Posterior contraction for conditionally Gaussian priors

Posterior contraction for conditionally Gaussian priors

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 8 oktober 2012 om 16.00 uur

door

René de Jonge

geboren te Zaanstad

Dit proefschrift is goedgekeurd door de promotor:

prof.dr. J.H. van Zanten

Posterior contraction for conditionally Gaussian priors /
by René de Jonge

A catalogue record is available from the Eindhoven University of Technology Library. ISBN: 978-90-386-3192-9

Contents

Chapter 1

Introduction

1.1 Non-parametric Bayesian inference

In mathematical statistics, observations in statistical experiments are usually assumed to be realizations of random variables $X_1, \ldots, X_n$ with a certain probability distribution P. This probability distribution is often assumed to be in some collection $\mathcal{P}$ of known probability distributions, called the statistical model. We say that the statistical model is indexed by a parameter θ from a parameter space Θ if we can write the model as $\mathcal{P} = \{P_\theta^{(n)} : \theta \in \Theta\}$. We usually assume that the distributions P_θ in the statistical model have probability densities p_θ with respect to some σ-finite measure μ on the sample space $\mathcal{X}$ of the observations.

In parametric statistics, the parameters θ are real numbers or finite vectors of real numbers. Statistical models can however also be indexed by parameters from infinite-dimensional parameter spaces. Statistical inference is then called non-parametric inference. An infinite-dimensional parameter is typically a function, such as a probability density or a regression function.

There are several possible ways of making inference about unknown parameters. We distinguish the frequentist and Bayesian paradigms. Frequentist statisticians assume that the distribution P is given by some fixed and unknown parameter θ_0 in the parameter space Θ. Bayesian statisticians in contrast assume that the uncertainty about the distribution P is described by a probability distribution on the parameter space Θ.

The Bayesian approach to statistical inference involves the choice by the Bayesian statistician of a probability distribution Π on the parameter space Θ. This so-called prior distribution represents the Bayesian statistician's belief about P before taking any observations into account. The Bayesian statistician then addresses the question how this belief should be updated once the observations are available. This updated belief is again given by a probability distribution on Θ. The

updated probability distribution is called the posterior distribution. More precisely, in the Bayesian setup, the law $P_\theta^{(n)}$ is viewed as the conditional distribution of $X = (X_1, \ldots, X_n)$ given that the prior Π has selected the parameter θ. Under regularity conditions, the pair (θ, X) then has a joint probability distribution

$$\mathbb{P}(\theta \in A, X \in B) = \int_A P_\theta^{(n)}(B)\, d\Pi(\theta).$$

The marginal distribution of X is then given by $\mathbb{P}(X \in B) = \int_\Theta P_\theta^{(n)}(B)\, d\Pi(\theta)$. The conditional distribution of θ given X is found from Bayes' formula

$$\mathbb{P}(\theta \in A | X \in B) = \frac{\mathbb{P}(\theta \in A, X \in B)}{\mathbb{P}(X \in B)} = \frac{\int_A P_\theta^{(n)}(B)\, d\Pi(\theta)}{\int_\Theta P_\theta^{(n)}(B)\, d\Pi(\theta)}.$$

Let $p_\theta^{(n)}$ be the probability density of $P_\theta^{(n)}$ with respect to the measure μ on $\mathcal{X}$. The preceding implies that the conditional probability $\Pi(\cdot|X)$, the posterior distribution, is given by

$$A \mapsto \Pi(A|X) = \frac{\int_A p_\theta^{(n)}(X)\, d\Pi(\theta)}{\int_\Theta p_\theta^{(n)}(X)\, d\Pi(\theta)}.$$

The posterior distribution can then be used to construct for instance estimators, credible sets and hypothesis tests for the parameter θ.

Asymptotic statistics deals with the behavior of statistical procedures as the number of observations n grows arbitrarily large. An asymptotic analysis of Bayesian procedures is possible in the frequentist's setup. This requires some notions to express the quality of a posterior distribution for making inference about the true parameter. A first important concept is that of posterior consistency. A sequence of posterior distributions $\Pi(\cdot|X_1, \ldots, X_n)$ is said to be consistent with respect to some metric d on the parameter space, if any fixed neighborhood $U_\delta = \{\theta \in \Theta : d(\theta, \theta_0) \leq \delta\}$ around the true parameter receives a posterior probability arbitrarily close to one as the number of observations tends to infinity. By this we mean that for any $\delta > 0$

$$\Pi(U_\delta | X_1, \ldots, X_n) \to 1 \text{ as } n \to \infty$$

either in P_0-probability or P_0-almost surely. Given posterior consistency, the next important concept is that of posterior contraction. The rate of posterior contraction is the smallest radius $\varepsilon_n \downarrow 0$ for which shrinking balls $U_{\varepsilon_n} = \{\theta \in \Theta : d(\theta, \theta_0) \leq \varepsilon_n\}$ still capture a posterior mass that converges to one as $n \to \infty$. By this we mean that

$$\Pi(U_{\varepsilon_n} | X_1, \ldots, X_n) \to 1 \text{ as } n \to \infty$$

either in P_0-probability or P_0-almost surely.

We consider in this work the asymptotic behavior of certain Bayesian non-parametric inference procedures based on an increasing number of observations. This is motivated by the use in practice of Bayesian procedures in non-parametric models. The Bayesian approach is a popular tool in applied statistics. Bayesian procedures can often be implemented using simulation methods such as the Markov Chain Monte Carlo methods. The conclusions of an applied Bayesian statistician make sense to a frequentist only if the results of such an approach are adequately justified by frequentistic performance guarantees, such as posterior consistency and rate of contraction results. For Bayesian inference procedures in finite-dimensional models, posterior consistency is typically not a problem. The Bernstein-von Mises theorem (see for instance van der Vaart [49]) implies that reasonable parametric prior distributions typically yield posteriors that contract around the correct finite-dimensional parameter at an optimal rate. This is however not true in the infinite-dimensional case. Prior distributions that look reasonable at first sight might exhibit posterior inconsistency. The posteriors of such priors might not even concentrate around a parameter at all. Examples of non-parametric priors that exhibit posterior inconsistency are given in e.g. Diaconis and Freedman [12, 13]. Even if a procedure is consistent, contraction rates may be sub-optimal, see [9].

These negative results have not limited the application of Bayesian procedures in non-parametric models with all kinds of non-parametric priors. The lack of theoretical justification of such methods does not necessarily mean that the conclusions are wrong, or even that non-parametric Bayesian procedures are impossible per se. Indeed, posterior consistency results for non-parametric priors have been obtained at least since Doob [15] and later by Schwartz [43], but the general understanding of the behavior of posterior distributions of non-parametric priors is still comparatively small. The use of non-parametric Bayesian procedures poses challenges for the mathematical statistician. It is necessary to study the performance of Bayes procedures that are used in practice and, if possible, to exhibit priors that are optimal in an appropriate sense.

In the present work we make a number of contributions in this respect. We obtain posterior rate of contraction results for two families of non-parametric priors that are applicable in a range of statistical problems, and we obtain a result about the posterior limiting distribution in a specific statistical setting. In particular, we exhibit priors that are optimal from the point of view of contraction rates and adaptation.

1.2 Bayesian asymptotics

Doob's consistency theorem [15] is the first well-known result about posterior consistency that applies to prior distributions on infinite-dimensional models.

This theorem says that the sequence of posterior distributions is consistent for every possible true parameter not contained in some null-set of the prior. If a prior is chosen for which the null-set contains the true parameter, then posterior consistency is not guaranteed nor excluded by Doob's result. A drawback of Doob's result is that it does not provide a means to check whether or not a parameter is contained in the prior-dependent exceptional set.

Examples of non-parametric prior distributions that exhibit posterior inconsistency have been obtained by for instance [11–14]. These counterexamples stress that using Bayesian inference procedures for non-parametric estimation problems is a delicate affair.

A first general posterior consistency result for non-parametric models was obtained by Schwartz [43]. This result asserts posterior consistency under two conditions on the prior and the model. The first condition is that the prior assigns sufficiently large probabilities to neighborhoods of the true parameter. We call this condition the prior mass condition. The second condition requires the existence of tests for testing the true parameter against the complements of neighborhoods around the true parameter. These conditions of Schwartz remain important conditions in many later general posterior contraction results. The more recent paper by Barron et al. [3] about posterior contraction in non-parametric models notes that the counterexample by Diaconis and Freedman [12] in fact fails to satisfy Schwartz' prior mass condition.

The existence of suitable tests to obtain posterior contraction are often found using a metric entropy condition on the model, which characterizes the size of the model by imposing that the model can be covered by a finite number of balls with any fixed radius. Posterior contraction results that impose a metric entropy condition on the model are for instance given by Ghosal et al. [21]. The latter paper does not merely establish posterior consistency, but actually concerns the rate of posterior contraction for general non-parametric priors. Other similar results include [20, 22] for general prior distributions and [17–19] for certain specific priors distributions.

Many recent results in Bayesian non-parametrics are concerned with posterior contraction rates of specific prior distributions and the question whether or not an optimal rate of posterior contraction can be achieved for the estimation of functions in a certain class of smooth functions.

Posterior distributions on non-parametric models give rise to function estimators. The possible rates of non-parametric estimation therefore put lower bounds on the posterior contraction rates that prior distributions are able to achieve. The optimal rate of posterior contraction therefore depends on the class of functions in which we assume the true parameter is contained. Consider for example the estimation of a smooth function in the sense of Hölder. This class of functions is specified by a smoothness level α. The minimax rate of estimation is

the optimal rate in some sense at which a function in this class can be estimated. The minimax rate of estimation by α-smooth functions of d variables is

$$\varepsilon_n = n^{-\frac{\alpha}{d+2\alpha}}$$

if n denotes the number of observations. This rate is also a lower bound on the posterior contraction rates of prior distributions defined on a statistical model given by such functions. The optimal rate of posterior contraction thus depends on the smoothness level of the class in which we assume the true function is contained.

To obtain the correct rate of estimation, the smoothness level of the prior should typically be chosen in accordance with the smoothness level of the true function. Because the true function is assumed to be unknown, it is desirable that the prior distribution does not to depend on the true smoothness level. A prior is called rate-adaptive if for every possible smoothness level of the true function, the posterior achieves the correct rate of contraction even though the prior itself does not depend on the true level.

There are some results in the literature that obtain adaptation of non-parametric Bayesian procedures, see for instance [4, 22, 23, 25, 33, 41, 51], but the number of results is limited. This thesis contributes to the area of non-parametric Bayesian adaptation. The two families of non-parametric prior distributions that we consider in this work are shown to be adaptive and near-optimal for the estimation of Hölder-smooth functions. The construction of these prior distributions starts with the definition of certain families of non-adaptive Gaussian process priors. In the following section we explain how an adaptive prior can be constructed from these families of Gaussian priors.

1.3 Gaussian process priors and beyond

A stochastic process can be viewed as a random element in a space of functions via its sample paths. The probability distribution of such a random element can be used as a prior distributions on functions, and hence as a prior distribution in non-parametric Bayesian inference. Important examples of such priors are the Gaussian process priors. The general posterior contraction rate results [21] were used by van der Vaart and van Zanten [50] to characterize the posterior contraction rates of Gaussian processes priors by a single condition on the concentration of the prior distribution.

The particular prior distributions that we consider in this work are built from Gaussian prior distributions. This allows us to use the machinery for Gaussian process priors reviewed in Section 2.5. The priors that we consider in Section 3.3 and Section 4.3 are in fact Gaussian process priors themselves. We first consider these Gaussian priors and determine the corresponding rates of posterior

contraction that follows from the results in [50]. We then turn to the definitive non-parametric prior distributions of interest in Section 3.5 and Section 4.5.

It turns out that the Gaussian priors under consideration achieve near optimal rates of posterior contraction for the class of functions on which they are defined. These priors however depend on the smoothness level of this class of functions. As such, they are not rate-adaptive. We would like to obtain rate-adaptive priors and this is the reason to look beyond the Gaussian priors.

Prior distributions in general may have tuning parameters. To distinguish them from the parameters of the statistical model, parameters of the prior are usually called hyper-parameters. Different values for the hyper-parameters correspond to different instances of the prior distribution, or, in other words, to different priors in the same family of prior distributions. We pursue rate-adaptive procedures by mixing different instances of the Gaussian priors. This is achieved by choosing a probability distribution on the tuning parameters. The prior distributions defined as such are referred to as conditionally Gaussian priors because the stochastic processes that define these priors are only Gaussian conditional on the values of the hyper-parameters.

A prior distribution constructed by mixing different instances of the same prior with respect to some probability distribution on the hyper-parameter is called a hierarchical prior. The conditionally Gaussian priors that we consider in this work are examples of hierarchical priors. A draw from a hierarchical prior can be obtained by first drawing a realization of the hyper-parameter and then drawing from the prior distribution that matches that particular hyper-parameter. The hierarchical prior is named after these consecutive steps in this procedure.

We thus construct hierarchical priors from our non-adaptive Gaussian priors. We show that the hierarchical priors are indeed rate-adaptive for suitable prior distributions on the hyper-parameters of the Gaussian prior distributions. In the following section we describe in more detail the specific contribution of each of the following chapters in this thesis.

1.4 Overview of this work

In Chapter 2 we review a number of results from the literature that will be of use in the subsequent chapters. In particular, we review a number of results that establish posterior contraction rates for general non-parametric prior distributions. These results can be applied to various statistical problems. We consider three statistical settings in particular. We then turn to prior distributions defined by stochastic processes. For each of the three statistical settings, we will reformulate the posterior contraction conditions for general priors into conditions for stochastic process priors. In the case of a Gaussian process prior, these conditions can be simplified into a single condition on the concentration of the Gaussian process.

We also review this result from the literature, as well as the Gaussian process machinery behind it. The latter will be of use for the conditional Gaussian process priors in the following chapters.

1.4.1 Chapter 3

In Chapter 3 we consider a family of process priors constructed using piecewise polynomial functions on the real line or some multi-dimensional Euclidean space. It is known that these so-called spline functions provide good approximations for Hölder-smooth functions. Splines can therefore be a useful tool for constructing prior distributions on smooth functions.

For an introduction to the splines that we use to construct our family of priors, we refer to Section 3.2. We build our priors from random tensor-product splines with independent Gaussian B-spline coefficients. We keep the order of the splines fixed and treat the number of knots as a hyper-parameter. The latter will at first be deterministic and later be endowed with a second, independent prior. As a result, the priors will be conditionally Gaussian process priors. We prove that a rate-optimal adaptive procedure for the inference about smooth multivariate functions is obtained in this way.

Ghosal et al. [21] obtain a rate-optimal procedure in the density estimation setting using a prior distribution on a log-spline model. If a sample is observed from an unknown density f on an interval, this result says that if $\log f$ is r-times continuously differentiable and uniformly bounded by a known constant, then posterior contraction of the order $n^{-r/(1+2r)}$ is achieved. This procedure is non-adaptive because it depends on knowledge of the smoothness level r of the unknown density f. Rate-adaptive results for spline priors have been obtained by Ghosal et al. [22] and Huang [25] in the density estimation case, using hierarchical priors that endow the dimension of the spline model with an additional prior, again assuming a uniform bound on f. The result in [25] for the density estimation case is accompanied by a similar result in a non-parametric regression context. The prior weights are chosen separately for each case, so the two settings are not treated in a unified manner. A joint feature of the approaches in [22] and [25] is that both the order and the knots of the splines are different among the finite dimensional models. This makes the priors rather involved.

Our approach and the results that we derive in Chapter 3 complement and extend the existing literature in a number of directions. Firstly, we do not fix a specific setting like density estimation to obtain results. Instead, we present general theorems about random spline processes that, in combination with existing general rate of posterior contraction results for specific settings, lead to concrete posterior contraction results in for instance density estimation, classification and regression. Secondly, we do not merely consider inference about functions of a

single variable, but also about functions of several variables. We show that suitable prior distributions on smooth multivariate functions can be constructed using tensor-product splines. Thirdly, we do not assume a known bound on the unknown function of interest. In our approach, we do not need to assume any uniform bound. This is a consequence of the fact that we use unbounded (namely Gaussian) prior weights on the B-spline coefficients. Lastly, we keep the order of the splines in the construction of our prior fixed at a certain value. Only the number of knots of the splines involved is viewed as a hyper-parameter. As a result our priors are simpler and conceivably also computationally more attractive.

Chapter 3 is based on the paper [28].

1.4.2 Chapter 4

In Chapter 4 we consider prior distributions on functions of one or more variables that are constructed using location-scale kernel mixtures. A discrete location-scale mixture of a fixed probability density p on $\mathbb{R}^d$ can be expressed as

$$x \mapsto \sum_{j=1}^{m} w_j \frac{1}{\sigma^d} p\Big(\frac{x - x_j}{\sigma}\Big), \tag{1.1}$$

where $m \in \mathbb{N}$ is called the grid size, the points $x_1, \ldots, x_m \in \mathbb{R}^d$ are called the grid locations and $\sigma > 0$ is called the bandwidth. The numbers $w_1, \ldots, w_m \geq 0$ are called the mixing weights, and typically satisfy $\sum w_j = 1$. The use of such mixtures of kernels is well established for the construction of non-parametric priors on probability densities. When p satisfies some regularity conditions, a wide class of probability densities can be well approximated by mixtures of the form (1.1). Obviously, a much wider class of functions is well approximated if we lift the restriction that the weights w_j should be nonnegative and sum up to 1. This suggests that location-scale mixtures might be attractive priors not just in the setting of density estimation, but for instance also in non-parametric regression.

The priors that we consider in Chapter 4 are constructed by choosing Gaussian priors on the mixing weights in the expression (1.1). We obtain general results for such priors, which can be used in a variety of statistical settings. To illustrate this we present rate of contraction results not just for non-parametric regression, but also for density estimation and classification settings. We will show that if the kernel and the priors on locations and scales are appropriately chosen, kernel mixture priors yield posteriors with good asymptotic properties. It is well known that for the estimation of an α-regular function of d variables, the best possible rate of convergence is of the order $n^{-\alpha/(d+2\alpha)}$ if n is the number of observations. We will prove that up to a logarithmic factor this optimal rate can be attained as the posterior contraction rate of location-scale mixture priors. More importantly, the near optimal rate can be achieved by a prior that does not depend on the

unknown smoothness level α of the regression function. In other words, we can obtain a fully rate-adaptive procedure. The bounds for the convergence rates that we will obtain depend crucially on the smoothness of the kernel p that is used. For kernels with only a finite degree of regularity, we get sub-optimal rates. We only obtain the correct near-optimal rate for kernels that are infinitely smooth, in the sense that they admit an analytic extension to a strip in complex space. The standard normal kernel is an example of an optimal choice in this respect. We also have to put mild conditions on the priors on the grid size m and the scale σ. In particular, the popular inverse gamma choice for the scale is included in our setup.

To prove adaptation to all smoothness levels, we use an idea of Rousseau [41] who establishes adaptation of an appropriate mixture of beta densities to all smoothness levels of a densities on the unit interval. The paper Kruijer et al. [31] employs the same idea to prove adaptation for kernel mixture priors for density estimation. We extend the technique to a multivariate setting. The paper Jonge and van Zanten [26] on which parts of Chapter 4 are based, was written at the same time and independently of the paper [31].

1.4.3 Chapter 5

In Chapter 5 we study the asymptotic behavior of the marginal posterior distribution of the error standard deviation in a non-parametric fixed design regression model with Gaussian errors. So we suppose we have observations $Y_1, \ldots, Y_n$ satisfying

$$Y_i = f_0(x_i) + \sigma_0 Z_i, \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are known elements of a general design space $\mathcal{X}$ and $Z_1, \ldots, Z_n$ are independent standard normal random variables. The variance of the observations $\sigma_0 > 0$ and the regression function $f_0 : \mathcal{X} \to \mathbb{R}$ are assumed to be unknown. We can make Bayesian inference about the parameters f and σ by endowing them with independent priors π_f and π_σ, respectively, and computing the resulting posterior distribution $\Pi(\cdot \,|\, Y_1, \ldots, Y_n)$. Although in most applied problems the main interest is in the regression function f, we are in Chapter 5 primarily interested in the asymptotic behavior of the marginal posterior distribution of the parameter σ.

In case the regression function f_0 is known and σ is the only unknown parameter in the problem, the classical Bernstein-von Mises (BvM) theorem asserts that under minimal regularity conditions, the posterior distribution of σ contracts around the true value σ_0 at the rate $n^{-1/2}$. See Section 5.2.1 for a precise statement. In Chapter 5 we investigate if and how this changes if the regression function f is in fact unknown. Roughly speaking, we show that if the rate of posterior contraction around the infinite-dimensional parameter f is fast enough, then the marginal

posterior distribution of σ has the same asymptotic behavior as in the case that f is known.

Our result can be viewed as a semiparametric Bernstein-von Mises theorem. In general, semiparametric BvM theorems deal with the asymptotic behavior of posterior distributions of finite-dimensional parameters in the presence of an infinite-dimensional nuisance parameter. Theorems of this type have recently been established by several authors, see for instance [5, 8, 10, 40, 44]. Our problem in fact fits into the general framework of Castillo [10] (up to minor adaptations) and we will use his results to derive our BvM theorem for the error standard deviation. As is explained in the cited papers, an important aspect of BvM results is that they allow to conclude that credible sets for the finite-dimensional parameter of interest, i.e. sets that receive a fixed amount α of posterior mass, are also asymptotic α-confidence sets in the frequentist sense. In other words, if a BvM theorem holds, then the posterior distribution correctly quantifies the uncertainty about the true value of the parameter.

This chapter is based on the paper [27].

Chapter 2

Rates in Bayesian non-parametrics

2.1 Introduction

In this chapter we consider the results that we use to obtain posterior contraction for the non-parametric prior distributions given in Chapter 3 and Chapter 4. These prior distributions are defined as the probability distributions of certain (conditionally) Gaussian processes with continuous sample paths. These sample paths can be used to parametrize the statistical models for the non-parametric density estimation, classification and fixed design regression settings as described in Section 2.2, and the probability distribution of the process seen as a random element in a function space thus defines a non-parametric prior on these models.

These priors are examples of stochastic process priors. In this chapter we obtain for such a process prior, a single set of conditions that implies posterior contraction in each of the three given settings simultaneously, by linking these conditions for each of the given settings to the conditions of certain general posterior contraction theorems already available in the literature.

For Gaussian stochastic process priors, such a unified approach has already been obtained in van der Vaart and van Zanten [50]. It was shown that this approach is possible for any mean-zero Gaussian process that takes values in a separable Banach space of functions. We include this result in this chapter, and use it in Section 3.4 and Section 4.4 to obtain posterior contraction rates for the specific Gaussian process priors defined in respectively Section 3.3 and Section 4.3. We also review the Gaussian process machinery behind this result, because it will be used to show posterior contraction rates for the conditional Gaussian process priors in Section 3.5 and Section 4.5. These rates are obtained by verifying the

conditions for posterior contraction of stochastic process priors introduced in this chapter.

The first condition for posterior contraction is called the prior mass condition. This condition requires that the prior distribution assigns enough probability mass to small neighborhoods of the truth. In the general posterior contraction theorems of Section 2.3, these neighborhoods are determined by the Kullback-Leibler numbers. In the process prior theorems, these neighborhoods are determined by the norm of the Banach space in which the process takes its values. The other conditions for posterior contraction require that the prior puts nearly all its mass on subsets of the model which are not too large in the sense that they can be covered by a finite number of balls of any fixed radius. These conditions are called the remaining mass condition and the metric entropy condition.

The concept of metric entropy is defined using the notion of covering numbers of sets in a metric space. The covering number of a set quantifies the size of this set in terms of the smallest number of balls of any fixed radius needed to cover the set.

Suppose A is a nonempty set in a metric space R with distance d. For any $a_0 \in A$ and $r > 0$, let $B(a_0, r)$ be the set of all $a \in A$ such that $d(a, a_0) < r$. Thus $B(a_0, r)$ is the ball of radius r around a_0. An ε-covering of A is a collection $\mathcal{B}$ of balls $B(\cdot, \varepsilon)$ of radius ε such that $A \subset \cup\{B : B \in \mathcal{B}\}$. The ε-covering number of A is the number of sets in a minimal ε-covering of A. This definition only makes sense if there exists such a finite covering of A. A set A is called totally bounded if it can be covered by finitely many balls of any fixed radius. The ε-covering numbers depend on the distance d, but they do not depend on the choice of the containing space R any further.

We can thus consider covering numbers for any combination of a fixed $\varepsilon > 0$, a distance d and a set A which is totally bounded with respect to d. Let $N(\varepsilon, A, d)$ be the ε-covering number of A with respect to the distance d as given above. The metric entropy of a set A is defined as $H(\varepsilon, A, d) = \log N(\varepsilon, A, d)$. We use the natural logarithm to define metric entropy, instead of the base-2 logarithm as used in for instance Kolmogorov and Tihomirov [29]. Obviously, both definitions of metric entropy are the same up to some constant factor.

In the general posterior contraction theorems, the radii of covering balls in the metric entropy condition are measured with respect to a distance appropriate to the model, for instance the Hellinger distance between probability densities. In the process prior theorems, this is again the distance induced by the Banach space norm.

The chapter is organized as follows. We first describe the three non-parametric statistical problems of interest in Section 2.2. We consider general rate of posterior contraction results from the literature in Section 2.3. Then, in Section 2.4, we turn to stochastic process priors. We give conditions that establish posterior contraction

of such priors in each of the three given statistical problem. In Section 2.5 we consider the Gaussian process machinery that will be useful to obtain posterior contraction rates for priors build from Gaussian process priors. In Section 2.6 we review the result that yields posterior rates of contraction for Gaussian process priors themselves.

2.2 Statistical models

We prove that our approach yields a posterior contraction rate in three particular statistical problems, namely non-parametric density estimation, classification and fixed design regression problems.

In the density estimation case, the goal of statistical inference is to determine the true probability density from a number of independent and identically distributed observations.

In the classification (or binary regression) problem, independent observations are considered that give a binary expression (e.g. 'true' or 'false') for a number of different covariate values. The binary regression function is the function that assigns to every possible covariate the probability of 'true'. The goal is to determine this function from the independent observations.

In a fixed design regression problem, the statistician observes the values of some function at a number of points in its domain, but the observations are perturbed with random measurement errors. The goal of the statistician is to determine the true regression function from the observations.

2.2.1 Density estimation

Consider independent and identically distributed random variables $X_1, \ldots, X_n$ taking values in a sample space $\mathcal{X}$. Assume that the probability distribution P_0 of such a variable X_i has a density p_0 with respect to some σ-finite measure μ on $\mathcal{X}$. The goal is to determine the unknown probability density function p_0.

In parametric statistics, we assume that P_0 is contained in some known family $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ of probability densities p_θ indexed by a parameter from a parameter space $\Theta \subset \mathbb{R}^k$. The goal is to determine the correct parameter θ_0 using the available observations $X_1, \ldots, X_1$. A possible approach could for instance be to estimate θ_0 by the maximum likelihood estimator $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$. The parameter $\hat{\theta}(x_1, \ldots, x_n)$ is defined as a location at which $\theta \mapsto p_\theta(x_1) \cdots p_\theta(x_n)$ is maximal.

In non-parametric statistics, the true density p_0 itself is being estimated, for instance using a kernel estimator

$$\hat{p}_n(x) = \sum_{i=1}^{n} \frac{1}{n\sigma} p\Big(\frac{x - X_i}{\sigma}\Big)$$

with p some probability density on $\mathcal{X}$ and $\sigma > 0$ a bandwidth. The statistical model is much larger than in the parametric density estimation problem. The parameter of the statistical model is now itself a probability density function.

In this work, we consider non-parametric Bayesian procedures. Let Π be a prior distribution on the set $\mathcal{P}$ of densities on $\mathcal{X}$. Given the observations $X_1, \ldots, X_n$ mentioned before, the posterior distribution $\Pi(\cdot|X_1, \ldots, X_n)$ on $\mathcal{P}$ is defined by

$$\Pi(A|X_1, \ldots, X_n) = \frac{\int_A \prod_{i=1}^n p(X_i)\, d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, d\Pi(p)}, \quad A \subset \mathcal{P}.$$

For some appropriate distance d on the statistical model $\mathcal{P}$, we are interested in the posterior mass of sets $A_n = \{p : d(p, p_0) \leq \varepsilon_n\}$ with $\varepsilon_n \to 0$. In the next section, we see that posterior contraction can be obtained with d for instance the Hellinger distance. The Hellinger distance $h(p, q)$ between two densities p and q is defined by

$$h^2(p, q) = \int_{\mathcal{X}} (\sqrt{p}(x) - \sqrt{q}(x))^2\, d\mu(x).$$

It is the distance between the square roots of p and q in L_2-sense.

In the next subsections, we introduce the classification and fixed design regression problems, and show what the distances are for which we obtain posterior contraction.

2.2.2 Classification

Let X be a random variable in $\mathcal{X}$ with a probability distribution G that has a density g relative to some σ-finite measure μ. Let Y be a random variable that takes values in $\{0, 1\}$. Assume that $Y|X = x$ has a Bernoulli distribution with probability of success

$$r_0(x) = \mathbb{P}(Y = 1|X = x), \quad x \in \mathcal{X}.$$

In other words, with $p_{r_0}(x, y)g(x) = r_0(x)^y(1 - r_0(x))^{1-y}g(x)$ the probability density of the joint distribution of the pair (X, Y), the marginal probability distribution of Y is given by

$$\mathbb{P}(Y = y) = \int_{\mathcal{X}} p_0(x, y)\, dG(x), \quad y \in \{0, 1\}.$$

We assume that the distribution G of the covariates is known. The goal is to determine the binary regression function r_0 or equivalently, the joint density p_0, from independent and identically distributed observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same distribution as (X, Y).

The distance between two densities p_r and p_s is measured using the L_2-distance

$$\|p_r - p_s\|_{2,G}^2 = \sum_{y \in \{0,1\}} \int_{\mathcal{X}} (p_r(x, y) - p_s(x, y))^2 g(x)\, d\mu(x).$$

The distance between the two corresponding binary regression functions r and s is measured using the L_2-distance

$$\|r-s\|_{2,G}^2 = \int_{\mathcal{X}} (r(x)-s(x))^2 g(x)\, d\mu(x).$$

These distances are actually equivalent. Because $p_s(x,0)-p_r(x,0) = r(x)-s(x) = p_r(x,1)-p_s(x,1)$, we in fact have

$$\|p_r - p_s\|_{2,G}^2 = 2\|r-s\|_{2,G}^2.$$

2.2.3 Fixed design regression

Suppose that $Y_1,\ldots,Y_n$ are independent random variables such that $Y_i = w_0(x_i)+e_i$ for some unknown function $w_0 : \mathcal{X} \to \mathbb{R}$, fixed and known elements $x_1,\ldots,x_n \in \mathcal{X}$ and unobservable independent random variables e_i. The goal in a non-parametric fixed design regression problem is to determine the function $w_0 : \mathcal{X} \to \mathbb{R}$ by observing the tuple $(Y_1,\ldots,Y_n)$. We think of this problem as observing a function w_0 at a number of given locations, where the observations have been contaminated with measurement errors. The measurement errors are not specific to the location of the measurement, and the goal is to remove all measurement errors using the various observations at the different locations.

We only consider Gaussian regression. In a Gaussian regression problem, the error variables e_i have a non-degenerate zero-mean Gaussian distribution with a known or unknown variance. As a consequence, the observation Y_i is a Gaussian random variable with mean $w_0(x_i)$ and some variance $\sigma_0^2 > 0$. In particular, the random variables Y_i are independent, not identically distributed.

We distinguish between the cases in which σ_0 is known and unknown. First suppose that σ_0 is known. For a function $w : \mathcal{X} \to \mathbb{R}$ from some collection C of such functions, let $p_w^{(n)}$ be the probability density of the distribution of $(Y_1,\ldots,Y_n)$ under the regression relation $Y_i = w(x_i)+e_i$ with w as the regression function. The model $\mathcal{P}^{(n)}$ consists of all densities $p_w^{(n)}$ with $w \in C$. This is non-parametric description of a finite-dimensional model. Indeed, two regression functions $v, w : \mathcal{X} \to \mathbb{R}$ that coincide on all design points give rise to the same probability distribution of $(Y_1,\ldots,Y_n)$, that is to say, we have $p_w^{(n)} = p_v^{(n)}$ for all $v, w : \mathcal{X} \to \mathbb{R}$ such that $v(x_i) = w(x_i)$ for all $i = 1,\ldots,n$. The dimension of the finite-dimensional model however changes with the number of available observations. The non-parametric interpretation is an obvious formulation if we want to let the sample size n tend to infinity. The distance $\|v-w\|_n$ between two regression functions $v, w : \mathcal{X} \to \mathbb{R}$ is given by the n-norm distance

$$\|v-w\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} (v(x_i)-w(x_i))^2.$$

This distance only takes the values of v and w at the fixed design points $x_1, \ldots, x_n$ into account. Every n-norm is bounded from above by the supremum norm.

If the error variance σ_0^2 is unknown, then the setup is slightly different. We add a variable $\sigma > 0$ to the parameter of the statistical model, which represents the standard deviation of the errors in the regression relation. The parameter of the statistical model takes the form $\eta = (w, \sigma)$ with $w : \mathcal{X} \to \mathbb{R}$ some regression function in C and σ some positive real number. For $\eta = (w, \sigma)$, let $p_\eta^{(n)}$ be the probability density of the distribution of $(Y_1, \ldots, Y_n)$ under the regression relation $Y_i = w(x_i) + \sigma Z_i$ in which the variables Z_i are independent standard Gaussians. The distance $\|\eta - \eta_0\|$ between the parameters $\eta = (w, \sigma)$ and $\eta_0 = (w_0, \sigma_0)$ is measured by

$$\|\eta - \eta_0\| = \|w - w_0\|_n + |\sigma - \sigma_0|$$

with $\|\cdot\|_n$ the n-norm as defined above.

2.3 Posterior contraction

Theorem 2.1 of Ghosal et al. [21] (see Theorem 2.1 below) gives general conditions for the posterior contraction at a certain rate around the truth of a prior distribution on a non-parametric statistical model in the case of independent and identically distributed observations. This result has been extended in various directions. Theorem 2.4 in the same paper specializes the theorem to a result that establishes parametric contraction rates $n^{-1/2}$ if the theorem is in fact applied to finite-dimensional models. In Theorem 2.1 of Ghosal and van der Vaart [19], the conditions of Theorem 2.1 in [21] are relaxed in the sense that they are formulated using two different rates ε_n and $\tilde{\varepsilon}_n$, which makes the conditions easier to verify. The posterior contraction rate is then the slower of the two rates. Ghosal and van der Vaart [20] generalize Theorem 2.4 in [21] to a result that also holds for non-i.i.d. observations, and in particular to independent but not identically distributed observations in Theorem 4 of [20].

The conditions for the finite dimensional result are more involved than the results that only apply to non-parametric models. Because we only consider non-parametric estimation, we will not consider the finite-dimensional result any further, because this would make the assumptions more complicated than needed. We consider independent observations, but the observations are not always identically distributed.

Let us now consider the posterior contraction statements. Let $\mathcal{P}$ be a collection of probability distributions with densities p relative to some measure on the sample space. Suppose that Π_n is a sequence of prior distributions on the collection $\mathcal{P}$ and that we observe an independent random sample $X_1, \ldots, X_n$ from some probability distribution $P_0 \in \mathcal{P}$ with density p_0. Denote by $\mathbb{E}_0$ the expectation with respect

to the underlying probability measure $\mathbb{P}_0$ (the probability measure $\mathbb{P}_0$ such that $P_0 = \mathbb{P}_0 \circ X_1^{-1}$) and consider the Kullback-Leibler type set

$$\mathrm{KL}(P, \varepsilon) = \{P : K(p, p_0) \leq \varepsilon^2, V(p, p_0) \leq \varepsilon^2\}.$$

with $K(p, p_0)$ and $V(p, p_0)$ respectively the Kullback-Leibler divergence and the second moment

$$K(p, p_0) = -\mathbb{E}_0 \log \tfrac{p}{p_0}(X_i) \quad \text{and} \quad V(p, p_0) = \mathbb{E}_0(\log \tfrac{p}{p_0}(X_i))^2. \tag{2.1}$$

We write $\log N(\varepsilon, \mathcal{P}_n, d)$ for the metric entropy of a subset $\mathcal{P}_n \subset \mathcal{P}$ with respect to ε and distance d. In the following, this distance d is either the Hellinger distance or the L_2 distance between densities.

Theorem 2.1 of Ghosal et al. [21], copied below as Theorem 2.1, asserts that under certain conditions on the model and the prior sequence, the sequence of posteriors contracts at a certain rate around the truth with respect to the Hellinger distance. As mentioned in [21], if the densities in the model are uniformly bounded, then the Hellinger distance can be replaced throughout the proof by the L_2 distance. Because the L_2 distance is in that case bounded from above by a multiple of the Hellinger distance, the conditions are less restrictive, but the assertion is weaker. So the theorem with the L_2 distance is really a different result. Both results are formulated by the following statement by allowing d to be either the Hellinger distance or the L_2-distance.

Theorem 2.1. *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ there exist sets $\mathcal{P}_n \subset \mathcal{P}$ such that*

$$\Pi_n(\mathrm{KL}(P_0, \varepsilon_n)) \geq \exp(-n\varepsilon_n^2) \tag{2.2}$$

$$\Pi_n(\mathcal{P} \backslash \mathcal{P}_n) \leq \exp(-5n\varepsilon_n^2) \tag{2.3}$$

$$\log N(\varepsilon_n, \mathcal{P}_n, d) \leq n\varepsilon_n^2. \tag{2.4}$$

Then for any sufficiently large constant L,

$$\Pi_n(P : d(P, P_0) \geq L\varepsilon_n | X_1, \ldots, X_n) \xrightarrow{P_0^n} 0.$$

This theorem can be used to obtain posterior contraction rates of sequences of Gaussian priors in non-parametric density estimation and classification problems via the general result in Theorem 2.18 of Section 2.6.

This result was extended in Theorem 2.1 of Ghosal and van der Vaart [19] to allow two rates in the conditions of the theorem. This result is as follows. Let d again be either the Hellinger distance or the L_2-distance.

Theorem 2.2. *Suppose that for a two positive sequence $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ with $n\varepsilon_n^2 \to \infty$ and $n\bar{\varepsilon}_n^2 \to \infty$ there exist sets $\mathcal{P}_n \subset \mathcal{P}$ such that*

$$\Pi_n(\mathrm{KL}(P_0, \tilde{\varepsilon}_n)) \geq \exp(-n\tilde{\varepsilon}_n^2) \tag{2.5}$$

$$\Pi_n(\mathcal{P} \backslash \mathcal{P}_n) \leq \exp(-5n\tilde{\varepsilon}_n^2) \tag{2.6}$$

$$\log N(\bar{\varepsilon}_n, \mathcal{P}_n, d) \leq n\bar{\varepsilon}_n^2. \tag{2.7}$$

Then for $\varepsilon_n = \tilde{\varepsilon}_n \vee \bar{\varepsilon}_n$ and any sufficiently large constant L,

$$\Pi_n(P : d(P, P_0) \geq L\varepsilon_n | X_1, \ldots, X_n) \xrightarrow{P_0^n} 0.$$

This result will be used to obtain posterior contraction rates for the conditionally Gaussian priors in non-parametric density estimation and classification problems via respectively Theorem 2.5 and Theorem 2.7 in Section 2.4.

To obtain a rate for the fixed design regression setting, we use the following posterior contraction result for independent but not identically distributed data. This result, given in Theorem 2.3 below, is a special case of Theorem 4 in [20].

Let $X_1, \ldots, X_n$ be a sequence of independent random variables and assume that X_i has a probability distribution $P_{0,i} \in \mathcal{P}$ with density $p_{0,i}$. The joint distribution of the observations is given by a product measure $P_0 \in \mathcal{P}^{(n)}$. Define

$$d_n^2(P, P_0) = \frac{1}{n} \sum_{i=1}^{n} \int (\sqrt{p_i}(x_i) - \sqrt{p_{0,i}}(x_i))^2 \, dx_i \tag{2.8}$$

as the average of the squared Hellinger distances (assuming densities are given relative to the Lebesgue measure as reference measure on the data) and the average Kullback-Leibler type set

$$\mathrm{KL}_n^*(P_0, \varepsilon) = \{P \in \mathcal{P}^{(n)} : \frac{1}{n} \sum_{i=1}^{n} K(p_i, p_{0,i}) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^{n} V_0(p_i, p_{0,i}) \leq \varepsilon^2\} \tag{2.9}$$

with

$$K(p_i, p_{0,i}) = -\mathbb{E}_0 \log \tfrac{p}{p_0}(X) \tag{2.10}$$

$$V_0(p_i, p_{0,i}) = \mathbb{E}_0\big(\log \tfrac{p}{p_0}(X) - \mathbb{E}_0 \log \tfrac{p}{p_0}(X)\big)^2. \tag{2.11}$$

Theorem 2.3. *Suppose that for a sequence of $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$, there exist sets $\mathcal{P}_n^{(n)} \subset \mathcal{P}^{(n)}$ such that*

$$\Pi_n(\mathrm{KL}_n^*(P_0, \varepsilon_n)) \geq \exp(-n\varepsilon_n^2/4) \tag{2.12}$$

$$\Pi_n(\mathcal{P}^{(n)} \backslash \mathcal{P}_n^{(n)}) \leq \exp(-3n\varepsilon_n^2) \tag{2.13}$$

$$\log N(\varepsilon_n, \mathcal{P}_n^{(n)}, d_n) \leq n\varepsilon_n^2. \tag{2.14}$$

Then

$$\Pi_n(P : d_n(P, P_0) \geq L_n \varepsilon_n | X_1, \ldots, X_n) \xrightarrow{P_0} 0$$

for every sequence $L_n \to \infty$.

As mentioned in [20], the average-like Hellinger distance d_n is bounded from above by the n-norm distance on the corresponding regression functions. It thus suffices to check the conditions with respect to $\|\cdot\|_n$ in order to show posterior contraction with respect to d_n. The average-like Hellinger distance is however not equivalent with the n-norm distance unless the regression functions are uniformly bounded. So, we cannot obtain contraction with respect to the n-norm distance from this result without an unattractive extra condition on the regression functions. However, in the fixed design regression setting, the average-like Hellinger distance d_n can be replaced throughout the proof by the n-norm distance, see Section 7.7 in [20], so this extra condition can actually be avoided.

This theorem can be used to obtain posterior contraction rates of Gaussian priors in (both cases of) the non-parametric fixed design regression setting in Section 2.2.3 via the general result in Theorem 2.18 of Section 2.6, cf. Theorem 3.3 in [50].

To obtain posterior contraction rates for the conditionally Gaussian priors, the result above is too strict for our purpose. We again want to relax the conditions to allow two different rates. We use the following theorem in the case of fixed design regression. We can again replace d_n throughout by the distance induced by the n-norm on the regression functions.

Theorem 2.4. *Suppose that for sequences* $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ *such that* $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n)^2 \to \infty$, *there exist sets* $\mathcal{P}_n^{(n)} \subset \mathcal{P}^{(n)}$ *such that*

$$\Pi_n(\mathrm{KL}_n^*(P_0, \tilde{\varepsilon}_n)) \geq \exp(-n\tilde{\varepsilon}_n^2/4) \tag{2.15}$$

$$\Pi_n(\mathcal{P}^{(n)} \backslash \mathcal{P}_n^{(n)}) \leq \exp(-3n\tilde{\varepsilon}_n^2) \tag{2.16}$$

$$\log N(\bar{\varepsilon}_n, \mathcal{P}_n^{(n)}, d_n) \leq n\bar{\varepsilon}_n^2. \tag{2.17}$$

Then

$$\Pi_n(P : d_n(P, P_0) \geq L_n(\tilde{\varepsilon}_n \vee \bar{\varepsilon}_n) | X_1, \ldots, X_n) \xrightarrow{P_0} 0$$

for every sequence $L_n \to \infty$.

2.4 Posterior contraction for stochastic process priors

Non-parametric estimation often means the estimation of a function. Because a stochastic process can be seen as a random element in a space of functions, and

hence as a prior distribution on this collection of functions, it is natural to consider stochastic process priors for Bayesian non-parametric inference. The two families of non-parametric prior distributions that we consider in the following chapters of this work are based on stochastic processes.

We write W for some stochastic process (or random field) with realizations in some Banach space of functions. Also, we view this process as a random element in this Banach space. If we refer to the process as a prior distribution, then we mean the probability distribution Π of the process seen as a random element in the Banach space. By the support of W we mean the support of this probability distribution, i.e. the smallest closed subset B of the Banach space which receives probability one under the probability distribution of the random element W.

The general posterior contraction results in Section 2.3 can be reformulated for stochastic process priors in terms of the Banach space in which the prior takes its values. These conditions are more or less the same for each of the three statistical problems mentioned in Section 2.2. This in fact also allows us to give a single set of conditions that we can use to obtain a posterior rate of contraction result in each of the three statistical problems.

The metric entropy conditions turn into metric entropy conditions on subsets of the Banach space with respect to the Banach space norm. The new remaining mass conditions are also formulated using subsets of the Banach space. For the prior mass conditions, the probability of a Kullback-Leibler type neighborhood around the truth is replaced by the probability of a small ball around the truth in the Banach space.

Given an arbitrary element w in the Banach space, we can consider the probability $\mathbb{P}(\|W - w\| \leq \varepsilon)$ that a realization of the prior belongs to an ε-ball around w with respect to the Banach space norm $\|\cdot\|$. For small $\varepsilon > 0$ such a probability is called a small ball probability of W. We assume that W has zero mean, and we thus refer to this probability with $w = 0$ as the centered small ball probability of W, and for any non-zero w we refer to it as a non-centered small ball probability.

The following results assume that the prior Π is obtained from a single stochastic process W. However, the results also hold if Π is replaced by a sequence Π_n of prior distributions corresponding to processes W_n. We always assume that W is a Borel measurable zero-mean random element in $C([0,1]^d)$ equipped with the supremum norm.

2.4.1 Density estimation

Let $\mathbb{B}$ be the Banach space $C([0,1]^d)$ of continuous functions on $[0,1]^d$ equipped with the supremum norm $\|\cdot\|_\infty$. Consider $\mathcal{P} = \{p_w : w \in \mathbb{B}\}$ for probability

densities p_w defined by

$$p_w(x) = \frac{e^{w(x)}}{\int_{[0,1]^d} e^{w(x)}\,dx}, \quad x \in [0,1]^d. \tag{2.18}$$

Suppose that W is a random element of $\mathbb{B}$. The probability distribution of W defines a prior distribution Π on $\mathcal{P}$ via the random variable p_W that takes values $p_w \in \mathcal{P}$ for realizations w of W.

Now suppose that we observe independent and identically distributed $X_1, \ldots, X_n$ from a positive density p_0. To estimate it, we put the prior Π on p and consider the corresponding posterior.

The following result gives conditions for the contraction of the posterior in terms of the Banach space in which the prior W takes its values. The theorem is established by linking the three conditions to the three conditions of Theorem 2.2, which then asserts the required posterior contraction statement. To link the conditions, we need a comparison of the Hellinger distance and Kullback-Leibler numbers with the supremum norm. This comparison is provided by Lemma 2.6 ahead.

Theorem 2.5. *Let Π be the distribution of p_W and let $w_0 = \log p_0$. If there exist sequences $\varepsilon_n \to 0$ and $\bar{\varepsilon}_n \to 0$ with $n(\varepsilon_n^2 \wedge \bar{\varepsilon}_n^2) \to \infty$ and, for every large enough constant C, measurable subsets $B_n \subset \mathbb{B}$ and a constant $D > 0$ such that*

$$\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \le Dn\bar{\varepsilon}_n^2, \tag{2.19}$$

$$\mathbb{P}(W \notin B_n) \le \exp(-Cn\varepsilon_n^2), \tag{2.20}$$

$$\mathbb{P}(\|W(x) - w_0(x)\|_\infty \le 2\varepsilon_n) \ge \exp(-n\varepsilon_n^2), \tag{2.21}$$

then, with h the Hellinger distance,

$$\Pi(p : h(p, p_0) \ge L(\varepsilon_n \vee \bar{\varepsilon}_n) | X_1, \ldots, X_n) \xrightarrow{P_0^n} 0$$

as $n \to \infty$, for every sufficiently large constant L.

Proof. We show that the three conditions imply the conditions (2.5)–(2.7), so that the required posterior contraction follows from Theorem 2.2.

The link between the prior mass conditions is based on the comparison of the Kullback-Leibler numbers (2.1) and supremum norm $\|\cdot\|_\infty$ in Lemma 2.6. It follows that for sufficiently large n, there exists a constant $A \ge 1$ such that both

$$K(p_w, p_0) \le \tilde{\varepsilon}_n^2 \quad \text{and} \quad V(p_w, p_0) \le \tilde{\varepsilon}_n^2$$

are satisfied if $\|w - w_0\|_\infty \le \tilde{\varepsilon}_n / A$. The probability that W maps into $\{w \in \mathbb{B} : \|w - w_0\|_\infty \le \tilde{\varepsilon}_n / A\}$ is thus bounded by the probability that p_W is in $\mathrm{KL}(p_0, \tilde{\varepsilon}_n)$. Thus, with $\tilde{\varepsilon}_n = 2A\varepsilon_n$, according to (2.21),

$$\Pi(\mathrm{KL}(p_0, \tilde{\varepsilon}_n)) \ge \mathbb{P}(\|W - w_0\|_\infty \le 2\varepsilon_n) \ge \exp(-n\varepsilon_n^2) \ge \exp(-n\tilde{\varepsilon}_n^2).$$

For B_n the given sieves, simply define the sets $\mathcal{P}_n = \{p_w \in \mathcal{P} : w \in B_n\}$ so that $\mathbb{P}(W \notin B_n) = \Pi(\mathcal{P}\backslash\mathcal{P}_n)$. It now follows from (2.20) that $\Pi(\mathcal{P}\backslash\mathcal{P}_n) \leq \exp(-5n\tilde{\varepsilon}_n)$ by choosing C large enough.

The link between the metric entropy conditions on the sieves follows from the relation between the Hellinger distance and the supremum norm $\|\cdot\|_\infty$. It follows from Lemma 2.6 that, for sufficiently large n, one has $h(p_v, p_w) \leq 4\bar{\varepsilon}_n$ for any $v, w \in \mathbb{B}$ such that $\|v - w\|_\infty \leq 2\bar{\varepsilon}_n$. From a minimal covering of B_n using balls of radius $2\bar{\varepsilon}_n$ with respect to the supremum norm, one can thus find a covering of $\mathcal{P}_n$ using balls of radius $4\bar{\varepsilon}_n$ with respect to the Hellinger distance. We thus find that

$$N(4\bar{\varepsilon}_n, \mathcal{P}_n, h) \leq N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty)$$

and hence, using (2.19), it follows that (2.7) is satisfied with some multiple of the present $\bar{\varepsilon}_n$. □

The proof is based on a comparison of the Hellinger distance and Kullback-Leibler numbers with the supremum norm. We use the following result, given as Lemma 3.1 in van der Vaart and van Zanten [50].

Lemma 2.6. *There exist constants $C, D \geq 0$ such that*

$$\begin{aligned} h(p_v, p_w) &\leq \|v - w\|_\infty e^{\|v-w\|_\infty/2} \\ K(p_v, p_w) &\leq C\|v - w\|_\infty^2 e^{\|v-w\|_\infty}(1 + \|v - w\|_\infty) \\ V(p_v, p_w) &\leq D\|v - w\|_\infty^2 e^{\|v-w\|_\infty}(1 + \|v - w\|_\infty)^2. \end{aligned}$$

2.4.2 Classification

Suppose that we observe independent and identically distributed pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with values in $\mathcal{X} \times \{0, 1\}$ such that $\mathbb{P}(Y_1 = 1|X_i = x) = r_0(x)$ for some binary regression function $r_0 : \mathcal{X} \to [0, 1]$.

Let $\mathbb{B}$ be the Banach space $C([0, 1]^d)$ of continuous functions on $[0, 1]^d$ equipped with the supremum norm $\|\cdot\|_\infty$. For Ψ the distribution function of the standard logistic distribution and for w realizations of a random element W in $\mathbb{B}$, we define functions $r_w : [0, 1]^d \mapsto (0, 1)$ by $r_w(x) = \Psi(w_x)$. Then the law of $\Psi(W)$ defines a prior Π on the model

$$\mathcal{P} = \{p_w(x, y) = r_w(x)^y(1 - r_w(x))^{1-y} : x \in [0, 1]^d, y \in \{0, 1\}\}.$$

Remember that $p_0(x, y) = r_0(x)^y(1 - r_0(x))^{1-y}$ is the true density of a pair of observations (X_i, Y_i) in the classification problem as described in Section 2.2.2 (relative to the probability distribution G on $[0, 1]^d$).

The following result gives conditions for the contraction of the posterior in terms of the Banach space in which the prior W takes its values. The theorem is

established by linking the three conditions to the three conditions of Theorem 2.2, which then asserts the required posterior contraction statement. To link the conditions, we need a comparison of the Hellinger distance and Kullback-Leibler numbers with the $L_2(G)$-distance. Remember that the $L_2(G)$ distance between p_v and p_w is defined by

$$\|p_v - p_w\|_{2,G}^2 = \sum_{y\in\{0,1\}} \int_{\mathcal{X}} (p_v(x,y) - p_w(x,y))^2 \, dG(x).$$

Note that $\|p_v - p_w\|_{2,G}^2 = 2\|r_v - r_w\|_{2,G}^2$, so the posterior contraction statement can also be formulated with respect to the binary regression function r_0 instead of the true probability density p_0 of the pairs of observations.

Theorem 2.7. *Let Π be the distribution of $\Psi(W)$ and let $w_0 = \Psi^{-1}(r_0)$. If there exist sequences $\varepsilon_n \to 0$ and $\bar\varepsilon_n \to 0$ with $n(\varepsilon_n^2 \wedge \bar\varepsilon_n^2) \to \infty$ and, for every large enough constant C, measurable subsets $B_n \subset \mathbb{B}$ and a constant $D > 0$ such that*

$$\log N(\bar\varepsilon_n, B_n, \|\cdot\|_{2,G}) \le Dn\bar\varepsilon_n^2, \tag{2.22}$$

$$\mathbb{P}(W \notin B_n) \le \exp(-Cn\varepsilon_n^2), \tag{2.23}$$

$$\mathbb{P}(\|W(x) - w_0(x)\|_{2,G} \le 2\varepsilon_n) \ge \exp(-n\varepsilon_n^2), \tag{2.24}$$

then

$$\Pi(r : \|r - r_0\|_{2,G} \ge L(\varepsilon_n \vee \bar\varepsilon_n) | X_1, Y_1 \ldots, X_n, Y_n) \xrightarrow{P_0^n} 0$$

as $n \to \infty$, for every sufficiently large constant L.

Proof. Let Π be the distribution of p_W. We show that

$$\Pi(p : \|p - p_0\|_{2,G} \ge L(\varepsilon_n \vee \bar\varepsilon_n) | X_1, Y_1 \ldots, X_n, Y_n) \xrightarrow{P_0^n} 0$$

by verifying the conditions of Theorem 2.2.

It follows from Lemma 2.8 that we have $K(p_w, p_0) \le \tilde\varepsilon_n^2$ and $V(p_w, p_0) \le \tilde\varepsilon_n^2$ if $\|w - w_0\|_{2,G} \le \tilde\varepsilon_n / A$ for some sufficiently large constant $A \ge 1$. Using (2.24), we conclude that for $\tilde\varepsilon_n = 2A\varepsilon_n$,

$$\Pi(\mathrm{KL}(p_0, \tilde\varepsilon_n)) \ge \mathbb{P}(\|W - w_0\|_{2,G} \le 2\varepsilon_n) \ge \exp(-n\tilde\varepsilon_n^2).$$

This verifies (2.5).

We again let $\mathcal{P}_n = \{p_w : w \in B_n\}$ so that $\mathcal{P}(W \notin B_n) = \Pi(\mathcal{P} \backslash \mathcal{P}_n)$. For $\tilde\varepsilon_n$ as above, condition (2.6) now follows from (2.23) by choosing C large enough.

The $L_2(G)$-distance between different p_w in the model can be bounded from above by some large enough multiple of the $L_2(G)$-distance of the corresponding indices w according to Lemma 2.8. As a consequence,

$$\log N(\bar\varepsilon_n, \mathcal{P}_n, \|\cdot\|_{2,G}) \le \log N(\bar\varepsilon_n, B_n, \|\cdot\|_{2,G}).$$

It now follows from (2.22) that (2.7) is satisfied for a sufficiently large multiple of the present $\bar\varepsilon_n$. □

We again linked to the corresponding conditions of Theorem 2.2. The distance d is not the Helling distance as in the previous case, but the $L_2(G)$-distance. We need the following results, obtained as a special case of Lemma 3.2 in [50].

Lemma 2.8. *There exists constants $C, D, E > 0$ such that*

$$\begin{aligned} \|p_v - p_w\|_{2,G} &\leq C\|v - w\|_{2,G} \\ K(p_v, p_w) &\leq D\|v - w\|_{2,G}^2 \\ V(p_v, p_w) &\leq E\|v - w\|_{2,G}^2. \end{aligned}$$

Instead of the standard logistic distribution function, we can also use the standard normal distribution function as link function Ψ. We then require that the conditions of Theorem 2.7 hold with the supremum norm $\|\cdot\|_\infty$ instead of the $L_2(G)$-norm. We actually turn to the supremum norm altogether in Section 2.4.4 ahead. The posterior contraction statement however remains in terms of the $L_2(G)$-distance on the densities. It follows from Lemma 3.2 in [50] that, in the case of a standard normal link function, both $K(p_v, p_w)$ and $V(p_v, p_w)$ are bounded from above by a multiple of $\|v - w\|_\infty$.

2.4.3 Fixed design regression

Let $(Y_1, \ldots, Y_n)$ be observations in the fixed design regression problem as described in Section 2.2.3, with design points $x_1, \ldots, x_n \in [0,1]^d$. First assume that the error standard deviation σ_0 is known.

Suppose that W is a random element in the Banach space $C([0,1]^d)$ equipped with the supremum norm. Given the covariates $x_1, \ldots, x_n \in [0,1]^d$, we consider the n-norm of a realization w of W

$$\|w\|_n^2 = \frac{1}{n}\sum_{i=1}^n w(x_i)^2.$$

For each n, the process W defines a prior distribution Π on the model

$$\mathcal{P}^{(n)} = \left\{ p_w^{(n)}(y) = \prod_{i=1}^n p_{w,i}(y_i) \right\}$$

with $p_{w,i}$ the probability densities of observations Y_i that satisfy the regression relation with the regression function w, that is to say the probability density of a Gaussian distribution with mean $w(x_i)$ and variance σ_0^2. Alternatively, we can say that the process W defines a prior on the regression functions $w : \mathcal{X} \to \mathbb{R}$ itself. Remember that w_0 is the true regression function, and that σ_0 is the true standard deviation of the error variables. The norm for which we obtain posterior contraction is the n-norm.

The following result gives conditions for the contraction of the posterior in terms of the Banach space in which the prior W takes its values. The theorem is established by linking the three conditions to the three conditions of Theorem 2.4, which then asserts the required posterior contraction statement. To link the conditions, we use a comparison of the Kullback-Leibler numbers and the distance induced by the n-norm in Lemma 2.10. The n-norm quantifies the distances between densities via the corresponding regression functions $w : \mathcal{X} \to \mathbb{R}$. The posterior contraction statement below is formulated with Π the distribution of W itself.

Theorem 2.9. *Let Π be the distribution of W. If there exist sequences $\varepsilon_n \to 0$ and $\bar{\varepsilon}_n \to 0$ with $n(\varepsilon_n^2 \wedge \bar{\varepsilon}_n^2) \to \infty$ and, for every large enough constant C, measurable subsets $B_n \subset \mathbb{B}$ and a constant $D > 0$ such that*

$$\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_n) \leq Dn\bar{\varepsilon}_n^2, \tag{2.25}$$

$$\mathbb{P}(W \notin B_n) \leq \exp(-Cn\varepsilon_n^2), \tag{2.26}$$

$$\mathbb{P}(\|W(x) - w_0(x)\|_n \leq 2\varepsilon_n) \geq \exp(-n\varepsilon_n^2), \tag{2.27}$$

then

$$\Pi(w : \|w - w_0\|_n \geq L_n(\varepsilon_n \vee \bar{\varepsilon}_n)|Y_1, \ldots, Y_n) \xrightarrow{P_0^{(n)}} 0$$

as $n \to \infty$, for every sequence $L_n \to \infty$.

Proof. Let Π be the distribution of $p_W^{(n)}$. We show that

$$\Pi(p_w^{(n)} : \|w - w_0\|_n \geq L_n(\varepsilon_n \vee \bar{\varepsilon}_n)|Y_1, \ldots, Y_n) \xrightarrow{P_0^{(n)}} 0$$

by verifying the conditions of Theorem 2.4.

It is immediately clear from Lemma 2.10 that the average Kullback-Leibler numbers in the definition of the Kullback-Leibler type neighborhood (2.9) are multiples of the squared n-norm distances between the indices. We have

$$\frac{1}{n}\sum_{i=1}^n K(p_{w,i}, p_{0,i}) = \frac{1}{2\sigma_0^2}\|w - w_0\|_n^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n V_0(p_{w,i}, p_{0,i}) = \frac{1}{\sigma_0^2}\|w - w_0\|_n^2$$

and thus

$$\Pi(\mathrm{KL}_n^*(P_0, \tilde{\varepsilon}_n)) \geq \mathbb{P}(\|W - w_0\|_n \leq \sigma_0\tilde{\varepsilon}_n).$$

For $\tilde{\varepsilon}_n$ a sufficiently large multiple of ε_n, this is again bounded from below by

$$\mathbb{P}(\|W - w_0\|_n \leq 2\varepsilon_n) \geq \exp(-n\varepsilon_n^2) \geq \exp(-n\tilde{\varepsilon}_n^2/4)$$

according to (2.27). This shows (2.15).

Define the sets $\mathcal{P}_n^{(n)} = \{p_w \in \mathcal{P}^{(n)} : w \in B_n\}$. As we have seen before, for this choice of sieve the link between the remaining mass conditions (2.26) and (2.16) follows by choosing the constant C large enough.

The average-like Hellinger distance d_n given in (2.8) satisfies $d_n(p_v, p_w) \leq \|v - w\|_n$. From this it follows that assumption (2.25) implies condition (2.17) of Theorem 2.4 with a sufficiently large multiple of the present $\tilde{\varepsilon}_n$.

We now obtain posterior contraction with respect to d_n. However, the assertion still holds if the distance d_n is replaced in the posterior contraction statement by the n-norm. Indeed, the distance d_n can be replaced by $\|\cdot\|_n$ throughout the proof of Theorem 2.4, as is the case for Theorem 2.3, as mentioned in [20]. So we obtain posterior contraction with respect to the n-norm. □

The Kullback-Leibler divergence (2.10) and the Kullback-Leibler variance (2.11) are given in the following lemma.

Lemma 2.10.

$$K(p_{w,i}, p_{0,i}) = \frac{1}{2\sigma_0^2}\big(w_0(x_i) - w(x_i)\big)^2$$
$$V_0(p_{w,i}, p_{0,i}) = \frac{1}{\sigma_0^2}\big(w_0(x_i) - w(x_i)\big)^2$$

Proof. Because $p_{w,i}(y_i)$ is proportional to $\exp(-\frac{1}{2\sigma_0^2}(y_i - w(x_i))^2)$,

$$-\log \tfrac{p_{w,i}}{p_{0,i}}(Y_i) = \tfrac{1}{2\sigma_0^2} w(x_i)^2 - \tfrac{1}{2\sigma_0^2} w_0^2 + \tfrac{1}{\sigma_0^2} w_0(x_i) Y_i - \tfrac{1}{\sigma_0^2} w(x_i) Y_i.$$

Taking the expectation with respect to $p_{0,i}$ substitutes $w_0(x_i)$ for Y_i. Completing the square gives the desired expression for $K(p_{w,i}, p_{0,i})$. For the Kullback-Leibler variance, note that

$$\log \tfrac{p_{w,i}}{p_{0,i}}(Y_i) + K(p_{w,i}, p_{0,i}) = \tfrac{1}{\sigma_0^2}(w_0(x_i) - w(x_i))(w(x_i) - Y_i)$$

and that $V_0(p_{w,i}, p_{0,i})$ is the expectation of its square. Because $\mathbb{E}_0(w(x_i) - Y_i)^2 = \sigma_0^2$ the result follows. □

If the error variance is unknown, then we also endow σ with a prior distribution. We assume that the prior on σ is supported on a given compact subinterval of $(0, \infty)$ that contains the true σ_0, and that this prior distribution has a Lebesgue density which is bounded away from zero. Together with the prior on w, this defines a total prior on the pair $\eta = (\sigma, w)$. We now denote by Π this total prior on the pair of parameters.

According to Theorem 3.3 in van der Vaart and van Zanten [50], under these assumptions on the prior for σ, the posterior distribution of the total prior contracts around the truth (σ_0, w_0) in the following sense under the same conditions as in the case in which the error variance was known.

Theorem 2.11. *Under the conditions of Theorem 2.9,*

$$\Pi((\sigma, w) : \|w - w_0\|_n + |\sigma - \sigma_0| \geq L_n(\varepsilon_n \vee \bar{\varepsilon}_n) | Y_1, \ldots, Y_n) \xrightarrow{P_0^{(n)}} 0$$

as $n \to \infty$ for every sequence $L_n \to \infty$.

2.4.4 Unified approach

Suppose that W is a Borel measurable zero-mean random element in $C([0,1]^d)$ equipped with the supremum norm $\|\cdot\|_\infty$ over $[0,1]^d$. The probability distribution of W can be used as a prior distribution in each of the preceding settings.

The uniform norm $\|\cdot\|_\infty$ is stronger than both $\|\cdot\|_{2,G}$ and $\|\cdot\|_n$. The conditions of Theorem 2.7 and Theorem 2.9 can therefore be verified using the supremum norm to reach the same conclusions. The set of conditions thus becomes

$$\log N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq Dn\bar{\varepsilon}_n^2, \tag{2.28}$$

$$\mathbb{P}(W \notin B_n) \leq \exp(-Cn\varepsilon_n^2), \tag{2.29}$$

$$\mathbb{P}(\|W(x) - w_0(x)\|_\infty \leq 2\varepsilon_n) \geq \exp(-n\varepsilon_n^2). \tag{2.30}$$

Although these conditions are stronger than the conditions of the preceding theorems, they might actually be easier to verify. Furthermore, we see that a posterior contraction result can actually be obtained in three different statistical problems simultaneously under the same conditions.

2.5 Gaussian process priors

In the previous section we have seen under which conditions on a stochastic process prior, we obtain posterior contraction around the truth for various statistical models simultaneously. In particular these results hold for Gaussian process priors. It turns out that in order to obtain posterior contraction in the case of Gaussian process priors, it suffices to have conditions with a single rate. In fact, the conditions obtained in Theorem 2.18 in Section 2.6.1 link to the general posterior contraction conditions of Theorem 2.1 for the density estimation and classification problems, and from Theorem 2.3 for the fixed design regression setting. This has been shown in van der Vaart and van Zanten [50].

Theorem 2.18 thus asserts the conditions for posterior contraction in terms of a Gaussian process and the separable Banach space in which it, seen as a random element in a space of functions, takes its values. This theorem replaces the posterior contraction conditions with another condition, the so-called concentration inequality. In Section 3.4 and Section 4.4 we obtain rates of posterior contraction for two families of Gaussian priors by verifying the concentration inequality.

In this section, we introduce the concepts and results needed to formulate and prove Theorem 2.18. We recall the concept of the reproducing kernel Hilbert space of a Gaussian process and show how it is related to the non-centered small ball probabilities of the prior. In this work, we do not consider the proof of Theorem 2.18 (a verification of a set of three conditions (2.34)–(2.36) similar to those seen in Section 2.4), but we want to introduce the Gaussian process

machinery behind this result in order to verify the conditions (2.28)–(2.30) for the non-Gaussian hierarchical priors in Section 3.5 and Section 4.5, which are in fact built from Gaussian process priors.

2.5.1 The reproducing kernel Hilbert space

We first recall the definition of the reproducing kernel Hilbert space attached to a zero-mean Gaussian process W with index set T. The reproducing kernel Hilbert space $\mathbb{H}$ is defined as the completion of the linear space of functions $t \mapsto \mathbb{E}W(t)H$ relative to the inner product

$$\langle \mathbb{E}W(\cdot)H_1, \mathbb{E}W(\cdot)H_2 \rangle_{\mathbb{H}} = \mathbb{E}H_1H_2,$$

where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i W(s_i)$ with $a_i \in \mathbb{R}$ and $s_i \in T$. Elements h of the reproducing kernel Hilbert space can be identified with functions $h(t)$ on T through the reproducing formula $h(s) = \langle h, \mathbb{E}W(\cdot)W(s) \rangle_{\mathbb{H}}$, with $\mathbb{E}W(\cdot)W(s) = K(s, \cdot)$ for K the covariance function of W.

We now consider the reproducing kernel Hilbert space of a finite sum Gaussian process. Let $\mathbb{B}$ be a Banach space of functions and let $b_1, \ldots, b_J$ be elements of $\mathbb{B}$. Let $Z_1, \ldots, Z_J$ be independent standard normal random variables. The process

$$W(x) = \sum_{j=1}^{J} Z_j b_j(x) \tag{2.31}$$

is a random element in the Banach space $\mathbb{B}$. For any $k \in \mathbb{N}$ and constants $a_1, \ldots, a_k \in \mathbb{R}$ and any choice of k locations $x_1, \ldots, x_k$, the linear combination

$$\sum_{i=1}^{k} a_i W(x_i) = \sum_{j=1}^{J} \big(\sum_{i=1}^{k} a_i b_j(x_i) \big) Z_j$$

is Gaussian. So W is a zero mean Gaussian process. The covariance function $K(x, y) = \mathbb{E}W(x)W(y)$ of such a Gaussian process is

$$K(x, y) = \sum_{j=1}^{J} b_i(x) b_j(y).$$

The RKHS in this case can be found as in Section 4 of the paper [54] by van der Vaart and van Zanten.

Lemma 2.12. *The reproducing kernel Hilbert space $\mathbb{H}$ of the process W in* (2.31) *consists of all functions $x \mapsto \sum_{j=1}^{J} w_j b_j(x)$ with $w_j \in \mathbb{R}$. The RKHS-norm of an element $h \in \mathbb{H}$ is given by $\|h\|_{\mathbb{H}}^2 = \inf_w \sum_{j=1}^{J} w_j^2$, where the infimum is taken over all $w \in \mathbb{R}^J$ for which the representation $h = \sum_{j=1}^{J} w_j b_j$ holds true.*

If $b_1, \ldots, b_J$ are linearly independent, then each $h \in \mathbb{H}$ has a unique representation. The infimum can then be removed from the expression for the RKHS-norm.

2.5.2 Small ball probabilities

Suppose that W is a zero-mean Gaussian random element in the Banach space $\mathbb{B}$ with norm $\|\cdot\|$. We consider the centered small ball probability $\mathbb{P}(\|W\| \leq \varepsilon)$ and define

$$\varphi_0(\varepsilon) = -\log \mathbb{P}(\|W\| \leq \varepsilon) \tag{2.32}$$

as minus the exponent of the centered small ball probability.

We now consider the small ball probability $\mathbb{P}(\|W - w\| \leq \varepsilon)$ around $w \in \mathbb{B}$. It is known that $w \mapsto \mathbb{P}(\|W - w\| \leq \varepsilon)$ is maximal at $w = 0$ for any fixed $\varepsilon > 0$. If $w \in \mathbb{B}$ is contained in the reproducing kernel Hilbert space of W, then even more is known. According to (4.16) of Kuelbs, Li and Linde [32]

$$\mathbb{P}(\|W - h\| \leq \varepsilon) \geq \exp(-\|h\|_{\mathbb{H}}^2/2)\mathbb{P}(\|W\| \leq \varepsilon)$$

or equivalently,

$$-\log \mathbb{P}(\|W - h\| \leq \varepsilon) \leq \frac{1}{2}\|h\|_{\mathbb{H}}^2 + \varphi_0(\varepsilon).$$

The following lemma, given as Lemma 5.3 in [54], extends this result to w contained in the support of W. For W in a separable Banach space $\mathbb{B}$, the support of W is equal to the closure of $\mathbb{H} \subset \mathbb{B}$ in the Banach space $\mathbb{B}$. We can thus think of a realization of W as a limit in $\mathbb{B}$ of a sequence of elements in the reproducing kernel Hilbert space. Consider the infimum

$$\varphi_{\inf}(\varepsilon) = \inf_{h \in \mathbb{H}: \|w-h\| \leq \varepsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2$$

for w in the support of W, and let

$$\varphi_w(\varepsilon) = \varphi_{\inf}(\varepsilon) + \varphi_0(\varepsilon). \tag{2.33}$$

Lemma 2.13. *For any w in the support of W and every $\varepsilon > 0$,*

$$\varphi_w(2\varepsilon) \leq -\log \mathbb{P}(\|W - w\| \leq 2\varepsilon) \leq \varphi_w(\varepsilon).$$

2.5.3 Centered small ball probabilities via metric entropy

In the previous section we considered the centered small ball probabilities of a Gaussian process. In this section, we see that these probabilities can be studied via the metric entropy of the unit ball of the reproducing kernel Hilbert space. Remember that the covering number $N(\varepsilon, A, d)$ of a set A in a metric space with

distance d is the minimum number of balls of radius ε needed to cover A. The metric entropy of A is defined as $\log N(\varepsilon, A, d)$.

Let $\mathbb{H}_1$ be the unit ball in the reproducing kernel Hilbert space of some zero-mean Gaussian random element W in a separable Banach space. We now consider the metric entropy of $\mathbb{H}_1$ with respect to the Banach space norm. The following lemma gives us a bound on the exponent $\varphi_0(\varepsilon)$ of the centered small ball probability of W in (2.32) if the metric entropy of $\mathbb{H}_1$ is suitably bounded from above. This result is an easy consequence of Theorem 1.2 in Li and Linde [34].

Lemma 2.14. *Suppose that for some $0 < \alpha < 2$ and some constant $K > 0$,*

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|) \leq K\varepsilon^{-\alpha}$$

for any sufficiently small $\varepsilon > 0$. Then there exists some constant C such that

$$\varphi_0(\varepsilon) \leq C\varepsilon^{-\frac{2\alpha}{2-\alpha}}$$

for any sufficiently small $\varepsilon > 0$.

The following lemma gives a similar result, only with different upper bounds. This result has been obtained in Corollary 2.4 of Aurzada et al. [1].

Lemma 2.15. *Suppose that for some $\gamma > 0$ and some constant K,*

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|) \leq K(\log 1/\varepsilon)^{\gamma}$$

for any $\varepsilon \leq 1$. Then there exists some constant C such that for any $\varepsilon \leq 1$,

$$\varphi_0(\varepsilon) \leq C(\log 1/\varepsilon)^{\gamma}.$$

2.5.4 Borell's inequality

Suppose that W is a zero-mean Gaussian random element in the separable Banach space $\mathbb{B}$. We considered the probability $\mathbb{P}(\|W - w\| \leq \varepsilon)$ that W maps into a small ball around some w in the support of W. The following result by Borell [7] deals with the probability that W maps into a ball around some $h \in \mathbb{H}$, for any h in the reproducing kernel Hilbert space for which the reproducing kernel Hilbert space norm is bounded by a constant L.

Let $\mathbb{H}_1$ be the unit ball in the reproducing kernel Hilbert space of W and let $\mathbb{B}_1$ be the unit ball in the Banach space. Let Φ be the cumulative distribution function of the standard Gaussian distribution.

Theorem 2.16. *For any $\varepsilon > 0$ and $L \geq 0$,*

$$\mathbb{P}(W \in L\mathbb{H}_1 + \varepsilon\mathbb{B}_1) \geq \Phi(\Phi^{-1}(\mathbb{P}(\|W\| \leq \varepsilon)) + L)$$

This result in the present form is given by Theorem 5.1 of [54]. The following variant of Borell's inequality is given in Proposition A.2.1 of van der Vaart and Wellner [53].

Theorem 2.17. *For any $x > 0$,*

$$\mathbb{P}(\|W\| \geq x) \leq 2 \exp\Big(- \frac{x^2}{8\mathbb{E}\|W\|^2} \Big).$$

This result is used in the proof of Theorem 5.3 for a process with values in the space $C^\alpha([0,1]^d)$ of Hölder functions, equipped with the Hölder norm $\|\cdot\|_\alpha$.

2.6 Posterior contraction for Gaussian priors

2.6.1 General result

In the previous section, we considered the reproducing kernel Hilbert space of a zero mean Gaussian process and showed that the non-centered small ball probabilities around w_0 can be calculated using the centered small ball probabilities and the RKHS norm of an approximation of w_0 by elements in the reproducing kernel Hilbert space. Moreover, the centered small ball probabilities can be calculated via the metric entropy of the reproducing kernel Hilbert space unit ball.

In the following theorem it is shown that in the case of a Gaussian prior, the conditions for posterior contraction around w_0 can be captured into a single condition which is basically the prior mass condition that we have seen before. This condition uses the non-centered small ball probability around w_0 to express that the prior should put at least a minimal amount of probability mass on small neighborhoods around the truth. In this case, we use the concentration function introduced in the previous section to impose this condition, and refer to this single condition as the concentration inequality. This result has been obtained in Theorem 2.1 of van der Vaart and van Zanten [50].

Theorem 2.18. *Let W be a Borel measurable zero-mean Gaussian random element in a separable Banach space $\mathbb{B}$ with norm $\|\cdot\|$. For any sequence ε_n that satisfies the concentration inequality $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ and any large enough contant C, there exists a measurable set $B_n \subset \mathbb{B}$ such that*

$$\mathbb{P}(\|W - w_0\| \leq \varepsilon_n) \geq \exp(-n\varepsilon_n^2) \tag{2.34}$$

$$\mathbb{P}(W \notin B_n) \leq \exp(-Cn\varepsilon_n^2) \tag{2.35}$$

$$\log N(\varepsilon_n, B_n, \|\cdot\|) \leq 6Cn\varepsilon_n^2 \tag{2.36}$$

The condition that C is large enough in this statement means that $C > 1$ and $e^{-Cn\varepsilon_n^2} < 1/2$. Note that for a sequence ε_n such that $n\varepsilon_n^2 \to \infty$ the latter condition is automatically satisfied for any large enough n.

The assertion of Theorem 2.18 can be linked to the three conditions for posterior contraction in the statistical models described in Section 2.2. For the density estimation and classification settings, these conditions are linked to the conditions Theorem 2.1 because the observations in these settings are independent and identically distributed. For the fixed design regression setting, they are linked to Theorem 2.3 because the observations are independent, but not identically distributed.

Although Theorem 2.18 will be used as an asymptotic result as $n \to \infty$, it is in fact a statement for every fixed n. The process W is therefore allowed to depend on n. The corresponding RKHS and concentration function then also depend on n.

2.6.2 Specific statistical settings

Let $\mathbb{B}$ be a separable Banach space that consists of functions on $\mathcal{X}$. Let $\|\cdot\|$ be the Banach space norm. The Banach space might for instance be the space $C([0,1])^d$ of continuous functions on $\mathcal{X} = [0,1]^d$ equipped with the supremum norm. Assume that W is a zero-mean Gaussian random element in $\mathbb{B}$ and that $w_0 : \mathcal{X} \to \mathbb{R}$ is in the support of W. Let φ_{w_0} be the concentration function of W.

We have seen how the prior distributions on the statistical models can be defined in the various statistical settings via the law of W. The posterior contraction results for the various settings in Theorems 2.19, 2.20 and 2.21 ahead can be summarized by saying that the posterior contracts at a rate ε_n around the true function w_0 if ε_n satisfies the concentration inequality $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$.

We now explain for each of the statistical settings what is exactly meant by the statement that the posterior contracts around the truth at a certain rate.

2.6.2.1 Density estimation

We consider a sample $X_1, \ldots, X_n$ from a probability density p_0 on the sample space $\mathcal{X}$. To make inference about the true p_0, we define a prior distribution Π on probability densities p and consider the posterior.

Let $\mathcal{P} = \{p_w : w \in \mathbb{B}\}$ for probability densities p_w defined by

$$p_w(x) = \frac{e^{w(x)}}{\int_{\mathcal{X}} e^{w(x)}\, dx}, \quad x \in \mathcal{X}. \tag{2.37}$$

Suppose that W is a Gaussian random element of $\mathbb{B}$. The distribution of W defines a prior distribution Π on $\mathcal{P}$ via the random variable p_W that takes values $p_w \in \mathcal{P}$ for realizations w of W. Let h be the Hellinger distance between

probability densities. The following theorem explains for the density estimation setting described in Section 2.2.1 what is meant by the posterior contraction that follows from Theorem 2.18.

Theorem 2.19. *Under the conditions of Theorem 2.18,*

$$\Pi(p : h(p, p_0) \geq L\varepsilon_n | X_1, \ldots, X_n) \xrightarrow{P_0^n} 0$$

as $n \to \infty$ for any sufficiently large constant L.

2.6.2.2 Classification

We consider i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i takes values in a sample space $\mathcal{X}$ and Y_i takes values in the set $\{0, 1\}$. The statistical problem is to determine the binary regression function $r_0(x) = \mathbb{P}(Y_i = 1 \,|\, X_i = x)$.

For Ψ the standard logistic or the standard normal distribution function, we define by $r_w(x) = \Psi(w_x)$ a function $r_w : \mathcal{X} \mapsto (0, 1)$ for each realization w of W. This defines a prior on $\mathcal{P} = \{p_w(x, y) = r_w(x)^y(1 - r_w(x))^{1-y} : x \in \mathcal{X}, y \in \{0, 1\}\}$. Let Π now be the prior distribution on binary regression functions given by the distribution of $\Psi(W)$.

The following theorem explains for the classification setting in Section 2.2.2 what is meant by the posterior contraction that follows from Theorem 2.18.

Theorem 2.20. *Under the conditions of Theorem 2.18,*

$$\Pi(r : \|r - r_0\|_{2,G} \geq L\varepsilon_n | X_1, Y_1, \ldots, X_n, Y_n) \xrightarrow{P_0^n} 0$$

as $n \to \infty$ for any sufficiently large constant L.

2.6.2.3 Fixed design regression

Suppose that we observe independent pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, where the x_i are known elements of a sample space $\mathcal{X}$ and the Y_i satisfy the regression relation $Y_i = w_0(x_i) + e_i$, for independent $N(0, \sigma_0^2)$-distributed error variables e_i. The aim is to estimate the regression function w_0. This is the fixed design regression problem as described in Section 2.2.3. In this setting, the process W itself defines a prior Π on the regression functions $w : \mathcal{X} \to \mathbb{R}$.

First assume that the variance σ_0^2 of the error variables is known. The following theorem explains what is meant by the posterior contraction that follows from Theorem 2.18.

Theorem 2.21. *Under the conditions of Theorem 2.18,*

$$\Pi(w : \|w - w_0\|_n \geq L\varepsilon_n | Y_1, \ldots, Y_n) \xrightarrow{P_0^{(n)}} 0$$

as $n \to \infty$ for any sufficiently large constant L.

If the error variance is unknown, then we also endow σ with a prior distribution as before. We assume that the prior on σ is supported on a given compact subinterval of $(0, \infty)$ that contains the true σ_0, and that this prior distribution has a Lebesgue density which is bounded away from zero. Together with the prior on w, this defines a total prior on the pair (σ, w). We now let Π be the total prior on the pair of parameters. The posterior contraction statement that follows from Theorem 2.18 is as follows.

Theorem 2.22. *Under the conditions of Theorem 2.18,*

$$\Pi((\sigma, w) : \|w - w_0\|_n + |\sigma - \sigma_0| \geq L\varepsilon_n | Y_1, \ldots, Y_n) \xrightarrow{P_0^{(n)}} 0$$

as $n \to \infty$ for any sufficiently large constant L.

Chapter 3

Posterior contraction for tensor-product spline priors

3.1 Introduction

In this chapter we consider prior distributions on functions of one or more variables that are constructed using so-called splines. A spline function is a piecewise polynomial function on either an interval of the real line or some multi-dimensional Euclidean space. Spline functions provide good approximations for Hölder smooth functions, see for instance De Boor [6] or Schumaker [42]. Therefore, splines can be a useful tool for constructing prior distributions on smooth functions.

There are a number of papers in the literature that obtain rates of estimation for smooth functions using splines, in particular ones using parametric log spline models in a density estimation setting. It was shown for instance by Stone [46] that a smooth probability density can be estimated at the minimax rate in a log spline model of growing dimension using a maximum likelihood estimator (MLE). This result was extended in [47] to the multivariate case. A Bayesian version of the former result was obtained by Ghosal, Ghosh and Van der Vaart [21]. They consider priors on densities that are constructed by postulating the same log spline model for the density as in Stone and putting an appropriate prior distribution on the coefficients in the so-called B-spline expansion. Ghosal, Ghosh and Van der Vaart [21] show that it is possible to attain the minimax rate as the posterior contraction rate if the log-density is bounded (by a known constant) and satisfies a smoothness condition. Specifically, the results state that in the case that a sample from an unknown univariate density f on an interval is observed, then if $\log f$ is uniformly bounded by a known constant and is r times continuously differentiable, a rate of convergence relative to the Hellinger metric (for the MLE in the case of

Stone [46] and for the posterior in the case of Ghosal, Ghosh and Van der Vaart [21]) of the optimal order $n^{-r/(1+2r)}$ can be attained. Stone [47] also obtains the optimal rate $n^{-r/(d+2r)}$ in the case that f is a d-variate density. The procedures in the cited papers are non-adaptive, in the sense that they rely on knowledge of the smoothness level r of the unknown density.

Rate-adaptive results for spline priors have been obtained by Huang [25] and by Ghosal, Lember and Van der Vaart [23]. The paper [23] deals with univariate density estimation again. Instead of letting the dimension J of the log-spline expansion tend to infinity with sample size in a deterministic manner, the "model index" J is viewed as a hyper-parameter and is endowed with an additional prior. Put differently, the density estimation problem is viewed as a model selection problem: a sequence of finite-dimensional log-spline models for the density is considered, each with their own (finite-dimensional) prior. Then appropriate prior weights are assigned to each of the models to obtain an overall prior for f. The resulting hierarchical prior does not depend on the regularity r of the density f and is rate-adaptive: it yields a posterior contraction rate of the order $n^{-r/(1+2r)}$ if $\log f$ is r times continuously differentiable. Huang [25] presents a very similar result, but with more complicated prior weights on the finite-dimensional models. This is accompanied by a similar result in a univariate nonparametric regression context. The two settings in [25] are not treated in a unified approach however. Priors weights for the models are chosen separately for each case.

A joint feature of the approaches of [25] and [23] is that both the order and the knots of the splines (see the next section for definitions of these notions) are changing between models. In view of the approximation properties of splines (see Section 3.2), allowing the orders of the splines to become arbitrarily large is indeed necessary when adaption to arbitrarily large smoothness levels is desired. On the other hand, it makes the priors rather involved and possibly less attractive from the computational perspective.

Our approach and the results we derive complement and extend the existing literature in a number of directions. First of all, we do not study specific settings like density estimation separately. Instead, we present general theorems about random spline processes (Theorems 3.6 and 3.10) that, in combination with existing general rate of contraction results for specific statistical settings (cf. Chapter 2) lead to concrete results for, for instance, density estimation, regression, or classification. As an illustration we work out the details for these three particular setting, but results for other nonparametric problems could be derived as well. In combination with the general theory in [36] for instance, results for nonparametric estimation problems in diffusion models could be obtained.

Secondly, we consider multivariate function estimation problems. Similar to what Stone [47] did for the frequentist approach, we show that sensible priors on multivariate functions can be constructed using tensor-product splines. We

prove that adaptive, rate-optimal procedures for multivariate function estimation problems can be obtained in this way.

Another difference concerns the fact the existing approaches in [21], [25] and [23] assume known uniform bounds on the log-density or the regression function that is being estimated, allowing the use of bounded priors on the B-spline coefficients. As is indicated in [23] this restriction could be removed by adding another hierarchical layer, treating the bound as an additional hyper-parameter. In our approach this is not necessary however and we do not need to assume any uniform bounds. This is a consequence of the fact that we use unbounded, namely Gaussian prior weights on the B-spline coefficients. In our rates we get additional logarithmic factors, which might in part be due to this issue.

Finally, we keep the order of the splines that we use fixed in the construction of the prior. Only the number of knots is viewed as a hyper-parameter, which we either send to infinity with sample size or endow with a prior. As a result our priors are simpler and conceivably also computationally more attractive. On the down side, with this approach we can not obtain adaption up to arbitrary high smoothness levels, but only up to the order of the splines that are used. Since we can freely choose this order however, we feel this is not a serious restriction.

As mentioned already, we build our spline priors from random splines with independent, Gaussian B-spline coefficients. We keep the order of the splines fixed and treat the number of knots as a hyper-parameter. The latter will be either deterministic, or endowed with a second, independent prior. As a result, the priors we construct will be (transformations of) Gaussian of conditionally Gaussian process priors. This allows us to use the rich machinery described in Chapter 2 for their analysis.

The remainder of this chapter is organized as follows. In Section 3.2 we review the notions of spline functions and B-splines, and formulate a result that gives a bound on the uniform distance between splines and a given smooth function. In Section 3.3 we define our spline prior with Gaussian coefficients. The connected reproducing kernel Hilbert space turns out to be the whole spline space, and the approximation result from Section 3.2 helps us to determine the concentration function. This allows us to obtain posterior contraction rates for the Gaussian random spline priors. In Section 3.4 we show that optimal posterior rates (up to logarithmic factors) can be achieved by letting the number of knots tend to infinity with the sample size in an appropriate way. In Section 3.5 we present a hierarchical procedure by choosing a prior distribution on the partition size hyper-parameter. We show that this hierarchical procedure also achieves a near-optimal rate of posterior contraction and adapts to the smoothness of the truth.

3.2 Preliminaries

We first introduce the concept of spline functions. We follow the definitions given by Schumaker [42]. We only consider a special setting in this chapter and refer to the book by Schumaker for an exhaustive treatment of the subject.

3.2.1 Spline functions on intervals

A spline function or *spline* is a piecewise polynomial function. Let us first consider spline functions defined on an interval. The domain of such a spline function can be partitioned into disjoint subintervals in such a way that the function coincides with a polynomial on every subinterval. Spline functions that share the same partition form a linear space. In the following we just speak of spline functions from a linear space and it should be understood that these splines always share the same partition.

A spline function is said to be of *order* q if all polynomials in its definition are of degree at most $q-1$. Without any further requirements, this set of piecewise polynomials is a linear space of dimension qm, where m is the number of partitioning intervals. Linear subspaces of lower dimension can be obtained by further imposing that adjacent polynomials are tied together smoothly at the knots of the partition.

In this chapter we use splines of order q that satisfy such a smoothness condition. We consider a space S_m of splines of order q on the unit interval that is partitioned into m subintervals of equal length. We first define

$$P_q = \Big\{ x \mapsto \sum_{k=0}^{q-1} c_k x^k, c_0, \ldots, c_{q-1} \in \mathbb{R} \Big\}$$

to be the space of polynomials of degree at most $q-1$. Let $y_j = j/m$ and denote the corresponding subintervals of $[0,1]$ by $I_j = [y_{j-1}, y_j)$ for $j = 1, \ldots, m-1$, and $I_m = [y_{m-1}, 1]$. A function $s : [0,1] \to \mathbb{R}$ is then defined to be in S_m if there exist polynomials $p_1, \ldots, p_m$ in P_q such that $s(x) = p_j(x)$ for $x \in I_j$ and, moreover, s is $q-2$ times continuously differentiable[1]. According to the terminology of [42], S_m is the space of *polynomial splines of order q with simple knots at the points* $1/m, 2/m, \ldots, (m-1)/m$. We will always take $q \geq 2$, so that all the splines in S_m are continuous functions.

The space S_m has dimension $q+m-1$, cf. Theorem 4.4 of [42]. A convenient basis of the space is given by the so-called *B-splines*. The exact definition of these functions (see Theorem 4.9 of Schumaker [42]) is of no importance to us. Important properties of B-splines are that they are nonnegative and supported on

[1] Here -1 times continuous differentiability is an empty condition and 0 times just means continuity.

relative small parts of the domain and that the sum of all B-splines at any given location equals one, i.e. they form a partition of unity: if we denote the B-splines by $B_1, \ldots, B_{q+m-1}$, then

$$\sum_{j=1}^{q+m-1} B_j(x) = 1$$

for all $x \in [0,1]$.

3.2.2 Tensor-product splines

Spline functions can also be defined on multi-dimensional domains using multivariate polynomials. One can construct linear spaces of such multivariate splines by taking tensor-products of the spline spaces mentioned above. This just means that we associate a direction with every linear space in the tensor product, that we introduce a different variable for each direction, and that we then multiply polynomials of a single variable defined on intervals to obtain multivariate polynomials defined on rectangles.

The space of tensor-product splines is spanned by the *tensor-product B-splines*, which are just products of the B-splines associated with the different directions. The dimension of the tensor-product space is thus found by multiplying the dimensions of the spline spaces from which it was constructed. The properties of univariate B-splines carry over to similar properties for their tensor-product analogues.

In the following we consider tensor-product splines from the d-fold tensor product space $\mathcal{S}_m = S_m \otimes \cdots \otimes S_m$ (d times), with S_m the space of univariate splines defined above. The tensor-product splines are thus defined on the unit cube $[0,1]^d$ in the Euclidean space of dimension d and this unit cube is partitioned into m^d equal cubes $I_{k_1} \times \ldots \times I_{k_d}$. On every such set the splines coincide with a polynomial of the form

$$\sum_{k_1=0}^{q-1} \cdots \sum_{k_d=0}^{q-1} c_{k_1,\ldots,k_d} x_1^{k_1} \cdots x_d^{k_d}. \tag{3.1}$$

The space $\mathcal{S}_m$ is of dimension $(q+m-1)^d$ and a basis is given by the tensor-product B-splines

$$B_j(x_1, \ldots, x_d) = B_{j_1}(x_1) \cdots B_{j_d}(x_d), \quad 1 \le j_i \le q+m-1.$$

From now on these multivariate B-splines are denoted by $B_1, \ldots, B_J$ for $J = (q+m-1)^d$. It is easy to see that we again have the partition of unity property

$$\sum_{j=1}^{J} B_j(x) = 1 \tag{3.2}$$

for all $x \in [0,1]^d$.

The total degree of a polynomial of the form (3.1) is the maximum of $k_1 + \ldots + k_d$ over all k for which the coefficient c_k is nonzero. The total degree of these polynomials is thus at most $d(q-1)$, but not any polynomial of total degree at most $d(q-1)$ is an element of $\mathcal{S}_m$. This is only true if the degree in each single variable $x_1, \ldots, x_d$ is at most $q-1$. In particular the polynomials of total *order* q are in $\mathcal{S}_m$, i.e. the polynomials of the form (3.1) with $c_k = 0$ if $|k| > q-1$. The approximating properties of such polynomials determine the approximating capabilities of the tensor-product splines in $\mathcal{S}_m$, see Lemma 3.1 ahead.

This approximation result is proved using a dual basis of the tensor-product space. Given a set of linear functionals $\lambda_j : \mathcal{S}_m \to \mathbb{R}$, we say that $\lambda_1, \ldots, \lambda_J$ is a dual basis of $\mathcal{S}_m$ if

$$\lambda_i(B_j) = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

for any $i, j = 1, \ldots, J$. For the spline $s \in \mathcal{S}_m$ given by

$$s = \sum_{j=1}^{J} a_j B_j \tag{3.3}$$

we have that $\lambda_j(s) = a_j$. Thus λ_j finds the coefficient belonging to the B-spline B_j.

3.2.3 Approximation properties

The following result describes how well splines in the space $\mathcal{S}_m$ can approximate functions with a smoothness level r that does not exceed the order q of the splines. We first explain what the appropriate notion of smoothness is in this situation.

Let $C([0,1]^d)$ be the space of continuous functions $f : [0,1]^d \to \mathbb{R}$ and denote the supremum norm of f over $[0,1]^d$ by $\|f\|_\infty$. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$, we define $|\alpha| = \alpha_1 + \ldots + \alpha_d$ and the partial derivative

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

For $r \in \mathbb{N}$, we define the Hölder space $C^r([0,1]^d)$ of all functions $f \in C([0,1]^d)$ with partial derivatives $D^\alpha f \in C([0,1]^d)$ for any $|\alpha| \leq r$, and we equip it with the norm

$$\|f\|_{C^r} = \|f\|_\infty + \sum_{\alpha : |\alpha| = r} \|D^\alpha f\|_\infty.$$

The lemma below gives an upper bound on the uniform distance of a function $f \in C^r([0,1]^d)$ and some spline in $\mathcal{S}_m$. The distance can be controlled by choosing the partition size m sufficiently large. The proof of the lemma is similar to the proof

of Theorem 12.7 in Schumaker [42]. We only need to apply the multidimensional Taylor expansion in Theorem 13.18 of Schumaker with a total Taylor expansion (13.33) in [42] instead of a tensor Taylor expansion (13.44), so that this expansion produces a polynomial of total order r.

Lemma 3.1. *For any $m, d, q \in \mathbb{N}$, $r \leq q$, and $f \in C^r([0,1]^d)$ there exists a spline $s \in \mathcal{S}_m$ and a constant $C > 0$ that only depends on d, q and r such that*

$$\|f - s\|_\infty \leq Cm^{-r} \sum_{\alpha:|\alpha|=r} \|D^\alpha f\|_\infty.$$

Proof. Let Q be the bounded linear operator in (12.29) of [42] that maps $C^r([0,1]^d)$ onto $\mathcal{S}_m$. It is given by $Qf(x) = \sum_{j=1}^J \lambda_j(f) B_j(x)$, for λ_j the (extensions of the) elements of the dual space of $\mathcal{S}_m$ given in Theorem 12.5 of [42]. Let H be a hypercube in the partition of $[0,1]^d$ and let $\|\cdot\|$ be the supremum over H. We will bound $\|f - Qf\|$ from above. It is obvious that $\|f - Qf\|_\infty$ is then bounded from above by the maximum of these bounds for the various cubes in the partition.

We have $\|Qf\| \leq C\|f\|$ for any $f \in C^r([0,1]^d)$ according to (12.31) of [42]. The constant C does not depend on the cube H as can be seen from (12.25) of [42], but it does depend on q. According to Theorem 13.18 of [42] there exists a polynomial $p = p_j$ of total order r such that $\|f - p\| \leq Dm^{-r} \sum_{\alpha:|\alpha|=r} \|D^\alpha f\|$ for some constant D that only depends on d, r and thus not on H. We have $Qp = p$ (see (12.30) in [42]) and hence $\|f - Qf\| \leq \|f - p\| + \|Q(f - p)\| \leq (C+1)\|f - p\|$. □

3.2.4 The size of a spline and its coefficients

In Section 3.3 we will use the fact that a smooth function can be approximated by a spline in $\mathcal{S}_m$ in the sense of Lemma 3.1. For our purposes, we do not need to know the approximating spline or its coefficients in full detail, but rather an expression that quantifies its size. We will use the following lemma, which states that the uniform norm of a spline is equivalent to the maximal norm of the vector of its B-spline coefficients.

Recall that the B-spline coefficients of a spline can be obtained from a dual basis of $\mathcal{S}_m$. We now assume that $\lambda_1, \ldots, \lambda_J$ is the dual basis given in Theorem 12.5 of [42]. Let $\|\lambda_j\|$ be the norm of the bounded linear functional λ_j. That is to say, $\|\lambda_j\|$ is the smallest constant K for which $|\lambda_j(s)| \leq K\|s\|_\infty$ holds for any $s \in \mathcal{S}_m$. Although $\max_{1 \leq j \leq J} \|\lambda_j\|$ depends on m, it can actually be replaced by a constant that does not depend on m, cf. Theorem 12.5 in [42].

Lemma 3.2. *Let $s \in \mathcal{S}_m$ be given by* (3.3)*. Then*

$$\|s\|_\infty \leq \max_{1 \leq j \leq J} |a_j| \leq \big(\max_{1 \leq j \leq J} \|\lambda_j\| \big) \|s\|_\infty \leq C\|s\|_\infty,$$

where $C > 0$ is a constant independent of m.

Proof. Because the B-splines are positive, $|s(x)| \leq \sum_{j=1}^{J} |a_j| B_j(x)$. Take the maximum of the absolute values $|a_j|$ outside the sum. The first inequality now follows from the partition of unity property (3.2). For the second inequality, use that $|a_j| \leq \|\lambda_j\| \|s\|_\infty$, by definition. The third inequality follows from Theorem 12.5 in [42]. □

3.3 Gaussian random splines

In this section we introduce and study a class of Gaussian processes that we will use to construct prior distributions for various statistical settings. The corresponding posterior contraction rates will be determined in Section 3.4. We use the tensor-product splines from the preceding section to define the stochastic process via its sample paths.

We have seen that the space $\mathcal{S}_m$ of tensor-product splines depends on two parameters q and m. The parameter q is the order of the splines and m quantifies the partition size. We fix some natural number $q \geq 2$ and from now on it will be understood that all splines are of order q. The remaining parameter m will simply be referred to as the partition size parameter.

Let $B_1, \ldots, B_J$ be the tensor-product B-spline basis of $\mathcal{S}_m$. Remember that in this notation it is hidden that not only the number of B-splines depends on the partition size m (we have $J = (m+q-1)^d$), but that also the B-splines themselves depend on this number. The sample paths of our process will be tensor-product splines in $\mathcal{S}_m$. In other words, the process can be seen as a random element in the tensor-product spline space $\mathcal{S}_m$.

For any $m \in \mathbb{N}$ we now define the Gaussian random element W^m in $\mathcal{S}_m$ as follows. Let $Z_1, \ldots, Z_J$ be independent, standard Gaussian random variables, and let W^m be the random process on $[0,1]^d$ defined by

$$W^m(x) = \sum_{j=1}^{J} Z_j B_j(x), \qquad x \in [0,1]^d. \tag{3.4}$$

We thus let W^m be a finite sum Gaussian process as in (2.31) in Section 2.5.1.

It follows from Lemma 2.12 that the reproducing kernel Hilbert space $\mathbb{H}^m$ of W^m consists of all splines of order q with respect to the given partition, and that the RKHS-norm of a such a spline is equal to the Euclidean norm of the vector of its B-spline coefficients. In other words, the reproducing Kernel Hilbert space of W^m is equal to the set $\mathcal{S}_m$ equipped with the norm $\|\cdot\|_{\mathbb{H}^m}$ given by

$$\Big\| \sum_{j=1}^{J} a_j B_j \Big\|_{\mathbb{H}^m}^2 = \sum_{j=1}^{J} a_j^2. \tag{3.5}$$

As we have seen in Chapter 2, the contraction rate of a posterior corresponding to a Gaussian process prior is determined by its concentration function, i.e. its

non-centered small ball probabilities around the truth. The concentration function can be determined from the centered small ball probabilities of the process in addition to a term that quantifies the size of an approximation of the truth in the reproducing kernel Hilbert space of the process. We study these two quantities in the next two subsections.

3.3.1 Centered small ball probabilities

The following lemma is a straightforward consequence of the definition of the process W^m and the basic properties of the B-splines.

Lemma 3.3. *For all $q, m \in \mathbb{N}$ such that $m \geq q-1$,*

$$\mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \geq (\varepsilon/2)^{2^d m^d}$$

for all $\varepsilon \in (0, 1/2)$.

Proof. By Lemma 3.2 and the fact that the random variables Z_j are independent and identically distributed we have

$$\mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \geq \mathbb{P}(\max|Z_j| \leq \varepsilon) = (\mathbb{P}(|Z_1| \leq \varepsilon))^J.$$

The probability $\mathbb{P}(|Z_1| \leq \varepsilon)$ is bounded from below by an area of width 2ε and height $\varphi(\varepsilon)$, with φ the probability density of a standard Gaussian random variable. Since $J = (q+m-1)^d \leq (2m)^d$ for $m \geq q-1$, it follows that for any $\varepsilon \in (0, 1/2)$ and any $q \geq 1$ and $m \geq q-1$,

$$(\mathbb{P}(|Z_1| \leq \varepsilon))^J \geq (2\varphi(1/2)\varepsilon)^{2^d m^d}$$

This proves the assertion, since $2\varphi(1/2) \geq 1/2$. □

3.3.2 Non-centered small ball probabilities

Consider $w_0 \in C^r([0,1]^d)$. The non-centered ball probability $\mathbb{P}(\|W^m - w_0\|_\infty \leq 2\varepsilon)$ is the probability that a realization of W^m ends up in a uniform ball of radius 2ε around w_0. This probability can be determined using the result in Section 2.5.2. In the following we actually derive a lower bound for the given probability using this approach. The result is presented in the next lemma.

Lemma 3.4. *Let $w_0 \in C^r([0,1]^d)$ for $r \leq q$. There exist constants $C, D > 0$ independent of m, such that for any $\varepsilon \in (0, 1/2)$ and any $m \in \mathbb{N}$ such that $Dm^{-r} \leq \varepsilon$,*

$$\mathbb{P}(\|W^m - w_0\|_\infty \leq 2\varepsilon) \geq \exp(-Cm^d \log(1/\varepsilon)).$$

Let $\varphi_{w_0}^m$ be the concentration function of W^m around w_0 as defined in (2.33). Then by Lemma 2.13,

$$\mathbb{P}(\|W^m - w_0\|_\infty \leq 2\varepsilon) \geq \exp(-\varphi_{w_0}^m(\varepsilon))$$

and a similar inequality holds for the upper bound. Now Lemma 3.4 is a consequence of the following result.

Lemma 3.5. *Let $w_0 \in C^r([0,1]^d)$ for $r \leq q$. There exist constants $C, D > 0$ independent of m, such that for any $\varepsilon \in (0, 1/2)$ and any $m \in \mathbb{N}$ such that $Dm^{-r} \leq \varepsilon$,*

$$\varphi_{w_0}^m(\varepsilon) \leq Cm^d \log(1/\varepsilon). \tag{3.6}$$

Proof. The concentration function is given by

$$\varphi_{w_0}^m(\varepsilon) = \inf_{h \in \mathbb{H}^m : \|w_0 - h\|_\infty \leq \varepsilon} \|h\|_{\mathbb{H}^m}^2 - \log \mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \tag{3.7}$$

in the present notation. The second term of the concentration function can be bounded from above using Lemma 3.3. For $\varepsilon \in (0, 1/2)$ we have

$$-\log \mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \leq 2^d m^d \log\Big(\frac{2}{\varepsilon}\Big). \tag{3.8}$$

As for the infimum part in (3.7), Lemma 3.1 shows that for every $m \in \mathbb{N}$ there exists a spline $s \in \mathcal{S}_m = \mathbb{H}^m$ such that $\|s - w_0\|_\infty \leq Dm^{-r}$, for $D > 0$ a constant that only depends on d, q, r and w_0. Now fix $\varepsilon \in (0, 1/2)$ and $m \in \mathbb{N}$ such that $Dm^{-r} \leq \varepsilon$. Then with s the spline above,

$$\inf_{h \in \mathbb{H}^m : \|w_0 - h\|_\infty \leq \varepsilon} \|h\|_{\mathbb{H}^m}^2 \leq \|s\|_{\mathbb{H}^m}^2.$$

Suppose that the spline $s \in \mathcal{S}_m$ is given by $s = \sum_{j=1}^J a_j B_j$. Then the squared RKHS-norm of s is given by (3.5) and satisfies

$$\|s\|_{\mathbb{H}^m}^2 = \sum_{j=1}^J a_j^2 \leq J\big(\max_{1 \leq j \leq J} |a_j|\big)^2.$$

We have seen in Lemma 3.2 that the absolute maximum $\max_{1 \leq j \leq J} |a_j|$ of the coefficients can be bounded from above by $C'\|s\|_\infty$ for some $C' > 0$ that does not depend on m. Note that by the triangle inequality and the fact that $Dm^{-r} \leq \varepsilon$, we have that $\|s\|_\infty \leq \|w_0\|_\infty + \varepsilon$. Since $J \leq (2m)^d$, we obtain an upper bound for $\|s\|_{\mathbb{H}^m}^2$ that can be written as a multiple of m^d. This concludes the proof. □

3.4 Posterior contraction for Gaussian spline priors

3.4.1 General result

The Gaussian spline processes W^m can be used to construct priors in various nonparametric statistical settings. In order for the priors to have large enough support to ensure for instance consistency, one has to either let the partition size parameter m tend to infinity with the sample size, or view it as a hyper-parameter that itself is estimated from the data. In this section we consider the former construction, leading to sequences of Gaussian process priors. We give bounds on the contraction rates of the corresponding posteriors. In the next section we investigate the possibility of endowing m with a prior distribution.

Let $m_n \to \infty$ be a sequence of natural numbers, fix an order $q \geq 2$ for the splines and consider the corresponding sequence W^{m_n} of Gaussian spline processes on $[0,1]^d$. For a natural number $r \leq q$ and $w_0 \in C^r([0,1])^d$, let $\varphi_{w_0}^{m_n}$ be the sequence of concentration functions defined by (3.7), with $\mathbb{H}^m$ the RKHS of the process W^m. The general theory of Gaussian process priors says that posterior contraction rates are obtained by solving the inequality

$$\varphi_{w_0}^{m_n}(\varepsilon_n) \leq n\varepsilon_n^2, \tag{3.9}$$

see Section 2.6. By Lemma 3.5 this inequality holds if

$$\begin{aligned} Cm_n^d \log m_n &\leq n\varepsilon_n^2, \\ Dm_n^{-r} &\leq \varepsilon_n, \end{aligned}$$

with $C, D > 0$ the constants from the statement of the lemma. The optimal solution of these inequalities is easily found and given in the following theorem.

Theorem 3.6. *In the setting described above, let* $m_n \sim (n/\log n)^{1/(d+2r)}$. *Then inequality (3.9) holds with* $\varepsilon_n \sim (n/\log n)^{-r/(d+2r)}$.

In combination with the results given in Section 2.6 this theorem immediately yields rate of contraction results for a number of important non-parametric statistical problems. We give details in the next section. Generally speaking, the results show that if the law of the Gaussian spline process W^{m_n} is used as a prior on an r-regular function of d variables, then with the choice $m_n \sim (n/\log n)^{1/(d+2r)}$ this leads to a posterior contraction rate of the order $n^{-r/(d+2r)}$, up to a logarithmic factor. This is typically the optimal rate for estimating an r-regular function of d variables, for instance in a minimax sense. Note however that through the partition size parameter m_n, the prior depends on the unknown smoothness level of the function of interest. Hence, the procedure is not rate-adaptive. In Section 3.5 we construct a hierarchical, conditionally Gaussian prior that does lead to adaption.

3.4.2 Gaussian regression

Suppose we observe independent pairs $(x_1, Y_1), \ldots, (x_1, Y_n)$, where the x_i are known elements of $[0,1]^d$ and the Y_i satisfy the regression relation $Y_i = w_0(x_i) + e_i$, for independent $N(0, \sigma_0^2)$-distributed error variables e_i. The aim is to estimate the regression function w_0.

In this case the spline process W^{m_n} can be used directly as a prior for w_0. If the standard deviation σ_0 of the errors is unknown, we endow it with a prior distribution as well, which we assume to be supported on a given interval $[a,b] \subset (0,\infty)$ that contains σ_0, with a Lebesgue density that is bounded away from zero. The total prior is denoted by Π_n.

We denote the corresponding posterior distribution by $\Pi_n(\cdot \,|\, Y_1, \ldots, Y_n)$. Let $\|w\|_n = \big(n^{-1}\sum_{i=1}^n w^2(x_i)\big)^{1/2}$ be the L_2-norm corresponding to the empirical distribution of the design points. We say that the posterior contracts at rate ε_n in this case if, for every sufficiently large L,

$$\Pi_n\Big((w,\sigma) : \|w - w_0\|_n + |\sigma - \sigma_0| \geq L\varepsilon_n \,|\, Y_1, \ldots, Y_n\Big) \xrightarrow{P_0^{(n)}} 0 \tag{3.10}$$

as $n \to \infty$.

Combining Theorems 3.6 and 2.21 yields the following result.

Theorem 3.7. *If $w_0 \in C^r([0,1]^d)$ for $r \leq q$ and $m_n \sim (n/\log n)^{1/(d+2r)}$, then the posterior contracts at the rate $\varepsilon_n \sim (n/\log n)^{-r/(d+2r)}$.*

3.4.3 Density estimation

After exponentiation and renormalization a Gaussian process can be used as a prior model for probability densities as well.

We consider a sample $X_1, \ldots, X_n$ from a continuous, positive density f_0 on the unit cube $[0,1]^d \subset \mathbb{R}^d$. As a prior distribution Π_n on f_0 we use the distribution of

$$x \mapsto \frac{e^{W^{m_n}(x)}}{\int_{[0,1]^d} e^{W^{m_n}(x)}\, dx}. \tag{3.11}$$

Let $\Pi_n(f \in \cdot \,|\, X_1 \ldots, X_n)$ denote the posterior distribution. We say that the posterior contracts at rate ε_n if, for every sufficiently large constant L, as $n \to \infty$,

$$\Pi_n\Big(f : h(f, f_0) \geq L\varepsilon_n \,|\, X_1, \ldots, X_n\Big) \xrightarrow{P_0^n} 0. \tag{3.12}$$

Here h is the Hellinger distance and the convergence is understood to be in probability under the (frequentist) assumption that $X_1, \ldots, X_n$ is a random sample from f_0.

Combining Theorems 3.6 and 2.19 yields the following result.

Theorem 3.8. *If $\log f_0 \in C^r([0,1]^d)$ for $r \leq q$ and $m_n \sim (n/\log n)^{1/(d+2r)}$, then the posterior contracts at the rate $\varepsilon_n \sim (n/\log n)^{-r/(d+2r)}$.*

3.4.4 Classification

As a last concrete case we consider binary regression, or classification. Here we have i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i takes values in the unit cube $[0,1]^d$ and Y_i takes values in the set $\{0,1\}$. The statistical problem is to estimate the binary regression function $r_0(x) = \mathbb{P}(Y_1 = 1 \,|\, X_1 = x)$.

As a prior Π_n on r_0 we use the law of the process $\Psi(W^{m_n})$, where $\Psi : \mathbb{R} \to (0,1)$ is the logistic or the normal distribution function. Let $\Pi_n\big(\cdot \,|\, (X_1, Y_1), \ldots, (X_n, Y_n)\big)$ denote the posterior and let $\|\cdot\|_{L_2(G)}$ be the L_2-norm relative to the marginal distribution G of X_1. We say that the posterior contracts at rate ε_n if, for every sufficiently large L,

$$\Pi_n(r : \|r - r_0\|_{L_2(G)} \geq L\varepsilon_n \,|\, X_1, Y_1, \ldots, X_n, Y_n) \xrightarrow{P_0^n} 0. \tag{3.13}$$

Combining Theorems 3.6 and 2.20 yields the following result.

Theorem 3.9. *If $\Psi^{-1}(r_0) \in C^r([0,1]^d)$ for $r \leq q$ and $m_n \sim (n/\log n)^{1/(d+2r)}$, then the posterior contracts at the rate $\varepsilon_n \sim (n/\log n)^{-r/(d+2r)}$.*

3.5 Adaptation using conditionally Gaussian priors

3.5.1 General result

In the previous section we have seen, for instance in the regression setting, that under a certain smoothness condition on the truth w_0, posterior contraction can be achieved at an optimal rate for an appropriate sequence of our Gaussian spline priors. We assumed that w_0 is contained in $C^r([0,1]^d)$ for a given $r \leq q$ and used the knowledge of the degree of regularity r to define a sequence of Gaussian priors via the partition size parameter m_n.

In practice however, the exact degree of smoothness is typically not known a-priori. Therefore, in this section we will only assume that for $q \geq 2$ fixed in advance, w_0 is contained in $C^r([0,1]^d)$ for r some unknown smoothness level such that $r \leq q$. In other words, we only assume a known upper bound on the smoothness. The aim now is to construct a prior independent of r such that the posterior achieves the same optimal rate as in the preceding section (perhaps up to a logarithmic factor) for every possible value of r. Such a procedure is said to adapt to the regularity of the truth up to the level q.

As before we take the Gaussian spline process W^m as the starting point for the definition of our priors. However, we now take a different approach to choosing m. In the Bayesian paradigm it is quite common to view unknown tuning parameters of this type as so-called hyper parameters and to endow them with a separate prior,

leading to hierarchical priors. We adopt this approach and show that if the prior on m is chosen carefully, we can achieve our goal of constructing a rate-adaptive procedure in this way.

Concretely, we define a new, conditionally Gaussian spline process W by setting $W = W^M$, for W^m the Gaussian process defined in (3.4) and M an independent $\mathbb{N}$-valued random variable. (Note there is a slight chance of confusion here, since the B-splines in the definition of W^m depend on m as well. In the definition of W, m also has to be substituted with the random M in those places.) This construction is hierarchical in the sense that a sample path of W is generated in two steps: first draw a realization m of the random variable M, then given m, draw a sample path of the Gaussian process W^m.

The hierarchical spline process can be used to construct priors for various statistical settings again. We consider our usual examples in the next subsection. The following general theorem about the process W will lead to the desired adaptive rate of contraction results.

Theorem 3.10. *Suppose that for every $m \geq 1$,*

$$C_1 \exp(-D_1 m^d \log^t m) \leq \mathbb{P}(M = m) \leq C_2 \exp(-D_2 m^d \log^t m) \tag{3.14}$$

for some constants $C_1, C_2, D_1, D_2, t \geq 0$. If $w_0 \in C^r([0,1]^d)$ for some integer $r \leq q$, then there exists for every constant $C > 0$, a constant $D > 0$ and measurable subsets B_n of $C([0,1]^d)$ such that

$$\mathbb{P}(\|W - w_0\|_\infty \leq 2\varepsilon_n) \geq \exp(-n\varepsilon_n^2), \tag{3.15}$$

$$\mathbb{P}(W \notin B_n) \leq \exp(-Cn\varepsilon_n^2), \tag{3.16}$$

$$\log N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq Dn\bar{\varepsilon}_n^2, \tag{3.17}$$

are satisfied for sufficiently large n, and for ε_n and $\bar{\varepsilon}_n$ given by

$$\varepsilon_n = c(n/\log^{1\vee t} n)^{-\frac{r}{d+2r}} \qquad \bar{\varepsilon}_n = n^{-\frac{r}{d+2r}} (\log n)^{\frac{(1\vee t)r}{d+2r} + (\frac{1-t}{2})_+}, \tag{3.18}$$

for $c > 0$ a large enough constant.

Combined with results from Chapter 2 this general theorem will lead to various results that state that in the different settings we consider, we will have posterior contraction at the rate $\varepsilon_n \vee \bar{\varepsilon}_n$ with this prior, provided that the true function has smoothness degree $r \leq q$. Hence, up to a logarithmic factor, the posteriors attain optimal convergence rates in this case. Moreover, since the prior does not depend on the unknown smoothness level r, we indeed obtain rate-adaptive procedures.

Note that condition (3.14) holds in particular, for $t = 0$, if M^d has a geometric distribution. The best rate $\varepsilon_n \vee \bar{\varepsilon}_n$ is obtained when t is chosen equal to 1. The resulting rate is $(n/\log n)^{-\frac{r}{d+2r}}$ in that case, which coincides with the rate obtained in Theorem 3.6 for the non-adaptive sequence of spline priors.

In our approach the order q of the splines remains fixed, contrary to for instance [25] or [23]. This keeps the priors simple and easy to deal with, but of course in practice q has to be chosen. From the theoretical perspective q can be chosen as large as one would like, although it might be chosen not too large for computational reasons. In practice, cubic splines ($q = 4$ in our notation) are a popular choice.

3.5.2 Results for specific statistical settings

Combined with general results presented in Chapter 2, Theorem 3.10 yields rate of contraction results for the hierarchical prior in the three different statistical settings described also in Section 3.4: density estimation, fixed design regression and classification. In this section we briefly state the respective results for these cases.

Consider first the regression setting described in Section 2.2.3. As prior on the regression function w_0 we now employ the law of the conditionally Gaussian process W, where the hyper prior on M is assumed to satisfy condition (3.14). The total prior on the pair (w_0, σ_0), with σ_0 the error standard deviation, is denoted by Π. As before we say that the corresponding posterior contracts at the rate ε_n if for all L large enough (3.10) holds as $n \to \infty$, with Π in the place of Π_n.

Combining Theorem 3.10 and Theorem 2.11 yields the following result for fixed design regression.

Theorem 3.11. *If the prior on m satisfies* (3.14) *and for the true regression function we have $w_0 \in C^r([0,1]^d)$ for $r \leq q$, then the posterior contracts at the rate*

$$n^{-\frac{r}{d+2r}} (\log n)^{\frac{(1\vee t)r}{d+2r} + (\frac{1-t}{2})_+}.$$

For density estimation we consider the setting described in Section 2.2.1 again. As prior on the density function f_0 we now employ the law Π of the random density

$$x \mapsto \frac{e^{W(x)}}{\int_{[0,1]^d} e^{W(x)}\, dx},$$

with W the conditionally Gaussian process, where the hyper prior on m is assumed to satisfy condition (3.14). We say that the corresponding posterior contracts at the rate ε_n if for all L large enough (3.12) holds as $n \to \infty$, with Π in the place of Π_n.

Combining Theorem 3.10 and Theorem 2.5 yields the following result for density estimation.

Theorem 3.12. *If the prior on m satisfies* (3.14) *and for the true density we have $\log f_0 \in C^r([0,1]^d)$ for $r \leq q$, then the posterior contracts at the rate*

$$n^{-\frac{r}{d+2r}} (\log n)^{\frac{(1\vee t)r}{d+2r} + (\frac{1-t}{2})_+}.$$

Finally, we consider the non-parametric classification problem described in Section 2.2.2. As prior on the binary regression function r_0 we employ the law Π of $\Psi(W)$, with Ψ the logistic or normal distribution function and W the conditionally Gaussian process, where the hyper prior on M is assumed to satisfy condition (3.14). We say that the corresponding posterior contracts at the rate ε_n if for all L large enough (3.13) holds as $n \to \infty$, with Π in the place of Π_n.

Combining Theorem 3.10 and Theorem 2.7 yields the following result for classification.

Theorem 3.13. *If the prior on m satisfies* (3.14) *and for the true binary regression function we have $\Psi^{-1}(r_0) \in C^r([0,1]^d)$ for $r \le q$, then the posterior contracts at the rate*

$$n^{-\frac{r}{d+2r}} (\log n)^{\frac{(1\vee t)r}{d+2r} + (\frac{1-t}{2})_+}.$$

3.5.3 Proof of the general Theorem 3.10

3.5.3.1 Prior mass condition (3.15)

Let $\varepsilon_n \to 0$ be given. Note that the inequality

$$\mathbb{P}(\|W - w_0\|_\infty \le 2\varepsilon_n) \ge \mathbb{P}(M = m)\mathbb{P}(\|W^m - w_0\| \le 2\varepsilon_n)$$

holds for any $m \ge 1$ by construction of W. According to Lemma 3.4 the second factor on the right is bounded from below by $\exp(-Cm_n^d \log m_n)$ for sufficiently large n and m_n such that $\varepsilon_n \ge Dm_n^{-r}$. The probability $\mathbb{P}(M = m_n)$ is bounded from below by $C_1 \exp(-D_1 m_n^d \log^t m_n)$ by assumption (3.14). We conclude that

$$\mathbb{P}(\|W - w_0\|_\infty \le 2\varepsilon_n) \ge C_1 \exp(-C_2 m_n^d \log^{1\vee t} m_n)$$

for some constants $C_1, C_2 > 0$. The inequalities

$$m_n^d \log^{1\vee t} m_n \lesssim n\varepsilon_n^2,$$
$$m_n^{-r} \lesssim \varepsilon_n,$$

are solved by $m_n \sim (n/\log^{1\vee t} n)^{1/(d+2r)}$ and ε_n as in (3.18). Condition (3.15) thus holds if the constant c in (3.18) is sufficiently large.

3.5.3.2 Construction of sieves B_n

Recall that $\mathbb{H}_1^m$ is the unit ball of the RKHS $\mathbb{H}^m$ of the Gaussian spline process W^m and $\mathbb{B}_1$ is the unit ball in the Banach space $C([0,1]^d)$. For $m \in \mathbb{N}$, let $B_n^m = L_n\mathbb{H}_1^m + \varepsilon_n\mathbb{B}_1$ for some k_n and L_n specified below, and $B_n = \bigcup_{m=1}^{k_n} B_n^m$.

In the next two subsections we show that conditions (3.16) and (3.17) are fulfilled if L_n and k_n satisfy certain inequalities. In Subsection 3.5.3.5 we show that these inequalities can be solved.

3.5.3.3 Remaining mass condition (3.16)

First note that the inequality

$$\mathbb{P}(W \notin B_n) \leq \sum_{m=1}^{k} \mathbb{P}(M = m)\mathbb{P}(W^m \notin B_n) + \mathbb{P}(M \geq k+1). \tag{3.19}$$

holds for any k by construction of W. Now take k equal to k_n as defined in the preceding subsection. By assumption (3.14) the tail probability $\mathbb{P}(M \geq k_n + 1)$ is bounded from above by a constant times the geometric series

$$\sum_{m \geq k_n+1} (\exp(-k_n^{d-1} \log^t k_n))^m \leq \exp(-k_n^d \log^t k_n).$$

So the tail probability is bounded by $\exp(-Cn\varepsilon_n^2)/2$ for large n if k_n is chosen such that $k_n^d \log^t k_n > Cn\varepsilon_n^2$, for C as in the assertion of the theorem.

We now show that

$$\mathbb{P}(W^m \notin B_n) \leq \exp(-Cn\varepsilon_n^2)/2$$

for any $m \leq k_n$, so that the first term on the right of (3.19) is also bounded by $\exp(-Cn\varepsilon_n^2)/2$. It follows from the construction of the sieve B_n that

$$\mathbb{P}(W^m \notin B_n) \leq \mathbb{P}(W^m \notin B_n^m)$$

for any $m \leq k_n$. By Borell's inequality (see Theorem 2.16),

$$\mathbb{P}(W^m \notin B_n^m) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W^m\|_\infty \leq \varepsilon_n)) + L_n).$$

A lower bound for the centered small ball probability $\mathbb{P}(\|W^m\|_\infty \leq \varepsilon_n)$ was given in Lemma 3.3. The lower bound provided by this lemma is a decreasing function of m. For every $m \leq k_n$ we thus have

$$\mathbb{P}(\|W^m\|_\infty \leq \varepsilon_n) \geq (\varepsilon/2)^{2^d k_n^d}.$$

For $y \in (0, 1/2)$ one has $\Phi^{-1}(y) \geq -\sqrt{(5/2)\log(1/y)}$. Apply this inequality with y equal to $(\varepsilon/2)^{2^d k_n^d}$ to find that

$$\mathbb{P}(W^m \notin B_n^m) \leq \Phi\Big(\sqrt{(5/2)2^d k_n^d \log(2/\varepsilon_n)} - L_n\Big)$$

for every $m \leq k_n$. Using the bound $\Phi(y) \leq \exp(-y^2/2)$ we obtain

$$\mathbb{P}(W^m \notin B_n) \leq e^{-\frac{1}{2}\left(L_n - \sqrt{(5/2)2^d k_n^d \log(2/\varepsilon_n)}\right)^2} \tag{3.20}$$

for every $m \leq k_n$. Hence if L_n and k_n are chosen such that

$$\frac{1}{2}\left(L_n - \sqrt{(5/2)2^d k_n^d \log(2/\varepsilon_n)}\right)^2 > Cn\varepsilon_n^2,$$

then the first term on the right of (3.19) is bounded by $\exp(-Cn\varepsilon_n^2)/2$ as well.

3.5.3.4 Proof of entropy condition (3.17)

Let $\bar{\varepsilon}_n$ be given by (3.18). Because B_n is a union of the sets B_n^m for $m = 1, \ldots, k_n$, its $2\bar{\varepsilon}_n$-covering number satisfies

$$N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq \sum_{m=1}^{k_n} N(2\bar{\varepsilon}_n, B_n^m, \|\cdot\|_\infty).$$

If $A_1, \ldots, A_N$ is a minimal covering of $\mathbb{H}_1^m$ using balls of radius $\bar{\varepsilon}_n/L_n$, then the sets $L_n A_i + \varepsilon_n \mathbb{B}_1$ are balls of radius $\bar{\varepsilon}_n + \varepsilon_n \leq 2\bar{\varepsilon}_n$ which cover B_n^m. This shows that

$$N(2\bar{\varepsilon}_n, B_n^m, \|\cdot\|_\infty) \leq N(\bar{\varepsilon}_n/L_n, \mathbb{H}_1^m, \|\cdot\|_\infty). \tag{3.21}$$

We now identify splines in $\mathbb{H}^m$ with points in $\mathbb{R}^J$ via the B-spline coefficients. Then $\mathbb{H}_1^m$ corresponds to the unit ball in $\mathbb{R}^J$ (see (3.5)). Moreover, for a spline $s = \sum a_j B_j$ in $\mathbb{H}^m$ we have that the uniform norm $\|s\|_\infty$ is bounded by the Euclidean norm $\|a\|$ of the vector of B-spline coefficients, by Cauchy-Schwarz and the basic properties of the B-splines. It follows that the covering number on the right of (3.21) is bounded by the $\bar{\varepsilon}_n/L_n$-covering number of the unit ball in $\mathbb{R}^J$ relative to the Euclidean distance. The latter is bounded from above by $(6L_n/\bar{\varepsilon}_n)^J$ according to e.g. Lemma 4.1 of Pollard [38].

We thus find

$$N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq k_n (6L_n/\bar{\varepsilon}_n)^{2^d k_n^d}$$

and consequently, if $L_n = O(n^p)$ for some $p > 0$, we have

$$\log N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq D k_n^d \log n$$

for some positive constant D. So if k_n is taken such that $k_n^d \log n$ is bounded by a multiple of $n\bar{\varepsilon}_n^2$, then condition (3.17) holds.

3.5.3.5 End of the proof of Theorem 3.10

The preceding subsections show that the proof of Theorem 3.10 is complete once we show that there exist sequences L_n and k_n such that

$$k_n^d \log^t k_n > C n \varepsilon_n^2 \tag{3.22}$$

$$\bar{\varepsilon}_n \geq \varepsilon_n \tag{3.23}$$

$$k_n^d \log n \leq C' n \bar{\varepsilon}_n^2 \tag{3.24}$$

$$L_n - \sqrt{(5/2) 2^d k_n^d \log(2/\varepsilon_n)} > \sqrt{2C n \varepsilon_n^2} \tag{3.25}$$

$$L_n = O(n^p), \tag{3.26}$$

where C is a given positive constant and p and C' may be chosen arbitrarily.

We have $n\varepsilon_n^2 = c^2 n^{d/(d+2r)} (\log n)^{(2r(1\vee t))/(d+2r)}$, hence (3.22) is fulfilled if

$$k_n^d = An^{\frac{d}{d+2r}} (\log n)^{\frac{2r(1\vee t)}{d+2r} - t},$$

with A a large enough positive constant. Conditions (3.23) and (3.24) are then fulfilled as well if C' is chosen large enough, by definition of the sequence $\bar{\varepsilon}_n$. Finally, conditions (3.25) and (3.26) are then easily taken care of by taking L_n to be a large enough power of n.

Chapter 4

Posterior contraction for location-scale mixture priors

4.1 Introduction

In this chapter we consider prior distributions on functions of one or more variables that are constructed using location-scale kernel mixtures. The use of such mixtures of kernels is well established for the construction of nonparametric priors on probability densities. The methodology is used in a variety of practical settings, and in recent years there has been substantial progress on the the mathematical, asymptotic theory for kernel mixture priors as well, cf. [16, 18, 19, 31, 48, 56]. At the present time we have a well-developed understanding of important aspects including consistency, convergence rates, rate-optimality, and adaptation properties. A similar, parallel development has taken place in the area of beta mixture priors, cf. [17, 30, 37, 41].

A discrete location-scale mixture of a fixed probability density p on $\mathbb{R}^d$ can be expressed as

$$x \mapsto \sum_{j=1}^{m} w_j \frac{1}{\sigma^d} p\Big(\frac{x - x_j}{\sigma}\Big), \tag{4.1}$$

where $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \mathbb{R}^d$, $w_1, \ldots, w_m \geq 0$ and $\sum w_j = 1$, and $\sigma > 0$. A prior on densities is obtained by putting prior distributions on m, the locations x_j, the scale σ and the weights w_j. When p satisfies some regularity conditions, a wide class of probability densities can be well approximated by mixtures of the form (4.1). This indicates that if the priors on the coefficients are suitably chosen, the resulting prior and posterior on probability densities can be expected to have good asymptotic properties. The cited papers give precise conditions under which this is indeed the case.

Obviously, a much wider class of functions is well approximated by mixtures of the form (4.1) if we lift the restriction that the weights w_j should be nonnegative and sum up to 1. This suggests that location-scale mixtures might be attractive priors not just in the setting of density estimation, but for instance also in nonparametric regression. Although this idea has been proposed in the applied literature, cf. e.g. [24, 45], it does not seem to have attracted a great deal of attention. The few examples do show however that the approach can yield quite satisfactory results.

In the paper [45], location-scale mixture priors are used in an astrophysical setting for the analysis of data from galatic radio sources. The statistical problem essentially boils down to a bivariate, nonparametric, fixed design regression problem. The use of a mixture prior is natural in that particular application because it reflects the idea that the function of interest, which describes the strength of the magnetic field caused by our planet and its "neighborhood" in space, is in fact an aggregate of contributions from a large number of locations, with different weights, which can be positive or negative.

Another reason for using a location-scale mixture prior in multivariate regression, instead of for instance the popular Gaussian squared exponential or Matérn priors, are computational advantages. Conditional on the gridsize m the prior only involves finitely many terms, so no artificial truncation or approximation is necessary for computation. As argued also in [45], the mixture prior allows to avoid the inversion or decomposition of non-trivial and often ill-behaved $n \times n$ matrices (with n the sample size), which can become cumbersome already for moderate sample sizes (cf. also the discussion in [2]). In the astrophysical application of [45], the sample size is of the order 1500 and it is shown that samples of this order can be dealt with effectively using kernel mixture priors.

On the theoretical side, little or nothing is known for kernel mixture priors in a regression setting. In this chapter we therefore take up the study of asymptotic properties, in order to assess the fundamental potential of the methodology and to provide a theoretical underpinning of its use in practice. We will show that if the kernel and the priors on locations and scales are appropriately chosen, kernel mixture priors yield posteriors with good asymptotic properties. It is well known that for the estimation of an α-regular function of d variables, the best possible rate of convergence is of the order $n^{-\alpha/(d+2\alpha)}$, where n is the number of observations available. We will prove that up to a logarithmic factor, this optimal rate can be attained with location-scale mixture priors. More importantly, the near optimal rate can be achieved by a prior that does *not* depend on the unknown smoothness level α of the regression function. In other words, we can obtain a fully rate-adaptive procedure.

The bounds for the convergence rates that we will obtain depend crucially on the smoothness of the kernel p that is used. For kernels with only a finite degree

of regularity, we get sub-optimal rates. We only obtain the optimal minimax rate (up to a logarithmic factor) for kernels that are infinitely smooth, in the sense that they admit an analytic extension to a strip in complex space. The standard normal kernel is an example of an optimal choice in this respect. We also have to put (mild) conditions on the priors on the grid size m and the scale σ. In particular, the popular inverse gamma choice for the scale is included in our setup.

Perhaps surprising is the fact that although we use a probability density p to construct the mixtures, we can still achieve adaptation to all smoothness levels. Intuition from kernel estimation might suggest that when p is a centered probability density, we have good approximation behaviour for regression functions with regularity at most 2, and that for more regular functions we should use higher order kernels. This turns out not to be the case however. To prove this fact we adapt an observation of J. Rousseau, who uses a similar idea to prove that for densities on the unit interval, using appropriate mixtures of beta densities yields adaptation to all smoothness levels, see [41]. In the present work we extend the technique to a multivariate setting (see Lemma 4.10 ahead). The paper [31], which was written at the same time and independently of the paper [26] on which this chapter is based, employs the same idea to prove adaptation for kernel mixture priors for density estimation.

The location-scale priors that we consider are (conditionally) Gaussian since we will put Gaussian priors on the mixing weights. This allows us to use the machinery for Gaussian process priors developed in [50, 54] and outlined in Chapter 2. We will obtain general results for (conditionally) Gaussian kernel mixture process prior, which can be used in a variety of statistical settings. To illustrate this we present rate of contraction results not just for nonparametric regression, which was our main motivation, but also for density estimation and classification settings.

The remainder of this chapter is organized as follows. In Section 4.3 we define a Gaussian process whose paths are location-scale kernel mixtures. We determine the reproducing kernel Hilbert space of the process and its centered and non-centered small ball probabilities. In Section 4.4 we determine in a unified manner the posterior contraction rate of the corresponding posterior in a variety of statistical settings using the general theory for Gaussian process priors. We see that it is possible for such a procedure to obtain an optimal rate of posterior contraction if the kernels in the mixture are chosen from a collection of infinitely regular kernels. In Section 4.5 we see that it is possible to construct a hierarchical procedure based on these Gaussian kernel mixtures which is also rate-optimal and adapts to the smoothness of the truth.

4.2 Auxiliary results

4.2.1 Definition of the spaces $C_R^\alpha(\mathcal{X})$ and $\mathcal{G}_\sigma$.

In Section 4.3.1 we will determine the centered small ball probabilities of the Gaussian process (4.6) via the connection with the metric entropy of the reproducing kernel Hilbert space unit ball (see Section 2.5.3) by embedding this unit ball into one of the two spaces defined in this section. The metric entropy for these spaces is considered in Section 4.2.2.

For a bounded $\mathcal{X} \subset \mathbb{R}^d$ and $\alpha > 0$, we define the space $C^\alpha(\mathcal{X})$ following van der Vaart and Wellner [53]. For any multi-index $k = (k_1, \ldots, k_d)$ consisting of integers $k_i \geq 0$, we define $|k| = k_1 + \ldots + k_d$ and let D^k be the $|k|$-th order differential operator

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}.$$

For r a positive integer, we say that a function $f : \mathcal{X} \to \mathbb{R}$ has bounded partial derivatives up to order r if $D^k f$ exists and is bounded for every multi-index k with $|k| \leq r$. For $\gamma \in (0, 1]$, a function $g : \mathcal{X} \to \mathbb{R}$ is Lipschitz of order γ if

$$\sup_{x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|^\gamma} < \infty.$$

Let $\underline{\alpha}$ be the largest integer strictly smaller than α. The space $C^\alpha(\mathcal{X})$ consists of all functions $f : \mathcal{X} \to \mathbb{R}$ with uniformly bounded partial derivatives up to order $\underline{\alpha}$ such that the highest order partial derivatives are uniformly Lipschitz of order $\alpha - \underline{\alpha}$. For $f \in C^\alpha(\mathcal{X})$ we let

$$\|f\|_\alpha = \max_{|k| \leq \underline{\alpha}} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{|k| \leq \underline{\alpha}} \sup_{x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}}.$$

For $R > 0$, we define $C_R^\alpha(\mathcal{X})$ as the space of functions $f \in C^\alpha(\mathcal{X})$ with $\|f\|_\alpha \leq R$.

For constants $K, \sigma > 0$ we define the space $\mathcal{G}_\sigma$ as follows. First, we define the strip $S_\sigma = \{z \in \mathbb{C}^d : |\operatorname{Im} z_j| \leq \sigma \text{ for } i = 1, \ldots, d\}$. The space $\mathcal{G}_\sigma$ consists of all analytic functions on S_σ which are bounded by $K\sigma^{-d}$.

4.2.2 Metric entropy of $C_R^\gamma([0,1]^d)$ and $\mathcal{G}_\sigma$.

The following lemma gives an upper bound on the metric entropy of $C_R^\gamma([0,1]^d)$ with respect to the supremum norm. The result is well known, see for instance Theorem 2.7.1 of van der Vaart and Wellner [53].

Lemma 4.1. *If $\gamma < \infty$, then*

$$\log N(\varepsilon, C^\gamma_{\sigma^{-(d+\gamma)}\|p\|_\gamma}([0,1]^d), \|\cdot\|_\infty) \leq K_0 \Big(\frac{1}{\varepsilon \sigma^{d+\gamma}}\Big)^{\frac{d}{\gamma}}$$

for all $\sigma, \varepsilon > 0$, with K_0 a constant independent of ε and σ.

The following lemma gives an upper bound on the metric entropy of $\mathcal{G}_\sigma$ with respect to the supremum norm over $[0,1]^d$.

Lemma 4.2. *There exist $\varepsilon_0, \sigma_0 > 0$ such that*

$$\log N(\varepsilon, \mathcal{G}_\sigma, \|\cdot\|_\infty) \le K_1 \frac{1}{\sigma^d} \Big(\log \frac{K_2}{\varepsilon \sigma^d} \Big)^{1+d}$$

for $\varepsilon \in (0, \varepsilon_0)$ and $\sigma \in (0, \sigma_0)$, with constants $K_1, K_2 > 0$ that do not depend on ε or σ. For $\sigma > \sigma_0$, it holds that

$$\log N(\varepsilon, \mathcal{G}_\sigma, \|\cdot\|_\infty) \le K_3 \Big(\log \frac{1}{\varepsilon} \Big)^{1+d}$$

for all $\varepsilon \in (0, \varepsilon_0)$, with $K_3 > 0$ a constant independent of ε and σ.

The statement is similar to the classical result given by Theorem 23 of [29], which gives the entropy for the class of analytic functions bounded by a constant on a strip in complex space. However, the proof of the present statement requires extra care to identify the role of σ, because it should not be considered as an irrelevant constant in our framework.

Proof. We prove the first statement of the lemma. The proof of the second statement is similar, and in fact easier. For $a, b > 0$, let $\mathscr{F}_{a,b}$ be the set of functions that are analytic on the strip $S_a = \{(z_1, \dots, z_d) \in \mathbb{C}^d : |\operatorname{Im} z_i| \le a \text{ for } i = 1, \dots, d\}$, and uniformly bounded by the constant b on that set. We first derive an entropy bound for $\mathscr{F}_{a,b}$ relative to the uniform norm $\|\cdot\|_\infty$ on the unit cube $[0,1]^d$.

We construct a net of piecewise polynomials. Fix an $r < a/2$ and let $R = 2r$. Let $t_1, \dots, t_n$ be a minimal r-net for the cube $[0,1]^d$ relative to the maximum norm. For $j = 1, \dots, n$, let $D_j = \{z \in \mathbb{C}^d : |\Re z_i - (t_j)_i| < r, |\operatorname{Im} z_i| < r \text{ for } i = 1, \dots, d\}$. Observe that the sets D_j cover $[0,1]^d$, that $D_j \subset S_a$ and that $n \le \text{const} \times (1/a)^d$ for a small enough.

Consider a function $f \in \mathscr{F}_{a,b}$. The function is analytic on D_j and hence, by the Cauchy formula, it holds that

$$f(z) = \sum_{k=0}^{\infty} \sum_{|l|=k} c_l (z - t_j)^l \tag{4.2}$$

for $z \in D_j$, where

$$c_l = \frac{1}{(2\pi i)^d} \oint_{C_1} \cdots \oint_{C_d} \frac{f(z)}{(z - t_j)^{|l|+1}} \, dz_1 \cdots dz_d,$$

with C_i a circle of radius R around the ith coordinate $(t_j)_i$ of t_j. (The second sum in (4.2) ranges over all $l \in \mathbb{N}_0^d$ such that $|l| = l_1 + \cdots + l_d = k$ and for $x \in \mathbb{R}^d$ and

$l \in \mathbb{N}_0^d$ we write x^l for $x_1^{l_1} \cdots x_d^{l_d}$.) Since the circles lie inside S_a by construction and f is bounded by b on S_a, we have

$$|c_l| \leq \frac{b}{R^{|l|}} \tag{4.3}$$

for every $l \in \mathbb{N}_0^d$. Consequently, we have, for a universal constant K_1,

$$\sup_{x \in D_j \cap [0,1]^d} \Big| f(x) - \sum_{k=1}^{m} \sum_{|l|=k} c_l (x - t_j)^l \Big| \leq b \sum_{k>m} \frac{k^{d-1}}{2^k} \leq b K_1 \Big(\frac{2}{3}\Big)^m.$$

It follows that

$$\log N(\varepsilon, \mathscr{F}_{a,b}, \|\cdot\|_\infty) \leq n \log N(\varepsilon, \mathcal{P}_m, \|\cdot\|_\infty),$$

where $\mathcal{P}_m$ is the collection of polynomials $p(x) = \sum_{k \leq m} \sum_{|l|=k} c_k x^k$ on $[-r, r]^d$, with m such that $bK_1(2/3)^m \leq \varepsilon$ and coefficients c_k satisfying (4.3). It is well known, and easily verified, that

$$\log N(\varepsilon, \mathcal{P}_m, \|\cdot\|_\infty) \leq m^d \log\Big(\frac{K_2 b}{\varepsilon}\Big)$$

for a universal constant $K_2 > 0$, see for instance the proof of Lemma 4.5 of [51]. We find that there exist constants $K_0, K_1 > 0$ such that

$$\log N(\varepsilon, \mathscr{F}_{a,b}, \|\cdot\|_\infty) \leq K_0 \frac{1}{a^d} \Big(\log \frac{K_1 b}{\varepsilon}\Big)^{d+1}$$

for $\varepsilon, a > 0$ small enough and b large enough. To complete the proof of the lemma, substitute $a = \sigma$ and $b = K\sigma^{-d}$. □

4.2.3 Centered small ball probabilities via metric entropy

The following results are used in Section 4.3.1 with the reproducing kernel Hilbert space unit ball $\mathbb{H}_1^{m,\sigma}$ embedded in respectively $C_R^\gamma([0,1]^d)$ and $\mathcal{G}_\sigma$ (for any m). Lemma 4.3 follows from Lemma 2.14. The required upper bound on the metric entropy has been obtained in Lemma 4.1.

Lemma 4.3. *Suppose that for some $\gamma > d/2$ and some constant K,*

$$\log N(\varepsilon, \mathbb{H}_1^{m,\sigma}, \|\cdot\|) \leq K \Big(\frac{1}{\varepsilon \sigma^{d+\gamma}}\Big)^{-d/\gamma}$$

for any sufficiently small $\varepsilon > 0$. Then there exists some constant C such that

$$-\log \mathbb{P}(\|W^{m,\sigma}\| \leq \varepsilon) \leq C \Big(\frac{1}{\varepsilon \sigma^{d+\gamma}}\Big)^{-\frac{2d}{2\gamma - d}}$$

for any sufficiently small ε.

Lemma 4.4 below follows by arguing as in the proof of Lemma 4.6 in van der Vaart and van Zanten [51]. It is used for $\mathbb{H}_1^{m,\sigma}$ with σ sufficiently small. The required upper bound on the metric entropy is provided by Lemma 4.2.

Lemma 4.4. *Suppose that for some constant K,*

$$\log N(\varepsilon, \mathbb{H}_1^{m,\sigma}, \|\cdot\|) \leq K \frac{1}{\sigma^d} \Big(\log \frac{1}{\varepsilon \sigma^d} \Big)^{1+d} \tag{4.4}$$

for any sufficiently small ε. Then for any $\sigma_0 > 0$ there exist some constants $C, \varepsilon_0 > 0$ such that

$$-\log \mathbb{P}(\|W^{m,\sigma}\| \leq \varepsilon) \leq C \frac{1}{\sigma^d} \Big(\log \frac{1}{\varepsilon \sigma^{1+d}} \Big)^{1+d} \tag{4.5}$$

for any $\sigma < \sigma_0$ and $\varepsilon < \varepsilon_0$.

For $\mathbb{H}_1^{\sigma}$ with $\sigma > \sigma_0$ we use Lemma 4.5 given below. This result follows trivially from Lemma 2.15. The required upper bound on the metric entropy is provided by Lemma 4.2.

Lemma 4.5. *Suppose that for some constant K,*

$$\log N(\varepsilon, \mathbb{H}_1^{m,\sigma}, \|\cdot\|) \leq K (\log 1/\varepsilon)^{1+d}$$

for any sufficiently small ε. Then there exists some constant C such that for any sufficiently small ε,

$$-\log \mathbb{P}(\|W^{m,\sigma}\| \leq \varepsilon) \leq C (\log 1/\varepsilon)^{1+d}.$$

4.3 Gaussian location-scale mixtures

In this section we define the Gaussian process that we will use later to construct prior distributions for different statistical settings. We determine the concentration function of the process so that we can compute posterior contraction rates for the corresponding Gaussian prior in Section 4.4. The paths of the process are given by certain location-scale kernel mixtures. The kernels in the mixture are equipped with an index γ which quantifies their smoothness. We will see that this quantity influences the rate of posterior contraction that we obtain for the corresponding mixture prior.

We first define the collection $\mathcal{P}_\gamma$ of γ-regular kernels. An integrable function p on $\mathbb{R}^d$ that integrates to one and has finite moments of every order is contained in the collection $\mathcal{P}_\gamma$ if it is uniformly Lipschitz on $\mathbb{R}^d$ and satisfies one of the following conditions, depending on whether $\gamma < \infty$ or $\gamma = \infty$:

- For $\gamma < \infty$: p belongs to $C^\gamma(\mathbb{R}^d)$.

- For $\gamma = \infty$: p is the restriction to $\mathbb{R}^d$ of a function that is defined on the set $S = \{(z_1, \ldots, z_d) \in \mathbb{C}^d : |\operatorname{Im} z_j| \leq 1 \text{ for } j = 1, \ldots, d\}$, and that is bounded and analytic on S.

Examples of kernels belonging to this collection $\mathcal{P}_\gamma$ for $\gamma < \infty$ are abundant. Using Fourier inversion it is not difficult to see that an integrable function p belongs to $\mathcal{P}_\infty$ if it has a characteristic function

$$\psi(\lambda) = \int_{\mathbb{R}^d} e^{i(\lambda,x)} p(x)\, dx$$

which is infinitely often differentiable at 0, which satisfies $\psi(0) = 1$, and which satisfies the exponential moment condition

$$\int_{\mathbb{R}^d} e^{\|\lambda\|} |\psi(\lambda)|\, d\lambda < \infty.$$

The prime example is the standard normal density on $\mathbb{R}^d$, which is easily seen to belong to $\mathcal{P}_\infty$. Note that we do not require that $p \geq 0$ in the definition of $\mathbb{P}_\gamma$. So in fact, higher order kernels are allowed as well.

We now introduce the Gaussian kernel mixture process. For fixed $m \in \mathbb{N}$ and $\sigma > 0$, we define the stochastic process $W^{m,\sigma}$ on $[0,1]^d$ by

$$W^{m,\sigma}(x) = \sum_{k \in \{1,\ldots,m\}^d} Z_k \frac{1}{m^{d/2}} \frac{1}{\sigma^d} p\big(\frac{x - k/m}{\sigma}\big), \tag{4.6}$$

for independent standard Gaussian random variables Z_k and a function $p : \mathbb{R}^d \to \mathbb{R}$ in the class of γ-regular kernels $\mathcal{P}_\gamma$ for some $\gamma > d/2$. With the restriction $\gamma > d/2$ we accomplish that the sum in (4.6) is well-defined if the sum is taken over all $k \in \mathbb{N}^d$ and this allows us to obtain bounds for the process $W^{m,\sigma}$ that are independent of m.

The following lemma describes the reproducing kernel Hilbert space of the process $W^{m,\sigma}$. It is an immediate consequence of Lemma 2.12.

Lemma 4.6. *The reproducing kernel Hilbert space $\mathbb{H}^{m,\sigma}$ of $W^{m,\sigma}$ consists of all the functions of the form*

$$h(x) = \sum_{k \in \{1,\ldots,m\}^d} w_k \frac{1}{\sigma^d} p\Big(\frac{x - k/m}{\sigma}\Big), \quad x \in [0,1]^d, \tag{4.7}$$

where the weights w_k range over the entire set of real numbers. The RKHS-norm is given by

$$\|h\|^2_{\mathbb{H}^{m,\sigma}} = m^d \min_w \sum_{k \in \{1,\ldots,m\}^d} w_k^2, \tag{4.8}$$

where the minimum is over all weights w_k for which the representation (4.7) holds true.

We remark that if the functions $x \mapsto p((x-k/m)/\sigma)$ on $[0,1]^d$ are linearly independent, then the representation (4.7) of an element of the RKHS is necessarily unique and hence the minimum in (4.8) can be removed. Although it is for our purpose not important that these functions are independent for every fixed σ and m, the following lemma gives a sufficient condition for the case $\gamma = \infty$.

Lemma 4.7. *Suppose that $p \in \mathcal{P}_\infty$ and that the characteristic function of p has no zeros in $\mathbb{R}^d$. Then for every $\sigma > 0$ and $m \in \mathbb{N}$, the functions $[0,1]^d \ni x \mapsto p((x-k/m)/\sigma)$, $k \in \{1,\dots,m\}^d$, are linearly independent.*

Proof. Suppose that for real constants c_k, we have

$$f(x) = \sum_{k\in\{1,\dots,m\}^d} c_k p\Big(\frac{x-k/m}{\sigma}\Big) = 0$$

for all $x \in [0,1]^d$. Then since p is analytic on a set containing $\mathbb{R}^d$, the function f in fact vanishes on all of $\mathbb{R}^d$. Hence, for $\lambda \in \mathbb{C}^d$ we have

$$\begin{aligned} 0 &= \int_{\mathbb{R}^d} e^{i(\lambda,x)} f(x)\,dx \\ &= \sum_{k\in\{1,\dots,m\}^d} c_k \int_{\mathbb{R}^d} e^{i(\lambda,x)} p\Big(\frac{x-k/m}{\sigma}\Big)\,dx \\ &= \sigma^d \psi(\sigma\lambda) \sum_{k\in\{1,\dots,m\}^d} c_k e^{i(\lambda,k/m)}, \end{aligned}$$

where ψ is the characteristic function of the density p. Hence, since ψ has no zeros on $\mathbb{R}^d$ by assumption and the functions $\lambda \mapsto e^{i(\lambda,k/m)}$ on $\mathbb{R}^d$ are linearly independent, it follows that $c_k = 0$ for all k. □

4.3.1 Centered small ball probabilities

We now consider the centered small ball probabilities of the process $W^{m,\sigma}$. The results in Section 2.5.3 allow us to determine these probabilities via the metric entropy of the unit ball in the reproducing kernel Hilbert space. To find an upper bound for its metric entropy, we embed this unit ball in an appropriate space of functions for which the metric entropy relative to the supremum norm is essentially known. We do this separately for both $\gamma < \infty$ and $\gamma = \infty$.

4.3.1.1 The case $\gamma < \infty$

First we consider the case $\gamma < \infty$. Let h be an element of $\mathbb{H}^{m,\sigma}$. By Lemma 4.6, it admits a representation (4.7) with the weights w_k such that $\|h\|^2_{\mathbb{H}^{m,\sigma}} = m^d \sum w_k^2$. If $p \in \mathcal{P}_\gamma$ with $\gamma < \infty$, we get that $h \in C^\gamma([0,1]^d)$ and

$$\|h\|_\gamma \le \sigma^{-(d+\gamma)} \|p\|_\gamma \|h\|_{\mathbb{H}^{m,\sigma}}. \tag{4.9}$$

Hence, we have $\mathbb{H}_1^{m,\sigma} \subset C_R^\gamma([0,1]^d)$ in this case, with $R = \sigma^{-(d+\gamma)}\|p\|_\gamma$.

In Lemma 4.1 we have seen an upper bound on the metric entropy of $C_R^\gamma([0,1]^d)$ with respect to the supremum norm. Lemma 2.14 is applicable in this case and gives us an upper bound for the centered small ball probability, see Lemma 4.3.

Lemma 4.8. *If $d/2 < \gamma < \infty$,*

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \le K_0 \Big(\frac{1}{\varepsilon\sigma^{d+\gamma}}\Big)^{\frac{2d}{2\gamma-d}}$$

for all $\varepsilon, \sigma > 0$, with K_0 a constant independent of ε and σ.

4.3.1.2 The case $\gamma = \infty$

For $\gamma = \infty$ and h as before, it follows from the assumptions on p that the function h is in fact well defined on $S_\sigma = \{z \in \mathbb{C}^d : \forall j \mid \operatorname{Im} z_j| \le \sigma\}$, is analytic on this set, and takes real values on $\mathbb{R}^d$. By the Cauchy-Schwarz inequality, it follows that

$$|h(z)|^2 \le \frac{1}{\sigma^{2d}} \Big(\sum_{k\in\{1,\dots,m\}^d} w_k^2 \Big) \Big(\sum_{k\in\{1,\dots,m\}^d} \big|p\big(\frac{z-k/m}{\sigma}\big)\big|^2 \Big).$$

The last factor in the right-hand side is bounded from above by a multiple of m^d on the set S_σ. Hence, we obtain

$$|h(z)| \le K\sigma^{-d}\|h\|_{\mathbb{H}^{m,\sigma}} \tag{4.10}$$

for every $z \in S_\sigma$, where the constant K only depends on the density p. Let $\mathcal{G}_\sigma$ be the set of all analytic functions on S_σ, uniformly bounded by $K\sigma^{-d}$ on that set, with K the same constant as in (4.10). The preceding shows that for the RKHS unit ball we have $\mathbb{H}_1^{m,\sigma} \subset \mathcal{G}_\sigma$ if $\gamma = \infty$.

In Lemma 4.2 we have seen an upper bound on the metric entropy of $\mathcal{G}_\sigma$ with respect to the supremum norm. Lemma 4.4 or Lemma 4.5 is applicable in this case and gives us an upper bound for the centered small ball probability.

Lemma 4.9. *If $\gamma = \infty$, there exist $\varepsilon_0, \sigma_0, K_1 > 0$, not depending on ε and σ, such that*

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \le K_1 \frac{1}{\sigma^d} \left(\log \frac{1}{\varepsilon\sigma^{1+d}} \right)^{1+d}$$

for all $\varepsilon \in (0, \varepsilon_0)$ and $\sigma \in (0, \sigma_0)$. For $\sigma \ge \sigma_0$,

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \le K_2 \Big(\log \frac{1}{\varepsilon} \Big)^{1+d}$$

for all $\varepsilon \in (0, \varepsilon_0)$, where $K_2 > 0$ is independent of ε and σ.

4.3.2 Non-centered small ball probabilities

We fix two numbers $0 < a < b < 1$ from this point on and define $\mathcal{X} = [a, b]^d$. In this subsection we view $W^{m,\sigma}$ as a random element in $C(\mathcal{X})$ and investigate its non-centered small ball probabilities around a given function $w_0 \in C^\alpha(\mathcal{X})$. The reason for considering a smaller set $\mathcal{X} \subset [0, 1]^d$ is that in order to obtain good enough approximations of the given function w_0 defined on $\mathcal{X}$ by location-scale mixtures of the kernel p, we also need kernels centered at points just outside the set $\mathcal{X}$. Results like Lemma 4.11 below with the supremum over the entire unit cube would only be possible under additional assumptions on the boundary behaviour of the function w_0.

According to Lemma 2.13, we have

$$-\log \mathbb{P}(\|W^{m,\sigma} - w_0\|_\infty < 2\varepsilon) \leq \varphi^{m,\sigma}_{w_0}(\varepsilon), \tag{4.11}$$

with the concentration function $\varphi^{m,\sigma}_{w_0}$ in (2.33) given by

$$\varphi^{m,\sigma}_{w_0}(\varepsilon) = \inf_{h \in \mathbb{H}^{m,\sigma} : \|h - w_0\|_\infty \leq \varepsilon} \|h\|^2_{\mathbb{H}^{m,\sigma}} - \log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon). \tag{4.12}$$

The centered small ball probability in the last term of (4.12) has been dealt with in the previous section. We now consider the question of estimating $w_0 \in C^\alpha(\mathcal{X})$ by elements of the reproducing kernel Hilbert space of the Gaussian process. To obtain a suitable approximation we first need an auxiliary result concerning the approximation of a smooth function f by convolutions.

4.3.2.1 General approximation result using convolutions

For a multi-index $k \in \mathbb{N}_0^d$ we write, as in the preceding chapter, $|k| = \sum_{i \leq d} k_i$ and D^k is the $|k|$-th order differential operator

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}.$$

For a function $f \in C^\alpha(\mathbb{R}^d)$ and $\sigma > 0$ we define the transform $T_{\alpha,\sigma} f$ of f by

$$T_{\alpha,\sigma} f = f - \sum_{j=1}^{\beta} \sum_{|k|=j} d_k \sigma^j (D^k f) \tag{4.13}$$

for d_k the sequence defined in (4.14) below, and β the largest integer strictly smaller than α.

Let $p_\sigma(x) = \sigma^{-d} p(x/\sigma)$. The next lemma gives a uniform bound on the difference of f and the convolution of p_σ with the transform of f.

Lemma 4.10. *For $\alpha, \sigma > 0$ and $f \in C^\alpha(\mathbb{R}^d)$ we have $\|p_\sigma * (T_{\alpha,\sigma} f) - f\|_\infty \leq K_6 \sigma^\alpha$, where $K_6 > 0$ is a constant independent of σ.*

The lemma is an extension of an idea of Rousseau [41], where a similar method is used to approximate arbitrary smooth densities by beta mixtures. The proof follows the same lines but is more involved in the present higher-dimensional case.

The right choice for the sequence d_n is what makes the proof work, as we will see below. To define this sequence, first let $m_k = \int y^k p(y)\,dy$ for $k \in \mathbb{N}_0^d$. Next, for $n \in \mathbb{N}_0^d$ we recursively define two collections of numbers c_n and d_n as follows. If $|n| = 1$ we put $c_n = 0$ and $d_n = -m_n/n!$. For $|n| \geq 2$, we define

$$c_n = -\sum_{\substack{n=l+k \\ |l|\geq 1, |k|\geq 1}} \frac{(-1)^{|k|}}{k!} m_k d_l \quad \text{and} \quad d_n = \frac{(-1)^{|n|} m_n}{n!} + c_n. \tag{4.14}$$

Note that the numbers c_n and d_n are well defined and that they only depend on the moments of p. We can now give the proof.

Proof of Lemma 4.10. The proof is by induction on β, which is the largest integer strictly smaller than α. If $\beta = 0$ then $\alpha \in (0, 1]$ and $T_{\alpha,\sigma} f = f$ and the statement of the claim is standard. To prove the induction step, suppose now that $\beta \geq 1$. By definition of $T_{\alpha,\sigma} f$ we have

$$\begin{aligned}
&(p_\sigma * T_{\alpha,\sigma} f - f)(x) \\
&\quad = \int p_\sigma(y)\Big(f(x-y) - f(x) - \sum_{j=1}^{\beta} \sum_{|k|=j} d_k \sigma^j (D^k f)(x-y)\Big)\,dy.
\end{aligned}$$

By Taylor's formula and the fact that $f \in C^\alpha$,

$$f(x-y) - f(x) = \sum_{j=1}^{\beta} \sum_{|k|=j} \frac{(-y)^k}{k!} (D^k f)(x) + R(x,y),$$

where $|R(x,y)| \leq C\|y\|^\alpha$. It follows that

$$\begin{aligned}
&(p_\sigma * T_{\alpha,\sigma} f - f)(x) \\
&\qquad = \int p_\sigma(y) R(x,y)\,dy \\
&\qquad\quad + \sum_{j=1}^{\beta} \sum_{|k|=j} \Big(\frac{1}{k!}(-1)^j (D^k f)(x) \sigma^j m_k - d_k \sigma^j \Big(p_\sigma * (D^k f)\Big)(x)\Big).
\end{aligned}$$

The first term on the right is easily seen to be bounded by a constant times σ^α. To see that this holds for the second term as well we use the induction hypothesis.

By definition of the constants c_k and d_k (see (4.14)), the second term can be written as

$$\sum_{j=1}^{\beta} \sum_{|k|=j} \Big(\frac{(-1)^j}{k!} \sigma^j m_k \Big(D_k^j f - p_\sigma * (D^k f)\Big)(x) - c_k \sigma^j \Big(p_\sigma * (D^k f)\Big)(x)\Big).$$

Now for $j \leq \beta$ and $|k| = j$, consider the decomposition

$$\begin{aligned} &D^k f - p_\sigma * (D^k f) \\ &\quad = \Big(D^k f - p_\sigma * (T_{\alpha-j,\sigma} D^k f)\Big) + \Big(p_\sigma * (T_{\alpha-j,\sigma} D^k f) - p_\sigma * (D^k f)\Big). \end{aligned}$$

Since $D^k f \in C^{\alpha-j}$, the induction hypothesis implies that the first term on the right is uniformly bounded by a constant times $\sigma^{\alpha-j}$. Combined with the first display of the paragraph, this shows that it suffices to show that

$$\sum_{j=1}^{\beta} \sum_{|k|=j} \Big(\frac{(-1)^j}{k!} \sigma^j m_k \Big(T_{\alpha-j,\sigma} D^k f - D^k f \Big) - c_k \sigma^j \Big(D^k f \Big) \Big) = 0$$

identically. Straightforward algebra shows that

$$T_{\alpha-j,\sigma} D^k f - D^k f = - \sum_{i=1}^{\beta-j} \sum_{|l|=i} d_l \sigma^i D^{k+l} f.$$

Hence,

$$\begin{aligned} &\sum_{j=1}^{\beta} \sum_{|k|=j} \frac{(-1)^j}{k!} \sigma^j m_k \Big(T_{\alpha-j,\sigma} D^k f - D^k f \Big) \\ &\qquad = - \sum_{j=1}^{\beta} \sum_{|k|=j} \sum_{i=1}^{\beta-j} \sum_{|l|=i} \frac{(-1)^j}{k!} m_k d_l \sigma^{i+j} D^{l+k} f \\ &\qquad = - \sum_{s=2}^{\beta} \sum_{|n|=s} \Big(\sum_{\substack{n=l+k \\ |l|\geq 1, |k|\geq 1}} \frac{(-1)^{|k|}}{k!} m_k d_l \Big) \sigma^s D^n f. \end{aligned}$$

By definition of the numbers c_n and d_n this equals

$$\sum_{s=1}^{\beta} \sum_{|n|=s} c_n \sigma^s D^n f,$$

and the proof is complete. □

4.3.2.2 Approximation in the RKHS

We now return to the question of approximating smooth functions w_0 by elements of the reproducing kernel Hilbert space. In the following lemma we give an upper bound on the uniform distance between w_0 and a certain element in the reproducing kernel Hilbert space, constructed in the proof using the transform $T_{\alpha,\sigma}$ introduced in (4.13).

Lemma 4.11. *For all $\sigma > 0$, $m \geq 1$ and $w_0 \in C^\alpha(\mathcal{X})$ there exists an $h \in \mathbb{H}^{m,\sigma}$ such that $\|h\|_{\mathbb{H}^{m,\sigma}} \leq K_7(1 \vee \sigma)$ and*

$$\sup_{x \in \mathcal{X}} |h(x) - w_0(x)| \leq \frac{K_8(1 \vee \sigma^{\beta+1})}{\sigma^{1+d} m^{\alpha-\beta}} + K_9 \sigma^\alpha, \tag{4.15}$$

for $K_7, K_8, K_9 > 0$ constants independent of σ and m and β the largest integer strictly smaller than α.

Proof. Since $\mathcal{X} = [a, b]^d \subset (0, 1)^d$ we can extend w_0 to all of $\mathbb{R}^d$ in such a way that the resulting function belongs to $C^\alpha(\mathbb{R}^d)$ and has support strictly inside $(0, 1)^d$. Using the operator $T_{\alpha,\sigma}$ introduced in (4.13), we define

$$h(x) = \sum_{k \in \{1,\dots,m\}^d} (T_{\alpha,\sigma} w_0)(k/m) \frac{1}{m^d} \frac{1}{\sigma^d} p\Big(\frac{x - k/m}{\sigma}\Big)$$

for $x \in [0, 1]^d$. By Lemma 4.6 it holds that $h \in \mathbb{H}^{m,\sigma}$ and

$$\|h\|^2_{\mathbb{H}^{m,\sigma}} \leq \frac{1}{m^d} \sum_{k \in \{1,\dots,m\}^d} \Big((T_{\alpha,\sigma} w_0)(k/m)\Big)^2 \leq \|T_{\alpha,\sigma} w_0\|^2_\infty.$$

It follows from the definition of $T_{\alpha,\sigma}$ that this is bounded by a constant times $(1 \vee \sigma^\beta)^2$.

We are left to show the bound for the approximation error in (4.15). By the triangle inequality,

$$\|h - w_0\|_\infty \leq \|h - p_\sigma * (T_{\alpha,\sigma} w_0)\|_\infty + \|p_\sigma * (T_{\alpha,\sigma} w_0) - w_0\|_\infty. \tag{4.16}$$

The first term on the right is the difference between the convolution $p_\sigma * T_{\alpha,\sigma} w_0$ and the corresponding Riemann sum. Using again the triangle inequality we get

$$\begin{aligned}
&|h(x) - (p_\sigma * T_{\alpha,\sigma} w_0)(x)| \\
&\quad \leq \sup_{\|y-z\|_\infty \leq 1/m} |T_{\alpha,\sigma} w_0(y) p_\sigma(x - y) - T_{\alpha,\sigma} w_0(z) p_\sigma(x - z)| \\
&\quad \leq \|T_{\alpha,\sigma} w_0\|_\infty \sup_{\|y-z\|_\infty \leq 1/m} |p_\sigma(x - y) - p_\sigma(x - z)| \\
&\qquad + \|p_\sigma\|_\infty \sup_{\|y-z\|_\infty \leq 1/m} |T_{\alpha,\sigma} w_0(y) - T_{\alpha,\sigma} w_0(z)|.
\end{aligned}$$

Now use the facts that $T_{\alpha,\sigma} w_0$ is bounded by a constant times $1 \vee \sigma^\beta$, p_σ is bounded by σ^{-d} times a constant, p is Lipschitz and the definition of $T_{\alpha,\sigma} w_0$ to see that

$$\|h - p_\sigma * T_{\alpha,\sigma} w_0\|_\infty \leq \frac{C_1(1 \vee \sigma^\beta)}{\sigma^{1+d} m} + \frac{C_2(1 \vee \sigma^\beta)}{\sigma^d m^{\alpha-\beta}} \leq \frac{C_3(1 \vee \sigma^{\beta+1})}{\sigma^{1+d} m^{\alpha-\beta}},$$

which covers the first term on the right of (4.16). Lemma 4.10 implies that the second term is bounded by a constant times σ^α. □

4.3.2.3 Non-centered small ball probabilities

In the previous subsections we have considered the two ingredients of the concentration function of the Gaussian process. By combining the approximation result in Lemma 4.11 with the appropriate bounds in Lemma 4.8 and Lemma 4.9 on the centered small ball probabilities, we obtain upper bounds of the concentration function (4.12) for both $\gamma < \infty$ and $\gamma = \infty$. This will allow us to determine posterior contraction rates for the corresponding Gaussian priors. As we have seen in Lemma 2.13, the concentration function around w_0 is equivalent to the non-centered small ball probability of the process around w_0. Let us therefore only give an upper bound for the latter.

In view of (4.12) and Lemma 4.11, let us consider $\varepsilon > 0$ that satisfies the condition

$$\frac{K_3}{\sigma^{1+d} m^{\alpha-\beta}} + K_4 \sigma^{\alpha} \leq \varepsilon < \varepsilon_0 \tag{4.17}$$

for certain $\varepsilon_0, K_3, K_4 > 0$ that do not depend on σ and m. The small ball probabilities around w_0 are bounded as follows for $\gamma < \infty$ and $\gamma = \infty$.

Lemma 4.12. *If $d/2 < \gamma < \infty$, then there exist ε_0, σ_0, K_1, K_2, K_3, $K_4 > 0$ such that for any m, any $\sigma \in (0, \sigma_0)$ and any ε that satisfies condition* (4.17)*,*

$$-\log \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < 2\varepsilon \Big) \leq K_1 + K_2 \Big(\frac{1}{\varepsilon \sigma^{d+\gamma}} \Big)^{\frac{2d}{2\gamma - d}}.$$

Lemma 4.13. *If $\gamma = \infty$, then there exist ε_0, σ_0, K_1, K_2, K_3, $K_4 > 0$ such that for any m, any $\sigma \in (0, \sigma_0)$ and any ε that satisfies condition* (4.17)*,*

$$-\log \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < 2\varepsilon \Big) \leq K_1 + K_2 \frac{1}{\sigma^d} \Big(\log \frac{1}{\varepsilon \sigma^{1+d}} \Big)^{1+d}.$$

4.4 Posterior contraction for Gaussian kernel mixture priors

4.4.1 General result

The Gaussian kernel mixture process $W^{m,\sigma}$ can be used to construct priors in various nonparametric statistical settings. In order to ensure consistency one has to let the scale parameter σ tend to 0 and let the partition size parameter m tend to infinity with the sample size, or estimate these hyper-parameters from the data. In this section we consider sequences of priors constructed by letting σ and m depend on the sample size in a deterministic manner. We give bounds on the contraction rates of the corresponding posteriors. In the next section we investigate the possibility of endowing m and σ with prior distributions.

Consider a sequence of positive numbers $\sigma_n \to 0$ and a sequence of natural numbers $m_n \to \infty$ and let W^n be the process that is obtained by substituting m and σ by m_n and σ_n in (4.6), hence

$$W^n(x) = \sum_{k \in \{1,\dots,m\}^d} Z_k \frac{1}{m_n^{d/2}} \frac{1}{\sigma_n^d} p\big(\frac{x - k/m_n}{\sigma_n}\big), \tag{4.18}$$

with the Z_k independent standard Gaussian random variables and $p \in \mathcal{P}_\gamma$. For $w_0 \in C^\alpha(\mathcal{X})$, $\alpha > 0$, let $\varphi_{w_0}^n$ be the corresponding concentration function (4.12). Here we view W^n as a Gaussian random element in $C(\mathcal{X})$ again, i.e.

$$\varphi_{w_0}^n(\varepsilon) = \inf_{h \in \mathbb{H}^{m_n,\sigma_n} : \sup_{x\in\mathcal{X}} |h(x) - w_0(x)| \le \varepsilon} \|h\|^2_{\mathbb{H}^{m_n,\sigma_n}} - \log \mathbb{P}\Big(\sup_{x \in \mathcal{X}} |W^{m_n,\sigma_n}| < \varepsilon \Big).$$

According to the general theory of Gaussian process priors, the posterior contraction rate for priors based on the law of the process W^n is obtained by solving the inequality

$$\varphi_{w_0}^n(\varepsilon_n) \le n\varepsilon_n^2, \tag{4.19}$$

see Section 2.6. By Lemma 4.12, this inequality holds in the case $d/2 < \gamma < \infty$ if

$$K_1 + K_2 \Big(\frac{1}{\varepsilon_n \sigma_n^{d+\gamma}} \Big)^{\frac{2d}{2\gamma - d}} \le n\varepsilon_n^2,$$
$$\frac{K_3}{\sigma_n^{1+d} m_n^{\alpha-\beta}} + K_4 \sigma_n^\alpha \le \varepsilon_n,$$

with $K_1, \dots, K_4$ the constants appearing in the lemma and β the largest integer strictly smaller than α. The optimal solution of these inequalities is easily found, i.e. the solution yielding the smallest possible sequence ε_n (in order) and we obtain the following theorem. We use the following notation:

$$d_\gamma = \frac{2d(d+\gamma)}{2\gamma - d}, \qquad \delta_\gamma = \frac{d}{2\gamma - d}. \tag{4.20}$$

Theorem 4.14. *In the setting described above, suppose that $d/2 < \gamma < \infty$. Let $m_n^{\alpha-\beta} \gtrsim n^{(1+\alpha+d)/(d_\gamma + 2\alpha(1+\delta_\gamma))}$ and $\sigma_n \sim n^{-1/(d_\gamma + 2\alpha(1+\delta_\gamma))}$. Then (4.19) holds with*

$$\varepsilon_n \sim n^{-\alpha/(d_\gamma + 2\alpha(1+\delta_\gamma))}.$$

Note that for the numbers defined in (4.20) we have $d_\gamma \to d$ and $\delta_\gamma \to 0$ as $\gamma \to \infty$ and consequently the exponent of $1/n$ in the expression for ε_n given in the theorem tends to $\alpha/(d + 2\alpha)$, which corresponds to the optimal minimax rate of convergence for estimating an α-smooth function of d variables.

It turns out that if we use an infinitely smooth kernel, i.e. $\gamma = \infty$, we indeed achieve the optimal rate, up to a logarithmic factor. In this case inequality (4.19)

holds, by Lemma 4.13, if the sequences ε_n, m_n and σ_n satisfy

$$K_1 + K_2 \frac{1}{\sigma_n^d} \left(\log \frac{1}{\varepsilon_n \sigma_n^{1+d}} \right)^{1+d} \leq n\varepsilon_n^2$$
$$\frac{K_3}{\sigma_n^{1+d} m_n^{\alpha-\beta}} + K_4 \sigma_n^\alpha \leq \varepsilon_n,$$

with $K_1, \ldots, K_4$ the constants appearing in the lemma and β the largest integer strictly smaller than α. Again it is straightforward to find the optimal solutions of these inequalities and we arrive at the following result.

Theorem 4.15. *In the setting described above, suppose that $\gamma = \infty$. Let $m_n^{\alpha-\beta} \gtrsim (n/\log^{1+d} n)^{(1+\alpha+d)/(d+2\alpha)}$ and $\sigma_n \sim (n/\log^{1+d} n)^{-1/(d+2\alpha)}$. Then (4.19) holds with*

$$\varepsilon_n \sim (n/\log^{1+d} n)^{-\alpha/(d+2\alpha)}.$$

In combination with the results given in Section 2.6 these theorems immediately yield rate of contraction results for various statistical settings. We give details in the next section. In particular, Theorem 4.15 will imply that if the law of the Gaussian kernel mixture process W^{m_n,σ_n} is used as a prior on an α-regular function of d variables, then if the kernel that is used is infinitely smooth in the sense that it belongs to $\mathcal{P}_\infty$, and we set $m_n^{\alpha-\beta} \sim (n/\log^{1+d} n)^{(1+\alpha+d)/(d+2\alpha)}$ and $\sigma_n \sim (n/\log^{1+d} n)^{-1/(d+2\alpha)}$, this leads to posterior contraction rates of the optimal order $n^{-\alpha/(d+2\alpha)}$, up to a logarithmic factor.

Note however that these Gaussian process priors depend on the unknown regularity α of the function that is being estimated through the choice of m_n and σ_n. In Section 4.5 we will show that if instead of choosing the hyper-parameters m and σ deterministically, we endow them with appropriate prior distributions, it is possible to obtain a rate-adaptive procedure.

4.4.2 Results for specific statistical settings

4.4.2.1 Gaussian regression

Consider the nonparametric regression setting described in Section 3.4.2, but now with the design points x_i all belonging to the space $\mathcal{X} = [a, b]^d \subset [0, 1]^d$ and $w_0 : \mathcal{X} \to \mathbb{R}$. The Gaussian process W^n defined by (4.18) can be used directly as a prior for w_0 in this case. We again also endow the error standard deviation, which we now denote by τ_0 to avoid confusion, with a prior distribution which we assume to be supported on a given compact interval in $(0, \infty)$ that contains τ_0, with a Lebesgue density that is bounded away from zero. We denote the total prior on (w_0, τ_0) by Π_n and write $\Pi_n(\cdot \mid Y_1, \ldots, Y_n)$ for the corresponding posterior distribution. We say that the posterior contracts at rate ε_n if the convergence in (3.10) holds as $n \to \infty$ for every sufficiently large constant L (with σ replaced

by τ). Combining the previous results of Theorem 4.14 and Theorem 4.15 with the general result in Theorem 2.21 yields the following result in the fixed design regression case.

Theorem 4.16. *Suppose that $w_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$.*

- *If $\gamma < \infty$ and*

$$m_n \sim n^{\frac{1+d+\alpha}{(d_\gamma+2\alpha(1+\delta_\gamma))(\alpha-\beta)}} \quad \text{and} \quad \sigma_n \sim n^{-\frac{1}{d_\gamma+2\alpha(1+\delta_\gamma)}}, \tag{4.21}$$

  then the posterior contracts at rate

$$\varepsilon_n \sim n^{-\frac{\alpha}{d_\gamma+2\alpha(1+\delta_\gamma)}}.$$

- *If $\gamma = \infty$ and*

$$m_n \sim (n/\log^{1+d} n)^{\frac{1+d+\alpha}{(d+2\alpha)(\alpha-\beta)}} \quad \text{and} \quad \sigma_n \sim (n/\log^{1+d} n)^{-\frac{1}{d+2\alpha}}, \tag{4.22}$$

  then the posterior contracts at rate

$$\varepsilon_n \sim (n/\log^{1+d} n)^{-\frac{\alpha}{d+2\alpha}}.$$

4.4.2.2 Density estimation

Next consider the nonparametric density estimation problem described in Section 3.4.3, but with f_0 a density on the space $\mathcal{X} \subset [0,1]^d$. As prior on the density function f_0 we use the law Π_n of the random density $x \mapsto ce^{W^n(x)}$, with c the renormalization constant $c = (\int_{\mathcal{X}} e^{W^n(x)}\,dx)^{-1}$. Let $\Pi_n(\cdot \,|\, X_1 \ldots, X_n)$ denote the posterior distribution. We say that the posterior contracts at rate ε_n if the convergence in (3.12) holds as $n \to \infty$ for every sufficiently large constant L. Let $w_0 = \log f_0$. Then by combining the previous Theorems 4.14 and 4.15 with Theorem 2.19, it follows that the statement of Theorem 4.16 is also true in this case.

4.4.2.3 Classification

Finallly, consider the classification setting in Section 3.4.4, where we now assume that the X_i take values in $\mathcal{X}$ and hence the binary regression function r_0 is a function on $\mathcal{X}$. As a prior on r_0 we use the law Π_n of $\Psi(W^n)$, with Ψ the logistic or normal distribution function. Let $\Pi_n(\cdot \,|\, X_1, Y_1, \ldots, X_n, Y_n)$ be the corresponding posterior. We say that the posterior contracts at rate ε_n if the convergence in (3.13) holds as $n \to \infty$ for every sufficiently large L. Let $w_0 = \Psi^{-1}(r_0)$. Then by combining the previous Theorems 4.14 and 4.15 with Theorem 2.20, it follows that the statement of Theorem 4.16 holds true in this case as well.

4.5 Adaptation using conditionally Gaussian priors

4.5.1 General result

In the previous section we saw that in several statistical settings involving an unknown smooth function w_0, posterior contraction can be achieved at an optimal rate for an appropriate sequence of Gaussian kernel mixture priors. However, we used the knowledge of the regularity of the function w_0 to construct the specific priors.

As in the preceding chapter, our next goal is to obtain a fully rate-adaptive procedure by viewing the tuning parameters σ and m in (4.6) as hyperparameters and endowing them with appropriate prior distributions. Again it turns out that using such hierarchical priors, it is indeed possible to obtain adaptive, rate-optimal procedures in this manner.

Concretely, the hierarchical priors will be based on the conditionally Gaussian kernel mixture process W that is defined as

$$W(x) = W^{M,\Sigma}(x) = \sum_{k\in\{1,\ldots,M\}^d} Z_k \frac{1}{M^{d/2}} \frac{1}{\Sigma^d} p\big(\frac{x - k/M}{\Sigma}\big), \qquad x \in [0,1]^d, \quad (4.23)$$

with independent, standard Gaussian random variables Z_k, independent also of the independent random variables M and Σ on, respectively, $\mathbb{N}$ and $(0,\infty)$. We assume that Σ has a Lebesgue density g. As before $p : \mathbb{R}^d \to \mathbb{R}$ is a function in the class of γ-regular kernels $\mathcal{P}_\gamma$, for some $\gamma \in (d/2, \infty]$. Note that by construction, we have that conditional on $M = m$ and $\Sigma = \sigma$, this is the Gaussian process $W^{m,\sigma}$ in (4.6).

The hierarchical kernel mixture process can be used to construct priors for various statistical settings again. The following general theorem about the process W will lead to the desired adaptive rate of contraction results. Recall that $\mathcal{X} = [a,b]^d$, for $0 < a < b < 1$.

Theorem 4.17. *Suppose that for some $C > 0$ and $s > 1$,*

$$\mathbb{P}(M \geq m) \geq C m^{-s} \quad (4.24)$$

and that g satisfies, for all σ in a neighborhood of zero,

$$D_1 \sigma^{-q} e^{-D_2 \sigma^{-d\gamma} (\log \frac{1}{\sigma})^r} \leq g(\sigma) \leq D_3 \sigma^{-q} e^{-D_4 \sigma^{-d\gamma} (\log \frac{1}{\sigma})^r} \quad (4.25)$$

for some $D_1, D_2, D_3, D_4 > 0$ and $q, r \geq 0$. Then there exists, for every constant $K > 1$, measurable subsets B_n of $C([0,1]^d)$ and a constant $L > 0$ such that, for

sufficiently large n,

$$\mathbb{P}(W \notin B_n) \leq e^{-Kn\varepsilon_n^2} \tag{4.26}$$

$$\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq Ln\bar{\varepsilon}_n^2 \tag{4.27}$$

$$\mathbb{P}(\sup_{x\in\mathcal{X}} |W(x) - w_0(x)| \leq \varepsilon_n) \geq e^{-n\varepsilon_n^2} \tag{4.28}$$

for $\bar{\varepsilon}_n$ *the rate*

$$c' n^{-\frac{\alpha(1-(d\delta_\gamma)/(2\gamma))}{(d_\gamma+2\alpha(1+\delta_\gamma))(1+d/(2\gamma))}} \qquad \text{if} \quad \gamma < \infty, \tag{4.29}$$

$$c' n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{r\vee(1+d)}{2+d/\alpha}+(\frac{1+d-r}{2})_+} \qquad \text{if} \quad \gamma = \infty \tag{4.30}$$

and ε_n *the rate*

$$c n^{-\frac{\alpha}{d_\gamma+2\alpha(1+\delta_\gamma)}} \qquad \text{if} \quad \gamma < \infty, \tag{4.31}$$

$$c n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{r\vee(1+d)}{2+d/\alpha}} \qquad \text{if} \quad \gamma = \infty, \tag{4.32}$$

for $c, c' > 0$ *large enough constants.*

As explained in more detail in the next section, this general theorem connects to the results giving sufficient conditions for having a certain posterior contraction rate in several statistical models, see Chapter 2. For priors based on the conditionaly Gaussian process W we will obtain posterior rates of the order $\varepsilon \vee \bar{\varepsilon}_n$.

Note that we only get an actual rate in the case $\gamma < \infty$ if $d\delta_\gamma < 2\gamma$. It is easy to verify that this is true if and only if $\gamma > (1+\sqrt{5})d/4 \approx (0.81)d$. In this case the exponent of $1/n$ in the rate tends to the optimal $\alpha/(d+2\alpha)$ as $\gamma \to \infty$.

For $\gamma = \infty$ we obtain the optimal rate $n^{-\alpha/(d+2\alpha)}$, up to a logarithmic factor. Moreover, we have a rate-adaptive procedure in this case, since the process W does not depend on the regularity of the function that is being estimated. The exponent of the log factor in the rate for $\gamma = \infty$ is minimal if $r = 1 + d$. The posterior rate is then $(n/\log n)^{-\alpha/(d+2\alpha)}$. For larger or smaller values of r, the rate is slightly worse. If Σ^d has an inverse gamma distribution, then condition (4.25) holds for $r = 0$, leading to the rate

$$n^{-\frac{\alpha}{d+2\alpha}} \log^{\frac{4\alpha+4\alpha d+d+d^2}{4\alpha+2d}} n.$$

4.5.2 Results for specific statistical settings

Combined with the general results presented in Chapter 2 , Theorem 4.17 yields rate of contraction results for priors based on the hierarchical kernel mixture process (4.23) in various statistical settings. In this section we briefly state the results for density estimation, fixed design regression and classification.

Consider first the regression setting described in Section 4.4.2.1. As prior on the regression function w_0 we now employ the law of the conditionally Gaussian process W, where the hyper priors on M and Σ are assumed to satisfy conditions (4.24) and (4.25), respectively. The total prior on the pair (w_0, τ_0), with τ_0 the error standard deviation, is denoted by Π. As before we say that the corresponding posterior contracts at the rate ε_n if, for all L large enough, (3.10) holds as $n \to \infty$, with Π in the place of Π_n (and σ replaced by τ).

Combining Theorem 4.17 and Theorem 2.11 yields the following result for fixed design regression. Recall the definitions (4.20) of d_γ and δ_γ.

Theorem 4.18. *Suppose that* $w_0 \in C^\alpha(\mathcal{X})$.

- *If* $\gamma \in (d/2, \infty)$, *then the posterior contracts at the rate*

$$n^{-\frac{\alpha(1-(d\delta_\gamma)/(2\gamma))}{(d_\gamma+2\alpha(1+\delta_\gamma))(1+d/(2\gamma))}}.$$

- *If* $\gamma = \infty$, *then the posterior contracts at the rate*

$$n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{r\vee(1+d)}{2+d/\alpha}+(\frac{1+d-r}{2})_+}.$$

For density estimation we consider the setting of Section 4.4.2.2 again. As prior on the density function f_0 we use the law Π of the random density

$$x \mapsto \frac{e^{W(x)}}{\int_{\mathcal{X}} e^{W(x)}\, dx},$$

where M and σ are assumed to satisfy conditions (4.24) and (4.25), respectively. We say that the posterior contracts at rate ε_n if the convergence in (3.12) holds as $n \to \infty$ for every sufficiently large constant L, with Π_n replaced by Π.

Combining Theorem 4.17 and Theorem 2.5 yields the following result for density estimation.

Theorem 4.19. *Suppose that* $\log f_0 \in C^\alpha(\mathcal{X})$.

- *If* $\gamma \in (d/2, \infty)$, *then the posterior contracts at the rate*

$$n^{-\frac{\alpha(1-(d\delta_\gamma)/(2\gamma))}{(d_\gamma+2\alpha(1+\delta_\gamma))(1+d/(2\gamma))}}.$$

- *If* $\gamma = \infty$, *then the posterior contracts at the rate*

$$n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{r\vee(1+d)}{2+d/\alpha}+(\frac{1+d-r}{2})_+}.$$

Finally, we again consider the non-parametric classification problem described in Section 4.4.2.3. As prior on the binary regression function r_0 we employ the

law Π of $\Psi(W)$, with Ψ the logistic or normal distribution function and W the conditionally Gaussian kernel mixture process, where M and σ are assumed to satisfy conditions (4.24) and (4.25), respectively. We say that the corresponding posterior contracts at the rate ε_n if for all L large enough (3.13) holds as $n \to \infty$, with Π in place of Π_n.

Combining Theorem 4.17 and Theorem 2.7 yields the following result in the classification setting.

Theorem 4.20. *Suppose that* $\Psi^{-1}(r_0) \in C^\alpha(\mathcal{X})$.

- *If* $\gamma \in (d/2, \infty)$, *then the posterior contracts at the rate*

$$n^{-\frac{\alpha(1-(d\delta_\gamma)/(2\gamma))}{(d_\gamma+2\alpha(1+\delta_\gamma))(1+d/(2\gamma))}}.$$

- *If* $\gamma = \infty$, *then the posterior contracts at the rate*

$$n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{r\vee(1+d)}{2+d/\alpha}+(\frac{1+d-r}{2})_+}.$$

4.5.3 Proof of Theorem 4.17

4.5.3.1 Prior mass condition (4.28)

Let $\lambda_m = \mathbb{P}(M = m)$. The probability $\mathbb{P}(\sup_{x\in\mathcal{X}} |W(x) - w_0(x)| \le \varepsilon)$ can be written as

$$\sum_{m=1}^{\infty} \lambda_m \int_0^\infty g(\sigma) \mathbb{P}\Big(\sup_{x\in\mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < \varepsilon \Big)\, d\sigma, \tag{4.33}$$

by conditioning on M and Σ.

First consider the case $\gamma < \infty$. According to Lemma 4.12, there exist constants $\varepsilon_0, C_1, C_2, C_3, C_4 > 0$ independent of σ and m, such that

$$-\log \mathbb{P}\Big(\sup_{x\in\mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < \varepsilon \Big) \le C_3 + C_4 \Big(\frac{1}{\varepsilon \sigma^{d+\gamma}} \Big)^{\frac{2d}{2\gamma-d}}$$

for any $\varepsilon < \varepsilon_0$ and any σ and m that satisfy

$$\frac{1}{2} C_1 \varepsilon^{1/\alpha} < \sigma < C_1 \varepsilon^{1/\alpha} \le 1 \quad \text{and} \quad m \ge C_2 \varepsilon^{-\frac{1+d+\alpha}{\alpha(\alpha-\beta)}}. \tag{4.34}$$

For $\varepsilon < \varepsilon_0$, the probability of interest can be bounded from below by restricting the sum and integral in (4.33) to the m and σ that satisfy (4.34). The lower bound on the integral becomes

$$\int_{\frac{1}{2}C_1\varepsilon^{1/\alpha}}^{C_1\varepsilon^{1/\alpha}} g(\sigma) \exp\Big(- C_4 \Big(\frac{1}{\varepsilon\sigma^{d+\gamma}} \Big)^{\frac{2d}{2\gamma-d}} \Big)\, d\sigma \ge C_5 \exp(-C_6 \varepsilon^{-\frac{\alpha+d+\gamma}{\alpha}\frac{2d}{2\gamma-d}})$$

by first substituting the lower bound for $g(\sigma)$ in (4.25) and then bounding the integrand using the bounds of the integration interval. We used that $\frac{\alpha+d+\gamma}{\alpha}\frac{2d}{2\gamma-d} > \frac{d_\gamma}{\alpha}$, which is clear from the definition of d_γ in (4.20). The lower bound on the integral does not depend on m and thus comes out of the sum. For $\varepsilon < \varepsilon_0$ we thus have

$$\mathbb{P}(\sup_{x\in\mathcal{X}}|W(x)-w_0(x)|\leq\varepsilon) \geq \mathbb{P}(M \geq C_2\varepsilon^{-\frac{1+d+\alpha}{\alpha(\alpha-\beta)}})C_5\exp(-C_6\varepsilon^{-\frac{\alpha+d+\gamma}{\alpha}\frac{2d}{2\gamma-d}}).$$

The assumption (4.24) implies that the lower bound on the tail probability of M is some positive power of ε and therefore also bounded from below by an exponential lower bound. We thus conclude that

$$\mathbb{P}(\sup_{x\in\mathcal{X}}|W(x)-w_0(x)|\leq\varepsilon) \geq C_7\exp(-C_8\varepsilon^{-\frac{\alpha+d+\gamma}{\alpha}\frac{2d}{2\gamma-d}})$$

for some constants $C_7, C_8 > 0$. It follows that condition (4.28) is fulfilled for ε_n some large enough multiple of $n^{-\frac{\alpha}{d_\gamma+2\alpha(1+\delta_\gamma)}}$.

The proof for $\gamma = \infty$ is similar. We find that there exist constants $C_5, C_6 > 0$ such that for $\varepsilon > 0$ small enough,

$$\mathbb{P}(\sup_{x\in\mathcal{X}}|W(x)-w_0(x)|\leq\varepsilon) \geq C_5\exp(-C_6\varepsilon^{-d/\alpha}\log^{r\vee(1+d)}(1/\varepsilon)).$$

It follows that condition (4.28) is satisfied in this case for ε_n equal to some large enough multiple of $n^{-\frac{\alpha}{d+2\alpha}}\log^t n$, provided that $t \geq (r\vee(1+d))/(2+d/\alpha)$. Choose t minimal.

4.5.3.2 Construction of sieves

We have seen the existence of appropriate sieves for Gaussian priors in Theorem 2.18. The proof in [50] explicitly constructs the sieves for a Gaussian prior from both the unit ball $\mathbb{B}_1$ in the Banach space in which the process takes it values, and from the unit ball $\mathbb{H}_1$ in the reproducing kernel Hilbert space connected to the Gaussian process. The relevant sieves are given by $L_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1$ for some appropriately chosen sequence L_n and with ε_n the posterior contraction rate.

We now construct in a similar way sieves for the present conditionally Gaussian prior. As opposed to the constuction in the previous chapter, in the construction of these sieves we will not be using the sieves that belong to the Gaussian process. Instead, the sieves will be constructed using the sets in which we have before embedded the RKHS unit ball. In both cases $\gamma < \infty$ and $\gamma = \infty$, this construction involves the choice of a sequence of radii R_n and a sequence L_n. The particular choices for these sequences are given in Subsection 4.5.3.5.

First consider the case $\gamma < \infty$. Recall from (4.9) that $\mathbb{H}_1^{m,\sigma} \subset C^\gamma_{\sigma^{-(d+\gamma)\|p\|_\gamma}}([0,1]^d)$ and hence $\mathbb{H}_1^{m,\sigma} \subset C^\gamma_{R^{-(d+\gamma)\|p\|_\gamma}}([0,1]^d)$ for any $\sigma \geq R$. This

motivates the choice of sieves

$$B_n = L_n C^{\gamma}_{R_n^{-(d+\gamma)}\|p\|_\gamma}([0,1]^d) + M_1\varepsilon_n\mathbb{B}_1$$

for some sequences L_n and R_n, and a large enough constant M_1. Here $\mathbb{B}_1$ is the unit ball of the space $C([0,1]^d)$ and ε_n is given by (4.32).

The sieves for the prior with $\gamma = \infty$ are constructed using $\mathcal{G}_\sigma$. Recall that $\mathcal{G}_\sigma$ is the set of all analytic functions defined on the strip $S_\sigma = \{z \in \mathbb{C}^d : \forall j \,|\operatorname{Im} z_j| \le \sigma\}$ that are bounded by $K_p\sigma^{-d}$ on S_σ, where K_p is a constant that only depends on the kernel p. We have seen that $\mathbb{H}_1^{m,\sigma} \subset \mathcal{G}_\sigma$. Note that $\mathcal{G}_{\sigma_1} \subset \mathcal{G}_{\sigma_2}$ if $\sigma_1 \ge \sigma_2$. We now define the sieves similar as in the case with $\gamma < \infty$. Let

$$B_n = L_n\mathcal{G}_{R_n} + M_1\varepsilon_n\mathbb{B}_1$$

for some sequences L_n and R_n, and a sufficiently large constant M_1. Here $\mathbb{B}_1$ is again the unit ball of the space $C([0,1]^d)$ and ε_n is in this case given by (4.32).

We will show that these sieves satisfy the properties stated in Theorem 4.17.

4.5.3.3 Remaining mass condition

We first verify the remaining mass condition (4.26) for $\gamma < \infty$. Let $C > 1$. By conditioning and restricting the integral to $\sigma \ge R_n$, we obtain

$$\mathbb{P}(W \notin B_n) \le \sum_{m=1}^{\infty} \lambda_m \int_{R_n}^{\infty} g(\sigma)\mathbb{P}(W^{m,\sigma} \notin B_n)\, d\sigma + \mathbb{P}(\Sigma < R_n).$$

We show that the first term on the right is bounded from above by $\exp(-Dn\varepsilon_n^2)$ for D a constant that we can choose as large as we like by choosing an appropriate large multiple M_1 of (4.32) for ε_n. For this, it suffices to show that $\mathbb{P}(W^{m,\sigma} \notin B_n)$ is bounded from above by $\exp(-Dn\varepsilon_n^2)$ for $\sigma \ge R_n$. Let $B_n^{m,\sigma} = L_n\mathbb{H}_1^{m,\sigma} + \varepsilon_n\mathbb{B}_1$. By construction, $B_n^{m,\sigma} \subset B_n$ for any $\sigma \ge R_n$ and any m. Hence

$$\mathbb{P}(W^{m,\sigma} \notin B_n) \le \mathbb{P}(W^{m,\sigma} \notin B_n^{m,\sigma}) \tag{4.35}$$

for any $\sigma \ge R_n$ and any m. By Borell-Sudakov (see Theorem 2.16), with Φ the standard normal distribution function and for $\sigma \ge R_n$,

$$\mathbb{P}(W^{m,\sigma} \notin B^{m,\sigma}) \le 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W^{m,\sigma}\|_\infty \le \varepsilon_n)) + L_n).$$

By Lemma 4.8 we have, for $\sigma \ge R_n$ and $R_n \le 1$,

$$\mathbb{P}(\|W^{m,\sigma}\|_\infty \le \varepsilon_n) \ge e^{-K_6 R^{-d_\gamma}\varepsilon_n^{-2d/(2\gamma-d)}}$$

for a constant $K_6 > 0$ and $\rho_n > 0$ small enough. Since $\Phi^{-1}(y) \ge -\sqrt{(5/2)\log(1/y)}$ for $y \in (0, 1/2)$, it follows that

$$\begin{aligned}\mathbb{P}(W^{m,\sigma} \notin B_n) &\le 1 - \Phi\Big(L_n - \sqrt{(5/2)K_6 R_n^{-d_\gamma}\varepsilon_n^{-2d/(2\gamma-d)}}\Big) \\ &\le e^{-\frac{1}{2}(L_n - \sqrt{(5/2)K_6 R_n^{-d_\gamma}\varepsilon_n^{-2d/(2\gamma-d)}})^2},\end{aligned}$$

for $\sigma \geq R_n$ and $L_n \geq \sqrt{(5/2)K_6 R_n^{-d_\gamma} \varepsilon_n^{-2d/(2\gamma-d)}}$. In Subsection 4.5.3.5 we will show that we can choose L_n and R_n such that

$$(L_n - \sqrt{(5/2)K_6 R_n^{-d_\gamma} \rho_n^{-2d/(2\gamma-d)}})^2 \geq Dn\varepsilon_n^2$$

so that $\mathbb{P}(W^{m,\sigma} \notin B_n)$ is indeed bounded from above by $\exp(-Dn\varepsilon_n^2)$.

We are left to bound $\mathbb{P}(\Sigma < R_n)$. For this we use the upper bound for $g(\sigma)$ in the assumption (4.25). A substitution $x = \sigma^{-1}$ then shows that

$$\mathbb{P}(\Sigma < R_n) \leq D_3 \int_{1/R_n}^{\infty} x^{q-2} e^{-D_4 x^{d_\gamma} (\log x)^r} \, dx.$$

According to Lemma 4.9 of [51], this is further bounded by

$$\frac{2D_3}{dD_4} \frac{(1/R_n)^{q-2-d_\gamma+1}}{(\log(1/R_n))^r} e^{-D_4 (1/R_n)^{d_\gamma} (\log(1/R_n))^r} \leq e^{-\frac{1}{2} D_4 (1/R_n)^{d_\gamma} (\log(1/R_n))^r}$$

for R_n small enough. In Subsection 4.5.3.5 we will show that the chosen R_n satisfies

$$\frac{1}{R_n^{d_\gamma}} \log^r \frac{1}{R_n} \geq M_1 n\varepsilon_n^2$$

so that $\mathbb{P}(\Sigma < R_n)$ is bounded from above by $\exp(-Dn\varepsilon_n^2)$, with D as before, by choosing M_1 sufficiently large.

We thus conclude that $\mathbb{P}(W \notin B_n) \leq 2\exp(-Dn\varepsilon_n^2)$. For sufficiently large D, this is bounded from above by $\exp(-Cn\varepsilon_n^2)$ for the given C, which was to be shown.

The remaining mass condition in the case $\gamma = \infty$ follows using the same conditioning argument as in the finite regularity case above. Arguing as before, we now get

$$\mathbb{P}(W^{m,\sigma} \notin B_n) \leq e^{-\frac{1}{2}(L_n - \sqrt{(5/2)K_6 R_n^{-d} (\log(1/(\varepsilon_n R_n^{1+d})))^{1+d}})^2}$$

for $\sigma \geq R_n$ and $L_n \geq \sqrt{(5/2)K_6 R_n^{-d} (\log(1/(\varepsilon_n R_n^{1+d})))^{1+d}}$. In Subsection 4.5.3.5 we will show that we can choose sequences L_n and R_n such that

$$\left(L_n - \sqrt{(5/2)K_6 R_n^{-d} (\log(1/(\varepsilon_n R_n^{1+d})))^{1+d}}\right)^2 \geq Dn\varepsilon_n^2.$$

Showing the bound on $\mathbb{P}(\Sigma < R_n)$ is the same as in the case with $\gamma < \infty$, but now with d instead of d_γ. We show in Subsection 4.5.3.5 that the chosen R_n satisfies

$$\frac{1}{R_n^d} \log^r \frac{1}{R_n} \geq M_1 n\varepsilon_n^2.$$

This is enough to complete the proof of the remaining mass condition.

4.5.3.4 Entropy condition

Let $\bar{\varepsilon}_n$ given in either (4.29) or (4.30). We verify the entropy condition (4.27).

Suppose first that $\gamma < \infty$. For the entropy of the sieve B_n we have in this case, because $\bar{\varepsilon}_n \geq M_1 \varepsilon_n$ for c' large enough,

$$N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq N(\bar{\varepsilon}_n, L_n C^\gamma_{R_n^{-(d+\gamma)}\|p\|_\gamma}([0,1]^d), \|\cdot\|_\infty).$$

Hence, by Lemma 4.1,

$$\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq K_1 \Big(\frac{L_n}{\bar{\varepsilon}_n R_n^{d+\gamma}} \Big)^{d/\gamma}.$$

In Subsection 4.5.3.5 we will show that for the chosen L_n and R_n, this is bounded from above by a constant times $n\bar{\varepsilon}_n^2$.

Let now $\gamma = \infty$. Arguing as before we have in this case by Lemma 4.2

$$N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq N(\bar{\varepsilon}_n / L_n, \mathcal{G}_{R_n}, \|\cdot\|_\infty) \leq K_1 \frac{1}{R_n^d} \Big(\log \frac{L_n}{\bar{\varepsilon}_n R_n^d} \Big)^{1+d}.$$

In Subsection 4.5.3.5 we will show that for the choices of R_n and L_n, this is bounded from above by a constant times $n\bar{\varepsilon}_n^2$.

4.5.3.5 End of the proof

We now finish the proof of Theorem 4.17 by choosing the appropriate sequences L_n and R_n for both $\gamma < \infty$ and $\gamma = \infty$.

In the case $\gamma < \infty$, we have to show we can choose R_n and L_n such that

$$\frac{1}{R_n^{d_\gamma}} \log^r \frac{1}{R_n} \geq M_1 n \varepsilon_n^2,$$

$$(L_n - \sqrt{(5/2) K_6 R_n^{-d_\gamma} \varepsilon_n^{-2d/(2\gamma - d)}})^2 \geq D n \varepsilon_n^2$$

and also $\bar{\varepsilon}_n \geq \varepsilon_n$ and

$$\Big(\frac{L_n}{\bar{\varepsilon}_n R_n^{d+\gamma}} \Big)^{d/\gamma} \leq const\ n\bar{\varepsilon}_n^2.$$

Observe that if we take

$$\frac{1}{R_n^{d_\gamma}} = M n^{\frac{d_\gamma + 2\alpha\delta_\gamma}{d_\gamma + 2\alpha(1+\delta_\gamma)}}$$

for a large enough constant M, then the first condition is satisfied. The second condition is then fulfilled if we choose

$$L_n^2 = N n^{\frac{d_\gamma + 4\alpha\delta_\gamma}{d_\gamma + 2\alpha(1+\delta_\gamma)}},$$

for N large enough. The inequalities for $\bar{\varepsilon}_n$ then hold as well.

For the case $\gamma = \infty$ we have to show we can choose R_n and L_n such that

$$\frac{1}{R_n^d} \log^r \frac{1}{R_n} \geq M_1 n\varepsilon_n^2,$$

$$(L_n - \sqrt{(5/2)K_6 R_n^{-d}(\log(1/(\varepsilon_n R_n^{1+d})))^{1+d}})^2 \geq Dn\varepsilon_n^2,$$

and also $\bar{\varepsilon}_n \geq \varepsilon_n$ and

$$\frac{1}{R_n^d}\Big(\log \frac{L_n}{\bar{\varepsilon}_n R_n^d}\Big)^{1+d} \leq const\ n\bar{\varepsilon}_n^2.$$

All these requirements are met if we take

$$\frac{1}{R_n^d} = Mn^{\frac{d}{d+2\alpha}} \log^v n$$

for a large enough constant M and

$$v = \frac{2(r \vee (1+d))}{2 + d/\alpha} - r$$

and L_n a large enough power of n.

Chapter 5

Semiparametric Bernstein–von Mises for the error standard deviation

5.1 Introduction

In this chapter we study the asymptotic behavior of the marginal posterior for the error standard deviation in a non-parametric, fixed design regression model with Gaussian errors, i.e. the setting described in Section 2.2.3. So we suppose we have observations $Y_1, \ldots, Y_n$ satisfying

$$Y_i = f_0(x_i) + \sigma_0 Z_i, \quad i = 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are known elements of a general design space $\mathcal{X}$, the variables $Z_1, \ldots, Z_n$ are independent, standard normal random variables and both the regression function $f_0 : \mathcal{X} \to \mathbb{R}$ and the error standard deviation $\sigma_0 > 0$ are unknown. We can then make Bayesian inference about the parameters f and σ by endowing them with independent priors π_f and π_σ, respectively, and computing the resulting posterior distribution $\Pi(\cdot \,|\, Y_1, \ldots, Y_n)$. Although in most applied problems the main interest is in the regression function f, we are in this chapter primarily interested in the asymptotic behavior of the marginal posterior distribution of the parameter σ.

The general rate of contraction result for fixed design regression obtained by Ghosal and Van der Vaart in [20] gives conditions under which the posterior for the regression function f contracts around the true f_0 at a certain rate ε_n as $n \to \infty$, under the assumption that σ_0 is known. As has been observed several times in the literature however (see e.g. [26, 50, 51]) it is relatively straightforward to extend

this result to the case that σ_0 is unknown, see also Section 2.4.3. In that case one also obtains a rate for the marginal posterior of σ. Specifically, the general results give conditions under which, for a given sequence $\varepsilon_n \to 0$, it holds that

$$\Pi\Big((f,\sigma) : \frac{1}{n}\sum_{i=1}^{n}(f-f_0)^2(x_i) + |\sigma-\sigma_0|^2 \geq M^2\varepsilon_n^2 \,|\, Y_1,\ldots,Y_n\Big) \stackrel{\mathbb{P}_0}{\to} 0 \qquad (5.1)$$

as $n \to \infty$, for every sufficiently large $M > 0$.

A result like (5.1) implies in particular that the marginal posterior for σ is asymptotically concentrated on an interval with length of the order ε_n around the true value σ_0. Now note that since ε_n is also a bound for the rate of contraction of the marginal posterior for f it is a "non-parametric rate" that will typically be slower than the parametric rate $n^{-1/2}$ if the space of regression functions that are considered is truly infinite-dimensional. If f_0 is for instance a general function of d variables with Hölder regularity β, then the optimal rate for estimating it is $n^{-\beta/(d+2\beta)}$. The rate bound ε_n for the one-dimensional parameter σ may therefore be rather crude and it is natural to ask whether in fact the actual rate of contraction for the marginal posterior for σ can be faster than the rate for the regression function f.

In the case that the regression function f is known and σ is the only unknown parameter in the problem, the classical Bernstein-von Mises (BvM) theorem asserts that under minimal regularity conditions, the posterior distribution of σ contracts around the true value σ_0 at the rate $n^{-1/2}$. Moreover, it says that the posterior law of $\sqrt{n}(\sigma-\sigma_0)$ behaves asymptotically like a normal distribution $N(\Delta_n, I_{\sigma_0}^{-1})$, with Δ_n a stochastically bounded sequence of random variables and I_{σ_0} the Fisher information for σ_0. The precise statement is recalled in the next section. In this chapter we investigate if and how this changes if the regression function f is in fact unknown. Roughly speaking we will show that if the rate ε_n for the infinite-dimensional parameter f is fast enough, then the marginal posterior distribution of σ has the same asymptotic behavior as in the case that f is known.

Our result can be viewed as a semiparametric Bernstein-von Mises theorem. In general, semiparametric BvM theorems deal with the asymptotic behavior of posterior distributions of finite-dimensional parameters in the presence of an infinite-dimensional "nuisance parameter". Theorems of this type have recently been established by several authors, see for instance [5, 10, 40, 44]. Our problem in fact fits into the general framework of Castillo [10] (up to minor adaptations) and we will use his results to derive our BvM theorem for the error standard deviation. As is explained in the cited papers, an important aspect of BvM results is that they allow to conclude that credible sets for the finite-dimensional parameter of interest, i.e. sets that receive a fixed amount α of posterior mass, are also asymptotic α-confidence sets in the frequentist sense. In other words, if a BvM theorem holds, the posterior distribution "correctly" quantifies the uncertainty

about the true value of the parameter.

After recalling the parametric BvM theorem in Section 5.2.1 we present our general semiparametric result for the error standard deviation in Section 5.2.2. It states that if the rate ε_n for the regression function f is fast enough in the sense that $n\varepsilon_n^4 \to 0$ and an entropy condition is satisfied for the space of regression function under consideration, then the BvM result holds for the marginal posterior of σ. Theorem 5.2 connects the result to the general contraction rate theorem for non-parametric regression given in Section 2.4.3. In Section 5.3 we consider the special case that the prior on f is Gaussian. In Theorem 5.3 conditions for semiparametric BvM are given in terms of the concentration function of the Gaussian prior, cf. Theorem 2.18 and the results of Section 2.6.2.3. We verify the conditions for two particular examples: a Matérn process prior and a Riemann-Liouville prior on f. The proof of our general theorem is given in Section 5.5.

5.2 General result

5.2.1 Prelude: parametric Bernstein–von Mises

The main result of this chapter is a semiparametric Bernstein–von Mises (BvM) theorem for the error standard deviation in a fixed design regression model. As a prelude we first consider the parametric case in which we observe variables $Y_1, \dots, Y_n$ satisfying

$$Y_i = f_0(x_i) + \sigma Z_i, \qquad i = 1, \dots, n,$$

for known covariates $x_i \in \mathcal{X}$ and independent standard normal random variables Z_i. We now assume that the regression function f_0 is *known*, so that the error standard deviation $\sigma > 0$ is the only unknown parameter. We denote its true value by σ_0. Observe that in this case we simply have a sample of size n from the $N(0, \sigma^2)$-distribution, given by the variables $X_i = Y_i - f_0(x_i)$, $i = 1, \dots, n$.

The BvM theorem in a smooth, parametric i.i.d. model like this one is classical. As an illustration and to connect to the semiparametric case studied ahead we briefly explain it. Let p_σ be the marginal density of X_i, $\ell_\sigma(x) = \log p_\sigma(x)$, $\dot{\ell}_\sigma(x) = \partial \ell_\sigma(x)/\partial\sigma$ and $\ddot{\ell}_\sigma(x) = \partial \dot{\ell}_\sigma(x)/\partial\sigma$. Then a Taylor expansion gives

$$\ell_\sigma(x) - \ell_{\sigma_0}(x) \approx (\sigma - \sigma_0)\dot{\ell}_{\sigma_0}(x) + \frac{1}{2}(\sigma - \sigma_0)^2 \ddot{\ell}_{\sigma_0}(x).$$

By the law of large numbers the average $-n^{-1}\sum_{i=1}^n \ddot{\ell}_{\sigma_0}(X_i)$ converges almost surely to the Fisher information $I_{\sigma_0} = -\mathbb{E}_0 \ddot{\ell}_{\sigma_0}(X_1) = \mathbb{V}\mathrm{ar}_{\sigma_0} \dot{\ell}_{\sigma_0}(X_1)$. It follows that for the full log-likelihood we have the so-called LAN approximation

$$\log \prod_{i=1}^n \frac{p_\sigma}{p_{\sigma_0}}(X_i) \approx -\frac{1}{2} I_{\sigma_0} \Big(n(\sigma - \sigma_0)^2 - 2\sqrt{n}(\sigma - \sigma_0)\Delta_n \Big),$$

where

$$\Delta_n = I_{\sigma_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\sigma_0}(X_i).$$

Note that by the central limit theorem, we have the weak convergence $\Delta_n \xrightarrow{d} N(0, I_{\sigma_0}^{-1})$ as $n \to \infty$.

If we now assume that σ_0 belongs to a compact subinterval of $(0, \infty)$ and on this interval we put a prior with a Lebesgue density π which is positive and continuous at θ_0, then for the corresponding posterior we have, for a Borel subset $B \subset \mathbb{R}$,

$$\Pi(\sqrt{n}(\sigma - \sigma_0) \in B \,|\, Y_1, \ldots, Y_n) = \frac{\int_{\sqrt{n}(\sigma-\sigma_0)\in B} \prod_{i=1}^{n} \frac{p_\sigma}{p_{\sigma_0}}(X_i)\pi(\sigma)\, d\sigma}{\int_{\mathbb{R}_+} \prod_{i=1}^{n} \frac{p_\sigma}{p_{\sigma_0}}(X_i)\pi(\sigma)\, d\sigma}.$$

By the LAN approximation, the integrands are approximately equal to a constant times

$$\pi(\sigma) \exp\Big(-\frac{1}{2} I_{\sigma_0} (\sqrt{n}(\sigma - \sigma_0) - \Delta_n)^2 \Big).$$

Making a change of variable $\sqrt{n}(\sigma - \sigma_0) = h$ we then see that the posterior probability that $\sqrt{n}(\sigma - \sigma_0)$ falls in the set B approximately equals $N(\Delta_n, I_{\sigma_0}^{-1})(B)$ for large n. This somewhat loose argumentation can be made precise and it can be shown that in probability, the total variation distance between the posterior distribution of $\sqrt{n}(\sigma - \sigma_0)$ and the $N(\Delta_n, I_{\sigma_0}^{-1})$-distribution vanishes as $n \to \infty$, cf. e.g. [49]. It is easily seen that in this case

$$\Delta_n = \frac{\sigma_0}{2\sqrt{n}} \sum_{i=1}^{n} (Z_i^2 - 1), \qquad I_{\sigma_0} = \frac{2}{\sigma_0^2}. \tag{5.2}$$

In the next section we state the semiparametric version of this result for the case that the regression function f is in fact unknown. It turns out that there is no loss of information for the error standard deviation and that under relatively mild conditions on the prior on the nonparametric part f, the asymptotic behavior of the marginal posterior for $\sqrt{n}(\sigma - \sigma_0)$ is the same as if f were known.

5.2.2 Semiparametric Bernstein–von Mises

Now suppose that we have observations $Y_1, \ldots, Y_n$ from the regression model

$$Y_i = f(x_i) + \sigma Z_i, \quad i = 1, \ldots, n, \tag{5.3}$$

with fixed and known design points $x_1, \ldots, x_n$ in the set $\mathcal{X}$, an *unknown* regression function $f : \mathcal{X} \to \mathbb{R}$, an unknown constant $\sigma > 0$, and with $Z_1, \ldots, Z_n$ independent standard Gaussian random variables. We assume that the true parameter (σ_0, f_0) belongs to the set $[a, b] \times \mathscr{F}$, for $0 < a < b < \infty$ and $\mathscr{F}$ a measurable space of functions on $\mathcal{X}$. The corresponding true distribution of the data is denoted by $\mathbb{P}_0$.

The log-likelihood is given by

$$\ell_n(\sigma, f; Y_1, \ldots, Y_n) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - f(x_i))^2.$$

We assume that for every n, the map $(\sigma, f, y) \mapsto \ell_n(\sigma, f; y_1, \ldots, y_n)$ is a measurable map on $[a,b] \times \mathscr{F} \times \mathbb{R}^n$. Note that this is the case for instance if $\mathcal{X}$ is a topological space and $\mathscr{F}$ is a measurable subset of the space of $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$, endowed with its Borel sigma-field.

To make Bayesian inference about f and σ we endow the pair (σ, f) with a product prior distribution of the form $\Pi = \pi_\sigma \times \pi_f$. Here π_σ is a distribution on $[a,b]$ with a positive and continuous Lebesgue density and π_f is a distribution on $\mathscr{F}$. In view of the measurability assumptions the corresponding posterior distribution is well defined and given by Bayes' formula. For A and B measurable subsets of $[a,b]$ and $\mathscr{F}$, respectively, the posterior measure of the set $A \times B$ is denoted by $\Pi(A \times B \,|\, Y_1, \ldots, Y_n)$ or by $\Pi(\sigma \in A, f \in B \,|\, Y_1, \ldots, Y_n)$.

The following theorem deals with the marginal posterior distribution of the parameter σ. It gives conditions under which we have, as in the case that f is known, that the posterior distribution of $\sqrt{n}(\sigma - \sigma_0)$ asymptotically behaves as an $N(\Delta_n, I_{\sigma_0}^{-1})$-distribution, where Δ_n and I_{σ_0} are as in (5.2). We still have the weak convergence

$$\Delta_n \xrightarrow{\mathrm{d}} N(0, I_{\sigma_0}^{-1})$$

under $\mathbb{P}_0$, by the central limit theorem.

The existing general contraction rate theorems (or, more precisely, their proofs) for fixed design regression give conditions under which the posterior contracts around the true parameter (σ_0, f_0). More precisely, for a sequence ε_n such that $n\varepsilon_n^2 \to \infty$ they give conditions under which there exist measurable subsets $\mathscr{F}_n$ growing to the whole space $\mathscr{F}$ such that

$$\Pi((\sigma, f) \in [a,b] \times \mathscr{F}_n : |\sigma - \sigma_0| + \|f - f_0\|_n \le \varepsilon_n \,|\, Y_1, \ldots, Y_n) \xrightarrow{\mathbb{P}_0} 1 \tag{5.4}$$

as $n \to \infty$, where the norm $\|\cdot\|_n$ is the L^2-norm associated with the empirical measure on the design points, i.e. $\|g\|_n^2 = n^{-1}\sum g^2(x_i)$. See Theorem 2.11 in Chapter 2. The case that σ_0 is known is covered by these general results as well. Following [10], we denote the posterior distribution for f in the model that σ_0 is known by $\Pi^{\sigma=\sigma_0}(\cdot \,|\, Y_1, \ldots, Y_n)$. In this notation, the general theory gives conditions under which

$$\Pi^{\sigma=\sigma_0}(f \in \mathscr{F}_n : \|f - f_0\|_n \le \varepsilon_n \,|\, Y_1, \ldots, Y_n) \xrightarrow{\mathbb{P}_0} 1 \tag{5.5}$$

as $n \to \infty$. The rate ε_n should be viewed as the contraction rate that is achieved for the non-parametric part of the statistical problem. The following theorem

states that if this rate is fast enough, namely $n\varepsilon_n^4 \to 0$, then under a typically mild additional entropy condition, we have the BvM result for the error standard deviation σ.

Theorem 5.1. *Consider positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and $n\varepsilon_n^4 \to 0$. If there exist measurable subsets $\mathscr{F}_n \subset \mathscr{F}$ such that (5.4) and (5.5) hold, and such that*

$$\int_0^{a\varepsilon_n} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\,\mathrm{d}\delta \to 0$$

holds for every $a > 0$, then, with Δ_n and I_{σ_0} given by (5.2),

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathscr{F} | Y_1, \ldots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \xrightarrow{\mathbb{P}_0} 0$$

as $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset [a, b]$.

Existing general theorems as exhibited in Chapter 2 give sufficient conditions on the prior π_f for (5.4) and (5.5) to hold. Indeed, Theorem 2.11 and the preceding result imply the following.

Theorem 5.2. *Consider positive numbers $\bar{\varepsilon}_n \geq \varepsilon_n$ such that $n\varepsilon_n^2 \to \infty$ and $n\bar{\varepsilon}_n^4 \to 0$. Suppose that for every $C_1 > 1$, there exist measurable subsets $\mathscr{F}_n \subset \mathscr{F}$ and a constant $C_2 > 0$ such that*

$$\pi_f(f : \|f - f_0\|_n \leq \varepsilon_n) \geq \exp(-n\varepsilon_n^2), \tag{5.6}$$

$$\pi_f(\mathscr{F} \backslash \mathscr{F}_n) \leq \exp(-C_1 n\varepsilon_n^2), \tag{5.7}$$

$$\log N(\bar{\varepsilon}_n, \mathscr{F}_n, \|\cdot\|_n) \leq C_2 n\bar{\varepsilon}_n^2, \tag{5.8}$$

$$\text{for all } a > 0\text{:} \quad \int_0^{a\bar{\varepsilon}_n} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\,\mathrm{d}\delta \to 0. \tag{5.9}$$

Then with Δ_n and I_{σ_0} given by (5.2),

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathscr{F} | Y_1, \ldots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \xrightarrow{\mathbb{P}_0} 0$$

as $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset [a, b]$.

Theorem 2.11 states that under conditions (5.6)–(5.8) we have the rate of contraction $\bar{\varepsilon}_n$ for the marginal posterior of the regression function f. It gives the same rate $\bar{\varepsilon}_n$ for the marginal posterior of the parameter σ, which is typically only a crude result. The theorem above states that under the additional assumptions $n\bar{\varepsilon}_n^4 \to 0$ and (5.9) we in fact have a rate $n^{-1/2}$ for σ and the marginal posterior is asymptotically normal.

Conditions (5.8) and (5.9) can sometimes be verified in one go by showing that for a constant $L > 0$,

$$\int_0^{\varepsilon} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\,\mathrm{d}\delta \leq L\sqrt{n}\varepsilon^2. \tag{5.10}$$

for all $\varepsilon > 0$ small enough. The fact that (5.9) is then satisfied follows immediately from the assumption $n\bar{\varepsilon}_n^4 \to 0$. For (5.8) we note that since $\delta \mapsto N(\delta, \mathscr{F}_n, \|\cdot\|_n)$ is decreasing, we have

$$\bar{\varepsilon}_n \sqrt{\log N(\bar{\varepsilon}_n, \mathscr{F}_n, \|\cdot\|_n)} \leq \int_0^{\bar{\varepsilon}} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\, \mathrm{d}\delta.$$

Hence if (5.10) holds, then $\log N(\bar{\varepsilon}_n, \mathscr{F}_n, \|\cdot\|_n) \leq L^2 n\bar{\varepsilon}_n^2$.

5.3 Gaussian priors on the regression function

We now specialize to the case that $\mathcal{X} = [0,1]^d$ for some $d \in \mathbb{N}$. As prior π_f on the regression function f we employ the law of a Gaussian random element W in the space $C([0,1]^d)$ of continuous functions on $[0,1]^d$. We denote the reproducing kernel Hilbert space (RKHS) of W by $\mathbb{H}$. For $f_0 \in C([0,1]^d)$ the true regression function, the concentration function is denoted by φ_{f_0}, that is to say

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon) \tag{5.11}$$

See Section 2.5 for these fundamental concepts.

The general theory for Gaussian process priors says that if $\varepsilon_n \to 0$ is such that $n\varepsilon_n^2 \to \infty$ and

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2, \tag{5.12}$$

then the marginal posteriors for f and σ contract at the rate ε_n around their true values, cf. Theorem 2.22. The following theorem states that if in addition the rate ε_n is fast enough and the sample paths of W are regular enough, then we have the BvM result for σ.

Recall that $C^\alpha([0,1]^d)$ is the space of functions $f : [0,1]^d \to \mathbb{R}$ with uniformly bounded partial derivatives up to order $\underline{\alpha}$ such that the partial derivatives of order $\underline{\alpha}$ are uniformly Lipschitz of order $\alpha - \underline{\alpha}$, where $\underline{\alpha}$ is the largest integer strictly smaller than α, see Section 4.2.1.

Theorem 5.3. *Suppose that W almost surely takes values in $C^\alpha([0,1]^d)$ for $\alpha > d/2$ and that (5.12) holds for numbers $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$, $n\varepsilon_n^4 \to 0$ and $n\varepsilon_n^{4\alpha/d} \to 0$. Then with Δ_n and I_{σ_0} given by (5.2),*

$$\sup_B \left| \Pi(\sqrt{n}(\sigma - \sigma_0) \in B, f \in \mathscr{F} | Y_1, \ldots, Y_n) - N(\Delta_n, I_{\sigma_0}^{-1})(B) \right| \overset{\mathbb{P}_0}{\to} 0$$

as $n \to \infty$, where the supremum is taken over all measurable subsets $B \subset [a,b]$.

Note that $n\varepsilon_n^{4\alpha/d} \to 0$ already follows from $n\varepsilon_n^4 \to 0$ if $\alpha \geq d$, so it is only an additional condition when $d/2 < \alpha < d$.

Proof. We verify the conditions of Theorem 5.2. By Theorem 2.18 and the fact that $\|\cdot\|_n \le \|\cdot\|_\infty$, condition (5.6) follows from (5.12). Now let $C > 1$ be given and define

$$\mathscr{F}_n = (K\sqrt{n}\varepsilon_n \mathbb{H}_1 + \varepsilon_n C_1) \cap L\sqrt{n}\varepsilon_n C_1^\alpha$$

for some appropriate constants K and L which will be further specified below, where $\mathbb{H}_1$ is the unit ball in the RKHS $\mathbb{H}$, C_1 is the unit ball in $C([0,1]^d)$ and C_1^α is the unit ball in the Hölder space $C^\alpha([0,1]^d)$.

Because the sieve $\mathscr{F}_n$ is an intersection, one has

$$\mathbb{P}(W \notin \mathscr{F}_n) \le \mathbb{P}(W \notin K\sqrt{n}\varepsilon_n \mathbb{H}_1 + \varepsilon_n C_1) + \mathbb{P}(W \notin L\sqrt{n}\varepsilon_n C_1^\alpha).$$

By Borell's inequality (see Theorem 2.16) and the fact that $\Phi^{-1}(y) \ge -\sqrt{(5/2)\log(1/y)}$ for small y, the first term on the right is bounded by $\exp(-(1/2)(K - \sqrt{5/2})^2 n\varepsilon_n^2)$, provided that $K \ge \sqrt{5/2}$. Since by assumption we can also view W as a Gaussian random element in the Hölder space $C^\alpha([0,1]^d)$, we have the inequality

$$\mathbb{P}(W \notin L\sqrt{n}\varepsilon_n C_1^\alpha) = \mathbb{P}(\|W\|_\alpha \ge L\sqrt{n}\varepsilon_n) \le 2\exp\Big(-\frac{L^2 n\varepsilon_n^2}{8\mathbb{E}\|W\|_\alpha^2}\Big)$$

for the second term, according to Theorem 2.17. If follows that by setting K and L large enough, we can ensure that condition (5.7) holds.

For the entropy conditions we first note that $\mathscr{F}_n \subset K\sqrt{n}\varepsilon_n \mathbb{H}_1 + \varepsilon_n C_1$ and thus we obtain (5.8) (with $\bar{\varepsilon}_n = \varepsilon_n$) in the same way as in the proof of Theorem 2.18 (see the proof of Theorem 2.1 in [50]). Because $\mathscr{F}_n \subset L\sqrt{n}\varepsilon_n C_1^\alpha$, we have

$$\log N(\delta, \mathscr{F}_n, \|\cdot\|_\infty) \le \log N(\delta/(L\sqrt{n}\varepsilon_n), C_1^\alpha, \|\cdot\|_\infty) \le M(L\sqrt{n}\varepsilon_n/\delta)^{\frac{d}{\alpha}}$$

for a constant $M > 0$, cf. Theorem 2.7.1 of [53]. Since $\alpha > d/2$, it follows that for $a > 0$ the entropy integral

$$\int_0^{a\varepsilon_n} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\,\mathrm{d}\delta$$

is bounded by a constant time $n^{\frac{d}{4\alpha}}\varepsilon_n$. This converges to zero by assumption, which shows that (5.9) holds. □

5.4 Examples: specific Gaussian priors

In this section we stay in the setting of the preceding one, so $\mathcal{X} = [0,1]^d$ for $d \in \mathbb{N}$ and we put a Gaussian prior on the regression function f, the law of a process W. We verify the conditions of Theorem 5.3 for two particular examples of a Gaussian prior on f. In the first example we choose a Matérn prior on a multivariate regression function. In the second example we consider the case $d = 1$ and choose a Riemann-Liouville type prior.

5.4.1 The Matérn prior

The Matérn process $(W_t : t \in [0,1]^d)$ with parameter $\alpha > 0$ is the zero mean Gaussian process with covariance function

$$\mathbb{E}W_s W_t = \int_{\mathbb{R}^d} e^{i\lambda^T(s-t)} m(\lambda)\,\mathrm{d}\lambda,$$

where the spectral density m is given by

$$m(\lambda) = \frac{1}{(1+\|\lambda\|^2)^{\alpha+d/2}}$$

with $\|\lambda\|$ the Euclidean norm on $\mathbb{R}^d$ and $\alpha > 0$. The Matérn process is a popular prior in Bayesian non-parametrics, see for instance Rasmussen and Williams [39] and the references therein. It is not difficult to see that there exists a version of the Matérn process that takes its values in $C^\gamma([0,1]^d)$ for any $\gamma < \alpha$, see van der Vaart and van Zanten [52].

For $\beta > 0$ a real number, the Sobolev space $H^\beta([0,1]^d)$ consists of all functions f on $[0,1]^d$ that can be extended to a function f on $\mathbb{R}^d$ with a Fourier transform $\hat{f}$ satisfying

$$\int |\hat{f}|^2 (1+\|\lambda\|^2)^\beta \, d\lambda < \infty.$$

Now suppose that for some $\beta > 0$, the true regression function is β-regular both in Hölder and Sobolev sense, i.e. $f_0 \in C^\beta([0,1]^d) \cap H^\beta([0,1]^d)$.

It is shown in Section IV of [52] that for such f_0 the inequality (5.12) holds for ε_n proportional to $n^{-(\alpha\wedge\beta)/(d+2\alpha)}$.

In this situation the conditions of Theorem 5.3 are satisfied if there exists a $\gamma < \alpha$ such that

$$\begin{aligned} n\varepsilon_n^2 &\to \infty, \\ n\varepsilon_n^4 &\to 0, \\ \gamma &> d/2, \\ n\varepsilon_n^{4\gamma/d} &\to 0. \end{aligned}$$

The first condition is always fulfilled. A $\gamma < \alpha$ such that the third and fourth conditions are verified exists as soon as $\alpha > d/2$ and $n\varepsilon_n^{4\alpha/d} \to 0$. Hence, the conditions of Theorem 5.3 are satisfied if $\alpha > d/2$ and $n\varepsilon_n^{4(1\wedge(\alpha/d))} \to 0$. Straightforward computations show that these requirements are fulfilled if and only if

$$\begin{aligned} \frac{\alpha}{d} &> \frac{1}{4} + \frac{1}{4}\sqrt{5}, \\ \frac{\beta}{d} &> \Big(\frac{1}{2} + \frac{d}{4\alpha}\Big) \wedge \Big(\frac{\alpha}{2d} + \frac{1}{4}\Big), \end{aligned} \tag{5.13}$$

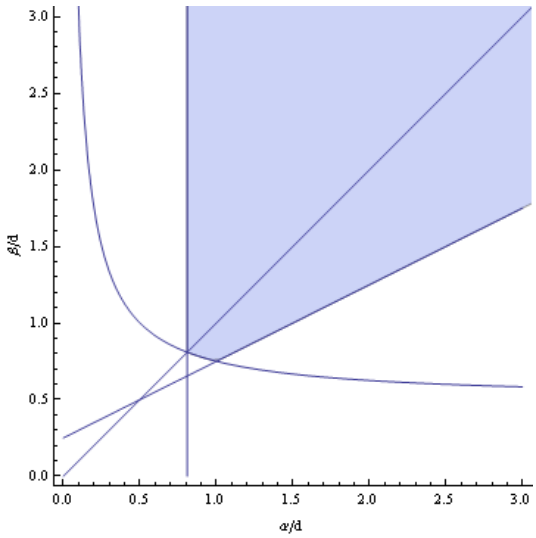

Figure 5.1: Values for the smoothness β of the true regression function f_0 and the regularity α of the Gaussian prior for which we have shown the BvM result holds.

and hence the BvM statement of Theorem 5.3 for the marginal posterior distribution of σ holds under these conditions.

The collection of α's and β's we found is sketched in Figure 5.1. The figure makes clear that for the BvM result to hold, it is not necessary to estimate the regression function f_0 at an optimal rate. In particular, it is not necessary that the smoothness α of the prior matches the smoothness β of the unknown regression function exactly. An arbitrary amount of undersmoothing ($\beta > \alpha$) is allowed and also some degree of oversmoothing ($\beta < \alpha$).

We note that it is not ruled out that the area for which BvM holds is actually larger than what we found. Our general theorems seem too crude however to shed more light on this issue. It is conceivable that more insight can be obtained by a more detailed analysis, tailored to the particular statistical problem and prior, in the spirit of [8].

5.4.2 A Riemann-Liouville type prior

In this subsection we consider the case $d = 1$, i.e. the true regression function is an unknown element $f_0 \in C[0, 1]$.

For $\alpha > 0$ and W a standard Brownian motion, the Riemann-Liouville process

with parameter α is defined by

$$R_t^\alpha = \int_0^t (t-s)^{\alpha-1/2}\,\mathrm{d}W_s.$$

It can be interpreted as the $(\alpha - 1/2)$-fold iterated integral of Brownian motion. The process R^α and its higher derivatives (if they exist) vanish at zero. In order to enlarge the class of functions that are well approximated by the process we modify it slightly, following [50]. Let $\underline{\alpha}$ be the biggest integer strictly smaller than α, and let $Z_1, \ldots, Z_{\underline{\alpha}+1}$ be independent standard normal random variables, independent of the Riemann-Liouville process R^α. Define the Riemann-Liouville-type process X as follows:

$$X_t = \sum_{k=0}^{\underline{\alpha}+1} Z_k t^k + R_t^\alpha.$$

The process $(X_t : t \in [0,1])$ is zero-mean Gaussian and can be seen as a random element in $C[0,1]$. For integer α the Riemann-Liouville process is simply a multiply integrated Brownian motion, which is a well-established prior in Bayesian non-parametrics. Its use goes back at least to Wahba [55].

Since Brownian motion has "regularity" 1/2, the Riemann-Liouville process with parameter α is expected to be "regular" of order α in an appropriate sense. Indeed it can be shown that the process R^α, and hence also the process X, has a version that take values in $C^\gamma[0,1]$ for all $\gamma < \alpha$, cf. Lifshits and Simon [35]. Upper bounds for the left hand side of (5.12) in this case are given in [10], see also [50]. If f_0 is in $C^\beta[0,1]$ for some $\beta \geq \alpha$, then the left hand side of (5.12) is bounded from above by a multiple of $\varepsilon_n^{-1/\alpha}$. For $\beta < \alpha$, the upper bound in [10] is $\varepsilon_n^{-(2\alpha-2\beta+1)/\beta}\log(1/\varepsilon_n)$. It follows that condition (5.12) is satisfied for ε_n a multiple of $(\log n/n)^{\beta/(1+2\alpha)}$ if $\beta < \alpha$ and for ε_n a multiple of $n^{-\frac{\alpha}{1+2\alpha}}$ if $\beta \geq \alpha$.

These conditions are almost the same as in the Matérn prior case. The log factor does not affect the pairs (α, β) for which the inequalities are true. We thus obtain that for the Riemann-Liouville prior as well, the BvM statement of Theorem 5.3 holds if the regularity β of the truth and the regularity α of the prior are related as in (5.13), for $d = 1$. Again, Figure 5.1 visualizes the set of α's and β's.

5.5 Proof of main result

In this section we give the proof of Theorem 5.1.

It is convenient to describe the model by the parameter (θ, f) with $\theta = 1/\sigma^2$. For this parametrization the log-likelihood is given by

$$\ell_n(\theta, f) = \frac{n}{2}\log\frac{\theta}{2\pi} - \frac{\theta}{2}\sum_{i=1}^n (Y_i - f(x_i))^2.$$

The first step in the proof is finding an appropriate expansion for the log-likelihood ratio $\Lambda_n(\theta, f) = \ell_n(\theta, f) - \ell_n(\theta_0, f_0)$. We define an inner product $\langle \cdot, \cdot \rangle_L$ on pairs (θ, f) of inverse variances and regression functions by

$$\langle (\theta, f), (\theta', f') \rangle_L = \frac{\theta\theta'}{2\theta_0^2} + \frac{\theta_0}{n} \sum_{i=1}^{n} f(x_i) f'(x_i).$$

The corresponding norm is denoted by $\| \cdot \|_L$, so

$$\|\theta, f\|_L^2 = \frac{\theta^2}{2\theta_0^2} + \theta_0 \|f\|_n^2.$$

Note that although it is not made explicit in the notation, the inner product and the norm obviously depend on the sample size n (and on the true parameter θ_0).

Straightforward manipulations yield the following lemma.

Lemma 5.4. *We have*

$$\Lambda_n(\theta, f) = -\frac{n}{2} \|\theta - \theta_0, f - f_0\|_L^2 + \sqrt{n} W_n(\theta - \theta_0, f - f_0) + R_n(\theta, f),$$

where

$$\begin{aligned} W_n(\theta, f) &= -\frac{\theta}{2\theta_0\sqrt{n}} \sum_{i=1}^{n} (Z_i^2 - 1) + \sqrt{\frac{\theta_0}{n}} \sum_{i=1}^{n} f(x_i) Z_i, \\ R_n(\theta, f) &= \frac{n}{2} \Big(\log\theta - \log\theta_0 - \frac{\theta - \theta_0}{\theta_0} + \frac{(\theta - \theta_0)^2}{2\theta_0^2} \Big) \\ &\quad - \frac{1}{2} n (\theta - \theta_0) \|f - f_0\|_n^2 + \frac{\theta - \theta_0}{\sqrt{\theta_0}} \sum_{i=1}^{n} (f(x_i) - f_0(x_i)) Z_i. \end{aligned}$$

We are now in the situation that we can apply Theorem 1 of [10]. Strictly speaking this theorem does not allow the dependence of the inner product $\langle \cdot, \cdot \rangle_L$ on n that we have, but inspection of Castillo's proof shows that this causes no problems. Since our LAN-norm has the property that the norm $\|\cdot, 0\|_L^2$ on $\mathbb{R}$ is independent of n, only minor adaptations of that proof are necessary. We note that our change of variables $\theta = 1/\sigma^2$ helps to establish a direct connection with the setup of [10], since the map W_n is linear in θ.

Castillo's theorem asserts that if there exists positive numbers δ_n such that $n\delta_n^2 \to \infty$ and measurable subsets $\mathscr{F}_n \subset \mathscr{F}$ such that

$$\Pi((\theta, f) \in [1/b^2, 1/a^2] \times \mathscr{F}_n : \|\theta - \theta_0, f - f_0\|_L \le \delta_n \,|\, Y_1, \ldots, Y_n) \xrightarrow{\mathbb{P}_0} 1, \quad (5.14)$$

$$\Pi^{\theta=\theta_0}(f \in \mathscr{F}_n : \|0, f - f_0\|_L \le \delta_n/\sqrt{2} \,|\, Y_1, \ldots, Y_n) \xrightarrow{\mathbb{P}_0} 1, \quad (5.15)$$

$$\sup_{\substack{(\theta,f) \in [1/b^2, 1/a^2] \times \mathscr{F}_n : \\ \|\theta - \theta_0, f - f_0\|_L \le \delta_n}} \frac{|R_n(\theta, f) - R_n(\theta_0, f)|}{1 + n(\theta - \theta_0)^2} \xrightarrow{\mathbb{P}_0} 0, \quad (5.16)$$

then

$$\sup_B \left| \Pi(\sqrt{n}(\theta-\theta_0) \in B, f \in \mathscr{F} | Y_1, \dots, Y_n) - N\Big(\frac{W_n(1,0)}{\|1,0\|_L^2}, \frac{1}{\|1,0\|_L^2}\Big)(B) \right| \overset{\mathbb{P}_0}{\to} 0. \tag{5.17}$$

The next step is to show that conditions (5.14)–(5.16) hold for δ_n equal to a constant times ε_n under the assumptions of Theorem 5.1.

Since $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$, we have $\|\theta-\theta_0, f-f_0\|_L \le C(|\theta - \theta_0| + \|f\|_n)$, for a constant $C > 0$ only depending on θ_0. Recalling that $\theta = 1/\sigma^2$ and that σ belongs to the compact interval $[a, b]$ we see that $\|\theta-\theta_0, f-f_0\|_L \le C'(|\sigma-\sigma_0| + \|f\|_n)$, for a constant $C' > 0$. It follows that under assumptions (5.4) and (5.5), conditions (5.14) and (5.15) hold for δ_n a multiple of ε_n.

Next we consider (5.16). Define $V_n = \{(\theta, f) \in [1/b^2, 1/a^2] \times \mathscr{F}_n : \|\theta-\theta_0, f-f_0\|_L \le \delta_n\}$. We consider the three terms in the definition of R_n in the statement of Lemma 5.4 separately. For $\theta_0 \in V_n$ it holds that $|\theta-\theta_0|$ is bounded by a multiple of δ_n. By Taylor's formula, the first term in the definition of R_n is $nO(|\theta-\theta_0|^3)$ for θ close to θ_0, and hence the first term is bounded by a multiple of $(1+n(\theta-\theta_0)^2)\delta_n$ on V_n. For the second term, note that $x \mapsto x/(1+nx^2)$ is maximal at $x = n^{-1/2}$, and equal to $n^{-1/2}/2$ at that point. It follows that

$$\sup_{(\theta,f)\in V_n} \frac{n|\theta-\theta_0|\|f-f_0\|_n^2}{1+n(\theta-\theta_0)^2} \le \frac{1}{2}\sqrt{n} \sup_{(\theta,f)\in V_n} \|f-f_0\|_n^2 \le \frac{\sqrt{n}\delta_n^2}{2\theta_0}.$$

Similarly, the supremum over V_n of the third term divided by $1+n(\theta-\theta_0)^2$ is bounded by

$$\frac{1}{2\sqrt{\theta_0}} \sup_{\substack{f\in\mathscr{F}_n \\ \sqrt{\theta_0}\|f-f_0\|_n \le \delta_n}} |\mathbb{G}_n f - \mathbb{G}_n f_0|,$$

where $\mathbb{G}_n$ is the Gaussian random map defined by

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(x_i) Z_i.$$

The norm $\|\cdot\|_n$ is precisely the natural semi-norm associated with the Gaussian process $\mathbb{G}_n$, in the sense that $\mathbb{E}_0(\mathbb{G}_n f - \mathbb{G}_n g)^2 = \|f-g\|_n^2$. Therefore, the well-known maximal inequality for sub-Gaussian processes, cf. e.g. [53], Corollary 2.2.8, implies that

$$\mathbb{E}_0 \sup_{\substack{f\in\mathscr{F}_n \\ \sqrt{\theta_0}\|f-f_0\|_n \le \delta_n}} |\mathbb{G}_n f - \mathbb{G}_n f_0| \le K \int_0^{\delta_n/\sqrt{\theta_0}} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\, \mathrm{d}\delta$$

for some constant $K > 0$. Altogether we conclude that the left-hand side of (5.16) is

$$O_{\mathbb{P}_0}\Big(\delta_n + \sqrt{n}\delta_n^2 + \int_0^{\delta_n/\sqrt{\theta_0}} \sqrt{\log N(\delta, \mathscr{F}_n, \|\cdot\|_n)}\, \mathrm{d}\delta\Big)$$

for $n \to \infty$. For δ_n a multiple of ε_n this is $o_{\mathbb{P}_0}(1)$ under the assumptions of the theorem, hence (5.16) holds as well.

We have now established that (5.17) holds under the conditions of Theorem 5.1. Next, observe that $\|1,0\|_L^2 = 1/(2\theta_0^2)$ and

$$\frac{W_n(1,0)}{\|1,0\|_L^2} = -\frac{\theta_0}{\sqrt{n}} \sum_{i=1}^{n} (Z_i^2 - 1) \xrightarrow{d} N(0, 2\theta_0^2)$$

under $\mathbb{P}_0$, by the central limit theorem. The statement of the theorem then follows by an application of Lemma 5.5 below, which gives a total variation version of the delta method, tailored to our situation. We apply the lemma with X_n a random variable which has the posterior distribution of θ as law, $x_0 = \theta_0$, $\mu_n = W_n(1,0)/\|1,0\|_L^2$, $\sigma^2 = 1/\|1,0\|_L^2 = 2\theta_0^2$ and $f(x) = 1/\sqrt{x}$. The lemma deals with the total variation distance between deterministic distributions. We can use it in our stochastic setting since $W_n(1,0)/\|1,0\|_L^2$ converges in distribution and hence is uniformly tight.

We denote the total variation distance between two probability measure μ and ν by $d_{TV}(\mu,\nu)$ and the law, or distribution of a random variable X by $\mathcal{L}(X)$.

Lemma 5.5. *Let X_n be a sequence of random variables such that*

$$d_{TV}(\mathcal{L}(\sqrt{n}(X_n - x_0)), N(\mu_n, \sigma^2)) \to 0, \tag{5.18}$$

for $x_0 \in \mathbb{R}$, $\sigma^2 > 0$ and μ_n a bounded sequence. Let $f : \mathbb{R} \to \mathbb{R}$ be a function that is twice continuously differentiable on a neighborhood of x_0 and $f'(x_0) \neq 0$. Then

$$d_{TV}(\mathcal{L}(\sqrt{n}(f(X_n) - f(x_0))), N(f'(x_0)\mu_n, (\sigma f'(x_0))^2)) \to 0.$$

Proof. We suppose for definiteness that $f'(x_0) > 0$. It follows from the assumptions on f that there exist neighborhoods U and V of x_0 and $f(x_0)$ such that f is an invertible (in this case increasing) bijection between U and V. The distribution $N(x_0 + \mu_n/\sqrt{n}, \sigma^2/n)$ concentrates around x_0 as $n \to \infty$. Hence, by (5.18), so does $\mathcal{L}(X_n)$ and hence the law $\mathcal{L}(f(X_n))$ concentrates around $f(x_0)$. Therefore, we only need to prove that

$$\sup_{B \subset V} |\mathbb{P}(f(X_n) \in B) - N(f(x_0) + \mu_n f'(x_0)/\sqrt{n}, (f'(x_0))^2\sigma^2/n)(B)| \to 0,$$

or, equivalently,

$$\sup_{A \subset U} |\mathbb{P}(X_n \in A) - N(f(x_0) + \mu_n f'(x_0)/\sqrt{n}, (f'(x_0))^2\sigma^2/n)(f(A))| \to 0.$$

Using (5.18), a change of variables and some straightforward algebra we then see that it suffices to show that

$$\int_U \Big| \frac{1}{\tau_n} \varphi\Big(\frac{f'(x_0)(x - x_0) - \delta_n)}{\tau_n} \Big) f'(x_0) - \frac{1}{\tau_n} \varphi\Big(\frac{f(x) - f(x_0) - \delta_n)}{\tau_n} \Big) f'(x) \Big| \, dx \to 0,$$

where φ denotes the standard normal density, $\delta_n = \mu_n f'(x_0)/\sqrt{n}$ and $\tau_n = \sigma f'(x_0)/\sqrt{n}$.

Consider the shrinking sets $U_n = \{x \in U : |x - x_0| \le K_n \tau_n\}$ for a sequence $K_n \to \infty$ such that $K_n^3 \tau_n \to 0$. For $x \in U_n^c$ it holds that $|f(x) - f(x_0)| \ge cK_n\tau_n$ for some $c > 0$ and hence

$$\int_{U_n^c} \frac{1}{\tau_n} \varphi\Big(\frac{f(x) - f(x_0) - \delta_n)}{\tau_n}\Big) f'(x)\, dx \le \int_{|z| > cK_n} \varphi(z - \mu_n/\sigma)\, dz \to 0.$$

Similarly,

$$\int_{U_n^c} \frac{1}{\tau_n} \varphi\Big(\frac{f'(x_0)(x - x_0) - \delta_n)}{\tau_n}\Big)\, dx \to 0.$$

Since φ is Lipschitz and f is twice continuously differentiable we have

$$\frac{1}{\tau_n} \int_{U_n} \Big|\varphi\Big(\frac{f'(x_0)(x - x_0) - \delta_n)}{\tau_n}\Big) - \varphi\Big(\frac{f(x) - f(x_0) - \delta_n)}{\tau_n}\Big)\Big|\, dx \lesssim K_n^3 \tau_n \to 0.$$

Finally, observe that by definition of U_n,

$$\begin{aligned} &\frac{1}{\tau_n} \int_{U_n} \varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big) |f'(x) - f'(x_0)|\, dx \\ &\quad \lesssim K_n \int_{U_n} \varphi\Big(\frac{f(x) - f(x_0) - \delta_n}{\tau_n}\Big)\, dx \lesssim K_n^2 \tau_n \to 0. \end{aligned}$$

The proof is completed by combining the convergence statements derived in this paragraph. □

Bibliography

[1] F. Aurzada, I. A. Ibragimov, M. A. Lifshits, and J. H. van Zanten. Small deviations of smooth stationary Gaussian processes. *Teor. Veroyatn. Primen.*, 53(4):788–798, 2008.

[2] Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(4):825–848, 2008.

[3] Andrew Barron, Mark J. Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2): 536–561, 1999.

[4] Eduard Belitser and Subhashis Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2): 536–559, 2003.

[5] P.J Bickel and B.J.K Kleijn. The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40(1):206–237, 2012.

[6] Carl de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001.

[7] Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30(2):207–216, 1975.

[8] Ismaël Castillo. Semiparametric Bernstein - von Mises theorem and bias, illustrated with Gaussian process priors. To appear in Sankhya A.

[9] Ismaël Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.*, 2:1281–1299, 2008.

[10] Ismaël Castillo. A semiparametric Bernstein - von Mises theorem for Gaussian process priors. *Probab. Theory Relat. Fields*, 152(1):53–99, 2012.

[11] Dennis D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923, 1993.

[12] Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–26, 1986.

[13] Persi Diaconis and David Freedman. On inconsistent Bayes estimates of location. *Ann. Statist.*, 14(1):68–87, 1986.

[14] Persi W. Diaconis and David Freedman. Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli*, 4(4):411–444, 1998.

[15] J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

[16] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158, 1999.

[17] Subhashis Ghosal. Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280, 2001.

[18] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007.

[19] Subhashis Ghosal and Aad W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001.

[20] Subhashis Ghosal and Aad W. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35:697–723, 2007.

[21] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[22] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. On Bayesian adaptation. In *Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002)*, volume 79, pages 165–175, 2003.

[23] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2: 63–89, 2008.

[24] Dave Higdon. Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer, London, 2002.

[25] Tzee-Ming Huang. Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593, 2004.

[26] R. de Jonge and J. H. van Zanten. Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320, 2010.

[27] R. de Jonge and J. H. van Zanten. Semiparametric Bernstein-von Mises for the error standard deviation. Preprint, 2012.

[28] R. de Jonge and J. H. van Zanten. Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. Preprint, 2012.

[29] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961.

[30] Willem Kruijer and Aad van der Vaart. Posterior convergence rates for Dirichlet mixtures of beta densities. *J. Statist. Plann. Inference*, 138(7): 1981–1992, 2008.

[31] Willem Kruijer, Judith Rousseau, and Aad van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257, 2010.

[32] James Kuelbs, Wenbo V. Li, and Werner Linde. The Gaussian measure of shifted balls. *Probab. Theory Related Fields*, 98(2):143–162, 1994.

[33] J. Lember and A.W. van der Vaart. On universal Bayesian adaptation. *Statistics and Decisions*, 25:127–152, 2007.

[34] Wenbo V. Li and Werner Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.

[35] Mikhail Lifshits and Thomas Simon. Small deviations for fractional stable processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 41(4):725–752, 2005.

[36] F. H. van der Meulen, A. W. van der Vaart, and J. H. van Zanten. Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli*, 12(5):863–888, 2006.

[37] Sonia Petrone and Larry Wasserman. Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(1):79–100, 2002.

[38] David Pollard. *Empirical processes: theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA, 1990.

[39] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006.

[40] Vincent Rivoirard and Judith Rousseau. Bernstein-von Mises theorem for linear functionals of the density. To appear in The Annals of Statistics.

[41] Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180, 2010.

[42] Larry L. Schumaker. *Spline functions: basic theory*. John Wiley & Sons Inc., New York, 1981. ISBN 0-471-76475-2. Pure and Applied Mathematics, A Wiley-Interscience Publication.

[43] Loraine Schwartz. On consistency of Bayes procedures. *Proc. Nat. Acad. Sci. U.S.A.*, 52:46–49, 1964.

[44] Xiaotong Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.*, 97(457):222–235, 2002.

[45] Margaret B. Short, David M. Higdon, and Philipp P. Kronberg. Estimation of Faraday rotation measures of the near galactic sky using Gaussian process models. *Bayesian Anal.*, 2(4):665–680, 2007.

[46] Charles J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990.

[47] Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–184, 1994.

[48] Surya T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā*, 68(1):90–110, 2006.

[49] A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[50] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36 (3):1435–1463, 2008.

[51] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

[52] Aad van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.*, 12:2095–2119, 2011.

[53] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[54] A.W. van der Vaart and J.H. van Zanten. *Reproducing Kernel Hilbert Spaces of Gaussian priors*, volume 3 of *IMS Collections*, pages 200–222. Institute of Mathematical Statistics, 2008.

[55] Grace Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40(3):364–372, 1978.

[56] Yuefeng Wu and Subhashis Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, 2:298–331, 2008.

Summary

Posterior contraction for conditionally Gaussian priors

The goal of statistics is to draw sensible conclusions from data. In mathematical statistics, observed data is assumed to be generated according to some unknown probability distribution. The aim is to find the unknown probability distribution using the available observations. In parametric statistics this is typically done by considering a finite-dimensional parametric family of probability distributions and estimating a parameter using the data. On the other hand, in non-parametric statistics one deals with infinite dimensional statistical models. The model is then described by some non-parametric parameter such as a probability distribution or a regression function.

In Bayesian statistics one makes inference by choosing a probability distribution on the statistical model. We distinguish between the prior distribution and the posterior distribution. These distributions represent the statistician's belief about the parameter before and after the data has become available. In the frequentist's setup however, the parameter is assumed to have some true value. An asymptotic analysis is then possible by considering the posterior measure of shrinking neighborhoods around the true parameter as the number of observations increases. We are interested in how fast the posterior concentrates around the true parameter.

In this thesis we consider two examples of a conditionally Gaussian process for the construction of a prior distribution on certain statistical models indexed by a function. The two examples that we consider are defined by choosing the paths of the process to be either tensor-product spline functions or location-scale kernel mixtures. The use of log-spline models and kernel mixtures to construct priors on probability densities is well-established in Bayesian non-parametrics. The use of Gaussian priors provides a unified approach to obtain rates of posterior contraction in various statistical settings. We consider density estimation, classification and fixed design regression settings.

If the true function is a function of d variables with smoothness level α in the sense of Hölder, then the optimal rate of posterior contraction is of the order

$n^{-\frac{\alpha}{d+2\alpha}}$ if n is the number of observations. We show that it is possible to construct Gaussian priors from either the spline functions or the kernel mixtures which actually achieve posterior contraction at a near optimal rate. These priors will however depend on α, an unknown characteristic of the function to be estimated.

We show that in both cases it is possible to define a new procedure, based on these Gaussian priors, which also achieves a near optimal rate of posterior contraction, but which itself does not depend on the level of smoothness of the function of interest. This procedure thus adapts to the smoothness level.

In the last chapter of this thesis, we focus on posterior contraction in the setting of fixed design regression with Gaussian errors. In this setting, the variance of the errors is a finite dimensional nuisance parameter which we can equip with a prior as well. The posterior contraction results imply in particular the concentration of posterior mass around this finite dimensional parameter at a non-parametric rate. We however know that posterior contraction in the finite-dimensional parameter case is typically faster: the optimal rate is $n^{-1/2}$. We show via a semi-parametric Bernstein-von Mises result that it is possible to achieve posterior contraction around the finite dimensional parameter at rate $n^{-1/2}$ if we equip the infinite dimensional parameter, the regression function f, as before with a Gaussian prior distribution.

Acknowledgments

Dankwoord

Dit proefschrift vormt het sluitstuk van vier jaar werk als promovendus in de mathematische statistiek, in het bijzonder op het gebied van de niet-parametrische Bayesiaanse statistiek. Realisatie van dit werk zou niet mogelijk zijn geweest zonder de betrokkenheid die ik heb ervaren van allen die ik heb leren kennen gedurende het traject

Ik wil eerst met name Harry van Zanten danken dat hij mij kennis heeft laten maken met niet-parametrische Bayesiaanse statistiek, en dat hij mij de kans heeft gegeven hierin promotieonderzoek te verrichten. Zijn suggesties en commentaar hebben een waardevolle bijdrage geleverd aan het verloop van het onderzoek en de manier waarop dit is vastgelegd in dit werk.

Verder bedank ik Eduard Belitser, Remco van der Hofstad, Frank van der Meulen, Rui Pires da Silva Castro, Judith Rousseau en Aad van der Vaart voor het aanvaarden van hun positie in de promotiecommissie en de voor de tijd die zij hebben genomen om dit proefschrift te lezen.

Ik dank mijn verscheidene kamergenoten voor het goede persoonlijke contact en evenzeer de 'leden' van de zogeheten Bayes club, en alle andere collega's bij wie altijd een plaatsje vrij was tijdens de lunch. Ook het contact met mede-promovendi van andere universiteiten op diverse bijeenkomsten zoals in Lunteren, Hilversum en Eindhoven heb ik altijd zeer gewaardeerd. Ik denk bovendien met veel plezier terug aan de maandelijkse culturele uitstapjes buiten het werk om met de collega's in Amsterdam en aan de kennismaking met Eindhoven.

Tot slot wil ik mijn ouders danken voor al de steun en liefde die zij mij hebben gegeven. Lieve mama en papa, ik ben jullie zeer dankbaar.

Curriculum Vitae

René de Jonge was born in Zaandam on December 7th, 1982. After he finished pre-university education at Het Zaanlands Lyceum in 2001, he started his studies in Mathematics and Statistics at the University of Amsterdam. He graduated in 2007 after finishing his master's thesis within the Stochastics group of the Mathematics department. In September 2007 he started working as a PhD candidate under the supervision of Harry van Zanten at the VU University in Amsterdam. He continued his work at the Eindhoven University of Technology from May 2009 onwards. The results of this work are presented in this dissertation.

ISBN 978-90-386-3192-9 Printed by Wöhrmann Print Service