Asymptotic results in nonparametric Bayesian function estimation

Kirichenko, A.

Citation for published version (APA):
Bayesian nonparametric methods are widely used in practical applications. They have numerous attractive features such as their philosophical appeal, conceptual simplicity, and ability to easily incorporate prior knowledge about the model. However, putting a prior on a large function space might result in erroneous estimates or suboptimal performance. Therefore, it is essential to study Bayesian procedures to gain insight about which priors to use and how their tuning affects the performance of these procedures. One way to do it is to take an asymptotic approach and to analyse Bayesian methods from the frequentist point of view by assuming that there exists a true underlying function and studying how fast a particular Bayesian procedure captures the truth as the number of observations goes to infinity.

In this thesis, we consider function estimation problems in two different statistical settings. First, we discuss regression and binary classification problems on large graphs, where the goal is to estimate a smooth function defined on the vertices of a graph. In the second setting, we aim to estimate the intensity of an inhomogeneous Poisson process from a realised point pattern. For both problems, we develop adaptive Bayesian procedures and study their asymptotic behaviour from the frequentist perspective. In particular, we derive contraction rates for our procedures and show that they are optimal in a minimax sense.
Asymptotic results in nonparametric Bayesian function estimation

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus, prof.dr.ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op 5 oktober 2017, te 12:00 uur

door

Alisa Kirichenko

geboren te Sint-Petersburg, Rusland
Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Promotor: prof.dr. J.H. van Zanten Universiteit van Amsterdam
Copromotor: dr. B.J.K. Kleijn Universiteit van Amsterdam
Overige leden: dr. A.J. van Es Universiteit van Amsterdam
prof.dr. S. Ghoshal North Carolina State University
prof.dr. M.C.M. de Gunst Vrije Universiteit Amsterdam
prof.dr. M.R.H. Mandjes Universiteit van Amsterdam
prof.dr. R. Nunez Queija Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Research supported by the Netherlands Organisation for Scientific Research (NWO).
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samenvatting</td>
<td>89</td>
</tr>
<tr>
<td>Author contributions</td>
<td>91</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>93</td>
</tr>
</tbody>
</table>
Introduction

This thesis is dedicated to the scientific field of mathematical statistics. In order to understand what kind of problems we are dealing with consider the following example. Figure 1 shows a component of the social network graph from the Framingham Heart Study conducted in 2000. The vertices of the graph represent people and the edges show ties among siblings, spouses, and friends. Vertex size is made proportional to people’s body mass index (BMI), and the vertex is coloured yellow if the BMI of the corresponding person is above 30 (which indicates obesity) and green otherwise. Christakis and Fowler [2007] show that the chance of a person becoming obese increases if his graph-neighbours (siblings/spouses or friends) became obese. In real life one wouldn’t have the value of BMI for every person, hence some of the vertices are going to be blank. Using statistical analysis we can try to fill in that data by predicting a person’s BMI based on the geometry of the social network and the information about other people in the graph.

The problem above is an example of function estimation problem, where the function of interest is the function on the vertices of the graph describing the BMI index. Of course, there are plenty of other scenarios in which the target function can be defined on different domains and take values in different spaces. In order to be able to make predictions a statistician chooses a model that incorporates his knowledge and assumptions about the underlying function. A fundamental problem in statistics is to develop models based on a sample of observations so that further analyses can be carried out. If the set of functions in the model is not very large (can be indexed by a subset of \( \mathbb{R}^d \)) such a model is called parametric. If the collection of functions considered is infinite dimensional the model is nonparametric. When there is enough information about the underlying function, parametric models are favourable. However, such models lack flexibility and are sometimes not robust enough in the sense that even slight contamination of the data by observations not following the particular parametric model might lead to incorrect conclusions. Nonparametric models are more flexible in that sense, and more robust against misspecification.

The study of nonparametric models started in the mid twentieth century. According to Prakasa Rao [1983], the first paper in the area of nonparametric function estimation is Rosenblatt [1956]. Since then nonparametric models have greatly evolved and gained tremendous popularity. A good introduction to the topic is the
Introduction

Figure 1: This is the largest connected component of the Framingham Heart Study Social Network conducted in 2000 with the total number of vertices equal to 2,200. The colour of the vertices indicates obesity (yellow for the BMI $> 30$, green for the BMI $\leq 30$), and the vertex size is proportional to the BMI.

In this thesis we adopt Bayesian nonparametric techniques for function estimation. Bayesian methods have become more and more common in the mainstream statistical literature in the last three decades. The essence of the Bayesian paradigm is the philosophical idea that the object of interest does not have one true value, but is rather perceived as a random object itself. The evidence about the true state of the world is then expressed in terms of degrees of belief. In order to make an inference a Bayesian statistician first puts a prior on the parameter class that represents his initial belief about the origin of the data. Central in the Bayesian framework is the posterior distribution which can be viewed as a data–amended version of the prior after the observations have been made. Bayesian methods can be appealing for philosophical reasons, their conceptual simplicity, ability to easily incorporate prior knowledge about the model, etc. One of the pragmatic reasons to use Bayesian techniques is that they can offer a form of regularisation to prevent overfitting. In overfitting a statistical model describes random error or noise instead of the underlying function. It occurs when a model is excessively complex. A way to prevent this from happening is to favour simple functions by introducing a penalty on more complex functions. In the Bayesian context this can be accomplished by assigning higher probabilities to functions that are considered to be more likely, for example, because of their smoothness properties. The list of books on Bayesian methods includes Berger [1993], Ghosh and Ramamoorthi [2003], and Ghosal and
However, putting a prior on a large function space makes it easy to obtain erroneous estimates or to have suboptimal performance (for more details see Section 1.5). Therefore, it is essential to study Bayesian procedures to gain insight about which priors to use and how their tuning affects the performance of these procedures. One way to do it is to take an asymptotic approach and to analyse Bayesian methods from the frequentist point of view by assuming that there exists a true underlying function and studying how fast a particular Bayesian procedure captures the truth as the number of observations goes to infinity. Among other factors the learning rate of the procedure is known to depend on the regularity of the underlying function. Indeed, it takes more observations to be able to capture the spikes of a rough function than to make an accurate estimate of a constant function. In practical applications the smoothness of the target functions is rarely known, hence it is desirable to devise rate-adaptive procedures that can attain fast learning rates across a wide range of regularities of the function.

In this thesis we develop Bayesian procedures and study their asymptotic behaviour in the context of two different statistical settings. First, we discuss regression and binary classification problems on large graphs, where the goal is to estimate a smooth function on the vertices of a graph. The second problem that we consider is the problem of intensity estimation of an inhomogeneous Poisson process from a realised point pattern.

This thesis starts by a somewhat broader introduction into Bayesian nonparametric statistics in Chapter 1. Chapters 2 and 3 cover the function estimation problem on large graphs. In Chapter 2 we define the general setting and the problems for which we derive minimax rates. In Chapter 3 we devise adaptive Bayesian procedures for function estimation problems in the aforementioned setting and we show that these procedures attain optimal rates of convergence. Chapter 4 studies the problem of Poisson intensity estimation and presents an adaptive Bayesian method that has an optimal rate of convergence.

In what follows we discuss the content of each chapter in more detail.

Bayesian nonparametric statistics (Chapter 1)

In Chapter 1 we cover the basic definitions and concepts of Bayesian nonparametric statistics with the focus on theoretical assessments of Bayesian methods. We present classic examples of nonparametric models and introduce a frequentist asymptotic approach to the evaluation of estimation procedures. We discuss minimax rates (the best possible learning rates) for the aforementioned models and give an example of an estimator that attains the minimax rate. Then we introduce the Bayesian approach to statistics and we discuss how performance of Bayesian procedures can be evaluated from the frequentist point of view. We consider two theorems that allow us to derive posterior convergence rates based on the information about the prior mass concentration and the complexity of the model. We finish the chapter by discussing ways to develop rate-adaptive procedures for function estimation in the cases when the regularity level of the target function is not known.
Function estimation on large graphs (Chapter 2 and Chapter 3)

There are various problems arising in modern statistics that involve making inference about a “smooth” function on a large graph. There are a number of possible scenarios: the graph can be given or it should be inferred first, there are regression or classification problems, sometimes it is only necessary to eliminate noise but sometimes only partial data is available. However, a lot of problems share two common features: the graph is usually large and the function of interest is somewhat smoothly varying along the edges of the graph. In these chapters we cover the case of undirected unweighted graphs without multiple edges or graph loops (edges that connect the vertex to itself). We don’t treat all possible scenarios exhaustively, but rather focus on two prototypical problems: regression and binary classification. Regression aims to recover a function on the vertices of the graph from noisy observations. In binary classification a label 0 or 1 is observed at each vertex and the object of interest is the “soft label” function $f$ whose value at a vertex $v$ is the probability of seeing a 1 at $v$.

In Chapter 2 we set up a general mathematical framework that allows us to study the performance of nonparametric function estimation methods on large graphs. We show that two major factors influence the difficulty of the problem. Firstly, despite the discrete structure of the data the smoothness of the target function, defined in a suitable manner, will have a crucial impact. Another aspect that will play a big role is the geometry of the graph. Indeed, the special cases of $d$-dimensional grids tell us that the complexity of the problem increases with the dimension.

We take an asymptotic approach and let the number of vertices of the graph grow asking how fast an estimation procedure can “learn” the underlying function of interest. We make two kinds of assumptions. Firstly, we assume that the target function has a certain degree of regularity, defined through the graph Laplacian, which is a matrix associated with the graph whose definition we recall in Chapter 2. The level of smoothness is not assumed to be known, as we are aiming to derive adaptive results. Secondly, we make an assumption on the asymptotic shape of the graph through a condition on the asymptotic behaviour of the spectrum of the Laplacian to tackle graphs that are sparse in nature and are “grid-like” in some sense. We provide the reader with various examples for which the assumption is satisfied. Additionally, in Chapter 2 we derive minimax rates for regression and classification problems on the graph in the described setting. The argumentation essentially follows the proof of the classical Pinsker’s Theorem (see for example, Tsybakov [2009]) with various modifications required due to the special nature of the problem.

In Chapter 3 we define Bayesian procedures using Laplacian regularisation to address regression and binary classification problems on large graphs in the same setting as in Chapter 2. We consider two types of priors on the space of functions on the graph. The first type performs regularisation using a power of the Laplacian, while the second uses an exponential function of the Laplacian. This can be seen as an analogue of Sobolev norm regularisation and the popular squared exponential prior on functions on Euclidean space, respectively. In both cases we consider
hierarchical priors that aim to achieve automatic adaptation to the regularity of the function of interest.

To study the performance of our Bayesian procedures we take an asymptotic perspective and let the number of vertices of the graph go to infinity. We address the question how fast the posterior distribution concentrates around the true function. We show that the first type of prior provides a procedure that is optimal in a minimax sense and adapts to the regularity of the true function up to a certain level. The second type is fully adaptive at the price of an extra log factor in the convergence rates.

We tackle both the full observation case and the case where only partial data is available. In the missing observation case we study MCaR (missing completely at random) mechanisms where the missingness of the observation does not depend on the values of the variables in the data. We additionally assume that the available data is somehow evenly distributed throughout the graph, meaning that there will be no large subgraphs for which no information is available. We show that in this setting slight modifications of the developed procedures achieve optimal rates and preserve their adaptation quality.

**Poisson intensity estimation (Chapter 4)**

Inhomogeneous Poisson processes are widely used models for count and point data in a variety of applied areas. Making an inference about the underlying intensity of a Poisson process from the available data is one of the most common problems in such applications. In this chapter we study a Bayesian approach to this problem. The present research is motivated by a hierarchical Bayes procedure for Poisson intensity estimation that was proposed by Adams et al. [2009]. These authors presented the first approach that is computationally fully nonparametric in the sense that it does not involve potentially inaccurate finite-dimensional approximations. In this method a prior that is a transformed Gaussian process with random scaling is studied with additional priors on the hyperparameter of the Gaussian process and on the multiplier. Simulation experiments and real data examples in Adams et al. [2009] show that the method can give very satisfactory results. However, there have been no results on the theoretical performance of this method. We again adopt asymptotic perspective and study the frequentist properties of this method. We show that with the appropriate choice of priors on the hyperparameters the procedure is fully adaptive and attains optimal rates of convergence.
Chapter 1

Bayesian nonparametric statistics

In this chapter we give an introduction to nonparametric statistical analysis with the focus on the theoretical assessment of Bayesian methods.

1.1 Statistical inference

Consider an observation $X$ which might be any mathematical object, such as a collection of numbers, vectors, matrices, functions, etc. Let $X$ belong to some space $\mathcal{X}$ that consist of other potential observations that we think we could have seen instead of $X$, meaning that we assume that $X$ has been a result of a random choice from the elements of $\mathcal{X}$. Denote by $P$ the distribution on $\mathcal{X}$ according to which this random choice has been made. This distribution is not known to us and usually is the object of interest. We restrict our possibilities by letting $P$ belong to a statistical model $\mathcal{P}$. The statistical model $\mathcal{P}$ is a collection of probability distributions that is usually parameterised using some parameter space $\Theta$

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}.$$ 

The assumptions about the data are incorporated in the model through the parameter $\theta$. The goal of a statistical inference is to learn different aspects of the probability distribution $P_\theta$ from the model that fits the data $X$ the best.

The parameter set $\Theta$ might be equal to a subset of Euclidean space, or it might be some infinite-dimensional space, for example, a space of functions. In the latter case we call the problem nonparametric.
1. Bayesian nonparametric statistics

1.2 Nonparametric models

In this thesis we focus on studying nonparametric models. Although, finite-dimensional models are simpler and attractive for computational convenience, they are often too restrictive. As more is assumed, when the assumptions are not correct, parametric models fail to deliver accurate estimates. Nonparametric models aim to make an estimation using fewer assumptions. These models are known to be more flexible and robust, which makes them attractive for practical applications. A broader introduction to nonparametric statistics can be found for instance in Wasserman [2004].

In function estimation problems it is common to consider nonparametric models with some assumption on the smoothness or regularity of the functional parameter \( \theta \in \Theta \), which can be defined in different fashions. In this thesis we use two kinds of smoothness spaces: the Sobolev and the Hölder spaces. Here we present the definitions of these spaces for functions defined on the interval \([0, 1]\). However, the notions can be generalised to functions on other domains as well.

Sobolev spaces are defined as follows. Let \( \psi_j \) denote an orthonormal basis of \( L^2[0, 1] \) and consider functions in \( L^2[0, 1] \) of the form

\[
\theta = \sum_{j=1}^{\infty} \theta_j \psi_j,
\]

where the \( \theta_j \) are the coefficients of \( \theta \) with respect to the basis \( \psi_j \). The Sobolev ball \( H^\beta(Q) \) of radius \( Q > 0 \) and regularity parameter \( \beta > 0 \) is given by

\[
H^\beta(Q) = \{ \theta \in L^2[0, 1] : \sum_{j=1}^{\infty} j^{2\beta} \theta_j^2 \leq Q^2 \}.
\]

When \( \beta \) is an integer and \( \psi_j \) is the classical Fourier basis, the Sobolev class is equal to the set of functions \( \theta : [0, 1] \to \mathbb{R} \) with \( \beta - 1 \) absolute continuous derivatives and with the \( \beta \)th derivative \( \theta^{(\beta)} \) satisfying

\[
\int_0^1 (\theta^{(\beta)}(t))^2 \, dt \leq \pi^{2\beta} Q^2.
\]

The Hölder space of order \( \beta \) on \([0, 1]\) consists of functions in \( C[0, 1] \) such that they have continuous derivatives up to order \( k = \lfloor \beta \rfloor \) and their \( k \)th derivative satisfies the Hölder condition with the exponent \( \alpha = \beta - k \). More precisely, the Hölder ball \( C^\beta(Q) \) of radius \( Q > 0 \) and the smoothness parameter \( \beta \) is defined as follows

\[
C^\beta(Q) = \{ \theta : \sup_{x, y \in [0, 1]} |\theta^{(k)}(x) - \theta^{(k)}(y)| \leq Q|x - y|^{\alpha}, \max_{j \leq k} \|\theta^{(j)}\|_\infty \leq Q \}.
\]

where \( \theta \in C[0, 1] \).
1.3. Frequentist asymptotics

In order to provide a better insight into typical nonparametric problems, we provide examples of the most basic nonparametric models: the Gaussian white noise model, nonparametric regression, and density function estimation. For a more detailed description and the discussion of the models see, for instance, Tsybakov [2009].

In the Gaussian white noise model the goal is to recover the function \( \theta_0 \in L^2[0,1] \) from an observed sample path \( X^{(n)} = \{X(t), t \in [0,1]\} \) satisfying the following stochastic differential equation

\[
dX(t) = \theta_0(t)dt + \frac{1}{\sqrt{n}}dW(t), \quad t \in [0,1],
\]

where \( W \) is the standard Brownian motion.

The regression model can be defined in various ways. In univariate nonparametric regression problem we observe \( n \) independent pairs of random variables \((X_i, Y_i)\) satisfying

\[ Y_i = \theta_0(X_i) + \xi_i, \]

where \( X_i \in [0,1], \xi_i \) given \( X_i \), say, are independent identically distributed random variables with mean zero, and \( \theta_0 \) is the function of interest. A particular case of this problem is regression with fixed design, where the \( X_i \) are deterministic points in \([0,1]\), for instance, the regular grid points \( i/n \). In the case of random \( X_i \) the problem is called regression with random design. It is commonly assumed for the true function to be smooth, for example, for it to belong to a Sobolev space.

In the density function estimation model we observe \( n \) identically distributed real-valued random variables with a common distribution that is absolutely continuous, for example, with respect to the Lebesgue measure on \([0,1]\). Let \( \rho_0 : \mathbb{R} \to [0,\infty) \) be the density function of this distribution. If we know a-priori that \( \rho_0 \) belongs to a parametrisable family with finite-dimensional parameter the problem becomes parametric. However, in many application cases there is no prior information on this function. Hence, it is common to have relatively weak restrictions, such as assuming that the density belongs some Hölder class, making the model nonparametric.

In the next section we introduce the asymptotic approach to comparing estimation procedures by looking at the maximum risks of the estimators. We also present minimax rates for the three models discussed above.

1.3 Frequentist asymptotics

One of the ways to assess the performances of the estimation methods is to take an asymptotic approach. Any statistical procedure can be indexed by the size of the sample used to calculate it, therefore producing a sequence of estimators. Properties of this sequence describe the behaviour of the estimation procedure when the sample size increases. An intuitively reasonable requirement for an estimation procedure is (asymptotic) consistency, which asks the outcome of the procedure with unlimited data to identify the underlying truth.
1. Bayesian nonparametric statistics

In order to define it mathematically we consider an observation \( X = X^{(n)} \) indexed by \( n \), where \( n \) is for instance the number of observations, or, as in the Gaussian white noise model, the signal to noise ratio. Assume that there exists a true distribution \( P_{\theta_0} \) according to which the data is generated. Let \( d \) be a semi-distance on the parameter space \( \Theta \). A sequence of estimators \( \hat{\theta}_n \) is called (asymptotically) consistent if

\[
d(\hat{\theta}_n, \theta_0) \underset{P_{\theta_0}}{\longrightarrow} 0, \text{ as } n \to \infty,
\]

where the convergence is in probability with respect to \( P_{\theta_0} \). We can measure the performance of an estimator \( \hat{\theta}_n \) of \( \theta \) using the maximum risk of this estimator on the set \( \Theta \), which is defined as follows

\[
r(\hat{\theta}_n) = \sup_{\theta \in \Theta} E_{\theta} l(\hat{\theta}_n, \theta),
\]

where \( E_{\theta} \) is the expectation with respect to the probability measure \( P_{\theta} \) and \( l : \Theta \times \Theta \to \mathbb{R} \) is a loss function. This function quantifies the amount by which the estimator deviates from the value of the true parameter. The choice of a meaningful loss function depends on the studied model and is far from obvious. We consider the commonly used loss function \( l = d^2 \).

The minimax risk associated with the statistical model \( \{P_{\theta}, \theta \in \Theta\} \) is given by

\[
R_n = \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} E_{\theta} l(\hat{\theta}_n, \theta),
\]

where the infimum is taken over all possible estimators. For a wide range of problems it is possible to establish an asymptotic lower and an upper bound on the minimax risk. An estimator attaining the lower bound up to a constant is called a rate optimal estimator. In problems where a smooth function is being estimated the minimax rate is influenced by the regularity of the true function. Pinsker’s theorem (see for instance Theorem 3.1 in Tsybakov [2009]) asserts that the minimax rate for the Gaussian white noise model is equal to \( n^{-2\beta/(2\beta+1)} \) with respect to the squared \( L^2 \) loss function, where \( \beta \) is the Sobolev smoothness of the target function.

**Theorem 1.3.1** (Pinsker’s theorem). Let \( \beta, Q > 0 \). Then for the Gaussian white noise model the minimax risk satisfies

\[
\lim_{n \to \infty} \inf_{\hat{\theta}_n} \sup_{\theta \in H^\beta(Q)} E_{\theta} n^{2\beta/(2\beta+1)} \|\hat{\theta}_n - \theta\|_2^2 = C^*,
\]

where the infimum is taken over all estimators and

\[
C^* = Q^{2/(2\beta+1)}(2\beta + 1)^{1/(2\beta+1)} \left( \frac{\beta}{\pi(\beta + 1)} \right)^{2\beta/(2\beta+1)}.
\]

Observe that in this case it is possible to determine the exact minimax rate including the constant in front of the exponent. Also, note that the theorem provides a lower and an upper bound on the risk. The latter is done by devising a projection estimator (called Pinsker’s estimator) that attains the minimax rate. To construct
the estimator we transfer the signal into a sequence of Fourier coefficients by taking the inner product of the sample path $X$ with the elements $\psi_j$ of the orthonormal basis of $L^2$. Given the model (1.1) the following infinite sequence of Gaussian observations is available to the statistician

$$X_j = \int_0^1 \psi_j(t) dX(t) = \theta_j + \frac{1}{\sqrt{n}} \xi_j,$$

where $\xi_j$ are independent standard Gaussian and the $\theta_j$ are the Fourier coefficients of $\theta_0$ with respect to the basis $\psi_j$. Then we can estimate the Fourier coefficients of the true parameter with the first $n^{1/(2\beta+1)}$ coefficients of the observed signal.

From Brown and Low [1996] and Nussbaum [1996] we know that both the nonparametric regression model and the density function estimation model are asymptotically equivalent to the Gaussian white noise model. In fact, for the regression or density function $\theta_0$ in the Sobolev class with a smoothness parameter $\beta > 0$ the minimax rate in those models is equal to $n^{-\beta/(2\beta+1)}$ up to a constant with respect to the $L^2$-norm (for more details see, for instance, Tsybakov [2009] and the references therein).

### 1.4 Bayesian approach

Contrary to the frequentist perspective, the Bayesian approach to statistics is based on the belief that there is no true fixed underlying parameter, but the parameter $\theta$ is random itself. It uses probability distribution on the parameter set to represent the belief of the statistician about the structure of the data. Consider a prior distribution $\Pi$ to be a probability measure on the parameter space $\Theta$. Then in the Bayesian setting the measure $P_\theta$ describes the conditional distribution of $X$ given the parameter value $\theta$. The prior can be interpreted as the degree of belief attached to subsets of the model before any observations has been made. Central in the Bayesian framework is the conditional distribution of $\theta$ given $X$. It is called the posterior distribution and can be viewed as a data–amended version of the prior after the observations are incorporated in the estimation procedure. If every distribution $P_\theta$ in the model $P$ admits a density $p_\theta$ with respect to some dominating measure, the posterior distribution can be described using Bayes formula

$$\Pi(A \mid X) = \frac{\int_A p_\theta(X)d\Pi(\theta)}{\int_\Theta p_\theta(X)d\Pi(\theta)}$$

for any measurable set $A \subset \Theta$. However, in many applications the straightforward computation of the posterior is not possible. A major reason for the popularity of Bayesian methods is the availability of sampling algorithms such as Markov chain Monte Carlo (MCMC) methods. These methods are designed for sampling from a probability distribution $P$ without having an explicit formula for it. They are based on the construction of Markov chains with the equilibrium distribution $P$. 
They help to draw from a posterior in the settings when it cannot be computed explicitly. More details on MCMC methods can be found, for instance, in the books Gilks et al. [1995] and Brooks et al. [2011].

Apart from philosophical reasons, Bayesian techniques are commonly used in practical applications due to their conceptual simplicity. Additionally, an attractive feature of such techniques for applications is the ability to incorporate knowledge about the parameter into a prior distribution. Moreover, Bayesian procedures can be considered as a form of regularisation which is performed to avoid overfitting. Specifically, a Bayesian method can give higher probabilities to the functions that are considered to be more likely, for example because of their smoothness properties. Finally, Bayesian procedures can be appealing from point of view of decision theory because of the complete class theorems. These theorems state that under mild conditions for any procedure there is a better (or at least not worse) Bayes procedure, and only those procedures are admissible (for more details see Ferguson [1967]). For a broader introduction to Bayesian methods see e.g. the books of Berger [1993], Ghosh and Ramamoorthi [2003], or Ghosal and van der Vaart [2017].

1.5 Bayesian asymptotics

In view of the ongoing debate between frequentist and Bayesian statisticians, it is of great interest to study the theoretical performance of Bayesian methods from the frequentist perspective. To assess the performance of Bayesian procedure we again take an asymptotic approach. A posterior distribution is called asymptotically consistent if most of its mass is concentrated in an arbitrary small neighbourhood $U$ of the truth for large $n$, i.e.

$$\Pi_n(U \mid X^{(n)}) \xrightarrow{P_{\theta_0}} 1,$$

where the convergence is in probability.

A classical result about posterior consistency is the Doob’s theorem (see Doob [1949]) stating that in the i.i.d. framework a Bayesian procedure with a prior $\Pi$ is $\Pi$–almost surely consistent. However, in practical applications it is not known whether the specific true parameter $\theta_0$ belongs to the prior–null set for a certain prior. A discussion and examples of inconsistency of posteriors can be found, for example, in Freedman [1963, 1965] and Diaconis and Freedman [1986a,b]. Among other things those results show that the null-set of inconsistency can be quite big.

Therefore, it is important to get sufficient conditions for consistency for a given parameter $\theta$. Such a result is presented in Schwartz [1965]. The paper studies the case of statistically separable models for which there exists an exponentially powerful test for $\theta = \theta_0$ against the hypothesis $\theta \not\in U$ for every neighbourhood $U$ of the true parameter $\theta_0$. The result shows that if in that case the prior gives positive mass to any small Kullback–Leibler neighbourhood of the true parameter, then the posterior distribution is consistent for this true parameter.

A more descriptive property of the posterior distribution is the rate of contraction
around the true parameter. The Bayesian procedure $\Pi_n$ has a contraction rate $\varepsilon_n$ if for $M > 0$ large enough

$$\Pi_n(\theta : d(\theta, \theta_0) \leq M\varepsilon_n | X^{(n)}) \xrightarrow{P_{\theta_0}} 1.$$ 

The contraction rate quantifies how quickly the mass of the posterior distribution concentrates around the true parameter $\theta_0$. Results on the rate of convergence for posterior mean in parametric models can be found for example in Le Cam [1973] and Ibragimov and Has’minskii [1981]. They show that under some regularity conditions, Bayesian procedures achieve the optimal rate of convergence $1/\sqrt{n}$ in that case.

However, matters are more complicated for nonparametric models. Recent results on the convergence rates of Bayesian procedures include Ghosal et al. [2000], Ghosal and van der Vaart [2007], and Shen and Wasserman [2001]. We discuss two theorems from Ghosal et al. [2000] and Ghosal and van der Vaart [2007] that allow to determine the rates of convergence based on information about the concentration of the prior and the complexity of the parameter space.

First, consider the case of independent and identically distributed observations. Let $X_1, \ldots, X_n$ be distributed according to some distribution $P_{\theta_0}$ with density $p_{\theta_0}$ with respect to some measure $\mu$ on the space $\mathcal{X}$. Let $\Pi_n$ be a sequence of prior probability measures on the set $\Theta$. Consider the Hellinger metric $d$ on $\Theta$ given by

$$d^2(\theta, \theta') = \int (\sqrt{p_\theta} - \sqrt{p_{\theta'}})^2 d\mu.$$ 

For two densities $f, g : \mathcal{X} \to [0, \infty)$ on a measurable space $(\mathcal{X}, \mu)$ define the Kullback–Leibler divergence to be

$$K(f, g) = \int f \log(f/g) d\mu.$$ 

Additionally, let

$$V_k(f, g) = \int f |\log(f/g)|^k d\mu, \quad V_{k,0}(f, g) = \int f |\log(f/g) - K(f, g)|^k d\mu.$$ 

Consider the following neighbourhoods of the true parameter

$$B(\theta_0, \varepsilon) = \{ \theta \in \Theta : K(p_{\theta_0}, p_\theta) \leq \varepsilon^2, V_2(p_{\theta_0}, p_\theta) \leq \varepsilon^2 \}.$$ 

For a set $\Theta_0 \subset \Theta$ define $N(\varepsilon, \Theta_0, d)$ to be the covering number, which is the minimal number of balls of radius $\varepsilon$ required to cover the set $\Theta_0$ with respect to the metric $d$. By Ghosal et al. [2000] the following theorem holds.

**Theorem 1.5.1.** Suppose for a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$, a constant
1. Bayesian nonparametric statistics

\( C > 0 \), and sets \( \Theta_n \subset \Theta \) the following conditions hold

\[
\log N(\varepsilon_n, \Theta_n, d) \leq n\varepsilon_n^2; \tag{1.2}
\]

\[
\Pi_n(\Theta \setminus \Theta_n) \leq e^{-n\varepsilon_n^2(C+4)}; \tag{1.3}
\]

\[
\Pi_n(B(\theta_0, \varepsilon_n)) \geq e^{-Cn\varepsilon_n^2}. \tag{1.4}
\]

Then \( \Pi_n(\theta : d(\theta, \theta_0) \geq M\varepsilon_n \mid X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} 0 \) for some \( M > 0 \), as \( n \to \infty \).

The remaining mass condition (1.3) can be understood as expressing that \( \Theta_n \) is almost as big as the support of the prior. The first and the third conditions of the theorem are the essential ones. Condition (1.2) ensures that the submodel \( \Theta_n \) is not too complex. The prior mass condition (1.4) requires the prior distribution to put a sufficient amount of mass in a small neighbourhood of the true parameter.

One of the generalisations of this theorem is studied in Ghosal and van der Vaart [2007], where the result is stated for the case of data being sampled independently, without the additional assumption of being identically distributed. Consider the observation vector \( X = (X_1, \ldots, X_n) \) of independent observations \( X_i \). In this case we take the measures \( P_{\theta} \) to be equal to the product measures \( \otimes_{i=1}^{n} P_{\theta,i} \) on a product space \( \otimes_{i=1}^{n} (\mathcal{X}_i, A_i) \). We assume that the distribution \( P_{\theta,i} \) of the \( i \)th component \( X_i \) has a density \( p_{\theta,i} \) relative to a \( \sigma \)-finite measure \( \mu_i \) on \( (\mathcal{X}_i, A_i) \).

Consider the average Hellinger distance \( d_n \) given by

\[
d_n^2(\theta, \theta') = \frac{1}{n} \sum_{i=1}^{n} \int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 d\mu_i.
\]

Additionally, consider a neighbourhood of the true parameter \( \theta_0 \) defined as follows

\[
B_n(\theta_0, \varepsilon, k) = \left\{ \theta : \frac{1}{n} \sum_{i=1}^{n} K(p_{\theta_0,i}, p_{\theta,i}) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{\theta_0,i}, p_{\theta,i}) \leq C_k \varepsilon^k \right\}.
\]

The following theorem of Ghosal and van der Vaart [2007] is an extension of Theorem 1.5.1.

**Theorem 1.5.2.** Let \( P_{\theta} \) be the product measures and \( d_n \) be the semi-metric defined above. Suppose that for a sequence \( \varepsilon_n \to 0 \) such that \( n\varepsilon_n^2 \to \infty \), some \( k > 1 \), all sufficiently large \( j \) and sets \( \Theta_n \subset \Theta \), the following conditions hold

\[
\log N(\varepsilon_n/36, \Theta_n, d_n) \leq n\varepsilon_n^2, \tag{1.5}
\]

\[
\frac{\Pi(\Theta \setminus \Theta_n)}{\Pi(B_n(\theta_0, \varepsilon_n, k))} = o\left(e^{-2n\varepsilon_n^2}\right), \tag{1.6}
\]

\[
\frac{\Pi(\theta \in \Theta_n : j\varepsilon_n < d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi(B_n(\theta_0, \varepsilon_n, k))} \leq e^{n\varepsilon_n^2j^2/4}. \tag{1.7}
\]

Then \( \Pi(\theta : d_n(\theta, \theta_0) \geq M\varepsilon_n \mid X^{(n)}) \xrightarrow{P_{\theta_0}} 0 \) for some \( M > 0 \), as \( n \to \infty \).
1.6. Adaptation

It is known that on the basis of the posterior distribution one can construct a frequentist point estimator that has a convergence rate at least of the same order as the posterior contraction rate. For example, according to Belitser and Ghosal [2003] an estimator equal to the centre of the smallest ball with posterior mass at least $3/4$ satisfies the requirement. Moreover, for a bounded metric $d$ with a convex square the posterior mean typically attains the convergence rate as well (see Ghosal et al. [2000]). Hence, using Theorems 1.5.1, 1.5.2 one can establish whether the posterior distribution achieves the optimal contraction rate around the true parameter. Results on this topic can be found, for instance, in Ghosal and van der Vaart [2017].

1.6 Adaptation

In Section 1.3 we encountered a minimax estimator that is constructed based on the knowledge of the regularity level of the target function. Such estimators are not particularly useful for practical applications, since often the regularity of the true function is not known in advance. Hence, it is desirable to develop procedures that are more flexible and that can attain the minimax rates across a wide range of regularity of parameters.

As we have seen, nonparametric classes are commonly characterised by a few hyperparameters that quantify different properties of the underlying function, such as its level of smoothness. In Theorem 1.3.1, as in many other cases, the minimax estimator relies on the knowledge of the value of such hyperparameters. For practical applications one would like to have an adaptive estimator that can attain minimax rates for a broad collection of the values of the hyperparameters. Such procedures have been thoroughly studied in the frequentist setting, see for example Bickel [1982], Efromovich and Pinsker [1996], Lepski and Spokoiny [1997].

The development of Bayesian procedures that are adaptive has only started relatively recently. To illustrate one of the ideas behind the construction of adaptive Bayesian estimators we present a random scaled squared exponential process prior studied, for example, in van der Vaart and van Zanten [2009]. The procedure employs a scale of priors indexed by a bandwidth parameter and adapts by making a data-dependent choice of the bandwidth. The squared exponential process is the centred Gaussian process $W = \{W_t, t \in \mathbb{R}\}$ with covariance function

$$E W_s W_t = e^{-|t-s|^2}.$$

This process is known to have a version with infinitely smooth sample paths. As a prior distribution for a function $\theta \in C[0, 1]$ we consider the law of the process

$$\{W_{At}, t \in [0, 1]\},$$

where $A$ is an independent random variable with a Gamma distribution. The inverse $1/A$ of the variable $A$ can be viewed as a bandwidth or length scale parameter.
1. Bayesian nonparametric statistics

The random variable $A$ plays the role of a scaling parameter that makes the prior suitable for less regular functions, when $A$ is large enough. The paper of van der Vaart and van Zanten [2009] shows that this Bayesian procedure attains optimal convergence rates (up to a log factor) in various settings for the parameter class $C^\beta(Q)$ with $Q, \beta > 0$. Observe that the prior itself does not depend on $\beta$ which makes the procedure fully rate-adaptive.

In general, adaptive Bayesian methods automatically choose the setting of hyperparameters by using either hierarchical or empirical Bayesian techniques. In this thesis we focus on hierarchical Bayesian methods that endow the hyperparameters with hyperpriors, making the Bayesian procedure multilevel. Examples of such procedures can be found in Belitser and Ghosal [2003], where an adaptive method is developed in the context of Gaussian white noise model. de Jonge and van Zanten [2010, 2012] studies adaptive procedure for the nonparametric regression problem. Full Bayesian methods in the context of density estimation are discussed, for example, in Ghosal et al. [2008], Lember and van der Vaart [2007], Kruijer et al. [2010], Rousseau [2010], and van der Vaart and van Zanten [2008a].

In this thesis we devise adaptive Bayesian procedures for two different nonparametric settings: function estimation on graphs and Poisson intensity estimation. We present hierarchical Bayesian procedures and study their asymptotic performance, showing that they attain (almost) optimal rates of convergence.
Chapter 2

The problem of function estimation on large graphs

2.1 Introduction

2.1.1 Learning a smooth function on a large graph

There are various problems arising in modern statistics that involve making inference about a “smooth” function on a large graph. The underlying graph structure in such problems can have different origins. Sometimes it is given by the context of the problem. This is typically the case, for instance, in the problem of making inference on protein interaction networks (e.g. Sharan et al. [2007]) or in image interpolation problems (Liu et al. [2014]). In other cases the graph is deduced from the data in a preliminary step, as is the case with similarity graphs in label propagation methods (e.g. Zhu and Ghahramani [2002]). Moreover, the different problems that arise in applications can have all kinds of different particular features. For example, the available data can be indexed by the vertices or by the edges of the graph, or both. Also, in some applications only partial data are available, for instance only part of the vertices are labeled (semi-supervised problems). Moreover, both regression and classification problems arise naturally in different applications.

Despite all these different aspects, many of these problems and the methods that have been developed to deal with them have a number of important features in common. In many cases the graph is relatively “large” and the function of interest can be viewed as “smoothly varying” over the graph. Consequently, most of the proposed methods view the problem as a high-dimensional or nonparametric estimation problem and employ some regularisation or penalisation technique that takes the geometry of the graph into account and that is thought to produce an appropriate bias-variance trade-off.

In this chapter we set up the mathematical framework that allows us to study the performance of such nonparametric function estimation methods on large graphs.
2. The problem of function estimation on large graphs

We do not treat all the variants exhaustively, instead we consider two prototypical problems: regression, where the function of interest $f$ is a function on the vertices of the graph that is observed with additive noise, and binary classification, where a label 0 or 1 is observed at each vertex and the object of interest is the “soft label” function $f$ whose value at a vertex $v$ is the probability of seeing a 1 at $v$. We assume the underlying graph is “large”, in the sense that it has $n$ vertices for some “large” $n$.

Despite the finite structure, it is intuitively clear that the “smoothness” of $f$, defined in a suitable manner, will have an impact on the difficulty of the problem and on the results that can be attained. Indeed, consider the extreme case of $f$ being a constant function. Then estimating $f$ reduces to estimating a single real number. In the regression setting, for instance, this means that under mild conditions the sample mean gives a $\sqrt{n}$-consistent estimator. In the other extreme case of a completely unrestricted function there is no way of making any useful inference. At best we can say that in view of the James-Stein effect we should employ some degree of shrinking or regularisation. However, if no further assumptions are made, nothing can be said about consistency or rates. We are interested in the question what we should do in the intermediate situation that $f$ has some “smoothness” between these two extremes.

Another aspect that will have a crucial impact on the problem, in addition to the regularity of $f$, is the geometry of the graph. Indeed, regular grids of different dimensions are special cases of the graphs we shall consider, and we know from existing theory that the best attainable rate for estimating a smooth function on a grid depends on the dimension of the grid. More generally, the geometry of the graph will influence the complexity of the spaces of “smooth” functions on the graph, and hence the performance of statistical or learning methods.

2.1.2 Asymptotic behaviour of estimators

To assess the performance of procedures we take an asymptotic perspective. We let the number of vertices of the graph grow and ask how fast an estimation procedure can "learn" the underlying function of interest. We derive minimax rates for regression and binary classification on the graph by providing a lower bound on them and presenting an estimator that achieves that rate. In order to do that we make two kinds of assumptions. Firstly, we assume that $f$ has a certain degree of regularity $\beta$, defined in suitable manner. The smoothness $\beta$ is not assumed to be known though, we are aiming at deriving adaptive results.

Secondly, we make an assumption on the asymptotic shape of the graph. In recent years, various theories of graph limits have been developed. Most prominent is the concept of the graphon, e.g. Lovász and Szegedy [2006] or the book of Lovász [2012]. More recently this notion has been extended in various directions, see, for instance, Borgs et al. [2014] and Chung [2014]. However, the existing approaches are not immediately suited in the situations we have in mind, which involve graphs that are sparse in nature and are “grid-like” in some sense. Therefore, we take an
2.2. Asymptotic geometry assumption on graphs

alternative approach and describe the asymptotic shape of the graph through a condition on the asymptotic behaviour of the spectrum of the Laplacian. To be able to derive concrete results we essentially assume that the smallest eigenvalues $\lambda_{i,n}$ of $L$ satisfy

$$\lambda_{i,n}^2 \asymp \left( \frac{i}{n} \right)^{2/r}$$

(2.1)

for some $r \geq 1$. Here we write $a_n \asymp b_n$ if $0 < \liminf a_n / b_n \leq \limsup a_n / b_n < \infty$. Very roughly speaking, this means that asymptotically, or “from a distance”, the graph looks like an $r$-dimensional grid with $n$ vertices. As we shall see, the actual grids are special cases (see Example 2.2.1), hence our results include the usual statements for regression and classification on these classical design spaces. However, the setting is much more general, since it is really only the asymptotic shape that matters. For instance, a 2 by $n/2$ ladder graph asymptotically also looks like path graph, and indeed we will see that it satisfies our assumption for $r = 1$ as well (Example 2.2.3). Moreover, the constant $r$ in (2.1) does not need to be a natural number. We will see, for example, at least numerically, that there are graphs whose geometry is asymptotically like that of a grid of non-integer “dimension” $r$ in the sense of condition (2.1).

We stress that we do not assume the existence of a “limiting manifold” for the graph as $n \to \infty$. We formulate our conditions and results purely in terms of intrinsic properties of the graph, without first embedding it in an ambient space. In certain cases in which limiting manifolds do exist (e.g. the regular grid cases) our type of asymptotics can be seen as “infill asymptotics” (Cressie [1993]). For a simple illustration, see Example 2.3.1. However, in applied settings (see, for instance, Example 2.2.7) it is typically not clear what a suitable ambient manifold could be, which is why we choose to avoid this issue altogether.

2.1.3 Organisation

The remainder of the chapter is organised as follows. In the next section we present our geometry assumption and give examples of graphs that satisfy it, either theoretically or numerically. In Section 2.3 we introduce Sobolev-type balls that are used to quantify the regularity of the function. In Section 2.4 we obtain minimax rates for regression and binary classification problems on large graphs. The mathematical proofs are given in Section 2.5.1 and 2.5.2.

2.2 Asymptotic geometry assumption on graphs

In this section we formulate our geometry assumption on the underlying graph and give several examples.
2. The problem of function estimation on large graphs

2.2.1 Graphs, Laplacians and functions on graphs

Let $G$ be a connected simple (i.e. no loops, multiple edges or weights), undirected graph with $n$ vertices labelled $1, \ldots, n$. Let $A$ be its adjacency matrix, i.e. $A_{ij}$ is 1 or 0 according to whether or not there is an edge between vertices $i$ and $j$. Let $D$ be the diagonal matrix with element $D_{ii}$ equal to the degree of vertex $i$. Let $L = D - A$ be the Laplacian of the graph. We note that strictly speaking, we will be considering sequences of graphs $G_n$ with Laplacians $L_n$ and we will let $n$ tend to infinity. However, in order to avoid cluttered notation, we will omit the subscript $n$ and just write $G$ and $L$ throughout.

A function $f$ on the (vertices of the) graph is simply a function $f : \{1, \ldots, n\} \rightarrow \mathbb{R}$. Slightly abusing notation we will write $f$ both for the function and for the associated vector of function values $(f(1), f(2), \ldots, f(n))$ in $\mathbb{R}^n$. We measure distances and norms of functions using the norm $\| \cdot \|_n$ defined by $\|f\|_n^2 = n^{-1} \sum_{i=1}^n f^2(i)$. The corresponding inner product of two functions $f$ and $g$ is denoted by

$$\langle f, g \rangle_n = \frac{1}{n} \sum_{i=1}^n f(i)g(i).$$

Again, in our results $n$ will be varying, so when we speak of a function $f$ on the graph $G$ we are, strictly speaking, considering a sequence of functions $f_n$. Also, in this case the subscript $n$ will usually be omitted.

The Laplacian $L$ is positive semi-definite and symmetric. It easily follows from the definition that its smallest eigenvalue is 0 (with eigenvector $(1, \ldots, 1)$). The fact that $G$ is connected implies that the second smallest eigenvalue, the so-called algebraic connectivity, is strictly positive (e.g. Cvetković et al. [2010]). We will denote the Laplacian eigenvalues, ordered by magnitude, by

$$0 = \lambda_{n,0} < \lambda_{n,1} \leq \lambda_{n,2} \leq \cdots \leq \lambda_{n,n-1}.$$

Again we will usually drop the first index $n$ and just write $\lambda_i$ for $\lambda_{n,i}$. We fix a corresponding sequence of eigenfunctions $\psi_i$, orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_n$.

2.2.2 Asymptotic geometry assumption

As mentioned in the introduction, we derive results under an asymptotic shape assumption on the graph, formulated in terms of the Laplacian eigenvalues. To motivate the definition we note that the $i$th eigenvalue of the Laplacian of an $n$-point grid of dimension $d$ behaves like $(i/n)^{2/d}$ (see Example 2.2.1 ahead). We will work with the following condition.

**Condition.** We say that the geometry condition is satisfied with parameter $r \geq 1$
2.2. Asymptotic geometry assumption on graphs

if there exist \( i_0 \in \mathbb{N}, \kappa \in (0,1] \) and \( C_1, C_2 > 0 \) such that for all \( n \) large enough,

\[
C_1 \left( \frac{i}{n} \right)^{2/r} \leq \lambda_i \leq C_2 \left( \frac{i}{n} \right)^{2/r}, \quad \text{for all } i \in \{i_0, \ldots, \kappa n\}. \tag{2.2}
\]

Note that this condition only restricts a positive fraction \( \kappa \) of the Laplacian eigenvalues, namely the \( \kappa n \) smallest ones. Moreover, we do not need restrictions on the first finitely many eigenvalues. We remark that if the geometry condition is fulfilled, then by adapting the constant \( C_1 \) we can ensure that the lower bound holds, in fact, for all \( i \in \{i_0, \ldots, n\} \). To see this, observe that for \( n \) large enough and \( \kappa n < i \leq n \)

\[
\lambda_i \geq \lambda_{\lfloor \kappa n \rfloor} \geq C_1 \left( \frac{\kappa n}{n} \right)^{2/r} \geq C_1 \left( \frac{\kappa}{2} \right)^{2/r} \left( \frac{i}{n} \right)^{2/r}.
\]

For the indices \( i < i_0 \) it is useful to note that we have a general lower bound on the first positive eigenvalue \( \lambda_1 \), hence on \( \lambda_2, \ldots, \lambda_{i_0} \) as well. Indeed, by Theorem 4.2 of Mohar [1991a] we have

\[
\lambda_1 \geq \frac{4}{\text{diam}(G)n} \geq \frac{4}{n^2}.	ag{2.3}
\]

Note that this bound also implies that our geometry assumption can not hold with a parameter \( r < 1 \), since that would lead to contradictory inequalities for \( \lambda_{i_0} \).

We first confirm that the geometry condition is satisfied for grids and tori of different dimensions.

**Example 2.2.1 (Grids).** For \( d \in \mathbb{N} \), a regular \( d \)-dimensional grid with \( n \) vertices can be obtained by taking the Cartesian product of \( d \) path graphs with \( n^{1/d} \) vertices (provided, of course, that this number is an integer). Using the known expression for the Laplacian eigenvalues of the path graph and the fact that the eigenvalues of products of graphs are the sums of the original eigenvalues, see, for instance, Theorem 3.5 of Mohar [1991b], we get that the Laplacian eigenvalues of the \( d \)-dimensional grid are given by

\[
4 \left( \sin^2 \frac{\pi i_1}{2n^{1/d}} + \cdots + \sin^2 \frac{\pi i_d}{2n^{1/d}} \right) \leq \frac{i_1^2 + \cdots + i_d^2}{n^{2/d}},
\]

where \( i_j = 0, \ldots, n^{1/d} - 1 \) for \( j = 1, \ldots, d \). By definition there are \( i + 1 \) eigenvalues less or equal than the \( i \)th smallest eigenvalue \( \lambda_i \). Hence, for a constant \( c > 0 \), we have

\[
i + 1 = \sum_{i_1^2 + \cdots + i_d^2 \leq c^2 n^{2/d} \lambda_i} 1.
\]

The sum on the right gives the number of lattice points in a sphere of radius \( R = cn^{1/d} \sqrt{\lambda_i} \) in \( \mathbb{R}^d \). For our purposes it suffices to use crude upper and lower bounds for this number. By considering, for instance, the smallest hypercube containing the sphere and the largest one inscribed in it, it is easily seen that the number of lattice points is bounded from above and below by a constant times
2. The problem of function estimation on large graphs

\[ R^d. \] We conclude that for the \( d \)-dimensional grid we have \( \lambda_i \approx (i/n)^{2/d} \) for every \( i = 0, \ldots, n - 1 \). In particular, the geometry condition is fulfilled with parameter \( r = d \).

**Example 2.2.2** (Discrete tori). For graph tori we can follow the same line of reasoning as for grids. A \( d \)-dimensional torus graph with \( n \) vertices can be obtained as a product of \( d \) ring graphs with \( n^{1/d} \) vertices. Using the known explicit expression of the Laplacian eigenvalues of the ring we find that the \( d \)-dimensional torus graph satisfies the geometry condition with parameter \( r = d \) as well.

The following lemma lists a number of operations that can be carried out on the graph without losing the geometry condition.

**Lemma 2.2.1.** Suppose that \( G = G_n \) satisfies the geometry assumption with parameter \( r \). Then the following graphs satisfy the condition with parameter \( r \) as well:

(i) The cartesian product of \( G \) with a connected simple graph \( H \) with a finite number of vertices (independent of \( n \)).

(ii) The graph obtained by augmenting \( G \) with finitely many edges (independent of \( n \)), provided it is a simple graph.

(iii) The graph obtained from \( G \) by deleting finitely many edges (independent of \( n \)), provided it is still connected.

(iv) The graph obtained by augmenting \( G \) with finitely many vertices and edges (independent of \( n \)), provided it is a simple connected graph.

**Proof.** (i). Say \( H \) has \( m \) vertices and let its Laplacian eigenvalues be denoted by \( 0 = \mu_0, \ldots, \mu_m \). Then the product graph has \( mn \) vertices and it has Laplacian eigenvalues \( \lambda_i + \mu_j, i = 0, \ldots, n - 1, j = 0, \ldots, m - 1 \) (see Theorem 3.5 of Mohar [1991b]). In particular, the first \( n \) eigenvalues are the same as those of \( G \). Hence, since \( G \) satisfies the geometry condition, so does the product of \( G \) and \( H \).

(ii) and (iii). These statements follow from the interlacing formula that asserts that if \( G + e \) is the graph obtained by adding the edge \( e \) to \( G \), then

\[
0 \leq \lambda_1(G) \leq \lambda_1(G + e) \leq \lambda_2(G) \leq \cdots \leq \lambda_{n-1}(G) \leq \lambda_{n-1}(G + e).
\]

See, for example, Theorem 3.2 of Mohar [1991b] or Theorem 7.1.5 of Cvetković et al. [2010].

(iv). Let \( v \) and \( e \) be a vertex and an edge that we want to connect to \( G \). Denote \( G_v \) a disjoint union of \( G \) and \( v \), and by \( G' \) the graph obtained by connecting edge \( e \) to \( v \) and an existing vertex of \( G \). By Theorem 3.1 from Mohar [1991b] we know that the eigenvalues of \( G_v \) are \( 0, 0, \lambda_1(G), \lambda_2(G), \ldots, \lambda_{n-1}(G) \). Using Theorem 3.2 of Mohar [1991b] we see that \( 0 = \lambda_0(G_v) = \lambda_0(G') \) and

\[
0 = \lambda_1(G_v) \leq \lambda_1(G') \leq \lambda_2(G_v) \leq \cdots \leq \lambda_{n-1}(G) \leq \lambda_n(G').
\]
The result follows from this observation. ■

Example 2.2.3 (Ladder graph). A ladder graph with \( n \) vertices is the product of a path graph with \( n/2 \) vertices and a path graph with 2 vertices. Hence, by part (i) of Lemma 2.2.1 and Example 2.2.1 it satisfies the geometry condition with parameter \( r = 1 \).

Example 2.2.4 (Lollipop graph). The so-called lollipop graph \( L_{m,n} \) is obtained by attaching a path graph with \( n \) vertices with an additional edge to a complete graph with \( m \) vertices. If \( m \) is constant, i.e. independent of \( n \), then according to parts (ii) and (iv) of the preceding lemma this graph satisfies the geometry condition with \( r = 1 \).

In the examples considered so far it is possible to verify theoretically that the geometry condition is fulfilled. In a concrete case in which the given graph is not of such a tractable type, numerical investigation of the Laplacian eigenvalues can give an indication as to whether or not the condition is reasonable and provide the appropriate value of the parameter \( r \). A possible approach is to plot \( \log \lambda_i \) against \( \log(i/n) \). If the geometry condition is satisfied with parameter \( r \), the \( \kappa \times 100\% \) left most points in this plot should approximately lie on a straight line with slope \( 2/r \), except possibly a few on the very left.

Our focus is not on numerics, but it is illustrative to consider a few numerical examples in order to get a better idea of the types of graphs that fit into our framework.

Example 2.2.5 (Two-dimensional grid, numerically). Figure 2.1 illustrates the suggested numerical approach for a two-dimensional, \( 20 \times 20 \) grid. The dashed line in the left panel is fitted to the left-most 35\% of the points in the plot, discarding the first three points on the left. In accordance with Example 2.2.1 this line has slope 1.0.

Example 2.2.6 (Watts-Strogatz ‘small world’ graph). In our second numerical example we consider a graph obtained as a realisation from the well-known random graph model of Watts and Strogatz [1998]. Specifically, we consider in the first step a ring graph with 200 vertices. In the next step every vertex is visited and the edges emanating from the vertex are rewired with probability \( p = 1/4 \), meaning that with probability 1/4 they are detached from the neighbour of the current vertex and attached to another vertex, chosen uniformly at random. In the right panel of Figure 2.2 a particular realisation is shown. Here we have only kept the largest connected component, which has 175 vertices in this case. On the left we have exactly the same plot as described in the preceding example for the grid case. The plot indicates that it is not unreasonable to assume that the geometry condition holds. The value of the parameter \( r \) deduced from the slope of the line equals 1.4 for this graph.

Example 2.2.7 (Protein interaction graph). In the final example we consider a graph obtained from the protein interaction graph of baker's yeast, as described in
2. The problem of function estimation on large graphs

Figure 2.1: Plot of $\log \lambda_i$ against $\log(i/n)$ for the $20 \times 20$ grid. Fitted line has slope 1.0, corresponding to $r = 2.0$ in the geometry assumption.

Figure 2.2: Plot of $\log \lambda_i$ against $\log(i/n)$ for the Watts-Strogatz graph in the right panel. Fitted line has slope 1.42, corresponding to $r = 1.4$ in the geometry assumption.
2.3 Smoothness assumption

Figure 2.3: Plot of $\log \lambda_i$ against $\log (i/n)$ for the protein interaction graph in the right panel. Fitted line has slope 0.94, corresponding to $r = 2.1$ in the geometry assumption.

detail in Section 8.5 of Kolaczyk [2009]. The graph, shown in the right panel of Figure 2.3, describes the interactions between proteins involved in the communication between a cell and its surroundings. Also for this graph it is true that with a few exceptions, the points corresponding to the 35% smallest eigenvalues lie approximately on a straight line. The same procedure as followed in the other examples gives a value $r = 2.1$ for the parameter in the geometry assumption.

2.3 Smoothness assumption

We define the “regularity” of the function using the graph Laplacian. Specifically, we will assume it belongs to a Sobolev-type ball of the form

$$H^\beta(Q) = \left\{ f : \left\langle f, (I + (n^{\frac{\beta}{2}}L)^\beta)f\right\rangle_n \leq Q^2 \right\}$$

for some $\beta, Q > 0$ (independent of $n$). The particular normalisation, which depends on the geometry parameter $r$, ensures non-trivial asymptotics.

It is illustrative to look at this assumption in a bit more detail in the simple case of the path graph.

Example 2.3.1 (Path graph). Consider a path graph $G$ with $n$ vertices, which we identify with the points $i/n$ in the unit interval, $i = 1, \ldots, n$. As seen in Example 2.2.1, this graph satisfies the geometry condition with parameter $r = 1$. Hence, in
this case the collection of functions $H^\beta(Q)$ is given by

$$H^\beta(Q) = \left\{ f : \langle f, (I + (n^2 L)^\beta)f \rangle_n \leq Q^2 \right\}.$$ 

To understand when a (sequence of) function(s) belongs to this space, say for $\beta = 1$, let $f_n$ be the restriction to the grid $\{i/n, i = 1, \ldots, n\}$ of a fixed function $f$ defined on the whole interval $[0, 1]$. The assumption that $f_n \in H^1(Q)$ then translates to the requirement that

$$\frac{1}{n} \sum_i f^2(i/n) + n \sum_{i \sim j} (f(i/n) - f(j/n))^2 \leq Q^2.$$

The first term on the left is a Riemann sum which approximates the integral $\int_0^1 f^2(x) \, dx$. If $f$ is differentiable, then for the second term we have, for large $n$,

$$n \sum_{i \sim j} (f(i/n) - f(j/n))^2 = n \sum_{i=1}^{n-1} (f((i+1)/n) - f(i/n))^2 \approx \frac{1}{n} \sum_i (f'(i/n))^2,$$

which is a Riemann sum that approximates the integral $\int_0^1 (f'(x))^2 \, dx$. Hence in this particular case the space of functions $H^1(Q)$ on the graph is the natural discrete version of the usual Sobolev ball

$$\left\{ f : [0, 1] \to \mathbb{R} : \int_0^1 (f^2(x) + f'^2(x)) \, dx \leq Q^2 \right\}.$$ 

Definition (2.4) is a way of describing “$\beta$-regular” functions on a general graph satisfying the geometry condition, without assuming the graph or the function on it are discretised versions of some “continuous limit”.

### 2.4 Minimax rates for function estimation problems on graphs

In this section we introduce two prototypical function estimation problems on the graphs: regression and binary classification problems. We assume that the underlying graph satisfies the geometry assumption with some $r \geq 1$. We show that the minimax rates over the balls $H^\beta(Q)$ for both problems depend on the “dimension” of the graph $r$ and the regularity $\beta$ of the target function. Specifically, we derive that the minimax rate behaves as $n^{-\beta/(2\beta + r)}$, when $n$ is large enough. Naturally the complexity of the problem increases with the growth of the dimension $r$. There is also the usual dependence on the regularity of the function, meaning that it is easier to estimate a function which varies along the edges of the graph more slowly than a more rough one.

The minimax results we derive consist of two parts. We obtain a lower bound
for the target family of functions and we construct a specific estimator for each problem such that its maximum risk is within a constant factor of the derived lower bound. However, the developed estimators are not rate-adaptive and thus have small applied value. In Chapter 3 we present a Bayesian approach to those problems and introduce the associated priors that have an (almost) minimax rate of convergence of posterior distribution and which are rate-adaptive.

Let $G$ be a connected simple undirected graph with vertices $1, \ldots, n$, satisfying the geometry assumption for $r \geq 1$. In the regression case we assume we have an observation set $Y = (Y_1, \ldots, Y_n)$ on the vertices of the graph such that

$$Y_i = f_0(i) + \sigma \xi_i,$$

where the $\xi_i$ are independent standard Gaussian, $\sigma > 0$ and $f_0$ is the function of interest.

Under these conditions we derive the following result.

**Theorem 2.4.1.** Suppose the geometry assumption holds for $r \geq 1$. Then for every $\beta > 0$

$$\inf_{\hat{f}} \sup_{f \in H^\beta(Q)} \mathbb{E}_f \left( \|\hat{f} - f\|_n^2 \right) \asymp n^{-\beta/(2\beta + r)},$$

where the infimum is taken over all estimators $\hat{f}$.

The theorem shows that the minimax rate for the regression problem on the graph is equal to $n^{-\beta/(2\beta + r)}$. We obtain an upper bound on the rate by constructing a projection estimator $\tilde{f}$ for which

$$\sup_{f \in H^\beta(Q)} \mathbb{E}_f \left( \|\tilde{f} - f\|_n^2 \right) \lesssim n^{-2\beta/(2\beta + r)}.$$

However, the choice of our estimator depends on the regularity $\beta$ of the true function.

In the binary classification problem we assume that the data $Y_1, \ldots, Y_n$ are independent $\{0, 1\}$-valued variables, observed at the vertices of the graph. In this case the goal is to estimate the binary regression function $\rho_0$, or “soft label function” on the graph, given by

$$\rho_0(i) = \mathbb{P}_0(Y_i = 1).$$

We employ a suitably chosen link function $\Psi : \mathbb{R} \to (0, 1)$. We assume that $\Psi$ is a differentiable function onto $(0, 1)$ such that $\Psi'(\Psi(1 - \Psi))$ is uniformly bounded, and $\Psi'(x) > 0$ for every $x \in \mathbb{R}$. Note that for instance the sigmoid, or logistic link $\Psi(f) = 1/(1 + \exp(-f))$ satisfies this condition. Under our conditions the inverse $\Psi^{-1} : (0, 1) \to \mathbb{R}$ is well defined. In this classification setting the regularity condition will be formulated in terms of $\Psi^{-1}(\rho_0)$. 

31
Theorem 2.4.2. Suppose the geometry assumption holds for \( r \geq 1 \). Consider the orthonormal basis \( \psi_j \) of the graph Laplacian (normalised with respect to the \( \| \cdot \|_n \)–norm). Assume that there exists \( C > 0 \) such that for every \( n \in \mathbb{N} \) and for all \( i = 1, \ldots, n \) and \( j = 0, \ldots, n - 1 \) it holds that \( |\psi_j(i)| \leq C \). Let \( \Psi : \mathbb{R} \to (0,1) \) be a differentiable onto \((0,1)\) such that \( \frac{\Psi'(\Psi^{-1}(\rho))}{\Psi(1 - \Psi)} \) is uniformly bounded and \( \Psi'(x) > 0 \) for every \( x \in \mathbb{R} \). Then for \( \beta \geq r/2 \)

\[
\inf_{\hat{\rho}} \sup_{\rho \in \{ \rho : \Psi^{-1}(\rho) \in H^\beta(Q) \}} \mathbb{E}_\rho \| \hat{\rho} - \rho \|_n^2 \approx n^{-2\beta/(2\beta + r)},
\]

where the infimum is taken over all estimators \( \hat{\rho} \).

According to the theorem the minimax rate for the classification problem is again \( n^{-\beta/(2\beta + r)} \). To obtain an upper bound on the minimax risk in the classification setting we use an estimator based on the projection estimator from the proof of the upper bound on the risk in Theorem 2.4.1. Again, the developed estimator is not rate-adaptive, which makes it impractical. In the next chapter we address this issue by devising Bayesian estimators that achieve (almost) minimax rate and which are adaptive.

Observe that the result is obtained for functions with regularity levels \( \beta \geq r/2 \). Additionally, we put an upper bound on the supremum norm for the normalised eigenvectors of the graph Laplacian. These conditions arise from translation of the norms between the set \( H^\beta(Q) \) and the set of soft label functions \( \{ \rho : \Psi^{-1}(\rho) \in H^\beta(Q) \} \).

We note that for \( d \)-dimensional grids the supremum norm of the normalised eigenvectors is bounded. In practical applications one can verify numerically whether the condition is satisfied.

Example 2.4.3 (Grids). By Merris [1998] the eigenvectors of the graph Laplacian of the path graph are as follows

\[
\tilde{\psi}_j(i) = \cos \left( \pi ij/n - \pi j/2n \right).
\]

The \( \| \cdot \|_n \)–norm of the jth eigenvector is given by

\[
\| \tilde{\psi}_j \|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \cos^2 \left( \pi ij/n - \pi j/2n \right) =
\]

\[
= \frac{1}{2} + \frac{1}{4n \sin(\pi j/n)} \sum_{i=1}^{n} 2 \sin(\pi j/n) \cos((2i - 1)\pi j/n).
\]

Using basic trigonometric computation we have for any \( x \in \mathbb{R} \)

\[
\sum_{i=1}^{n} 2 \sin x \cos(2ix - x) = \sum_{i=1}^{n} (\sin 2ix - \sin(2ix - 2x)) = \sin 2nx.
\]
2.5. Proofs

Hence, for any $j = 1, \ldots, n - 1$

$$\|\tilde{\psi}_j\|_n^2 = \frac{1}{2} + \frac{1}{4n} \sin \frac{2\pi j}{n} = \frac{1}{2}.$$ 

Notice that $\|\tilde{\psi}_0\|_n^2 = 1$. Then for the normalised with respect to the $\|\cdot\|_n$-norm eigenvectors $\psi_j$ of the graph Laplacian of the path graph there exists a constant $C > 0$ such that $|\psi_i(j)| \leq C$ for all $n \in \mathbb{N}$, $i = 1, \ldots, n$ and $j = 0, \ldots, n - 1$.

From e.g. Merris [1998] we know that the eigenvectors of a Cartesian product of two graphs are equal to the Kronecker products of pairs of eigenvectors associated with the Laplacians of those graphs. Then the eigenvectors of the graph Laplacian for the 2–dimensional grid with $n^2$ vertices are given by $\psi_i \psi_j$, $i, j = 1, \ldots, n$. Hence, the supremum norm of the normalised eigenvectors is bounded. The grids of higher dimensions retain the property, since they are equal to Cartesian products of path graphs.

2.5 Proofs

2.5.1 Proof of Theorem 2.4.1

2.5.1.1 Preliminaries

Let $Y = (Y_1, \ldots, Y_n)$, $\xi = (\xi_1, \ldots, \xi_n)$, and let $\psi_i$ be the orthonormal eigenfunctions of the graph Laplacian. Denote $\tilde{\xi}_i = (\xi, \psi_i)_n$ and observe that the $\tilde{\xi}_i$ are centred Gaussian with

$$\mathbb{E}\tilde{\xi}_i \tilde{\xi}_j = \frac{1}{n} \delta_{ij}.$$ 

The inner products $Z_i = (Y, \psi_i)_n$ satisfy the following relation for $i = 0, \ldots, n - 1$

$$Z_i = (Y, \psi_i)_n = f_i + \sigma \xi_i,$$

where $f_i$ are coefficients in the decomposition of the target function $f_0 = \sum_{i=0}^{n-1} f_i \psi_i$.

Additionally, consider the decomposition of an estimator $\hat{f} = \sum_{i=0}^{n-1} \hat{f}_i \psi_i$. Then

$$\|\hat{f} - f_0\|_n^2 = \left\langle \sum_{i=0}^{n-1} (\hat{f}_i - f_i) \psi_i, \sum_{i=0}^{n-1} (\hat{f}_i - f_i) \psi_i \right\rangle_n = \sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2.$$ 

Hence, the minimax rates for the original problem are of the same order as the minimax rates for the problem of recovering $f = (f_0, \ldots, f_{n-1})$, given the observations

$$Z_i = f_i + \varepsilon \xi_i, \quad (2.7)$$

where $\xi_i$ are independent standard Gaussian and $\varepsilon = \frac{\sigma}{\sqrt{n}}$. To avoid confusion we define general ellipsoids on the space of coefficients for an arbitrary sequence $a_j > 0$
2. The problem of function estimation on large graphs

and some finite constant $Q > 0$

$$B_n(Q) = \{ f \in \mathbb{R}^n : \sum_{j=0}^{n-1} a_j^2 f_j^2 \leq Q \}. \quad (2.8)$$

For a function $f$ in the Sobolev-type ball $H^\beta(Q)$ its vector of coefficients belongs to $B_n(Q)$ with

$$a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r}, \quad j = 0, \ldots, n - 1.$$  

In order to prove the theorem it is sufficient to show that

$$\inf_f \sup_{f \in B_n(Q)} \mathbb{E}_f \left( \sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2 \right) \asymp n^{-2\beta/(2\beta + r)}. \quad (2.9)$$

We are going to follow the proof of Pinker’s theorem (see for example Tsybakov [2009]) which studies a similar case in the setting of the Gaussian white noise model on the interval $[0,1]$. The proof requires some modifications arising from the nature of our problem. The main differences from the Pinsker’s result are that we only have $n$ observations and that the definition of $B_n(Q)$ is slightly different than usual.

In order to proceed we first consider the problem of obtaining minimax rates in the class of linear estimators. We introduce Pinsker’s estimator and present the linear minimax lemma showing that Pinsker’s estimator is optimal in the class of linear estimators. The risk of a linear estimator $\tilde{f}(l) = (l_1 Z_1, \ldots, l_n Z_n)$ with $l = (l_1, \ldots, l_n) \in \mathbb{R}^n$ is given by

$$R(l, f) = \mathbb{E}_f \left( \sum_{j=0}^{n-1} (\hat{f}_j - f_j)^2 \right) = \sum_{j=0}^{n-1} ((1 - l_j)^2 f_j^2 + \varepsilon^2 l_j^2).$$

For large $n$ we introduce the following equation with respect to the variable $x$

$$\frac{\varepsilon^2}{x} \sum_{j=0}^{n-1} a_j (1 - xa_j)_+ = Q. \quad (2.10)$$

Suppose, there exists a unique solution $x$ of (2.10). For such a solution, define a vector of coefficients $l'$ consisting of entries

$$l'_j = (1 - xa_j)_+.$$

The linear estimator $\tilde{f} = \tilde{f}(l')$ is called the Pinsker estimator for the general ellipsoid $B_n(Q)$. The following lemma that appears as Lemma 3.2 in Tsybakov [2009], shows that the Pinsker estimator is a linear minimax estimator.

**Lemma 2.5.1** (Linear minimax lemma). Suppose that $B_n(Q)$ is a general ellipsoid...
2.5. Proofs

defined by (2.8) with \( Q > 0 \) and a positive set of coefficients \( \{a_j\}_{j=1}^n \). Suppose there exists a unique solution \( x \) of (2.10) and suppose that

\[
S = \varepsilon^2 \sum_{j=0}^{n-1} l'_j < \infty. \tag{2.11}
\]

Then the linear minimax risk satisfies

\[
\inf_{l \in \mathbb{R}^n} \sup_{f \in B_n(Q)} R(l, f) = \sup_{f \in B_n(Q)} R(l', f) = S. \tag{2.12}
\]

In order to follow the steps the proof of Pinsker’s Theorem we present a technical lemma that studies the asymptotic properties of the Pinsker’s estimator in our setting.

**Lemma 2.5.2.** Consider the ellipsoid \( B_n(Q) \) defined by (2.8) with \( Q > 0 \) and

\[
a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r}, \quad j = 0, \ldots, n - 1.
\]

Then, as \( n \to \infty \), we have the following

(i) There exists a solution \( x \) of (2.10) which is unique and satisfies

\[
x \asymp n^{-\beta/(2\beta+r)}.
\]

(ii) The weighted sum (2.11) of the coefficients of the Pinsker’s estimator satisfies

\[
S \asymp n^{-2\beta/(2\beta+r)}.
\]

(iii) For \( \varepsilon = \frac{s}{\sqrt{n}} \) define \( v_j = \frac{\varepsilon^2(1-xa_j)}{xa_j} \). Then

\[
\max_{j=0,\ldots,n-1} v_j^2 a_j^2 = O \left( n^{-r/(2\beta+r)} \right).
\]

**Proof.** (i) According to Lemma 3.1 from Tsybakov [2009] for large enough \( n \) and for an increasing positive sequence \( a_j, \ j = 0 < \ldots, n - 1 \) with \( a_n \to +\infty \), as \( n \to \infty \), there exists a unique solution of (2.10) given by

\[
x = \frac{\varepsilon^2 \sum_{j=0}^{N-1} a_j}{Q + \varepsilon^2 \sum_{j=0}^{N-1} a_j^2},
\]

where

\[
N = \max \left\{ m : \varepsilon^2 \sum_{j=0}^{m-1} a_j(a_m - a_j) < Q \right\} < +\infty.
\]
2. The problem of function estimation on large graphs

Consider \( N \) defined above. Denote

\[
A_m = \varepsilon^2 \sum_{j=0}^{m-1} a_j (a_m - a_j) = n^{\beta/r-1} \sigma^2 \sum_{j=0}^{m-1} \sqrt{\left(1 + \lambda_j^{2\beta/r} n^{2\beta/r}\right) \left(\lambda_m^{2\beta/r} n^{2\beta/r} - \lambda_j^{2\beta/r} n^{2\beta/r}\right)}.
\]

Since the geometry condition on the graph is satisfied for \( j = i_0, \ldots, \kappa n \) the eigenvalues of the graph Laplacian can be bounded in a following way

\[
C_1 \left(\frac{j}{n}\right)^{2/r} \leq \lambda_j \leq C_2 \left(\frac{j}{n}\right)^{2/r}.
\]

Then for some \( K_1 > 0 \)

\[
A_m \geq n^{-1} \sigma^2 C_1 \sum_{j=1}^{m} j^{\beta/r} \left(C_1 m^{\beta/r} - C_2 j^{\beta/r}\right) \geq K_1 n^{-1} m^{\beta/r} \sum_{j=1}^{m} j^{\beta/r}.
\]

Hence, there exists \( K_2 > 0 \) such that for all

\[
m > K_2 n^{r/(2\beta+r)}
\]

it holds that \( A_m > Q \). In a similar manner we can show that there exists \( K_3 > 0 \) such that for all

\[
m < K_3 n^{r/(2\beta+r)}
\]

it holds that \( A_m < Q \).

That leads us to the conclusion that \( N \geq n^{r/(2\beta+r)} \). Then equation (2.10) has a unique solution that satisfies

\[
x = \frac{\sigma^2}{n \left(Q + \frac{\sigma^2}{n} \sum_{j=0}^{N-1} a_j^2\right)} \sum_{j=0}^{N-1} a_j \approx \frac{1}{n} N^{1+\beta/r} \approx n^{-\beta/(2\beta+r)}.
\]

(ii) Since \( G \) satisfies the geometry assumption, we deduce from (i) that for some \( K_4 > 0 \)

\[
l_j' \approx \left(1 - K_4 n^{-\beta/(2\beta+r)} j^{\beta/r}\right)_{+}, \text{ for } j = i_0, \ldots, N.
\]

For \( j = 0, \ldots, i_0 - 1 \) we bound \( l_j' \) from above by 1. Then

\[
S \approx n^{-1} i_0 + n^{-1} \sum_{j=i_0}^{N-1} \left(1 - K_4 n^{-\beta/(2\beta+r)} j^{\beta/r}\right)_{+} \approx n^{-1} N \approx n^{-2\beta/(2\beta+r)}.
\]
2.5. Proofs

(iii) Note that for $j > N$ we have $v^2_j = 0$. We also know that $a_N < \frac{1}{x}$. Then

$$v^2_j a_j^2 = \frac{\sigma^2 a_j (1 - xa_j)}{nx} \leq \frac{\sigma^2 a_N}{nx} \leq \frac{\sigma^2}{nx^2}.$$ 

Hence, as $n \to \infty$,

$$\max_{j=0, \ldots, n-1} v^2_j a_j^2 = O\left(n^{-r/(2\beta+r)}\right).$$

This finishes the proof of the lemma. ■

2.5.1.2 Proof of the upper bound on the risk

Recall that we only need to provide the upper bound in (2.9). Consider the Pinsker’s estimator $\hat{f} = (\lambda_0 Z_0, \ldots, l'_{n-1} Z_{n-1})$ with

$$l'_{j} = (1 - xa_j)_{+},$$

where $a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r}$ and $x$ is a unique solution of (2.10). Using Lemma 2.5.1 and Lemma 2.5.2 we conclude that the Pinsker’s estimator satisfies

$$\sup_{f \in B_n(Q)} \mathbb{E}_f \left( \sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2 \right) \lesssim n^{-2\beta/(2\beta+r)}.$$ 

2.5.1.3 Proof of the lower bound on the risk

We follow the steps of the general reduction scheme for obtaining minimax rates (see e.g. Chapter 3 of Tsybakov [2009] for more details). First, we reduce the considered parameter space. We show that it is sufficient to only take into account the first $N$ coefficients in the decomposition of the target function, where

$$N = \max \left\{ m : \varepsilon^2 \sum_{j=0}^{m-1} a_j (a_m - a_j) < Q \right\}$$

with $a_j^2 = 1 + \lambda_j^{2\beta/r} n^{2\beta/r}$.

Denote $R_n$ to be the minimax risk

$$R_n = \inf_{\hat{f}} \sup_{f \in B_n(Q)} \mathbb{E}_f \left( \sum_{i=0}^{n-1} (\hat{f}_i - f_i)^2 \right).$$
2. The problem of function estimation on large graphs

For the coefficients \( a_j \) define

\[
B_n(Q, N) = \{ f^{(N)} = (f_0, \ldots, f_{N-1}, 0, \ldots, 0) \in \mathbb{R}^n : \sum_{j=0}^{N-1} a_j^2 f_j^2 \leq Q \}.
\]

Observe that \( B_n(Q, N) \subseteq B_n(Q) \). Then

\[
R_n \geq \inf_{\hat{f}^{(N)} \in B_n(Q, N)} \sup_{f^{(N)} \in B_n(Q, N)} \mathbb{E}_f \sum_{j=0}^{N-1} (\hat{f}_j - f_j)^2.
\]

Following the steps of the proof of Pinsker’s theorem, we argue that it is sufficient to bound the Bayes risk instead of the minimax risk. Indeed, consider the density \( \mu(f^{(N)}) = \prod_{j=0}^{N-1} \mu_s(f_j) \) with respect to the Lebesgue measure on \( \mathbb{R}^N \). Here \( s_j = (1 - \delta) v_j^2 \) for some \( \delta \in (0, 1) \) and \( \mu_{\sigma} \) denotes the density of the Gaussian distribution with mean 0 and variance \( \sigma^2 \). By (2.13) we can bound the minimax risk from below by the Bayes risk

\[
R_n \geq \inf_{\hat{f}^{(N)} \in B_n(Q, N)} \sup_{f^{(N)} \in B_n(Q, N)} \mathbb{E}_f \sum_{j=0}^{N-1} (\hat{f}_j - f_j)^2 \mu(f^{(N)}) df^{(N)} \geq I^* - r^*,
\]

where

\[
I^* = \inf_{\hat{f}^{(N)} \in B_n(Q, N)} \sum_{j=0}^{N-1} \int_{B_n(Q, N)} \mathbb{E}_f (\hat{f}_j - f_j)^2 \mu(f^{(N)}) df^{(N)};
\]

\[
r^* = \sup_{f^{(N)} \in B_n(Q, N)} \sum_{j=0}^{N-1} \int_{B_n(Q, N)^c} \mathbb{E}_f (\hat{f}_j - f_j)^2 \mu(f^{(N)}) df^{(N)}
\]

with \( B(Q, N)^c = \mathbb{R}^N \setminus B_n(Q, N) \). From the proof of Pinsker’s theorem we get the following bounds

\[
I^* \gtrsim S,
\]

\[
r^* \lesssim \exp \left\{ -K \left( \max_{j=0, \ldots, n-1} v_j^2 a_j^2 \right)^{-1} \right\}
\]

for some \( K > 0 \). Using the results of Lemma 2.5.2 we conclude that

\[
R_n \gtrsim n^{-2\beta/(2\beta + r)}.
\]
2.5. Proofs

2.5.2 Proof of Theorem 2.4.2

2.5.2.1 Proof of the upper bound on the risk

We define the estimator that gives us an upper bound on the minimax risk based on the estimator \( \tilde{f} \), which has been introduced in subsection 2.5.1.2 of the proof of Theorem 2.4.1. Consider the estimator

\[
\tilde{\rho} = \Psi \left( \sum_{i=0}^{N-1} \tilde{f}_i \psi_i \right).
\]

By the reasoning given in the aforementioned subsection and using the properties of the link function, we can see that

\[
\sup_{\rho \in \mathcal{H}_\beta(\mathcal{Q})} E \rho \| \tilde{\rho} - \rho \|_2^2 \lesssim \sup_{f \in \mathcal{B}_n(\mathcal{Q})} E f \left( \sum_{i=0}^{n-1} (\tilde{f}_i - f_i)^2 \right) \lesssim n^{-2\beta/(2\beta + r)}.
\]

2.5.2.2 Proof of the lower bound on the risk

The proof of the lower bound on the risk is based on the corollary of Fano’s lemma (for more details see Corollary 2.6 in Tsybakov [2009]). Observe that by Markov’s inequality for any soft label functions \( \rho_0, \ldots, \rho_M \) there exists \( C > 0 \) such that

\[
\inf_{\hat{\rho}} \sup_{\rho \in \mathcal{H}_\beta(\mathcal{Q})} E \rho \| \tilde{\rho} - \rho \|_2^2 \gtrsim \inf_{\rho} \max_{\rho \in \{\rho_0, \ldots, \rho_M\}} P \rho \left( \| \tilde{\rho} - \rho \|_n \gtrsim n^{-2\beta/(2\beta + r)} \right).
\]

Consider probability measures \( P_0, P_1, \ldots, P_M \) corresponding to the soft label functions \( \rho_0, \ldots, \rho_M \). For a test \( \phi : \mathbb{R}^n \to \{0, 1, \ldots, M\} \) define the average probability of error by

\[
\overline{p}_M(\phi) = \frac{1}{M+1} \sum_{j=0}^{M} P_j(\phi \neq j).
\]

Additionally, let

\[
\overline{p}_M = \inf_{\phi} \overline{p}_M(\phi).
\]

From the general scheme for obtaining lower bounds for minimax risk (for more detail see Chapter 2 of Tsybakov [2009]) we know that if \( \rho_0, \ldots, \rho_M \) are such that for any pair \( i, j \in \{0, \ldots, M\} \)

\[
\| \rho_i - \rho_j \|_n \gtrsim n^{-2\beta/(2\beta + r)}, \text{ when } i \neq j,
\]

then

\[
\inf_{\hat{\rho}} \max_{\rho \in \{\rho_0, \ldots, \rho_M\}} P \rho \left( \| \tilde{\rho} - \rho \|_2^2 \gtrsim C n^{-2\beta/(2\beta + r)} \right) \gtrsim \overline{p}_M.
\]
Also, Corollary 2.6 in Tsybakov [2009] states that if $P_0, P_1, \ldots, P_M$ satisfy

$$\frac{1}{M+1} \sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M, \quad (2.16)$$

for some $0 < \alpha < 1$, then

$$p_M \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$ 

Here $K(\cdot, \cdot)$ is the Kullback–Leibler divergence. Hence, if we construct the probability measures $P_0, P_1, \ldots, P_M$ corresponding to some soft label functions $\rho_0, \ldots, \rho_M$ for which (2.15), (2.16) holds, we will have

$$\inf_{\hat{\rho}} \sup_{\rho \in \{\rho : \Psi^{-1}(\rho) \in H^{\beta}(Q)\}} \mathbb{E}_{\rho} n^{2\beta/(2\beta+r)} \|\hat{\rho} - \rho\|^2 \gtrsim \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$ 

If $M \to \infty$, as $n \to \infty$, the required result follows.

Let $N = n^{r/(2\beta+r)}$ and let $\psi_0, \ldots, \psi_{N-1}$ to be an orthonormal eigenbasis of the graph Laplacian $L$ with respect to the $\|\cdot\|_n$–norm. For $\delta > 0$ and $\theta = (\theta_1, \ldots, \theta_N) \in \{-1\}^N$ define

$$f_\theta = \delta N^{-(2\beta+r)/2r} \sum_{j=0}^{N-1} \theta_j \psi_j.$$ 

We will select $M$ vectors of coefficients $\theta^{(j)}$ such that the probability measures corresponding to $\rho_j = \Psi(f_{\theta^{(j)}})$ will satisfy (2.16), where $\Psi$ is the link function.

Observe that for small enough $\delta > 0$ functions $f_\theta$ belong to the class $H^\beta(Q)$. Indeed, using the geometry assumption we obtain

$$\left\langle f_\theta, (I + (n^{\frac{\beta}{2}}L)^\beta) f_\theta \right\rangle_n = \delta^2 N^{-(2\beta+r)/r} \sum_{j=0}^{N-1} (1 + n^{2\beta/r} \lambda_j^\beta) \leq \delta^2 N^{-(2\beta+r)/r} \left( N + C_2 t_0^{(2\beta+r)/r} + C_2 \sum_{j=i_o}^{N} j^{2\beta/r} \right) \leq K_1 \delta^2$$

for some constant $K_1 > 0$.

We pick a subset $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$ of $\{-1\}^N$ such that for any pair $i, j \in \{1, \ldots, M\}$ such that $i \neq j$ the vectors from the subset were sufficiently distant from each other

$$d_h(\theta^{(i)}, \theta^{(j)}) \gtrsim N, \quad (2.17)$$

where $d_h(\theta, \theta') = \sum_{j=0}^{N-1} \mathbf{1}_{\theta_j = \theta'_j}$ is the Hamming distance. By the Varshamov–Gilbert bound (see for example Lemma 2.9 in Tsybakov [2009]) we know that there
2.5. Proofs

exist such a subset \( \{ \theta^{(1)}, \ldots, \theta^{(M)} \} \), and the size \( M \) of this subset satisfies

\[ M \geq b^N \]

for some \( 1 < b < 2 \). Let \( \theta^{(0)} = (0, \ldots, 0) \in \mathbb{R}^N \). We define a set of probability measures \( \{ P_0, \ldots, P_M \} \) by setting \( P_j = P_{\rho_j} \), where \( \rho_j = \Psi(f_{\theta^{(j)}}) \).

In order to show that the \( P_j \) satisfy (2.16) we present a technical lemma. In the classification setting the Kullback–Leibler divergence \( K(\cdot, \cdot) \) satisfies

\[ K(P_\rho, P'_\rho) = \sum_{i=1}^n \left( \rho(i) \log \frac{\rho(i)}{\rho'(i)} - (1 - \rho(i)) \log \frac{1 - \rho(i)}{1 - \rho'(i)} \right). \]

**Lemma 2.5.3.** If \( \frac{\Psi'}{\Psi(1 - \Psi)} \) is bounded, then there exist \( c > 0 \) such that for any \( v_1, v_2 \in \mathbb{R}^n \) we have

\[ K(P_{\Psi(v_1)}, P_{\Psi(v_2)}) \leq nc\|v_1 - v_2\|^2. \]

**Proof.** For every \( x \in \mathbb{R} \) consider the function \( g_x : \mathbb{R} \to \mathbb{R} \) defined as

\[ g_x(y) = \Psi(x) \log \frac{\Psi(x)}{\Psi(y)} + (1 - \Psi(x)) \log \frac{1 - \Psi(x)}{1 - \Psi(y)}. \]

We see that \( g'_x(y) = \frac{\Psi'(y)}{\Psi(y)(1 - \Psi(y))}(\Psi(y) - \Psi(x)) \). Then by Taylor’s theorem we can see that

\[ |g_x(y)| \leq \sup_{v \in [x,y] \cup [y,x]} \left| \frac{\Psi'(v)}{\Psi(v)(1 - \Psi(v))} \right| \sup_{v \in [x,y] \cup [y,x]} |\Psi'(v)|(x - y)^2. \]

The statement of the lemma follows. \( \blacksquare \)

By Lemma 2.5.3, we obtain for some \( K_2 > 0 \)

\[ K(P_j, P_0) \leq K_2 n \|f_{\theta^{(j)}} - 0\|^2_n = 4K_2 \delta^2 n N^{-2\beta/r}, \]

since

\[ \|f_\theta - f_{\theta'}\|^2_n = 4\delta^2 N^{-(2\beta + r)/r} d_h(\theta, \theta'). \tag{2.18} \]

Observe that this bound does not depend on \( j \). Hence,

\[ \frac{1}{M + 1} \sum_{j=1}^M K(P_j, P_0) \leq K_2 \delta^2 n N^{-2\beta/r} = K_2 \delta^2 \log M. \]

We can choose \( \delta > 0 \) to be small enough such that the condition (2.16) is satisfied with some \( 0 < \alpha < 1 \).

To finish the proof of the theorem we need to show that \( \rho_0, \ldots, \rho_M \) satisfy (2.15).
From (2.17) and (2.18) we get
\[ \| f_{\theta^{(i)}} - f_{\theta^{(j)}} \|_n \gtrsim n^{-\beta/(2\beta+r)}. \]

Moreover, by the assumption of the theorem we have for any \( j = 1, \ldots, M \)
\[ \max_{i=1,\ldots,n} |f_{\theta^{(j)}}(i)| \lesssim N^{1-(2\beta+r)/2r} \max_{i=1,\ldots,n} |\psi_j(i)| \lesssim N^{(r-2\beta)/2r}. \]

For \( \beta \geq r/2 \) the norm is then bounded by some constant, which does not depend on \( n \) or \( j \). Hence, there exists \( K_3 \geq 0 \) such that for every \( i = 1, \ldots, n \) and every \( j = 1, \ldots, M \)
\[ |f_{\theta^{(j)}}(i)| \leq K_3. \]

Observe that since \( \Psi'(x) \neq 0 \) for any \( x \in \mathbb{R} \), there exists \( K_4 > 0 \) such that for any \( x, y \in [-K_3, K_3] \)
\[ |\Psi(x) - \Psi(y)| \geq K_4 |x - y|. \]

Thus, for any pair \( i, j \in \{1, \ldots, M\} \) such that \( i \neq j \)
\[ \| \rho_i - \rho_j \|_n = \| \Psi(f_{\theta^{(i)}}) - \Psi(f_{\theta^{(j)}}) \|_n \gtrsim n^{-\beta/(2\beta+r)}. \]

This completes the proof of the theorem.
Chapter 3

Bayesian function estimation on large graphs

3.1 Introduction

In the previous chapter we have discussed the general mathematical framework that allows us to study the performance of nonparametric function estimation methods on large graphs. We have also obtained minimax rates for two prototypical problems on graphs: the regression and the binary classification. In this chapter we continue studying the same setting and introduce a Bayesian approach to these problems. We develop Bayesian procedures that involve Laplacian regularisation and show that these procedures turn out to be (almost) optimal in a minimax sense.

Several approaches to learning functions on graphs that have been explored in the literature involve regularisation using the Laplacian matrix associated with the graph (see, for example, Belkin et al. [2004], Smola and Kondor [2003], Ando and Zhang [2007], Zhu et al. [2003], Huang et al. [2011]). The graph Laplacian \( L = D - A \), where \( A \) is the adjacency matrix of the graph and \( D \) is the diagonal matrix with the degrees of the vertices on the diagonal), when viewed as a linear operator, acts on a function \( f \) on the graph as

\[
Lf(i) = \sum_{j \sim i} (f(i) - f(j)),
\]

where we write \( i \sim j \) if vertices \( i \) and \( j \) are connected by an edge. Several related operators are routinely employed as well, for instance, the normalised Laplacian \( \tilde{L} = D^{-1/2}LD^{-1/2} \). We will continue to work with \( L \), but much of the story goes through if \( L \) is replaced by such a related operator, after minor adaptations.

Clearly, the Laplacian norm of \( f \) quantifies how much the function \( f \) varies when moving along the edges of the graph. Therefore, several papers have proposed regularisation or penalisation using this norm, as well as generalisations involving
Bayesian function estimation on large graphs

powers of the Laplacian or other functions, for instance, exponential functions. See, for example, Belkin et al. [2004] or Smola and Kondor [2003] and the references therein. There exist only few papers that study theoretical aspects of the performance of such methods. We mention, for example, Belkin et al. [2004], in which a theoretical analysis of a Tikhonov regularisation method is conducted in terms of algorithmic stability. Johnson and Zhang [2007] and Ando and Zhang [2007] consider sub-sampling schemes for estimating a function on a graph.

However, the existing papers have different viewpoints than ours and do not study how the performance depends on (the combination of) graph geometry and function regularity.

We investigate Bayesian regularisation approaches, where we consider two types of priors on functions on graphs. The first type performs regularisation using a power of the Laplacian. This can be seen as the graph analogue of Sobolev norm regularisation of functions on “ordinary” Euclidean spaces. The second type of priors uses an exponential function of the Laplacian. This can be viewed as the analogue of the popular squared exponential prior on functions on Euclidean space or its extension to manifolds, as studied by Castillo et al. [2014]. In both cases we consider hierarchical priors with the aim of achieving automatic adaptation to the regularity of the function of interest.

To assess the performance of our Bayes procedures we take an asymptotic perspective. We let the number of vertices of the graph grow and ask at what rate the posterior distribution concentrates around the unknown function \( f \) that generates the data. We make the same assumption on the geometry of the graph and the smoothness of the true function in terms of Sobolev-type balls as in the previous chapter. We show that the proposed methods appear to be (almost) optimal in a minimax sense and adaptive to the regularity of the true function.

Additionally, we consider the case of missing observations, where only a part of observations is available. In order to tackle the problem we assume that the available data is somehow uniformly distributed on the graph. We study two missing mechanisms: coin flipping and sampling with replacement. In the coin flipping mechanism for every vertex we independently flip a coin and see the observation with probability \( p_n \). In the sampling with replacement we sample \( m_n = np_n \) points from the complete data set, allowing repetition of the vertices. We are interested in the case where the available sample size constitutes a small part of the whole set, so we allow \( p_n \to 0 \). Although we only consider part of the data points, we use a distance on functions that takes all \( n \) points into account, i.e. we study a form of generalisation performance.

The remainder of the chapter is organised as follows. In Section 3.2 we introduce two families of priors on functions on graphs. We present theorems that quantify the amount of mass that the priors put on neighbourhoods of “smooth” functions and quantify the complexity of the priors in terms of metric entropy. In Section 3.3 these results are used to derive theorems about posterior contraction for the nonparametric regression and the binary classification problems on graphs. In Section 3.4 we treat the missing observation case, introducing the modified families of priors and deriving
3.2. General results on prior concentration

posterior contraction rates in the same setting of nonparametric regression and binary classification. Mathematical proofs are given in Section 3.5.

3.2 General results on prior concentration

We consider two different priors on functions on graphs. The first corresponds to regularisation using a power of the Laplacian, the second one uses an exponential function of the Laplacian. In this section we present two general results which quantify both the mass that these priors give to shrinking $\| \cdot \|_n$-neighbourhoods of a fixed function $f_0$, and the complexity of the support of the priors measured in terms of metric entropy. In the next section we combine these results with known results from Bayesian nonparametrics theory to deduce convergence rates and adaptation for nonparametric regression and classification problems on graphs.

Our results assume that the geometry condition (2.2) holds for some $r \geq 1$ and the regularity of the function is quantified by Sobolev-type balls defined in (2.4). The first family of priors we consider penalise the higher order Laplacian norm of the function of interest. This corresponds to using a Gaussian prior with a power of the Laplacian as precision matrix (inverse of the covariance). (We note that since the Laplacian always has 0 as an eigenvalue, it is not invertible. We remedy this by adding a small multiple of the identity matrix $I$ to $L$.) The larger the power of the Laplacian used, the more “rough” functions on the graph are penalised. The power is regulated by a hyperparameter $\alpha > 0$ which can be seen as describing the “baseline regularity” of the prior. To enlarge the range of regularities for which we obtain good contraction rates in the statistical results, we add a multiplicative hyperparameter which we endow with a suitable hyperprior. In (3.2) we assume standard exponential distribution, but inspection of the proof shows that the range of priors for which the result holds is actually larger. To keep the exposition clean we omit these details. Observe that in the following we denote by $N(\| \cdot \|^{-\alpha}_B)$ the minimal number of balls of $\| \cdot \|$-radius $\varepsilon$ needed to cover $B$ for some $\varepsilon > 0$ and a norm $\| \cdot \|$ on a set $B$.

**Theorem 3.2.1** (Power of the Laplacian). Suppose the geometry assumption holds for $r \geq 1$. Let $\alpha > 0$ be fixed and assume that $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $0 < \beta \leq \alpha + r/2$. Let the random function $f$ on the graph be defined by

$$
c \sim \text{Exp}(1)
$$

$$
f | c \sim N(0, ((n/c)^2/r (L + n^{-2} I))^{\alpha + r/2})^{-1}).
$$

Then there exists a constant $K_1 > 0$ and for all $K_2 > 1$ there exist Borel measurable
3. Bayesian function estimation on large graphs

subsets $B_n$ of $\mathbb{R}^n$ such that for every sufficiently large $n$

$$P(\|f - f_0\|_n < \varepsilon_n) \geq e^{-K_1 n\varepsilon_n^2},$$  \hspace{1cm} (3.4)

$$P(f \notin B_n) \leq e^{-K_2 n\varepsilon_n^2},$$ \hspace{1cm} (3.5)

$$\log N(\varepsilon_n, B_n, \|\cdot\|_n) \leq n\varepsilon_n^2,$$ \hspace{1cm} (3.6)

where $\varepsilon_n$ is a multiple of $n^{-\beta/(2\beta+r)}$.

Note that in this theorem we obtain the rate $n^{-\beta/(2\beta+r)}$ for all $\beta$ in the range $(0, \alpha + r/2]$. In the statistical results presented in Section 3.3 this translates into rate-adaptivity up to regularity level $\alpha + r/2$. So by putting a prior on the multiplicative scale we achieve a degree of adaptation, but only up to an upper bound that is limited by our choice of the hyperparameter $\alpha$. A possible solution is to consider other functions of the Laplacian instead of using a power of $L$ in the prior specification. Here we consider usage of an exponential function of the Laplacian. We include a “lengthscale” or “bandwidth” hyperparameter that we endow with a prior as well for added flexibility. This prior can be seen as the analogue of the prior used in Castillo et al. [2014] in the context of function estimation on manifolds, which in turn is a generalisation of the popular squared exponential Gaussian prior used for estimation functions on Euclidean domains (e.g. Rasmussen and Williams [2006]). However, we stress again that we do not rely on an embedding of our graph in a manifold or the existence of a “limiting manifold”.

**Theorem 3.2.2 (Exponential of the Laplacian).** Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $\beta > 0$. Let the random function $f$ on the graph be defined by

$$c \sim \text{Exp}(1)$$

$$f \mid c \sim N(0, ne^{-(n/c)^{2/r}L}).$$ \hspace{1cm} (3.7)

Then there exists a constant $K_1 > 0$ and for all $K_2 > 1$ there exist Borel subsets $B_n$ of $\mathbb{R}^n$ such that for every sufficiently large $n$

$$P(\|f - f_0\|_n < \varepsilon_n) \geq e^{-K_1 n\varepsilon_n^2},$$ \hspace{1cm} (3.9)

$$P(f \notin B_n) \leq e^{-K_2 n\varepsilon_n^2},$$ \hspace{1cm} (3.10)

$$\log N(\tilde{\varepsilon}_n, B_n, \|\cdot\|_n) \leq n\tilde{\varepsilon}_n^2,$$ \hspace{1cm} (3.11)

where $\varepsilon_n = (n / \log^{1+r/2} n)^{-\beta/(2\beta+r)}$ and $\tilde{\varepsilon}_n = \varepsilon_n \log^{1/2+r/4} n$.

3.3 Function estimation on graphs

In this section we translate the general Theorems 3.2.1 and 3.2.2 into results about posterior contraction in nonparametric regression and binary classification problems
3.3. Function estimation on graphs

on graphs. Since the arguments needed for this translation are very similar to those in earlier papers, we omit full proofs and just give pointers to the literature.

3.3.1 Nonparametric regression

As before we let \( G \) be a connected simple undirected graph with vertices \( 1, 2, \ldots, n \). In the regression case we assume that we have observations \( Y_1, \ldots, Y_n \) at the vertices of the graph, satisfying

\[
Y_i = f_0(i) + \sigma \xi_i,
\]

where \( f_0 \) is the function on \( G \) that we want to make inference about, \( \xi_i \) are independent standard Gaussian, and \( \sigma > 0 \). We assume that the underlying graph satisfies the geometry assumption with some parameter \( r \geq 1 \). As prior on the regression function \( f \) we then employ one of the two priors described by (3.2)–(3.3) or (3.7)–(3.8). If \( \sigma \) is unknown, we assume it belongs to a compact interval \([a, b] \subset (0, \infty)\) and endow it with a prior with a positive, continuous density on \([a, b]\), independent of the prior on \( f \).

For a given prior \( \Pi \), the corresponding posterior distribution on \( f \) is denoted by \( \Pi(\cdot | Y_1, \ldots, Y_n) \). For a sequence of positive numbers \( \varepsilon_n \to 0 \) we say that the posterior contracts around \( f_0 \) at the rate \( \varepsilon_n \) if for all large enough \( M > 0 \),

\[
\Pi(f : \|f - f_0\|_n \geq M \varepsilon_n | Y_1, \ldots, Y_n) \xrightarrow{P_{f_0}} 0
\]
as \( n \to \infty \). Here the convergence is in probability under the law \( P_{f_0} \) corresponding to the data generating model (3.12).

The usual arguments allow us to derive the following statements from Theorems 3.2.1 and 3.2.2. See, for instance, van der Vaart and van Zanten [2008a] or de Jonge and van Zanten [2013] for details.

**Theorem 3.3.1** (Nonparametric regression). Suppose the geometry assumption holds for \( r \geq 1 \). Assume that \( f_0 \in H^\beta(Q) \) for \( \beta, Q > 0 \).

1. (Power of the Laplacian.) If the prior on \( f \) is given by (3.2)–(3.3) for \( \alpha > 0 \) and \( \beta \leq \alpha + r/2 \), then the posterior contracts around \( f_0 \) at the rate \( n^{-\beta/(r+2\beta)} \).

2. (Exponential of the Laplacian.) If the prior on \( f \) is given by (3.7)–(3.8), then the posterior contracts around \( f_0 \) at the rate \( n^{-\beta/(r+2\beta)} \log^\kappa n \) for some \( \kappa > 0 \).

Observe that since the priors do not use knowledge of the regularity \( \beta \) of the regression function, we obtain rate-adaptive results. For the power prior the range of regularities that we can adapt to is bounded by \( \alpha + r/2 \), where \( \alpha \) is the hyperparameter describing the “baseline regularity” of the prior. In the case of the exponential prior the range is unbounded. This comes at the modest cost of having an additional logarithmic factor in the rate.
We can derive the analogous results in the classification problem in which we assume that the data $Y_1,\ldots,Y_n$ are independent $\{0,1\}$-valued variables, observed at the vertices of the graph. In this case the goal is to estimate the binary regression function $\rho_0$, or “soft label function” on the graph, given by

$$\rho_0(i) = \mathbb{P}_0(Y_i = 1).$$

We consider priors on $\rho$ constructed by first defining a prior on a real-valued function $f$ by (3.2)–(3.3) or (3.7)–(3.8) and then setting $\rho = \Psi(f)$, where $\Psi : \mathbb{R} \to (0,1)$ is a suitably chosen link function. We again assume that $\Psi$ is a strictly increasing, differentiable function onto $(0,1)$ such that $\Psi'/(\Psi(1 - \Psi))$ is uniformly bounded. Also in this case we denote the posterior corresponding to a prior $\Pi$ by $\Pi(\cdot | Y_1,\ldots,Y_n)$ and we say that the posterior contracts around $\rho_0$ at the rate $\varepsilon_n$ if for all large enough $M > 0$,

$$\Pi(\rho : \|\rho - \rho_0\|_n \geq M \varepsilon_n | Y_1,\ldots,Y_n) \xrightarrow{P_{\rho_0}} 0$$

as $n \to \infty$.

To derive the following result from Theorems 3.2.1 and 3.2.2 we can argue along the lines of the proof of Theorem 3.2 of van der Vaart and van Zanten [2008a]. Some adaptations are required, since in the present case we have fixed design points. However, the necessary modifications are straightforward and therefore omitted.

**Theorem 3.3.2** (Classification). Suppose the geometry assumption holds for $r \geq 1$. Let $\Psi : \mathbb{R} \to (0,1)$ be onto, strictly increasing, differentiable and suppose that $\Psi'/(\Psi(1 - \Psi))$ is uniformly bounded. Assume that $\Psi^{-1}(\rho_0) \in H^\beta(Q)$ for $\beta,Q > 0$.

(i) (Power of the Laplacian.) If the prior on $p$ is given by the law of $\Psi(f)$, for $f$ given by (3.2)–(3.3) for $\alpha > 0$ and $\beta \leq \alpha + r/2$, then the posterior contracts around $\rho_0$ at the rate $n^{-\beta/(r+2\beta)}$.

(ii) (Exponential of the Laplacian.) If the prior on $p$ is given by the law of $\Psi(f)$, for $f$ given by (3.7)–(3.8), then the posterior contracts around $\rho_0$ at the rate $n^{-\beta/(r+2\beta)} \log \kappa n$ for some $\kappa > 0$.

### 3.4 Function estimation on graphs with missing observations

The previous section covered function estimation on large graphs in the full observation setting. However, it is often an intermediate step in function recovery, since a lot of practical applications involve making inference about a function in the situation when only partial data is available. As an example, consider the classification problem on the protein interaction graph in Figure 3.1 studied in
3.4. Function estimation on graphs with missing observations

Hartog and van Zanten [2016]. A vertex of the graph is labelled red if the corresponding protein is involved in so-called intracellular signalling cascades (ICSC), which is a specific form of communication, and blue otherwise. The goal is to recover the missing labels marked white on this graph. Despite the fact that some observations are missing, we would still like to recover function as good as possible with respect to the $\| \cdot \|_n$-norm that takes the values at all the vertices into account. The convergence rates obtained for this case can be then viewed as a quantitative characteristic of generalisation performance. We again assume that the underlying graph satisfies the geometry assumption for some $r \geq 1$ and that the target function has a smoothness level $\beta > 0$ in terms of the Sobolev type balls, defined in Sections 2.2.2 and 2.3.

Certainly, an estimator on the graph will not be able to recover the target function based on incomplete data set as efficiently as in the full observation case. In order to study how having partial data influences the performance of estimators we again take an asymptotic approach and investigate how the rates of convergences are affected by missing data. Suppose we only observe a small proportion of the full data sample, meaning that we see (on average) $np_n$ points out of a complete sample of $n$ points with $p_n$ potentially going to zero as $n \to \infty$. Recall that in the full observation case the minimax rates for regression and binary classification on large graphs were equal to $n^{-\beta/(2\beta+r)}$. It would be natural to expect that in the missing observation case the rates will be transformed into $(np_n)^{-\beta/(2\beta+r)}$, since we only see $np_n$ points. In this chapter we show that the contraction rates for the modifications of the Bayesian estimators developed in Section 3.2 are equal to the aforementioned value in the missing observation case. The minimax rates are not known for the present situation. However, a particular case of our problem with

Figure 3.1: Protein-protein interaction graph. The red vertices in the graph are involved in ICSC, the blue vertices are not involved, and the label of the white vertices is not known.
3. Bayesian function estimation on large graphs

\( r = 1 \) can be linked to the regression setting on the interval \([0, 1]\) with missing observations, studied in Efromovich [2011]. As shown in the aforementioned paper, the minimax rate for this problem is \((np_n)^{-\beta/(2\beta+1)}\). Hence, it appears that the rates obtained in this section cannot be fundamentally improved.

The way in which the missing observations are distributed throughout the graph plays a crucial role in estimation. We do not treat all the possible cases but rather consider a prototypical problem that satisfies the following two assumptions. First, we assume that the missingness does not depend on the values of variables in the data (we consider missing completely at random (MCAR) mechanisms, the general classification of types of missing mechanisms is given in Little and Rubin [2002]). We also assume that the available data is somehow uniformly distributed on the graph. We implement these ideas in two missing mechanisms: a coin flipping scheme, where we see the observation at a certain vertex independently of the others with some probability \(0 < p_n < 1\), and sampling with replacement of \(p_n n\) data points.

In the next section we discuss the missing mechanisms that we study in more detail. Then we present general theorems about the convergence rates of posterior distribution in the regression and binary classification settings with missing observations. The derived theorems are similar and based on the works Ghosal et al. [2000], Ghosal and van der Vaart [2007], and Ghosal and van der Vaart [2001]. However, the missing observation setting requires to control the prior mass and the entropy of balls of larger diameter compared to the full observation case. We introduce a modification of the power prior defined by (3.2)–(3.3) and we consider the exponential prior defined by (3.7)–(3.8) in the missing observation setting. For those priors we derive convergence rates of posterior distributions for regression and binary classification problems on the graph and show that the procedures remain rate-adaptive.

### 3.4.1 Missing mechanisms

We introduce the missing mechanisms for the problem of recovering a function \(g_0\) defined on the vertices of the graph \(G\). In the regression setting \(g_0\) is equal to the regression function \(f_0\), while in the classification setting the function of interest a soft label function \(\rho_0\). Consider a full observation set \((Y_1, \ldots, Y_n)\) on the vertices \(1, \ldots, n\) of the graph \(G\).

For missing at random by coin flipping mechanism we consider \(V_i\) to be independent Bernoulli random variables with success probability \(p_n > 0\). Next, we define

\[
Z_i = \begin{cases} 
\star, & \text{if } V_i = 0; \\
Y_i, & \text{if } V_i = 1.
\end{cases}
\]  

(3.13)

The goal is to estimate \(g_0\) from the observations \(Z_1, \ldots, Z_n\) as accurately as possible with respect to the \(\| \cdot \|_n\) norm

\[
\|g_1 - g_2\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (g_1(i) - g_2(i))^2.
\]
3.4. Function estimation on graphs with missing observations

For a given prior $\Pi$, the corresponding posterior distribution is denoted by $\Pi(\cdot \mid Z_1, \ldots, Z_n)$ and we say that the posterior contracts at rate $\varepsilon_n > 0$ if, for every sufficiently large constant $M$, an $n \to \infty$

$$\Pi(g : \|g - g_0\|_n \geq M\varepsilon_n \mid Z_1, \ldots, Z_n) \xrightarrow{P_{g_0}} 0.$$ 

Here the convergence is in probability under the law $P_{g_0}$ corresponding to the data generating model (3.13).

For the sampling with replacement missing mechanism we only see $m_n = p_n n$ pairs of observations $(X_i, Z_i)$, where

$$X_i \sim \text{Uniform}\{1, \ldots, n\}, \quad (3.14)$$
$$Z_i \mid X_i \sim Y_{X_i}. \quad (3.15)$$

The goal is again to estimate $g_0$ with as accurately as possible with respect to the $\| \cdot \|_n$-norm, having only $m_n$ pairs of observations available.

For a given prior $\Pi$, the corresponding posterior distribution in this case is denoted by $\Pi(\cdot \mid (X_1, Z_1), \ldots, (X_{m_n}, Z_{m_n}))$ and we say that the posterior contracts at rate $\varepsilon_n > 0$ if, for every sufficiently large constant $M$, and $n \to \infty$

$$\Pi(g : \|g - g_0\|_n \geq M\varepsilon_n \mid (X_1, Z_1), \ldots, (X_{m_n}, Z_{m_n})) \xrightarrow{P_{g_0}} 0.$$ 

In the regression problem $g_0 = f_0$ and the observations $Y_i$ satisfy standard regression relation

$$Y_i = f_0(i) + \sigma \xi_i,$$

where $\sigma > 0$ and $\xi_i$ are independent standard Gaussian.

In the binary classification problem we independently observe values $Y_i \in \{0, 1\}$ at the vertices of the graph, and the goal is to estimate the binary regression function $g_0 = \rho_0$, or “soft label function”, given by

$$\rho_0(i) = P_0(Y_i = 1).$$

For the classification problem we only consider priors on $\rho$ constructed by first defining a prior on a real-valued function $f$ and then setting $\rho = \Psi(f)$, where $\Psi : \mathbb{R} \to (0, 1)$ is as usual a strictly increasing, differentiable function onto $(0,1)$ such that $\Psi'/(\Psi(1 - \Psi))$ is uniformly bounded. Additionally, we denote $f_0$ to be equal to $\Psi^{-1}(\rho_0)$.

3.4.2 General theorems for regression and classification

General results on the convergence rates of posterior distribution, such as Theorem 1.5.2, cannot be directly applied in the case of missing observations. In this section we provide analogous results that are based on the aforementioned theorem with slight modifications allowing us to accommodate the missing data in the case of
regression and classification. We show that in order to derive a rate of convergence the prior mass and the entropy should be controlled for balls of larger diameter than for the same expressions in the full observation case. Observe that these requirements turn out to be more demanding than in the original theorem.

**Theorem 3.4.1.** Consider the sequences $\varepsilon_n, \varepsilon_n^*, \bar{\varepsilon}_n \to 0$, such that $n\varepsilon_n^2 \to \infty$, $\varepsilon_n^* \asymp p_n^{-1/2}\varepsilon_n$, $\bar{\varepsilon}_n \asymp \varepsilon_n (\log n)^k$ for $k \geq 0$. Additionally suppose that for some constant $K > 0$ and for all $L > 1$ there exist Borel measurable subsets $B_n$ of $\mathbb{R}^n$ for which the following is true, when $n$ is large enough

$$\Pi(f : \|f - f_0\|_n < \varepsilon_n^*) \geq e^{-Kn\varepsilon_n^2},$$  

(3.16)

$$\Pi(f \notin B_n) \leq e^{-Ln\varepsilon_n^2},$$  

(3.17)

$$\log N(\varepsilon_n^*, B_n, \| \cdot \|_n) \leq n\bar{\varepsilon}_n^2.$$  

(3.18)

Then the posterior contracts at the rate $\varepsilon_n^*(\log n)^k$ for both regression and classification problems and for both missing mechanisms.

The theorem can be proved using general Theorems 1.5.1, 1.5.2. However, the direct application of the theorems is not possible, so we translate the given setting in order for it to match the conditions of the aforementioned theorems. Then we translate the obtained results back to the original setting of the theorem.

### 3.4.3 General results on prior concentration and convergence rates for regression and classification

We again consider two families of priors: the modified power prior and the exponential prior. We present theorems analogous to Theorems 3.2.1, 3.2.2, providing lower bounds on the mass that these priors give to shrinking $\| \cdot \|_n$-neighbourhoods of a fixed function $f_0$, and the complexity of the support of the priors, measured in terms of metric entropy. We show that the developed priors satisfy the new, more strict conditions, which we require for obtaining convergence rates in the missing observations setting. In the next section we combine these results with the general theorems from Section 3.4.2 to deduce convergence rates and adaptation for nonparametric regression and classification problems on graphs with missing observations.

Consider a connected undirected simple graph $G$ with vertices labeled $\{1, \ldots, n\}$ that satisfies the geometry assumption with $r \geq 1$.

**Theorem 3.4.2.** Suppose the geometry assumption holds for $r \geq 1$. Let $\alpha > 0$ be fixed and assume that $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $0 < \beta \leq \alpha + r/2$. Let the random function $f$ on the graph be defined by

$$c \sim \text{Exp}(p_n^r/(2\alpha + r))$$  

(3.19)

$$f|c \sim N(0, (((n/c)^{2/r}(L + n^{-2}I))^{\alpha+r/2})^{-1}).$$  

(3.20)
3.4. Function estimation on graphs with missing observations

Then there exists a constant $K_1 > 0$ and for all $K_2 > 1$ there exist Borel measurable subsets $B_n$ of $\mathbb{R}^n$ such that for every sufficiently large $n$,

\begin{align}
\mathbb{P}(\|f - f_0\|_n < \varepsilon_n^*) &\geq e^{-K_1 \varepsilon_n^2}, \\
\mathbb{P}(f \notin B_n) &\leq e^{-K_2 \varepsilon_n^2}, \\
\log N(\varepsilon_n^*, B_n, \|\cdot\|_n) &\leq n \varepsilon_n^2,
\end{align}

(3.21) (3.22) (3.23)

where $\varepsilon_n \asymp p_n^{r/2(2\beta+r)} n^{-\beta/(2\beta+r)}$ and $\varepsilon_n^* \asymp (np_n)^{-\beta/(2\beta+r)}$.

Observe that the scaling parameter in the power prior is distributed according to $\text{Exp}(p_n^{r/(2\alpha+r)})$, in contrast to the full observation case, where it had a standard exponential distribution. This increases the variance of the prior which is necessary to compensate for the uncertainty arising from the missing data.

Interestingly, as shown in the following theorem, the exponential prior doesn’t require any modifications in order to accommodate missing data.

**Theorem 3.4.3.** Suppose the geometry assumption holds for $r \geq 1$. Assume that $f_0 \in H^\beta(Q)$ for some $Q > 0$ and $\beta > 0$. Let the random function $f$ on the graph be defined by

\begin{align}
c &\sim \text{Exp}(1) \\
f | c &\sim N(0, ne^{-\varepsilon_n^2/c}\|c\|_L).
\end{align}

(3.24) (3.25)

Then there exists a constant $K_1 > 0$ and for all $K_2 > 1$ there exist Borel measurable subsets $B_n$ of $\mathbb{R}^n$ such that for every sufficiently large $n$,

\begin{align}
\mathbb{P}(\|f - f_0\|_n < \varepsilon_n^*) &\geq e^{-K_1 \varepsilon_n^2}, \\
\mathbb{P}(f \notin B_n) &\leq e^{-K_2 \varepsilon_n^2}, \\
\log N(\varepsilon_n^*, B_n, \|\cdot\|_n) &\leq n \varepsilon_n^2,
\end{align}

(3.26) (3.27) (3.28)

where $\varepsilon_n \asymp p_n^{r/2(2\beta+r)} n^{-\beta/(2\beta+r)}(\log n)^{(1+r/2)\beta/(2\beta+r)}$, $\varepsilon_n^* \asymp p_n^{-1/2} \varepsilon_n$, and $\bar{\varepsilon}_n \asymp \varepsilon_n (\log n)^k$ for some $k \geq 0$.

3.4.4 Bayesian procedure for function estimation on graphs with missing observations

In the regression case we assume the full observation set $Y = (Y_1, \ldots, Y_n)$ on the vertices of the graph to be such that

\begin{align}
Y_i = f_0(i) + \sigma \xi_i,
\end{align}

(3.29)

where $\xi_i$ are independent standard Gaussian, $\sigma > 0$ and $f_0$ is the function of interest.

We assume that we effectively see (on average) only $np_n$ data points uniformly distributed on the graph by one of the missing mechanisms described in Section
3. Bayesian function estimation on large graphs

3.4.1. As prior on the regression function \( f \) we then employ one of the two priors described by (3.19)–(3.20) or (3.24)–(3.25). If \( \sigma \) is unknown, we assume it belongs to a compact interval \([a, b] \subset (0, \infty)\) and endow it with a prior with a positive, continuous density on \([a, b]\), independent of the prior on \( f \).

In the binary classification problem we assume that the data \( Y_1, \ldots, Y_n \) are independent \( \{0, 1\} \)-valued variables, observed at the vertices of the graph. In this case the goal is to estimate the binary regression function \( \rho_0 \), or “soft label function” on the graph, given by

\[
\rho_0(i) = \mathbb{P}_0(Y_i = 1).
\]

We again assume to see only partial data according to one of the missing mechanisms described in Section 3.4.1. We consider a suitably chosen link function \( \Psi : \mathbb{R} \rightarrow (0, 1) \). We assume that \( \Psi \) is a strictly increasing, differentiable function onto \((0, 1)\) such that \( \Psi'(1 - \Psi) \) is uniformly bounded. We consider priors on \( \rho \) constructed by first defining a prior on a real-valued function \( f \) by (3.19)–(3.20) or (3.24)–(3.25) and then setting \( \rho = \Psi(f) \), where \( \Psi : \mathbb{R} \rightarrow (0, 1) \) is a suitably chosen link function.

Combining the results of the theorems stated in this section we derive the following result.

**Theorem 3.4.4.** Suppose the geometry assumption holds for \( r \geq 1 \). Let \( \alpha > 0 \) be fixed and assume that in the regression problem \( f_0 \in H^\beta(Q) \) for some \( Q > 0 \) and \( \beta > 0 \). For the classification problem we assume \( \Psi^{-1}(\rho_0) \in H^\beta(Q) \) for some \( Q > 0 \) and \( \beta > 0 \). Then for both problems on the graph

(i) For the power priors based on (3.19)–(3.20) the posterior contracts around the truth at the rate \((np_n)^{-\beta/(2\beta+r)}\) for both missing mechanisms under the additional restriction \( \beta \leq \alpha + r/2 \).

(ii) For the exponential priors based on (3.24)–(3.25) the posterior contracts around the truth at the rate \((np_n)^{-\beta/(2\beta+r)}(\log n)^{k'}\) for both missing mechanisms for some \( k' > 0 \) and for all \( \beta > 0 \).

Since the developed priors do not use knowledge of the regularity \( \beta \) of the true function, we again obtained rate-adaptive results with the same remarks about the log factor as in the full observation case. For the power prior the range of regularities that we can adapt to is bounded by \( \alpha + r/2 \), where \( \alpha \) is the hyperparameter describing the “baseline regularity” of the prior. In the case of the exponential prior the range is unbounded. This comes at the modest cost of having an additional logarithmic factor in the rate. Lower bounds for the rates of estimation are not known for the present situation. However, a particular case of regression on a path graph can be related to the case of the regression problem on the \([0, 1]\) interval with missing data, studied in Efroimovich [2011]. The minimax rate obtained in that paper is equal to \((np_n)^{-\beta/(2\beta+1)}\), which coincides with our rates for \( r = 1 \). Thus, it appears that the rates cannot be fundamentally improved. However, a detailed investigation of lower bounds in this setting is necessary to obtain proper insight into this matter.
3.5 Proofs

Recall that we identify functions on the graph with vectors in $\mathbb{R}^n$. In both cases we have that given $c$, the random vector $f$ is a centred $n$-dimensional Gaussian random vector. We view this as a Gaussian random element in the space $(\mathbb{R}^n, \| \cdot \|_n)$. The corresponding RKHS $\mathbb{H}^c$ is the entire space $\mathbb{R}^n$, and the corresponding RKHS-norm is given by

$$\| h \|_{\mathbb{H}^c}^2 = h^T \Sigma_c^{-1} h,$$

where $\Sigma_c$ is the covariance matrix of $f|c$. (See e.g. van der Vaart and van Zanten [2008b] for the definition and properties of the RKHS.) Recall that the $\psi_i$ are the eigenfunctions of $L$, normalised with respect to the norm $\| \cdot \|_n$. They are then also eigenfunctions of $\Sigma_c^{-1}$ in both cases. We denote the corresponding eigenvalues by $\mu_i$.

The Gaussian $N(0, \Sigma_c)$ admits the series representation

$$\sum \zeta_i \psi_i / \sqrt{n \mu_i},$$

where $\zeta_1, \ldots, \zeta_n$ are standard normal variables. In particular the functions $\psi_i / \sqrt{n \mu_i}$ form an orthonormal basis of the RKHS $\mathbb{H}^c$. Hence, the ordinary $\| \cdot \|_n$-norm and the RKHS-norm of a function $h$ with expansion $h = \sum h_i \psi_i$ are given by

$$\| h \|_n^2 = \sum_{i=0}^{n-1} h_i^2, \quad \| h \|_{\mathbb{H}^c}^2 = n \sum_{i=0}^{n-1} \mu_i h_i^2.$$ (3.31)

We denote the unit ball of the RKHS by $\mathbb{H}_1^c = \{ h \in \mathbb{H}^c : \| h \|_{\mathbb{H}^c} \leq 1 \}$.

3.5.1 Proof of Theorem 3.2.1

In this case $\Sigma_c^{-1} = ((n/c)^{2/r} (L + n^{-2} I))^{\alpha + r/2}$ is the precision matrix of $f$ given $c$ and the eigenvalues of $\Sigma_c^{-1}$ are given by

$$\mu_i = \left( \left( \frac{n}{c} \right)^{2/r} \left( \lambda_i + \frac{1}{n^2} \right) \right)^{\alpha + r/2}.$$ (3.32)

3.5.1.1 Proof of (3.4)

By Lemma 5.3 of van der Vaart and van Zanten [2008b], it follows from Lemmas 3.5.1 and 3.5.2 ahead that under the conditions of the theorem and for $\varepsilon = \varepsilon_n = n^{-\beta/(r+2\beta)}$ and $c = c_n$ satisfying $\sqrt{n \varepsilon_n^{(\beta-\alpha)/\beta}} \leq c_n^{(\alpha+2r)/r} \leq 2 \sqrt{\varepsilon_n^{(\beta-\alpha)/\beta}}$, we have

$$- \log P(\| f - f_0 \|_n | c) \lesssim \varepsilon_n^{-r/\beta}.$$
3. Bayesian function estimation on large graphs

By conditioning, it is then seen that
\[
\mathbb{P}(\|f - f_0\|_n < \varepsilon_n) \geq e^{-K_0\varepsilon_n^{-r/\beta}} \int_{(\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta})^{r/(\alpha+r/2)}}^{(2\sqrt{n}\varepsilon_n^{(\beta-\alpha)/\beta})^{r/(\alpha+r/2)}} e^{-x} dx \geq e^{-K_1\varepsilon_n^{-r/\beta}},
\]
for constants \(K_0, K_1 > 0\).

**Lemma 3.5.1.** For \(n\) large enough and \(\varepsilon > 0\) and \(\varepsilon\sqrt{n}/c^{(\alpha+r/2)/r}\) small enough,
\[
- \log \mathbb{P}(\|f\|_n \leq \varepsilon \mid c) \lesssim \left(\frac{c^{(\alpha+r/2)/r}}{\varepsilon \sqrt{n}}\right)^{\frac{\varepsilon}{\alpha}}.
\]

**Proof.** By the series representation (3.30) we have
\[
\mathbb{P}(\|f\|_n \leq \varepsilon \mid c) = \mathbb{P}\left(\sum_{i \leq i_0} \frac{\zeta_i^2}{n\mu_i} \leq \varepsilon^2\right).
\]
Recall from Section 2.2.2 that we can assume without loss of generality that we have the lower bounds
\[
\lambda_i \geq C_1\left(\frac{1}{n}\right)^2, \quad 1 \leq i \leq i_0, \quad (3.33)
\]
\[
\lambda_i \geq C_1\left(\frac{1}{n}\right)^{2/r}, \quad i > i_0. \quad (3.34)
\]
These translate into lower bounds for the \(\mu_i\) from which it follows that for \(\varepsilon > 0\),
\[
\mathbb{P}(\|f\|_n^2 \leq 2\varepsilon^2 \mid c) \geq \mathbb{P}\left(\sum_{i \leq i_0} \frac{\zeta_i^2}{n\mu_i} \leq \varepsilon^2, \sum_{i > i_0} \frac{\zeta_i^2}{n\mu_i} \leq \varepsilon^2\right)
\]
\[
\geq \mathbb{P}\left(\sum_{1 < i \leq i_0} \frac{\zeta_i^2}{n\mu_i} \leq (C_1 p e^{-2p/r} n^{(2\alpha+2r-2pr)/r})\varepsilon^2\right) \mathbb{P}\left(\sum_{i > i_0} \frac{\zeta_i^2}{n^{2p/r}} \leq C_1 p e^{-2p/r} n\varepsilon^2\right),
\]
where we write \(p = \alpha + r/2\). By Corollary 4.3 from Dunker et al. [1998], the last factor in the last line is bounded form below by
\[
\exp\left(\right. - \text{const} \times \left( e^{-p/r} \varepsilon \sqrt{n}\right)^{-r/\alpha} \left.),
\]
provided \(e^{-p/r}\varepsilon \sqrt{n}\) is small enough. By the triangle inequality and independence, the first factor is bounded from below by
\[
\left(\mathbb{P}(\|\zeta_1\| \leq i_0^{1/2} C_1^{p/2} e^{-p/r} n^{(\alpha+r-pr)/r} \varepsilon)\right)^{i_0}.
\]
Since \(r \geq 1\), we have \(e^{-p/r} n^{(\alpha+r-pr)/r} \varepsilon = O(e^{-p/r} \varepsilon \sqrt{n})\). Hence, for \(e^{-p/r} \varepsilon \sqrt{n}\) small
enough the probability is further bounded from below by
\[ \text{const} \times \left( c^{-p/r} n^{(\alpha + r - pr)/r \varepsilon} \right)^{i_0}. \]

Combining the bounds for the separate factors we find that, for \( c^{-p/r} \varepsilon \sqrt{n} \) small enough,
\[ -\log \mathbb{P}(\|f\|_n^2 \leq 2 \varepsilon^2 | c) \lesssim \log \left( \frac{c^{p/r}}{n^{(\alpha + r - pr)/r \varepsilon}} \right) + \left( \frac{c^{p/r}}{\varepsilon \sqrt{n}} \right)^{r/\alpha}. \]

Since \( r \geq 1 \) the first term on the right is smaller than a constant times the second one if \( c^{-p/r} \varepsilon \sqrt{n} \) is small enough.

\[ \text{Lemma 3.5.2.} \]
Let \( f \in H^\beta(Q) \) for \( \beta \leq \alpha + r/2 \). For \( \varepsilon > 0 \) such that \( \varepsilon \to 0 \) as \( n \to \infty \) and \( 1/\varepsilon = o(n^{\beta/r}) \) and \( n \) large enough,
\[ \inf_{h \in \mathcal{H}^c} \|h - f\|_{n, \leq \varepsilon} \|h\|_{H^c n}^2 \lesssim n c^{-(2\alpha + r)}/\varepsilon - 2(\alpha - \beta)/r. \] (3.35)

\[ \text{Proof.} \] We use the expansion \( f = \sum f_i \psi_i \), with \( \psi_i \) the orthonormal eigenfunctions of the Laplacian. In terms of the coefficients the smoothness assumption reads
\[ \sum (1 + n^{2\beta/r} \lambda_i^\beta) f_i^2 \leq Q^2. \] Now for \( I \) to be determined below, consider \( h = \sum_{i \leq I} f_i \psi_i \).

In view of (3.33)–(3.34) we have, for \( I \) large enough,
\[ \|h - f\|_n^2 = \sum_{i > I} f_i^2 \leq \frac{Q^2}{1 + n^{2\beta/r} \lambda_i^\beta} \leq Q^2 C_1^{-\beta} I^{-2\beta/r}. \]

Setting \( I = \text{const} \times \varepsilon^{-r/\beta} \), we get \( \|h - f\|_n \leq \varepsilon \). By (3.31), the RKHS-norm of \( h \) satisfies
\[ \|h\|_{H^c n}^2 = n \sum_{i \leq I} \left( (n/c)^{2/r} (\lambda_i + n^{-2}) \right)^{\alpha + r/2} f_i^2 \]
\[ \lesssim n c^{-2p/r} Q^2 + c^{-2p/r} Q^2 n^{2 + 2(\alpha - \beta)/r} \lambda_i^\beta. \]

The condition on \( \varepsilon \) ensures that for the choice of \( I \) made above and \( n \) large enough, \( i_0 \leq I \leq \kappa n \). Hence, by (3.33)–(3.34), \( \|h\|_{H^c n}^2 \) is bounded by a constant times the right-hand side of (3.35).

\[ \text{3.5.1.2 Proof of (3.5) and (3.6)} \]
Define \( B_n = M_n \mathbb{H}_1^n + \varepsilon_n \mathbb{B}_1 \), where \( \mathbb{B}_1 \) is the unit ball of \((\mathbb{R}^n, \|\cdot\|_n)\), \( \varepsilon_n = n^{-\beta/(r+2\beta)} \) again and \( c_n, M_n \) are the sequences to be determined below. By Lemma 3.5.3 we have
\[ \log N(2\varepsilon_n, B_n, \|\cdot\|_n) \leq \log N(\varepsilon_n/M_n, \mathbb{H}_1^n, \|\cdot\|_n) \lesssim c_n \left( \frac{M_n}{\varepsilon_n \sqrt{n}} \right)^{\frac{r}{p}}. \]
where \( p = \alpha + r/2 \) again. For \( M_n = M \sqrt{n\varepsilon_n^2} \) and \( c_n^{p/r} = N \sqrt{n\varepsilon_n^{(\beta-\alpha)/\beta}} \) this is bounded by a constant times \( n\varepsilon_n^2 \), which proves (3.6).

It remains to show that for given \( K_2 > 1 \), the constants \( M \) and \( N \) can be chosen such that (3.5) holds. We have

\[
\mathbb{P}(f \notin B_n) \leq \int_0^{c_n} \mathbb{P}(f \notin M_nB_1^n + \varepsilon_nB_1^n \mid c) e^{-c} dc + \int_{c_n}^{\infty} e^{-c} dc.
\]

For \( c \leq c_n \) we have the inclusion \( B_1^n \subseteq B_1^n \). Hence, by the Borell–Sudakov inequality, we have for \( c \leq c_n \) that

\[
\mathbb{P}(f \notin B_n \mid c) \leq \mathbb{P}(f \notin M_nB_1^n + \varepsilon_nB_1^n \mid c)
\]

\[
\leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|f\|_n \leq \varepsilon_n \mid c) + M_n))
\]

\[
\leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|f\|_n \leq \varepsilon_n \mid c_n) + M_n)),
\]

where \( \Phi \) is the distribution function of the standard normal distribution. By Lemma 3.5.1 the small ball probability in this expression is for \( c_n^{p/r} = N \sqrt{n\varepsilon_n^{(\beta-\alpha)/\beta}} \) bounded from below by \( \exp(-K\varepsilon_n^{-r/\beta}) \) for some \( K > 0 \). Using the bound \( \Phi^{-1}(y) \geq -((5/2)\log(1/y))^{1/2} \) for \( y \in (0, 1/2) \), it follows that for \( c \leq c_n \),

\[
\mathbb{P}(f \notin B_n \mid c) \leq 1 - \Phi(M_n - K^r\varepsilon_n^{-r/(2\beta)})
\]

for some \( K' > 0 \). For \( M_n \) a large enough multiple of \( \varepsilon_n^{-r/(2\beta)} \), this probability is bounded by \( \exp(-K_2\varepsilon_n^{-r/\beta}) = \exp(-K_2n\varepsilon_n^2) \).

**Lemma 3.5.3.** For \( n \) large enough and \( c, \varepsilon > 0 \) we have

\[
\log N(\varepsilon, \| \cdot \|_n) \lesssim c\left(\frac{1}{\varepsilon \sqrt{n}}\right)^{\frac{r}{\alpha+r/2}}.
\]

**Proof.** By expanding the RKHS elements in the eigenbasis of the Laplacian and taking into account the relations (3.31) we see that the problem is to bound the entropy \( \log N(\varepsilon, A, \| \cdot \|) \), where

\[
A = \left\{ x \in \mathbb{R}^n : n \sum_{i=0}^{n-1} ((n/c)^{2/r}(\lambda_i + n^{-2})^{\alpha+r/2}x_i^2 \leq 1 \right\}.
\]

Using the bounds (3.33)–(3.34), it follows that with \( p = \alpha + r/2 \) and \( R = c^{p/r}n^{-(\alpha+r)/r} \) we have the inclusions

\[
A \subseteq \left\{ x \in \mathbb{R}^n : \sum_{i=0}^{n-1} \lambda_i^{p}x_i^2 \leq R^2 \right\}
\]

\[
\subseteq \left\{ x \in \mathbb{R}^n : \sum_{i \leq i_0} x_i^2 \leq C_1^{-p}n^{2p}R^2, \sum_{i > i_0} x_i^{2p/r}x_i^2 \leq C_1^{-p}n^{2p/r}R^2 \right\}.
\]

58
3.5. Proofs

By using the well-known entropy bounds for balls in $\mathbb{R}^i_0$ and ellipsoids in $\ell^2$ we deduce from this that for $\varepsilon > 0$,

$$\log N(2\varepsilon, A, \| \cdot \|) \lesssim \log \left( \frac{n^p R}{\varepsilon} \right) + \left( \frac{n^p R}{\varepsilon} \right)^{r/p} \lesssim \left( \frac{n^p R}{\varepsilon} \right)^{r/p}.$$ 

The proof is completed by recalling the expressions for $p$ and $R$. 

3.5.2 Proof of Theorem 3.2.2

In this case the eigenvalues of $\Sigma^{-1} c$ are given by

$$\mu_i = n^{-1} e^{(n/c)^2/r} \lambda_i.$$ 

3.5.2.1 Proof of (3.9)

By Lemma 5.3 of van der Vaart and van Zanten [2008b], it follows from Lemmas 3.5.4 and 3.5.5 ahead that under the conditions of the theorem and for $\varepsilon = \varepsilon_n = (n/\log^{1+r/2} n)^{-\beta/(r+2\beta)}$ and $n\varepsilon^2/\log^{1+r/2} n \leq c \leq 2n\varepsilon^2/\log^{1+r/2} n$, we have

$$-\log P(\|f - f_0\|_n \leq \varepsilon | c) \lesssim c \log^{1+r/2} \frac{c}{\varepsilon^2} + e^{Kc^{-2/r} \varepsilon^{-2/\beta}} \lesssim n\varepsilon^2.$$ 

By conditioning, similar as in the previous case, we find that $-\log P(\|f - f_0\|_n \leq \varepsilon) \lesssim n\varepsilon^2$ as well.

**Lemma 3.5.4.** If $\varepsilon \to 0$, $c$ is bounded away from 0 and $c/\varepsilon^2 \to \infty$, then

$$-\log P(\|f\|_n \leq \varepsilon | c) \lesssim c \log^{1+r/2} \frac{c}{\varepsilon^2}.$$ 

**Proof.** Again the series representation of the Gaussian law of $f | c$ gives $P(\|f\|_n \leq \varepsilon | c) = P(\sum \xi_i^2 \leq \varepsilon^2)$, where the $\xi_i$ are independent standard normal random variables. By the lower bounds (3.33)–(3.34), it follows that

$$P(\sum \xi_i^2 \leq \varepsilon^2) \geq P\left( \sum_{i \leq i_0} e^{-C_1 n^{(2-2\gamma)/r} \varepsilon^{2/r}} \xi_i^2 \leq \varepsilon^2 \right) P\left( \sum_{i \geq 1} e^{-C_1 n^{(2-2\gamma)/r} \varepsilon^{2/r}} \xi_i^2 \leq \varepsilon^2 \right).$$ 

The first probability in the last line is bounded from below by

$$\left( P(\xi_1 < e^{-1/2} c^{(1/2)} C_1 n^{(2-2\gamma)/r} \varepsilon^{2/r}) \right)^{i_0}.$$ 

Since the quantity on the right of the inequality in this probability becomes arbitrarily small under the conditions of the lemma, this is further bounded from below by a constant times $\varepsilon^{i_0} \exp(i_0((1/2)C_1 n^{(2-2\gamma)/r} \varepsilon^{2/r})).$
3. Bayesian function estimation on large graphs

For the second probability we use Theorem 6.1 of Li and Shao [2001]. This asserts that if $a_k > 0$ and $\sum a_k < \infty$, then as $\varepsilon \to 0$

$$
\mathbb{P}\left(\sum a_i \xi_i^2 \leq \varepsilon^2\right) \approx \frac{1}{\sqrt{4\pi \sum (a_i \gamma_a)^2}} e^{\varepsilon^2 \gamma_a - (1/2) \sum \log(1+2a_i \gamma_a)},
$$

(3.37)

where $\gamma_a = \gamma_a(\varepsilon)$ is uniquely determined, for $\varepsilon > 0$ small enough, by the equation

$$
\varepsilon^2 = \sum \frac{a_i}{1 + 2a_i \gamma_a}.
$$

(3.38)

We apply (3.37) with $a_i = \exp(-C_1 (i/c)^{2/r})$.

We first determine bounds for $\gamma_a$. Note that in our case the terms in the sum $S$ on the right of (3.38) are decreasing in $i$. It follows that we have the bounds

$$
\int_1^\infty \frac{1}{e^{C_1(x/c)^{2/r} + 2\gamma_a}} dx \leq S \leq \int_0^\infty \frac{1}{e^{C_1(x/c)^{2/r} + 2\gamma_a}} dx.
$$

A change of variables shows that the integral on the right equals

$$
\frac{cr}{2C_1^{r/2}} \int_0^\infty \frac{t^{r/2-1}}{e^t + 2\gamma_a} dt = \frac{c-r \Gamma(r/2)}{4\gamma_a C_1^{r/2}} \text{Li}_{r/2}(-2\gamma_a),
$$

where $\text{Li}_s(z)$ denotes the polylogarithm. By Wood [1992],

$$
\frac{\text{Li}_{r/2}(-2\gamma_a)}{\log^{r/2} 2\gamma_a} \to -\frac{1}{\Gamma(r/2 + 1)}
$$

as $\gamma_a \to \infty$. Hence for large $\gamma_a$, we have the upper bound $S \leq \text{const} \times c \gamma_a^{-1} \log^{r/2} \gamma_a$.

It is easily seen that we have a lower bound of the same order, so that

$$
\varepsilon^2 \approx \frac{c \log^{r/2} \gamma_a}{\gamma_a}.
$$

Under our condition that $\varepsilon^2/c \to 0$ this holds if and only if

$$
\gamma_a \approx \frac{c}{\varepsilon^2} \log^{r/2} \frac{c}{\varepsilon^2}.
$$

Next we consider the sums appearing on the right of (3.37). To bound $\sum \log(1 + 2a_i \gamma_a) \leq \sum \log(1 + 2\exp(-C_1 (i/c)^{2/r}) \gamma_a)$ we consider the index $I = c(\log \gamma_a/C_1)^{r/2}$, which is determined such that $a_I \gamma_a = 1$. Note that for $m > 0$, we have $a_m \gamma_a = a_I^{m^{2/r}} \gamma_a = a_I^{1-m^{2/r}}$. We first split up the sum, writing

$$
\sum \log(1 + 2a_i \gamma_a) = \sum_{i < I} \log(1 + 2a_i \gamma_a) + \sum_{i \geq I} \log(1 + 2a_i \gamma_a)
$$

60
3.5. Proofs

The first sum on the right is bounded by a multiple of $I \log \gamma_a$. The second one we split into blocks of length $I$. This gives

$$\sum_{i \geq I} \log(1 + 2a_i \gamma_a) \leq I \sum_{m \geq 1} \log(1 + 2\gamma_a^{1-m/2}) \lesssim I.$$  

Hence, we have $\sum \log(1 + 2a_i \gamma_a) \lesssim c \log(1 + 2a_i \gamma_a)$. For the other sum appearing in (3.37) we have

$$\sum \left( \frac{2a_i \gamma_a}{1 + 2a_i \gamma_a} \right) \leq \sum \frac{2a_i \gamma_a}{1 + 2a_i \gamma_a} = 2\gamma_a \varepsilon^2.$$  

The proof is completed by combining all the bounds we have found. 

Lemma 3.5.5. Suppose that $f \in \mathcal{H}_\beta(Q)$ for some $\beta, Q > 0$. For $\varepsilon > 0$ such that $\varepsilon \to 0$ as $n \to \infty$ and $1/\varepsilon = o(n^{\beta/r})$ and $c > 0$,

$$\inf_{h \in \mathcal{H}_c : \|h - f\|_n \leq \varepsilon} \|h\|_{\mathcal{H}_c}^2 \lesssim e^{K c^{-2/\varepsilon} \varepsilon^{-2/\beta}}$$  

for $n$ large enough, where $K > 0$ is a constant.

Proof. We use an expansion $f = \sum f_i \psi_i$, with $\psi_i$ the orthonormal eigenfunctions of the Laplacian. We saw in the proof of Lemma 3.5.2 that if we define $h = \sum_{i \leq I} f_i \psi_i$ for $I = \text{const} \times \varepsilon^{-r/\beta}$, then $\|h - f\|_n \leq \varepsilon$. By (3.31), the RKHS-norm of $h$ satisfies in this case

$$\|h\|_{\mathcal{H}_c}^2 = \sum_{i \leq I} e^{(n/c)^2/r \lambda_i} f_i^2 \leq Q^2 e^{(n/c)^2/r \lambda_I}.$$  

The condition on $\varepsilon$ ensures that for the choice of $I$ made above and $n$ large enough, $i_0 \leq I \leq \kappa n$. Hence, by (3.33)–(3.34), $\|h\|_{\mathcal{H}_c}^2$ is bounded by a constant times the right-hand side of (3.39). 

3.5.2.2 Proof of (3.10)–(3.11)

Define $B_n = M_n \mathbb{H}_1^{c_n} + \varepsilon_n \mathbb{B}_1$, where $\varepsilon_n$ is as above and $M_n$ and $c_n$ are determined below.

For (3.10) we first note again that

$$\mathbb{P}(f \notin B_n) \leq \int_0^{c_n} \mathbb{P}(f \notin M_n \mathbb{H}_1^{c_n} + \varepsilon_n \mathbb{B}_1 \mid c) e^{-c} dc + \int_{c_n}^\infty e^{-c} dc.$$  

Exactly as in the proof of (3.5), the Borell–Sudakov inequality implies that for $c \leq c_n$,

$$\mathbb{P}(f \notin B_n \mid c) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|f\|_n \leq \varepsilon_n \mid c_n) + M_n)).$$  

61
3. Bayesian function estimation on large graphs

By Lemma 3.5.4 the small ball probability on the right is lower bounded by

\[ \exp \left( -Kc_n \log^{1+r/2} \frac{c_n}{\varepsilon_n^2} \right). \]

It follows that for \( c \leq c_n \),

\[ \mathbb{P}(f \notin B_n | c) \leq 1 - \Phi \left( M_n - K' \sqrt{c_n \log^{1+r/2} \frac{c_n}{\varepsilon_n^2}} \right) \]

for some \( K' > 0 \). For a given \( K_2 > 0 \), choosing \( M_n \) a large multiple of \((c_n \log^{1+r/2}(c_n/\varepsilon_n^2))^{1/2}\) we find that, for large \( n \),

\[ \mathbb{P}(f \notin B_n) \leq e^{-K'' c_n \log^{1+r/2} \frac{c_n}{\varepsilon_n^2}} + e^{-c_n} \leq 2e^{-c_n}. \]

If \( K_2 > 0 \) is a given constant, then for \( c_n \) a large enough multiple of \( n\varepsilon_n^2 \), this is bounded by \( \exp(-K_2 n \varepsilon_n^2) \).

For these choices of \( M_n \) and \( c_n \), Lemma 3.5.6 implies that the entropy satisfies, for any \( \tilde{\varepsilon}_n \geq \varepsilon_n \),

\[ \log N(2\tilde{\varepsilon}_n, B_n, \| \cdot \|) \leq \log N(2\varepsilon_n, B_n, \| \cdot \|) \lesssim c_n \left( \log \frac{M_n}{\varepsilon_n} \right)^{1+r/2}. \]

This proves that (3.11) holds for \( \tilde{\varepsilon}_n = \varepsilon_n \log^{1/2+r/4} n \).

**Lemma 3.5.6.** Let \( \varepsilon, c > 0 \) be such that \( c \log^{r/2}(1/\varepsilon) \to \infty \) as \( n \to \infty \). Then for \( n \) large enough,

\[ \log N(\varepsilon, H^c, \| \cdot \|) \lesssim c \log^{1+r/2} \left( \frac{1}{\varepsilon} \right). \]

**Proof.** We need to bound the metric entropy of the set

\[ A = \{ x \in \mathbb{R}^n : \sum_{i=0}^{n-1} e^{(n/c)^{2/r} \lambda_i x_i^2} \leq 1 \}, \]

relative to the Euclidean norm \( \| \cdot \| \). Set \( I = (2/C_1)^{r/2} c \log^{r/2}(1/\varepsilon) \). Under the assumption of the lemma this is larger than \( i_0 \), hence by (3.33)–(3.34) we have \( \exp(-(n/c)^{2/r} \lambda_I) \leq \varepsilon^2 \). It follows that if for \( x \in A \) we define the projection \( x^I \) by \( x^I = (x_1, \ldots, x_I, 0, 0, \ldots) \), then

\[ \| x - x^I \|^2 = \sum_{i > I} x_i^2 \leq e^{-(n/c)^{2/r} \lambda_I} \sum_{i > I} e^{(n/c)^{2/r} \lambda_i} x_i^2 \leq \varepsilon^2. \]

Moreover, we have \( \| x^I \| \leq 1 \). By the triangle inequality, it follows that if the points \( x_1, \ldots, x_N \) form an \( \varepsilon \)-net for the unit ball in \( \mathbb{R}^I \), then the points \( \overline{x}_1, \ldots, \overline{x}_N \) in \( \mathbb{R}^n \) obtained by appending zeros to the \( x_j \) form a \( 2\varepsilon \)-net for \( A \). Hence, \( N(2\varepsilon, A, \| \cdot \|) \lesssim \varepsilon^{-I} \). The proof is completed by recalling the expression for \( I \). 

\[ \blacksquare \]
3.5. Proofs

3.5.3 Proof of Theorem 3.4.1

The idea of the proof is to use general results for obtaining convergence rates for posterior distributions formulated in Theorems 1.5.1, 1.5.2. However, straightforward application of the theorems does not provide the desired results, hence we are going to reformulate the setting and the conditions of the theorem, use Theorems 1.5.1, 1.5.2, and then translate the obtained result back into original setting. In the sampling with replacement setting the observed data is identically distributed, which allows us to use Theorem 1.5.1. This is a special case of function estimation with random design and with less data available. For the coin flipping setting, we reformulate the problem in a such a way that it becomes a special case of the function estimation problem with fixed design. That allows us to use Theorem 1.5.2. The main challenge of the proof is to show that the distances employed in Theorems 1.5.1, 1.5.2 can be appropriately related to the $\| \cdot \|_n$-norm that is used in the conditions and conclusions of the theorem.

Observe that Theorems 1.5.1, 1.5.2 can only be applied for the case $k = 0$, when $\varepsilon_n = \tilde{\varepsilon}_n$. When $k \neq 0$, one has to use a generalised version of the Theorem 1.5.2, which can be directly deduced from Theorem 2.1 in Ghosal and van der Vaart [2001]. For simplicity from now on we assume that $k = 0$. Also, we omit the proof for the sampling with replacement missing mechanism in the classification setting, since this case can be handled in the same manner as the regression setting.

3.5.3.1 Proof for sampling with replacement in the regression setting

Since the observations are identically distributed, we can apply Theorem 1.5.1 for the following set up. Consider the space $\{1, \ldots, n\} \times \mathbb{R}$ with the product $\sigma$-algebra. Define the measure $\mu$ such that for $A \subseteq \{1, \ldots, n\}$ and $B \in \mathcal{B}(\mathbb{R})$

$$\mu(A \times B) = \sum_{i=1}^{n} \mathbf{1}_{i \in A} \lambda(B),$$

where $\lambda$ is the standard Lebesgue measure. Then for every $i = 1, \ldots, m_n$ the distribution of $(X_i, Z_i)$ is given by the probability measure $P_{f_0}$, where for a function $f \in \mathbb{R}^n$ on the vertices of the graph the probability measure $P_f$ is given by

$$P_f(A \times B) = \frac{1}{n} \sum_{j \in A} \int_B \varphi_{f(j), \sigma^2}(x) dx,$$
with $\varphi_{a,\sigma^2}$ the density of a Gaussian random variable with mean $a$ and variance $\sigma^2$. The Hellinger distance $d(f_1, f_2)$ between $P_{f_1}$ and $P_{f_2}$ satisfies

$$d^2(P_{f_1}, P_{f_2}) = \int_{\mathbb{R}} \left( \sqrt{\frac{dP_{f_1}}{d\mu}} - \sqrt{\frac{dP_{f_2}}{d\mu}} \right)^2 d\mu = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \left( \sqrt{\varphi_{f_1(i),\sigma^2(x)}} - \sqrt{\varphi_{f_2(i),\sigma^2(x)}} \right)^2 dx.$$  

Using the properties of $\varphi_{a,\sigma^2}$ we obtain that

$$d(P_{f_1}, P_{f_2}) \asymp \|f_1 - f_2\|_n. \quad (3.40)$$

In order to use Theorem 1.5.1 we consider the Kullback–Leibler divergence given by

$$-\int \log(dP_f/dP_{f_0}) dP_{f_0} = -\frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \log \frac{\varphi_{f(i),\sigma^2(x)}}{\varphi_{f_0(i),\sigma^2(x)}} \varphi_{f_0(i),\sigma^2(x)} dx \lesssim \|f - f_0\|_n^2.$$  

Additionally, one can see that

$$\int (\log(dP_f/dP_{f_0}))^2 dP_{f_0} \lesssim \|f - f_0\|_n^2.$$  

For more details on the computations see Ghosal and van der Vaart [2007]. Define the following sets in $\mathbb{R}^n$

$$B(f_0, \varepsilon) = \left\{ f : -\int \log(dP_f/dP_{f_0}) dP_{f_0} \leq \varepsilon^2, \int (\log(dP_f/dP_{f_0}))^2 dP_{f_0} \leq \varepsilon^2 \right\}.$$  

Using (3.16)-(3.18) we obtain that for some constant $K > 0$, and for all $L > 1$ the following holds

$$\Pi(B(f_0, \varepsilon)) \geq e^{-Kn_n(\varepsilon_n^*)^2}, \quad (3.41)$$

$$\Pi(f \notin B_n) \leq e^{-Ln_n(\varepsilon_n^*)^2}, \quad (3.42)$$

$$\log N(\varepsilon_n^*, B_n, d) \leq m_n(\varepsilon_n^*)^2. \quad (3.43)$$

Notice that (3.41)-(3.43) match one to one the conditions of Theorem 1.5.1. Thus, there exists an $M > 0$ such that

$$\Pi \left( f : d(P_f, P_{f_0}) \geq M\varepsilon_n^* | (X_1, Z_1), \ldots, (X_{m_n}, Z_{m_n}) \right) \rightarrow 0.$$  

Using (3.40) we derive the desired result.
3.5. Proofs

3.5.3.2 Proof for the coin flipping mechanism in the regression setting

We are going to apply Theorem 1.5.2 to the following set up. Consider the space 
\( R^* = R \cup \{\ast\} \) endowed with the \( \sigma \)-algebra generated by the Borel sets in \( R \) and the 
set \( \{\ast\} \). For every measurable subset \( B \) of \( R^* \) set

\[
\lambda^*(B) = \lambda(B \cap R) + \delta(B \cap \{\ast\}),
\]

where \( \delta \) is a counting measure and \( \lambda \) is the standard Lebesgue measure on \( R \). In 
this setting \( Z_1, Z_2, \ldots, Z_n \) are independent observations distributed according to 
the following densities (with respect to \( \lambda^* \))

\[
Z_i \sim g^*_{0, i} = p_n \phi_{f_0(i), \sigma^2} + (1 - p_n) \mathbb{1}_{\{\ast\}},
\]

where \( \phi_{a, \sigma^2} \) denotes a Gaussian density of \( N(a, \sigma^2) \). For a function \( f \in R^n \) on the 
vertices of the graph consider a vector of densities \( g^*_{f, i} \) by setting

\[
g^*_{f, i} = p_n \phi_{f(i), \sigma^2} + (1 - p_n) \mathbb{1}_{\{\ast\}}.
\]

The average Hellinger distance between such vectors of densities is given by

\[
\left( d^*_n(g^*_f, g^*_{f'}) \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \int_{R^*} \left( \sqrt{g^*_{f, i}} - \sqrt{g^*_{f', i}} \right)^2 d\lambda^* = \frac{p_n}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \left( \sqrt{\phi_{f(i), \sigma^2}} - \sqrt{\phi_{f'(i), \sigma^2}} \right)^2 d\lambda,
\]

where \( f, f' \in R^n \) are two functions on the graph. We obtain that

\[
d_n^*(g^*_f, g^*_f') \asymp \sqrt{p_n} \| f - f' \|_n. \tag{3.44}
\]

Additionally, we define the following sets

\[
B^*_n(f_0, \varepsilon) = \left\{ f \in R^n : \frac{1}{n} \sum_{i=1}^{n} K^*(g^*_{f_0, i}, g^*_{f, i}) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^{n} V^*(g^*_{f_0, i}, g^*_{f, i}) \leq C\varepsilon^2 \right\},
\]

where for two densities \( f^*, g^* \) defined on \( R^* \)

\[
K^*(f^*, g^*) = \int f^* \log(f^*/g^*) d\lambda^*,
\]

\[
V^*(f^*, g^*) = \int f^* (\log(f^*/g^*) - K(f^*, g^*))^2 d\lambda^*.
\]
We are going to show that the sets 
\[ G_n^* = \{ g_f^* = (g_{f,1}^*, \ldots, f_{1,n}^*), \text{for } f \in B_n \} \]

satisfy the following conditions

\[
\log N(\epsilon_n/36, G_n^*, d_n^*) \leq n\epsilon_n^2, \quad (3.45)
\]

\[
\frac{\Pi(f \in \mathbb{R}^n : g_f^* \notin G_n^*)}{\Pi(B_n(f_0, \epsilon_n))} = o\left( e^{-2n\epsilon_n^2} \right), \quad (3.46)
\]

\[
\Pi(B_n(f_0, \epsilon_n)) \geq e^{-n\epsilon_n^2}. \quad (3.47)
\]

Then by Theorem 1.5.2 we have 
\[ \Pi(f : d_n^*(g_f^*, g_{f_0}^*) \geq M\epsilon_n \mid X^{(n)}) \xrightarrow{P_{f_0}} 0 \] for some \( M > 0 \). The statement of the theorem follows from (3.44).

Recall the definitions of the discrepancy measures \( K(\cdot, \cdot) \) and \( V(\cdot, \cdot) \) for two probability densities on \( \mathbb{R} \):

\[
K(f, g) = \int f \log(f/g) d\lambda,
\]

\[
V(f, g) = \int (f \log(f/g) - K(f, g))^2 d\lambda.
\]

It is known (see for example Ghosal and van der Vaart [2007]) that there exists constants \( A, B \) such that

\[
K^*(g_{f_0,0}, g_{f_0,1}) = p_n K(\varphi_{f_0,0}, \varphi_{f_0,1}) \leq A (f_0(i) - f(i))^2,
\]

\[
V^*(g_{f_0,0}, g_{f_0,1}) = p_n V(\varphi_{f_0,0}, \varphi_{f_0,1}) \leq B (f_0(i) - f(i))^2.
\]

Then using (3.16) we obtain

\[
-\log \Pi(B_n^*(f_0, \epsilon_n, k)) \gtrsim -\log \Pi( f : \|f - f_0\|^2_n \leq p_n^{-1/2} \epsilon_n^2) \leq Kn\epsilon_n^2,
\]

since \((\epsilon_n^*)^2 = p_n^{-1/2} \epsilon_n^2\). Moreover, by (3.44) and (3.18) the entropy can be bounded as follows

\[
\log N(\epsilon_n, G_n^*, d_n^*) \approx \log N(p_n^{-1/2} \epsilon_n, B_n, \| \cdot \|_n) \lesssim n\epsilon_n^2.
\]

Also, by (3.17)

\[
\Pi(f \in \mathbb{R}^n : g_f^* \notin G_n^*) = \Pi(f \notin B_n) \leq e^{-Ln\epsilon_n^2}.
\]

Observe that (3.48)–(3.50) match one to one the conditions (3.45)-(3.47). It concludes the proof of the theorem for this case.

### 3.5.3.3 Proof for coin flipping mechanism in the classification setting

We prove the result using Theorem 1.5.2 in the following setting. Consider the space \( \{0, 1, \star\} \) with the counting measure \( \lambda_c^* \). Notice that \( Z_1, \ldots, Z_n \) are independent
3.5. Proofs

observations distributed according to the following densities

\[ U_i \sim \rho_0^*(i) = p_n \left[ \rho_0(i) \mathbb{1}(1) + (1 - \rho_0(i)) \mathbb{1}(0) \right] + (1 - p_n) \mathbb{1}(\ast). \]

For a function \( f \in \mathbb{R}^n \) on the vertices of the graph define densities \( \rho_f, \rho_f^* \) such that \( \rho_f(i) = \Psi(f(i)) \) and

\[ \rho_f^*(i) = p_n \left[ \rho_f(i) \mathbb{1}(1) + (1 - \rho_f(i)) \mathbb{1}(0) \right] + (1 - p_n) \mathbb{1}(\ast). \]

Then the average Hellinger distance is given by

\[
(d_n^*(\rho_f^*, \rho_{f'}^*))^2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( \sqrt{p_n \rho_f(i)} - \sqrt{p_n \rho_{f'}(i)} \right)^2 + \left( \sqrt{p_n(1 - \rho_f(i))} - \sqrt{p_n(1 - \rho_{f'}(i))} \right)^2 \right] \approx p_n \| \sqrt{\rho_f} - \sqrt{\rho_{f'}} \|^2_n.
\]

Thus, using the properties of the link function \( \Psi \) we obtain

\[
(d_n^*(\rho_f^*, \rho_{f'}^*))^2 \approx p_n \| \rho_f - \rho_{f'} \|^2_n. \tag{3.51}
\]

Moreover, for every \( v_1, v_2 \in \mathbb{R}^n \)

\[ ||\sqrt{\Psi(v_1)} - \sqrt{\Psi(v_2)}||_n \lesssim ||v_1 - v_2||_n. \]

Hence, we have the following relation between \( d_n^*(\rho_f^*, \rho_{f'}^*) \) and \( ||f - f'||_n \)

\[ (d_n^*(\rho_f^*, \rho_{f'}^*))^2 \lesssim p_n ||f - f'||_n^2. \]

Applying a similar argument and Lemma 2.5.3 we get that

\[
\frac{1}{n} \sum_{i=1}^{n} K^*(\rho_{f_0}^*(i), \rho_f^*(i)) \lesssim p_n \frac{1}{n} \sum_{i=1}^{n} K(\Psi(f_0(i)), \Psi(f(i))) \lesssim p_n \|f_0 - f\|^2_n,
\]

where \( f_0 = \Psi^{-1}(\rho_0) \). Additionally,

\[
\frac{1}{n} \sum_{i=1}^{n} V^*(\rho_{f_0}^*(i), \rho_f^*(i)) \lesssim p_n \|f_0 - f\|^2_n.
\]

Define the sets \( G_n^* = \left\{ \rho_f^*, \text{ for } f \in B_n \right\} \). Also, consider the following sets

\[
B_n^*(\rho_0, \epsilon) = \Psi \left( \left\{ f : \frac{1}{n} \sum_{i=1}^{n} K^*(\rho_{f_0}^*(i), \rho_f^*(i)) \leq \epsilon^2, \frac{1}{n} \sum_{i=1}^{n} V^*(\rho_{f_0}^*(i), \rho_f^*(i)) \leq C\epsilon^2 \right\} \right).
\]

67
3. Bayesian function estimation on large graphs

By (3.16)–(3.18) we have

\[
\log N(\varepsilon, G_n^*, d_n^*) \lesssim n\varepsilon_n^2,
\]

\[
\frac{\Pi(f \in \mathbb{R}^n : \rho_f^* \notin G_n^*)}{\Pi(B_n^*(\rho_0, \varepsilon_n))} \lesssim e^{-(L-K)n\varepsilon_n^2},
\]

\[
\Pi(B_n^*(\rho_0, \varepsilon_n)) \gtrsim e^{-Kn\varepsilon_n^2}.
\]

Then using Theorem 1.5.2 we obtain

\[
\Pi(f \in \mathbb{R}^n : d_n^*(\rho_f^*, \rho^*) \geq M\varepsilon_n | X^{(n)}) \xrightarrow{P} 0.
\]

The desired result follows from (3.51).

3.5.4 Proofs of Theorem 3.4.2 and Theorem 3.4.3

The proofs of Theorem 3.4.2 and 3.4.3 go along the lines of the proof of Theorem 3.2.1 and 3.2.2 with slight modifications due to the differences in the prior and in the required result. We only provide the outline of the proof of Theorem 3.4.2, since Theorem 3.4.3 can be derived in a similar manner.

3.5.4.1 Proof of (3.21)

We know from Lemma 3.5.1 that for large \( n \) and when \((\varepsilon\sqrt{n})/c^{(\alpha+r/2)/r}\) is small enough, there is a following bound on the centred small balls

\[
-\log P(\|f\|_n \leq \varepsilon) \lesssim \left(\frac{c^{(\alpha+r/2)/r}}{\varepsilon\sqrt{n}}\right)^{r/\alpha}.
\]

Note that the condition is satisfied for \( n \) large enough, since

\[
\frac{\varepsilon_n^* \sqrt{n}}{c^{(\alpha+r/2)/r}} \approx (\varepsilon_n^*)^{\alpha/\beta} \to 0.
\]

To handle the bias, we recall from Lemma 3.5.2 that for \( \varepsilon \to 0 \) such that \( \varepsilon^{-1} = o(n^{\beta/r}) \) we obtain

\[
\inf_{h \in \mathbb{H}^c : \|h - f\|_n \leq \varepsilon} \|h\|_{\mathbb{H}^c}^2 \lesssim nc^{-(2\alpha+r)/r} \varepsilon^{-(2(\alpha-\beta)+r)/\beta}.
\]

Then for \( c \) such that \( c^{(\alpha+r)/r} \approx \sqrt{n}(\varepsilon_n^*)^{1-\alpha/\beta} \) we have that

\[
-\log P(\|f - f_0\|_n \leq \varepsilon^*_n | c) \lesssim n\varepsilon_n^2.
\]

By integrating \( c \) out we deduce that for some constant \( K_1 > 0 \)

\[
\mathbb{P}(\|f - f_0\|_n \leq \varepsilon^*_n) \gtrsim e^{-K_1 n\varepsilon_n^2}.
\]
3.5. Proofs

3.5.4.2 Proof of (3.22)-(3.23)

To prove (3.23) we recall that by Lemma 3.5.1 for small $\varepsilon$
\[ \log N(\varepsilon, \mathbb{H}_1^c, \| \cdot \|_n) \lesssim c(\varepsilon \sqrt{n})^{-r/(\alpha+r/2)}. \]

We use slightly altered sieves
\[ B_n = M_n \mathbb{H}_1^{c_n} + \varepsilon_n^* B_1, \]
where for $M$ and $C_0$ to be determined below
\[ c_n = C_0 n^{r/(2\alpha+r)} (\varepsilon_n^*)^{r(1-\alpha/\beta)/(\alpha+r/2)}, \]
\[ \varepsilon_n^* = (np_n)^{-\beta/(2\beta+r)}, \]
\[ M_n = M \sqrt{n \varepsilon_n}, \]
\[ \varepsilon_n = p_n^{r/2(2\beta+r)} n^{-\beta/(2\beta+r)}. \]

Then by Lemma 3.5.3 we get
\[ \log(2\varepsilon_n^*, B_n, \| \cdot \|_n) \lesssim c_n \left( \frac{M_n}{\sqrt{n \varepsilon_n^*}} \right)^{r/(\alpha+r/2)} \approx n \varepsilon_n^2. \]

In order to prove (3.22) we follow the steps of the proof of Theorem 3.4.2. First, we notice that for $c \leq c_n$ we have $\mathbb{H}_1^c \subset \mathbb{H}_1^{c_n}$. Then for $c \leq c_n$ it holds that
\[ \mathbb{P}(f \notin B_n | c) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|f\|_n \leq \varepsilon_n^* | c_n)) + M_n) \leq 1 - \Phi(M_n - K'(\varepsilon_n^*)^{-r/2\beta}). \]

We can set $M$ big enough, such that the righthand side is bounded by $e^{-K_2 n \varepsilon_n^2}$. Then we can choose $C_0$ such that
\[ \mathbb{P}(f \notin B_n) \leq \int_0^{c_n} \mathbb{P}(f \notin B_n | c) e^{-p_n^{r/(2\alpha+r)} c} dc + \int_{c_n}^{c} e^{-p_n^{r/(2\alpha+r)} c} dc \leq e^{-K_2 n \varepsilon_n^2}. \]
Chapter 4

Poisson intensity estimation

4.1 Introduction

Inhomogeneous Poisson processes are widely used models for counting the occurrence of certain events in a variety of applied areas. A typical task in applications is to learn the underlying intensity function of a Poisson process from a realised point pattern. In this chapter we consider nonparametric Bayesian approaches to this problem. These do not assume a specific parametric form of the intensity function and produce posterior distributions which do not only give an estimate of the intensity, e.g. through the posterior mean or mode, but also give a measure of the remaining uncertainty through the spread of the posterior.

Several papers have explored nonparametric Bayesian approaches in this setting. An early reference is Møller et al. [1998], who study log-Gaussian priors. Gugushvili and Spreij [2013] recently considered Gaussian processes (GP) combined with different, non-smooth link functions. Kernel mixtures priors are considered in Kottas and Sansó [2007]. Spline-based priors are used in DiMatteo et al. [2001] and Belitser et al. [2015].

The present study is motivated by a method that is not covered by earlier theoretical papers, namely the method of Adams et al. [2009]. These authors presented the first approach that is computationally fully nonparametric in the sense that it does not involve potentially inaccurate finite-dimensional approximations. The method involves a prior on the intensity that is a random multiple of a transformed GP. Both the hyperparameters of the GP and the multiplicative constant are endowed with priors as well, resulting in a hierarchical Bayes procedure (details in Section 4.2.3). Simulation experiments and real data examples in Adams et al. [2009] show that the method can give satisfactory results.

The aim of this chapter is to advance the theoretical understanding of the method of Adams et al. [2009], which they named “Sigmoidal Gaussian Cox Process” (SGCP). It is by now well known both from theory and practice that nonparametric Bayesian methods need to be tuned very carefully to produce good results. An
4. Poisson intensity estimation

unfortunate choice of the prior or incorrectly tuned hyperparameters can easily result in procedures that give misleading results or that make suboptimal use of the information in the training data. See for instance Diaconis and Freedman [1986b], or the more recent paper van der Vaart and van Zanten [2011] and the references therein.

A challenge in this problem (and in nonparametric function estimation in general) is to devise a procedure that avoids overfitting and underfitting. The difficulty is that the appropriate degree of “smoothing” depends on the (unknown) regularity of the intensity function according to which the data is generated. Indeed, intuitively it is clear that if the function is very smooth then to learn the intensity at a certain location we can borrow more information from neighbouring points than if it is very rough. Ideally we want to have a procedure that automatically uses the appropriate degree of smoothing, i.e. that adapts to regularity.

To address this issue theoretically we again take an asymptotic approach. We assume that we have \( n \) independent sets of training data, produced by Poisson processes, say, on the \( d \)-dimensional domain \( S = [0,1]^d \), with the same intensity function \( \lambda_0 : S \to [0,\infty) \). We aim to construct a learning procedure that achieves an optimal convergence rate, irrespective of the regularity level of the intensity. In the problem at hand it is known that if \( \lambda_0 \) has regularity \( \beta > 0 \), then the best rate that any procedure can achieve is of the order \( n^{-\beta/(d+2\beta)} \). See Kutoyants [1998] or Reynaud-Bouret [2003] for minimax results in the context of the Poisson process model considered in this chapter.

Note that the smoothness degree is unknown to us, so we can not use it in the construction of the procedure, but still we want that the posterior contracts around \( \lambda_0 \) at the rate \( n^{-\beta/(d+2\beta)} \), as \( n \to \infty \), if \( \lambda_0 \) is \( \beta \)-smooth. We prove that with appropriate priors on the hyperparameters, the SGCP approach of Adams et al. [2009] attains this optimal rate (up to a logarithmic factor). It does so for every regularity level \( \beta > 0 \), so it is fully rate-adaptive.

In order to study contraction rates for Gaussian and conditionally Gaussian priors we use the mathematical framework developed in van der Vaart and van Zanten [2008a] and van der Vaart and van Zanten [2009]. We also use an extended version of the general result for Bayesian inference for 1-dimensional Poisson processes from Belitser et al. [2015]. The reasoning is mainly similar to that of van der Vaart and van Zanten [2009]. However, due to the presence of a link function and a random multiplicative constant in the SGCP model (see Section 4.2 ahead) their results are not applicable in the present setting and additional mathematical arguments are required to derive the contraction rates.

The remainder of this chapter is organised as follows. In the next section we describe the Poisson process observation model and the SGCP prior model, which together determine a full hierarchical Bayesian model. The main result about the performance of the SGCP approach is presented and discussed in Section 4.3. Mathematical proofs are given in Section 4.4.
4.2 The SGCP model

4.2.1 Observation model

We assume we observe \( n \) independent copies \( N^1, \ldots, N^n \) of an inhomogeneous Poisson process with the intensity function \( \lambda_0 \) on the \( d \)-dimensional unit cube \( S = [0, 1]^d \) (adaptation to other domains is straightforward). Formally every \( N^i \) is a counting measure on subsets of \( S \). The intensity function is a (Lebesgue integrable) function \( \lambda_0 : [0, 1]^d \rightarrow [0, \infty) \) with the property that given \( \lambda_0 \), every \( N^j \) is a random counting measure on \([0, 1]^d\) such that \( N^j(A) \) and \( N^j(B) \) are independent if the sets \( A, B \subset [0, 1]^d \) are disjoint and the number of points \( N^j(B) \) falling in the set \( B \) has a Poisson distribution with mean \( \int_B \lambda_0(s) \, ds \). If we want to stress that the probabilities and expectations involving the observations \( N^j \) depend on \( \lambda_0 \), we use the notations \( \mathbb{P}_{\lambda_0} \) and \( \mathbb{E}_{\lambda_0} \), respectively. We note that instead of considering observations from \( n \) independent Poisson processes with intensity \( \lambda_0 \), one could equivalently consider observations from a single Poisson process with intensity \( n \lambda_0 \).

4.2.2 Prior model

The SGCP model introduced in Adams et al. [2009] postulates a-priori that the intensity function \( \lambda \) is of the form

\[
\lambda(s) = \lambda^* \sigma(g(s)), \quad s \in S, \tag{4.1}
\]

where \( \lambda^* > 0 \) is an upper bound on \( \lambda \), \( g \) is a GP indexed by \( S \) and \( \sigma \) is the sigmoid, or the logistic function on the real line, defined by \( \sigma(x) = (1 + e^{-x})^{-1} \). In the computational section of Adams et al. [2009] \( g \) is modelled as a GP with squared exponential covariance kernel and zero mean, with a prior on the length scale parameter. The hyperparameter \( \lambda^* \) is endowed with an independent gamma prior.

In the mathematical results presented here we allow a bit more flexibility in the choice of the covariance kernel of the GP, the link function \( \sigma \) and the priors on the hyperparameters. We assume that \( g \) is a zero-mean, homogenous GP with covariance kernel given in spectral form by

\[
\mathbb{E}g(s)g(t) = \int e^{-i\langle \xi, \ell(t-s) \rangle} \mu(\xi) \, d\xi, \quad s, t \in S, \tag{4.2}
\]

where \( \ell > 0 \) is an (inverse) length scale parameter and \( \mu \) is a spectral density on \( \mathbb{R}^d \) such that the map \( a \mapsto \mu(a\xi) \) on \((0, \infty)\) is decreasing for every \( \xi \in \mathbb{R}^d \) and that satisfies

\[
\int e^{\delta\|\xi\|} \mu(d\xi) < \infty \tag{4.3}
\]

for some \( \delta > 0 \) (the Euclidean inner product and norm in \( \mathbb{R}^d \) are denoted by \( \langle \cdot, \cdot \rangle \) and \( \| \cdot \| \), respectively). Note that, in particular, the centred Gaussian spectral
4. Poisson intensity estimation

density satisfies this condition and corresponds to the squared exponential kernel

$$\mathbb{E}g(s)g(t) = e^{-\ell^2\|t-s\|^2}.$$ 

We endow the length scale parameter $\ell$ with a prior with density $p_\ell$ on $[0, \infty)$ that satisfies the following bounds. We assume there exist positive constants $C_1, D_1, C_2, D_2$, nonnegative constants $p, q$, and every sufficiently large $x > 0$ such that

$$C_1 x^p \exp(-D_1 x^d \log^q x) \leq p_\ell(x) \leq C_2 x^p \exp(-D_2 x^d \log^q x). \quad (4.4)$$

This condition is, for instance, satisfied if $\ell^d$ has a gamma distribution, which is a common choice in practice. Note, however, that the technical condition (4.4) is only a condition on the tail of the prior on $\ell$. Furthermore, on the upper bound parameter $\lambda^*$ we put a prior satisfying an exponential tail bound. Specifically, we use a positive, continuous prior density $p_{\lambda^*}$ on $[0, \infty)$ such that for some $c_0, C_0, \kappa > 0$,

$$\int_{\lambda_0}^{\infty} p_{\lambda^*}(x) \, dx \leq C_0 e^{-c_0 \lambda^0} \quad (4.5)$$

for all $\lambda_0 > 0$. Note that this condition is fulfilled if we place a gamma prior on $\lambda^*$. Finally, we use a strictly increasing, infinitely smooth link function $\sigma : \mathbb{R} \to (0, 1)$ in (4.1) that satisfies

$$|\sqrt{\sigma(x)} - \sqrt{\sigma(y)}| \leq c|x - y| \quad (4.6)$$

for all $x, y \in \mathbb{R}$. This condition is in particular fulfilled for the sigmoid function employed by Adams et al. [2009]. It holds for other link functions as well, for instance for the cumulative distribution function of the standard normal distribution.

4.2.3 Full hierarchical model

With the assumptions made in the preceding section, the full hierarchical specification of the prior and observation model can then be summarised as follows:

$$\ell \sim p_\ell \quad \text{(satisfying (4.4))}$$
$$\lambda^* \sim p_{\lambda^*} \quad \text{(satisfying (4.5))}$$
$$g | \ell, \lambda^* \sim \text{GP with kernel given by (4.2)--(4.3)}$$
$$\lambda | g, \ell, \lambda^* \sim \text{defined by (4.1), with smooth } \sigma \text{ satisfying (4.6)}$$
$$N_1, \ldots, N_n | \lambda, g, \ell, \lambda^* \sim \text{independent Poisson processes with intensity } \lambda.$$

Note that under the prior several quantities are, by construction, independent. Specifically, $\ell$ and $\lambda_*$ are independent, and $g$ and $\lambda^*$ are independent.

The main results of the chapter concern the posterior distribution of the intensity function $\lambda$, i.e. the conditional distribution $\lambda | N_1, \ldots, N_n$. We denote the prior on $\lambda$ by $\Pi$ and the posterior by $\Pi(\cdot | N_1, \ldots, N_n)$. In this setting the Bayes’ formula
4.3 Main result

asserts that

\[ \Pi(\lambda \in B \mid N^1, \ldots, N^n) = \frac{\int_B p(N^1, \ldots, N^n \mid \lambda) \Pi(d\lambda)}{\int p(N^1, \ldots, N^n \mid \lambda) \Pi(d\lambda)}, \tag{4.7} \]

where the likelihood is given by

\[ p(N^1, \ldots, N^n \mid \lambda) = \prod_{i=1}^n e^{\int_S \lambda(x)N^i(dx) - \int_S (\lambda(x) - 1) dx} \]

(see, e.g., Kutoyants [1998]).

4.3 Main result

Consider the prior and observations model described in the preceding section and let \( \Pi(\cdot \mid N^1, \ldots, N^n) \) be the corresponding posterior distribution of the intensity function \( \lambda \).

The following theorem describes how quickly the posterior distribution contracts around the true intensity \( \lambda_0 \) according to which the data is generated. The rate of contraction depends on the smoothness level of \( \lambda_0 \). This is quantified by assuming that \( \lambda_0 \) belongs to the H"older space \( C^\beta[0, 1]^d \) for \( \beta > 0 \). By definition a function on \([0, 1]^d\) belongs to this space if it has partial derivatives up to the order \([\beta]\) and if the \([\beta]\)th order partial derivatives are all H"older continuous of the order \(\beta - [\beta]\). Here \([\beta]\) denotes the greatest integer strictly smaller than \(\beta\). The rate of contraction is measured in the \(L^2\)-distance between the square roots of intensities. This is a natural statistical metric in this problem, as it can be shown that in this setting the Hellinger distance between the models with intensity functions \(\lambda_1\) and \(\lambda_2\) is equivalent to \(\min \{\|\sqrt{\lambda_1} - \sqrt{\lambda_2}\|_2, 1\} \) (see Belitser et al. [2015]). Here \(\|f\|_2\) denotes the \(L^2\)-norm of a function on \(S = [0, 1]^d\), i.e. \(\|f\|_2^2 = \int_S f^2(s) ds\).

**Theorem 4.3.1.** Suppose that \(\lambda_0 \in C^\beta([0, 1]^d)\) for some \(\beta > 0\) and that \(\lambda_0\) is strictly positive. Then for all sufficiently large \(M > 0\),

\[ \mathbb{E}_{\lambda_0} \Pi(\lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \geq Mn^{-\beta/(d+2\beta)} \log \rho n \mid N^1, \ldots, N^n) \to 0 \tag{4.8} \]

as \(n \to \infty\), for some \(\rho > 0\).

The theorem asserts that if the intensity \(\lambda_0\) that generates the data is \(\beta\)-smooth, then, asymptotically, all the posterior mass is concentrated in (Hellinger) balls around \(\lambda_0\) with a radius that is up to a logarithmic factor of the optimal order \(n^{-\beta/(d+2\beta)}\). Since the procedure does not use the knowledge of the smoothness level \(\beta\), this indeed shows that the method is rate-adaptive, i.e. the rate of convergence adapts automatically to the degree of smoothness of the true intensity. Let us mention once again that the conditions of the theorem are in particular fulfills if in (4.1) parameter \(\lambda^*\) is gamma–distributed, \(\sigma\) is the sigmoid (logistic) function, and \(g\) is a squared exponential GP with length scale \(\ell\), where \(\ell^d\) is a gamma variable.
4. Poisson intensity estimation

4.4 Proofs

In this section we present the proof of Theorem 4.3.1. To prove the theorem we employ an extended version of a result from Belitser et al. [2015] that gives sufficient conditions for having (4.8) in the case $d = 1$, cf. their Theorem 1. Adaptation to the case of a general $d \in \mathbb{N}$ is straightforward. To state the result we need some (standard) notation and terminology. For a set of positive functions $\mathcal{F}$ we write $\mathcal{F}^c$ for its complement and $\sqrt{\mathcal{F}} = \{\sqrt{f}, f \in \mathcal{F}\}$. For $\varepsilon > 0$ and a norm $\| \cdot \|$ on $\mathcal{F}$, let $N(\varepsilon, \mathcal{F}, \| \cdot \|)$ be the minimal number of balls of radius $\varepsilon$ with respect to norm $\| \cdot \|$ needed to cover $\mathcal{F}$. The uniform norm $\|f\|_\infty$ of a function $f$ on $S$ is defined, as usual, as $\|f\|_\infty = \sup_{s \in S} |f(s)|$. The space of continuous functions on $S$ is denoted by $C(S)$.

Let $\Pi$ now be a general prior on the intensity function $\lambda$ and let $\Pi(\cdot | N_1, \ldots, N^n)$ be the corresponding posterior (4.7). As usual, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

**Theorem 4.4.1.** Assume that $\lambda_0$ is bounded away from 0. Suppose that for positive sequences $\delta_n, \delta_n \to 0$ such that $n(\delta_n \wedge \delta_n)^2 \to \infty$ as $n \to \infty$ and constants $c_1, c_2 > 0$, it holds that for all $L > 1$, there exist subsets $\mathcal{F}_n \subset C(S)$ and a constant $c_3$ such that

$$1 - \Pi(\mathcal{F}_n) \leq e^{-Ln\delta_n^2},$$

$$\Pi(\lambda : \|\lambda - \lambda_0\|_\infty \leq \delta_n) \geq c_1 e^{-nc_2\delta_n^2},$$

$$\log N(\delta_n, \sqrt{\mathcal{F}_n}, \| \cdot \|_2) \leq c_3 n\delta_n^2.$$  

Then for $\varepsilon_n = \delta_n \vee \delta_n$ and all sufficiently large $M > 0$,

$$\mathbb{E}_{\lambda_0} \Pi(\lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \geq M\varepsilon_n | N_1, \ldots, N^n) \to 0$$

as $n \to \infty$.

We note that this theorem has a form that is commonly encountered in the literature on contraction rates for nonparametric Bayes procedures. The so-called “prior mass condition” (4.10) requires that the prior puts sufficient mass near the true intensity function $\lambda_0$ according to which the data is generated. The “remaining mass condition” (4.9) and the “entropy condition” (4.11) together require that “most” of the prior mass should be concentrated on so-called “sieves” $\mathcal{F}_n$ that are not too large in terms of their metric entropy. The sieves grow as $n \to \infty$ and in the limit they capture all the posterior mass.

In the subsequent subsections we show that the prior defined in Section 4.2.3 fulfils the conditions of this theorem for $\delta_n = n^{-\beta/(2\beta+d)}(\log n)^{k_1}$ and $\delta_n^2 = n^{-\beta/(2\beta+d)}(\log n)^{d(1+q)/2 + 2k_1}$, with $L_1 > 0$ and $k_1 = (1 + d) \vee q)/(2 + d/\beta)$. The proofs are based on earlier work, especially from van der Vaart and van Zanten [2009], in which results like (4.9)–(4.11) have been derived for the GP’s like $g$. Here we extend and adapt these results to deal with the additional link function $\sigma$ and
the prior on the maximum intensity $\lambda^*$. 

### 4.4.1 Prior mass condition

In this section we show that with $\lambda^*$, $\sigma$, and $g$ specified in Section 4.2.3 and $\lambda_0 \in C^\beta(S)$, we have

$$P(\|\lambda^* \sigma(g) - \lambda_0\|_{\infty} \leq \delta_n) \geq c_1 e^{-n c_2 \delta_n^2} \quad (4.13)$$

for constants $c_1, c_2 > 0$ and $\delta_n$ as defined above.

The link function $\sigma$ is strictly increasing and smooth, hence it has a smooth inverse $\sigma^{-1} : (0,1) \to \mathbb{R}$. Define the function $w_0$ on $S$ by

$$w_0(s) = \sigma^{-1}\left(\frac{\lambda_0(s)}{2\|\lambda_0\|_{\infty}}\right), \quad s \in S,$$

so that $\lambda_0 = 2\|\lambda_0\|_{\infty} \sigma(w_0)$. Since the function $\lambda_0$ is positive and continuous on the compact set $S$, it is bounded away from 0 on $S$, say $\lambda_0 \geq a > 0$. It follows that $\lambda_0(s)/2\|\lambda_0\|_{\infty}$ varies in the compact interval $[a/2\|\lambda_0\|_{\infty}, 1/2]$ as $s$ varies in $S$, hence $w_0$ inherits the smoothness of $\lambda_0$, i.e. $w_0 \in C^\beta(S)$.

Now observe that for $\varepsilon > 0$,

$$P(\|\lambda^* \sigma(g) - \lambda_0\|_{\infty} \leq 2\varepsilon) = P(\|\lambda^* - 2\|\lambda_0\|_{\infty}\sigma(g) + 2\|\lambda_0\|_{\infty}(\sigma(g) - \sigma(w_0))\|_{\infty} \leq 2\varepsilon) \geq P(\|\lambda^* - 2\|\lambda_0\|_{\infty}\| \leq \varepsilon)P(\|\sigma(g) - \sigma(w_0)\|_{\infty} \leq \varepsilon/2\|\lambda_0\|_{\infty}).$$

Since $\lambda^*$ has a positive continuous density the first factor on the right is bounded from below by a constant times $\varepsilon$. Since the function $\sqrt{\sigma}$ is Lipschitz by assumption, the second factor is bounded from below by $P(\|g - w_0\|_{\infty} \leq c\varepsilon)$ for a constant $c > 0$. By Theorem 3.1 in van der Vaart and van Zanten [2008b] we have the lower bound

$$P(\|g - w_0\|_{\infty} \leq \delta_n) \geq e^{-n \delta_n^2},$$

with $\delta_n$ as specified above. The proof of (4.13) is now easily completed.

### 4.4.2 Construction of sieves

Let $\mathbb{H}^f$ be the RKHS of the GP $g$ with covariance (4.2) and let $\mathbb{H}_1^f$ be its unit ball (see van der Vaart and van Zanten [2008b] for background on these notions). Let $\mathbb{B}_1$ be the unit ball in $C[0,1]^d$ relative to the uniform norm. Define

$$\mathcal{F}_n = \bigcup_{\lambda \leq \lambda_n} \lambda \mathcal{G}_n,$$
4. Poisson intensity estimation

where

\[ G_n = \left[ M_n \sqrt{\frac{\theta_n}{\gamma_n}} H_1^n + \varepsilon_n B_1 \right] \cup \left[ \bigcup_{a \leq \gamma_n} (M_n H_1^n) + \varepsilon_n B_1 \right], \]

and \( \lambda_n, M_n, \gamma_n, r_n, \text{ and } \varepsilon_n \) are sequences to be determined later. In the next two subsections we study the metric entropy of the sieves \( F_n \) and the prior mass of their complements.

### 4.4.3 Entropy

Since \( \sqrt{\sigma} \) is bounded and Lipschitz we have, for \( a, b \in [0, \lambda_n] \), some \( c > 0 \), and \( f, g \in G_1 \)

\[ \| \sqrt{a\sigma(f)} - \sqrt{b\sigma(g)} \|_\infty \leq |\sqrt{a} - \sqrt{b}| + c \sqrt{\lambda_n} \| f - g \|_\infty. \]

Since \( |\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|} \) for \( a, b > 0 \), it follows that for \( \varepsilon > 0 \)

\[ N(2\varepsilon \sqrt{\lambda_n}, \sqrt{F_n}, \| \cdot \|_2) \leq N(\varepsilon \sqrt{\lambda_n}, [0, \lambda_n], \sqrt{\cdot}) N(\varepsilon/c, G_n, \| \cdot \|_\infty), \]

and hence

\[ \log N(2\varepsilon \sqrt{\lambda_n}, \sqrt{F_n}, \| \cdot \|_2) \lesssim \log \left( \frac{1}{\varepsilon} \right) + \log N(\varepsilon/c, G_n, \| \cdot \|_\infty). \]

By formula (5.4) from van der Vaart and van Zanten [2009],

\[ \log N(3\varepsilon_n, G_n, \| \cdot \|_\infty) \leq K r_n^d \left( \log \frac{d^{1/4} M_n^{3/2} \sqrt{2\tau r_n}}{\varepsilon_n^{3/2}} \right)^{1+d} + 2 \log \frac{2M_n \sqrt{\| \mu \|}}{\varepsilon_n}, \]

for \( \| \mu \| \) the total mass of the spectral measure \( \mu \), \( \tau^2 \) the second moment of \( \mu \), a constant \( K > 0 \), \( \gamma_n = \varepsilon_n / (2\sqrt{dM_n}) \), \( r_n > A \) for some constant \( A > 0 \), and given that the following relations hold

\[ d^{1/4} M_n^{3/2} \sqrt{2\tau r_n} > 2\varepsilon_n^{3/2}, \quad M_n \sqrt{\| \mu \|} > \varepsilon_n. \] (4.14)

By substituting \( \eta_n = \varepsilon_n \sqrt{\lambda_n} \) we get that for some constants \( K_1 \) and \( K_2 \),

\[ \log N(2\eta_n, \sqrt{F_n}, \| \cdot \|_2) \lesssim K_1 r_n^d \left( \log \frac{\lambda_n^{3/4} M_n^{3/2} d^{1/4} \sqrt{2\tau r_n}}{\eta_n^{3/2}} \right)^{1+d} + K_2 \log \frac{\lambda_n^{1/2} M_n}{\eta_n}, \]

when \( M_n > 1 \). In terms of \( \eta \) the conditions (4.14) can be rewritten as

\[ d^{1/4} M_n^{3/2} \lambda_n^{3/4} \sqrt{2\tau r_n} > 2\eta_n^{3/2}, \quad M_n \lambda_n^{1/2} \sqrt{\| \mu \|} > \eta_n. \] (4.15)
4.4. Proofs

So we conclude that we have the entropy bound

$$\log N(\tilde{\eta}_n, \sqrt{\mathcal{F}_n}, \| \cdot \|_2) \lesssim n\tilde{\eta}_n^2$$

for sequences $\lambda_n, M_n, r_n$ and $\tilde{\eta}_n$ satisfying (4.15) and

$$K_1 r_n^d \left( \log \frac{\lambda_n^{3/4} M_n^{3/2} d^{1/4} \sqrt{2\pi} r_n}{\tilde{\eta}_n^{3/2}} \right)^{1+d} < n\tilde{\eta}_n^2, \quad K_2 \log \frac{\lambda_n^{1/2} M_n}{\tilde{\eta}_n} < n\tilde{\eta}_n^2. \quad (4.16)$$

4.4.4 Remaining mass

By conditioning we have

$$\mathbb{P}(\lambda^* \sigma(g) \not\in \mathcal{F}_n) = \int_0^\infty \mathbb{P}(\lambda \sigma(g) \not\in \mathcal{F}_n) p_{\lambda^*}(\lambda) d\lambda \leq \int_0^{\lambda_n} \mathbb{P}(\lambda \sigma(g) \not\in \mathcal{F}_n) p_{\lambda^*}(\lambda) d\lambda + \int_{\lambda_n}^\infty p_{\lambda^*}(\lambda) d\lambda.$$

By (4.5) the second term is bounded by a constant times $\exp(-c_0 \lambda_n^\kappa)$. For the first term, note that for $\lambda \leq \lambda_n$ we have

$$\lambda^{-1} \bigcup_{\lambda' \leq \lambda_n} \lambda' \sigma(G_n) \supset \sigma(G_n).$$

Hence, $\mathbb{P}(\lambda \sigma(g) \not\in \mathcal{F}_n) \leq \mathbb{P}(g \not\in G_n)$. From (5.3) in van der Vaart and van Zanten [2009] we obtain the bound

$$\mathbb{P}(g \not\in G_n) \leq \frac{K_3 r_n^{p-d+1} e^{-D_2 r_n^d \log q} r_n}{\log q r_n} + e^{-M_n^2/8}$$

for some $K_3 > 0$, $\varepsilon_n < \varepsilon_0$ for a small constant $\varepsilon_0 > 0$, and $M_n, r_n$ and $\varepsilon_n$ satisfying

$$M_n^2 > 16 K_4 r_n^d (\log (r_n/\varepsilon_n))^{1+d}, \quad r_n > 1, \quad (4.17)$$

where $K_4$ is some large constant. It follows that $\mathbb{P}(g \not\in G_n)$ is bounded above by a multiple of $\exp(-L n \tilde{\eta}_n^2)$ for a given constant $L$ and $\tilde{\eta}_n = \lambda_n \varepsilon_n$, provided $M_n, r_n, \gamma_n$ and $\varepsilon_n$ satisfy (4.17) and

$$D_2 r_n^d \log q r_n \geq 2L n \tilde{\eta}_n^2, \quad r_n^{p-d+1} \leq e^{L n \tilde{\eta}_n^2}, \quad M_n^2 \geq 8L n \tilde{\eta}_n^2. \quad (4.18)$$

Note that in terms of $\tilde{\eta}_n$, (4.17) can be rewritten as

$$M_n^2 > 16 K_4 r_n^d (\log (r_n \lambda_n / \tilde{\eta}_n))^{1+d}, \quad r_n > 1. \quad (4.19)$$
4. Poisson intensity estimation

We conclude that if (4.19),(4.18) holds and
\[ c_0 \lambda_n^\kappa > L n \tilde{\eta}_n^2, \tag{4.20} \]
then
\[ \mathbb{P}\left( \frac{\lambda^\ast \sigma(g \not\in \mathcal{F}_n)}{\lambda} \right) \lesssim e^{-L n \tilde{\eta}_n^2}. \]

4.4.5 Completion of the proof

In the view of the preceding it only remains to show that \( \tilde{\eta}_n, \eta_n, \rho_n, M_n \) can be chosen such that relations (4.15), (4.16), (4.18), (4.19) and (4.20) hold.

One can see that it is true for \( \tilde{\eta}_n = \delta_n \) and \( \eta_n = \delta_n \) described in the theorem, with \( r_n, M_n, \lambda_n \) as follows
\[
\begin{align*}
  r_n &= L_2 n^{1/(2\beta+d)} (\log n)^{2k_1/d}, \\
  M_n &= L_3 n^{d/(4\beta+2d)} (\log n)^{2k_1+(d+1)/2}, \\
  \lambda_n &= L_4 n^{d/\kappa(2\beta+d)} (\log n)^{4k_1/\kappa}
\end{align*}
\]
for some large constants \( L_2, L_3, L_4 > 0 \).
References


References


82
References


Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications **19**, 533–597. 60


References


85


Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. 21

Summary

Asymptotic results in nonparametric Bayesian function estimation

Nonparametric models are widely used in practical applications and involve at least one infinite-dimensional parameter of interest which is commonly a function or a measure. A standard example of such a model would be a density function estimation problem, in which the goal is to recover the density function corresponding to a certain distribution by looking at the data drawn from it.

In the last decades Bayesian nonparametric methods became extremely popular for numerous reasons, such as their philosophical appeal and conceptual simplicity. However, until recently there was little fundamental mathematical understanding of such procedures. The essence of the Bayesian paradigm is the philosophical idea that the parameter of interest in a statistical model does not have one true value, but it is rather perceived as a random object itself. The evidence about the true state of the world is then expressed in terms of degrees of belief. In Bayesian statistics in order to get an estimator one has to put a prior distribution on a parameter space. This can be interpreted as one's prior opinion about the nature of the data. Central in the Bayesian framework is the posterior distribution which can be viewed as an updated version of the prior belief given the evidence.

It is sensible to evaluate the robustness of Bayesian procedures used in nonparametric settings in order to gain insight about which priors to use and how their tuning affects the performance of the procedures. One way to do it is to take an asymptotic approach and to study the Bayesian procedures from frequentist perspective by assuming that there exists a true underlying parameter and studying how well it could be approximated by the estimator as the data sample grows. With the growth of the data sample a good estimator should get in smaller and smaller neighbourhoods of the truth. Moreover, it is desirable for the estimator to be able to have a fast “learning” rate (in terms of the size of this neighbourhood) for every parameter in the class. However, it is well known that this rate is bounded from below by a minimax rate.

This thesis is focused on function estimation problems. For such problems the minimax rate, among other things, depends on the smoothness of the target
function. For instance, in a particular case of a regression problem on the interval \([0, 1]\) it is clear that the rougher the target function is, the more data points we need in order to be able to accurately approximate the behaviour of the function. In most real world applications the smoothness of the function is not known in advance, so it is desirable to develop adaptive procedures that can adapt to all levels of smoothness in the parameter class.

In this thesis we develop Bayesian procedures and study their asymptotic behaviour in the context of two different statistical settings. First, we discuss regression and binary classification problems on large graphs, where the goal is to estimate a smooth function on the vertices of a graph. The second problem that we consider is the problem of intensity estimation of inhomogeneous Poisson process from a realised point pattern.

Chapter 1 introduces the main notions and concepts of Bayesian nonparametric statistics. In Chapter 2 we present a framework for studying the performance of methods for nonparametric function estimation on large graphs. We propose assumptions on the geometry of the underlying graph and the regularity of the function formulated in terms of the Laplacian of the graph. We also derive minimax rates for the regression and classification problems on large graph within the introduced framework. In Chapter 3 we exhibit nonparametric Bayes function estimation procedures in the graph setting. We study rescaled Gaussian priors and we show that the procedures based on these priors achieve good convergence rates and that they are rate–adaptive. We present the results for the cases of full and missing observations. Chapter 4 is dedicated to estimating the intensity function of an inhomogeneous Poisson process. We study the Bayesian procedure developed by Adams et al. [2009]. We show that their SGCP approach to learning intensity functions enjoys very favourable theoretical properties, provided the priors on the hyperparameters are chosen appropriately.
Samenvatting

Asymptotische resultaten in het niet-parametrisch Bayesiaans schatten van functies

Niet-parametrische modellen worden veel gebruikt in praktische toepassingen en brengen ten minste één oneindig-dimensionele parameter met zich mee, meestal in de vorm van een functie of een maat. Een standaard voorbeeld is het schatten van een onbekende dichtheidsfunctie aan de hand van trekkingen uit een bepaalde verdeling.

In de laatste decennia zijn niet-parametrische Bayesiaanse modellen bijzonder populair geworden om verschillende redenen. Zo zijn deze modellen vaak conceptueel eenvoudig en bevatten ze een zekere filosofische aantrekkingskracht. Tot voor kort waren er weinig fundamentele resultaten voor de bijbehorende schattingsprocedures. De essentie van het Bayesiaanse denkkader is het idee dat de parameter die men wil schatten in het statistische model niet een vaste, echte waarde heeft, maar zelf ook een toevalsvariabele is. De waarde van deze parameter in de echte wereld wordt uitgedrukt in waarschijnlijkheidstermen. Om binnen de Bayesiaanse statistiek tot een schatter te komen dient men eerst een a-prioriverdeling te kiezen op de parametterruimte. Het is gebruikelijk dat deze wordt geïnterpreteerd als a-priori vermoeden over de parameter. Centraal in het Bayesiaanse raamwerk is de a-posteriori verdeling, die kan worden geïnterpreteerd als een herziene versie van het a-priori vermoeden, gegeven de observaties.

Het ligt voor de hand om de robuustheid van niet-parametrische Bayesiaanse procedures te evalueren om inzicht te krijgen in welke a-prioriverdelingen bruikbaar zijn en hoe hun afstelling de prestatie beïnvloedt. Een manier om dit te doen, is een asymptotische aanpak waarin de Bayesiaanse procedures vanuit een frequentistisch oogpunt bestudeerd worden. Dit wordt gedaan door aan te nemen dat er een echte waarde bestaat voor de onderliggende parameter en te onderzoeken hoe goed deze benaderd kan worden naarmate de steekproef groter wordt. Voor een goede schatter geldt dat deze, naarmate de steekproef groter wordt, zich concentreert rond de echte waarde. Daarnaast is het wenselijk dat de schatter zich snel concentreert voor elke parameter in een bepaalde klasse. Het is bekend dat deze snelheid begrensd wordt door de zogenoemde “minimax rate”.
Dit proefschrift behandelt het schatten van functies. Voor dergelijke problemen hangt de minimax rate onder andere af van de gladheid van de te schatten functie. Bijvoorbeeld, in het geval van regressie op het interval $[0,1]$ is het duidelijk dat naarmate de te schatten functie minder glad is, er meer data nodig is om het gedrag van de functie nauwkeurig te benaderen. Normaal gesproken is de gladheid van deze functie onbekend en derhalve is het wenselijk om adaptieve procedures te ontwikkelen die zich aan kunnen passen aan verschillende gladheidsgraden.

In dit proefschrift ontwikkelen we Bayesiaanse procedures en bestuderen we hun asymptotische gedrag in twee verschillende statistische contexten. Eerst behandelen we regressie en binaire classificatie op grote grafen, met als doel een gladde functie op de knooppunten van de graaf te schatten. Ten tweede kijken we naar het schatten van de intensiteitsfunctie van een inhomogene Poisson proces.

Hoofdstuk 1 introduceert de belangrijkste begrippen en concepten van niet-parametrisch Bayesiaanse statistiek. In hoofdstuk 2 presenteren we een raamwerk voor het bestuderen van de prestaties van niet-parametrische methodes om functies te schatten op grote grafen. We presenteren aannames over de geometrie van de onderliggende graaf en de regulariteit van de functie geformuleerd in termen van de Laplaciaan van de graaf. Daarnaast leiden we minimax rates af voor regressie- en classificatieproblemen voor grote grafen binnen het geïntroduceerde raamwerk. In hoofdstuk 3 worden niet-parametrische Bayesiaanse functieschatters gepresenteerd voor grafen. We bestuderen geschalde Gaussische a-prioriverdelingen en laten zien dat procedures gebaseerd op deze a-prioriverdelingen goede convergentiesnelheden hebben en dat zij zich aanpassen aan de onbekende gladheid van de functie. We presenteren de resultaten voor zowel het geval van volledige en van onvolledige observaties. Hoofdstuk 4 behandelt het schatten van de intensiteitsfunctie van een inhomogene Poisson proces. We bestuderen de Baysiaanse procedure ontwikkeld door Adams et al. [2009] en laten zien dat hun SGCP aanpak erg gunstige theoretische eigenschappen heeft, gegeven dat de a-prioriverdeling van de hyperparameters juist gekozen zijn.
Author contributions

This thesis is based on the following two published articles and one article in preparation:


Each of the authors equally contributed to each of the articles.
Acknowledgments

I am thankful to all the people who supported me during my time as a PhD student at University of Amsterdam. Foremost, I want to express my deepest gratitude to my supervisor, Harry van Zanten. He was a constant source of inspiration and it was a great honour for me to work with him during the last four years. I appreciate his immediate availability for me at times when I needed help as well as the space that he gave me to conduct the research at my own pace. Harry’s cheerful character and genuine eagerness to do mathematics made working with him a very enjoyable experience.

I wish to express my sincere appreciation to NWO (De Nederlandse Organisatie voor Wetenschappelijk Onderzoek) that funded this PhD work. I am grateful to the members of my defence committee, Bert van Es, Subhashis Ghosal, Mathisca de Gunst, Michel Mandjes, and Sindo Nunez Queija, for taking their time to read this thesis. I also thank the director of KdVI Eric Opdam and my co-promoter Bas Kleijn.

I am very grateful to Yakov Nikitin who introduced me to the fascinating field of mathematical statistics, guided me during my master’s program, and supported me in finding a PhD position.

Furthermore, I would like to thank all my friends and colleagues who made my time in Amsterdam so pleasurable. I am especially grateful to my paranymps, Jarno and Ana. Jarno’s enthusiasm and dedication to every activity he is engaged in made the atmosphere at the university lively. He always made time for me and supported me. Ana is a person that I can turn to when I am struggling. Her constant optimism, deeply caring character, and her unconditional faith in me helped me to get through the most difficult times. I am also thankful to my officemate Jan for his kindness, passion for mathematics, and for all the delightful conversations we had over the years.

This dissertation has greatly benefitted from my numerous talks with Botond whose mathematical abilities and life philosophy I deeply admire. I would like to thank him for always being available if I had any questions and for his insightful comments on my work. I thank Denis for being there for me to discuss mathematical computations, polish my presentations, improve my writing skills, and assist in practically every other matter that I ever needed help with. I am also grateful to Aram for supporting me during the final stage of writing my PhD, for commenting
on my writing, and for motivating me in every aspect of my life. I would like to mention Claudia, who inspired me greatly with her joyful spirit and inexhaustible energy. Her perspective on the challenges of academia and her useful advice on academic writing encouraged me to overcome the difficulties I encountered on my way.

I thank Kiyoshi Takahase Segundo for providing a picture for the cover of my thesis and Vladimir Levitin for its design.

Finally, I thank my family for their unconditional love and support. Words cannot express the amount of help I received from my parents, and I will be forever be grateful for all the sacrifices they made on my behalf. They have always been a great example of diligence, hard work, and loving kindness. Their keen passion for education and constant self development gave me the strength to enroll and complete my PhD.