + \sum_{j \in N} x^{\mathrm{c}}_{ji}(t - \kappa_{ji}) + x^{\mathrm{r}}_{ji}(t - \kappa_{ji}), \quad \forall i \in N, t \in \mathcal{T}, \tag{3.1}$$

which force the the vehicles departing station i at time t, $\sum_{j \in N} x^{\mathrm{c}}_{ij}(t) + x^{\mathrm{r}}_{ij}(t)$, to be equal to the initial number of vehicles in station i,$\phi(t)$, plus vehicles entering the station in time interval t, $\sum_{j \in N} x^{\mathrm{c}}_{ji}(t - \kappa_{ji}) + x^{\mathrm{r}}_{ji}(t - \kappa_{ji})$. The constraint keeps track of where vehicles are in the model and travel times in between stations, κ_{ij}. In the model presented in this thesis the travel times are assumed to be constant.

Imbalance

The imbalance dynamics models the difference between passenger requests and vehicles in each station. Ideally, the imbalance is zero at all times, i.e. there is a perfect match between the number of travel demand and vehicles,

$$\lambda_{ij}(t) - x^{c}_{ij}(t) = 0, \forall i, j \in N, t \in \mathcal{T}. \tag{3.2}$$

However, if there are more customers than vehicles, constraint Eq. (3.2) is violated. Therefore this constraint needs to be relaxed to ensure feasibility, which is done by introducing the slack variable $s_{ij}(t)$,

$$s_{ij}(t) = \lambda_{ij}(t) - x^{\mathrm{c}}_{ij}(t) \quad \forall i, j \in N, t \in \mathcal{T}. \tag{3.3}$$

If $s_{ij}(t) > 0$, i.e. there are more request then available vehicles, the remaining request should be served at a later time step. Hence, we carry on $s_{ij}(t)$ to the next time step if $t > t_0$,

$$s_{ij}(t+1) = s_{ij}(t) + \lambda_{ij}(t+1) - x^{\mathrm{c}}_{ij}(t+1) \quad \forall i, j \in N, t \in [t_0 + 1, T + t_0], \tag{3.4a}$$

$$s_{ij}(t_0) = \lambda_{ij}(t_0) - x^{\mathrm{c}}_{ij}(t_0) \qquad \forall i, j \in N, \tag{3.4b}$$

The state $x_{ij}^{c}(t)$ cannot be larger then the number of travel request since it represent only vehicles that drives customer, i.e. the imbalance should be greater or equal to zero,

$$s_{ij}(t) \geq 0, \quad \forall i, j \in N, t \in [t_0, T + t_0]. \tag{3.5}$$

The combination of constraint Eq. (3.5) and that $s_{ij}(t)$ is a integer decision variable, gives that

$$s_{ij}(t) \in \mathbb{N}, \quad \forall i, j \in N, t \in [t_0, T + t_0].$$

3.2 Gaussian Process regression

Gaussian Process Regression (GPR) is a type of machine learning algorithm used for regression and probabilistic classification. It models the relationship between input features and output targets as a Gaussian distribution, allowing for the prediction of output values and estimation of uncertainty. GPR can be used for time series prediction by modeling the temporal dependencies between consecutive time steps as a covariance function. The covariance function captures the similarity between time steps, allowing for the prediction of future time steps based on the observed historical data. GPR can provide not only point estimates for the future values but also a probabilistic estimate of the uncertainty around the predictions. This makes GPR a powerful tool for time series prediction, especially in cases where the underlying process is complex and non-linear [29].

The concept of GPR can be understood from a functions perspective, also known as the function-space view, as discussed in [24]. Imagine a black box system with input $\mathbf{t}$ and output $\lambda = f(t)$, where $f(t)$ is an unknown function. We have a collection of past input and output data, called the training data set $\mathcal{D} = (\mathbf{t}_i, \lambda_i)|i = 1, ..., n$. There are many functions that could fit this data. GPR utilizes a probabilistic approach to find the best fit among these functions. This is achieved by assigning a multivariate probability distribution to the entire function-space. This distribution allows us to predict with confidence.

The goal of GPR is to find the underlying multivariate distribution based on prior knowledge and a training data set. This distribution is assumed to be a multivariate normal distribution, so the estimated output follows a

normal distribution $\lambda_1, ..., \lambda_n \sim \mathcal{N}(\mu(\mathbf{t})_{i,..,n}, \mathbf{\Sigma})$, where $\mathbf{\Sigma_{i,j}} = Cov(\lambda_i, \lambda_j) = k(t_i, t_j)$ is the covariance function (kernel) and $\mu(\mathbf{t})$ is the mean function. Thus, the Gaussian process is defined by its mean and covariance functions $f(t) \sim \mathcal{GP}(\mu(\mathbf{t}), k(\mathbf{t}, \mathbf{t}'))$. Kernels are selected based on prior knowledge of the data, such as smoothness (e.g. radial basis function kernel) or periodicity (e.g. periodic kernel). If we assume a smooth function the radial basis function kernel (RBF) can be used

$$k_{\text{RBF}}(\mathbf{t}, \mathbf{t}') = \exp\left(- \frac{\|\mathbf{t} - \mathbf{t}'\|^2}{2l^2} \right), \tag{3.6}$$

where l is the lengthscale hyperparameter. If the data is periodic a periodic kernel is proposed

$$k_{\text{Periodic}}(\mathbf{t}, \mathbf{t}') = \exp\left(- 2\frac{\sin^2\left(\frac{\pi}{p}(\mathbf{t} - \mathbf{t}')\right)}{l^2} \right), \tag{3.7}$$

where p is the period and l the lengthscale hyperparameter. The sum and multiplication of two kernels are also kernels [24].

When the kernels have been selected the hyperparameters are trained on the dataset by maximizing the log-marginal likelihood [24]. The log marginal likelihood is given by

$$\log p(\mathbf{y}|X, \theta) = -\frac{1}{2}\mathbf{y}^\top \mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{n}{2}\log(2\pi), \tag{3.8}$$

where $\mathbf{\Sigma}$ is the covariance matrix,

$$\mathbf{\Sigma_{n,n}} = \begin{pmatrix} k_{1,1} & k_{1,2} & \cdots & k_{1,n} \\ k_{2,1} & k_{2,2} & \cdots & k_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n,1} & k_{n,2} & \cdots & k_{n,n} \end{pmatrix}. \tag{3.9}$$

A gradient method is used to find the hyperparameters that maximizes the log marginal likelihood, i.e. the partial derivatives of Eq. (3.8) with respect to the hyperparameters are computed:

$$\frac{\partial}{\partial\theta_i}\log p(\mathbf{y}|X, \theta) = -\frac{1}{2}\mathbf{y}^\top \mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial\theta_i}\mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\text{tr}(\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial\theta_i}) \tag{3.10}$$

The computational complexity of GPR training mainly arises from the need to invert the covariance matrix $\mathbf{\Sigma}$, which has a computational complexity of $\mathcal{O}(n^3)$. Once the kernels and mean function have been tuned, future predictions can be made using the conditional probability of the posterior distribution.

$$\hat{\mu}(t^*) = \mathbf{k}_*{}^\top \mathbf{\Sigma}^{-1}\mathbf{y}, \tag{3.11}$$

$$\hat{\sigma}^2(t^*) = k(t_*, t_*) - \mathbf{k}_*{}^\top \mathbf{\Sigma}^{-1}\mathbf{k}_*, \tag{3.12}$$

where $\mathbf{\Sigma}$ is the covariance matrix for the training data, $\mathbf{k}_*$ is the vector of covariances between t^* and the n training points, where $\mathbf{\Sigma}_n$ is the noise matrix. The predictive mean is used as the estimated output and the predictive covariance provides a measure of uncertainty in the prediction.

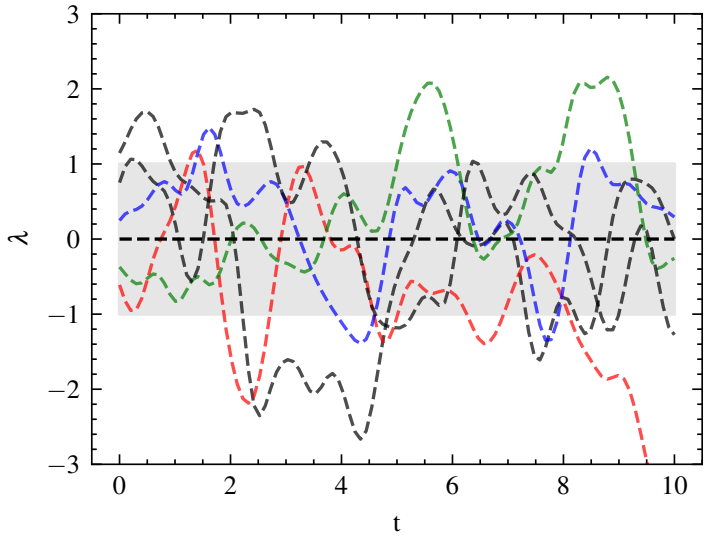

Figure 3.2: Samples from prior distribution of the locally periodic kernel.

Locally periodic kernels (which are considered in Paper 1) are a combination of RBF and periodic kernels. The periodic kernel assumes perfect correlation between data points separated by Np units, i.e. $\mathbf{t} - \mathbf{t}' = Np$, where N is an integer. However, this strict periodicity assumption is not applicable to most stochastic functions, including travel demand data, which exhibit some periodicity but not in a strict manner, such as daily commuter patterns that vary in time and extent. The locally periodic kernel allows the periodic component

to change over time, making it a more appropriate choice for travel demand prediction. Fig. 3.2 illustrates samples from the locally periodic kernel's prior, highlighting local periodicity with variability over time.

3.3 Chance Constrained Optimization

Recent advancements in convex optimization and improvements in computing power have made it possible to efficiently solve large optimization problems. This, along with the ability to collect and store large amounts of data, has led to a shift from traditional to data-driven optimization methods.

Different methods exist for solving convex optimization problems with uncertainty, including robust optimization, which can be formulated as follows:

$$\begin{aligned} \underset{x \in X}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g(x, \xi) \leq 0 \quad \forall \xi \in \Xi \end{aligned} \tag{3.13}$$

where ξ is a random variable. In robust optimization, all possible cases are considered, which may result in a conservative solution. Chance constraint optimization is a method that only considers a certain percentage of the random variable ξ by providing a probability guarantee to the constraint. This is formulated as:

$$\begin{aligned} \underset{x \in X}{\text{minimize}} \quad & c^\intercal x \\ \text{subject to} \quad & \mathbb{P}(h(x, \xi) \leq 0) \geq 1 - \epsilon \end{aligned} \tag{3.14}$$

where the probability of the constraint being fulfilled is greater than $1 - \epsilon$. The concept of chance constraint optimization was introduced by Charnes and Cooper in 1959 [52]. However, computing the probability of uniformly distributed variables is an NP-hard problem, making CCO programs computationally intractable and in need of approximation [53].

Distributional robust optimization is a method that reforms the chance constraint into a tractable form. By using known information about the distribution $\mathbb{P}$, the goal is to minimize the set of possible distributions, called the ambiguity set $\mathscr{P}$,

$$\begin{aligned} \underset{x \in X}{\text{minimize}} \quad & c^\intercal x \\ \text{subject to} \quad & \mathbb{P}(h(x, \delta) \leq 0) \geq 1 - \epsilon, \forall \mathbb{P} \in \mathscr{P}. \end{aligned} \tag{3.15}$$

If the mean and variance of a distribution are known, there are ways to estimate the probability distribution. The most popular of these is Chebyshev inequality, which was first presented in [54]:

Theorem 1 (Chebyshev inequality[54])**:** *Let $X \in \mathbb{R}$ be a random variable with finite mean μ and finite non-zero variance σ^2. Then for any real number $k > 0$,*

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \begin{cases} \frac{1}{k^2} & k > 1 \\ 1 & otherwise \end{cases} \tag{3.16}$$

Chebyshev inequality states that for any probability distribution, no more than $\frac{1}{k^2}$ of the values can be more than k standard deviations from the mean. This gives a conservative estimate of the distribution, considering all possible distributions in the ambiguity set.

If the stochastic variables enters the chance constraint in an affine way, it is a special case of the chance constraint and is referred to as a separable chance constraint [55]. A separable chance constraint with known probability distribution can be reformulated as a deterministic constraint. The separable chance constraint can be formulates as,

$$\mathbb{P}(h(x) \geq A\xi) \geq 1 - \epsilon, \tag{3.17}$$

where $h(x)$ is a deterministic function and A is a constant matrix describing how the stochastic variable ξ enters the constraint. By using the cumulative distribution function of ξ, $\mathbf{F}_\xi(z) \coloneqq \mathbb{P}_\delta(\epsilon \leq z)$, the separable constraint (Eq. (3.17)) can be simplified to the following deterministic constraint,

$$\mathbf{F}_\xi(h(x)) \geq 1 - \epsilon \tag{3.18}$$

Chance Constraint Model Predictive Control

In this thesis we use GPR to predict stochastic and non-linear travel demand. The uncertainty bounds on the travel demand prediction are used to formulate a chance constraint of constraint Eq. (3.4),

$$\begin{aligned} &\mathbb{P}_{ij}\left(s_{ij}(t+1) = s_{ij}(t) + \lambda_{ij}(t+1) - x^{\mathrm{c}}_{ij}(t+1) \leq k\right) \geq 1 - \epsilon \\ &\qquad \forall i, j \in N, t \in [t_0, T + t_0], \end{aligned} \tag{3.19}$$

where the constant k is an upper bound on the imbalance. The GPR gives a mean prediction, μ, and a confidence bound on the prediction, σ, where the confidence is assumed to follow a Gaussian distribution. We can therefore use the cumulative distribution function for a Gaussian distribution, which is defined as

$$F(1-\epsilon;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{1-\epsilon} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz. \tag{3.20}$$

We can therefore formulate the chance constraint (Eq. (3.19)) as a deterministic constraint,

$$\begin{aligned} &s_{ij}(t+1) + k + x_{ij}^{\mathrm{c}}(t+1) - s_{ij}(t) \geq F^{-1}_{\lambda_{ij}(t+1)}\left(1-\epsilon;\hat{\mu},\hat{\sigma}\right), \\ &\quad \forall i,j \in N, t \in [t_0, T+t_0], \end{aligned} \tag{3.21}$$

To optimize the dispatching and rebalancing of vehicles we formulate the following chance constrained model predictive control (CCMPC),

$$\underset{x_{ij}^{\mathrm{r}}, s_{ij}}{\text{minimize}} \sum_{t=t_0}^{T+t_0} \sum_{j,i=1}^{N} c_{ij}^{\mathrm{r}}(t) x_{ij}^{\mathrm{r}}(t) + c_\lambda(t) s_{ij}(t) \tag{3.22a}$$

subject to

$$Eqs.(3.4\mathrm{b}), (3.19), (3.21) \tag{3.22b}$$

$$x_{ij}^{\mathrm{r}}, s_{ij}, x_{ij}^{\mathrm{c}}(t) \in \mathbb{N} \qquad \forall i,j \in N, t \in [t_0, T+t_0], \tag{3.22c}$$

where $c_{ij}^{\mathrm{r}}(t)$ and $c_\lambda(t)$ are the cost of rebalancing respectively of the cost of imbalance. The objective of the CCMPC is to find the optimal rebalancing strategy that minimizes the rebalance distance (empty milages) and the imbalance (service quality). If the imbalance is kept at zero the pickup time will depend on the size of the stations. More and smaller station will lead to a shorter pickup time.

Mixed Integer Linear Program

The CCMPC Eq. (3.22) is a Mixed-Integer Linear Programming (MILP), which is a type of optimization problem that involves both continuous and discrete variables. The objective function and constraints are linear, but some/all

of the variables are restricted to be integers. This makes MILP a more complex problem than Linear Programming (LP), which only involves continuous variables. MILP can be used to model a wide range of real-world problems, including scheduling, resource allocation, and logistics. MILP problems can be solved using specialized algorithms such as branch-and-bound and branch-and-cut, which can be implemented in software such as CPLEX ([56]) and Gurobi ([57]). MILPs can be challanging to solve due to there combinatorial nature. Some specific challenges include:

- *Complexity*: The number of possible solutions grows exponentially with the number of integer variables, making the problem intractable for large instances.

- *NP-hardness*: Many MILP problems are known to be NP-hard, which means that no algorithm can solve them in polynomial time, in the worst case.

- *Local optima*: MILP solvers can get stuck in locally optimal solutions, which may not be the global optimal solution.

- *Symmetry*: MILP problems often have a large number of symmetric solutions, which can make it difficult for the solver to find the optimal one.

Totally Unimodular Problems

There are certain cases where the optimal solution of the LP relaxation of an MILP is guaranteed to be integral. Consider the following MILP,

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & c^\top \mathbf{x} \\ \text{subject to} \quad & \\ & A\mathbf{x} \leq b \\ & \mathbf{x} \in \mathbb{N}. \end{aligned}$$

If the A matrix is totally unimodular (TU) then the LP relaxation will always have one integral solution [58]. The following theorem and proposition can be used to check if a matrix A is TU,

Theorem 2: *A matrix A is TU if det(B)* $\in \{0, +1, -1\}$ *for every square submatrix B of A [59]*

Proposition 1. *Let* $\mathbf{A} \in \{-1, 0, 1\}^{n \times m}$*. If every column of A has at most one 1 and at most one -1, then A is totally unimodular [60].*

The following proposition states that if matrix A is Totally Unimodular (TU), then the basic solution is integral:

Proposition 2. *Let* $\mathbf{A} \in \mathbb{Z}^{n \times m}$ *be totally unimodular having rank m and* $b \in \mathbb{Z}^m$*. Then every basic solution of* $A\mathbf{x} = b$ *is integral (i.e. in* $\mathbf{Z^n}$*) [61].*

We will now prove that Eq. (3.22) is Totally Unimodular. Let $\mathbf{x}$ be the vector of all decision variables $s_{ij}(t)$, $x^{\mathrm{c}}ij(t)$ and $x^{\mathrm{c}}ij(t)$. Since the decision variables in Eqs. (3.21),(3.4),(3.1) appear as additions or subtractions, all entries in matrix A will either be 1, -1, or 0. Since $x^{\mathrm{c}}_{ij}(t)$ is the only decision variable that appears both in Eq. (3.21) and Eq. (3.1), but with different signs, each column of A will contain at most one 1 and one -1, making the A matrix Totally Unimodular. As a result, the Mixed Integer Linear Program (MILP) in Eq. (3.22) can be solved as a Linear Program (LP).

CHAPTER 4

Summary of included papers

This chapter provides a summary of the included papers.

4.1 Paper A

Sten Elling Tingstad Jacobsen, Anders Lindman, Balázs Kulcsár
A Predictive Chance Constraint Rebalancing Approach to Mobility-on-Demand Services
Submitted to *Elsevier Communication in Transportation Research* in Jan. 2023 `https://arxiv.org/abs/2209.03214`.

This paper considers the problem of supply-demand imbalances in Mobility-on-Demand (MoD) services. These imbalances occur due to uneven stochastic travel demand and can be mitigated by proactively rebalancing empty vehicles to areas where the demand is high. To achieve this, we propose a method that takes into account uncertainties of predicted travel demand while minimizing pick-up time and rebalance mileage for autonomous MoD ride-hailing. More precisely, first travel demand is predicted using Gaussian Process Regression (GPR) which provides uncertainty bounds on the prediction. We then for-

mulate a stochastic model predictive control (MPC) for the autonomous ride-hailing service and integrate the demand predictions with uncertainty bounds. In order to guarantee constraint satisfaction in the optimization under estimated stochastic demand prediction, we employ a probabilistic constraining method with user-defined confidence interval, using Chance Constrained MPC (CCMPC). The benefits of the proposed method are twofold. First, travel demand uncertainty prediction from data can naturally be embedded into the MoD optimization framework, allowing us to keep the imbalance at each station below a certain threshold with a user-defined probability. Second, CCMPC can be relaxed into a Mixed-Integer-Linear-Program (MILP) and the MILP can be solved as a corresponding Linear-Program, which always admits an integral solution. Our transportation simulations show that by tuning the confidence bound on the chance constraint, close to optimal oracle performance can be achieved, with a median customer wait time reduction of 4% compared to using only the mean prediction of the GPR.

CHAPTER 5

Concluding Remarks and Future Work

5.1 Concluding Remarks

Transportation is an essential brick stone in the building and running of modern societies. However, the increased demand for transportation due to urbanization and a shift in the vehicle industry towards carbon-neutral vehicles and production implies several challenges. To solve these challenges, the use of technological advancement and the development of new transportation services are essential. In this thesis, we have argued that Autonomous Mobility-on-Demand (AMoD) services have the potential to transform the transportation sector by offering affordable and convenient transportation services. The potential lies in the fact that these services are very flexible regarding how and where they can operate and that they can be centrally controlled. The flexibility of AMoD makes these systems more dynamic and stochastic compared to traditional public transportation services. Hence, several new challenges need to be investigated, such as the problem of stochastic supply-demand imbalance in AMoD services. This thesis has provided a methodology that addresses these challenges. The methodology developed in this thesis can be used for simulating the potential of AMoD for future transportation systems.

In more detail, we have proposed a predictive chance constraint rebalancing approach for autonomous mobility-on-demand (AMoD) services, which is applied to the use case of ride-hailing. We first introduce a commonly used model for this service where the service area is discretized into smaller areas called stations. The model consists of constraints for the imbalance and vehicle conservation. Based on the model, a model predictive controller (MPC) is formulated with the multi-objective to minimize vehicle rebalance distance and the imbalance in each station. The travel demand is predicted using Gaussian Process regression (GPR). In contrast to other proposed prediction methods, GPR is superior for small data sets and provides a confidence bound on the prediction. We account for uncertainties in the travel demand prediction by formulating a chance constraint MPC (CCMPC). The CCMPC is relaxed using the GPR prediction and the separable model. The proposed algorithm was benchmarked using the high-fidelity transport simulator AMoDeus and real taxi data from San Francisco [12]. Our results show the importance of incorporating the confidence bound of the demand prediction. By tuning the confidence bound, the median wait time is reduced by 4% compared to using only the mean prediction of the GPR. We showed that the CCMPC performs close to optimal performance and is significantly better than a reactive controller. The performance and computational efficiency of the proposed method imply that it would be helpful for real-time control.

5.2 Future Work

In this section, we would like to discuss possible future directions of this PhD project.

Charging and state of charge constraints

A natural future direction of this PhD project is to include charging and range limitation constraint into the proposed model. Since most future vehicles are predicted to be battery electric it is important to consider such constraints. There are several papers on the topic but simplified energy consumption models have been used and to the best of our knowledge no research have been done on using stochastic model predictive control (SMPC) that are scalable. Questions that are important to answer are,

- What car features are optimal for electric AMoD? Such as battery size, charging speed and energy efficiency.
- Where should charging station be placed and how many?
- How can a SMPC be formulized for real-time control?
- How can the electricity price be considered in the objective function?

Ride-pooling

In this thesis, it is assumed that each vehicle can only pick up and drop off one passenger at a time. However, the potential for pooling passengers together remains an area of investigation. Previous studies have shown that empty vehicle mileages can be reduced through ride-pooling with two passengers [13]. The feasibility of pooling more than two passengers, however, is an open question.

Endogenous congestion models

Most existing papers on control of AMoD services, including this paper, does not consider endogenous congestion models. There are a few papers that have looked at endogeneous congestion models and how AMoD could affect congestion [62]. However, there are still a few unanswered research questions:

1. Methods to reduce congestion, for example ride-pooling or integration of AMoD with public transportation.
2. Algorithms that can be operated in real-time and that can model congestion.

Competing AMoD service providers

In this thesis, we have assumed that there is only one AMoD operator in the city. However, it is most likely that there will be several AMoD operators competing for the same customers. What is the optimal dispatching and rebalancing strategy when there is competition? One important aspect here would be different pricing strategies and being able to offer a reliable service. Game theory models could be important to consider for such systems.

References

[1] E. E. Agency. "Greenhouse gas emissions by source sector." (2019), [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/ENV_AIR_GGEDV_447/default/table?lang=en (visited on 01/28/2023).

[2] Z. Zhu, Z. Li, Y. Liu, H. Chen, and J. Zeng, "The impact of urban characteristics and residents' income on commuting in china," *Transportation research part D: transport and environment*, vol. 57, pp. 474–483, 2017.

[3] R. Godfrey and M. Julien, "Urbanisation and health," *Clinical Medicine*, vol. 5, no. 2, p. 137, 2005.

[4] G. L. Ooi, "Challenges of sustainability for asian urbanisation," *Current opinion in environmental sustainability*, vol. 1, no. 2, pp. 187–191, 2009.

[5] J. Eliasson and M. Börjesson, "Costs and benefits of parking charges in residential areas," *Transportation Research Part B: Methodological*, vol. 166, pp. 95–109, 2022.

[6] A. Nilsson, G. Schuitema, C. J. Bergstad, J. Martinsson, and M. Thorson, "The road to acceptance: Attitude change before and after the implementation of a congestion tax," *Journal of environmental psychology*, vol. 46, pp. 1–9, 2016.

[7] S. Shaheen, B. Cohen Adam andYelchuru, and S. Sarkhili, "Mobility on demand operational concept report," U.S. Department of Transportation, Tech. Rep., 2017.

[8] K. Heineke, B. Kloss, T. Möller, and C. Wiemuth, *Shared mobility: Where it stands, where it's headed*, Surfline.com, Ed., 2021.

[9] G. D. Erhardt, S. Roy, D. Cooper, B. Sana, M. Chen, and J. Castiglione, "Do transportation network companies decrease or increase congestion?" *Science Advances*, vol. 5, no. 5, eaau2670, 2019.

[10] H. Becker, F. Becker, R. Abe, *et al.*, "Impact of vehicle automation and electric propulsion on production costs for mobility services worldwide," *Transportation Research Part A: Policy and Practice*, vol. 138, pp. 105–126, 2020, ISSN: 0965-8564.

[11] M. Schwall, T. Daniel, T. Victor, F. Favaro, and H. Hohnhold, *Waymo public road safety performance data*, 2020.

[12] C. Ruch, S. Hörl, and E. Frazzoli, "Amodeus, a simulation-based testbed for autonomous mobility-on-demand systems," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3639–3644.

[13] M. Tsao, D. Milojevic, C. Ruch, M. Salazar, E. Frazzoli, and M. Pavone, "Model predictive control of ride-sharing autonomous mobility-on-demand systems," in *2019 International conference on robotics and automation (ICRA)*, IEEE, 2019, pp. 6665–6671.

[14] B. K., "Future on-demand systems," Presented at ITSC 2022 Workshop on Co-Design and Coordination of Future Mobility Systems, 2022.

[15] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, Mar. 1955.

[16] R. Zhang, F. Rossi, and M. Pavone, "Model predictive control of autonomous mobility-on-demand systems," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016.

[17] R. Zhang and M. Pavone, "Control of robotic mobility-on-demand systems: A queueing-theoretical perspective," *CoRR*, vol. abs/1404.4391, 2014.

[18] M. Pavone, S. L. Smith, E. Frazzoli, and D. Rus, "Robotic load balancing for mobility-on-demand systems," *The International Journal of Robotics Research*, vol. 31, no. 7, pp. 839–854, 2012.

[19] S. Banerjee, C. Riquelme, and R. Johari, "Pricing in ride-share platforms: A queueing-theoretic approach," *Available at SSRN 2568258*, 2015.

[20] R. Iglesias, F. Rossi, R. Zhang, and M. Pavone, "A bcmp network approach to modeling and controlling autonomous mobility-on-demand systems," *The International Journal of Robotics Research*, vol. 38, no. 2-3, pp. 357–374, 2019.

[21] R. Zhang, F. Rossi, and M. Pavone, "Analysis, control, and evaluation of mobility-on-demand systems: A queueing-theoretical approach," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 1, pp. 115–126, 2018.

[22] A. Braverman, J. Dai, X. Liu, and L. Ying, "Empty-car routing in ridesharing systems," *Operations Research*, vol. 67, Sep. 2016.

[23] K. Treleaven, M. Pavone, and E. Frazzoli, "Asymptotically optimal algorithms for one-to-one pickup and delivery problems with applications to transportation systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2261–2276, 2013.

[24] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005, ISBN: 026218253X.

[25] R. Iglesias, F. Rossi, K. Wang, *et al.*, "Data-driven model predictive control of autonomous mobility-on-demand systems," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6019–6025.

[26] M. Tsao, R. Iglesias, and M. Pavone, "Stochastic model predictive control for autonomous mobility on demand," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3941–3948.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667.

[28] X. Geng, Y. Li, L. Wang, *et al.*, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 3656–3663.

[29] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20 110 550, 2013.

[30] G. Kim, Y.-S. Ong, C. K. Heng, P. S. Tan, and N. A. Zhang, "City vehicle routing problem (city vrp): A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1654–1666, 2015.

[31] H. N. Psaraftis, M. Wen, and C. A. Kontovas, "Dynamic vehicle routing problems: Three decades and counting," *Networks*, vol. 67, no. 1, pp. 3–31, 2016.

[32] B. Eksioglu, A. V. Vural, and A. Reisman, "The vehicle routing problem: A taxonomic review," *Computers & Industrial Engineering*, vol. 57, no. 4, pp. 1472–1483, 2009.

[33] G. Zardini, N. Lanzetti, M. Pavone, and E. Frazzoli, "Analysis and control of autonomous mobility-on-demand systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. 1, Nov. 2021.

[34] L. Zhao and A. A. Malikopoulos, "Enhanced mobility with connectivity and automation: A review of shared autonomous vehicle systems," *IEEE Intelligent Transportation Systems Magazine*, vol. 14, no. 1, pp. 87–102, 2020.

[35] N. Agatz, A. Erera, M. Savelsbergh, and X. Wang, "Optimization for dynamic ride-sharing: A review," *European Journal of Operational Research*, vol. 223, no. 2, pp. 295–303, 2012.

[36] S. Narayanan, E. Chaniotakis, and C. Antoniou, "Shared autonomous vehicle services: A comprehensive review," *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 255–293, 2020.

[37] A. Mourad, J. Puchinger, and C. Chu, "A survey of models and algorithms for optimizing shared mobility," *Transportation Research Part B: Methodological*, vol. 123, pp. 323–346, 2019.

[38] M. Z. Spivey and W. B. Powell, "The dynamic assignment problem," *Transportation science*, vol. 38, no. 4, pp. 399–419, 2004.

[39] C. Ruch, J. Gächter, J. Hakenberg, and E. Frazzoli, "The +1 method: Model-free adaptive repositioning policies for robotic multi-agent systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 3171–3184, 2020.

[40] F. Miao, S. Han, S. Lin, *et al.*, "Data-driven robust taxi dispatch under demand uncertainties," *IEEE Transactions on Control Systems Technology*, Sep. 2017.

[41] X. Guo, N. Caros, and J. Zhao, "Robust matching-integrated vehicle rebalancing in ride-hailing system with uncertain demand," *Transportation Research Part B Methodological*, vol. 150, Jun. 2021.

[42] L. He, Z. Hu, and M. Zhang, "Robust repositioning for vehicle sharing," *Manufacturing and Service Operations Management*, vol. 22, Apr. 2019.

[43] Z. T. Qin, H. Zhu, and J. Ye, "Reinforcement learning for ridesharing: An extended survey," *Transportation Research Part C: Emerging Technologies*, vol. 144, p. 103 852, 2022.

[44] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, "The rich and the poor: A markov decision process approach to optimizing taxi driver revenue efficiency," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 2329–2334.

[45] T. Verma, P. Varakantham, S. Kraus, and H. C. Lau, "Augmenting decisions of taxi drivers through reinforcement learning for improving revenues," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 27, 2017, pp. 409–417.

[46] Y. Jiao, X. Tang, Z. T. Qin, *et al.*, "Real-world ride-hailing vehicle repositioning using deep reinforcement learning," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103 289, 2021.

[47] Z. Shou, X. Di, J. Ye, H. Zhu, H. Zhang, and R. Hampshire, "Optimal passenger-seeking policies on e-hailing platforms using markov decision process and imitation learning," *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 91–113, 2020.

[48] X. Yu, S. Gao, X. Hu, and H. Park, "A markov decision process approach to vacant taxi routing with e-hailing," *Transportation Research Part B: Methodological*, vol. 121, pp. 114–134, 2019.

[49] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo–simulation of urban mobility: An overview," in *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*, ThinkMind, 2011.

[50] A. Horni, K. Nagel, and K. W. Axhausen, *The Multi-Agent Transport Simulation MATSim*. London, GBR: Ubiquity Press, 2016, ISBN: 1909188751.

[51] P. Group. "Ptv visum." (2023), [Online]. Available: `https://www.myptv.com/en/mobility-software/ptv-visum` (visited on 02/01/2023).

[52] A. Charnes and W. W. Cooper, "Chance constraint optimization," *Management Science*, vol. 6, no. 1, pp. 73–79, 1959.

[53] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, 2006.

[54] P. Chebyshev., "Des valeurs moyennes," *Journal de Mathématiques pures et Appliquées*, 1867.

[55] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming. Modeling and theory*. Society for Industrial and Applied Mathematics, Jan. 2009.

[56] I. I. Cplex, "V12. 8: User's manual for cplex," *IBM*, 1987.

[57] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, 2023.

[58] A. J. Hoffman and J. B. Kruskal, "Integral boundary points of convex polyhedra," in *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 49–76.

[59] L. De Giovanni, "Methods and models for combinatorial optimization heuristis for combinatorial optimization," 2017.

[60] P. Seymour, "Decomposition of regular matroids," *Journal of Combinatorial Theory, Series B*, vol. 28, no. 3, pp. 305–359, 1980, ISSN: 0095-8956.

[61] C. K. "Math 5801 linear optimization — lecture notes." (2021), [Online]. Available: `https://people.math.carleton.ca/~kcheung/math/notes/MATH5801/07/7_3_total_unimodularity.html` (visited on 02/01/2023).

[62] F. Rossi, R. Iglesias, M. Alizadeh, and M. Pavone, "On the interaction between autonomous mobility-on-demand systems and the power network: Models and coordination algorithms," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 1, pp. 384–397, 2019.