Final Thesis

Bayesian adaptation using conditionally Gaussian priors

By Botond Szabó
Applied Mathematics
Eötvös Loránd Tudomány Egyetem
Department of Probability Theory and Statistics
2010

Supervisors:
Prof.dr. V. Prokaj
Prof.dr. J. H. van Zanten
Prof.dr. A.W. van der Vaart

Contents

Preface

This master thesis has been prepared at Vrije Universiteit Amsterdam between October, 2008 and June, 2009 and has been extended and corrected in the Eötvös Loránd Tudományegyetem between September, 2009 and June, 2010.

I would like to thank Harry van Zanten for his patient, his helpfulness and for the fruitful discussions we had during the time I had been working on the master project. It has been a pleasure to work with him.

Furthermore I would like to thank to Vilmos Prokaj for careful reading of the manuscript in the thesis, for his technical help on using LaTeX and finally for the fruitful conversations on the topic.

I am also thankful to Aad van der Vaart for the helpful answers to my questions.

1 Introduction

Consider the problem of estimating a probability density p_0 relative to a dominating measure μ on a sample space $(\mathcal{X}, \mathcal{A})$ based on an iid sample $X_1, \ldots, X_n$ from this density. Usually we use $\mathbb{R}$, $\mathbb{R}^d$, $[0,1]$ or $[0,1]^d$ as $\mathcal{X}$. In our case $\mathcal{X} = [0,1]$ and μ is the Lebesgue measure. Assume $p_0 > 0$ and $\log p_0$ is bounded. Gaussian processes can be adopted to construct prior distributions on infinite-dimensional statistical settings. For example they can be used in nonparametric density estimation as a prior distribution on a collection of probability densities (relative to the above mentioned μ). Let W be a Borel measurable, zero-mean Gaussian random element in $L^\infty(\mathcal{X})$, then define the prior distribution as

$$p_W(x) = \frac{e^{W_x}}{\int e^{W_x}\,\mu(\mathrm{d}x)}. \tag{1.1}$$

These functions are normalized and nonnegative, since the Gaussian process is exponentiated, so we get probability densities. We will refer to this prior distribution as Π.

According to Bayes' rule, the prior distribution and the iid sample yield a posterior distribution. We use the frequentist set up, which means that the data is sampled from a fixed "true distribution". The corresponding posterior distribution often contracts around the fixed true distribution, which is referred to as *posterior consistency*. We study the *rate of contraction* of the posterior distribution, which is the radius of the balls, where the posterior distribution puts most of its mass, as a function of n. The formal definition is that the posterior distribution has *rate of contraction at least* ε_n, if for a sufficiently large constant M

$$\Pi_n(p : d(p, p_0) \geq M\varepsilon_n | X_1, \ldots, X_n) \to 0 \quad \text{in probability as } n \to \infty,$$

where d is the Hellinger or total variation distance. The posterior distribution is the random measure given by

$$\Pi_n(B|X_1, X_2, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\,\mathrm{d}\Pi(p)}{\int \prod_{i=1}^n p(X_i)\,\mathrm{d}\Pi(p)}. \tag{1.2}$$

Since the posterior distribution puts all of its mass on the support of the prior, consistency can hold only if the parameter w_0, defining the true distribution of the data, belongs to this support.

The Hölder space of order $\alpha > 0$ is denoted by $C^\alpha[0,1]$. It is a collection of functions $f \in C[0,1]$ that have $\lfloor \alpha \rfloor$ continuous derivatives for $\lfloor \alpha \rfloor$ the biggest integer strictly smaller then α with the $\lfloor \alpha \rfloor$th derivative $f^{(\alpha)}$ being Lipschitz continuous of order β, for all $0 < \beta < \alpha - \lfloor \alpha \rfloor$. We will also refer to the elements of $C^\alpha[0,1]$ as α-regular functions.

In the simplest case, we would like to estimate α-regular density functions, where $\alpha > 0$ is a known parameter. By good choice of the prior distribution in form of (1.1) we can achieve the contraction rate $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$, for examples see the fourth chapter of [8].

Let $X_1, X_2, ..., X_n$ be an iid sample from density $p \in \mathcal{P}$, where $\mathcal{P}$ is some set of probability densities. The *minimax rate* for estimating p is up to a constant multiplier the square root of

$$\inf_{\hat{p}_n} \sup_{p \in \mathcal{P}} \mathbb{E}_p\big(d^2(p, \hat{p}_n)\big),$$

where the infimum is taken over all density function estimator $\hat{p}_n = \hat{p}_n(\cdot | X_1, ..., X_n)$ and d is a metric, for example, the Hellinger or total variation distance.

It is known that the minimax rate for estimating an α-regular density p_0 on $[0,1]^d$ is $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+d}}$, see the introduction part of [3]. Hence by good choice of the prior distribution we can construct estimators, with the help of the posterior distribution (1.2), which achieves the minimax rate.

We deal in this paper with the generalization of this problem, when the regularity of the density p_0 is unknown. In this case we use a set of models $\mathcal{P}^\alpha$ indexed by α, where $\alpha \in A \subset \mathbb{R}^+$ and the model $\mathcal{P}^\alpha$ can be used to estimate α-regular functions. Assume that for all $\alpha > \beta$ $(\alpha, \beta \in A)$ $\mathcal{P}^\alpha \subset \mathcal{P}^\beta$. We can put a prior on the α values and let the data decide the correct one through the corresponding posterior distribution. This Bayesian procedure fits within the framework of *adaptive estimation*, which focuses on constructing estimators that automatically choose a best model from a given set of models. An estimator on a given set of models is called *rate adaptive* if it attains the same rate of convergence that would have been attained if only the best model had been used. For example when the model $\mathcal{P}_\alpha$ contains the α-regular densities on $[0,1]^d$ and the set of models contains all of the β-regular densities, where $\beta > 0$, then an estimator is called rate adaptive, if it attains the rate $n^{-\alpha/(2\alpha+d)}$ whenever the true density is α-smooth, for any $\alpha > 0$. Our main goal is to construct a rate adaptive Bayesian estimation. For this we have

to give a prior distribution on the set of models. For simplicity this prior is discrete, and to achieve rate adaptiveness, it varies with the sample size.

This is an area of intensive research, see eg. [3], where a similar result to our main theorem was proved and applied to LogSpline models.

This paper is organized as follows. In Chapter 2 the notion of Reproducing Kernel Hilbert Space, the Riemann-Liouville process and the rescaled Gaussian process are discussed. Chapter 3 prepares the ground for our main theorem which is given in Chapter 4. The last chapter is for examples.

1.1 Notation

The ε-covering numbers of a metric space $(\mathcal{P}, d)$, denoted by $N(\varepsilon, \mathcal{P}, d)$, shows how many ε balls are needed to cover $\mathcal{P}$. The *ε-packing number* $D(\varepsilon, \mathcal{P}, d)$ of $\mathcal{P}$ is the supremum number of points in $\mathcal{P}$, such that the distance between every pair of points is at least ε. These two definitions are related by the inequalities

$$N(\varepsilon, \mathcal{P}, d) \leq D(\varepsilon, \mathcal{P}, d) \leq N(\frac{\varepsilon}{2}, \mathcal{P}, d).$$

Since we are only interested in rates of convergence, the additional constant $\frac{1}{2}$ is not important, hence we can replace the ε-covering number with the ε-packing number. The set of centers of ε-balls covering $\mathcal{P}$ is called an *ε-net*.

We use the Hellinger, total variation, L^2 distances, Kullback-Leibler divergence $K(\cdot, \cdot)$ and $V(\cdot, \cdot)$. For two probability densities p and q relative to the measure μ they are defined as

$$\begin{aligned}
h(p,q) &= \sqrt{\int |\sqrt{p} - \sqrt{q}|^2 \, \mathrm{d}\mu}, \\
\|p - q\|_1 &= \int |p - q| \, \mathrm{d}\mu \\
\|p - q\|_2 &= \sqrt{\int |p - q|^2 \, \mathrm{d}\mu}, \\
K(p,q) &= \int p \log \frac{p}{q} \, \mathrm{d}\mu, \\
V(p,q) &= \int \log^2 \left(\frac{p}{q}\right) p \, \mathrm{d}\mu.
\end{aligned}$$

Except of the L^2 distance they don't depend on the dominating measure μ.

For a random element W in $L^\infty(\mathcal{X})$ we define the *concentration function* of W around $w \in L^\infty(\mathcal{X})$

$$\varphi_w(\varepsilon) = -\log P(\|W - w\|_\infty < \varepsilon), \tag{1.3}$$

where $\|\cdot\|_\infty$ is the supremum norm.

Let $L(p)$ denote the n-sample likelihood ratio

$$L(p) = \prod_{i=1}^{n} \frac{p}{p_0}(X_i). \tag{1.4}$$

Furthermore, we use the notation $\mathbb{H}_1$ for the unit ball on the Hilbert space $\mathbb{H}$.

2 Reproducing Kernel Hilbert Space

In this section we give first the definition and some properties of the Reproducing Kernel Hilbert Space [1] and after that we give two examples [9], which will be applied in the fifth section.

2.1 RKHS in general

Let $W = (W_t : t \in T)$ be a zero-mean Gaussian stochastic process on the probability space $(\Omega, \mathcal{U}, P)$. In our paper we deal with the $T = [0, 1]$ case. The finite-dimensional distribution of such a process is determined by the covariance function $K : T \times T \mapsto \mathbb{R}$, defined by

$$K(s, t) = \mathbb{E}W_s W_t.$$

For a Gaussian process W we isometrically assign a function space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called Reproducing Kernel Hilbert Space (RKHS). The RKHS is important for us, because it determines the support and the "geometry" of the Gaussian measure, given by the Gaussian process W_t. Next, we give the classical definition and after that the definition for centered stochastic process of the RKHS. Let $\mathbb{H}$ be a Hilbert space of functions defined on T. A function

$$K : T \times T \to \mathbb{C}$$

is a *reproducing kernel* of the Hilbert space $\mathbb{H}$ if and only if

$$\begin{aligned} \forall t \in T, \quad & K(., t) \in \mathbb{H} \quad \text{and} \\ \forall t \in T, \forall \varphi \in \mathbb{H} \quad & \langle \varphi, K(., t) \rangle = \varphi(t), \end{aligned}$$

where the last condition is called "the reproducing property". A Hilbert space of complex-valued functions which possesses a reproducing kernel is called a *Reproducing Kernel Hilbert Space*. Next we give two examples for RKHS.

Example 2.1. Let $\mathbb{H}$ be a finite dimensional complex vector space of functions with basis $(f_1, f_2, ..., f_n)$ and define the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ by the numbers

$$\langle f_i, f_j \rangle_{\mathbb{H}} := g_{i,j},$$

where $G = (g_{i,j})$ is an arbitrary hermitian and positive definite matrix. Then $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ is a Hilbert space. Let $G^{-1} = (g^{ij})$ denote the inverse of the matrix

G. Then the bivariate function

$$K(x,y) = \sum_{i,j=1}^{n} \bar{g}^{i,j} f_i(x) \bar{f}_j(y)$$

is the reproducing kernel of the Hilbert space $\mathbb{H}$, hence $\mathbb{H}$ is an RKHS.

Example 2.2. Let $T = [0,1]$ and $\mathbb{H} = \{f : f(0) = 0,\ f \text{ is absolute continuous},\ f' \in L^2[0,1]\}$. $\mathbb{H}$ is a Hilbert space with the inner product

$$\langle \varphi, \psi \rangle = \int_{[0,1]} \varphi' \bar{\psi}' d\lambda.$$

Then $\mathbb{H}$ has reproducing kernel $K(x,y) = \min\{x,y\}$, hence $\mathbb{H}$ is an RKHS.

The next theorem gives an alternative characterization of a Reproducing Kernel Hilbert Space.

Theorem 2.3 (Theorem 1 of [1]). *A Hilbert space $\mathbb{H}$ of complex valued functions on T has reproducing kernel if and only if all the evaluation functionals δ_t, $t \in T$, are continuous on $\mathbb{H}$.*

A function $K : T \times T \to \mathbb{C}$ is called a *positive type function* (or *positive definite function*) if

$$\forall n \geq 1,\ \forall (a_1, ..., a_n) \in \mathbb{C}^n,\ \forall (x_1, ..., x_n) \in T^n,$$
$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \bar{a}_j K(x_i, x_j) \in \mathbb{R}^+,$$

where $\mathbb{R}^+$ denotes the set of nonnegative real numbers. The linkage between reproducing kernels and positive type functions:

Lemma 2.4 (Lemma 2 of [1]). *Any reproducing kernel is a positive type function.*

Theorem 2.5 (Moore-Aronszajn Theorem (Theorem 3 of [1])). *Let K be a positive type function on $T \times T$. There exists only one Hilbert space $\mathbb{H}_K$ of functions on T with K as a reproducing kernel.*

Another property of the RKHS is that if $\mathbb{H}_1$ and $\mathbb{H}_2$ are two RKHS's on T_1 and T_2, then $\mathbb{H}_1 \otimes \mathbb{H}_2$ is a RKHS on $T_1 \times T_2$.

Next we give the definition of the Hilbert space generated by a centered process. Let's take an arbitrary square integrable centered process $\{X_t : t \in$

$T\}$ and denote $\mathcal{L}(X)$ the linear space spanned by the variables X_t, $t \in T$. Let $\bar{\mathcal{L}}(X)$ be the closure in $L^2(\Omega, \mathcal{A}, P)$ of $\mathcal{L}(X)$. $\bar{\mathcal{L}}(X)$ is equipped with the inner product induced by that of $L^2(\Omega, \mathcal{A}, P)$

$$\langle U, V \rangle_{L^2(\Omega,\mathcal{A},P)} = E(UV).$$

We call the space $\bar{\mathcal{L}}(X)$ the *Hilbert space generated by the centered process* X. Then the RKHS generated by the centered process X is the following:

Theorem 2.6 (Loéve's theorem (Theorem 35 of [1])). *The Hilbert space* $\bar{\mathcal{L}}(X)$ *generated by the centered process* $\{X_t, t \in T\}$ *with covariance function* K *is congruent to the RKHS* $\mathbb{H}_K$

2.2 The Riemann-Liouville process

For $\alpha > 0$ and W a standard Brownian motion, the Riemann-Liouville process with parameter $\alpha > 0$, $RL(\alpha)$ for short, is defined as

$$R_t^\alpha = \int_0^t (t-s)^{\alpha-\frac{1}{2}} \, \mathrm{d}W_s, \qquad t \in [0,1].$$

The process R_t^α is zero-mean Gaussian process, with continuous sample paths. Let's define the *α-order Riemann-Liouville fractional integral of f* for a measurable function f as

$$I_{0+}^\alpha f(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s) \, \mathrm{d}s.$$

Hence we can also view R_t^α as a multiple of the $(\alpha - \frac{1}{2})$-fractional integral of the "derivative of the Brownian motion $(\mathrm{d}W_t)$". So for $\alpha \geq \frac{1}{2}$ the Riemann-Liouville process is equal to $\Gamma(\alpha + \frac{1}{2}) I_{0+}^{\alpha-\frac{1}{2}} W$. The transformation I_{0+}^α maps β-regular functions into $\alpha+\beta$ regular functions, see [7]. Since Brownian motion is $\frac{1}{2}$-regular, the $RL(\alpha)$ process is a good model for α-regular densities.

Theorem 2.7 (Theorem 4.2 of [8]). *The RKHS of the Riemann-Liouville process with parameter* $\alpha > 0$ *is* $\mathbb{H} = I_{0+}^{\alpha+\frac{1}{2}}(L_2[0,1])$ *and the RKHS-norm is given by*

$$\|I_{0+}^{\alpha+\frac{1}{2}} f\|_{\mathbb{H}} = \frac{\|f\|_2}{\Gamma(\alpha + \frac{1}{2})}.$$

We can use the $RL(\alpha)$ process for approximating $C^\alpha[0,1]$ functions which have $[\alpha]$ vanishing derivatives at zero, where $[\alpha]$ means the floor of α. By

adding random elements in polynomial form, the trajectories of the process won't have $[\alpha]$ vanishing derivatives necessarily so we define the process

$$X_t^\alpha = \sum_{i=0}^{[\alpha]} \frac{t^i}{i!} Z_i + \frac{1}{\Gamma(\alpha + \frac{1}{2})} R_t^\alpha,$$

where $Z_0, \ldots, Z_{[\alpha]}, R_t^\alpha$ are independent, Z_i's are standard normal variables and R_t^α is an $RL(\alpha)$ process.

Theorem 2.8 (Theorem 4.3 of [8]). *For all $\alpha > 0$ the support of the process X_t^α is the whole space $C[0,1]$. For any $w \in C^\alpha[0,1]$, the concentration function of X_t^α satisfies $\varphi_w(\varepsilon) = O(\varepsilon^{-\frac{1}{\alpha}})$ as $\varepsilon \searrow 0$.*

By Theorem 2.8 the concentration function of X_t^α satisfies that $\varphi_w(\varepsilon) = O(\varepsilon^{-\frac{1}{\alpha}})$ so the inequality

$$\varphi_w(\varepsilon_{n,\alpha}) \leq n\varepsilon_{n,\alpha}^2 \tag{2.1}$$

holds with $\varepsilon_{n,\alpha} = cn^{-\frac{\alpha}{2\alpha+1}}$ which is exactly the minimax rate of estimating α-regular functions ($w_0 \in C^\alpha[0,1]$).

The RKHS of the process X_t^α, where $k = \alpha - \frac{1}{2} \in \mathbb{N}$, is the set of functions $f : [0,1] \mapsto \mathbb{R}$ that are k-times differentiable, their kth-derivatives are absolutely continuous and square integrable, equipped with the inner product

$$\langle f, g \rangle_{\mathbb{H}_\alpha} = \sum_{i=0}^{k} f^{(i)}(0) g^{(i)}(0) + \int_0^1 f^{(k+1)}(s) g^{(k+1)}(s)\, \mathrm{d}s,$$

which defines the norm

$$\|f\|_{\mathbb{H}_\alpha} = \sum_{i=0}^{k} \left(f^{(i)}(0)\right)^2 + \int_0^1 \left(f^{(k+1)}(s)\right)^2 \mathrm{d}s, \tag{2.2}$$

see Section 10 of [9].

2.3 Rescaled multiple integrated Brownian motion

In this subsection we use the results of [10].

Our first goal is to give a prior distribution on α-regular functions for all $\alpha \in (0, k+1]$, for a k large enough. The main idea is to rescale a smooth enough process, with rescaling constants depending on the sample size. Given a fixed Gaussian process $(W_t : t \geq 0)$ indexed by the positive time axis and scaling constants $c_n > 0$ we use the rescaled sample path

$$t \mapsto W_{t/c_n}, \qquad t \in [0,1]$$

as a prior model for a given function $w_0 : [0,1] \mapsto \mathbb{R}$. By rescaling we change the appearance of the process by stretching or shrinking its sample paths. There are two cases, if $c_n \to \infty$ then we stretch the sample path $t \mapsto W_t$ on the short interval $[0, 1/c_n]$ to $[0,1]$. Typically this has the effect of smoothing the sample paths. In the second case scaling factors $c_n \to 0$ use the sample path on the long interval $[1, 1/c_n]$ and shrink this to interval $[0,1]$. This usually makes the prior rougher.

We will use self similar processes as smooth prior processes in this section. If W is the k-fold integrated Brownian motion (plus an independent polynomial part) then using W as prior on the $(k+1/2)$-regular functions yields an optimal convergence rates for the posterior, see [8]. We will see that for any $\alpha \in (0, k+1]$ there exist scaling sequences c_n such that choosing the prior based on the rescaled process W_{t/c_n}, the posterior will give the optimal contraction rate whenever the true density is α-regular. If $\alpha \in (0, k+1/2)$ then we have $c_n \to 0$, else if $\alpha \in (k+1/2, k+1]$ then we have $c_n \to \infty$. Finally for $\alpha = k+1/2$ we don't need rescaling.

Consider an α-order, self-similar, zero-mean Gaussian process $(W_t : t \geq 0)$, which means that the processes $(c^\alpha W_{t/c} : t \geq 0)$ and $(W_t : t \geq 0)$ are equal in distribution for every $c > 0$. So the rescaled process $(W_{t/c} : t \geq 0)$ has the same law as the process $(c^{-\alpha} W_t : t \geq 0)$. Hence the RKHS and small ball probabilities are equal in the time axis and in the vertical axis rescaled processes. We use the notation $\mathbb{H}_W$ for the RKHS of the W_t process and $\varphi_0(\varepsilon; W) = -\log P(\|W\| \leq \varepsilon)$ for the exponent in the small ball probability. The RKHS of the process $W^c = (W_{t/c} : 0 \leq t \leq 1)$ is the set of functions $\mathbb{H}_W$ equipped with the norm $\|h\|_{W^c} = c^\alpha \|h\|_W$. The centered small ball exponent of W^c satisfies $-\log P(\|W^c\| \leq \varepsilon) = \varphi_0(c^\alpha \varepsilon; W)$.

We deal with the k-fold integrated Brownian motion, which has k derivatives at 0 equal to 0, hence its RKHS satisfies similar condition. So a better choice of the prior is to add an independent polynomial part to the process. We consider the process

$$V_t^{c,a} = (I_{0+}^k B)_{t/c} + \frac{1}{\sqrt{a}} \sum_{i=0}^{k} Z_i \frac{t^i}{i!} \tag{2.3}$$

for scaling factors $c, a > 0$, B a standard Brownian motion and independent standard normal variables $Z_0, Z_1, \ldots, Z_k$, independent of B.

Theorem 2.6 of [10] gives a centered small deviation bound for the process

$V^{c,a}$, and describes the approximation of smooth functions by elements of its RKHS $\mathbb{H}^{c,a}$.

Theorem 2.9. *The process $V_t^{c,a}$ given in (2.3) satisfies for $\varepsilon > 0$ small enough,*

$$-\log P(\sup_{0\le t\le 1}|V_t^{c,a}| \le 2\varepsilon) \lesssim \big(\frac{1}{c^{k+1/2}\varepsilon}\big)^{\frac{1}{k+1/2}} + k\log\frac{1}{\sqrt{a}\varepsilon}.$$

Moreover, for $w \in C^\beta[0,1]$ and $\beta \le k+1$,

$$\inf\{\|h\|^2_{\mathbb{H}^{c,a}} : \|h-w\|_\infty \le \varepsilon\} \lesssim c^{2k+1}\big(\frac{1}{\varepsilon}\big)^{(2k+2-2\beta)/\beta} + a\big(\frac{1}{\varepsilon}\big)^{((2k-2\beta)/\beta)\vee 0}.$$

where the notation $\lesssim$ is used for "smaller then or equal to a universal constant times" and P is the probability defined by the process $V_t^{c,a}$.

In the proof of Theorem 2.9 ([10]) it was shown that the RKHS of the process $V_t^{c,a}$, which is the Sobolev space $H^{k+1}[0,1]$ of functions $h : [0,1] \to \mathbb{R}$ that are k-times continuously differentiable with absolutely continuous kth derivative that is the integral of function $h^{(k+1)} \in L_2[0,1]$, equipped with the norm with square

$$\|h\|^2_{\mathbb{H}^{c,a}} = c^{2k+1}\|h^{(k+1)}\|_2^2 + a\sum_{i=0}^{k} h^{(i)}(0)^2.$$

Theorem 3.2 from the same paper gives for all $\alpha \in (0, k+1]$, for an arbitrary k, a sequence of processes, which can be used in the prior distribution, such that the posterior distribution achieves the minimax rate.

Theorem 2.10. *For $\alpha > 0$ and $k \in \mathbb{N}$, let V_t^n be the modified k-fold integrated Brownian motion defined in (2.3) with the scaling constant c replaced by*

$$c_{n,\alpha} = n^{\frac{\alpha-(k+\frac{1}{2})}{(k+\frac{1}{2})(1+2\alpha)}}$$

and a replaced by a sequence of $a_{n,\alpha}$ satisfying

$$a_{n,\alpha} \le n^{\frac{1+2\alpha-2k}{1+2\alpha}}.$$

Define the prior Π_n^α as in (1.1), with the Gaussian process V^n . Then if $\log p_0 \in C^\alpha[0,1]$ and $\alpha \le k+1$, we have

$$\mathbb{E}_0\Pi^\alpha(p : h(p,p_0) > M\varepsilon_{n,\alpha}|X_1,\dots,X_n) \to 0$$

for all M large enough, where $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{1+2\alpha}}$ and h is the Hellinger distance.

So we can use the rescaled processes in (1.1) to estimate α-regular functions for all $\alpha > 0$ and it achieves the minimax rate in the Bayesian estimation procedure.

Remark 2.11. In the proof of Theorem 2.10 it was shown that the concentration function $\varphi_w(\varepsilon)$ for $w \in C^\alpha[0,1]$ satisfies

$$\varphi_w(\varepsilon_{n,\alpha}) \leq n\varepsilon_{n,\alpha}^2,$$

where $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{1+2\alpha}}$.

3 Preliminary results

Recall from the first chapter that $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$ is the minimax rate for estimating α-regular functions on the interval [0,1]. By good choice of the Gaussian process W^α we get a prior distribution Π^α according to (1.1), such that the posterior contraction rate achieves the minimax rate. Our main goal is to give a minimax estimation for an α-regular function, where $\alpha > 0$ is an unknown parameter.

First we give a hierarchical prior on the densities. Let λ_n be a discrete probability distribution on a finite subset of $\mathbb{R}^+$. We construct a new, hierarchical prior Π_n such that it first chooses α according to λ_n and next p according to Π^α for the chosen α:

$$\Pi_n = \sum_{\alpha \in B_n} \lambda_n(\alpha) \Pi^\alpha.$$

The corresponding posterior distribution is given by

$$\begin{aligned} \Pi_n(B|X_1, \dots, X_n) &= \frac{\int_B \prod_{i=1}^n p(X_i) \, \mathrm{d}\Pi_n(p)}{\int \prod_{i=1}^n p(X_i) \, \mathrm{d}\Pi_n(p)} \\ &= \frac{\int \int_B \prod_{i=1}^n p(X_i) \, \mathrm{d}\Pi^\alpha(p) \, \lambda_n(\mathrm{d}\alpha)}{\int \int \prod_{i=1}^n p(X_i) \, \mathrm{d}\Pi^\alpha(p) \, \lambda_n(\mathrm{d}\alpha)}. \end{aligned} \tag{3.1}$$

In section 2.2 and section 2.3 different choices of W^α Gaussian processes were given such that the concentration function φ_w^α defined by W^α satisfies

$$\varphi_w(\varepsilon_{n,\alpha}) \leq n\varepsilon_{n,\alpha}^2$$

for every $w \in C^\alpha[0,1]$, where $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$.

In the following, four conditions are given which are applied in the next theorem.

1. For all $\alpha > 0$ there exists a set $\mathcal{W}^\alpha \subset L^\infty$ and a sequence of positive numbers $\varepsilon_{n,\alpha} \searrow 0$, such that $n\varepsilon_{n,\alpha}^2 \nearrow \infty$ and for every $w \in \mathcal{W}^\alpha$

$$\varphi_w(\varepsilon_{n,\alpha}) \leq n\varepsilon_{n,\alpha}^2 \quad \text{if n is large enough,} \tag{3.2}$$

   where $\varphi_w(\varepsilon)$ is the concentration function of W^α defined in (1.3). Furthermore assume that for all $0 < \alpha < \beta$

$$\frac{\varepsilon_{n,\alpha}}{\varepsilon_{n,\beta}} \to \infty. \tag{3.3}$$

2. For all $0 < \alpha < \beta$, where $\alpha, \beta \in \mathbb{N}$ and for all $c_1, c_2 > 0$

$$\frac{\lambda_n(\alpha)}{\lambda_n(\beta)} e^{c_1 n\varepsilon_{n,\beta}^2} \Pi^\alpha(c_2\varepsilon_{n,\beta} < h(p, p_o) \leq c_2\varepsilon_{n,\alpha}) \to 0 \quad \text{as } n \to \infty.$$

3. For all $\beta > 0$ there exists a constant $\delta > 0$ not depending on n such that

$$\lambda_n(\beta) > \delta e^{-n\varepsilon_{n,\beta}^2} \sum_{\alpha \geq \beta} \lambda_n(\alpha).$$

4. For all $0 < \alpha < \beta$

$$\mathbb{H}_1^\beta \subseteq K\mathbb{H}_1^\alpha,$$

where $\mathbb{H}_1^\alpha$ is the unit ball of $\mathbb{H}^\alpha$.

Conditions 2 and 3 are satisfied if we choose the prior distribution λ_n as

$$\lambda_n(\alpha) \sim c_\alpha e^{-n\varepsilon_{n,\alpha}^2} \tag{3.4}$$

with an arbitrary convergent sequence of positive numbers c_α, regardless of the prior Π^α.

Theorem 3.1. *Assume that Conditions 1, 2, 3 and 4 hold. If* $\log p_0 \in \mathcal{W}^\beta$ *for some* $\beta \in \mathbb{N}$, *then*

$$\mathbb{E}_0 \Pi_n(p : h(p, p_0) > M\varepsilon_{n,\beta} | X_1, \ldots, X_n) \to 0$$

for M *large enough.*

From this theorem it follows that by good choice of the Gaussian processes W^α for all $\alpha \in \mathbb{N}$ and with the special choice of λ_n, given in (3.4), we can construct an estimator (for example):

$$\hat{p}_0 = \arg\max_q \Pi_n(p : h(p, q) < M\varepsilon_{n,\beta} | X_1, \ldots, X_n),$$

which achieves the minimax rate, whenever we estimate β-regular functions, with an unknown $\beta \in \mathbb{N}$ parameter. It is true, since the intersection of the balls with radius $M\varepsilon_{n,\beta}$ around p_0 and $\hat{p}_0$ is not empty. In the next chapter we will give the generalization of Theorem 3.1, where the same statement holds for $\beta \in \mathbb{R}^+$, so not only for natural numbers.

3.1 The fundamental theorem

Theorem 2.1 of [2] is a very important theorem in this topic and we will refer to it as the fundamental theorem.

Theorem 3.2 (The fundamental theorem). *Let Π_n be the sequence of prior probability measures supported on some set of probability measures $\mathcal{P}$. Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$, we have*

$$\log D(\varepsilon_n, \mathcal{P}_n, d) \leq n\varepsilon_n^2; \tag{3.5}$$

$$\Pi_n(\mathcal{P}_n^c) \leq e^{-n\varepsilon_n^2(C+4)}; \tag{3.6}$$

$$\Pi_n\big(p : K(p_0, p) \leq \varepsilon_n^2, V(p_0, p) \leq \varepsilon_n^2\big) \geq e^{-Cn\varepsilon_n^2}. \tag{3.7}$$

Then for sufficiently large M, we have that

$$\Pi_n(p : h(p, p_0) \geq M\varepsilon_n | X_1, \ldots, X_n) \to 0, \quad \text{in } (P_0)^n\text{-probability},$$

where h is the Hellinger metric.

We give some explanation to the conditions. The first and the third conditions of the theorem are the essential ones. Condition (3.5) says that $\mathcal{P}_n$ is not too big. It is true for all $\varepsilon_n' > \varepsilon_n$ as soon as it is true for ε_n, so it is a restriction on the minimal value of ε_n. When the densities are uniformly bounded, we can use L_2-distance also, which is bounded above by the multiple of the Hellinger distance. If the densities are uniformly bounded and uniformly bounded away from zero, then the Hellinger, total variation and L_2-distance are equivalent. Condition (3.6) says that $\mathcal{P}_n$ is a good approximation of the prior. Condition (3.6), combined with condition (3.5), can be interpreted as saying that a part of $\mathcal{P}$ that barely receives prior mass $\mathcal{P}_n^c$ is not small. Condition (3.7) is the other main determinant of the posterior rate given by the theorem. It says that the prior measure puts some sufficient amount of mass in the small neighborhood of the true density p_0. This condition is also satisfied for all $\varepsilon_n' > \varepsilon_n$, if it is satisfied for ε_n, hence it is another restriction on a minimal value of ε_n.

The assertion of the theorem says that the posterior mass outside of a ball of radius proportional to ε_n is approximately zero in-probability. The in-probability statement can be improved to an almost sure assertion under stronger conditions.

In the proof of Theorem 3.2 the existence of tests $\varphi_n : \mathcal{X}^n \mapsto [0, 1]$ were applied. We use the word test as it is used in randomized hypothesis testing

for the function giving the rejection probability. These φ_n's are tests of p_0 versus $\{p : d(p, p_0) > \varepsilon\}$ the complement of the ball of radius ε around p_0. For every pair p_0 and p_1 in the model $\mathcal{P}$ there exist tests φ_n such that, for some universal constant K

$$\begin{aligned} \mathbb{E}_0^n \varphi_n &\leq \exp(-Knd^2(p_0, p_1)), \\ \sup_{p:\, d(p,p_1)<d(p_0,p_1)/2} \mathbb{E}_p^n(1-\varphi_n) &\leq \exp(-Knd^2(p_0, p_1)), \end{aligned}$$

where d could be both the Hellinger and the total variation distance.

With the help of this tests it was shown in the proof of the fundamental theorem ([2]) that:

Lemma 3.3. *Assume that the conditions of the fundamental theorem hold. Then there exist a sequence of test φ_n and a set of events $Q_n \subset \mathcal{X}^n$, such that*

$$\begin{aligned} \mathbb{E}_0 \Pi_n(p : h(p, p_0) > M\varepsilon_n | X_1, \ldots, X_n)\varphi_n &\leq 2e^{-Kn\varepsilon_n^2}, \\ \mathbb{E}_0 \Pi(p : h(p, p_0) > M\varepsilon_n | X_1, \ldots, X_n)(1-\varphi_n) I_{Q_n} &\leq 2e^{-2n\varepsilon_n^2}, \\ \mathbb{E}_0 I_{Q_n^c} &\leq \frac{1}{n\varepsilon_n^2}. \end{aligned}$$

In the proof of the fundamental theorem Lemma 8.1 of [2] was applied also which we will need later.

Lemma 3.4. *For every $\varepsilon > 0$ and probability measure Π on the set*

$$\Big\{ p :\ K(p_0, p) \leq \varepsilon^2,\, V(p_0, p) \leq \varepsilon^2 \Big\}$$

we have, for every $C > 0$,

$$P_0\Big(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, \mathrm{d}\Pi(p) \leq \exp(-(C+1)n\varepsilon^2) \Big) \leq \frac{1}{C^2 n\varepsilon^2}.$$

In the proof of our main theorem we will use a version of Borell's inequality given in [8]. Let W be a Borel measurable, zero-mean Gaussian random element in $L^\infty(\mathcal{X})$. Recall, that for the concentration function φ_0 of W

$$e^{-\varphi_0(\varepsilon)} = P(W \in \varepsilon \mathbb{B}_1),$$

where $\mathbb{B}_1$ is the unit ball of $(L^\infty(\mathcal{X}), \|\cdot\|_\infty)$.

Theorem 3.1 (Theorem 5.1 of [8]). *For any $\varepsilon > 0$ and $M \geq 0$,*

$$P(W \in \varepsilon \mathbb{B}_1 + M\mathbb{H}_1) \geq \Phi(\Phi^{-1}(e^{-\varphi_0(\varepsilon)}) + M),$$

where Φ is the distribution function of the standard normal distribution and $\mathbb{H}_1$ is the unit ball of the RKHS of W.

3.2 Determination of constants

In the proofs of the following theorems the constant multipliers are hidden, hence we copy their proofs and extend them with the computation of a sufficiently large constants. First we give Lemma 8 of [4].

Lemma 3.5. *For every pair of probability densities p and q we have*

$$\mathbb{E}_p \log p/q \le h^2(p,q)\big(2 + \log \|p/q\|_\infty\big), \tag{3.8}$$

$$\mathbb{E}_p \log^2(p/q) \le h^2(p,q)\big(3 + \log \|p/q\|_\infty\big)^2. \tag{3.9}$$

Proof. In both inequality there is nothing to prove if $p = q$, that is $h(p,q) = 0$. So we can and do assume that $h(p,q) > 0$.

Let $\tilde{r} : (0,\infty) \to \mathbb{R}$ be the function $\tilde{r}(x) = 2\frac{x-1-\log(x)}{(x-1)^2}$ if $x \neq 1$ and $\tilde{r}(1) = 1$. Then $\tilde{r}$ is continuous. Put also $r(x) = \tilde{r}(\sqrt{x})$. With this notation $\log x = 2(\sqrt{x}-1) - r(x)(\sqrt{x}-1)^2$ or equivalently

$$\log \frac{1}{x} = 2(1-\sqrt{x}) + r(x)(\sqrt{x}-1)^2. \tag{3.10}$$

The following representation of $\tilde{r}$ is useful:

$$\frac{1}{2}\tilde{r}(x) = \int_0^1 \frac{y}{(x-1)y+1}dy = \int_0^1 \frac{y}{xy+1-y}dy.$$

From this, it is easily seen that

(i) $\tilde{r}$ and hence also r is non-negative and decreasing.

(ii) $\tilde{r}$ is concave, since $\tilde{r}' < 0$ for $x > 0$. This can be seen from

$$\frac{1}{2}\tilde{r}'(x) = -\int_0^1 \frac{y^2}{(yx+(1-y))^2}dy.$$

Another way to express the derivative $\tilde{r}'(x)$

$$\begin{aligned}\frac{1}{2}\tilde{r}'(x) &= \left(\frac{x-1-\log(x)}{(x-1)^2}\right)' \\ &= \frac{1-\frac{1}{x}}{(x-1)^2} - 2\frac{1}{x-1}\frac{1}{2}\tilde{r}(x) \\ &= -\frac{1}{x} + \frac{1}{1-x}\left(\tilde{r}(x)-1\right).\end{aligned}$$

If $x \in (0,1]$ then $\tilde{r}(x) \ge 1$, hence $\frac{1}{2}\tilde{r}'(x) + 1/x > 0$ and therefore $2\log(x) + \tilde{r}(x)$ is increasing. Hence on $(0,1)$ we have the following estimate also

$$\begin{aligned}&\tilde{r}(x) \le 1 - 2\log(x), \\ &r(x) = \tilde{r}(x^{1/2}) \le 1 - 2\log x^{1/2} = 1 + \log\frac{1}{x}.\end{aligned}$$

(iii) For $b \geq 1$ the function $x^b \tilde{r}(x)$ is increasing. Hence $x^b r(x)$ is also increasing for $b \geq 2$.

Remark 3.6. Using the concavity of $\tilde{r}(x)+2\log(x)$ and the fact that $\tilde{r}'(1) = 4/3$ we also obtain the next sharper estimate on $(0,1)$

$$-2 + 3x + \log\left(\frac{1}{x}\right) \leq \tilde{r}(x) \leq -\frac{1}{3} + \frac{4}{3}x + \log\left(\frac{1}{x}\right).$$

Using (3.10) with $X = q/p$ and integrating with respect to p we obtain that

$$\mathbb{E}_p \log(p/q) = \mathbb{E}_p \log 1/X = h^2(p,q) + \mathbb{E}_p\Big(r(X)\big(\sqrt{X}-1\big)^2\Big).$$

To estimate the last term on the right take $\varepsilon \in (0,1)$ and split the integral according to $X > \varepsilon$ or $X \leq \varepsilon$, and use that

$$r(X)(\sqrt{X}-1)^2 \chi_{(X>\varepsilon)} \leq r(\varepsilon)(\sqrt{X}-1)^2$$

since r is decreasing by Property (ii).

For the other part take $b > 2$ then using Property (i) of r

$$r(X)(\sqrt{X}-1)^2 \chi_{(X\leq\varepsilon)} \leq \frac{X^b r(X^b)}{X^b}(\sqrt{X}-1)^2 \chi_{(X\leq\varepsilon)} \leq \varepsilon^b r(\varepsilon) X^{-b}.$$

Taking expectation we obtain that

$$\begin{aligned}\mathbb{E}_p \log(p/q) &\leq h^2(p,q)\Big[1 + r(\varepsilon)\Big(1 + \frac{\varepsilon^b \mathbb{E}_p (p/q)^b}{h^2(p,q)}\Big)\Big] \\ &\leq h^2(p,q)\Big[1 + \Big(1 + \log\frac{1}{\varepsilon}\Big)\Big(1 + \frac{\varepsilon^b \mathbb{E}_p (p/q)^b}{h^2(p,q)}\Big)\Big].\end{aligned}$$

To finish the proof of the first part of the statement choose $\varepsilon_b < 1$ depending on b such that $\varepsilon_b^b \mathbb{E}_p (p/q)^b \to 0$ as $b \to \infty$ and $1/\varepsilon_b \to \|p/q\|_\infty > 1$. The inequality $\|p/q\|_\infty > 1$ follows from the facts that $p \neq q$ and both p and q are probability densities. With this choice we obtain

$$\mathbb{E}_p \log(p/q) \leq h^2(p,q)(2 + \log\|p/q\|_\infty)$$

This proves (3.8).

To prove the second inequality, note that

$$\Big(\log\frac{1}{x} + 2(\sqrt{x}-1)\Big)^2 = r^2(x)(\sqrt{x}-1)^4.$$

Substituting $X = q/p$ and integrating on the set $\{X \le c\}$ with $c \ge 4$ we get by applying the Cauchy Schwartz inequality

$$\mathbb{E}_p(\chi_{(X\le c)}\log^2 X) + 4H^2 - 4\mathbb{E}_p(\chi_{(X\le c)}\log^2 X)^{1/2}H \le \mathbb{E}_p\big(r^2(X)(\sqrt{X}-1)^4\chi_{(X\le c)}\big), \quad (3.11)$$

where $H^2 = \mathbb{E}_p\big(\chi_{(X\le c)}(\sqrt{X}-1)^2\big)$. Denote the right hand side of the inequality by A and put $z = \mathbb{E}_p(\chi_{(X\le c)}\log^2 X)^{1/2}$. Then (3.11) can be written as

$$z^2 - 4zH + 4H^2 - A \le 0$$

This gives that

$$z \le \frac{4H + \sqrt{16H^2 - 16H^2 + 4A}}{2} = 2H + \sqrt{A} \quad (3.12)$$

A can be bounded as above, by taking $\varepsilon \in (0,1)$ and $b \ge 2$ and splitting the integral according to $X \ge \varepsilon$ or $X < \varepsilon$. Here we use that $c \ge 4$ hence $(\sqrt{X}-1)^4\chi_{(X\le c)} \le (\sqrt{X}-1)^2(\sqrt{c}-1)^2\chi_{(X\le c)}$. Thus

$$\begin{aligned} A &\le \mathbb{E}_p r^2(X)(\sqrt{X}-1)^4\chi_{(X\le c)} \\ &\le (\sqrt{c}-1)^2H^2r^2(\varepsilon)\Big(1 + \frac{\varepsilon^{2b}\mathbb{E}_p(p/q)^{2b}}{H^2}\Big) \\ &\le (\sqrt{c}-1)^2H^2\big(1+\log(1/\varepsilon)\big)^2\Big(1 + \frac{\varepsilon^{2b}\mathbb{E}_p(p/q)^{2b}}{H^2}\Big) \end{aligned}$$

We choose again a sequence ε_b such that $1/\varepsilon_b \to \|p/q\|_\infty$ and $\varepsilon_b^{2b}\mathbb{E}_p(p/q)^{2b} \to 0$ when $b \to \infty$. This gives

$$A \le (\sqrt{c}-1)^2H^2(1 + \log\|p/q\|_\infty)$$

Plugging this into (3.12) with $c = 4$ yields

$$\begin{aligned} \mathbb{E}_p(\chi_{(q\le 4p)}\log^2 p/q) &\le H^2\left(2 + (1 + \log\|p/q\|_\infty)\right)^2 \\ &\le H^2\left(3 + \log\|p/q\|_\infty\right)^2. \end{aligned}$$

To finish the proof we use that $\sup_{x\ge 4}\log^2 x/(\sqrt{x}-1)^2 \le \log^2 4 \le 2$. This gives that

$$\mathbb{E}_p\chi_{(X\ge 4)}\log^2 X \le 2\mathbb{E}_p\chi_{(X\ge 4)}(\sqrt{X}-1)^2.$$

Since $H^2 + \mathbb{E}_p\chi_{(X\ge 4)}(\sqrt{X}-1)^2 = h^2(p,q)$ we obtain that

$$\begin{aligned} \mathbb{E}_p\log^2 p/q &\le H^2\left(3 + \log\|p/q\|_\infty\right)^2 + 2\mathbb{E}_p\big(\chi_{(X\ge 4)}(\sqrt{X}-1)^2\big) \\ &\le h^2(p,q)\left(3 + \log\|p/q\|_\infty\right)^2. \end{aligned}$$

□

In the next lemma we use the notation $p_w = ce^w$, where c is the normalizing constant.

Lemma 3.7 (Lemma 3.1 of [8]). *For bounded measurable functions $v, w : \mathcal{X} \mapsto \mathbb{R}$, we have the following:*

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

where the notation $\lesssim$ is used for "smaller then or equal to a universal constant times".

Next we prove a version of the previous lemma.

Lemma 3.8. *For any measurable functions $v, w : \mathcal{X} \mapsto \mathbb{R}$, for which $\|v - w\|_\infty < 1/3 \log(16/9)$, we have the following:*

$$\begin{aligned} K(p_v, p_w) &\le 4\|v - w\|_\infty^2; \\ V(p_v, p_w) &\le 16\|v - w\|_\infty^2. \end{aligned}$$

Proof. We use that $\log \|p/q\|_\infty = \|\log(p/q)\|_\infty$, (3.8) and the following result:

$$\Big\| \log \frac{p_v}{p_w} \Big\|_\infty \le 2\|v - w\|_\infty.$$

From these and Lemma 3.5 we get that

$$\begin{aligned} \mathbb{E}_p \log \frac{p_v}{p_w} &\le h^2(p_v, p_w)\Big(2 + \log \Big\| \frac{p_v}{p_w} \Big\|_\infty\Big) \\ &\le \|v - w\|_\infty^2 e^{\|v-w\|_\infty}(2 + 2\|v - w\|_\infty) \end{aligned} \tag{3.13}$$

It is easy to check that $e^x \le \sqrt{2}$ if $x \le 0.5 \log 2$ and $(1+x) < \sqrt{2}$ if $x < \sqrt{2} - 1$. So if $\|v - w\|_\infty < 1/3 \log(16/9)$, then from (3.13) we can see that

$$\mathbb{E}_p \log \frac{p_v}{p_w} \le 4\|v - w\|_\infty^2.$$

We prove the second part of the lemma in the same way. We can check that $e^x \le (16/9)^{\frac{1}{3}}$ if $x \le 1/3 \log(16/9)$ and $(1 + x) < (16/9)^{\frac{1}{3}}$ if $x < (16/9)^{\frac{1}{3}} - 1$. Hence for all v, $w : \mathcal{X} \mapsto \mathbb{R}$ for which $\|v - w\|_\infty < 1/3 \log(16/9)$ holds,

$$\begin{aligned} V(p_v, p_w) &\le h^2(p_v, p_w)\Big(3 + \log \Big\| \frac{p_v}{p_w} \Big\|_\infty\Big)^2 \\ &\le \|v - w\|_\infty^2 e^{\|v-w\|_\infty}(3 + 2\|v - w\|_\infty)^2 \\ &\le 9\|v - w\|_\infty^2 e^{\|v-w\|_\infty}(1 + \|v - w\|_\infty)^2 \\ &\le 16\|v - w\|_\infty^2. \end{aligned}$$

□

The next theorem gives us a construction of a sequence of sets B_n, which helps us to construct a sequence of sets $\mathcal{P}_n$ for Theorem 3.2 above.

Theorem 3.9 (Theorem 2.1 of [8]). *Let W be a Borel measurable, zero-mean, Gaussian random element in a common separable Banach space $(\mathbb{B}, \|\cdot\|)$, with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and let w_0 be contained in the closure of $\mathbb{H}$ in $\mathbb{B}$. For a sequence $\varepsilon_n > 0$ satisfying (3.2) for φ_{w_0} given by (1.3) and any $C > 1$ with $\exp[-Cn\varepsilon_n^2] < 1/2$, there exists a measurable set $B_n \subset \mathbb{B}$ such that*

$$\log N(3\varepsilon_n, B_n, \|\cdot\|) \leq 6Cn\varepsilon_n^2, \tag{3.14}$$

$$P(W \notin B_n) \leq \exp[-Cn\varepsilon_n^2], \tag{3.15}$$

$$P(\|W - w_0\|^2 < 4\varepsilon_n^2) \geq \exp[-n\varepsilon_n^2]. \tag{3.16}$$

The method of construction is presented in the next chapter, in the proof of our main theorem. So the $\mathcal{P}_n$ sets are constructed as

$$\mathcal{P}_n = \{\frac{e^{w_x}}{\int e^{w_x}\,\mathrm{d}x} : w \in B_n\}.$$

The three assertion of this theorem can be matched one by one with the conditions of the fundamental theorem. The main difference between them is that in (3.14), (3.15) and (3.16) the norm of the Banach space stands, while in the fundamental theorem the assumptions are in terms of metrics appropriate to the statistical problem under consideration (in our case this is usually the Hellinger or the total variation metric). Hence we need Lemma 3.7 or Lemma 3.8 to ensure that our above constructed $\mathcal{P}_{n,\alpha}$ sets satisfy the conditions of the fundamental theorem.

Finally we give Theorem 3.1 of [8].

Theorem 3.10. *Let W be a Borel measurable, zero-mean tight Gaussian random element in $L^\infty(\mathcal{X})$. Suppose that $w_0 = \log p_0$ is contained in the support of W and let φ_{w_0} be the function in (1.3) with $\|\cdot\|$ the uniform norm on L^∞. Furthermore assume that ε_n satisfies (3.2). Then the posterior distribution relative to the prior Π satisfies*

$$\mathbb{E}_0\Pi(p_w : h(p_w, p_{w_0}) > M\varepsilon_n | X_1, X_2, \ldots, X_n) \to 0$$

for a sufficiently large constant M, which is independent from the choice of W.

Remark 3.11. In the proof of Theorem 3.10 in [8] it was shown that the three conditions of the fundamental theorem are satisfied. So if the conditions of Theorem 3.10 are satisfied, then we can apply the results from Lemma 3.3.

4 Main Theorem

In this chapter we give a hierarchical prior, for which the posterior contraction rate is rate adaptive. Let $\{W^\alpha : \alpha \in A\}$ denote the collection of zero mean Gaussian random elements, where $A \subseteq (0, \infty)$ and define for all $\alpha \in A$ the prior distribution Π^α according to (1.1) with W^α. Then define $\varepsilon_{n,\alpha}$ for all $\alpha \in A$, such that it satisfies (3.2), where the concentration function is given by (1.3). Furthermore assume that $n\varepsilon_{n,\alpha}^2 = n^{g(\alpha)}$, where $g : [0, \infty) \mapsto (0, \infty)$ is a strictly decreasing continuous function. By good choice of λ_n distribution we generalize Theorem 3.1.

Theorem 4.1. *There is a hierarchical model, depending on the sample size, such that the rate of contraction is rate adaptive. More precisely assume that for all $\alpha < \beta$ $(\alpha, \beta \in A)$*

$$\mathbb{H}_1^\beta \subseteq K\mathbb{H}_1^\alpha \tag{4.1}$$

holds, where W^α is zero mean Gaussian random element and $\mathbb{H}_1^\alpha$ is the unit ball of the RKHS of W^α. Moreover, assume that $\varepsilon_{n,\alpha}$ satisfies (3.2) for all $w \in \mathcal{W}^\alpha$ for all $\alpha \in A$ and $n\varepsilon_{n,\alpha}^2 = n^{g(\alpha)}$, with a strictly decreasing continuous function $g : [0, \infty) \mapsto (0, \infty)$. Then there exists a prior distribution λ_n such that the hierarchical posterior distribution (3.1) satisfies

$$\mathbb{E}_0 \Pi_n(p : h(p, p_0) > M\varepsilon_{n,\beta} | X_1, \ldots, X_n) \to 0 \tag{4.2}$$

if $\log p_0 \in \mathcal{W}^\beta$ for $\beta \in A$ and M is large enough.

Proof. First we give the construction of λ_n.

Lemma 4.2. *Let $A \subset (0, \infty)$ be the set of α parameters and $g : [0, \infty) \mapsto (0, \infty)$ be a strictly decreasing function. For an arbitrary constant $H > 0$ we can construct a finite set $B_n \subset A$, such that*

$$\#B_n \leq \frac{2g(0)}{\log H} \log n \tag{4.3}$$

and there exists $\tilde{\alpha}_n \geq \max B_n$, such that for n large enough we have

$$n^{g(\tilde{\alpha}_n)} - \log\log n \to \infty$$

Moreover, for all $\beta \in A$ there exists a sequence $\beta_n \in B_n$ such that for n large enough we have

$$0 \le g(\beta_n) - g(\beta) < \log H / \log n \tag{4.4}$$

and for all $\alpha \in B_n \cap (0, \beta_n)$

$$g(\alpha) - g(\beta_n) \ge \frac{\log H}{2 \log n} \tag{4.5}$$

Let B_n be the set obtained from Lemma 4.2 with the constant H sufficiently large, $H = 130^2 + 1$ is an appropriate choice, and define λ_n as

$$\lambda_n(\alpha) = \frac{e^{-n\varepsilon_{n,\alpha}^2} I_{B_n}(\alpha)}{\sum_{\alpha \in B_n} e^{-n\varepsilon_{n,\alpha}^2}}. \tag{4.6}$$

Assume that $\log p_0 \in \mathcal{W}^\beta$ with some $\beta \in A$. Then by Lemma 4.2 there exists a sequence β_n satisfying (4.4). Next we show that

$$\mathbb{E}_0 \Pi(p : h(p, p_0) > M' \varepsilon_{n,\beta_n} | X_1, \dots, X_n) \to 0 \tag{4.7}$$

if M' is large enough. This formula is the same as (4.2) except β is replaced with β_n. After showing this we are finished, since from (4.4)

$$\log H^{\frac{1}{2}} \ge \log \varepsilon_{n,\beta_n} - \log \varepsilon_{n,\beta}$$

holds. Hence from this and (4.7)

$$\begin{aligned} &\mathbb{E}_0 \Pi(p : h(p, p_0) > M' H^{\frac{1}{2}} \varepsilon_{n,\beta} | X_1, \dots, X_n) \le \\ &\qquad \mathbb{E}_0 \Pi(p : h(p, p_0) > M' \varepsilon_{n,\beta_n} | X_1, \dots, X_n) \to 0. \end{aligned}$$

Denoting by D the set $\{p : h(p, p_0) \ge M' \varepsilon_{n,\beta_n}\}$, we have

$$\begin{aligned} \Pi(D | X_1, \dots, X_n) &= \frac{\sum_\alpha \lambda_n(\alpha) \int_D L(p) \Pi^\alpha(dp)}{\sum_\alpha \lambda_n(\alpha) \int L(p) \Pi^\alpha(dp)} \\ &\le \frac{\sum_{\alpha < \beta_n} \lambda_n(\alpha) \int_D L(p) \Pi^\alpha(dp)}{\sum_\alpha \lambda_n(\alpha) \int L(p) \Pi^\alpha(dp)} + \Pi'(D | X_1, \dots, X_n) \\ &= I + II, \end{aligned}$$

where

$$\Pi' = \sum_{\alpha \ge \beta_n} \lambda_n'(\alpha) \Pi^\alpha, \tag{4.8}$$

with

$$\lambda'_n(\alpha) = \frac{\lambda_n(\alpha)}{\sum_{\alpha \geq \beta_n} \lambda_n(\alpha)}. \tag{4.9}$$

We study I and II separately. The main difference between the proof of Theorem 3.1 and the proof of this theorem takes place in this part. In Theorem 3.1 the number of terms in expression I is uniformly bounded in n, while in our case it may go to infinity with n.

Lemma 4.3. *For λ_n given in* (4.6) *and $L(p)$ defined in* (1.4)

$$\mathbb{E}_0 \frac{\sum_{\alpha < \beta_n} \lambda_n(\alpha) \int_D L(p) \Pi^\alpha(dp)}{\sum_\alpha \lambda_n(\alpha) \int L(p) \Pi^\alpha(dp)} \longrightarrow 0 \quad \text{as } n \to \infty,$$

where $D = \{p : h(p, p_0) \geq M' \varepsilon_{n,\beta_n}\}$, β_n is given by (4.4) *and* $\log p_0 \in \mathcal{W}^\beta$.

The following lemma ensures that the expected value of the expression II also goes to 0.

Lemma 4.4. *For the hierarchical prior distribution Π'_n given by* (4.8)

$$\mathbb{E}_0 \Pi'_n(p : h(p, p_0) \geq M' \varepsilon_{n,\beta_n} | X_1, \ldots, X_n) \longrightarrow 0 \quad \text{as } n \to \infty,$$

where β_n is given by (4.4) *and* $\log p_0 \in \mathcal{W}^\beta$.

Finally one can see that from Lemma 4.3 and Lemma 4.4 the convergence (4.7) holds. □

Remark 4.5. Assume that A has no accumulation point and there exists a constant $\delta > 0$ such that for every distinct pair of $\alpha, \beta \in A$ $|\alpha - \beta| > \delta$ holds. In this case we can use Theorem 3.1 (with small modifications), instead of using Theorem 4.1, since the expression I contains only uniformly bounded amount of points for all n. The estimation of II is proved in a similar way in both of the theorems. The application of Theorem 3.1 makes the $\lambda_n(\alpha)$ prior distribution simpler, because we do not have to deal with the B_n sets, but we can say that the support of the prior λ_n is the whole set A for all n.

Next we give the proofs of the lemmas used in the proof of Theorem 4.1.

Proof of Lemma 4.2. In the proof we apply Example 2.9 of [6]. Choose $\widetilde{\alpha}_n \nearrow \infty$ such that $n^{g(\widetilde{\alpha}_n)} - \log\log n \to \infty$. Let

$$B_n \subset (0, \widetilde{\alpha}_n] \cap A$$

be such that $\{g(\alpha)\log n : \alpha \in B_n\}$ is a $\log H$-net in the set $\{g(\alpha)\log n : \alpha \in (0, \widetilde{\alpha}_n] \cap A\}$ and there doesn't exist $\alpha_1, \alpha_2 \in B_n$ $(\alpha_1 \neq \alpha_2)$, such that $|g(\alpha_1) - g(\alpha_2)| < \frac{\log H}{2\log n}$. We can choose the set B_n for example by greedy algorithm. Since $\tilde{\alpha}_n \nearrow \infty$, $g(\beta)\log n$ is in the set for n large enough. Define the sequence β_n as

$$\beta_n = \max\{\alpha \leq \beta : \alpha \in B_n\}.$$

In this case $|g(\beta_n) - g(\beta)| < \log H/\log n \Longrightarrow \beta_n \to \beta$. One can easily see that $\#B_n < \frac{2g(0)}{\log H}\log n$. □

Proof of Lemma 4.3. Let $D = D_1 \cup D_2$, where $D_1 = \{p : h(p, p_0) \geq M\varepsilon_{n,\alpha}\}$ and $D_2 = \{p : M\varepsilon_{n,\beta_n} < h(p, p_0) \leq M\varepsilon_{n,\alpha}\}$. We can write

$$\frac{\sum_{\alpha<\beta_n}\lambda_n(\alpha)\int_D L(p)\Pi^\alpha(dp)}{\sum_\alpha \lambda_n(\alpha)\int L(p)\Pi^\alpha(dp)} \leq \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} \frac{\int_{D_1} L(p)\Pi^\alpha(dp)}{\int L(p)\Pi^\alpha(dp)} + \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} \frac{\lambda_n(\alpha)\int_{D_2} L(p)\Pi^\alpha(dp)}{\lambda_n(\beta_n)\int L(p)\Pi^{\beta_n}(dp)}.$$

From Remark 3.11 there exist tests φ_n^α and a set of events $Q_n \subset \mathcal{X}^n$, such that

$$\mathbb{E}_0\Pi^\alpha(p : h(p, p_0) > M\varepsilon_{n,\alpha}|X_1, \ldots, X_n)\varphi_n^\alpha \leq 2e^{-Kn\varepsilon_{n,\alpha}^2}, \quad (4.10)$$

$$\mathbb{E}_0\Pi^\alpha(p \in Q_n : h(p, p_0) > M\varepsilon_{n,\alpha}|X_1, \ldots, X_n)(1 - \varphi_n^\alpha)I_{Q_n} \leq 2e^{-2n\varepsilon_{n,\alpha}^2}, \quad (4.11)$$

$$\mathbb{E}_0 I_{Q_n^c} \leq \frac{1}{n\varepsilon_{n,\alpha}^2}, \quad (4.12)$$

where K is an universal constant. From the seventh section of [2] in the Hellinger distance's case this universal constant is $K = \frac{1}{2}$. One can easily see that for an arbitrary constant $c > 0$, and for enough large x

$$2e^{-cx} \leq \frac{1}{x}. \quad (4.13)$$

Specially for $c = 2$ it is true for all x and for $c = \frac{1}{2}$ for $x \geq 0.715$. From assumption $n\varepsilon_{n,\alpha}^2 = n^{g(\alpha)}$, where $g(\alpha) > 0$ for all $\alpha > 0$, $n\varepsilon_{n,\alpha}^2 > 1$ for all $n \in \mathbb{N}$. So for all $\alpha > 0$ and $n \in \mathbb{N}$ (4.13) holds with $x = n\varepsilon_{n,\alpha}^2$ and $c = 2$ or $c = \frac{1}{2}$. Hence from (4.10), (4.11), (4.12) and (4.13) one can easily see

$$\mathbb{E}_0\frac{\int_{D_1} L(p)\Pi^\alpha(dp)}{\int L(p)\Pi^\alpha(dp)} = \mathbb{E}_0\Pi^\alpha(p : h(p, p_0) > M\varepsilon_{n,\alpha}|X_1, \ldots, X_n) \leq \frac{3}{n\varepsilon_{n,\alpha}^2}. \quad (4.14)$$

From (4.3), (4.14) and the inequality $\varepsilon_{n,\alpha} > \varepsilon_{n,\beta_n}$ for all $\alpha < \beta_n$

$$\begin{aligned}\mathbb{E}_0 \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} \frac{\int_{D_1} L(p)\Pi^\alpha(dp)}{\int L(p)\Pi^\alpha(dp)} &\le \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} \frac{3}{n\varepsilon_{n,\alpha}^2} \\ &\le \frac{6g(0)}{\log H}\frac{\log n}{n\varepsilon_{n,\beta_n}^2} \\ &\le \frac{6g(0)}{\log H}\frac{\log n}{n^{g(\beta)}}\end{aligned}$$

holds, where the right side goes to 0 as $n \nearrow \infty$.

Next, we study the second term. From Lemma 3.4 and the fact that $n\varepsilon_{n,\beta_n}^2 \to \infty$, it follows that

$$\begin{aligned}\int L(p)\Pi^{\beta_n}(\mathrm{d}p) \ge& e^{-2\cdot 2^6 n\varepsilon_{n,\beta_n}^2} \\ &\times \Pi^{\beta_n}\Big(p:\ K(p_0,p) \le 2^6\varepsilon_{n,\beta_n}^2,\ V(p_0,p) \le 2^6\varepsilon_{n,\beta_n}^2\Big) \quad (4.15)\end{aligned}$$

on an event A_n with P_0-probability tending to 1. From Theorem 3.9 and Lemma 3.8 for an arbitrary $w_0 \in \mathcal{W}^{\beta_n}$, which for $\log p_0 = w_0$,

$$\begin{aligned}&\Pi^{\beta_n}\Big(p:\ K(p_0,p) \le 2^6\varepsilon_{n,\beta_n}^2,\ V(p_0,p) \le 2^6\varepsilon_{n,\beta_n}^2\Big) \\ &\quad \overset{Lemma\,3.8}{\ge} P(4\|W-w_0\|_\infty^2 \le 2^6\varepsilon_{n,\beta_n}^2, 16\|W-w_0\|_\infty^2 \le 2^6\varepsilon_{n,\beta_n}^2) \\ &\quad = P(\|W-w_0\|_\infty \le 2\varepsilon_{n,\beta_n}) \\ &\quad \overset{(3.16)}{\ge} e^{-n\varepsilon_{n,\beta_n}^2}. \quad (4.16)\end{aligned}$$

Hence by (4.15) and (4.16)

$$\begin{aligned}\int L(p)\Pi^{\beta_n}(\mathrm{d}p) &\ge e^{-2^7 n\varepsilon_{n,\beta_n}^2} e^{-n\varepsilon_{n,\beta_n}^2} \\ &= e^{-(2^7+1)n\varepsilon_{n,\beta_n}^2}\end{aligned}$$

on the event A_n. In the following inequality we apply Fubini's theorem, the fact that $\mathbb{E}_0 L(p) = 1$ and the definition of distribution λ_n. Hence

$$\begin{aligned}\mathbb{E}_0 \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} \frac{\lambda_n(\alpha)\int_{D_2} L(p)\Pi^\alpha(dp)}{\lambda_n(\beta_n)\int L(p)\Pi^{\beta_n}(dp)} I_{A_n} &\le \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} \frac{\lambda_n(\alpha)}{\lambda_n(\beta_n)} e^{(2^7+1)n\varepsilon_{n,\beta_n}^2}\Pi^\alpha(D_2) \\ &\le \sum_{\substack{\alpha<\beta_n \\ \alpha\in B_n}} e^{(2^7+1+1)n\varepsilon_{n,\beta_n}^2 - n\varepsilon_{n,\alpha}^2} \\ &\le \frac{2g(0)}{\log H} e^{-(H^{\frac{1}{2}}-2^7-2)n\varepsilon_{n,\beta_n}^2}\log n, \quad (4.17)\end{aligned}$$

where the last inequality comes from (4.5), because for all $\alpha \in B_n \wedge (0, \beta_n)$:

$$\frac{1}{4}\log H + \log \varepsilon_{n,\beta_n} \leq \log \varepsilon_{n,\alpha}$$
$$\varepsilon_{n,\beta_n} H^{\frac{1}{4}} \leq \varepsilon_{n,\alpha}$$
$$-n\varepsilon_{n,\beta_n}^2 H^{\frac{1}{2}} \geq -n\varepsilon_{n,\alpha}^2.$$

Finally from the definition of H ($H = 130^2 + 1 > (2^7+2)^2$) we get that the right hand side of (4.17) goes to 0. □

Proof of Lemma 4.4. Theorem 3.2 says that if the following conditions hold for $\varepsilon'_{n,\beta_n} = 8\varepsilon_{n,\beta_n}$, where $\varepsilon'_{n,\beta_n} \searrow 0$, $n(\varepsilon'_{n,\beta_n})^2 \to \infty$, and $\mathcal{P}_{n,\beta_n}$ measurable sets

$$\log N(\varepsilon'_{n,\beta_n}, \mathcal{P}_{n,\beta_n}, h) \leq c_1 n(\varepsilon'_{n,\beta_n})^2, \tag{4.18}$$
$$\Pi'(\mathcal{P}^c_{n,\beta_n}) \leq c_2 e^{-(c_3+4)n(\varepsilon'_{n,\beta_n})^2}, \tag{4.19}$$
$$\Pi'\big(p:\ K(p_0,p) \leq (\varepsilon'_{n,\beta_n})^2,\ V(p_0,p) \leq (\varepsilon'_{n,\beta_n})^2\big) \geq e^{-c_3 n(\varepsilon'_{n,\beta_n})^2}, \tag{4.20}$$

then

$$\mathbb{E}_0\Pi'(p : h(p,p_0) \geq M^*\varepsilon'_{n,\beta_n}|X^{(n)}) = \mathbb{E}_0\Pi'(D|X^{(n)}) \to 0,$$

with $M' = 8M^*$. So our goal is to verify these conditions.

Since $\beta_n \to \beta$, there exists $N > 0$, such that for all $n > N$: $\varepsilon'_{n,\beta_n} \leq \varepsilon'_{n,\frac{\beta}{2}} \to 0$ as $n \to \infty$. From assumption $n(\varepsilon'_{n,\beta_n})^2 \geq n(\varepsilon'_{n,\beta})^2$, where the right hand side goes to infinity, so $n(\varepsilon'_{n,\beta_n})^2 \to \infty$.

Lemma 4.6. *There is a $\delta > 0$ such that we have*

$$\lambda_n(\beta) > \delta e^{-n\varepsilon^2_{n,\beta}} \sum_{\substack{\alpha \geq \beta \\ \alpha \in B_n}} \lambda_n(\alpha) \tag{4.21}$$

for all n and $\beta \in B_n$.

According to (4.9) Lemma 4.6 says that exists $\delta > 0$ such that $\lambda'_n(\beta_n) > \delta \exp\{-n(\varepsilon'_{n,\beta_n})^2\}$. From this, (4.16) and $\varphi_{w_0^\alpha}(\varepsilon'_{n,\alpha}) \leq n(\varepsilon'_{n,\alpha})^2$ we get that

$$\begin{aligned}
&\Pi'\Big(p:\ K(p_0,p) \leq (\varepsilon'_{n,\beta_n})^2,\ V(p_0,p) \leq (\varepsilon'_{n,\beta_n})^2\Big) \\
&\quad = \Pi'\Big(p:\ K(p_0,p) \leq 2^6\varepsilon^2_{n,\beta_n},\ V(p_0,p) \leq 2^6\varepsilon^2_{n,\beta_n}\Big) \\
&\quad \geq \lambda'_n(\beta_n)\Pi^{\beta_n}\Big(p:\ K(p_0,p) \leq 2^6\varepsilon^2_{n,\beta_n},\ V(p_0,p) \leq 2^6\varepsilon^2_{n,\beta_n}\Big) \\
&\quad \geq \lambda'_n(\beta_n)\mathbb{P}(\|W^{\beta_n} - w_0\|_\infty \leq 2\varepsilon_{n,\beta_n}) \\
&\quad \geq \delta e^{-2n\varepsilon^2_{n,\beta_n}} \\
&\quad = \delta e^{-n2^{-5}(\varepsilon'_{n,\beta_n})^2}
\end{aligned}$$

for some $\delta > 0$ independent of n. Hence the prior mass condition (4.20) is satisfied. We choose $\mathcal{P}_{n,\beta_n}$ equal to

$$\mathcal{P}_{n,\beta_n} = \{p_w : w \in \frac{\varepsilon'_{n,\beta_n}}{4}\mathbb{B}_1 + M_n^{\beta_n}\mathbb{H}_1^{\beta_n}\},$$

with $\mathbb{B}_1$ the unit ball in $L^\infty(\mathcal{X})$ and for all $\alpha > 0$

$$M_n^\alpha = 2K\Phi^{-1}(1 - e^{-(c_3+4)n(\varepsilon'_{n,\alpha})^2}),$$

with Φ the cdf of the standard normal distribution. By the proof of Theorem 2.1 of [8] we can verify, that these sets satisfy the right entropy condition (4.18), see below.

Let $h_1, \ldots, h_n$ be contained in $M_n^{\beta_n}\mathbb{H}_1^{\beta_n}$ and be $\varepsilon'_{n,\beta_n}/2$-separated for the norm $\|\cdot\|$, such that doesn't exist $h \in M_n^{\beta_n}\mathbb{H}_1^{\beta_n}$, which for $|h_j - h| > \varepsilon'_{n,\beta_n}/2$ for every $j \in \{1, ..., n\}$. Then the $\|\cdot\|$-balls $h_j + (\varepsilon'_{n,\beta_n}/4)\mathbb{B}_1$ of radius $\varepsilon'_{n,\beta_n}/4$ around these h_j points are disjoint, hence

$$\begin{aligned} 1 &\geq \sum_{j=1}^N P(W^{\beta_n} \in h_j + \frac{\varepsilon'_{n,\beta_n}}{4}\mathbb{B}_1) \\ &\geq \sum_{j=1}^N e^{-\frac{1}{2}\|h_j\|^2_{\mathbb{H}^{\beta_n}}} P(W^{\beta_n} \in \frac{\varepsilon'_{n,\beta_n}}{4}\mathbb{B}_1) \\ &\geq N e^{-\frac{1}{2}(M_n^{\beta_n})^2} e^{-\varphi_0(\varepsilon'_{n,\beta_n}/4)}, \end{aligned} \tag{4.22}$$

where the second inequality follows from 4.16 of [5]. Since the $h_j + (\varepsilon'_{n,\beta_n}/2)\mathbb{B}_1$ balls cover $M_n^{\beta_n}\mathbb{H}_1^{\beta_n}$

$$N(\varepsilon'_{n,\beta_n}/2, M_n^{\beta_n}\mathbb{H}_1^{\beta_n}, \|\cdot\|) \leq N \leq e^{\frac{1}{2}(M_n^{\beta_n})^2 + \varphi_0(\varepsilon'_{n,\beta_n}/4)},$$

where the last inequality follows from (4.22). By its definition, any point of the set $B_n = \{(\varepsilon'_{n,\beta_n}/4)\mathbb{B}_1 + M_n^{\beta_n}\mathbb{H}_1^{\beta_n}\}$ is within distance $\varepsilon'_{n,\beta_n}/4$ of some point of $M_n^{\beta_n}\mathbb{H}_1^{\beta_n}$. From the definition of ε'_{n,β_n}: $\varepsilon'_{n,\beta_n}/4 > \varepsilon_{n,\beta_n}$, hence (3.2) holds with $\varepsilon'_{n,\beta_n}/4$ also. These implies that

$$\begin{aligned} \log N\Big(\frac{3}{4}\varepsilon'_{n,\beta_n}, B_n, \|\cdot\|\Big) &\leq \log N\Big(\frac{\varepsilon'_{n,\beta_n}}{2}, M_n^{\beta_n}\mathbb{H}_1^{\beta_n}, \|\cdot\|\Big) \\ &\leq \frac{1}{2}{M_n^{\beta_n}}^2 + \varphi_0(\varepsilon'_{n,\beta_n}/4) \\ &\leq 5(c_3+4)K^2 n(\varepsilon'_{n,\beta_n})^2 + n(\varepsilon'_{n,\beta_n}/4)^2 \\ &\leq (5c_3+21)K^2 n(\varepsilon'_{n,\beta_n})^2 \end{aligned} \tag{4.23}$$

by the definition of $M_n^{\beta_n}$ if $e^{-(c_3+4)n(\varepsilon'_{n,\beta_n})^2} < \frac{1}{2}$, because $\Phi^{-1}(y) \geq -\sqrt{\frac{5}{2}\log\frac{1}{y}}$ and is negative for every $y \in (0, \frac{1}{2})$. From the proof of Theorem 3.1 of [8]

$$\log N(\varepsilon'_{n,\beta_n}, \mathcal{P}_{n,\beta_n}, h) \leq \log N\big((3/4)\varepsilon'_{n,\beta_n}, B_n, \|\cdot\|\big) \tag{4.24}$$

holds. From (4.23) and (4.24) we get that

$$\log N(\varepsilon'_{n,\beta_n}, \mathcal{P}_{n,\beta_n}, h) \leq (5c_3+21)K^2 n(\varepsilon'_{n,\beta_n})^2,$$

which is the same up to a constant multiplier as (4.18), with $c_1 = 6(c_3+4)K^2$.

By concentration assumption (4.1) and Borell's inequality (Theorem 3.1)

$$\begin{aligned}\Pi^\alpha(\mathcal{P}^c_{n,\beta_n}) &= \mathbb{P}(W^\alpha \notin (\varepsilon'_{n,\beta_n}/4)\mathbb{B}_1 + M_n^{\beta_n}\mathbb{H}_1^{\beta_n}) \\ &\leq \mathbb{P}(W^\alpha \notin (\varepsilon'_{n,\beta_n}/4)\mathbb{B}_1 + \frac{M_n^{\beta_n}}{K}\mathbb{H}_1^{\alpha}) \\ &\leq 1 - \Phi(c_n + \frac{M_n^{\beta_n}}{K})\end{aligned}$$

for $\alpha \geq \beta_n$, where $c_n = \Phi^{-1}\big(\Pi^\alpha\big((\varepsilon'_{n,\beta_n}/4)\mathbb{B}_1\big)\big)$. Since $\varepsilon'_{n,\alpha} \leq \varepsilon'_{n,\beta_n}$, we have $c_n \geq \Phi^{-1}\big(\Pi^\alpha\big((\varepsilon'_{n,\alpha}/4)\mathbb{B}_1\big)\big)$ and $M_n^\alpha \leq M_n^{\beta_n}$. From the following expressions

$$\begin{aligned}-\frac{1}{2K}M_n^{\beta_n} \leq -\frac{1}{2K}M_n^\alpha &= \Phi^{-1}(e^{-(c_3+4)n(\varepsilon'_{n,\alpha})^2}) \\ &\leq \Phi^{-1}(e^{-n(\varepsilon'_{n,\alpha}/4)^2}) \leq \Phi^{-1}\big(\Pi^\alpha\big((\varepsilon'_{n,\alpha}/4)\mathbb{B}_1\big)\big) \leq c_n\end{aligned} \tag{4.25}$$

we can see that $c_n \geq -\frac{1}{2K}M_n^{\beta_n}$, hence $c_n + \frac{1}{K}M_n^{\beta_n} \geq \frac{1}{2K}M_n^{\beta_n}$. The inequality (4.25) follows from $P\big(W^\alpha \notin (\varepsilon'_{n,\alpha}/4)\mathbb{B}_1\big) = e^{-\varphi_0(\varepsilon'_{n,\alpha}/4)}$ and $\varphi_0(\varepsilon'_{n,\alpha}/4) \leq n(\varepsilon'_{n,\alpha}/4)^2$. So we can write that

$$\begin{aligned}\Pi^\alpha(\mathcal{P}^c_{n,\beta_n}) &\leq 1 - \Phi(c_n + \frac{M_n^{\beta_n}}{K}) \\ &\leq 1 - \Phi(\frac{M_n^{\beta_n}}{2K}) = e^{-(c_3+4)n(\varepsilon'_{n,\beta_n})^2}.\end{aligned}$$

Hence the last condition (4.19) is also satisfied. □

Proof of Lemma 4.6. We use the notation

$$k_n = \sum_{\alpha \in B_n} e^{-n\varepsilon^2_{n,\alpha}}$$

during the proof. An upper bound can be formulated for the right hand side

of (4.21)

$$e^{-n\varepsilon_{n,\beta}^2}\sum_{\substack{\alpha\geq\beta\\ \alpha\in B_n}}\lambda_n(\alpha)\leq e^{-n\varepsilon_{n,\beta}^2}\sum_{\alpha\in B_n}\lambda_n(\alpha)$$
$$\leq\frac{2g(0)}{\log H}e^{-n\varepsilon_{n,\beta}^2}\frac{\log n}{k_n}e^{-n\varepsilon_{n,\widetilde{\alpha}_n}^2}$$
$$=\frac{e^{-n\varepsilon_{n,\beta}^2}}{k_n}\frac{2g(0)}{\log H}e^{-n^{g(\widetilde{\alpha}_n)}}\log n. \tag{4.26}$$

From the definition of $\widetilde{\alpha}_n$ the convergence

$$\log\frac{\log n}{e^{n^{g(\widetilde{\alpha}_n)}}}\to-\infty$$

holds. So if n is big enough, then $\big(2g(0)/\log H\big)e^{-n^{g(\widetilde{\alpha}_n)}}\log n<1$ and from this and (4.26)

$$e^{-n\varepsilon_{n,\beta}^2}\sum_{\substack{\alpha\geq\beta\\ \alpha\in B_n}}\lambda_n(\alpha)\leq\frac{e^{-n\varepsilon_{n,\beta}^2}}{k_n}=\lambda_n(\beta).$$

Choosing δ sufficiently small (4.21) holds for all n. □

5 Examples

In this chapter we give two possible constructions of the models $\{\mathcal{P}^\alpha\colon \alpha \in A\}$, where $A \subset \mathbb{R}^+$ and a λ_n distribution on the models for which the Bayesian estimation is rate adaptive.

5.1 The Extended Riemann-Liouville process

In the first example we deal with the

$$X_t^{(\alpha)} = \sum_{i=0}^{k} \frac{t^i}{i!} Z_i + (I_{0+}^k W)_t$$

process, where $k = \alpha - 1/2 \in \mathbb{N}$. In section 2.2 and section 3.2 we have shown that we can apply the Riemann-Liouville process to estimate α-regular functions, where α is known ($\log p_0 \in \mathcal{W}^\alpha = C^\alpha[0,1]$), and the posterior contraction rate is the minimax rate. We would like to extend our estimation procedure for unknown α. Our goal is to show that the posterior contraction rate is rate adaptive in the special case when our true density is α-regular, where $\alpha - 1/2 \in \mathbb{N}$. First we prove that (4.1) holds for this we need the following lemma.

Lemma 5.1. *For all $\alpha, \beta \in \mathbb{N}$, $\alpha < \beta$ and for all f β-times differentable, $f^{(\beta)}$ absolut continuous, square integrable functions:*

$$\|f\|_\alpha \leq \sqrt{e}\|f\|_\beta,$$

where

$$\|f\|_\alpha = \sum_{i=0}^{\alpha} f^{(i)}(0)^2 + \int_0^1 f^{(\alpha+1)}(s)^2 \,\mathrm{d}s.$$

Proof. First we will show that for all $\alpha \in \mathbb{N}$

$$\|f\|_1 \leq \sqrt{e}\|f\|_\alpha.$$

The following inequality holds

$$\begin{aligned}(a_0 + a_1 + \frac{a_2}{2!} + \cdots + \frac{a_k}{k!})^2 &\leq (1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{k!})(a_0^2 + a_1^2 + \cdots + a_k^2) \\ &\leq e(a_0^2 + a_1^2 + \cdots + a_k^2), \end{aligned} \tag{5.1}$$

because

$$2\frac{a_i}{i!}\frac{a_j}{j!} = 2\frac{a_i}{\sqrt{i!j!}}\frac{a_j}{\sqrt{i!j!}} \leq (\frac{a_i}{\sqrt{i!j!}})^2 + (\frac{a_j}{\sqrt{i!j!}})^2 \leq \frac{a_i^2}{j!} + \frac{a_j^2}{i!}.$$

For an arbitrary $s \in (0,1)$

$$\begin{aligned}
f^{(2)}(s) &= f^{(2)}(0) + \int_0^s f^{(3)}(z_1)\,\mathrm{d}z_1 \\
&= f^{(2)}(0) + \int_0^s f^{(3)}(0) + \int_0^{z_1} f^{(4)}(z_2)\,\mathrm{d}z_2\,\mathrm{d}z_1 \\
&= f^{(2)}(0) + sf^{(3)}(0) + \int_0^s \int_0^{z_1} f^{(4)}(0) + \int_0^{z_2} f^{(5)}(z_3)\,\mathrm{d}z_3\,\mathrm{d}z_2\,\mathrm{d}z_1 = ... \\
&= f^{(2)}(0) + sf^{(3)}(0) + \frac{s^2}{2!}f^{(4)}(0) + \cdots + \frac{s^{\alpha-2}}{(\alpha-2)!}f^{(\alpha)}(0) \\
&\quad + \int_0^s \int_0^{z_1} ... \int_0^{z_{\alpha-2}} f^{(\alpha+1)}(z_{\alpha-1})\,\mathrm{d}z_{\alpha-1}... \,\mathrm{d}z_2\,\mathrm{d}z_1 \\
&= f^{(2)}(0) + sf^{(3)}(0) + \frac{s^2}{2!}f^{(4)}(0) + \cdots + \frac{s^{\alpha-2}}{(\alpha-2)!}f^{(\alpha)}(0) \\
&\quad + \frac{1}{(\alpha-1)!}\int_0^s f^{(\alpha+1)}(z_{\alpha-1})(s-z_{\alpha-1})^{\alpha-1}\,\mathrm{d}z_{\alpha-1}.
\end{aligned}$$

Hence

$$\begin{aligned}
f^{(2)}(s)^2 &\overset{(5.1)}{\leq} e\Bigg(f^{(2)}(0)^2 + s^2 f^{(3)}(0)^2 + s^4 f^{(4)}(0)^2 + \ldots + s^{2(\alpha-2)} f^{(\alpha)}(0)^2 \\
&\qquad + \Big(\int_0^s f^{(\alpha+1)}(z)(s-z)^{\alpha-1}\,\mathrm{d}z \Big)^2 \Bigg) \\
&\overset{Jensen,\, s\leq 1, CSB}{\leq} e\Bigg(f^{(2)}(0)^2 + f^{(3)}(0)^2 + f^{(4)}(0)^2 + \ldots + f^{(\alpha)}(0)^2 \\
&\qquad + \int_0^s f^{(\alpha+1)}(z)^2\,\mathrm{d}z \int_0^s (s-z)^{2(\alpha-1)}\,\mathrm{d}z \Bigg) \\
&\overset{s\in(0,1)}{\leq} e\Bigg(f^{(2)}(0)^2 + f^{(3)}(0)^2 + f^{(4)}(0)^2 + \ldots + f^{(\alpha)}(0)^2 \\
&\qquad + \int_0^1 f^{(\alpha+1)}(z)^2\,\mathrm{d}z \Bigg). \qquad (5.2)
\end{aligned}$$

With the help of the above we can finish the proof, since

$$\begin{aligned}\|f\|_1^2 &= f(0)^2 + f^{(1)}(0)^2 + \int_0^1 f^{(1)}(z)\,\mathrm{d}z \\ &\overset{(5.2)}{\leq} e\left(f(0)^2 + f^{(1)}(0)^2 + f^{(2)}(0)^2 + \ldots + f^{(\alpha)}(0)^2 + \int_0^1 f^{(\alpha+1)}(z)^2\,\mathrm{d}z \right) \\ &= e\|f\|_\alpha^2.\end{aligned}$$

Applying this result to $f^{(\alpha-1)}$ and $\alpha' = \beta - \alpha$ we get that

$$\begin{aligned}\|f\|_\alpha^2 &= \sum_{i=0}^{\alpha-2} f^{(i)}(0)^2 + f^{(\alpha-1)}(0)^2 + f^{(\alpha)}(0)^2 + \int_0^1 f^{(\alpha+1)}(s)^2\,\mathrm{d}s \\ &\leq e\sum_{i=0}^{\alpha-2} f^{(i)}(0)^2 + e\Big(\sum_{i=\alpha-1}^{\beta} f^{(i)}(0)^2 + \int_0^1 f^{(\beta+1)}(z_t)^2\,\mathrm{d}z_t \Big) \\ &= e\|f\|_\beta^2.\end{aligned}$$

□

The norm $\|f\|_{\mathbb{H}^\alpha}$ given in (2.2) is equal to the norm $\|f\|_{\alpha-1/2}$ for all $f \in \mathbb{H}^\alpha$ and $\alpha \in \mathbb{N}$. Hence from Lemma 5.1 for all $\alpha < \beta$, where $\alpha - 1/2$, $\beta - 1/2 \in \mathbb{N}$

$$\sqrt{e}\mathbb{H}_1^\alpha \supseteq \mathbb{H}_1^\beta.$$

With λ_n defined in (4.6) the conditions of Theorem 4.1 hold, hence we get the minimax rate in contraction.

5.2 The Rescaled process

In the second example we use in the construction of the prior distribution the process

$$V_t^{n,\alpha} = (I_{0+}^k B)_{t/c_{n,\alpha}} + \frac{1}{\sqrt{a_{n,\alpha}}} \sum_{i=0}^{k} Z_i \frac{t^i}{i!},$$

where $a_{n,\alpha}$ and $c_{n,\alpha}$ are given in section 2.3. In section 2.3 and section 3.2 we have shown that for a known parameter $\alpha > 0$ we can estimate α-regular functions applying this process and we will achieve the minimax rate in the Bayesian procedure ($\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$). We would like to extend this procedure for unknown parameter α. For this we take first a k large enough constant and after that, with the help of Theorem 4.1, we show that for all unknown $\alpha \in (0, k+1]$ the posterior contraction rate is $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$. So we want to apply Theorem 4.1, but for this we have to check first that its conditions hold.

From the RKHS of the process $V_t^{c_{n,\alpha},a_{n,\alpha}}$ and the definition of the sequences $a_{n,\alpha}, c_{n,\alpha}$ one can see that for all $k \in \mathbb{N}_0$ and for all $\beta > \alpha$ $(\beta, \alpha \in (0, k+1])$ the inequalities

$$c_{n,\alpha} \leq c_{n,\beta},$$
$$a_{n,\alpha} \leq a_{n,\beta}$$

hold. Hence the condition

$$\mathbb{H}_1^\alpha \supseteq \mathbb{H}_1^\beta$$

is satisfied. From the proof of Theorem 2.10 one can see that (3.2) holds for $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$.

Finally we give the prior distribution on the models. The prior Π_n^α defined by the substitution of the process $V_t^{n,\alpha}$ into the prior distribution (1.1) gives the minimax contraction rate $\varepsilon_{n,\alpha} = n^{-\frac{\alpha}{2\alpha+1}}$, when $\log p_0 \in C^\alpha[0,1]$. With probability $\lambda_n(\alpha)$, defined in (4.6), we choose the prior Π_n^α, which was defined by the help of the process $V_t^{n,\alpha}$.

Now we can apply Theorem 4.1, since all of its conditions hold. The theorem says that for all $\beta \in (0, k+1]$ and $\log p_0 \in C^\beta[0,1]$ the hierarchical posterior convergence rate is $\varepsilon_{n,\beta} = n^{-\frac{\beta}{1+2\beta}}$, which is exactly the minimax rate.

A more general approach is to rescale the smooth Gaussian distribution with an independent random variable A. This random variable could be for example the gamma distribution. It is an extension of the example to derive a similar result to the process

$$V_t^\alpha = (I_{0+}^k B)_{Ct} + \frac{1}{\sqrt{A}} \sum_{i=0}^{k} Z_i \frac{t^i}{i!},$$

where A and C are independent random variables and are independent from the B Brownian motion also.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[2] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[3] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89, 2008.

[4] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007.

[5] James Kuelbs, Wenbo V. Li, and Werner Linde. The Gaussian measure of shifted balls. *Probab. Theory Related Fields*, 98(2):143–162, 1994.

[6] Jüri Lember and Aad van der Vaart. On universal Bayesian adaptation. *Statist. Decisions*, 25(2):127–152, 2007.

[7] Stefan G. Samko, Anatoly A. Kilbas, and Oleg I. Marichev. *Fractional integrals and derivatives*. Gordon and Breach Science Publishers, Yverdon, 1993. Theory and applications, Edited and with a foreword by S. M. Nikolskiĭ, Translated from the 1987 Russian original, Revised by the authors.

[8] 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.

[9] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 200–222. Inst. Math. Statist., Beachwood, OH, 2008.

[10] A.W. van der Vaart and J.H. van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.*, 1:433–448 (electronic), 2007.