A Data Driven Prognostic Model for the Availability of Parking

Matthias Wissel

Supervisor: Prof. Dr. Felix Krahmer
Advisor: Dr. Frank-Dieter Filbir
Submission Date: 15th September 2016
I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Garching, 15th September 2016
Abstract

Zusammenfassung


Abstract

In this thesis, we propose a time-dependent nonhomogeneous Markov model for predicting free parking spaces in urban areas. Of particular interest is the parameter estimation based on a given data set. To this end, we extend the framework of empirical risk minimization to matrix-valued functions and use matrix-valued reproducing kernel Hilbert spaces to approximate the time-dependent generator of the Markov process. Also, an overview of related issues and further research opportunities is given.
Acknowledgements

I want to use the opportunity to thank the people who have helped in the process of this thesis and supported me along the way.

Above all, my thanks go to Dr. Frank Filbir for advising me while working on this thesis and for making it possible to choose this very interesting topic for my research. I especially want to thank him for the numerous insightful talks and discussions, his friendly ear for my questions and his advice and encouragement over the time working on this thesis and my studies in general. I could not have asked for a better guide throughout this time.

I also thank Prof. Dr. Felix Krahmer for supervising this thesis.

Furthermore, I would like to thank Dr. Daniel Kotzor for introducing me to this area of research and for offering his expert opinion whenever needed.

Moreover, I want to thank my fellow students, in particular Björn Händler, David Melching, Florian Speicher, Rebecca Bürkle, David Matthäus, and Serghei Mogoreanu, for proofreading this work and the helpful discussions we had over the last months.

Last but not least, I want to say thank you to my family for their neverending support during my studies and especially while writing this thesis.
### Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>number of parking segments</td>
</tr>
<tr>
<td>$E_k$</td>
<td>parking segment $k$, $k = 1, \ldots, K$</td>
</tr>
<tr>
<td>$N_k$</td>
<td>number of parking spots in segment $E_k$, $k = 1, \ldots, K$</td>
</tr>
<tr>
<td>$\mathbb{P}$</td>
<td>probability measure</td>
</tr>
<tr>
<td>$X(t)$</td>
<td>stochastic process</td>
</tr>
<tr>
<td>$S$</td>
<td>state space of the process</td>
</tr>
<tr>
<td>$\mathbb{E}[\cdot]$</td>
<td>expectation value</td>
</tr>
<tr>
<td>$P(t, s)$</td>
<td>matrix of transition probabilities</td>
</tr>
<tr>
<td>$Q(t)$</td>
<td>(time-dependent) generator matrix</td>
</tr>
<tr>
<td>$\lambda_i$</td>
<td>birth rates, $i = 0, \ldots, N$</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>death rates, $i = 0, \ldots, N$</td>
</tr>
<tr>
<td>$X$</td>
<td>non-empty set</td>
</tr>
<tr>
<td>$\mathcal{Y}$</td>
<td>closed subset of the space of matrices</td>
</tr>
<tr>
<td>$\mathcal{M}$</td>
<td>closed subset of the space of matrices</td>
</tr>
<tr>
<td>$\mathcal{T}(n)$</td>
<td>space of tridiagonal matrices of dimension $n$</td>
</tr>
<tr>
<td>$\mathcal{D}$</td>
<td>data set, so subset of $X \times \mathcal{Y}$</td>
</tr>
<tr>
<td>$M$</td>
<td>number of data points</td>
</tr>
<tr>
<td>$\mathbb{D}$</td>
<td>empirical measure depending on the data set $\mathcal{D}$</td>
</tr>
<tr>
<td>$L$</td>
<td>loss function</td>
</tr>
<tr>
<td>$\mathcal{R}_{L, P}$</td>
<td>risk functional w.r.t. distribution $\mathbb{P}$ and loss $L$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>regularization parameter</td>
</tr>
<tr>
<td>$\mathcal{R}_{L, P, \alpha}$</td>
<td>regularized risk functional w.r.t distribution $\mathbb{P}$, loss $L$ and parameter $\alpha &gt; 0$</td>
</tr>
<tr>
<td>$\mathcal{L}^0(X, \mathcal{M})$</td>
<td>space of measurable functions $f: X \to \mathcal{M}$</td>
</tr>
<tr>
<td>$\mathcal{L}^p(X, \mathcal{M}, \mu)$</td>
<td>space of $p$-integrable matrix-valued functions w.r.t. a measure $\mu$</td>
</tr>
<tr>
<td>$\mathcal{L}^p(X, \mu)$</td>
<td>space of $p$-integrable real-valued functions w.r.t. a measure $\mu$</td>
</tr>
<tr>
<td>$\mathbf{B}(\mathcal{M}, \mathcal{Z})$</td>
<td>space of bounded linear operators from $\mathcal{M}$ to $\mathcal{Z}$</td>
</tr>
<tr>
<td>$\mathbf{B}(\mathcal{M})$</td>
<td>space of bounded linear operators from $\mathcal{M}$ to $\mathcal{M}$</td>
</tr>
<tr>
<td>$\mathbf{B}_+(\mathcal{M})$</td>
<td>space of non-negative bounded linear operators from $\mathcal{M}$ to $\mathcal{M}$</td>
</tr>
<tr>
<td>$A^*$</td>
<td>adjoint of an operator $A$</td>
</tr>
<tr>
<td>$K(x, t)$</td>
<td>matrix-valued kernel</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>reproducing kernel Hilbert space of $\mathcal{M}$-valued functions</td>
</tr>
</tbody>
</table>
Contents

5.2.6. Estimating Time-Dependent Parameters Based on the Work of William A. Massey ........................................ 61
5.2.7. Compartment Model ........................................ 62
5.2.8. Purely Data-Driven Approach .............................. 63
5.2.9. Real Time Information ........................................ 63
5.2.10. Determine the Occupancy of a Parking Segment by Data From One Car .................................. 63

A. Appendix .......................................................... 65

Bibliography .......................................................... 70
1. Introduction and Overview

Mobility is becoming a more and more crucial topic for people all over the world. Especially in growing urban areas, the task of going from point A to point B and the question of how to do that conveniently can be quite difficult to answer. Even with modern public transportation systems and concepts like park and ride or special lanes for bicycles, for a lot of people the answer is still, going by car. Opposed to the flexibility a car provides in various aspects, there are also major set backs when going by car, no matter the question if it is from a car sharing service or a personally owned one. Besides red lights and traffic jams, the other major issue in terms of mobility by car is parking, or to be more precise, finding a free parking spot. This last topic is what we want to focus on in this thesis.

Just finding a free parking spot is already a big challenge on its own, in particular if the place you need to go to is in the city center or another urban area with a lot of traffic and only little space to park your vehicle. The trend to even further urbanization will make this problem a lot worse over the next years. Moreover, this is not only a problem concerning a single driver. It is not easy to precisely define park search traffic and there are various different definitions in the literature but, loosely speaking, it characterizes unnecessary traffic after the destination is already reached. In [1] the mentioned studies find that depending on the used definition and the investigated area, the proportion of park search traffic goes as high as 44%. So, parking is also an issue concerning the amount of traffic in general and therefore, of course, the environment. Until the market introduction of self-driving vehicles this problem will not go away or even get better, since more and more people all over the world want to benefit from the mobility features a car provides.

To attenuate this problem a little bit, a prediction system for free parking spots for the driver would be very helpful. So for example, if we are only 10 to 15 minutes away from our final destination the desired system can recommend a place close by where we are likely to find a free parking spot without having to look for it by cruising around the area for several minutes. Moreover, the navigation system could also adjust our route accordingly. Going by the desired destination before getting the car to the parking spot would of course also be possible, in case we have passengers or want to check the way back from the parking position. In any case, having a system telling us where we can conveniently find a free parking space would make life a lot easier.

Therefore, a lot of work has been and is still put into the development of such a prediction system. Already in the work of Frank A. Haight in 1963 on Mathematical Theories of Traffic Flow [2], besides the previous mentioned issues of mobility, we can find a whole section about parking as well, where the situation in a specific street is described via a Markov model, which we will use as well. Since parking prediction is a very complex topic also several other approaches have been discussed and analyzed. We want to give a brief overview of the available literature and methods used to tackle this problem.
1. Introduction and Overview

1.1. Related Work

In this part of the thesis we will present an overview of research also concerned with the issue of parking. Some references to literature about the mathematical tools and techniques we will use in this thesis will be given in the respective sections. As the topic of parking prediction is very broad, we will see different approaches and also different issues that are discussed and worked on. We will also see that most of the research is very recent due to the developments in sensor availability and therefore access to necessary information about parking behaviour.

As briefly mentioned before, in [2] Haight addresses the issue of parking from a mathematical point of view. He proposed a queueing model which, as we will see, is closely related to a Markov chain approach. The model parameters are assumed to be constant over time and already known. The main focus is on calculating the probabilities of a certain amount of parking spots in a segment to be free for different queueing systems.

Caliskan et al. [3] make use of that same model, so a continuous-time homogeneous Markov model to describe the situation in a specific parking lot. The parameters of the model are again constant over time and provided by the investigated parking lot. The model is evaluated by simulations based on a study done by city officials in Brunswick, Germany. Apart from the over time constant parameters this model is closest to what we will study in this thesis. Also, once again, the parameter estimation is neglected. In our studies we will mostly focus on a way to determine those parameters.

Now we will come to approaches very different from the one we will pursue.

In [4], Zheng et al. compare different techniques for parking prediction in car parks of smart cities. This means the filling levels of parking segments are provided by sensors which are part of the infrastructure. The used methods are regression trees, support vector regression and neural networks.

Caicedo et al. [5] use data from intelligent parking reservation (IPR) systems for their forecasting approach. The forecast is also aimed at IPR systems and is supposed to help drivers to select the best parking lot for their preferences, based on availability and also parking fees. For the forecast a combination of discrete choice model with previously calibrated parameters for the allocation of parking requests and a real-time availability forecast algorithm is used. The forecast algorithm uses historical as well as real-time data. It processes the allocated parking requests from the discrete choice model and simulated future parking duration to come to a forecast about the availability of a parking spot in the considered parking structures.

In [6] Ji et al. show a way to improve parking guidance information systems, i.e. off-street parking, by including a forecasting ability instead of just providing the information about the current situation. They are therefore addressing the issue of short-term availability of parking spaces and compare a wavelet neural network approach with methods based on the largest Lyapunov exponent. So, they consider this situation as a dynamical system. Their research is also based on data coming from the parking structures.

In [7] Beheshti et al. present a hybrid modeling approach for traffic prediction in urban simulations. According to the authors, this approach can also be used for parking prediction. The hybrid model is based on an agent system to bootstrap the proposal distribution which is then used in a Markov Chain Monte Carlo (MCMC) algorithm for Bayesian inference. As data
source a survey among university students about their daily mobility routines is used.

Hössinger et al. [8] worked on a comparison of real-time data sources and their capability to be used for occupancy prediction of short-term parking zones. In their evaluation, they made use of mobile phone data, car park counts and traffic flow volumes as sources of information. The occupancy was calculated in two different ways. First just as an overall average and then as an average over the day for different time intervals.

Vlahogianni et al. [9] propose a parking prediction system for smart cities based on on-street parking sensors. The prediction system is based on a neural network approach for the time series analysis of the occupancy rates. This approach is then combined with a hazard based model, which in general is used to gain a statistical representation of time to event data. In this case for the duration of free parking spots being still available.

Pullola et al. [10] present a way to find parking spots with an intelligent GPS-based navigation system. The idea is to maximize the probability that the parking lot still has one free spot when the driver arrives at the destination among the parking options close by. The availability of a parking spot is modeled by a Poisson process that incorporates historic and real-time data. The data for testing their approach is simulating a downtown area of Ottawa.

Last but not least, Coric et al. [11] address an issue very closely related to parking prediction. The aim of their research is to provide a method to generate a map of legal parking spaces by using the parking sensor data of cars. This is an important issue because, as we will point out in Section 2.6, it is not obvious how many parking spots are in a certain area and who is allowed to park there. The authors try to solve this problem using the historical data of many cars going to and from the parking segments various times. So, this work gives the foundation for on-street parking prediction where information like capacity of the lot are not previously given.

Summarizing this overview we see that major issues are time-dependence of the model and the data source. Dealing with off-street parking, where the infrastructure provides accurate information about the occupancy or filling level of the parking lot is a lot more common so far. This is understandable since gathering the necessary information about an on-street parking segment is a big challenge on its own.

1.2. Problem Description

After reviewing some of the approaches and issues discussed in the literature, we want to give a more precise description of how we want to tackle the problem of parking prediction and which techniques we will focus on.

The aim of this thesis is to analyze and further extend the often used stochastic model of Markov chains for the availability of free parking spots in urban areas. Mainly, this approach has been used for predictions via simulations and with already knowing the model parameters. Also, the possibility that the parameters of arriving and leaving cars might change over time throughout the day has been mostly neglected. We therefore aim to extend the Markov chain model to a time-dependent ODE system, so a time-dependent nonhomogeneous chain, and gain the predictions about the occupancy rate of a parking segment from there by studying the limiting behaviour of the chain. But, for this time-dependent behaviour the classical notions of
stationary distributions have to and will be relaxed. To make the proposed model work properly, the main issue is the estimation of the model parameters, here a time-dependent matrix-valued function. This will be our primary task to work on. To that end, the framework of empirical risk minimization and reproducing kernel Hilbert spaces needs to be extended to matrix-valued functions. This includes the study of loss functions and their corresponding risks in the matrix-valued case as well as the construction of kernels of that type which are appropriate for the task of estimating our parameter function. We will also study how the reproducing kernel Hilbert spaces can be used to provide a representation of the minimizer in the context of empirical risk minimization.

We will now give a short overview about how we seek to proceed with the different parts of our problem and how this thesis is structured.

1.3. Overview

The upcoming parts of the thesis are organized as follows:

**Chapter 2** describes the general setting and formalizes the environment we work in for further mathematical analysis. Then the basic model is introduced and two different interpretations of it are discussed. Also, the issue of what data we work with and how it can be gathered is addressed. Afterwards we give a brief guideline of which steps need to be taken for the whole procedure from data gathering to prediction using the described approach. So, this whole part aims to present the big picture of our approach before we go into more, sometimes technical, details in the next chapters. This part of the thesis ends with the description of some related problems to the field of predicting parking availability.

**Chapter 3** provides the mathematical framework for the prediction approach. We start with the formal introduction of the Markov model. Then we continue with the notions for the matrix-valued regression problem which we face when estimating the parameters for the Markov model. Therefore we will talk about loss functions and their associated risks, matrix-valued reproducing kernel Hilbert spaces and their usage for risk minimization.

**Chapter 4** is concerned with estimating the parameters for the introduced Markov model based on a given data set, making use of the framework of empirical risk minimization. This will be done for constant and time-dependent parameters to emphasize the difference in terms of complexity that comes along with the time-dependent aspect. Also, in the time-dependent case a kernel for the needed matrix-valued reproducing kernel Hilbert space is constructed and the respective optimization problems are derived.

**Chapter 5** sums up the results and gives an outlook on the various open questions that give opportunity for further research. This list includes mathematical issues, as well as numerical questions. Also, model assumptions are discussed and a possible choice for a completely different model is given. Finally, we briefly talk about a modeless, so a purely data-driven, attempt for prediction and also how to include real time information.
2. Model Description

In this part of the thesis we will introduce the basic model we will use to describe a parking segment and see in which ways we can interpret it for our purpose. Also, we want to address the format of the available data, how we get from the raw data to a prediction and some related issues that come up while working on the prediction of available parking spaces. The rigorous mathematical definition and analysis of the model which we will introduce in the following sections and the objects and notions connected with it will be done in Chapter 3.

2.1. General Setting

To come up with a mathematical model to describe our task we first need to understand the circumstances and formalize the environment we are working in.

For parking a car we basically have two different options available. The first one is to use parking structures and large parking lots in special locations of a wider area, which is also referred to as off-street parking. The second option is roadside parking, i.e. using the available parking spots alongside the roads. Since we want our prediction to be accurate on a local level, we need to divide certain roadside parking options even further because some of the roads are too long or have different parking rules. This means different groups of drivers are allowed to park in specified areas at different times of the day or week, like for local residents. The reason we might want to divide a long road in several parking segments is that if we are looking for a parking spot at one end of the road and the prediction tells us that there are free spaces available in that segment, i.e. alongside the whole road, then we might not find any free spaces at the part of the road we are looking for a space since the free spaces are for example at the other end of that street.

To describe this in a convenient way, we say our area of interest is divided in $K$ parking segments $E_k$, $k = 1, \ldots, K$. Each of these segments $E_k$ provides $N_k$, $k = 1, \ldots, K$, parking spots or slots.

For the prediction, we now need a stochastic model which allows us to calculate how many of the total amount of available parking spaces are occupied with a certain probability. Favorably that model is able to account for time-dependent behaviour and allows us to incorporate the different traffic intensities which occur throughout the day. In general, we can expect to have higher volatility of the process in the morning and in the evening when a lot of people go to and from work. Also, the working days during the week will differ from the weekend. In a restaurant district we might have higher intensities during lunch time and in the evening, whereas in a school area we would expect to see the intensity peaks when school starts and is over. We can think of more circumstances but these few examples already show how important the time-dependent aspect of our approach will be.

We now come to the stochastic model that is at the center of our approach.
2. Model Description

2.2. Markov Model

To describe the behaviour in a parking segment we will use a Markov process \( \{ X(t) \}_{t \in \mathbb{R}_{\geq 0}} \) as a mathematical model. More precisely we will use a time-continuous Markov chain with a finite state space. We therefore need to define the state space \( S \), the time-dependent infinitesimal generator matrix \( Q(t) \) and the initial probability distribution \( P(t_0) \) on the state space for each parking segment \( E_k \). The crucial part here is the time-dependence of the infinitesimal generator, which makes our resulting Markov chain nonhomogeneous or non-stationary. We will see that this complicates the analysis in several ways when Markovian models are introduced in more detail in Section 3.1.

As seen in the few examples before, the behaviour can be different depending not only on time, but also due to the location and the surrounding area. So for each parking segment \( E_k \) we will have a state space \( S_k = \{0, ..., N_k\} \), a time-dependent infinitesimal generator matrix \( Q_k(t) \in \mathbb{R}^{N_k+1 \times N_k+1} \) for \( t \in [0, T] \) some time interval and an initial probability distribution \( P_k(t_0) \).

Although the behaviour can vary in different locations, these variations due to location are always reflected in the time-dependent behaviour. So, we use the same time-dependent model for each location, but with different time-dependent parameters. And, since the model is similar for every parking segment and for sake of notation we will drop the index \( k \) to distinguish between the parking segments \( E_k \) in this thesis. For a real life application on the other hand, we would need to consider all the different parking segments \( E_k, k = 1, ..., K \), one by one.

Describing the situation in each parking segment with a Markovian model means that our state space represents the number of occupied parking spaces. So, a change of the state of the process means that there is a changed number of cars occupying the available parking spaces in a particular segment. Assuming the Markov property (see Definition 3.4), which is the essential feature of a Markov process, for the behaviour of the occupancy of a segment is reasonable since the probability of cars arriving or leaving the segment is only subject to the number of cars currently parked there and not to how many there have been before. So, only the current state of the process is of interest, which is exactly what the Markov property states. The transition probability matrix \( P(t) \in \mathbb{R}^{N+1 \times N+1} \) gives the probabilities of how likely it is for the process to change from one state to another in a small time interval \( \Delta t \). This, of course, depends on the number of cars arriving or leaving in a given time interval. For a general Markov process the time-dependent generator matrix has the form

\[
Q(t) = \begin{bmatrix}
q_{0,0}(t) & \cdots & q_{0,N}(t) \\
\vdots & \ddots & \vdots \\
q_{N,0}(t) & \cdots & q_{N,N}(t)
\end{bmatrix},
\]

where \( q_{i,j}(t) \) are real valued functions fulfilling the so called q-Property, which we will define properly in Definition 3.6.

In our case, the state of the process changes only by one each time, i.e. one car parks in or one car parks out. Due to this we can interpret this model both as a birth-death process or as
a queueing problem. We will elaborate on these interpretations in Section 2.3. This also results in a tridiagonal structure for the transition rate matrix, also called generator matrix \( Q(t) \). The transition rates represent the parameters of the random variable \( X(t) \) describing cars coming to the parking segment to park, denoted by \( \lambda_i(t), i = 0, \ldots, N-1 \), and already parked cars to leave the segment, denoted by \( \mu_j(t), j = 1, \ldots, N \). In whole, we get a Markov process \( \{X(t)\}_{t \in \mathbb{R}_0^+} \) and corresponding state transition probabilities \( P(\Delta t)_{ij} = \mathbb{P}[X(t + \Delta t) = j | X(t) = i], i, j = 0, 1, 2, \ldots, N \), which together with the time-dependent infinitesimal generator matrix \( Q(t) \) lead to the system of differential equations, first introduced in [12] and now known as the Kolmogorov forward equation,

\[
\dot{P}(t) = P(t) \cdot Q(t) \quad P(t_0) = I, \tag{2.2}
\]

where \( I \) is the identity matrix,

\[
Q(t) =
\begin{bmatrix}
-\lambda_0(t) & \lambda_0(t) & & & \\
\mu_1(t) & -(\lambda_1 + \mu_1)(t) & \lambda_1(t) & & \\
& \mu_2(t) & -(\lambda_2 + \mu_2)(t) & \lambda_2(t) & \\
& & \ddots & \ddots & \ddots \\
& & & \mu_{N-1}(t) & -(\lambda_{N-1} + \mu_{N-1})(t) & \lambda_{N-1}(t) \\
& & & & \mu_N(t) & -\mu_N(t)
\end{bmatrix},
\tag{2.3}
\]

and \( \lambda_i, \mu_j : [0, T] \rightarrow \mathbb{R}_{\geq 0}, i = 0, \ldots, N-1, j = 1, \ldots, N \), are positive, real valued functions, representing the time-dependent transition rates during the considered time interval.

So the infinitesimal generator matrix represents how volatile our system is, i.e. how fast changes in the transition probabilities occur. This is due to the fact that the entries of the generator matrix work as factors on the transition probabilities and this product of transition probabilities and transition rates forms the derivative of the transition probability matrix, so their rate of change.

We now want to include two more assumptions into our model concerning the transition rates \( \lambda_i, \mu_j \) and how the current state of the process affects these rates:

**Arrival- or Birth-rate**, the time-dependent rate of arrivals of cars to the parking segment is independent of the number of cars which already park there, so we have \( \lambda_i(t) = \lambda(t) \) for all \( i = 0, \ldots, N-1 \).

**Service- or Death-rate**, the time-dependent rate of departures of cars is linearly related to the number of cars already parking in the segment, which yields \( \mu_i(t) = i \cdot \mu(t) \) for all \( i = 1, \ldots, N \).

The terminology is borrowed from the possible model interpretations which we will focus on in more detail in Section 2.3. Having these transition rates leads to the observation that the interarrival times of the cars as well as the service times, i.e. the parking durations, are exponentially distributed (see [13]). We can also find a justification for our assumption of linear dependence of the service rates \( \mu_i(t) \) on the state of the process there.
2. Model Description

These assumptions lead to the observation that our generator matrix $Q(t)$ is of the form

$$Q(t) = \begin{bmatrix} -\lambda(t) & \lambda(t) \\ \mu(t) & -(\lambda + \mu)(t) & \lambda(t) \\ 2\mu(t) & -(\lambda + 2\mu)(t) & \lambda(t) \\ \cdots & \cdots & \cdots \\ i\mu(t) & -(\lambda + i\mu)(t) & \lambda(t) \\ \cdots & \cdots & \cdots \\ N\mu(t) & -N\mu(t) \end{bmatrix}.$$ (2.4)

In theory this looks quite nice so far. The problem now is to actually determine the generator matrix, i.e. the parameters of our model. As mentioned before this part is often neglected in previous work about free parking spot prediction. To do this we will apply a kernel approach for matrix-valued functions. This is necessary since we have to handle the whole system of the Kolmogorov equations and in general can not reduce it to a 1-dimensional problem. The mathematical tools for this are described in Section 3.2. We will discuss the parameter estimation in detail in Chapter 4. For this approach to work we will assume to be given a data set consisting of pairs of time step and transition probability matrix of the process $X(t)$, i.e.

$$D = \{(t_1, P(t_1)), \ldots, (t_M, P(t_M))\} \subset [0, T] \times \mathbb{R}^{N+1 \times N+1}.$$ (2.5)

How to possibly attain this data set is briefly described in Section 2.5. Some general issues about the given and needed data, as well as their gathering, are addressed in Section 2.4.

Knowing the parameters of the model we now want to answer the question that is crucial for our application: What is the probability of the Markov chain being in a certain state $i$ at a given time $t^*$? So we are looking for $P[X(t^*) = i]$. This means, we want to know how many parking spots are occupied at a given time with a certain probability. To do that we turn to investigating the limiting behaviour of the chain, i.e. stationary distributions and their generalizations. Why is this an appropriate way to do this and what does it mean in our setting?

As pointed out in [14], already in the constant case we study long term limiting behaviour of the process, i.e. the chain, but this is not because we are solely and entirely interested in what happens in the distant future. We do this rather to use the long term behaviour as a description of the present and near future, since we assume that the system is already in an equilibrium or steady state.

With time-varying rates however, this is no longer appropriate. Doing the same as in the constant parameters case, we would approximate the current behaviour of our process with the limiting behaviour which is based on parameters that have not happened yet. So, the kind of asymptotics resulting from limits as $t \to \infty$ need to be modified.

Therefore, we assume that the rates of our generator $Q(t)$ change slowly enough over time for the process $X(t)$ to achieve an equilibrium before a too significant change in the rates occurs. To formalize this, we introduce an additional parameter $\varepsilon > 0$ and go from looking at $Q(t)$ to $Q(\varepsilon t)$. The corresponding differential equation then reads

$$\frac{\partial P^\varepsilon(t)}{\partial t} = P^\varepsilon(t)Q(\varepsilon t).$$ (2.6)
With this setup, we will take the limit as $t \to \infty$, as usual, but simultaneously we let $\varepsilon \to 0$ with $\tau = \varepsilon t$ for some fixed $\tau > 0$. The corresponding probability distribution $P^\varepsilon(\tau)$ will then solve the forward equation

$$\frac{\partial P^\varepsilon(\tau)}{\partial \tau} = P^\varepsilon(\tau) \frac{Q(\tau)}{\varepsilon}. \quad (2.7)$$

With the introduction of the additional parameter $\varepsilon$, we now have two different time scales for the generator and hence the Markov chain. We have the regular time scale, corresponding to $t$, and the stretched time scale $t/\varepsilon$.

Using the results about perturbed systems, like Equation (2.7), provided in [15] we will calculate the so called quasi-stationary distribution, a generalization of the classical stationary distribution. This distribution provides us with information we are looking for, so about the state of the chain or to be more precise the probability of each state of the chain over time. Therefore, we can, according to the calculated probability, conclude how many free parking spots there are in the parking segment of interest.

Since the described Markov model is quite general, there are different ways to interpret it. So, in the following section we want to have a closer look at two possible interpretations, as a queueing problem with finite number of service devices and as a birth-death process.

### 2.3. Interpretations of the Model

To get a better feeling how the model represents our situation in one specific parking segment and what the different parameters stand for, we will now discuss two more specific models which are closely connected to each other and Markov models in general. So, they both give an equivalent description of our problem but still allow us to gain more insight due to the fact that both have been extensively studied in various other areas of application.

#### 2.3.1. As M/M/c - Queue

Queueing problems occur in various fields and are also a whole area of study on their own. For a closer look we refer to [13], [16] or [17], although, therein the time-dependent case is not discussed. These kinds of models are very popular, in particular in telecommunication to represent servers and clients. But also every ticket counter for a public event or the cashier at a supermarket are situations we can think of. Here we just want to give an explanation of how our Markov model can be seen as a queueing problem and what kind of queueing problem it is. To do that we will use the basic elements of queueing problems described in the very beginning of [13], so how we can categorize different queueing systems. The basic elements are the following:

**The input process**, a description of the number or frequency of arriving customers in the system during a time interval or the time interval between successive arrivals via a random variable.

**The service mechanism**, covers aspects including the number of service stations, the number of customers that is being served at a certain time and how long the service takes. For each of these features random variables can be used.
2. Model Description

The system capacity, describes the amount of customers being able to wait in line. This might be infinite.

The queue discipline, combines all factors regarding the rules of conduct of the system, like the rule by which the customers are served, for example "first come first serve".

By identifying these elements in our problem we will gain a good insight about what kind of queueing problem we need to look at for further analysis. To distinguish between different queueing models the Kendall notation, initially introduced in [18], is quite common, so we will make use of it as well.

The general description of a queueing problem in this form then reads \( A/S/c/K/N/D \), where

- \( A \) denotes the arrival process and putting \( M \) or \( D \) represents Poisson/Markovian or deterministic input process, respectively,
- \( S \) denotes the service time distribution, and we can put \( M \) or \( D \) as in the arrival process,
- \( c \) denotes the number of service channels or stations, i.e. putting a one here represents a single service station and putting \( n \), accordingly, represents \( n \) service stations,
- \( K \) denotes the capacity of the system, so a limit of customers in the system, including the one currently in service, but it can be infinite, which is the default and the symbol is omitted,
- \( N \) denotes the size of the overall population, so how many customers are able to address the queueing system, here as well this can be infinite, which is the default and then again the symbol is omitted,
- \( D \) denotes the queueing discipline, so if there is a system of prioritizing customers and if there is, which one, the default is FIFO, so first in first out, also known as first come first serve, then the symbol is omitted.

To classify our model we start by deciding what kind of input process we have. Our interarrival times are exponentially distributed, as we discussed before, so we have a Poisson process at hand. The same goes for the service times, so in both cases we get the symbol \( M \). Next, we consider the number of service stations. These are in our case represented by the number of parking spots. So for a segment we have \( N \) of those. The capacity of our system is infinite since we can consider the traffic system, so all the roads, as potential waiting room. Also, the size of the overall population of our system is infinite, since in general every random car could come by the considered parking segment trying to find a free spot. We assumed no prioritizing in our model, so we also consider the standard here.

This makes our considered model a \( M/M/c \) queueing problem in the Kendall notation, where we have in particular \( c = N \). Various work has covered \( M/M/c \) queueing systems like [13] or [19], which we can use to gain further insights about our model. Looking at \( M/M/c \) queues we see that they can be described again as Markov processes with a generator matrix as in Equation (2.4).
2.3.2. As Birth and Death Process

Birth and death processes are a special kind of continuous-time Markov processes. We refer to Chapter 10 of [20] for an introduction as part of continuous-time Markov chains or to [21] for a more sophisticated analysis. In [17] they are studied as well. Their key feature is the fact, that transitions between states only occur in the way that the state variable goes up by one, i.e. a "birth", or it goes down by one, i.e. a "death". As the nomenclature already suggests, birth and death processes are widely used in biology and related fields like epidemiology or demography. They are mainly used to describe the population of a system like bacteria evolution or the spread of a disease in a population. Looking at Section 2.3.1, we can see that there is also a very close link to queueing theory. So the waiting line at a super market cashier can also be modeled as a birth and death process.

Birth and death processes are characterized by their transition rates, so the birth rates \( \{\lambda_i\}_{i=0}^{\infty} \) and death rates \( \{\mu_i\}_{i=1}^{\infty} \). If the state, so the size of the population, is limited, as in our case, then there are only finitely many birth and death rates, i.e. we have \( \{\lambda_i\}_{i=0}^{N-1} \) and \( \{\mu_j\}_{j=1}^{N} \). In our case these rates are not constants but time-dependent positive functions. This then resembles the structure of the generator matrix we described in Section 2.2, especially Equation (2.3). Our particular choice for the birth and death rates is just a more specific case of a birth and death process but we can nevertheless understand our situation as one of this problem class.

The cars arriving at the parking segment and actually parking there are the births and the cars leaving, so parking out, are the deaths. Therefore the number of parked cars in the considered segment represents the size of the population in the system.

Next, we turn to the issue of what data we can work with and which kind of data is actually necessary for our model.

2.4. Available vs. Needed Data

In this section we will focus on the data needed for the Markov process \( X(t) \) in each parking segment. Other issues, regarding additional information needed for the model and parking prediction in general, will be briefly discussed in Section 2.6 and are omitted here.

We distinguish two types of information:

- The first one being the time, location and type of a parking event. A parking event triggered by a car can either be departing from a particular parking segment or arriving and parking at a certain segment. Looking at the whole drive of a car from A to B it is of course both, but we will look at each event individually, so from the point of view of each respective segment.

- The second part of the data is the information of how many slots of the respective parking segment are occupied, so how many cars are there.

From these types of information we obtain the desired samples from the stochastic process \( X(t) \) for a given segment \( E_k \), so in total a time series \( \{X(t_0), \ldots, X(t_M)\} \) for each segment \( E_k \). How gathering these information can be achieved and what problems arise with the second type of information will be discussed in Section 2.4.1. But, even if we manage to gather this data,
2. Model Description

unfortunately this is still not enough for our parameter estimation approach to be applied. We need the information about the transition probabilities from one state of the process $X(t)$ to another one after a small time step $\Delta t$, which is described in the entries of our matrix $P(t)$, via $P(\Delta t)_{ij} = P[X(t + \Delta t) = j|X(t) = i]$. But without knowing the underlying probability distribution beforehand this is not possible right away.

To go on with our approach we assume in this thesis that we already have a data set $D = \{(t_0, P(t_0)), (t_1, P(t_1)), \ldots, (t_M, P(t_M))\} \subset [0, T] \times \mathbb{R}^{N+1 \times N+1}$ in the form of pairs of time and belonging transition probability matrix. Work on methods for obtaining this data from the stochastic process is briefly mentioned in Section 2.5.

2.4.1. Data Gathering

Now we want to address how we can actually collect some of the needed information. We will omit all the technical issues here, like the sensor accuracy, the frequency of and necessary architecture for sending the data from the car to the server and so forth. These aspects play a very important role in this process but we will rather have a look at what can be quite easily gathered and where problems occur due to the type of information and the available ways to get it.

We distinguish three ways to gather our information:

- The first way is by car, i.e. we use cars and their sensors to collect data like GPS locations and send this data to a server where we can use it for further analysis. This is done by most modern cars which use some kind of connected car service with car-to-x communication. A study by McKinsey found that until the year 2020 about one in five cars will be connected to the internet [22]. As the development in this area advances, there will be a lot more information gathered by the cars that can be used, like data collected by the sensors of automatic parking assistance systems.

- The second way is by infrastructure. For example the parking segments themselves might be able to gather and send information. This way of data collection is in general only available in parking structures or restricted parking lots which track the number of cars going in and out. Roadside parking, which is a more common way of parking a car does usually not offer this kind of information.

Although more and more companies try to make it easier and cheaper for cities to deploy this kind of technology coupled with the infrastructure, the first way, i.e. data collection by car, is much more common at the moment.

- Developments over the last few years have brought a third way of data gathering. It might be considered as by infrastructure as well but we will keep it separate since it does not fit the classical perception we have of infrastructure. This way of data gathering is via platform, so software architectures, like SCARiT [23], where users can report free parking spaces, or offer their own private parking spots to the public for several hours of the day. This trend is of course closely related to the spread of smartphones and the possibility to easily provide useful information to a broad user base. Other platforms like parkpocket [24] gather information about parking structures and provide services like reservation, routing and live data about the occupancy rate of the parking structure.
2.5. From Data to Prediction

After distinguishing the ways of data collection, we now want to see how we can use those to gather our two basic types of information:

The first part of the data, time, location and type of the parking event, is less of a challenge. By evaluating the GPS information sent by the cars to servers of the respective car company, or some other service provider, it is possible to determine where and when a car is started (i.e. parked out) and also where and when it is turned off (i.e. parked in). Of course, for the GPS data a mapping process is necessary to determine which parking segment was used by the car. Some of the issues that are related to this process are mentioned in Section 2.6.

The second part of the data, the number of occupied parking spots in a particular segment, is a lot harder to get. Of course, if the infrastructure is providing this information on a sufficiently frequent basis we are done. The same goes for suitable platform solutions. If, on the other hand, we only have the information from one car available then already this issue becomes quite tricky. We might be able to use the parking sensors of the car to see if there are objects, maybe other cars, in front of or behind it. Then, it is still not clear how many other cars there are. The same applies if there is free space. How much of it is in the segment? For how many cars does it suffice to park? Newer premium models have sensors to scan the side of the road for free space to possibly park in due to an automatic parking assistant. The problem here is that these sensors only work up to a certain moderate speed limit and close proximity to the roadside, in general up to around 35km/h and 1.5m away from the roadside [25], [26], [27].

We therefore can get only very few pieces of information about the number of occupied spaces. If the car finally stopped, i.e. parked, at a segment, we know that there was at least one free spot. Also, if the car leaves the segment we know that there is at least one free space right now. Unfortunately, this is not very accurate information about the state of the underlying process.

In general, getting the information about the state of the process $X(t)$, i.e. how many slots are occupied, by just the data from one car is a challenging and still open problem on its own, which is beyond the scope of this thesis. We will give a broader overview of other remaining open issues in the field in Section 5.2.

2.5. From Data to Prediction

In this section we want to give an overview of the whole procedure that is necessary end up with a prediction about free parking spots. So, how might we get from data about the time $t$ and the respective state of the process $X(t)$ for a particular parking segment to a prediction of how likely it is that there is a free parking spot.

Assume we have a data set containing timestamps and the respective states of the process, i.e.

$$\{(t_1, X(t_1)), (t_2, X(t_2)), \ldots, (t_M, X(t_M))\} \subseteq [0, T] \times \{0, 1, \ldots, N\}, \tag{2.8}$$

so a realization of the path the process took. Depending on the amount of data of this form we have, there are different ways to come to the type of data we need to proceed with our approach. As pointed out in Section 2.4, we need information about the transition probability $P(\Delta t)_{ij} = \mathbb{P}[X(t+\Delta t) = j | X(t) = i]$, $i, j = 0, \ldots, N$, instead of just the state of the process $X(t)$. 

13
2. Model Description

Assuming $M$ is reasonably big, we can make use of statistical tools such as Maximum Likelihood estimation methods. In [21] and [28] the authors have described ways to process this information into the necessary form, using expectation maximization algorithms. Methods known from time series analysis, like those described in [29] could be used as well.

Also, if we have an amount of data large enough to use statistical tools we can make use of the technique developed in [30] where a product limit estimator is proposed to estimate the transition probabilities of a nonhomogeneous Markov chain with finite state space. This exactly fits the features of our model.

If we only got a small amount of data available then we will have to use a different approach. We can probably estimate some of the entries, i.e. probability values, for the transition probability matrix, but we might have some missing entries. This can be due to lack of data all together or just not enough data for a sufficient estimate. But, we also have some additional information about the properties of the transition probability matrices. In particular, we know that each row has to sum up to one, since we are talking about a probability matrix. So, in that case a possible way out are matrix completion methods, initially described in [31]. These methods can be used for various problems, the probably most famous one being the so called Netflix problem [32].

After having obtained a data set of the form

$$D = \left\{ (t_0, P(t_0)), (t_1, P(t_1)), (t_2, P(t_2)), \ldots, (t_M, P(t_M)) \right\} \subset [0, T] \times \mathbb{R}^{N+1 \times N+1},$$

the next step depends on weather we assume our generator matrix $Q(t)$ to be constant or time-dependent. Both cases are described in detail in Chapter 4. In Section 3.2 we rigorously introduce the necessary mathematical framework. We will only give a short overview here.

In the constant parameters case, we will use the fact that the Kolmogorov Forward equation,

$$\dot{P}(t) = P(t) \cdot Q \quad P(t_0) = I,$$

has the analytic solution $P(t) = \exp(-tQ)$, where $\exp(\cdot)$ denotes the matrix exponential function. With that we can formulate an optimization problem, for example by using a least squares approach, to determine the best fit of our parameters $\lambda$ and $\mu$ for the given data set. This is subject of Section 4.1.

If we assume our generator matrix to be time-dependent, formulating the optimization problem becomes more complicated. Again, we want to find the best fit for our data set. Due to the lack of a closed form solution we need to specify a function space in which we aim to find the best fitting function for our data set. To help with this and the representation of the solution of the minimization problem in that function space we choose a kernel approach to determine the time-dependent generator matrix $Q(t)$. A preliminary problem here is that we cannot do this immediately with the given data set. To adjust to this, we therefore discretize the derivative in the Kolmogorov forward equation (Equation (2.2)) and solve a preliminary optimization problem to obtain data samples from the matrix-valued generator function. So, we end up with a modified or preprocessed data set

$$D_{\text{mod}} = \left\{ (t_1, Q(t_1)), \ldots, (t_M, Q(t_M)) \right\}.$$
functions. More details will be shown in Section 4.2.

Now that the parameters of our Markov model are calculated and we know the generator matrix, we can come to the prediction of the state of the chain at a given time. For this we make use of the results on perturbed systems and their limiting behaviour provided in [15] which we will recapitulate in the end of Section 3.1. In practical applications this evaluation will be done by an ODE solver, but the mathematical analysis gives us requirements for our solution to exist and what to expect of it.

After we have pointed out a way from data about the process to an actual prediction, we now want to address some additional problems and questions that we neglected so far while on that journey.

2.6. Related Problems

Other issues related to the general quest of predicting available parking spaces arise when the previously explained model needs to be filled with information.

The first issue is, we assume the number $N_k$ of available parking spots in each segment is known. This might be the case for the parking structures. But, for roadside parking it is often unknown, even to the respective city administrations. Of course, in many situations we can just go there and count the number of slots available. Although it would be a tedious task to do this in a city like Munich, or even worse one of the worlds mega cities like New York, even that is not always possible because on some streets there are no single parking spots specified. There is just space provided for parking alongside the road. So the number of available spots is depending on the types of cars parking there, i.e. their length or width, depending on the geometry of the area, or how much space is left between cars that are already parked there when new cars arrive at the parking segment, so how organized people park their cars.

Another issue is the parking rules specified for a given segment $E_k$. This of course also influences the available parking spots $N_k$ in general but also might just restrict the parking segment to a specific group of car owner’s or type of cars, for example local residents or electric cars. These pieces of information also are not stored in a global database but at best at different administrative entities. Also the various variables in these rules, i.e. type of car, type of driver, time of the day, day of the week, make them quite complicated to integrate into our model.

The third important issue is traffic regulations. These play a role when whole segments are not available, for example because of roadworks, but are most crucial for the question of how to divide the city into parking segments. Segments to be connected is probably most convenient but not necessary in general. The choice here will influence how hard it is to map the GPS data from the cars and the resulting parking events to a certain parking segment.

Going along with this is the general question of how to best divide the area or city in parking segments. There are a lot of ways to do this and the chosen method might have a huge impact on the results of the prediction. Also, the accuracy based on location is of course closely related to the size of the parking segments. If we choose our parking segments to be entire areas or city
2. Model Description

districts then we cannot expect to get a prediction about how easy it is to find a parking spot in a certain street. We remind of the example of a long street we already discussed. The same problem occurs if the parking segments turn to whole areas.

After we laid out the general model for our approach and addressed some related topics alongside our main question, we will now come to the more formal part of this thesis. In the next part we will rigorously define and derive the mathematical notions and tools necessary for the model and processes we described so far.
3. Mathematical Framework

This part of the thesis will provide the mathematical foundations for our goal to estimate time-dependent parameters of the stochastic model we introduced in Chapter 2.

3.1. Markov Chains

Markov chains in general are a huge area and a lot of different aspects and properties can be discussed and analyzed. We will focus on the main ideas and concepts needed for our model. For a broader view of the topic, especially with regards to homogeneous Markov processes, we refer to [33] or [20]. The generator of the Markov process plays an essential role in our studies so we will follow the approach of introducing stochastic processes proposed in [15]. This is also supported by the fact that we are mainly interested in nonhomogeneous Markov processes, which are often neglected in the standard literature.

We start of by introducing the basic notions we need. And the first object we want to define is a stochastic process.

**Definition 3.1 (Stochastic Process, State Space and Jump Process)**

Let \((X, \mathcal{F}, \mathbb{P})\) be a probability space and \(S\) a Polish space with Borel \(\sigma\)-algebra \(\mathcal{B}(S)\). Then a \(S\)-valued **stochastic process** is a totally ordered collection of \(S\)-valued random variables on \(X\).

\[
\{X(t): t \geq 0\},
\]

where each \(X(t)\) is a \(S\)-valued random variable on \(X\). We call \(S\) the **state space** of the process.

As a special case of a stochastic process, we call a right-continuous stochastic process with piecewise constant sample path a **jump process**.

For our application the state space \(S\) will always be finite. So, we will assume from now on that \(S\) is a finite space.

The next concept we need for our analysis is the notion of filtrations.

**Definition 3.2 (Filtration and adapted process)**

Let \((X, \mathcal{F}, \mathbb{P})\) be a probability space. A collection of \(\sigma\)-algebras

\[
\{\mathcal{F}_t: t \geq 0\}
\]

or short \(\{\mathcal{F}_t\}\), is called **filtration** if

\[
\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F} \quad \text{for all } s \leq t.
\]

A process \(X(t)\) is said to be **adapted** to the filtration if, for every \(t \geq 0\), \(X(t)\) is a \(\mathcal{F}_t\)-measurable random variable.
3. Mathematical Framework

Now we come to martingales, which will allow us to introduce the concept of generators later on.

**Definition 3.3 (Martingale)**
A stochastic process \( \{X(t) : t \geq 0\} \) is said to be a **martingale** on a probability space \((X, \mathcal{F}, \mathbb{P})\) with respect to a filtration \(\{\mathcal{F}_t\}\) if

(i) for each \( t \geq 0 \), \( X(t) \) is \( \mathcal{F}_t \)-measurable, i.e the process is adapted to the filtration,

(ii) \( \mathbb{E}[|X(t)|] < \infty \), and

(iii) \( \mathbb{E}[X(t)|\mathcal{F}_s] = X(s) \) with probability one for all \( t \geq s \),

where \( \mathbb{E}[\cdot] \) is the expectation value.

Since a central part of our approach is a time-continuous Markov chain we now want to define what we understand by that. We start with the essential property of a Markov chain, the Markov property.

**Definition 3.4 (Markov Property)**
We say that a stochastic process \( \{X(t) : t \geq 0\} \) defined on a probability space \((X, \mathcal{F}, \mathbb{P})\), with values in a finite space \( S = \{0, \ldots, N\} \), the state space of the process, has the **Markov property** if and only if, for any finite set \( 0 \leq t_1 < t_2 < \cdots < t_n < t_n+1 \) of times and corresponding set \( i_1, i_2, \ldots, i_{n-1}, i, j \) of states in \( S \) with \( \mathbb{P}[X(t_n) = i, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1] > 0 \), we have

\[
\mathbb{P}[X(t_{n+1}) = j|X(t_n) = i, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1] = \mathbb{P}[X(t_{n+1}) = j|X(t_n) = i].
\]

We now introduce the notion of continuous-time Markov chains, as well as some properties and closely related objects of them.

**Definition 3.5 (Continuous-Time Markov Chain and transition matrix)**
A jump process \( \{X(t) : t \geq 0\} \) is called **continuous-time Markov chain** if it fulfills the Markov property Equation (3.4).

We denote by \( p_{i,j}(t, s) \) the transition probability \( \mathbb{P}[X(t) = j|X(s) = i] \) for any \( i, j \in S \) and \( t \geq s \geq 0 \). We use \( P(t, s) \) for the matrix \( (p_{i,j}(t, s))_{i,j=0}^N \), named the **transition matrix** or **transition function** of the Markov chain \( X(t) \), and postulate that

\[
\lim_{t \to s^+} p_{i,j}(t, s) = \delta_{i,j},
\]

where \( \delta_{i,j} = 1 \) if \( i = j \) and 0 otherwise.

If for all \( s, t \) such that \( 0 \leq s \leq t \) and all \( i, j \in S \) the conditional probability \( \mathbb{P}[X(t) = j|X(s) = i] \) appearing on the right-hand side of Equation (3.4) depends only on \( t - s \) and not on \( s \) and \( t \) individually, we say that the process \( \{X(t) : t \geq 0\} \) is **homogeneous** or has **stationary transition probabilities**. Otherwise we call the process **nonhomogeneous**.

In the case of a homogeneous Markov chain we have

\[
\mathbb{P}[X(t) = j|X(s) = i] = \mathbb{P}[X(t-s) = j|X(0) = i],
\]

and the transition function reduces to

\[
p_{i,j}(t) := \mathbb{P}[X(t) = j|X(0) = i] \quad \text{for all } i, j \in S, \quad t \geq 0.
\]
From this definition it follows that, for $0 \leq s \leq \xi \leq t$,
\begin{align*}
p_{i,j}(t, s) &\geq 0, \quad i, j \in S, \\
\sum_{j \in S} p_{i,j}(t, s) &= 1, \quad i \in S, \\
p_{i,j}(t, s) &= \sum_{k \in S} p_{i,k}(\xi, s)p_{k,j}(t, \xi), \quad i, j \in S.
\end{align*}
(3.8)

Usually Equation (3.8) is referred to as Chapman-Kolmogorov equation.

Since we want to incorporate time-dependent aspects in our model we will mostly focus on nonhomogeneous Markov chains.

Before we finally come to the definition of a generator of a Markov chain we need to specify some properties of the object we want to introduce later.

**Definition 3.6 (q-Property)**

Let $Q(t) = (q_{i,j}(t))$, for $t \geq 0$ and $i, j \in S$. Then $Q(t)$ satisfies the q-Property, if

(i) $q_{i,j}(t)$ is Borel measurable for all $i, j \in S$ and $t \geq 0$,

(ii) $q_{i,j}(t)$ is uniformly bounded, that is, there exists a constant $c$ such that $|q_{i,j}(t)| \leq c$, for all $i, j \in S$ and $t \geq 0$,

(iii) $q_{i,j}(t) \geq 0$ for $j \neq i$ and $q_{i,i}(t) = -\sum_{j \neq i} q_{i,j}(t)$, $t \geq 0$.

Now we will define what we understand by a generator of a Markov process.

**Definition 3.7 (Generator)**

A matrix $Q(t)$, $t \geq 0$, is an infinitesimal generator, or simply generator, of a stochastic process $X(t)$ if it satisfies the q-Property, and for all bounded real-valued functions $f$ defined on $S$, the process
\begin{equation}
f(X(t)) - \int_0^t \sum_{j \in S} q_{X(\xi),j}(\xi)f(j) \, d\xi
\end{equation}
(3.9)
is a martingale.

For a finite state space $S$, as we will use it, according to [15] this can be simplified with the following lemma. The proof is also provided in [15].

**Lemma 3.8**

Let $S = \{0, \ldots, N\}$. Then $X(t) \in S$, $t \geq 0$, is a Markov chain generated by $Q(t)$ if and only if, the process
\begin{equation}
(I_{\{X(t)=0\}}, \ldots, I_{\{X(t)=N\}}) - \int_0^t (I_{\{X(\xi)=1\}}, \ldots, I_{\{X(\xi)=N\}})Q(\xi) \, d\xi
\end{equation}
(3.10)
is a martingale.
In [15] the authors also describe a way to construct a Markov chain from the time-dependent generator \( Q(t) \) by a piecewise-deterministic process approach. The connection between the constructed process, the generator and the Kolmogorov differential equations is given by the next theorem.

**Theorem 3.9**

Let \( Q(t) \) satisfy the q-Property for \( t \geq 0 \).

Then

(i) The process \( X(\cdot) \) constructed by the piecewise-deterministic process approach described in Section 2.4.1 of [15] is a Markov chain.

(ii) The process

\[
(3.11) \quad f(X(t)) - \int_0^t \sum_{j \in S} q_{X(\xi),j}(\xi)f(j) \, d\xi
\]

is a martingale for any uniformly bounded function \( f(\cdot) \) on \( S \). Thus \( Q(t) \) is indeed the generator of \( X(\cdot) \).

(iii) The transition matrix \( P(t,s) \) satisfies the forward differential equation

\[
(3.12) \quad \frac{\partial P(t,s)}{\partial t} = P(t,s) \cdot Q(t), \quad t \geq s, \\
P(s,s) = I,
\]

where \( I \) is the identity matrix.

(iv) Assume further that \( Q(t) \) is continuous in \( t \). Then \( P(t,s) \) also satisfies the backward differential equation

\[
(3.13) \quad \frac{\partial P(t,s)}{\partial s} = -Q(s) \cdot P(t,s), \quad t \geq s, \\
P(t,t) = I.
\]

The proof can again be found in [15] and cited literature therein. The system in Equation (3.12) is known as Kolmogorov forward equation, like we also described it in Section 2.2.

We will now give another useful result for our study of the behaviour of a Markov chain and the associated solutions to the forward differential equation. It will allow us to have a probabilistic interpretation of just these solutions.

So, let \( Q(t) \in \mathbb{R}^{N+1 \times N+1} \) be a generator and \( X(t) \) a finite-state Markov chain with state space \( S = \{0, \ldots, N\} \) and generator \( Q(t) \). We denote the probability distribution of the underlying chain at time \( t \) by the row vector

\[
(3.14) \quad p(t) = \left( \mathbb{P}[X(t) = 0], \ldots, \mathbb{P}[X(t) = N] \right) \in \mathbb{R}^{1 \times N+1}.
\]
By Theorem 3.9 we see, that \( p(t) \) is a solution of the forward equation

\[
\frac{\partial p(t)}{\partial t} = p(t) \cdot Q(t),
\]

(3.15)

\( p(0) = p^0 \) such that \( p^0_i \geq 0 \) for each \( i \), and \( \sum_{i=0}^{N} p^0_i = 1 \),

where \( p^0 = (p^0_0, \ldots, p^0_N) \) and \( p^0_i \) denotes the \( i \)-th component of \( p^0 \). By this we see that studying the probability distribution is equivalent to examining the solution of Equation (3.15). From this we get the following result.

**Lemma 3.10**

*The solution \( p(t) \) of Equation (3.15) satisfies*

\[
0 \leq p_i(t) \leq 1 \quad \text{and} \quad \sum_{i=0}^{N} p_i(t) = 1.
\]

(3.16)

A justification of this can be found in [15].

Next, we want to introduce the necessary notions to study the long-term behaviour of a Markov chain. This is essential to give an answer to the question of free parking spaces, as we discussed in Section 2.2.

In the classical setup with a time constant generator matrix we need irreducibility of the process or the generator and the therefore existing stationary distribution for that. For more details in the constant case see [20]. Due to the fact that we are dealing with time-dependent generators and therefore nonhomogeneous Markov chains the notion of stationary distributions to determine long-term behaviour is no longer sufficient. Also, the notion of irreducibility needs to be generalized a little bit. We start by defining weak irreducibility.

**Definition 3.11** (Weak Irreducibility)

We call a generator \( Q(t) \) weakly irreducible if, for each fixed \( t \geq 0 \), the system of equations

\[
\nu(t) \cdot Q(t) = 0, \quad \sum_{i=0}^{N} \nu_i(t) = 1,
\]

(3.17)

has a unique solution \( \nu(t) = (\nu_0(t), \ldots, \nu_N(t)) \in \mathbb{R}^{1 \times N+1} \) and \( \nu_i(t) \geq 0 \) for all \( i = 0, \ldots, N \).

Next, we introduce the generalization of stationary distributions.

**Definition 3.12** (Quasi-Stationary Distribution)

For \( t \geq 0 \), \( \nu(t) \) is called a quasi-stationary distribution if it is a solution of Equation (3.17) satisfying \( \nu(t) \geq 0 \).

We can see that from Definition 3.11 and Definition 3.12 the existence of a quasi-stationary distribution is immediate if the generator is weakly irreducible. The additional requirement of \( \nu(t) \geq 0 \) in Definition 3.12 allows the probabilistic interpretation we need for our application.
3. Mathematical Framework

purpose, although this is also possible due to Lemma 3.10.

After generalizing the notions we need to study long-term behaviour of a Markov chain, we would like to modify our forward differential equation in a way that this is actually possible also in the time-dependent case. For that we follow Section 4.2 of [15].

Let $Q(t) \in \mathbb{R}^{N+1 \times N+1}$ be a generator, $\varepsilon > 0$ a small parameter and suppose that $X^\varepsilon(t)$ is a finite-state Markov chain with state space $S = \{0, \ldots, N\}$ generated by $Q^\varepsilon(t) = Q(t)/\varepsilon$. Again we denote the probability distribution of the underlying chain at time $t$ by $p^\varepsilon(t) = (\mathbb{P}[X^\varepsilon(t) = 0], \ldots, \mathbb{P}[X^\varepsilon(t) = N]) \in \mathbb{R}^{1 \times N+1}$. With Theorem 3.9 we get that $p^\varepsilon(t)$ is the solution to the forward equation

$$
\frac{\partial p^\varepsilon(t)}{\partial t} = \frac{1}{\varepsilon} p^\varepsilon(t) \cdot Q(t),
$$

(3.18)

where $p^0 = (p^0_0, \ldots, p^0_N)$ and $p^0_i$ denotes the $i$-th component of $p^0$. This gives us the possibility of studying the probability distribution by examining the solution of Equation (3.18).

When studying this kind of perturbed system we have the problem that when we consider the situation $\varepsilon \to 0$, the chains behaviour gets more and more volatile with $\varepsilon$ getting smaller and it will be difficult to analyze the fast changing process $X^\varepsilon(t)$. However, as we just pointed out, we can circumvent this problem by looking at the limit properties instead. So, we look at a type of average in contrast to the process on its own. Therefore, we need to understand the asymptotic properties of $p^\varepsilon(t)$, in particular the limit behaviour as $\varepsilon \to 0$. We will see that there exists a crucial difference between the case when our generator matrix $Q(t) = Q$ is a constant matrix to the time-dependent case. This is based on the existence of an analytic solution of Equation (3.18) in the constant case. We will also revisit this issue in Chapter 4.

As proposed in [15], for the time-dependent case we divide our time interval $[0, T]$ into two parts. One part will be for $t$ in an $\varepsilon$ neighbourhood of 0 and the other one for $t$ being bounded away from 0. As the behaviour of $p^\varepsilon(t)$ differs significantly in these two parts we make use of the matched asymptotic expansion. The expansion will also prove useful in understanding the connection between the quasi-stationary distribution and its constant case counterpart, the traditional stationary distribution.

To derive results about the asymptotic expansions of the probability distribution $p^\varepsilon(t)$ we need to make two essential assumptions:

(1) Given $0 < T < \infty$, for each $t \in [0, T]$, $Q(t)$ is weakly irreducible, that is, the system of equations

$$
f(t) \cdot Q(t) = 0,
$$

(3.19)

$$
\sum_{i=0}^{N} f_i(t) = 1
$$

has a unique nonnegative solution.
3.1. Markov Chains

For some \( n \in \mathbb{N} \), \( Q(\cdot) \) is \((n+1)\)-times continuously differentiable on \([0, T]\), and \((\partial^{n+1}/\partial t^{n+1})Q(\cdot)\) is Lipschitz on \([0, T]\).

The second assumption gives us sufficient smoothness of the generator matrix \( Q(t) \) for the asymptotic expansion.

The following lemma will help us understand the kind of results we will get from Theorem 3.14 in terms of the approximations to the probability distribution \( p^\varepsilon(\cdot) \). It is taken from [15] where the proof is provided as well.

**Lemma 3.13**

Let \( Q \in \mathbb{R}^{N+1 \times N+1} \) be a generator of a (homogeneous or stationary) finite-state Markov chain that is weakly irreducible. Consider the forward matrix differential equation

\[
\frac{\partial P(t)}{\partial t} = P(t) \cdot Q, \quad P(0) = I,
\]

where \( P(t) \in \mathbb{R}^{N+1 \times N+1} \) and \( I \) the identity matrix.

Then we have \( P(t) \to \bar{P} \) as \( t \to \infty \) and

\[
\left| \exp(Qt) - \bar{P} \right| \leq c \exp(-\gamma t) \quad \text{for some } \gamma, c > 0,
\]

where \( \bar{P} = 1(\nu_0, \ldots, \nu_N) \), \( 1 = (1, \ldots, 1)' \in \mathbb{R}^{N+1 \times 1} \) and \((\nu_0, \ldots, \nu_N)\) is the quasi-stationary distribution of the Markov process with generator \( Q \).

When going to the time-dependent case we now seek an asymptotic expansion of our probability distribution \( p^\varepsilon(\cdot) \) of the form

\[
p^\varepsilon(t) = \Phi^\varepsilon_n(t) + \Psi^\varepsilon_n \left( \frac{t}{\varepsilon} \right) + e^\varepsilon_n(t),
\]

where \( e^\varepsilon_n(t) \) is the remainder,

\[
\Phi^\varepsilon_n(t) = \varphi_0(t) + \varepsilon \varphi_1(t) + \cdots + \varepsilon^n \varphi_n(t),
\]

and

\[
\Psi^\varepsilon_n \left( \frac{t}{\varepsilon} \right) = \psi_0 \left( \frac{t}{\varepsilon} \right) + \varepsilon \psi_1 \left( \frac{t}{\varepsilon} \right) + \cdots + \varepsilon^n \psi_n \left( \frac{t}{\varepsilon} \right),
\]

with the functions \( \varphi_i(\cdot) \) and \( \psi_i(\cdot) \) to be determined later. For these functions we get the following properties:

**Theorem 3.14**

Let assumptions (I) and (II) be satisfied. Denote the unique solution of Equation (3.18) by \( p^\varepsilon(\cdot) \).

Then we can construct two sequences of functions \( \varphi_i(\cdot) \) and \( \psi_i(\cdot) \), \( 0 \leq i \leq n \), such that

(i) \( \varphi_i(\cdot) \) is \((n + 1 - i)\)-times continuously differentiable on \([0, T]\),

(ii) for each \( i = 0, \ldots, n \), there is a \( \gamma > 0 \) such that

\[
\left| \psi_i \left( \frac{t}{\varepsilon} \right) \right| \leq c \exp \left( -\frac{\gamma t}{\varepsilon} \right),
\]

with the functions \( \varphi_i(\cdot) \) and \( \psi_i(\cdot) \) to be determined later. For these functions we get the following properties:
(iii) the following estimate holds:

\[
\sup_{t \in [0,T]} \left| p^\varepsilon(t) - \sum_{i=0}^{n} \varepsilon^i \varphi_i(t) - \sum_{i=0}^{n} \varepsilon^i \psi_i \left( \frac{t}{\varepsilon} \right) \right| \leq c \varepsilon^{n+1}.
\]

The constructive prove can be found in [15]. We will just derive a corollary from this result, also presented in [15], that shows the connection of the expansion, in particular \( \varphi_0(t) \), to the quasi-stationary distribution \( \nu(t) \).

**Corollary 3.15**

Suppose \( Q(\cdot) \) is continuously differentiable on \([0, T]\), which satisfies assumption (I), and \((\partial/\partial t)Q(\cdot)\) is Lipschitz on \([0, T]\).

Then for all \( t > 0 \),

\[
\lim_{\varepsilon \to 0} p^\varepsilon(t) = \nu(t) = \varphi_0(t),
\]

i.e., \( p^\varepsilon(\cdot) \) converges to the quasi-stationary distribution.

By introducing a second time scale for the chain and together with Theorem 3.14 and Corollary 3.15 we got the results we need to justify using the quasi-stationary distribution as an indicator of the state of the chain in current time or the near future. They also provide us not only with the convergence itself but also with information about the convergence rate.

After we laid the groundwork for our model and how we will use it we now come to the closer investigation of how to determine the time-dependent generator matrix \( Q(t) \) from our data samples. Therefore, in the upcoming section we will introduce the necessary notions for the kernel approach to learn matrix-valued functions.

### 3.2. Kernel Approach

In this section we will lay out the framework for estimating the parameters of our model, i.e. the generator matrix, with a kernel approach. To do that we first introduce loss functions for matrices and their associated risks in Section 3.2.1. Afterwards we present the theory of reproducing kernel Hilbert spaces for matrix- or vector-valued functions in Section 3.2.2. Finally, in Section 3.2.3 we will see how we can combine these two parts to gain results about risk minimization and representation of our solution.

In the beginning we will introduce some notation and nomenclature we will make use of in the following sections.

So, let \((X, \mathcal{F})\) be a measurable space. Elements of \( X \) will be denoted by small letters, i.e. \( x, t \). By \((\mathbb{R}^{n \times m}, \langle \cdot, \cdot \rangle_\mathcal{M})\), we denote the Hilbert space of real-valued matrices of dimension \( n \) times \( m \). By the inner product we also get the induced norm for the Hilbert space via \( \| \cdot \|_\mathcal{M} := \sqrt{\langle \cdot, \cdot \rangle_\mathcal{M}} \).

Since we have a finite dimensional space the choice of the norm is not as crucial, due to the fact that all norms are equivalent in this case. Let \( \mathcal{M} \subseteq \mathbb{R}^{n \times m} \) be a closed subspace of the space of real matrices. This will be the target space of the functions we intend to study, i.e. \( f : X \to \mathcal{M} \). Also, let \( \mathcal{Y} \subseteq \mathbb{R}^{n \times m} \) be a closed subset of the space of matrices. From this set we will get the
3.2. Kernel Approach

samples of the target space, as part of the data set. Their elements will be denoted by capital letters, i.e. $A, B$. Furthermore we define the matrices $E_{i,j}$ componentwise as $(E_{i,j})_{k,l} = 1$ if $(k,l) = (i,j)$, $k = 1, \ldots, n$, $l = 1, \ldots, m$, and zero otherwise, which form a basis for the space of real matrices. We assume $E_{i,j} \in \mathcal{M}$. By not considering all of the $E_{i,j}$ we can also get a basis of certain subspaces.

In terms of notions concerning measurability we always use the Borel $\sigma$-algebra induced by the norm of the target space, i.e. $\mathcal{B}(\mathcal{M})$ for the target space $\mathcal{M}$, and the Lebesgue measure if not specified differently. The product measure is defined on the space $\mathcal{X} \times \mathcal{Y}$, $\mathcal{F} \otimes \mathcal{B}(\mathcal{Y})$ with the corresponding product $\sigma$-algebra. We denote the space of matrix-valued measurable functions $f : \mathcal{X} \to \mathcal{M}$ by $L^0(\mathcal{X}, \mathcal{M})$. We will also use the Banach spaces of $p$-integrable, matrix-valued functions which we denote by $(L^p(\mathcal{X}, \mathcal{M}, \mu), \| \cdot \|_{\mathcal{M}, p})$, $1 \leq p \leq \infty$, where

\[
\|f\|_{\mathcal{M}, p} := \left( \int_{\mathcal{X}} \|f(x)\|_{\mathcal{M}}^p \, d\mu(x) \right)^{\frac{1}{p}} \quad \text{and} \\
\|f\|_{\mathcal{M}, \infty} := \sup_{x \in \mathcal{X}} \|f(x)\|_{\mathcal{M}}.
\]

In case our target space is $\mathbb{R}$ we have the standard $L^p$ spaces of functions $f : \mathcal{X} \to \mathbb{R}$ with respect to a measure $\mu$, denoted by $(L^p(\mathcal{X}, \mu), \| \cdot \|_p)$, $1 \leq p \leq \infty$, where

\[
\|f\|_p := \left( \int_{\mathcal{X}} |f(x)|^p \, d\mu(x) \right)^{\frac{1}{p}} \quad \text{and} \\
\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|,
\]

as widely used in the literature on measure theory.

When we use the introduced framework in Chapter 4 we will choose $\mathcal{X} = [0, T] \subset \mathbb{R}$. Also $\mathcal{Y}$ and $\mathcal{M}$ will be specified more precisely later on (see Chapter 4), but we will give the results in this more general setting. In particular, the use of $\mathcal{M}$ as a real subspace of $\mathbb{R}^{n \times m}$ is intended to emphasize that the results we are providing here are given for general Hilbert spaces in the literature. So we can easily adapt them also for subspaces, which we will need for our application to estimate the generator matrix $Q(t)$ of our Markov process.

3.2.1. Loss Functions and Associated Risks

Now, we want to closer investigate the behaviour of loss functions for matrices and how their properties, like continuity, convexity, integrability and differentiability relate to the properties of the corresponding risk functional. To do this we follow, in outline, Chapter 2 about loss functions and their risks in [34]. The general aim, as stated before, is to find a function $f : \mathcal{X} \to \mathcal{M}$ such that for $(x, \mathcal{Y}) \in \mathcal{X} \times \mathcal{Y}$ the matrix $f(x)$ is a good approximation of $\mathcal{Y}$ at $x$. First, we define what we want to understand by "good".

**Definition 3.16 (Matrix-valued loss function)**

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and $\mathcal{Y} \subseteq \mathbb{R}^{n \times m}$ closed.

Then $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{M} \to [0, \infty)$ is called **matrix-valued loss function** or just **loss** if it is measurable.
3. Mathematical Framework

So, the loss function \( L(x, Y, f(x)) \) can be interpreted as the cost or loss of predicting \( Y \) by \( f(x) \) if we observe \( x \). Or, the smaller the value \( L(x, Y, f(x)) \), the better is the choice of our approximating function \( f \) in regard to \( L \). A basic example for a loss function for matrices is the squared Frobenius-Norm \( \| \cdot \|_F^2 \) which can be used to adapt the least squares approach for the matrix-valued case. We will give some more details about this in Chapter 4.

Since we want our approximation or prediction to be good in average, we need a small average loss for, so far, unknown observations \((x, Y)\). Therefore we use the following definition.

**Definition 3.17 (L-risk)**

Let \( L: X \times Y \times M \to [0, \infty) \) be a loss function and \( P \) be a probability measure on \( X \times Y \).

Then, for a measurable function \( f: X \to M \) the **L-risk** is defined by

\[
R_{L,P}(f) := \int_{X \times Y} L(x, Y, f(x)) \, dP(x, Y).
\]  
(3.30)

The function \((x, Y) \mapsto L(x, Y, f(x))\) is measurable by assumption and since it is also non-negative the integral over \( X \times Y \) defined above always exists, but it is not necessarily finite.

The next lemma shows that we can also rewrite the L-risk as a double integral.

**Lemma 3.18**

Let \( L: X \times Y \times M \to [0, \infty) \) be a loss function and \( P \) be a probability measure on \( X \times Y \).

Then, for a measurable function \( f: X \to M \) we have the following equality for the L-risk

\[
R_{L,P}(f) = \int_{X \times Y} L(x, Y, f(x)) \, dP(x, Y)
\]

\[
= \int_X \int_Y L(x, Y, f(x)) \, dP(Y|x) \, dP_X(x).
\]  
(3.31)

**Proof.** The assumed measurability and non-negativity of \( L \) guarantee the existence of the integral. Also since our space \( Y \) is closed Theorem A.1 ensures the existence of the conditional probability \( P(Y|x) \) we used in the inner integral. The equality follows from Fubini-Tonelli’s theorem. \( \square \)

For a given data set \( D := \{(x_i, Y_i) \mid i = 1, \ldots, M\} \subset X \times Y \) we use the measure

\[
D = \frac{1}{M} \sum_{i=1}^M \delta_{(x_i, Y_i)},
\]  
(3.32)

where \( \delta_{(x_i, Y_i)} \) denotes the Dirac measure at \((x_i, Y_i)\). By this we have an empirical measure supported at the points of the given data set and therefore get the **empirical L-risk**

\[
R_{L,D}(f) = \frac{1}{M} \sum_{i=1}^M L(x_i, Y_i, f(x_i)).
\]  
(3.33)

By the law of large numbers we get that \( R_{L,D}(f) \) is close to \( R_{L,P}(f) \), i.e. \( |R_{L,D}(f) - R_{L,P}(f)| \) is small with high probability for measurable \( f \) and \( R_{L,P}(f) \)<\( \infty \).

As we interpret \( L \) as our cost function we are looking for the minimum risk.
Definition 3.19 (Bayes risk and Bayes decision function)

Let \( L : X \times Y \times \mathcal{M} \rightarrow [0, \infty) \) be a loss function and \( \mathbb{P} \) a probability measure on \( X \times Y \).

Then the minimal \( L \)-risk

\[
R^*_L : = \inf \{ R_{L, \mathbb{P}}(f) \mid f : X \rightarrow \mathcal{M} \text{ measurable} \}
\]

is called **Bayes risk** with respect to \( L \) and \( \mathbb{P} \).

Also, a measurable \( f^*_L : X \rightarrow \mathcal{M} \) with

\[
R_{L, \mathbb{P}}(f^*_L) = R^*_L
\]

is called **Bayes decision function**.

Finding the function \( f^*_L \) is not directly possible. A first complication we have to deal with is that in general we do not know the distribution \( \mathbb{P} \) that generates the input and output pairs, i.e. the value pairs \((x, Y)\). To deal with that problem we introduced the empirical risk \( R_{L, \mathbb{D}}(f) \) based on a given data set \( \mathbb{D} \) in Equation (3.33). As already pointed out, the law of large numbers shows that this risk provides an approximation of \( R_{L, \mathbb{P}}(f) \) for every single \( f : X \rightarrow \mathcal{M} \), solving

\[
\inf_{f : X \rightarrow \mathcal{M}} R_{L, \mathbb{D}}(f),
\]

but it does not give us an approximate minimizer of \( R_{L, \mathbb{P}}(\cdot) \). This can be observed when we consider the function that is equal to \( Y_i \) in \( x_i \) for all \( i = 1, \ldots, M \), and zero everywhere else. This function solves Equation (3.36) but is a very poor choice to approximate Equation (3.34) since it shows an extreme example of a phenomenon called **overfitting**. This basically describes the situation when the function we are looking for is modeling the output values in the data set too closely and is therefore not very useful for future data, i.e. providing a prediction.

To prevent this problem we choose a smaller set \( \mathfrak{S} \) of functions \( f : X \rightarrow \mathcal{M} \) than \( L^0(X, \mathcal{M}) \) that we expect to give a good approximation of \( R_{L, \mathbb{P}}(\cdot) \) (Equation (3.34)). So instead of minimizing \( R_{L, \mathbb{D}}(\cdot) \) over all measurable functions we only minimize over the chosen set of functions \( \mathfrak{S} \), i.e.

\[
\inf_{f \in \mathfrak{S}} R_{L, \mathbb{D}}(f).
\]

This approach is called **empirical risk minimization** (ERM) and tends to produce approximate solutions of the counterpart with infinite samples, i.e.

\[
\inf_{f \in \mathfrak{S}} R_{L, \mathbb{P}}(f),
\]

so the Bayes risk restricted to the chosen set of functions \( \mathfrak{S} \). Of course, choosing this set of functions is critical in terms of the **approximation error**

\[
R^*_{L, \mathbb{P}, \mathfrak{S}} := \inf_{f \in \mathfrak{S}} R_{L, \mathbb{P}}(f),
\]

\[
R^*_{L, \mathbb{P}} - R^*_{L, \mathbb{P}, \mathfrak{S}}.
\]

Also, computational feasibility of Equation (3.37) is an issue that needs to be taken into account. These two issues are briefly addressed in Section 5.2, but will not be in the focus of this thesis.

After introducing the basic objects for our studies and giving an overview of what we seek to achieve, we now want to investigate how properties of the loss function transfer to the associated risk. We want to remind that measurability of a vector- or matrix-valued function is a
3. Mathematical Framework

We first come to the measurability of risks. The following lemma shows under which conditions this is the case.

**Lemma 3.20 (Measurability of risks)**

Let \( L: X \times Y \times \mathcal{M} \to [0, \infty) \) be a loss function and \( \mathcal{G} \subset L^0(X, \mathcal{M}) \) a subset equipped with a complete and separable metric \( d \) and corresponding \( \sigma \)-algebra. Let \( d \) dominate pointwise convergence, i.e.

\[
\lim_{k \to \infty} d(f_k, f) = 0 \quad \text{implies} \quad \lim_{k \to \infty} f_k(x) = f(x), \quad x \in X
\]

for all \( f, f_k, \in \mathcal{G} \).

Then the evaluation map

\[ \mathcal{G} \times X \to \mathcal{M} \]

\[ (f, x) \mapsto f(x) \]

is measurable, and therefore the map \( (x, Y, f) \mapsto L(x, Y, f(x)) \) defined on \( X \times Y \times \mathcal{G} \) is also measurable. Given a distribution \( \mathbb{P} \) on \( X \times Y \) the risk functional \( R_{L, \mathbb{P}}: \mathcal{G} \to [0, \infty] \) is measurable.

**Proof.** We will use the arguments componentwise where necessary. We have that \( d \) dominates pointwise convergence, so for fixed \( x \in X \) the \( \mathcal{M} \)-valued map \( f \mapsto f(x) \) defined on \( \mathcal{G} \) is componentwise continuous, i.e. continuous with respect to \( d \). Also \( \mathcal{G} \subset L^0(X, \mathcal{M}) \) implies that for fixed \( f \) the \( \mathcal{M} \)-valued map \( x \mapsto f(x) \) defined on \( X \) is componentwise measurable, i.e. measurable. By the Lemma of Carathéodory (Theorem A.2) we get the first assertion.

This implies that the map \( (x, Y, f) \mapsto (x, Y, f(x)) \) is measurable and we obtain the second assertion.

Assertion three follow’s from Fubini-Tonelli’s Theorem.

For example, a metric \( d \) induced via \( d(f, g) := \| f - g \|_{M, \infty} \) by the supremum norm for matrix-valued functions, dominates pointwise convergence if we choose \( \mathcal{G} = C^0_b(X, \mathcal{M}) \subset L^\infty(X, \mathcal{M}) \), the set of continuous and bounded functions. So not only the choice of the set of functions \( \mathcal{G} \) is crucial, but also that we can equip it with the right kind of metric. We will see later on, that metrics induced by the norms of reproducing kernel Hilbert spaces, which we will discuss in Section 3.2.2, fulfill this property as well.

Now we will investigate the property of convexity for loss functions.

**Definition 3.21 ((Strict) convexity of a loss function)**

A loss \( L: X \times Y \times \mathcal{M} \to [0, \infty) \) is called (strictly) convex if \( L(x, Y, \cdot): \mathcal{M} \to [0, \infty) \) is (strictly) convex for all \( x \in X \) and \( Y \in \mathcal{Y} \).

The next lemma shows that convexity of the loss function implies convexity of the corresponding risk.

**Lemma 3.22 (Convexity of risks)**

Let \( L: X \times Y \times \mathcal{M} \to [0, \infty) \) be a (strictly) convex loss function on \( X \times Y \).

Then \( R_{L, \mathbb{P}}: L^0(X, \mathcal{M}) \to [0, \infty) \) is (strictly) convex.
3.2. Kernel Approach

Proof. Let \( f_1, f_2 \in \mathcal{L}^0(X, \mathcal{M}) \) and \( \beta \in [0, 1] \). Then we have

\[
\mathcal{R}_{L,p}(\beta f_1 + (1 - \beta) f_2) = \int_{X \times \mathcal{Y}} L(x, Y, \beta f_1(x) + (1 - \beta) f_2(x)) \, d\mathbb{P}(x, Y)
\]

\[
\leq \int_{X \times \mathcal{Y}} \beta L(x, Y, f_1(x)) + (1 - \beta) L(x, Y, f_2(x)) \, d\mathbb{P}(x, Y)
\]

\[
= \beta \mathcal{R}_{L,p}(f_1) + (1 - \beta) \mathcal{R}_{L,p}(f_2).
\]

(3.42)

The same arguments hold for strict convexity. \( \square \)

Another property we would like to have a closer look at is continuity. We start with the definition.

Definition 3.23 (Continuity of a loss function)

A loss \( L: X \times \mathcal{Y} \times \mathcal{M} \to [0, \infty) \) is called continuous if \( L(x, Y, \cdot): \mathcal{M} \to [0, \infty) \) is continuous for all \( x \in X \) and \( Y \in \mathcal{Y} \).

If our loss function is continuous, then we have for \((f_k)_{k \in \mathbb{N}} \subset \mathcal{L}^0(X, \mathcal{M}) \) with \( f_k(x) \to f(x) \) that \( L(x, Y, f_k(x)) \to L(x, Y, f(x)) \).

However, this does not imply convergence of the integrals and thus \( \mathcal{R}_{L,p}(f_k) \to \mathcal{R}_{L,p}(f) \). The following lemma gives a weaker result that always holds.

Lemma 3.24 (Lower semi-continuity of risks)

Let \( L: X \times \mathcal{Y} \times \mathcal{M} \to [0, \infty) \) be a continuous loss function, \( \mathbb{P} \) a probability distribution on \( X \times \mathcal{Y} \), \((f_k)_{k \in \mathbb{N}} \subset \mathcal{L}^0(X, \mathcal{M}) \), \( f \in \mathcal{L}^0(X, \mathcal{M}) \) with

\[
\lim_{k \to \infty} \mathbb{P}_X\left\{ x \in X : \|f_k(x) - f(x)\|_{\mathcal{M}} \geq \varepsilon \right\} = 0
\]

for all \( \varepsilon > 0 \) (i.e. convergence in probability).

Then

\[
\mathcal{R}_{L,p}(f) \leq \liminf_{k \to \infty} \mathcal{R}_{L,p}(f_k).
\]

(3.44)

Proof. For any accumulation point of a sequence we can choose a subsequence which converges to this point, in particular, there exists \((f_{k_l})_{l \in \mathbb{N}}\) of \((f_k)_{k \in \mathbb{N}}\) such that

\[
\lim_{l \to \infty} \mathcal{R}_{L,p}(f_{k_l}) = \liminf_{k \to \infty} \mathcal{R}_{L,p}(f_k).
\]

(3.45)

The chosen subsequence \((f_{k_l})_{l \in \mathbb{N}}\) still converges to \( f \) in probability. From probability theory (see e.g. [33]) we know that convergence in probability implies that there exists a subsequence that converges almost surely, so there exists \((f_{k_{l_i}})_{i \in \mathbb{N}}\) such that

\[
\|f_{k_{l_i}}(x) - f(x)\|_{\mathcal{M}} \to 0 \quad \text{for } \mathbb{P}_X\text{-almost all } x \in X.
\]

(3.46)

This sub-subsequence fulfills Equation (3.45) as well as the pointwise convergence, i.e. Equation (3.46). For convenience in notation we will denote it with \((f_{k_i})_{i \in \mathbb{N}}\).
We have $L(x,Y,f_k(x)) \to L(x,Y,f(x))$ for $\mathbb{P}$-almost all $(x,Y) \in X \times Y$ by continuity of $L$ and therefore by Fatou’s lemma

$$R_{L,\mathbb{P}}(f) = \int_{X \times Y} \lim_{l \to \infty} L(x,Y,f_k(x)) \, d\mathbb{P}(x,Y)$$

$$\leq \liminf_{l \to \infty} \int_{X \times Y} L(x,Y,f_k(x)) \, d\mathbb{P}(x,Y)$$

$$= \liminf_{l \to \infty} R_{L,\mathbb{P}}(f_k)$$

$$= \liminf_{k \to \infty} R_{L,\mathbb{P}}(f_k),$$

where the last equality is due to Equation (3.45).

Now, we want to see what is necessary to have a majorant for $L(\cdot,\cdot,f_k(\cdot))$ to apply Lebesgue’s theorem and get $R_{L,\mathbb{P}}(0) < \infty$ and we get $R_{L,\mathbb{P}}^* < \infty$.

Next, we investigate the continuity of Nemitski loss based risks.

**Definition 3.25 (Nemitski loss (of order) / $\mathbb{P}$-integrable Nemitski loss)**

*We call a loss $L : X \times Y \times M \to [0,\infty)$ a Nemitski loss if there exist* $b : X \times Y \to [0,\infty)$ *measurable and $h : [0,\infty) \to [0,\infty)$ monotonously increasing such that*

$$L(x,Y,C) \leq b(x,Y) + h(C) \quad \text{for all } (x,Y,C) \in X \times Y \times M.$$  

*Furthermore, $L$ is a Nemitski loss of order $p \in (0,\infty)$ if there is a constant $c > 0$ such that*

$$L(x,Y,C) \leq b(x,Y) + c \|C\|^p_M \quad \text{for all } (x,Y,C) \in X \times Y \times M.$$  

*Finally, if $\mathbb{P}$ is a distribution on $X \times Y$, $b \in L^1(X \times Y, \mathbb{P})$ and $L$ satisfies Equation (3.50), then $L$ is called a $\mathbb{P}$-integrable Nemitski loss function.*

For $\mathbb{P}$-integrable Nemitski loss functions of order $p$ and $f \in L^p(X,M)$ we have

$$R_{L,\mathbb{P}}^*(f) = \int_{X \times Y} L(x,Y,f(x)) \, d\mathbb{P}(x,Y)$$

$$\leq \int_{X \times Y} b(x,Y) + c \|f(x)\|^p_M \, d\mathbb{P}(x,Y)$$

$$= \|b\|_1 + \|f\|_{M,p} < \infty.$$  

Thus $R_{L,\mathbb{P}}^*(0) < \infty$ and we get $R_{L,\mathbb{P}}^* < \infty$.  

Next, we investigate the continuity of Nemitski loss based risks.
Lemma 3.26 (Continuity of risks)

Let $\mathbb{P}$ be a distribution on $X \times Y$ and $L: X \times Y \times \mathcal{M} \to [0, \infty)$ be a continuous, $\mathbb{P}$-integrable Nemitski loss. Then the following hold:

(i) Let $f_k: X \to \mathcal{M}$ be a sequence of bounded, measurable functions with $\|f_k\|_{\mathcal{M}, \infty} \leq c$ for all $k \geq 1$. If $(f_k)_{k \in \mathbb{N}}$ converges $\mathbb{P}_X$-almost surely to $f: X \to \mathcal{M}$, i.e.

$$\lim_{k \to \infty} \mathbb{P}_X \left( \left\{ x \in X : \sup_{l \geq k} \| f_l(x) - f(x) \|_{\mathcal{M}} > \varepsilon \right\} \right) = 0,$$

for all $\varepsilon > 0$. Then

$$\lim_{k \to \infty} R_{L, \mathbb{P}}(f_k) = R_{L, \mathbb{P}}(f).$$

(ii) $R_{L, \mathbb{P}}: L^\infty(X, \mathcal{M}, \mathbb{P}_X) \to [0, \infty)$ is well defined and continuous.

(iii) If $L$ is of order $p \in [1, \infty)$, then $R_{L, \mathbb{P}}: L^p(X, \mathcal{M}, \mathbb{P}_X) \to [0, \infty)$ is well defined and continuous.

Proof. (i) We have $f_k(x) \to f(x)$ for $\mathbb{P}_X$-a.e. $x \in X$, thus $f$ is measurable and $\|f\|_{\mathcal{M}, \infty} \leq c$. The continuity of $L$ gives

$$\lim_{k \to \infty} \mathbb{P}_X \left( \left\{ x \in X : \sup_{l \geq k} \| f_l(x) - f(x) \|_{\mathcal{M}} > \varepsilon \right\} \right) = 0,$$

for $\mathbb{P}$-almost all $(x, Y) \in X \times Y$.

Also we have

$$\left| L(x, Y, f_k(x)) - L(x, Y, f(x)) \right| \leq 2b(x, Y) + h\left( \| f_k(x) \|_{\mathcal{M}} \right) + h\left( \| f(x) \|_{\mathcal{M}} \right)$$

$$\leq 2b(x, Y) + 2h(c),$$

for all $(x, Y) \in X \times Y$ and all $k \geq 1$.

Since $2b(\cdot, \cdot) + 2h(c)$ is $\mathbb{P}$-integrable, we constructed the necessary majorant and can use dominated convergence, which together with

$$\left| R_{L, \mathbb{P}}(f_k) - R_{L, \mathbb{P}}(f) \right| \leq \int_{X \times Y} \left| L(x, Y, f_k(x)) - L(x, Y, f(x)) \right| d\mathbb{P}(x, Y) \xrightarrow{k \to \infty} 0$$

gives the assertion.

(ii) We have

$$L(x, Y, C) \leq b(x, Y) + h(\| C \|_{\mathcal{M}}) \quad \text{for all } (x, Y, C) \in X \times Y \times \mathcal{M}.$$

Therefore

$$R_{L, \mathbb{P}}(f) \leq \int_{X \times Y} b(x, Y) + h\left( \| f(x) \|_{\mathcal{M}} \right) d\mathbb{P}(x, Y) < \infty,$$

i.e. $R_{L, \mathbb{P}}$ is well defined on $L^\infty(X, \mathcal{M})$. The continuity follows from (i).
(iii) Since $L$ is a $\mathbb{P}$-integrable Nemitski loss function of order $p$ we have $\mathcal{R}_{L,\mathbb{P}}(f) < \infty$, hence $\mathcal{R}_{L,\mathbb{P}}$ is well defined by Equation (3.52).

To prove continuity, let $(f_k) \subset L^p(X,\mathcal{M})$ be a convergent sequence with limit $f \in L^p(X,\mathcal{M})$. Convergence in $L^p$ implies convergence in probability, due to the Chebychev-Markov-Inequality (see Theorem 5.11 of [33]). We apply Lemma 3.24 and get

$$
\mathcal{R}_{L,\mathbb{P}}(f) \leq \liminf_{k \to \infty} \mathcal{R}_{L,\mathbb{P}}(f_k).
$$

(3.60)

Now, consider $\widetilde{L}(x,Y,C) := b(x,Y) + c\|C\|_{\mathcal{M}} - L(x,Y,C)$, which is a continuous loss function. Hence, Lemma 3.24 yields

$$
\|b\|_1 + c\|f\|_{\mathcal{M},p} - \mathcal{R}_{L,\mathbb{P}}(f) = \mathcal{R}_{L,\mathbb{P}}(f)
\leq \liminf_{k \to \infty} \mathcal{R}_{L,\mathbb{P}}(f_k)
$$

(3.61)

$$
= \liminf_{k \to \infty} \left(\|b\|_1 + c\|f_k\|_{\mathcal{M},p} - \mathcal{R}_{L,\mathbb{P}}(f_k)\right).
$$

By continuity of $\|\cdot\|_{\mathcal{M},p}$ on $L^p(X,\mathcal{M})$ we obtain

$$
c\|f\|_{\mathcal{M},p} - \mathcal{R}_{L,\mathbb{P}}(f) \leq \|\cdot\|_{\mathcal{M},p} - \liminf_{k \to \infty} \mathcal{R}_{L,\mathbb{P}}(f_k).
$$

(3.62)

Making use of $(f_k)$ being convergent yields,

$$
c\|f\|_{\mathcal{M},p} - \mathcal{R}_{L,\mathbb{P}}(f) \leq \|f\|_{\mathcal{M},p} - \liminf_{k \to \infty} \mathcal{R}_{L,\mathbb{P}}(f_k).
$$

(3.63)

This is equivalent to

$$
\liminf_{k \to \infty} \mathcal{R}_{L,\mathbb{P}}(f_k) \leq \mathcal{R}_{L,\mathbb{P}}(f),
$$

(3.64)

which finishes the proof.

Next, we turn to a more quantitative concept of continuity for loss functions.

**Definition 3.27** (*Local Lipschitz continuity* of a loss function)

A loss $L: X \times \mathcal{Y} \times \mathcal{M} \to [0, \infty)$ is called **locally Lipschitz continuous** if for all $a \geq 0$ there exists a constant $c_a \geq 0$ such that

$$
\sup_{x \in X, y \in \mathcal{Y}} |L(x,Y,C) - L(x,Y,C')| \leq c_a \|C - C'\|_{\mathcal{M}}
$$

holds, where $\|C\|_{\mathcal{M}}, \|C'\|_{\mathcal{M}} \leq a$.

Moreover, let

$$
|L|_{1,a} := \inf \{ c_a : \text{Equation (3.65) holds} \},
$$

(3.66)

then $L$ is called **Lipschitz continuous** if and only if $|L|_{1} := \sup_{a \geq 0} |L|_{1,a} < \infty$.

As the next lemma shows, Lipschitz continuity carries over from the loss function to the risk functional.
Lemma 3.28 (Lipschitz continuity of risks)
Let \( L: X \times Y \times M \to [0, \infty) \) be a locally Lipschitz continuous loss and \( \mathbb{P} \) a probability distribution on \( X \times Y \). Then for all \( a \geq 0 \) and all \( f, g \in \mathcal{L}(X, M) \) with \( \|f\|_{M, \infty} \leq a \) and \( \|g\|_{M, \infty} \leq a \) we have
\[
\|R_{L, \mathbb{P}}(f) - R_{L, \mathbb{P}}(g)\| \leq |L|_{a,1} \cdot \|f - g\|_{M,1}.
\]

Proof. We use the Lipschitz continuity of the loss function and get
\[
\|R_{L, \mathbb{P}}(f) - R_{L, \mathbb{P}}(g)\| \leq \int_{X \times Y} \left| L(x, Y, f(x)) - L(x, Y, g(x)) \right| \, d\mathbb{P}(x, Y)
\]
(3.68)
which is the desired result. \( \square \)

At last, we want to consider differentiability of risks. Therefore we, analogously to vector analysis, introduce different notions of differentiability of loss functions in the next definitions. We start with the total derivative. As for the other properties we studied, the differentiability of the loss function is also only concerned with the last argument of the loss.

Definition 3.29 (Total derivative of a loss function)
A loss \( L: X \times Y \times M \to [0, \infty) \) is called (totally) differentiable at \( C^* \in M \) if for all \( (x, Y) \in X \times Y \) there exists a linear mapping \( A_{(x,Y)}: M \to \mathbb{R} \) such that
\[
\lim_{C \to C^*} \frac{L(x, Y, C) - L(x, Y, C^*) - A_{(x,Y)}(C - C^*)}{\|C - C^*\|_M} = 0.
\]
(3.69)
We call \( A_{(x,Y)} = DL(x, Y, C^*) \) the total derivative or differential of \( L \) at \( C^* \in M \) for all \( (x, Y) \in X \times Y \).

If \( L \) is differentiable for all \( C^* \in M \) we call \( L \) differentiable in \( M \).

Next, we have a look at directional derivatives since we will use them to proof our upcoming lemma.

Definition 3.30 (Directional derivative of a loss function)
A loss \( L: X \times Y \times M \to [0, \infty) \) is said to have a directional derivative at \( C \in M \) in direction \( V \in \mathbb{R}^{n \times m} \) if the limit
\[
\lim_{h \to 0} \frac{L(x, Y, C + hV) - L(x, Y, C)}{h}
\]
(3.70)
exists for all \( (x, Y) \in X \times Y \) and we denote the directional derivative by
\[
D_V L(x, Y, C) = \lim_{h \to 0} \frac{L(x, Y, C + hV) - L(x, Y, C)}{h}.
\]
(3.71)
Just as in the case of vector calculus we get the special case of the partial derivatives if we choose \( V = E_{ij} \), i.e. a basis vector, for the direction.

Before we finally go to the result about differentiability of risks we introduce the gradient of a loss function. This also goes along the same idea as in vector calculus.
Definition 3.31 (Gradient of a loss function)
Let \( L : X \times Y \times \mathcal{M} \to [0, \infty) \) be loss function differentiable at \( C \in \mathcal{M} \). Then the gradient of \( L \) at \( C \in \mathcal{M} \) for all \((x, Y) \in X \times Y\), denoted by \( \nabla L(x, Y, C) \), is defined by

\[
\nabla L(x, Y, C), V \rangle_{\mathcal{M}} = D_L(x, Y, C)(V) \quad \text{for all } V \in \mathbb{R}^{n \times m}.
\]

In case we have the matrices \( E_{ij}, i = 1, \ldots, n, j = 1, \ldots, m \), as an orthonormal basis of \( \mathcal{M} \) then the gradient is formed as the sum of the partial derivatives times the basis vectors, i.e.

\[
\nabla L(x, Y, C) = \sum_{i=1}^{n} \sum_{j=1}^{m} D_{E_{ij}} L(x, Y, C) \cdot E_{ij} = \begin{bmatrix} D_{E_{11}} L(x, Y, C) & \cdots & D_{E_{1m}} L(x, Y, C) \\ \vdots & \ddots & \vdots \\ D_{E_{n1}} L(x, Y, C) & \cdots & D_{Emn} L(x, Y, C) \end{bmatrix} \in \mathbb{R}^{n \times m}.
\]

We also notice that for the case of loss functions we keep the relationship between gradient and directional derivative known from vector calculus, i.e.

\[
\nabla L(x, Y, C), V \rangle_{\mathcal{M}} = D_V L(x, Y, C),
\]

at \( C \in \mathcal{M} \) and for direction \( V \in \mathbb{R}^{n \times m} \).

In general the property of differentiability does not immediately transfer from the loss function to the risk but for some Nemitski loss functions we can acquire the differentiability of the associated risk.

Lemma 3.32 (Differentiability of risks)
Let \( \mathbb{P} \) be a probability distribution on \( X \times Y \) and \( L : X \times Y \times \mathcal{M} \to [0, \infty) \) be a totally differentiable loss such that both \( L \) and \( ||\nabla L||_{\mathcal{M}, \infty} \) are \( \mathbb{P} \)-integrable Nemitski losses.

Then the risk functional \( R_{L, \mathbb{P}} : L^\infty(X, \mathcal{M}) \to [0, \infty) \) is Fréchet differentiable and its derivative at \( f \in L^\infty(X, \mathcal{M}) \) is the bounded linear operator \( R'_{L, \mathbb{P}}(f) : L^\infty(X, \mathcal{M}) \to \mathbb{R} \) given by

\[
R'_{L, \mathbb{P}}(f) g = \int_{X \times Y} g(x) ||g(x)||_{\mathcal{M}} \cdot \langle \nabla L(x, Y, f(x)), g(x) \rangle_{\mathcal{M}} \, d\mathbb{P}(x, Y), \quad g \in L^\infty(X, \mathcal{M}).
\]

For a definition of Fréchet differentiability see Definition A.3

Proof. First, we show that the expression in our risk functional,

\[
||g(x)||_{\mathcal{M}} \cdot \langle \nabla L(x, Y, f(x)), g(x) \rangle_{\mathcal{M}}
\]

is measurable. For that we notice that the second factor is equal to the directional derivative as pointed out in Equation (3.74). Then we have that the directional derivative is the pointwise limit of measurable functions, it is therefore measurable. As the pointwise product of measurable functions the expression in whole is measurable.
3.2. Kernel Approach

Now, let $f \in L^\infty(X, M)$ and $(f_n) \subset L^\infty(X, M)$ be a sequence with $f_n \neq 0$, $n \geq 1$ and $\lim_{n \to \infty} \|f_n\|_{M, \infty} = 0$. Without loss of generality, we also assume for later use that $\|f_n\|_{M, \infty} \leq 1$ for all $n \geq 1$.

Next, we have a look at what we need to show for $\mathcal{R}'_{L, p}$ to be our Fréchet derivative. We want the following to converge, to zero:

\begin{equation}
\frac{1}{\|f_n\|_{M, \infty}} \left| \mathcal{R}_{L, p}(f + f_n) - \mathcal{R}_{L, p}(f) - \mathcal{R}'_{L, p}(f)f_n \right| \leq \int_{X \times Y} \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x)) - \|f_n(x)\|_M \cdot \left( \nabla L(x, Y, f(x)), f_n(x) \right)_M}{\|f_n\|_{M, \infty}} \right| \, d\mathbb{P}(x, Y).
\end{equation}

We now make use of the fact that $\|f(x)\|_M \leq \|f_n\|_{M, \infty}$ for all $x \in X$ and get

\begin{equation}
\int_{X \times Y} \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x)) - \|f_n(x)\|_M \cdot \left( \nabla L(x, Y, f(x)), f_n(x) \right)_M}{\|f_n(x)\|_M} \right| \, d\mathbb{P}(x, Y)
\end{equation}

\begin{align*}
&\leq \int_{X \times Y} \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x)) - \|f_n(x)\|_M \cdot \left( \nabla L(x, Y, f(x)), f_n(x) \right)_M}{\|f_n(x)\|_M} \right| \, d\mathbb{P}(x, Y) \\
&= \int_{X \times Y} \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x))}{\|f_n(x)\|_M} \right| \, d\mathbb{P}(x, Y)
\end{align*}

To show the convergence of this expression we will construct a majorant and use dominated convergence.

To do this, for $(x, Y) \in X \times Y$ and $n \geq 1$ we define

\begin{equation}
G_n(x, Y) := \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x))}{\|f_n(x)\|_M} \right| - \left( \nabla L(x, Y, f(x)), f_n(x) \right)_M,
\end{equation}

if $\|f_n(x)\|_M \neq 0$, and $G_n(x, Y) := 0$ otherwise. With this we get

\begin{equation}
\frac{1}{\|f_n\|_{M, \infty}} \left| \mathcal{R}_{L, p}(f + f_n) - \mathcal{R}_{L, p}(f) - \mathcal{R}'_{L, p}(f)f_n \right| \leq \int_{X \times Y} G_n(x, Y) \, d\mathbb{P}(x, Y)
\end{equation}

Also, the definition of $G_n(x, Y)$ and of the directional derivative of $L$ yield

\begin{equation}
\lim_{n \to \infty} G_n(x, Y) = 0,
\end{equation}

for $(x, Y) \in X \times Y$. Moreover, we have by triangle inequality

\begin{equation}
G_n(x, Y) = \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x))}{\|f_n(x)\|_M} \right| - \left( \nabla L(x, Y, f(x)), f_n(x) \right)_M
\end{equation}

\begin{align*}
&\leq \left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x))}{\|f_n(x)\|_M} \right| + \left( \nabla L(x, Y, f(x)), f_n(x) \right)_M.
\end{align*}
3. Mathematical Framework

As a next step, we will calculate upper bounds for the two expressions of Equation (3.82). For that notice that by the use of the mean value theorem for functions of several variables, Theorem A.4, we get that for \((x, Y) \in X \times \mathcal{Y}, \|f_n(x)\|_\mathcal{M} \neq 0\) there exists \(g_n(x, Y)\) such that \(\|g_n(x, Y)\|_\mathcal{M} \in [0, \|f_n(x)\|_\mathcal{M}]\) and

\[
L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x)) = \left( \nabla L(x, Y, f(x) + g_n(x, Y)), f_n(x) \right)_\mathcal{M}.
\]

Applying this and the Cauchy-Schwarz inequality to the first expression of Equation (3.82) yields

\[
\left| \frac{L(x, Y, f(x) + f_n(x)) - L(x, Y, f(x))}{\|f_n(x)\|_\mathcal{M}} \right| = \frac{\left| \left( \nabla L(x, Y, f(x) + g_n(x, Y)), f_n(x) \right)_\mathcal{M} \right|}{\|f_n(x)\|_\mathcal{M}} \leq \left| \nabla L(x, Y, f(x) + g_n(x, Y)) \right|_\mathcal{M} \cdot \|f_n(x)\|_\mathcal{M} = \|\nabla L(x, Y, f(x) + g_n(x, Y))\|_\mathcal{M}.
\]

For the second expression we also use the Cauchy-Schwarz inequality and that for all \(x \in X\) we have \(\|f_n(x)\|_\mathcal{M} \leq \|f_n\|_{\mathcal{M}, \infty} \leq 1\). This gives

\[
\left| \left( \nabla L(x, Y, f(x)), f_n(x) \right)_\mathcal{M} \right| \leq \left| \nabla L(x, Y, f(x)) \right|_\mathcal{M} \cdot \|f_n(x)\|_\mathcal{M} \leq \|\nabla L(x, Y, f(x))\|_\mathcal{M}.
\]

Since we assumed the norm of the gradient, \(\|\nabla L\|_{\mathcal{M}, \infty}\), to be a \(\mathbb{P}\)-integrable Nemitski loss there exist \(b \in L^1(X \times \mathcal{Y}, \mathbb{P})\) and \(h : [0, \infty) \to [0, \infty)\) monotonously increasing such that

\[
\|\nabla L(x, Y, C)\|_\mathcal{M} \leq b(x, Y) + h(\|C\|_\mathcal{M}) \quad \text{for all } (x, Y, C) \in X \times \mathcal{Y} \times \mathcal{M}.
\]

Using this together with \(\|f_n\|_{\mathcal{M}, \infty} \leq 1, n \geq 1\), for both expressions of Equation (3.82), we get

\[
G_n(x, Y) \leq b(x, Y) + h(\|f(x) + g_n(x, Y)\|_\mathcal{M}) + b(x, Y) + h(\|f(x)\|_\mathcal{M}) \leq 2b(x, Y) + h(\|f\|_{\mathcal{M}, \infty} + \|f_n\|_{\mathcal{M}, \infty}) + h(\|f\|_{\mathcal{M}, \infty}) \leq 2b(x, Y) + h(\|f\|_{\mathcal{M}, \infty} + 1) + h(\|f\|_{\mathcal{M}, \infty}),
\]

for all \((x, Y) \in X \times \mathcal{Y}\) and \(n \geq 1\). From Equation (3.80) and Equation (3.81) together with Lebesgue’s theorem, we obtain the desired result.

After we investigated the properties of loss functions and their associated risks, in the next section we will go to studying reproducing kernel Hilbert spaces, their kernels and how these two notions are connected.
3.2. Kernel Approach

3.2.2. Matrix-Valued Reproducing Kernel Hilbert Spaces

In the upcoming part we give some necessary definitions and results about reproducing kernel Hilbert spaces, in short RKHS, for matrix-valued functions. The main difference to the scalar case, which is extensively discussed and described in [34] and various references therein, is that the output space is no longer the underlying field \( \mathbb{R} \) or \( \mathbb{C} \) but the space of real matrices of dimension \( n \times m \), \( \mathbb{R}^{n \times m} \), or a subspace thereof, which we denoted by \( \mathcal{M} \). The case of a general Hilbert space as target space was already discussed in even more generality in [35], where the results are given for locally convex spaces, and regained attention with the work of Micchelli and Pontil [36] and Carmelli, De Vito and Toigo [37], about vector-valued or operator-valued kernels, where we adopt our main results from. Also the work of Aronszajn [38] already gave a very important result for this field, which we will also provide in this more general framework. In his dissertation Schrödl [39] gives a nice summary of the theory of operator-valued reproducing kernel Hilbert spaces as well as possible applications.

Additionally to the notation and nomenclature we introduced already, we will clarify some more objects that are frequently used in the following sections.

Let \( \mathcal{H} \) be a linear space of functions on \( X \) with values in \( \mathcal{M} \). We also assume that \( \mathcal{H} \) is a Hilbert space, with inner product \( \langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R} \). We denote the set of all bounded linear operators from a space \( \mathcal{M} \) to another space \( \mathcal{Z} \) by \( \mathcal{B}(\mathcal{M}, \mathcal{Z}) \). In case we have \( \mathcal{M} = \mathcal{Z} \), we just write \( \mathcal{B}(\mathcal{M}) \). The corresponding norms will be denoted by \( \| \cdot \|_{\mathcal{B}(\mathcal{M}, \mathcal{Z})} \) and \( \| \cdot \|_{\mathcal{B}(\mathcal{M})} \), respectively. The elements of sets of operators will be written as bold capital letters, i.e. \( A, \Delta \). Also, we will use \( \mathcal{B}_{+}(\mathcal{M}) \) to denote the cone of non-negative, bounded linear operators, i.e.

\[
A \in \mathcal{B}_{+}(\mathcal{M}) \quad \text{if and only if} \quad (AY, Y)_{\mathcal{M}} \geq 0, \quad Y \in \mathcal{M}.
\]

We also assume the elements of \( \mathcal{B}_{+}(\mathcal{M}) \) to be symmetric. Finally, we denote the adjoint of an operator \( A \in \mathcal{B}(\mathcal{M}) \) by \( A^{*} \).

We will go with the canonical way to open up the field of reproducing kernel Hilbert spaces. In the scalar as in the vector-valued case this is by defining a RKHS and go from there to the kernel and its properties. To do this we will follow in outlines the work presented in [37]. Notice that this is not the only way to define a RKHS. Another one is also presented in [37], giving a different point of view on the subject. As briefly mentioned before the results in [37], [36], [35] and [39] are for an even more general setting. We will adapt them for our case of matrix-valued functions \( f : X \rightarrow \mathcal{M} \).

**Definition 3.33 (Matrix-valued Reproducing Kernel Hilbert Space (RKHS))**

We say that a Hilbert space \( \mathcal{H} \subseteq \{ f \mid f : X \rightarrow \mathcal{M} \} \) is a matrix-valued reproducing kernel Hilbert space (RKHS) when for any \( x \in X \) the linear functional

\[
\Delta_{x} : \mathcal{H} \rightarrow \mathcal{M}, \quad \Delta_{x}(f) := f(x)
\]

is continuous, that is, for every \( x \in X \) there exists a constant \( c_{x} > 0 \) such that

\[
\| \Delta_{x} f \|_{\mathcal{M}} = \| f(x) \|_{\mathcal{M}} \leq c_{x} \| f \|_{\mathcal{H}}.
\]
To describe the kernel of the reproducing kernel Hilbert space, let $\Delta^*_x : \mathcal{M} \rightarrow \mathfrak{H}$ be the adjoint of $\Delta_x$.

Now, for $x, t \in X$ we define

\[ K(x, t) := \Delta_x \Delta^*_t. \]

(3.91)

In the following theorem we will check if $K(x, t)$ fulfills the properties we expect from a kernel, so in particular the reproducing property and positive definiteness.

**Theorem 3.34**

Let $\mathfrak{H}$ be a matrix-valued reproducing kernel Hilbert space with reproducing kernel $K : X \times X \rightarrow B(\mathcal{M})$.

Then the following properties hold:

(i) We have $K(x, t) \in B(\mathcal{M})$, $K(x, t) = K(t, x)^*$ and $K(x, x) \in B_+(\mathcal{M})$.

(ii) For any $N \in \mathbb{N}$, $\{z_i : 1 \leq i \leq N\} \subset \mathbb{C}$, $\{x_j : 1 \leq j \leq N\} \subset X$, we have

\[ \sum_{i=1}^{N} \sum_{j=1}^{N} z_i z_j (K(x_i, x_j)Y, Y)_{\mathcal{M}} \geq 0, \]

(3.92)

for all $Y \in \mathcal{M}$, i.e. positive definiteness.

(iii) For all $x, t \in X$,

\[ \|K(x, t)\|_{B(\mathcal{M})} \leq \|K(x, x)\|_{B(\mathcal{M})}^{\frac{1}{2}} \|K(t, t)\|_{B(\mathcal{M})}^{\frac{1}{2}} \]

(3.93)

holds.

(iv) We have the reproducing property, i.e.

\[ \Delta^*_x Y = K(\cdot, t)Y, \]

for all $Y \in \mathcal{M}$.

**Proof.** (i) Since $\Delta_x$ is continuous, i.e. bounded, by definition, we know from functional analysis that $\Delta^*_x$ is also bounded. The composition of bounded operators is again bounded, hence $K(x, t) \in B(\mathcal{M})$.

With the properties for adjoint operators we get

\[ K(x, t) = \Delta_x \Delta^*_t \]

(3.95)

\[ = (\Delta_t \Delta^*_x)^* \]

\[ = K(t, x)^*. \]

For the last property we notice,

\[ (\Delta_x \Delta^*_x Y, Y)_{\mathcal{M}} = (\Delta^*_x Y, \Delta^*_x Y)_{\mathfrak{H}} \]

(3.96)

\[ = \|\Delta^*_x Y\|_{\mathfrak{H}}^2 \geq 0, \]

for all $Y \in \mathcal{M}$. 

38
(ii) To prove positive definiteness, let \( N \in \mathbb{N}, \{z_i : 1 \leq i \leq N\} \subset \mathbb{C}, \{x_j : 1 \leq j \leq N\} \subset X \) and \( Y \in \mathcal{M} \). We then have

\[
\sum_{i=1}^{N} \sum_{j=1}^{N} z_i \overline{z_j} \left( K(x_i, x_j)Y, Y \right)_{\mathcal{M}} = \sum_{i=1}^{N} \sum_{j=1}^{N} z_i \overline{z_j} \left( \Delta x_i^* Y, \Delta x_j^* Y \right)_{\mathcal{Y}} \\
= \left| \sum_{j=1}^{N} \Delta x_j^* Y \right|^2_{\mathcal{Y}} \\
= \left| \sum_{j=1}^{N} z_j \Delta x_j^* Y \right|^2_{\mathcal{Y}} \\
\geq 0.
\]

(3.97)

(iii) Let \( x, t \in X \). We have

\[
\left\| K(x, t) \right\|^2_{B(\mathcal{M})} \leq \left\| \Delta x \Delta t^* \right\|^2_{B(\mathcal{M})} \\
\leq \left\| \Delta x \right\|^2_{B(\mathcal{Y}, \mathcal{M})} \cdot \left\| \Delta t^* \right\|^2_{B(\mathcal{M}, \mathcal{Y})} \\
= \left\| \Delta x \Delta t^* \right\|^2_{B(\mathcal{M})} \cdot \left\| \Delta x \Delta t^* \right\|_{B(\mathcal{M})} \\
= \left\| K(x, x) \right\|_{B(\mathcal{M})} \left\| K(t, t) \right\|_{B(\mathcal{M})} \\
\]

where the penultimate equality is due to the norm equality of adjoint operators and their compositions (see e.g. [40]).

(iv) To check the reproducing property we first define \( \delta_{(x, Y)} : \mathcal{Y} \to \mathbb{R} \) via

\[
\delta_{(x, Y)}(f) := \left( f(x), Y \right)_{\mathcal{M}} = \left( \Delta x f, Y \right)_{\mathcal{M}}.
\]

By the Riesz-Fréchet representation theorem (see, e.g. [41]) we get that there exists \( H(\cdot, x, Y) \) such that

\[
\delta_{(x, Y)}(f) = \left< f, H(\cdot, x, Y) \right>_{\mathcal{Y}}.
\]

(3.100)

We also get that \( H \) is linear in its third variable, i.e. \( H(\cdot, x, Y) = H(\cdot, x)Y \). Therefore we have

\[
\left( \Delta x f, Y \right)_{\mathcal{M}} = \left< f, H(\cdot, x)Y \right>_{\mathcal{Y}},
\]

for all \( f \in \mathcal{Y} \) and \( Y \in \mathcal{M} \). From this we get

\[
H(\cdot, x)Y = \Delta x^* Y.
\]

(3.101)

(3.102)

It remains to show that \( H \) is equal to the kernel \( K \). For that, we notice

\[
H(t, x)Y = \Delta t (H(t, x)Y) \\
= \Delta t \Delta x^* Y \\
= K(t, x)Y
\]

(3.103)
3. Mathematical Framework

holds for all \( x, t \in X \) and \( Y \in \mathcal{Y} \). Hence,

\[
H(t, x) = K(t, x)
\]

and

\[
\Delta^*_x Y = K(\cdot, x) Y.
\]

From this theorem and its proof we can draw some additional immediate conclusions:

1. If \( \mathcal{H} \) is a RKHS then the corresponding kernel \( K \) is uniquely determined, which can be seen from part (iv) of the proof.

2. As mentioned in the remark after the proof of Lemma 3.20 the norm \( \| \cdot \|_{\mathcal{H}} \) of the RKHS \( \mathcal{H} \) induces a metric that dominates pointwise convergence. This can be seen by Definition 3.33 and the fact that

\[
\| \Delta^*_x \|_{B(\mathcal{H}, \mathcal{M})} = \| \Delta^*_x \|_{B(\mathcal{M}, \mathcal{H})} = \| K(x, x) \|_{B(\mathcal{M})}^{\frac{1}{2}},
\]

for all \( x \in X \) and \( f \in \mathcal{H} \).

   This means, if we have a sequence \( (f_n)_{n \in \mathbb{N}} \) that converges to \( f \) in \( \mathcal{H} \) then \( (f_n(x))_{n \in \mathbb{N}} \) converges to \( f(x) \) in \( \mathcal{M} \), i.e. pointwise convergence.

Now, we come to the reverse statement in terms of uniqueness. Similar to the theorem from Aronszajn in the scalar-valued case, also in the operator-valued case it can be shown that a positive definite, operator-valued kernel \( K: X \times X \to B(\mathcal{M}) \) generates a (up to an isometry) unique reproducing kernel Hilbert space, which has \( K \) as reproducing kernel. As pointed out, we can therefore also start this theoretical framework with a kernel, instead of the definition of a reproducing kernel Hilbert space. We state the result here for sake of completeness, the proof can be found in [37] and follows the construction known as Gelfand-Naimark-Segal construction.

**Theorem 3.35**

Let \( K: X \times X \to B(\mathcal{M}) \) be a positive matrix-valued kernel.

Then there exists a (up to an isometry) unique reproducing kernel Hilbert space \( \mathcal{H} \) of \( \mathcal{M} \)-valued functions, which has \( K \) as its reproducing kernel.

Next, we want to give two results we will need to provide the adaption of the well known representer theorem presented in Section 3.2.3. The first of these results is concerned with the connection between measurable kernels and measurable functions in the corresponding reproducing kernel Hilbert space. It was presented in [37], as Proposition 6, where the proof can be found as well. The authors of [37] emphasize the importance of the target space of studied functions, in our case \( \mathcal{M} \), being separable. This is due to the fact that if it was not separable we would deal with two different versions of measurability, weak and strong measurability. Since we are dealing with \( \mathcal{M} \subset \mathbb{R}^{n \times m} \), which is separable, these two versions coincide and we do not need to make this distinction. The same issue occurs if the reproducing kernel Hilbert space \( \mathcal{H} \) is not separable, so we will assume this as well. We adapt the result to fit our purpose of matrix-valued functions.
Theorem 3.36
Let $X$ be a measurable space and $\mathcal{H}$ a separable reproducing kernel Hilbert space with reproducing kernel $K$.

Then the following statements are equivalent:

(i) the elements of $\mathcal{H}$ are measurable functions $f : X \to \mathcal{M}$,

(ii) the map $K : X \times X \to B(\mathcal{M})$ is measurable,

(iii) for all $x \in X$, the map $t \mapsto K(t, x) \in B(\mathcal{M})$ is measurable.

The second result we will need presents insight in regards to boundedness of the kernel and the elements of $\mathcal{H}$. It is again taken from [37], see Proposition 8. Before we get to it we need to give another definition to introduce a special notion of boundedness.

Definition 3.37 (p-bounded kernel)
Let $1 \leq p \leq \infty$. A measurable kernel $K : X \times X \to B(\mathcal{M})$ is said to be $p$-bounded if the following hold:

(i) for almost all $x \in X$

\[
\int_X \| K(x, t)^* Y \|_{\mathcal{M}}^q \, dt < \infty \quad \text{for all } Y \in \mathcal{M},
\]

(ii) for all $g \in L^p(X, \mathcal{M})$, the map

\[
x \mapsto w \cdot \int_X K(x, t) g(t) \, dt
\]

is in $L^q(X, \mathcal{M})$, where $w \cdot \int_X f(t) \, dt$ denotes the (weak) Pettis integral (see [42]).

Now, this enables us to continue with the announced result.

Theorem 3.38
Let $\mathcal{H}$ be a separable reproducing kernel Hilbert space of measurable functions. Fix $1 \leq p \leq \infty$, then the following are equivalent:

(i) the elements of $\mathcal{H}$ belong to $L^p(X, \mathcal{M})$,

(ii) the reproducing kernel $K$ of $\mathcal{H}$ is $q$-bounded with $q = \frac{p}{p-1}$.

If one of the above conditions hold, we also have that the embedding

\[
i : \mathcal{H} \to L^p(X, \mathcal{M})
\]

is continuous.

We pointed out in Section 2.2 and Section 2.5 that for our application we need to estimate a matrix-valued function, so a function mapping to the space of real matrices $\mathbb{R}^{n \times m}$ or a subspace thereof, which we denoted by $\mathcal{M}$. To be able to do that we need the appropriate kernel, which by Theorem 3.35 will give us a corresponding function space. So far we studied the general properties of these kernels. Now, we want to go a little closer to our application. From our
results so far we can see that in the case of matrix-valued functions, given \( x, t \in X \) we have that \( K(x, t) \in \mathcal{B}(\mathcal{M}) \), so a bounded linear mapping from \( \mathcal{M} \) to \( \mathcal{M} \). This mapping can be written in the form of a matrix of dimension \( m \cdot n \times m \cdot n \), i.e.

\[
K(x, t) = \begin{bmatrix}
K_{1,1}(x, t) & \cdots & K_{1,mn}(x, t) \\
\vdots & \ddots & \vdots \\
K_{mn,1}(x, t) & \cdots & K_{mn,mn}(x, t)
\end{bmatrix}, \quad x, t \in X
\]  

(3.110)

where \( K_{ij} : X \times X \to \mathbb{R}, 1 \leq i, j \leq m \cdot n \) are scalar-valued kernels. This way is chosen in [39]. Since scalar-valued kernels are easier to access than matrix-valued kernels as a whole, we can use this representation to construct matrix-valued kernels.

Another way to write the map \( K(x, t) \) is via a tensor representation, which is a little more convenient when dealing with matrix-valued functions. Choosing a basis \( \{ e_i \}_{i=1}^n \) for \( \mathbb{R}^n \) and \( \{ e'_j \}_{j=1}^m \) for \( \mathbb{R}^m \) we can form the basis of the tensor product space \( \mathbb{R}^n \otimes \mathbb{R}^m \) via \( \{ e_i \otimes e'_j \}_{i,j=1}^{n,m} \). If we choose the standard basis for \( \mathbb{R}^n \) and \( \mathbb{R}^m \), respectively, then the basis of the tensor space equals the matrices \( \{ E_{ij} \}_{i,j=1}^{n,m} \) we defined earlier on, so the basis of the space of real matrices. When using the tensor notation we see that the bounded linear map \( K(x, t) \) is a tensor of 4th order, i.e. for an element \( B \in \mathcal{M} \) we have

\[
(K(x, t)B)_{i_1,j_1} = \sum_{i_2=1}^n \sum_{j_2=1}^m K_{i_1,j_1,i_2,j_2}(x, t)b_{i_2,j_2} \quad x, t \in X,
\]

(3.111)

for \( i_1 = 1, \ldots, n \) and \( j_1 = 1, \ldots, m \).

So, in the remainder of this section we want to have a brief look at how the relationship among the scalar-valued kernels \( K_{i_1,j_1,i_2,j_2}(x, t) \) needs to be, to end up with a positive definite matrix-valued kernel. Chapter 2 of [39] gives a nice summary of construction methods for vector-valued functions which we will adapt to our case of matrix-valued functions. Since both are finite dimensional spaces the main difference is notation and the ordering of the basis of the objects we are dealing with, so the results given in [39] still hold in our case.

We will start by giving a result that proposes a very elementary way to construct a positive definite kernel for a corresponding RKHS that is finite dimensional. The result is taken from [39] and adapted for the case of functions mapping to the space of real matrices, as we need it for our application.

**Theorem 3.39**

Let \( f_d : X \to \mathbb{R}^n \otimes \mathbb{R}^m \) be functions mapping to the Hilbert space of real matrices and let \( a_d \in \mathbb{R}^+ \) be positive constants for \( 1 \leq d \leq D < \infty \), then

\[
K(x, t) = \sum_{d=1}^M a_d(f_d(t)^t)_{\mathcal{M}^{\prime}} f_d(x)
\]

(3.112)

is a positive definite matrix-valued kernel of the reproducing kernel Hilbert space \( \mathcal{S} = \text{span}\{ f_d : 1 \leq d \leq D \} \).
3.2. Kernel Approach

The proof in matrix notation can be found in [39]. We will just have a look at how we can write this kernel as a 4th order tensor. So, we need to compute \( K(x, t)e_{i_1} \otimes e'_{j_2} \):

\[
K(x, t)e_{i_1} \otimes e'_{j_2} = \sum_{d=1}^{D} a_d(e_{i_1} \otimes e'_{j_2}, f_d(t)) f_d(x)
\]

\[
= \sum_{d=1}^{D} a_d(e_{i_1} \otimes e'_{j_2}, \sum_{i_2=1}^{n} \sum_{j_2=1}^{m} f_{d}^{i_2,j_2}(t)e_{i_2} \otimes e'_{j_2}) \mathcal{M} \sum_{i_2=1}^{n} \sum_{j_2=1}^{m} f_{d}^{i_2,j_2}(x)e_{i_2} \otimes e'_{j_2}
\]

\[
= \sum_{i_2=1}^{n} \sum_{j_2=1}^{m} K_{i_1,j_1,i_2,j_2}(x, t)e_{i_2} \otimes e'_{j_2},
\]

where \( K_{i_1,j_1,i_2,j_2}(x, t) := \sum_{d=1}^{D} a_d f_{d}^{i_2,j_2}(t)(e_{i_1} \otimes e'_{j_2}, e_{i_2} \otimes e'_{j_2}) \mathcal{M} f_{d}^{i_2,j_2}(x) \) and \( f_{d}^{i_2,j_2}(x), f_{d}^{i_2,j_2}(t) \) denote the entries of the matrices \( f_d(x), f_d(t) \) respectively.

For the next construction we again make use of arbitrary functions to generate our kernel. This time, we additionally need a scalar-valued kernel to get the matrix-valued kernel, since the positive definiteness of the scalar-valued kernel ensures that we have positive definiteness of the matrix-valued kernel as well.

**Theorem 3.40**

Let \( k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) be a positive definite scalar-valued kernel and \( f_{i,j} : X \rightarrow \mathbb{R}^d \) vector-valued functions, \( 1 \leq i \leq n, 1 \leq j \leq m \), then the matrix-valued function \( K : X \times X \rightarrow \mathcal{B}(\mathcal{M}) \) defined via

\[
K_{i_1,j_1,i_2,j_2}(x, t) = k(f_{i_1,j_1}(x), f_{i_2,j_2}(t)), \quad x, t \in X, \quad 1 \leq i_1, i_2 \leq n, \quad 1 \leq j_1, j_2 \leq m,
\]

is a positive definite matrix-valued kernel.

Another handy result for constructing kernels is the fact that the element-wise product of two matrix-valued kernels, is again a matrix-valued kernel, which is covered in the next theorem. This statement is based on the results by Aronszajn and Schur that the elementwise product of positive semidefinite matrices is positive definite again, see [38] and [43].

**Theorem 3.41**

Let \( K', K'' : X \times X \rightarrow \mathcal{B}(\mathcal{M}) \) be positive definite matrix-valued kernels.

Then the function \( K : X \times X \rightarrow \mathcal{B}(\mathcal{M}) \) defined via the element wise product, i.e.,

\[
K_{i_1,j_1,i_2,j_2}(x, t) = K_{i_1,j_1,i_2,j_2}'(x, t) \cdot K_{i_1,j_1,i_2,j_2}''(x, t), \quad x, t \in X, \quad 1 \leq i_1, i_2 \leq n, \quad 1 \leq j_1, j_2 \leq m,
\]

is a positive definite matrix-valued kernel as well.
3. Mathematical Framework

The proof is again provided in [39] in matrix notation.

The last result we want to mention here concerning the construction of matrix-valued kernels is the analogue to the result in the scalar-valued case that convex combinations of positive definite kernels give again a positive definite kernel. In the matrix-valued case this still holds true if we use positive definite operators and scalar-valued kernels.

**Theorem 3.42**

Let \( k_d : X \times X \to \mathbb{R} \), \( 1 \leq d \leq D \), \( D \in \mathbb{N} \) be positive definite scalar-valued kernels. Moreover, let \( A_d \in \mathcal{B}_+ (\mathcal{M}) \), \( 1 \leq d \leq D \).

Then the function \( K : X \times X \to \mathcal{B}(\mathcal{M}) \) given by

\[
K(x, t) = \sum_{d=1}^{D} A_d k_d(x, t)
\]

is a positive definite matrix-valued kernel.

The idea of the proof is already given in [39] but we will elaborate a little on it.

**Proof.** The symmetry property is obvious by the definition of \( \mathcal{B} (\mathcal{M}) \) and scalar-valued kernels.

For positive definiteness we first notice that we can look at each term of the sum individually. Also, for a symmetric, positive definite operator \( A \in \mathcal{B}_+ (\mathcal{M}) \) and a scalar-valued kernel \( k : X \times X \to \mathbb{R} \) we have that

\[
(A k(x, t) Y, Z)_\mathcal{M} = (A Y, Z)_\mathcal{M} k(x, t)
\]

By the results of Schur [43], we get that \( A k(x, t) \) is positive definite, since \( (A \cdot, \cdot)_\mathcal{M} \) and \( k(\cdot, \cdot) \) are both positive definite. Now, given arbitrary \( N \in \mathbb{N} \), \( \{x_i : 1 \leq i \leq N\} \subset X \) and \( \{Y_j : 1 \leq j \leq N\} \subset \mathcal{M} \), we have that

\[
\sum_{i=1}^{N} \sum_{j=1}^{N} (A k(x_j, x_i) Y_i, Y_j)_\mathcal{M} = \sum_{i=1}^{N} \sum_{j=1}^{N} (A Y_i, Y_j)_\mathcal{M} k(x_j, x_i)
\]

holds true, which gives the desired result.

Note that the elements of \( \mathcal{B}(\mathcal{M}) \) can all be represented by a 4th order tensor. So here the entries \( K_{i_1,j_1,i_2,j_2} (x, t) \) are given by \( \sum_{d=1}^{D} A_d^{i_1,j_1} A_d^{i_2,j_2} k_d(x, t) \).

For our application this result will prove particularly useful since it allows us to control the structure of our kernel via the positive definite matrices in Equation (3.116) and therefore get mappings for the necessary subspace \( \mathcal{M} \subseteq \mathbb{R}^{n \times m} \).

As already mentioned, in the literature there are many more ways to construct matrix- or operator-valued kernels. In [36] for example a way via so called entire functions is described.

In the next section, we want to give the results which let us benefit from the efforts we put in so far. We will see how reproducing kernel Hilbert spaces can be used to solve risk minimization problems for given function classes and regularized versions thereof also in the matrix-valued case. For the scalar-valued case we refer to Chapter 5 of [34].
3.2.3. Risk Minimization and Regularization

In this part of the thesis we will study how the two objects, the risk functional $R_{L,P}$ and the reproducing kernel Hilbert space $\mathcal{H}$, we introduced in the last two sections can be connected. So we want to give results about how we can use reproducing kernel Hilbert spaces to obtain functions fitting our data $D = \{(x_i, Y_i) : i = 1, \ldots, M\} \subset X \times Y$ in a good way, so with minimal risk.

In Definition 3.19 we specified the so called Bayes risk $R_{L,P}$, so the best risk value we can achieve for a given loss function $L : X \times Y \times M \to [0, \infty)$ and a probability distribution $P$ on $X \times Y$. Theoretically this is a nice and clean setup but it has some setbacks if we want to use it in a practical setting.

A first issue is the probability distribution $P$ on $X \times Y$, which is usually not available. The only thing we have to get an insight in this distribution is the data set $D$. So we will replace the general $L$-risk $R_{L,P}$ by the empirical $L$-risk $R_{L,D}$, which is subject to our given data set $D$ and was given in Equation (3.33).

The next issue is that the Bayes risk is defined as the infimum over all measurable functions, which is a too large space for our purpose. Even if we use the empirical $L$-risk we will run into problems due to overfitting of the data. We therefore need to find a good replacement for the space of matrix-valued measurable functions $L^0(X,M)$, which is neither too big, nor too small. We also want to prevent underfitting of course.

A still quite general but good choice are the matrix-valued reproducing kernel Hilbert spaces $\mathcal{H}$ we just introduced in Section 3.2.2.

Finally, we might not want to fit our function exactly to the data, because of noise or outliers, and because we do not have an interpolation problem at hand. To account for this we consider a regularized version of the risk functional, so for $\alpha > 0$, define

$$R_{L,P,\alpha}(f) := \int_{X \times Y} L(x, Y, f(x)) \, dP(x, Y) + \alpha \|f\|_{\mathcal{H}}^2$$

$$= R_{L,P}(f) + \alpha \|f\|_{\mathcal{H}}^2,$$

where $\| \cdot \|_{\mathcal{H}}$ denotes the norm in the reproducing kernel Hilbert space $\mathcal{H}$.

Analogously, we consider the empirical version of the regularized risk, i.e.

$$R_{L,D,\alpha}(f) := \frac{1}{M} \sum_{i=1}^{M} L(x_i, Y_i, f(x_i)) + \alpha \|f\|_{\mathcal{H}}^2$$

$$= R_{L,D}(f) + \alpha \|f\|_{\mathcal{H}}^2.$$

Together with these two regularizations come adaptions of the minimization problems leading to the respective Bayes decision functions $f_{L,P,\alpha}^*$ and $f_{L,D,\alpha}^*$ (see Definition 3.19), so we have

$$R_{L,P,\alpha}(f_{L,P,\alpha}^*) := \inf \{R_{L,P,\alpha}(f) : f \in \mathcal{H}\}$$

and

$$R_{L,D,\alpha}(f_{L,D,\alpha}^*) := \inf \{R_{L,D,\alpha}(f) : f \in \mathcal{H}\}.$$
3. Mathematical Framework

respectively.

Now that we specified what we are looking for, we will start off by answering questions about existence, uniqueness and representation of the Bayes decision function $f_{L,D,\alpha}^{\star}$.

In terms of uniqueness we can easily reproduce the result from the scalar-valued case given by Theorem 5.1 in [34].

**Theorem 3.43**

Let $L: \mathcal{X} \times \mathcal{Y} \times \mathcal{M} \to [0, \infty)$ be a convex loss function, $\mathcal{H}$ a separable RKHS of a measurable kernel and $\mathbb{P}$ a probability distribution on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{R}_{L,P}(f) < \infty$ for some $f \in \mathcal{H}$.

Then for all $\alpha > 0$ the Bayes decision function $f_{L,D,\alpha}^{\star}$ is unique.

**Proof.** Assume we have two minimizers $f_1, f_2 \in \mathcal{H}$ with $f_1 \neq f_2$ of $\mathcal{R}_{L,P,\alpha}$. By the polarization identity we get

\begin{equation}
\|f_1 + f_2\|_{\mathcal{H}}^2 + \|f_1 - f_2\|_{\mathcal{H}}^2 = 2\|f_1\|_{\mathcal{H}}^2 + 2\|f_2\|_{\mathcal{H}}^2
\end{equation}

which is equivalent to

\begin{equation}
\left\| \frac{1}{2}(f_1 + f_2) \right\|_{\mathcal{H}}^2 + \left\| \frac{1}{2}(f_1 - f_2) \right\|_{\mathcal{H}}^2 > 0
\end{equation}

which gives

\begin{equation}
\left\| \frac{1}{2}(f_1 + f_2) \right\|_{\mathcal{H}}^2 < \frac{1}{2}\|f_1\|_{\mathcal{H}}^2 + \frac{1}{2}\|f_2\|_{\mathcal{H}}^2.
\end{equation}

We now use the convexity of $L$ and consequently by Lemma 3.22 of $\mathcal{R}_{L,P}$ to get

\begin{equation}
\mathcal{R}_{L,P,\alpha} \left( \frac{1}{2}(f_1 + f_2) \right) = \mathcal{R}_{L,P} \left( \frac{1}{2}(f_1 + f_2) \right) + \alpha \left\| \frac{1}{2}(f_1 + f_2) \right\|_{\mathcal{H}}^2
\end{equation}

\begin{equation}
\leq \frac{1}{2} \mathcal{R}_{L,P}(f_1) + \frac{1}{2} \mathcal{R}_{L,P}(f_2) + \alpha \left\| \frac{1}{2}(f_1 + f_2) \right\|_{\mathcal{H}}^2
\end{equation}

\begin{equation}
< \frac{1}{2} \mathcal{R}_{L,P}(f_1) + \frac{1}{2} \mathcal{R}_{L,P}(f_2) + \alpha \left( \frac{1}{2}\|f_1\|_{\mathcal{H}}^2 + \frac{1}{2}\|f_2\|_{\mathcal{H}}^2 \right)
\end{equation}

\begin{equation}
= \frac{1}{2} \mathcal{R}_{L,P,\alpha}(f_1) + \frac{1}{2} \mathcal{R}_{L,P,\alpha}(f_2)
\end{equation}

\begin{equation}
= \mathcal{R}_{L,P,\alpha}(f_1) = \mathcal{R}_{L,P,\alpha}(f_2),
\end{equation}

which contradicts the assumption on $f_1$ or $f_2$, respectively, being the minimizer. \hfill \Box

Concerning existence we have the following result, which is an adaption of the scalar-valued case presented in Theorem 5.2 of [34].

**Theorem 3.44**

Let $L: \mathcal{X} \times \mathcal{Y} \times \mathcal{M} \to [0, \infty)$ be a convex loss function and $\mathbb{P}$ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ such that $L$ is a $\mathbb{P}$-integrable Nemitski loss. Moreover, let $K$ be a bounded, measurable kernel on $\mathcal{X} \times \mathcal{X}$ with corresponding separable RKHS $\mathcal{H}$.

Then, for all $\alpha > 0$ there exists at least one minimizer $f_{L,P,\alpha}^{\star}$.
Proof. Since our kernel $K$ is measurable and $\mathcal{H}$ is a separable RKHS we get by Theorem 3.36 that the elements $f \in \mathcal{H}$ are also measurable. By the boundedness of the kernel, Theorem 3.38 gives that the elements $f \in \mathcal{H}$ are also bounded and the embedding $\iota: \mathcal{H} \to L^\infty(X,\mathcal{M})$ is continuous. Due to the convexity and Lemma A.5 we get from the fact that $L(x,Y,C) < \infty$ for all $(x,Y,C) \in X \times Y \times \mathcal{M}$ that $L$ is continuous. Together with Lemma 3.26 we get that $\mathcal{R}_{L,P}: L^\infty(X,\mathcal{M}) \to [0,\infty)$ is continuous, and because of the continuity of the embedding $\iota: \mathcal{H} \to L^\infty(X,\mathcal{M})$ so is $\mathcal{R}_{L,P}: \mathcal{H} \to [0,\infty)$. By Lemma 3.22 we get that $\mathcal{R}_{L,P}$ is convex. The mapping $f \mapsto \alpha \|f\|^2_{\mathcal{H}}$ is also convex and continuous which leads to $\mathcal{R}_{L,P,\alpha}$ being also continuous and convex.

Consider now the set
\begin{equation}
\mathfrak{A} := \{ f \in \mathcal{H} : \alpha \|f\|^2_{\mathcal{H}} + \mathcal{R}_{L,P}(f) \leq c \},
\end{equation}
where $c := \mathcal{R}_{L,P}(0)$. Since $0 \in \mathfrak{A}$ we have $\mathfrak{A} \neq \emptyset$. Also if $f \in \mathfrak{A}$ we know $\alpha \|f\|^2_{\mathcal{H}} \leq c$ and therefore $\mathfrak{A} \subset (\frac{c}{\alpha})^\frac{1}{2} B_1(0)$, where $B_1(0) := \{ f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1 \}$, the unit ball in $\mathcal{H}$. This makes $\mathfrak{A}$ a non-empty bounded subset and Theorem A.6 gives the existence of a minimizer.

The main result concerning solutions $f^*_L,D,\alpha$ to our risk minimization problem is provided by the representer theorem, which we will also adapt to our setting from Theorem 5.5 in [34].

Theorem 3.45
Let $L: X \times Y \times \mathcal{M} \to [0,\infty)$ be a convex loss function and $\mathcal{D} := ((x_1,Y_1), \ldots, (x_M,Y_M)) \in (X \times Y)^M$. Furthermore, let $\mathcal{H}$ be the corresponding RKHS to the measurable and bounded kernel $K: X \times X \to B(\mathcal{M})$.

Then for all $\alpha > 0$ there exists a unique minimizer $f^*_L,D,\alpha$ of $\mathcal{R}_{L,D,\alpha}$. Also, there exist $C_1, \ldots, C_M \in \mathcal{M}$ such that
\begin{equation}
f^*_L,D,\alpha = \sum_{i=1}^M K(\cdot, x_i) C_i.
\end{equation}

Before we go to the proof of the theorem we want to give a short insight about how this representer theorem fits together with the one we know from the scalar-valued case. In the scalar-valued setting we have a kernel $k: X \times X \to \mathbb{R}$ which is itself an element of the corresponding reproducing kernel Hilbert space $\mathcal{H}$ if we fix one variable, i.e. $k_x = k(\cdot, x) \in \mathcal{H}$. Our representer in this setup has the following form
\begin{equation}
f^*_L,D,\alpha = \sum_{i=1}^M c_i k(\cdot, x_i) = \sum_{i=1}^M c_i k_x \quad c_1, \ldots, c_M \in \mathbb{R}.
\end{equation}

In the generalized representer theorem the coefficients are no longer elements of the underlying field but rather elements of the space we want our function to map to, so in our case the space of matrices $\mathcal{M}$. The kernel is now an operator mapping these elements to an element in the reproducing kernel Hilbert space $\mathcal{H}$, so an element of $B(\mathcal{M}, \mathcal{H})$.

We can get the same view on the representer theorem in the scalar case if we consider $k_x$ as a bounded linear operator $k_x: \mathbb{R} \to \mathcal{H}, c \mapsto ck_x$.

Now we come to the proof of Theorem 3.45.
Proof. There are three different properties to show now: uniqueness of the solution, the existence and the representation specified in Equation (3.128).

Uniqueness: Can be shown like in the proof of Theorem 3.43.

Existence: By the convexity of $L$ and Lemma A.5 we get the continuity of the loss function $L$. Since we have a bounded kernel we can make use of property (iii) of Theorem 3.34 to get pointwise convergence out of the convergence in norm of the RKHS, i.e.

$$\lim_{n \to \infty} \mathcal{R}_{L,D}(f_n) = \frac{1}{M} \sum_{i=1}^{M} \lim_{n \to \infty} L(x_i, f_n(x_i))$$

holds.

From here we can show the existence as in the proof of Theorem 3.44.

Representation: We first set $X := \{x_i: 1 \leq i \leq M\}$ and define

$$\mathcal{H}_{X'} := \text{span}\left\{ \sum_{i=1}^{M} K(\cdot, x_i)C_i: C_i \in \mathcal{M} \right\}.$$  

Due to Theorem 3.35 and the construction of the RKHS corresponding to a specified kernel in the proof we know that $\mathcal{H}_{X'}$ is the corresponding RKHS to the restricted kernel $K_{|X' \times X'}: X' \times X' \to B(\mathcal{M})$. We also know from our results about existence and uniqueness from before that there is a minimizer $f_{L,D,X',\mathcal{H}_{X'}} \in \mathcal{H}_{X'}$ to the risk $\mathcal{R}_{L,D,X',\mathcal{H}_{X'}} \to [0, \infty)$, so in the corresponding RKHS $\mathcal{H}_{X'}$ to the restricted kernel $K_{|X' \times X'}$.

Now, let $\mathcal{H}_{X'}^\perp$ denote the orthogonal complement of $\mathcal{H}_{X'}$ in $\mathcal{H}$. For $f \in \mathcal{H}_{X'}^\perp$, we have

$$\langle (f(x_i), Y)_{\mathcal{M}} = \langle f, K_{X',Y}\rangle_{\mathcal{H}} = 0 \text{ for all } i = 1, \ldots, M \ Y \in \mathcal{M},$$

which implies

$$f(x_i) = 0 \text{ for all } i = 1, \ldots, M.$$  

Let $P_{X'}: \mathcal{H} \to \mathcal{H}$ be the orthogonal projection onto $\mathcal{H}_{X'}$. We now make use of the decomposition $f = P_{X'} f + (I_d - P_{X'}) f$ and get

$$\mathcal{R}_{L,D}(f) = \mathcal{R}_{L,D}(P_{X'} f + (I_d - P_{X'}) f)$$

$$= \frac{1}{M} \sum_{i=1}^{M} L(x_i, Y, f_{X'} f(x_i) + (I_d - P_{X'}) f(x_i))$$

$$= \mathcal{R}_{L,D}(P_{X'} f),$$

for all $x \in X$.
so

\[(3.136)\]
\[\mathcal{R}_{L,D}(f) = \mathcal{R}_{L,D}(P_X f).\]

From functional analysis we know that \(\|P_X f\|_\beta \leq \|f\|_\beta\). Using these two facts we get

\[\begin{align*}
\mathcal{R}^*_\alpha &= \inf_{f \in \mathcal{H}} \{\alpha \|f\|_\beta^2 + \mathcal{R}_{L,D}(f)\} \\
&\leq \inf_{f \in \mathcal{H}} \{\alpha \|f\|_\beta^2 + \mathcal{R}_{L,D}(f)\} \\
&= \inf_{f \in \mathcal{H}} \{\alpha \|P_X f\|_\beta^2 + \mathcal{R}_{L,D}(P_X f)\} \\
&\leq \inf_{f \in \mathcal{H}} \{\alpha \|f\|_\beta^2 + \mathcal{R}_{L,D}(f)\} \\
&= \mathcal{R}^*_\alpha,
\end{align*}\]

and therefore actually have equality in this chain of inequalities. This means that \(f^*_\alpha\) minimizes \(\alpha \|f\|_\beta^2 + \mathcal{R}_{L,D}(f)\) in \(\mathcal{H}\) and by the uniqueness of the minimizer we have that \(f^*_\alpha = f^*_L\) and the representation follows from the definition of \(\mathcal{H}_X\).

\[\square\]

This result about the representation of our solution will allow us to make an ansatz for the function we want to estimate in the respective risk functional and thereby compute the coefficients \(C_i \in \mathcal{M}\). This will be described in more detail and adapted to our specific problem in Section 4.2.

In [36] and also in [39] another way of deriving the representer theorem is proposed. They first introduce minimal norm interpolation problems and provide conditions that are necessary to solve them. After relaxing the interpolation problem to a regression problem as a special version of a Tikhonov regularization, they go to general loss functions. The proof of the representer theorem in this setting is also brought back to the case of an interpolation problem. In [36] they also give a result concerning differentiable loss functions and the representation of the minimizer which does not need convexity any more.

Now that we presented the necessary framework for our parameter estimation, in the next part of the thesis we will further specify our setting and use the tools we derived in this part to construct the appropriate kernel for our endeavor and formulate the optimization problem to find the solution to our risk functional.
4. Parameter Estimation

In this section we want to introduce the approach of how we seek to estimate our parameters for the aforementioned Markov model. Especially, the question of how we can determine the generator matrix $Q(t)$ from the available data is of particular interest.

Various approaches can be used to do that. Most of them can be looked at in the view of our framework of risk minimization. We will apply an extended form of the well known kernel approach. The main difference of our situation to the scenarios in which this method is commonly used is that we need to approximate a matrix-valued function and therefore need the expanded framework of matrix-valued reproducing kernel Hilbert spaces introduced in Section 3.2.2.

We will briefly recap the setup we have:

We have a particular parking segment $E$ with at maximum $N$ free parking spaces. So, the stochastic process $X(t)$ describing the number of occupied spaces has the finite state space $S = \{0, \ldots, N\}$. The behaviour of the cars arriving and leaving the parking segment is described by a Markov model with generator matrix $Q(t)$, which can be either constant or time-dependent.

The generator matrix is of the form of Equation (2.4), i.e.

$$Q(t) = \begin{bmatrix}
    -\lambda(t) & \lambda(t) & & & \\
    \mu(t) & -(\lambda + \mu)(t) & \lambda(t) & & \\
    2\mu(t) & -(\lambda + 2\mu)(t) & \lambda(t) & \cdots & \\
    & \ddots & \ddots & \ddots & \ddots & \\
    & & \ddots & \ddots & \ddots & \\
    i\mu(t) & -(\lambda + i\mu)(t) & \lambda(t) & \cdots & \\
    & \ddots & \ddots & \ddots & \ddots & \\
    & & \ddots & \ddots & \ddots & \\
    & & & \ddots & \ddots & \ddots & \\
    N\mu(t) & -N\mu(t) & & & & \\
\end{bmatrix}$$

with $\lambda, \mu: [0, T] \to \mathbb{R}_{\geq 0}$ positive, real valued functions, representing the time-dependent transition rates during the considered time interval. As this is the function we want to estimate we can fix the target space $M$ of our functions as $T(N + 1) := \{A \in \mathbb{R}^{N+1 \times N+1}: a_{ij} = 0 \text{ for } i-j > 1, i, j = 1, \ldots N + 1\}$ the space of tridiagonal matrices of dimension $N + 1$. Since we gave our results in a quite general setting so far, we also choose a specific inner product for $(\cdot, \cdot)_M$ now. We will use the Frobenius scalar product, i.e.

$$\langle A, B \rangle_F := \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} a_{ij} b_{ij}.$$ 

The transition probabilities of the Markov process $\{X(t)\}_{t \in \mathbb{R}_{\geq 0}}$ are given by $P(\Delta t)_{ij} = \mathbb{P}[X(t + \Delta t) = j | X(t) = i]$, $i, j = 0, 1, 2, \ldots, N$. These together with the generator matrix $Q(t)$ lead to the Kolmogorov forward equation

$$\dot{P}(t) = P(t) \cdot Q(t), \quad P(t_0) = I.$$
4. Parameter Estimation

For the estimation of the parameters, i.e. the generator matrix \( Q(t) \), we need data points from this matrix-valued function. To obtain these, we assume we are given a data set

\[
\mathcal{D} = \left\{ (t_0, P(t_0)), (t_1, P(t_1)), (t_2, P(t_2)), \ldots, (t_M, P(t_M)) \right\} \subset [0, T] \times \mathbb{R}^{N+1 \times N+1},
\]

containing pairs of a timestamp and the corresponding transition probability matrix of the Markov model. How we can gather this data set was discussed in Section 2.4. Since the way to get from this data set to the information we need about the generator matrix \( Q(t) \) to apply techniques of risk minimization is different for the constant and time-dependent case, we will now distinguish two different situations:

**Constant parameters**, i.e. we assume our transition rates to be constant, so for the generator matrix we have \( Q(t) = Q \in \mathbb{R}^{N+1 \times N+1} \) for all \( t \in [0, T] \).

**Time-dependent parameters**, i.e. the transition rates are time-dependent and therefore our generator matrix is a matrix-valued function of \( t \in [0, T] \), which changes over time.

We will start with the easier case of constant parameters.

4.1. Estimation of Constant Parameters

In this section we will only consider the case of constant parameters. We therefore have \( Q(t) = Q \in \mathbb{R}^{N+1 \times N+1} \) for all \( t \in [0, T] \) and therefore the Kolmogorov forward equation reduces to

\[
\dot{P}(t) = P(t) \cdot Q, \quad P(t_0) = I, \quad P(t) \in \mathbb{R}^{N+1 \times N+1}.
\]

This system of ordinary differential equations is well studied and known to have a closed form solution. The solution is given by

\[
P(t) = P(t_0) \cdot \exp(tQ),
\]

where \( \exp(\cdot) \) is the matrix exponential function.

We now use the framework we set up in Section 3.2.1 and Section 3.2.3, to estimate our parameter matrix \( Q \). We choose the loss function \( L: \mathbb{R}^{N+1 \times N+1} \times \mathbb{R}^{N+1 \times N+1} \to [0, \infty) \) given by \( L(Y, C) = \|Y - C\|_F^2 \), i.e. the squared Frobenius norm which results in a matrix-valued equivalent to the well known least squares approach.

For a given data set

\[
\mathcal{D} = \left\{ (t_0, P(t_0)), (t_1, P(t_1)), (t_2, P(t_2)), \ldots, (t_M, P(t_M)) \right\} \subset [0, T] \times \mathbb{R}^{N+1 \times N+1},
\]

we consider the empirical risk \( \mathcal{R}_{L,D} \). In the case of constant parameters, an obvious choice for the function space for the optimization problem to attain a good approximation to the Bayes risk \( \mathcal{R}_{L,P} \) is given by the closed form solution to the system of differential equations and we only need to optimize the parameters, i.e. the matrix \( Q \in \mathbb{R}^{N+1 \times N+1} \). Therefore the optimization problem yields

\[
\min_{Q \in \mathbb{R}^{N+1 \times N+1}} \frac{1}{M + 1} \sum_{i=0}^{M} \left\| \exp(t_i Q) - P(t_i) \right\|_F^2.
\]
4.1. Estimation of Constant Parameters

Depending on the assumptions we make on the matrix $Q$ this optimization problem changes in terms of the constraints on the space we optimize in. Explicitly solving these kinds of optimization problems is beyond the scope of this thesis, we therefore refer to the literature, for example see [44], [45] or [46].

A first, and for our problem very reasonable, assumption we want to put in is that $Q$ is a generator matrix, i.e.

fulfills the $q$-Property from Definition 3.6, and that we only have transitions from one state to the state next to it, i.e.

only one birth or one death per time step $\Delta t$ which results in the tridiagonal structure. Formalizing this a little more, with these restrictions on the matrix $Q$ we get

$$\min_{Q \in \mathcal{G}} \frac{1}{M+1} \sum_{i=0}^{M} \left\| \exp(t_i Q) - P(t_i) \right\|_F^2,$$

where $\mathcal{G} := \{Q \in \mathcal{T}(N + 1) : q_{ij} \geq 0$ for $i \neq j$ and $q_{ii} = -\sum_{j \neq i} q_{ij} \}$, the space of possible generator matrices for the birth and death process which fulfill the $q$-Property from Definition 3.6.

Since we made even more assumptions on the structure of our generator matrix $Q$ we want to incorporate them as well. Let the matrix $Q$ have a structure as in Equation (2.4) but with constant birth- and death-rates, i.e.

$\lambda(t) = \lambda \geq 0$ and $\mu(t) = \mu \geq 0$ for all $t \in [0, T]$.

We want to describe this structure by decomposing our matrix $Q$ into a sum of two matrices which are responsible for the part of the birth- and death-rate respectively. So let

$$Q = \lambda Q_\lambda + \mu Q_\mu,$$

where

$Q_\lambda = \begin{bmatrix}
-1 & 1 & & & \\
0 & -1 & 1 & & \\
& 0 & -1 & 1 & \\
& & \ddots & \ddots & \ddots \\
& & & 0 & -1 & 1 \\
& & & & 0 & 0
\end{bmatrix} \in \mathbb{R}^{N+1 \times N+1}$

and

$Q_\mu = \begin{bmatrix}
0 & 0 & & & \\
1 & -1 & 0 & & \\
2 & -2 & 0 & & \\
& \ddots & \ddots & \ddots \\
& & N-1 & -(N-1) & 0 \\
& & & N & -N
\end{bmatrix} \in \mathbb{R}^{N+1 \times N+1}.$

Modifying our optimization problem from Equation (4.8) according to the proposed decomposition we get

$$\min_{\lambda, \mu \in \mathbb{R}_{\geq 0}} \frac{1}{M+1} \sum_{i=0}^{M} \left\| \exp(t_i (\lambda Q_\lambda + \mu Q_\mu)) - P(t_i) \right\|_F^2.
4. Parameter Estimation

In this way we are in the much more comfortable position of optimizing in the space $\mathbb{R}^2$, instead of the space of matrices $\mathbb{R}^{N+1 \times N+1}$, so we reduced the degrees of freedom in the optimization problem immensely.

After we had a look at the case of constant parameters, we will now go to the more interesting and for our application crucial case of time-dependent parameters.

4.2. Estimation of Time-Dependent Parameters

Opposed to the case with constant generator matrix $Q(t) = Q$ for all $t \in [0, T]$, there is no closed form solution for the time-dependent case which is desired for our model. This means the data set we have is not immediately of use to us. We need to connect the information provided by the transition probability matrices to the time-dependent generator matrix. To do that, we use again the Kolmogorov forward equation

$$\dot{P}(t) = P(t) \cdot Q(t), \quad P(t_0) = I, \quad P(t) \in \mathbb{R}^{N+1 \times N+1}. \tag{4.14}$$

To get a data set we can work with from this equation and the given data set we therefore substitute the derivative by the differential quotient and get

$$\frac{1}{t_{j+1} - t_j} (P(t_{j+1}) - P(t_j)) = Q(t_{j+1}) \cdot P(t_{j+1}), \quad j = 0, 1, \ldots, M - 1. \tag{4.15}$$

If we have a $P(t_{j+1})$ that is invertible, we get

$$\frac{1}{t_{j+1} - t_j} (P(t_{j+1}) - P(t_j)) \cdot P(t_{j+1})^{-1} = Q(t_{j+1}), \quad j = 0, 1, \ldots, M - 1. \tag{4.16}$$

With this procedure we get the needed samples from our matrix-valued function, i.e.

$$\{Q(t_1), \ldots, Q(t_M)\} \subset \mathbb{R}^{N+1 \times N+1}. \tag{4.17}$$

But, there are two issues that need to be addressed here:

1. due to the approximation of the derivative by the differential quotient the resulting matrix might not be tridiagonal any more.

2. we need $P(t_{j+1})$ to be invertible. This need not be the case for every sample scheme.

To overcome these two problems we modify our approach a little bit. First, we rewrite Equation (4.15) as

$$(I - (t_{j+1} - t_j)Q(t_{j+1})) \cdot P(t_{j+1}) = P(t_j), \quad j = 0, 1, \ldots, M - 1. \tag{4.18}$$

Then we consider the minimization problem

$$\min_{Q(t) \in \mathcal{T}(N+1)} \|Q(t_{j+1})\|_F \quad \text{s.t.} \quad \text{Equation (4.18) holds,} \tag{4.19}$$

where $\mathcal{T}(N+1) \subset \mathbb{R}^{N+1 \times N+1}$ is again the space of tridiagonal matrices and $\|\cdot\|_F$ is the Frobenius norm. We could put even more information about the structure of the generator matrix $Q(t)$ into
the constraints of our optimization problem, like fulfilling the q-Property (see Definition 3.6), but for the moment we go with this weaker constraints. In any case, this procedure gives us a kind of data samples of the form

\[(4.20) \quad \mathcal{D}_{\text{mod}} = \left\{ (t_1, Q(t_1)), \ldots, (t_M, Q(t_M)) \right\} \subset [0, T] \times \mathcal{T}(N + 1), \]

which we need to continue. Together with the framework we already used and discussed in Section 4.1 and Section 3.2 respectively, we can now estimate the time-dependent generator matrix for the model of the respective parking segment.

As a first step to do that, we construct an appropriate matrix-valued kernel. In Section 3.2 we discussed kernels, which map to the space of real matrices or a subspace thereof which we denoted by \( \mathcal{M} \). But, since we made more assumptions on our generator matrix \( Q(t) \) in Section 2.2 that resulted, amongst other things, in a tridiagonal structure, we have to adapt the kernel to this. So, we want to construct a kernel that allows us to estimate functions mapping to the space of tridiagonal matrices. We therefore set \( \mathcal{M} = \mathcal{T}(N + 1) \). By Theorem 3.35, the result will be a corresponding RKHS \( \mathcal{H} \) of functions mapping to the space of tridiagonal matrices \( \mathcal{T}(N + 1) \) which we can use as the function space for the risk minimization problem, as described in the beginning of Section 3.2.1. We also know that our timestamp samples are from an interval \([0, T] \subset \mathbb{R} \), so we set \( X = [0, T] \subset \mathbb{R} \). The kernel we want to construct is therefore a mapping \( K : [0, T] \times [0, T] \rightarrow \mathcal{B}(\mathcal{T}(N + 1)) \). Besides the structure property of the target space of the function we seek to estimate, we want the to be constructed kernel to be able to modify all of the entries of the matrix-valued function individually. This is to make sure we get a good approximation to the given data set.

One possibility for a kernel capable of covering these requirements is an adaption of the diagonal kernels proposed in [36]. In Section 3.2.2 we studied several ways to construct matrix-valued kernels. To construct this kind of diagonal kernel Theorem 3.42 will prove particularly useful. The diagonal nature of the kernel will be structure preserving, so when looking at Theorem 3.45 we see that given coefficients of the correct structure, we will get a minimizer of the risk minimization problem that is a tridiagonal-valued function.

To apply Theorem 3.42 we need to specify the number \( D \in \mathbb{N} \) of scalar-valued kernels we want to use, the scalar-valued kernels themselves and the positive definite structure preserving tensors of 4th order. Since we want every matrix entry to be influenced individually, we choose \( D = (N + 1)^2 \). Hence, we need scalar-valued kernels \( k_{ij} : [0, T] \times [0, T] \rightarrow \mathbb{R}, \ i, j = 1, \ldots, N + 1 \), so one kernel for each entry of the matrix-valued function we want to approximate. Not all of these kernels have to be different, but this way of construction gives us the possibility to be that flexible. The last ingredients needed for the construction method of Theorem 3.42 are the 4th order tensors \( A_{ij} \in \mathcal{B}_+(\mathcal{T}(N + 1)) \), \( i, j = 1, \ldots, N + 1 \). For these tensors we first define \( E_{i_1,j_1,i_2,j_2} \in \mathcal{B}(\mathcal{M}) \) as

\[(4.21) \quad \left( E_{i_1,j_1,i_2,j_2}\right)_{i_1'j_1'i_2'j_2'} := \begin{cases} 1 & (i_1,j_1,i_2,j_2) = (i_1',j_1',i_2',j_2') \\ 0 & (i_1,j_1,i_2,j_2) \neq (i_1',j_1',i_2',j_2') \end{cases} \]

If we now consider the tensors \( E_{i,j,i,j} \) we first notice that \( E_{i,j,i,j} \in \mathcal{B}(\mathcal{T}(N + 1)) \). Also, we have \( E_{i,j,i,j} \in \mathcal{B}_+(\mathcal{T}(N + 1)) \) for all \( i, j = 1, \ldots, N + 1 \). To see this, let \( Y \in \mathcal{T}(N + 1) \) and the
4. Parameter Estimation

entries be denoted by $y_{ij}$, then

\[(4.22) \quad (E_{i,j,i,j}Y,Y)_F = (\bar{Y}, Y)_F,\]

where $\bar{Y} = (\tilde{y}_{kl}) = y_{ij}$ if $(k, l) = (i, j)$ and zero otherwise. This then gives

\[(4.23) \quad (\bar{Y}, Y)_F = \sum_{k=1}^{N+1} \sum_{l=1}^{N+1} \tilde{y}_{kl} y_{kl} = y_{ij} y_{ij} = y_{ij}^2 \geq 0.\]

So, by Theorem 3.42 we get a positive definite matrix-valued kernel fitting our requirements via

\[(4.24) \quad K(x, t) = \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} E_{i,j,i,j}^k k_{i,j}(x, t).\]

This kernel can be interpreted as a diagonal matrix of dimension $(N + 1)^2 \times (N + 1)^2$ if we choose to columnwise reorder the target space matrix to a column vector and consider linear maps in between those kind of vectors.

The next step to estimating the generator matrix $Q(t)$ is to formulate the regression problem. We therefore choose again the squared Frobenius norm as loss function $L: Y \times M \rightarrow [0, \infty)$, i.e.

\[L(Y, C) := \|Y - C\|_F^2.\]

Given the modified data set $D_{mod} = \{(t_1, Q(t_1)), \ldots, (t_M, Q(t_M))\} \subset [0, T] \times \mathcal{T}(N + 1)$ we constructed earlier (see Equation (4.19)), we use the regularized empirical risk $R_{L,D,\alpha}$ to approximate the regularized Bayes risk $R_{L,P,\alpha}$, where we will use the reproducing kernel Hilbert space $\mathcal{H}$ corresponding to the kernel given by Equation (4.24) as function space to find the minimizer in. We therefore get

\[(4.26) \quad R_{L,P,\alpha}^* = \inf_{f \in \mathcal{H}} \frac{1}{M} \sum_{m=1}^{M} \|f - Q(t_m)\|_F^2 + \alpha \|f\|_{\mathcal{H}}^2.\]

Now, we use the representer theorem (Theorem 3.45), which gives $f_{L,D,\alpha}^*(t) = \sum_{i=1}^{M} K(t, t_i) C_i$ with $C_1, \ldots, C_M \in \mathcal{T}(N + 1)$, and get the minimization problem

\[(4.27) \quad \min_{C_1, \ldots, C_M \in \mathcal{T}(N + 1)} \frac{1}{M} \sum_{m=1}^{M} \left\| \sum_{i=1}^{M} K(t_m, t_i) C_i - Q(t_m) \right\|_F^2 + \alpha \left\| \sum_{i=1}^{M} K(t, t_i) C_i \right\|_{\mathcal{H}}^2.\]

For explicitly solving this problem we refer again to the literature (\cite{44,45,46}). If we want to incorporate more of our assumptions on the generator matrix we made in Section 2.2 into this approach, we need to start again with the kernel. So, this is also a lot more difficult than in the case with constant parameters.
4.2. Estimation of Time-Dependent Parameters

After formulating the optimization problems for our parameter estimation, in the next and final part of the thesis we want to summarize our results and give an overview of open problems and remaining research opportunities related to this thesis. For example, we will see that incorporating more or other model assumptions into the kernel, like the row sum of the generator matrix being zero or the q-Property in general, is one of these open questions.
5. Summary and Outlook

In this final part we want to briefly summarize what the results of this thesis are and, since we started with a very broad and general question into our research, give a little more extended overview of open questions and problems that need to or can be addressed in future research.

5.1. Results

In this thesis we presented one possible model for the problem of parking prediction and described the procedure to go from raw data about the process to probabilities of the filling levels in the segment (see Chapter 2). After describing the setup for the model, we rigorously introduced the necessary mathematical framework to analyze it and also provided a general method for the parameter fitting of the proposed model. In the course of this, we adapted the results about loss functions and their risk functionals to the case of matrix-valued functions (see Chapter 3). Also the necessary results about minimizing the risk functional when using reproducing kernel Hilbert space were provided. We then used these mathematical tools to derive optimization problems that allow us to calculate the parameters of the proposed Markov model based on a given data set (see Chapter 4). Finally, we will give an overview of open issues regarding the proposed model and also parking prediction in general in Section 5.2.

As far as we know the approach of estimating the parameters of the Markov model with methods based on reproducing kernel Hilbert spaces has not been used before, in particular not for the purpose of free parking space prediction. To our knowledge, the results about loss functions have not been studied in this detail for the matrix-valued case in the literature. By making use of the derived results we, in conclusion, extended a classical approach (see [2]) to describe the behaviour of parking cars to a data driven model for the prediction of free parking spaces in urban areas.

In the next and final section of this thesis we will point out some open problems that provide the opportunity for further research. There are even more issues to address for a real world application of this approach, especially concerning technical difficulties, like the transmission of the information between the cars and the backend. But we hope to give an insight in some of the open questions related to pure mathematics, modeling or numerical analysis.

5.2. Open Problems

In this section, we want to give an overview of open questions, not pursued approaches and remaining research possibilities in the context of this thesis and parking prediction in general. There are several issues that could not be addressed but we want to give the reader the opportunity to have a closer look on their own. Most of these issues bring along several more questions. We do not attempt to cover all of them, but we will try to give an idea what troubles and opportunities might come along with these approaches.
5. Summary and Outlook

We will start with open mathematical issues that are connected with the model we presented in this thesis. Afterwards, we will go to more general questions that need to be addressed in the context of free parking space prediction and therefore are more related to the modelling of the problem itself. Also the issue of real time information will be briefly discussed.

5.2.1. Incorporate Model Assumptions in Kernel and Corresponding RKHS

We first want to address an open issue that is mainly mathematical and has great influence on the result of the estimation of the generator matrix $Q(t)$. We have seen in Section 3.2.3 with the representer theorem (Theorem 3.45), the application of it in Section 4.2 and when constructing the kernel that all assumptions we made on the structure of the generator matrix $Q(t)$ have to be encoded in the subspace $\mathcal{M} \subset \mathbb{R}^{n \times m}$ we want our function to map to and in the kernel. So far, we made sure that our solution $f^*_{L,D,\alpha}$ to the risk minimization $R_{L,D,\alpha}$ maps to the space of tridiagonal matrices $\mathcal{T}(n)$. Other information, some that are necessary for the framework of Markov processes, like the q-Property from Definition 3.6, and others that come from the modeling, like the linear dependence of the death-rate to the state of the process $X(t)$, have not been incorporated yet. Also the irreducibility of the resulting matrix-valued function $Q(t)$ from Definition 3.11 needs to be satisfied.

So, in general we need to gain a better insight in the influence of the kernel on the properties of the target space $\mathcal{M}$ and how to make sure we get and keep the properties we need for our model to work.

5.2.2. Approximation Quality of the RKHS for a Specific Kernel

A more general issue that was not addressed in this thesis is concerned with the approximation properties of the corresponding RKHS of a chosen kernel. So, how does for example a certain smoothness of the kernel influence the elements of the RKHS. Results of this kind are given in Section 4.3 and 4.4 of [34] for the scalar-valued case. Depending on how general we look at this issue, especially with regard to the target space $\mathcal{M}$ of our functions, not even the necessary tools we need to derive results like this are obvious. For a finite dimensional space like we discussed, we might be able to make use of the machinery of Fourier analysis, but for a general Hilbert space of operators this is not immediately possible anymore.

5.2.3. Numerical and Computational Issues

A whole bunch of questions regarding the application of our framework have been neglected in this thesis. After constructing a kernel and making sure it has and preserves the properties we want, we need to compute the solution of the minimization problem we end up with. This problem of course depends on the choice of the loss function $L$ and the size of the data set along with many more things. Also the choice of the kernel can be varied and needs to be evaluated. But in any case, we want to make sure that we can actually compute the solution to it in a feasible amount of time for example. Numerical stability is another issue we did not address so far. The question of fast and problem specific algorithms will also come up with the growth of the data set.

We now turn to open problems that are more model related than specific for our presented approach.
5.2.4. Model Evaluation

An important issue concerning modeling in general is, of course, the evaluation of the described model against real life situations. So, does the model reproduce the behaviour that can be observed at a parking segment in the city. Together with this goes the choice of the loss function. In Chapter 4, we chose the Frobenius norm as a loss function to derive the optimization problem, but it is not clear that this is the best choice for the problem. We also constructed a possible kernel in this process but here as well are still a lot of not investigated possibilities. Evaluating all this is quite a big challenge since, as we described in Section 2.5, there is not only the prediction on which we mainly focused in this thesis, but also several other issues that influence the final result.

As a start for the evaluation the information from parking structures, which are more precise and easier to access than for on-street parking could be used. Besides the fact that we want to make sure our model reproduces the real life situation in this process we also need to check whether the assumptions we made throughout the modelling hold up and which ones might need to be relaxed or completely changed.

5.2.5. M/M/c vs. M/M/c/K Queues

One of the model assumptions that might need to be changed is the type of queue we are looking at. We briefly described in Section 2.3.1 that our model can be interpreted as a queueing problem. We classified it as an M/M/c queue, so a queue with infinite capacity for waiting customers or in our case cars. This is certainly a valid option, but we would like to consider the possibility that this is not entirely accurate. While looking for a parking spot one driver does not necessarily wait at a particular parking segment until another person pulls out and there is a free spot. The driver goes on and looks at the next parking segment on the way. If the driver sees other cars circling the block, also waiting for a free spot, the driver might consider going to a whole different area. This suggests to use a queue of the M/M/c/K type, with a finite amount of spots in the waiting line. When we consider the special case of M/M/c/c this is also called a loss system. Another aspect that we might consider in this line of questioning is the rule by which the waiting cars are processed. If we assume that a driver looking for a parking spot moves from one parking segment to the next if there are no free spots available, then ”first come first serve” might not be correct, but rather ”last come first serve”.

So, this open problem comes down to a question of modeling. What therefore remains to be done is to take both possible models and compare them in terms of how they fit our observations. As the availability of extensive data was a problem from the beginning, this remains another open challenge.

5.2.6. Estimating Time-Dependent Parameters Based on the Work of William A. Massey

Often when it comes to Markov chains or queuing models the results given in the literature are restricted to constant parameters instead of time-dependent ones. In contrast, William A. Massey et al. have, as a result of their research, published results about time-dependent queues, especially in the context of telecommunication systems in their various works (see for example
5. Summary and Outlook

We hope that their analysis can be used for a different approach to estimate the time-dependent parameters in our queueing model, given the most basic data set we assumed, of various states of the process $X(t)$ for each parking segment over time.

5.2.7. Compartment Model

Although the Markov model leading to a queueing problem, or more general, a birth and death process is a pretty obvious choice, which has been used for quite some time for this application in the literature \cite{2} \cite{3}, we already pointed out that the data we can gather right away does not perfectly fit our need. Therefore, we want to describe a different model that might offer another way for free parking space prediction.

Compartment models are most commonly used in ecology and also in biology, biomedicine, physiology and genetics. But, as it turns out, they might also be very well suited to be applied to traffic flow problems such as parking prediction \cite{50}. For an introduction into the topic we refer to \cite{51} and \cite{52}. So in this section we will shortly describe how we can use a compartment model for the problem of free parking space prediction and what upsides as well as downsides this approach brings along.

There are different types of compartment models which are distinguished by the structure in between the compartments. The model we deem best suited for parking prediction is a mammillary model. The key point of a mammillary model is the fact that there is one central compartment that is connected to all the other compartments, which are not interconnected on their own. In our case, the central compartment represents the traffic network, so the streets and highways. The other compartments take the place of our parking segments $E_k$. To make a prediction about parking availability, we then need to get an insight about the filling state of each compartment over time. The filling state is governed by the in- and outgoing flow of each compartment. We get information about the flows and the filling states in two ways. The first way is, if it is available, directly by the state of the filling state process of the compartment, so the state of $X(t)$ as in the Markov model we discussed. The difference to the Markov model, and what makes the compartment model very appealing, is that we can also make use of the information we get easily by a car, i.e. where it starts and leaves a compartment and where it parks and enters a compartment. These events are data points for the compartment model that provide information about the in- and outgoing flow of the compartment. So, we can use the information we can easily gather in a more direct way. From here we could derive probabilities of the filling level of the compartments which directly correlate to the amount of available parking spots, if we know the amount $N_k$ of cars that can park in that compartment.

On the other hand, the compartment model also induces some new problems we have to deal with. Problems that did not occur in that way with the Markov model approach. One issue is that the compartment model approach leads to one big system of equations for all the compartments and their input and output flows together. So, the number of parking segments will be the dimension of the resulting matrix system. An issue that goes along with that will be that we need information on all the flows of the compartments at the same time. Even if we accumulate data to every minute, or every 5 minutes, that will not be easy to accomplish. Especially, since not every parking segment or compartment is frequented the same way. Here we might have to make use of matrix completion methods again. Another way to circumvent this problem can be
to divide the city into smaller areas. Of course, we then have to analyze how this influences the results of the prediction and how these smaller areas interact with each other.

To sum up, we will experience upsides as well as downsides in the case of using a compartment model to describe our problem. Another approach, that therefore might seem even more appropriate, is completely omitting the model and works purely with data.

5.2.8. Purely Data-Driven Approach

So far, we presented and discussed various approaches involving a certain model that we think describes the problem and can therefore give further insight about the observed phenomenon. As we have seen, there are several uncertainties going along with the modeling of parking, independent of which model we actually choose.

Therefore, a completely different approach to the task of predicting free parking spaces that omits the model entirely might be very appealing. Especially, the high level of complexity with the various challenges going along with them suggests a less sophisticated but more problem oriented approach. This might be done by using a purely data-driven method for the prediction. Although this lacks the additional benefit of getting an insight in the causality behind the observed phenomenon, the solution we get might be more accurate since we do not fit the data to a beforehand unknown but assumed model. Also, this gives us the possibility to additionally make use of a lot more different kinds of information, i.e. parking ticket data, type of area, public transport availability and so on. In [53] we can find a comprehensive list of influence factors for parking that we can use to find more data sources. According to a Q&A interview on one of their investors website this approach is used by parknav - Street Parking Reinvented, see [54] and [55], with quite some success.

5.2.9. Real Time Information

Even though the prediction of free parking spaces based on historical data is already a big advantage to the current system of just going to the final destination and then starting to look around for a free spot, an aspect the driver would even more benefit from is real time information. So another key feature to the solution of finding a free parking spot is combining the prediction of the historic data with real time information. Future developments in regards of sensor availability and accuracy might become useful to solve this problem.

A method that might become useful in this context is the Kalman filter or its time-continuous version, the Kalman-Bucy filter. For more details about these methods we refer to [56], [57], [58] and [59].

5.2.10. Determine the Occupancy of a Parking Segment by Data From One Car

As mentioned in Section 2.4.1, the process of gaining information about the state of the process from just one car is still an unsolved issue. It is of course closely related to the kind of car we have at hand and especially what sensors it brings with it.

Of course, this listing is not even close to being complete, but we hope that it helped, together with Section 2.6, to get a better understanding of the various types of issues involved with parking.
5. Summary and Outlook

prediction. Also it shows the many remaining research opportunities in this field, be it in terms of modeling and evaluation or pure mathematics.
A. Appendix

Here, we state the additional results used throughout this thesis. For the proofs we refer to the literature.

In this section we use, \((X, \mathcal{F})\) a measurable space, \(Y\) a metric space and \(P : \mathcal{F} \otimes \mathcal{B}(Y) \to [0, \infty]\) where \(\mathcal{B}(Y)\) is the Borel \(\sigma\)-algebra, if not otherwise stated.

**Theorem A.1** (Regular conditional distribution)
Let \((X, \mathcal{F})\) a measurable space, \(Y\) a metric space and \(P : \mathcal{F} \otimes \mathcal{B}(Y) \to [0, \infty]\). Then there is a mapping \(P(\cdot | \cdot) : \mathcal{B}(Y) \times X \to [0, 1]\) with the following properties:

(i) \(P(\cdot | x)\) is a probability measure on \(\mathcal{B}(Y)\) for all \(x \in X\).

(ii) the mapping \(x \mapsto P(B | x)\) is measurable for all \(B \in \mathcal{B}(Y)\).

(iii) \(P(A \times B) = \int_A P(B | x) dP_X(x)\)

\(P(\cdot | \cdot)\) is called regular conditional distribution. And \(P_X(A) = P(\pi^{-1}(A))\) where \(\pi : X \times Y \to X\) is the projector.

**Theorem A.2** (Lemma of Caratheodory)
Let \((X, \mathcal{F})\) be a measurable space, \(Z\) be a Polish space equipped with its Borel \(\sigma\)-algebra, and \(h : X \times Z \to \mathbb{R}\) be a map.

Then \(h\) is measurable if the following two conditions are satisfied:

(i) \(h(x, \cdot) : Z \to \mathbb{R}\) is continuous for all \(x \in X\).

(ii) \(h(\cdot, z) : X \to \mathbb{R}\) is measurable for all \(z \in Z\).

**Definition A.3** (Fréchet differentiable)
Let \(X, Y\) be normed spaces and \(U \subset X\) be an open subset of \(X\). An operator \(A : U \to Y\) is called Fréchet differentiable at \(f \in U\) if there exists a bounded, linear operator \(A'_f : X \to Y\) such that

\[
\lim_{\|h\|_X \to 0} \frac{\|A(f + h) - A(f) - A'_f h\|_Y}{\|h\|_X} = 0.
\]

**Theorem A.4** (Mean Value Theorem for Functions in Several Variables)
Let \(U \subset \mathbb{R}^n\) an open, connected subset of \(\mathbb{R}^n\) and let \(f : U \to \mathbb{R}\) a differentiable function. Moreover, let \(x, y \in U\) such that \(\{z \in \mathbb{R}^n : z = (1 - \beta)x + \beta y, \beta \in [0, 1]\} \subset U\).

Then there exists a \(z \in U\) such that

\[
f(y) - f(x) = (\nabla f(z), y - x),
\]

where \((\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}\) is the scalar product on \(\mathbb{R}^n\).

**Lemma A.5** (Continuity of convex functions)
Let \(f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}\) be a convex function and \(\text{Dom} f := \{t \in \mathbb{R} : f(t) < \infty\}\).

Then we have:
A. Appendix

(i) \( f \) is continuous at all \( t \in \text{IntDom} f \).

(ii) If \( f \) is lower semi-continuous, then \( f|\text{Dom} f \) is continuous.

**Theorem A.6** (Existence of minimizers)

Let \( E \) be a reflexive Banach space and \( f : E \to \mathbb{R} \cup \{\infty\} \) be a convex and lower semi-continuous map. If there exists a \( c > 0 \) such that \( \{x \in E : f(x) \leq c\} \) is non-empty and bounded, then \( f \) has a global minimum, i.e., there exists an \( x_0 \in E \) with

\[
f(x_0) \leq f(x) \quad x \in E.
\]

Moreover, if \( f \) is strictly convex, then \( x_0 \) is the only element minimizing \( f \).
Bibliography


