Scalability and uncertainty of Gaussian processes

Hadji, M.A.

Scalability and uncertainty of Gaussian processes

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 25 januari 2023
klokke 11.15 uur

door

Mohamed Amine Hadji

geboren te Algiers
in 1993

Promotor:
Prof. dr. A.W. van der Vaart (Technische Universiteit Delft)

Co-promotor:
Dr. B.T. Szabó (Universitá Bocconi)

Promotiecommissie:
Prof. dr. F.A.J de Haas
Prof. dr. P.D. Grünwald
Prof. dr. I. Castillo (Sorbonne Université)
Dr. P.J. de Andrade Serra (Vrije Universiteit Amsterdam)
Dr. H.N. Kekkonen (Technische Universiteit Delft)

Contents

CHAPTER 1
Introduction

The main theme of this thesis is the theoretical study of Gaussian processes as a tool in Bayesian nonparametric statistics. We are interested in the frequentist properties of Bayesian nonparametric techniques in an asymptotic regime.

This chapter will give a general introduction to Bayesian nonparametric statistics. After a brief preamble on statistical science, we will compare the two main statistical approaches: frequentism and Bayesianism. Our work will consist of a combination of both, since we will study the Bayesian techniques from a frequentist perspective, specifically consistency, convergence rates, uncertainty quantification and adaptation. These properties will be studied in the context of non-parametric problems, that is to say models with few modeling constraints. We will focus on the non-parametric regression model and its idealized version, the signal-in-white-noise model.

One of the main theme of the thesis is scalability of Bayesian techniques. Indeed, these computation-hungry techniques rapidly become intractable as the number of observations grows. This issue led to the introduction of distributed Bayesian methods in order to decrease the computational complexity of the techniques. The later chapters of this thesis will focus on such distributed methods and their properties. They will also be briefly introduced in this chapter.

§1.1 Statistical science

Statistics can be defined as the study of data, which covers its collection, organization, analysis, interpretation and presentation. The main focus of this thesis will be mathematical statistics, which follows the framework of probability theory. Moreover, the theoretical results will be illustrated with simulated data.

In order to understand both probability theory and statistics, it is useful to start with a definition of modeling. A *model* describes and explains a system with the help of mathematical concepts and language. Even though models are helpful in most disciplines, they are all considered as oversimplification of the described system. As George Box (Box, 1976) famously said "all models are wrong, but some are useful". While in probability theory, one studies the behavior of data following a specific and known model, in theoretical statistics, one follows the opposite paths: one tries to gather as much information from the data in order to make inference on the model.

More formally, one can think of a probability distribution P as a model for generating an observation X. The classical statistical approach is to assume that our model $\mathcal{P}$ is a collection of probability distributions, and the end goal is to infer properties

of an element of $\mathcal{P}$ which could best describe the observation.

§1.2 Frequentist and Bayesian inference

Inference in statistics consists of establishing and evaluating propositions about the process by which the data is generated. Different paradigms of statistical inference coexist. This section will focus on two important paradigms: *frequentist inference* and *Bayesian inference*.

§1.2.1 Frequentist inference

Frequentist inference is associated with the frequentist interpretation of probability; we consider that the probability of an event is the frequency of appearance of this event if it were possible to repeat the same experiment independently ad infinitum. One of the main characteristics of frequentism is that frequentist modeling views the data X as the realization of a random variable following some fixed probability distribution P_0 belonging to some model $\mathcal{P}$. Often, the statistical model $\mathcal{P}$ is defined as

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

where the probability distributions are indexed by a parameter θ and Θ is the set of all possible parameters, and we suppose the existence of a θ_0 such that $P_0 = P_{\theta_0}$. Basically, if θ_0 were known, the work of a frequentist statistician would be finished.

Among the most widely used tools of frequentist inference, one can think of the likelihood function. The *likelihood function* is a function of the parameter which measures how well a particular parameter of the model fits a given sample of data. Heuristically, it makes sense that parameters whose likelihoods are large are preferable, since they verify that the probability of the observed data occurring is high.

§1.2.2 Bayesian inference

The Bayesian approach is quite different from the frequentist one. It is based on the Bayesian interpretation of probability which stipulates that a probability is merely a quantification of a personal belief or a reasonable expectation. Bayesian statisticians generally believe that there is no fixed underlying parameter, but that the parameter itself is the realization of a random variable. More formally, the parameter θ has a probability distribution Π on the parameter space Θ called the prior distribution. This distribution represents the degree of belief the statistician attaches to possible parameters explaining the model. Then, the distribution P_θ describes the distribution of the data X conditionally on the value of the parameter θ. This belief is then updated once we have access to the data. If the model is dominated, which means all possible distributions P_θ are absolutely continuous with respect to a common measure μ and have corresponding densities p_θ, then this operation gives us the posterior distribution of the parameter

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

for a measurable set $A \subset \Theta$. The integrand in the numerator of the previous fraction represents the joint distribution of the parameter and the data, which can be obtained by multiplying the prior distribution of the parameter by the likelihood function, while the denominator represents the marginal distribution of the data.

Even though the Bayesian approach is quite different from the frequentist approach, it can be useful in certain situations. Indeed, Bayesian techniques benefit from conceptual simplicity in practice; after assigning a prior to the parameter, one needs only to compute the posterior to make relevant inference. Besides, previous expertise on a problem can be incorporated in the prior distribution of a parameter. Although this implied subjectivity can be seen as a drawback compared to the frequentist approach, subjectivity can never be completely absent from statistical modeling and Bayesian subjectivity might appear more natural. Moreover, once the statistician has access to the posterior distribution, all inference can be done exhaustively. Finally, from a decision theoretical point of view, there is a Bayesian procedure (or a limit of Bayesian procedures) which performs uniformly better than any other procedure according to the "complete class theorem", and only those procedures are admissible, see for instance (Balder et al., 1983) and references therein. For a more exhaustive introduction to Bayesian methods, see the books (Bernardo and Smith, 1994), (Robert, 2001), and (Ghosal and Van der Vaart, 2017).

One of the advantages of the Bayesian framework is the simplicity with which we can quantify uncertainty. As the parameter is a random variable whose posterior distribution depends on the data, we can construct sets containing the parameter with high posterior probability. It is interesting to note that this is a probability statement about the unknown parameter; thus, those sets are quite different from frequentist confidence sets which provide a probability statement about the sets themselves.

After the introduction of both paradigms, the following section will be focused on a hybrid approach where Bayesian methods are used, but given a frequentist interpretation.

§1.2.3 Frequentist Bayes

Since Bayesian inference always starts with the choice of a prior which represents subjective belief of an expert, if two experts do not share the exact same subjective belief, the outcomes of their analyses can vary even if they shared the methodology. This situation seldom occurs within the frequentist community, because of the shared assumption that there exists a single true distribution explaining how the data has been generated. Although the Bayesian paradigm differs significantly from the frequentist one, it can be interesting to see Bayesian inference as a mere different frequentist method. Indeed, one can still assume that the data X follows a fixed probability distribution P_0 belonging to some model $\mathcal{P}$, and then study the posterior distribution $\Pi(\cdot|X)$ of the parameter θ as a random measure which depends on the prior and the "true" distribution P_0. On one hand, this hybrid view allows statisticians to enjoy the relative simplicity of implementation of Bayesian methods, for instance the built-in uncertainty quantification; on the other hand, it gives frequentist guarantees for those same methods.

In the parametric case, where the parameter θ describing the model has a finite dimensional structure, the well-known Bernstein-von Mises theorem establishes that

under mild regularity conditions Bayesian inference is asymptotically equivalent to frequentist inference. More precisely, the theorem states that the posterior becomes asymptotically normal centered around the maximum likelihood or any efficient estimator with variance equal to the inverse of the Fisher information. Despite the strength of this result, its limitation to the parametric and semi-parametric case requires statisticians to find other ways to study non-parametric Bayesian methods.

§1.3 Non-parametric asymptotic statistics

As established, non-parametric statistical techniques are based on models with as few assumptions as possible. Technically, a *non-parametric model* does rely on a parameter θ, but this parameter is infinite dimensional; it can for instance be an infinite sequence or a function. Non-parametric models are attractive thanks to their flexibility and robustness. The books (Wasserman, 2006), (Tsybakov, 2009) and (Giné and Nickl, 2016) provide a more detailed introduction to non-parametric statistics.

In non-parametric statistics, it is common to consider some assumption on the smoothness or regularity of the parameter $\theta \in \Theta$. In this thesis, we consider frequently used smoothness spaces: the Sobolev and the Hölder spaces. These spaces will be defined for functions defined on $[0,1]$, but the idea can be expanded to functions on other domains as well.

Let $(\phi_i)_{i\geq 1}$ denote an orthonormal basis of $L^2([0,1])$ (i.e. measurable functions that are square-integrable in $[0,1]$) and consider functions on $L^2([0,1])$ in the form

$$\theta = \sum_{i=1}^{+\infty} \theta_i \phi_i,$$

where θ_i is the ith coefficient in the series expansion. The Sobolev ball $S^\beta(M)$ with smoothness parameter $\beta > 0$ and radius $M > 0$ is defined as

$$S^\beta(M) = \left\{ \theta \in L^2([0,1]) : \sum_i \alpha_i \theta_i^2 \leq M^2 \right\},$$

where $\alpha_i \asymp i^{2\beta}$. In this thesis, we will also consider a related regularity space of function: hyper-rectangles

$$\Theta^\beta(M) = \{\theta \in L^2([0,1]) : \sup_i i^{2\beta}\theta_i^2 \leq M\}.$$

For integer regularity parameters β and the classical Fourier basis, it is also possible to represent the Sobolev class as the set of $\beta - 1$ times differentiable functions $[0,1]$ with absolute continuous derivatives and with a βth derivative satisfying

$$\int_0^1 (\theta^\beta(t))^2 dt \leq M\pi^{2\beta}.$$

The Hölder space of smoothness $\beta > 0$ on $[0,1]$ denoted by $C^\beta([0,1])$, on the other hand, contains all functions θ with $\lfloor\beta\rfloor$ (the integer part of β) derivatives and their last continuous derivatives verifies the Hölder condition

$$\sup_{x,y\in[0,1]} |\theta^{\lfloor\beta\rfloor}(x) - \theta^{\lfloor\beta\rfloor}(y)| \leq M|x-y|^\alpha$$

with $\alpha = \beta - \lfloor\beta\rfloor$ and $M > 0$. The corresponding Hölder ball with radius $M > 0$ is defined as

$$H^\beta(M) = \left\{ \begin{array}{ll} \theta \in C^\beta([0,1]) : & \sup_{x,y\in[0,1]} |\theta^{\lfloor\beta\rfloor}(x) - \theta^{\lfloor\beta\rfloor}(y)| \leq M|x-y|^{\beta-\lfloor\beta\rfloor}, \\ & \sup_{k\leq\beta} \|\theta^{(k)}\|_\infty \leq M \end{array} \right\}.$$

The asymptotic approach to assess the performance of a statistical technique can be described intuitively. Indeed, the outcomes of any statistical procedure can be arranged in a sequence indexed by the size of the sample used during this procedure. The theoretical study of this sequence allows to understand the behavior of the procedure when the size of the data grows, and ideally the larger our data sample is, the better the procedure performs. Mathematically, if we observe the data $X^{(n)}$, generated from the distribution P_{θ_0}, and our outcome is a random variable called point estimator $T(X^{(n)})$, given by a measurable function T, then we are interested in the study of the sequence $(T(X^{(n)}))_{n\in\mathbb{N}}$. Since this thesis focuses on the frequentist Bayes approach, we will introduce some frequentist properties of Bayesian asymptotics: consistency, contraction rates and coverage of credible sets.

§1.3.1 Consistency

The first frequentist property one could check when applying Bayesian techniques is the consistency of the posterior. *Posterior consistency* simply means that the posterior distribution $\Pi(.|X^{(n)})$ puts most of its mass around a smaller and smaller neighborhood of the true parameter θ_0. Formally, it means that for all $\varepsilon > 0$,

$$\Pi(\theta : d(\theta, \theta_0) < \varepsilon | X^{(n)}) \overset{P_{\theta_0}}{\to} 1,$$

where d is a given metric and the convergence is in probability under the true parameter. Posterior consistency is a necessary frequentist property to assess the performance of a Bayesian method, but it is relatively weak and non-informative.

In non-parametric models, posterior consistency is not always satisfied even if the prior distribution covers the true parameter; it is possible that some posterior distribution does not concentrate around the true parameter even when any neighborhood of this parameter has positive mass, see for instance (Freedman, 1963), (Diaconis and Freedman, 1986) and (Kim and Lee, 2001). Fortunately, Doob's theorem and Schwartz's theorem provide a robust result to assess the consistency of non-parametric Bayesian models. Doob's consistency theorem (Doob, 1949) states that if the true parameter θ_0 is not in some null-set of the prior, then the posterior distribution is consistent. While topologically, prior null-sets can be large, Schwartz's theorem (Schwartz, 1965) gives a more extensive result. It assesses that posterior consistency over the whole parameter space is verified under two conditions: the prior mass condition which requires that the prior assigns sufficiently large probabilities to neighborhoods of the true parameter, and the test condition which necessitates the existence of a sequence of tests which separates the true parameter from the complements of neighborhoods around the parameter.

Although consistency is important in evaluating non-parametric Bayesian techniques, since it is not considered strong enough, we will focus on a more descriptive property: rate of contraction.

§1.3.2 Contraction rates

When estimating a parameter θ, the frequentist asymptotic way to evaluate the estimation $\hat{\theta}_n$, which is a measurable function of the data, is not only to verify that the sequence $(\hat{\theta}_n)_{n\in\mathbb{N}}$ converges (in probability or almost surely) to the true value θ_0 uniformly over the possible "truths", but also to evaluate the risk of this estimator. Mathematically, we first choose a loss function $L : \Theta \times \Theta \to (0, \infty)$ and then define the maximum risk of our estimator as

$$r(\hat{\theta}_n) = \sup_{\theta\in\Theta} E_\theta L(\hat{\theta}_n, \theta),$$

where the expectation E_θ is taken with respect to the probability measure P_θ. This function represents the amount an estimator deviates from the true parameter in the worst case scenario. The choice of a loss function impacts directly the choice of the best estimator. Once the maximum risk over the class Θ has been defined, the goal is to find an estimator which minimizes this risk. Such an estimator is called the minimax estimator, and the risk associated to it is the minimax risk

$$R_n = \inf_{\hat{\theta}_n} r(\hat{\theta}_n),$$

where the infimum is taken over all measurable functions of the data. Thanks to Markov's and Chebyshev's inequality, it can be shown that the minimax estimator $\hat{\theta}^*_n$ has the best convergence rate $\varepsilon_n^2 := R_n$, which means that

$$\sup_{\theta\in\Theta} P_\theta \left(d(\hat{\theta}^*_n, \theta) > M_n \varepsilon_n \right) \to 0,$$

for any $M_n \to +\infty$, and that for all δ_n such that $\delta_n = o(\varepsilon_n)$ there exists at least one $\theta^* \in \Theta$ such that

$$P_{\theta^*} \left(d(\hat{\theta}^*_n, \theta^*) > L\delta_n \right) \nrightarrow 0$$

for all $L > 0$. The minimax risk and estimators have been studied for a large variety of problems. We refer to (Lehmann and Casella, 1998) for more details about the topic.

The analogous property to convergence rate for Bayesian techniques is the rate of contraction. The *contraction rate* ε_n of a Bayesian procedure is a sequence converging to 0 in n such that

$$\Pi_n \left(\theta : d(\theta, \theta_0) \leq \varepsilon_n | X^{(n)} \right) \overset{P_{\theta_0}}{\to} 1.$$

This rate quantifies how quickly the posterior mass concentrates around the true parameter θ_0. In regular parametric models, the posterior contraction rate can always achieve the same order of the optimal frequentist convergence rate. This result follows from the Bernstein-von Mises theorem. However, it is not yet clear if such a result could also exist in the non-parametric case.

(Ghosal et al., 2000), (Ghosal and van der Vaart, 2007), (van der Vaart and van Zanten, 2008) and (Ghosal and Van der Vaart, 2017) address this issue. Let us consider the iid case where $X_1, ..., X_n$ are n random variable distributed according to P_{θ_0}. We consider the Hellinger metric d on the parameter space given by

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

where p_θ and $p_{\theta'}$ are the densities of P_θ and $P_{\theta'}$ with respect to the measure μ. We also define the entropy number (denoted by $N(\varepsilon, B, d)$), which represents the number of ε-radius balls required to cover the set B with respect to a given metric d.

Theorem 1.3.1 (Theorem 2.1 of (Ghosal et al., 2000))**.** *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and subsets $\Theta_n \subset \Theta$ of the parameter set, we have*

$$\begin{aligned}
&\log N(\varepsilon_n, \Theta_n, d) \leq n\varepsilon_n^2, \\
&\Pi(\theta \notin \Theta_n) \leq \exp(-(C+4)n\varepsilon_n^2), \\
&\Pi\Big(\theta : P_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}} \leq \varepsilon_n^2, P_{\theta_0} \log^2 \frac{p_\theta}{p_{\theta_0}} \leq \varepsilon_n^2\Big) \geq \exp(-Cn\varepsilon_n^2).
\end{aligned}$$

Then for sufficiently large M, we have that $\Pi(\theta : d(\theta, \theta_0) \geq M\varepsilon_n | X^{(n)}) \to 0$ in P_{θ_0}-probability.

The conditions can be understood as follows: the first condition, the entropy condition, implies that the sets Θ_n are not too large; the second condition, the remaining mass condition, implies that prior distribution puts enough mass on these sets; the last condition, the prior mass condition, requires the prior to put enough mass in a small neighborhood of the true distribution.

We consider that the contraction rate is optimal if it is equal to the minimax convergence rate. It is relatively easy to construct an estimator $\hat{\theta}$ from the posterior which attains the same convergence rate as the posterior construction. Indeed, we can for instance take the center of the smallest ball with posterior mass greater than $1/2$, see (Ghosal et al., 2000) and (Ghosal and Van der Vaart, 2017). Therefore, finding a posterior achieving optimal contraction rate leads directly to a Bayesian estimator with minimax convergence rate.

§1.3.3 Bayesian uncertainty quantification

Once the posterior distribution is computed, one can derive Bayesian point estimates. Those are simply functions of the data which minimize an integrated loss function with respect to the posterior measure. Despite their usefulness, point estimates do not provide information about the uncertainty of our inference on θ_0. For that reason, it is also interesting to generate a set of possible values for our parameter.

In the frequentist case, these sets are known as confidence sets. *Confidence sets* are set-valued estimators $\hat{C}_n(X^{(n)})$ to which the true parameter is most likely to belong, according to the data. More formally, we define a confidence set $\hat{C}_n$ of level $1-\alpha$ as

$$\inf_{\theta \in \Theta} P_\theta(\theta \in \hat{C}_n) \geq 1 - \alpha.$$

Asymptotically, confidence sets should also verify

$$\liminf_{n \to \infty} \inf_{\theta \in \Theta} P_\theta(\theta \in \hat{C}_n) \geq 1 - \alpha.$$

These sets have asymptotically uniform confidence level $1-\alpha$. In other words, they have the same confidence level asymptotically even in the worst case scenario for the true parameter.

Nonetheless, the construction of confidence sets is not easy when the model is computationally complex. In the Bayesian approach, we talk instead of *credible sets*, which accumulate most of the posterior probability mass. Since it is possible to sample from the posterior, the construction of approximate credible sets thanks to Markov chain Monte-Carlo (MCMC) methods is relatively simpler than the construction of confidence sets. Formally, credible sets $\tilde{C}_n$ of level $1-\alpha$ are sets such that

$$\Pi_n(\theta \in \tilde{C}_n | X^{(n)}) \geq 1 - \alpha.$$

Contrary to confidence sets, credible sets are subjective in nature because of the influence of the prior distribution in the posterior. Although confidence sets and credible sets are asymptotically equivalent in the parametric case under some regularity assumptions thanks to the Bernstein-von Mises theorem, the question of their equivalence in the non-parametric case has not been entirely answered.

Actually, it is known through several negative examples that Bayesian credible sets do not converge automatically to frequentist confidence sets when the model is infinite-dimensional, see for instance (Cox, 1993). This phenomenon encourages a further analysis of the behavior of credible sets from a frequentist point of view. It appears in (Knapik et al., 2011), (Castillo and Nickl, 2013), (Szabo et al., 2015), (Sniekers and van der Vaart, 2015b), (Belitser, 2017) and (Szabo et al., 2017) that in some situations, Bayesian credible sets ensure good coverage of the true parameter, that is to say that the credible sets are asymptotically likely to contain the true parameter according to the true distribution. Moreover, (Rousseau and Szabo, 2020) studied the asymptotic coverage of Bayesian credible sets in a general non-parametric framework. The general intuition is that Bayesian credible sets have good coverage when the prior generates parameters that are slightly less smooth than the true parameter.

§1.4 Gaussian processes

The examination of Bayesian techniques from a frequentist perspective requires a good understanding of the prior distribution because it influences heavily the outcome of a Bayesian statistical procedure. In this work, we mainly treat Gaussian process priors. *Gaussian processes* are stochastic processes, i.e. collections of random variables indexed by time or space which can be viewed as random elements in a function space verifying the following condition: every finite collection of the evaluation of the process at different times has a multivariate normal distribution. We refer to (Rasmussen and Williams, 2006) for a more comprehensive introduction to Gaussian processes and their applications.

A Gaussian process $W = \{W_t : t \in T\}$ indexed by a set T is characterized by its mean $\mu : T \to \mathbb{R}$ and its covariance function $K : T \times T \to \mathbb{R}$ given by $\mu(t) = EW_t$ and $K(s,t) = \mathrm{Cov}(W_t, W_s)$, respectively. Generally, the set T is a subset of $\mathbb{R}^d$ so that the function $t \mapsto W_t$ belongs to the space of real-valued functions on T. As priors, we typically take centered Gaussian processes, i.e. their mean $\mu(t)$ is set to be zero; thus, their behaviors can be understood entirely through their covariance functions. Moreover, thanks to Mercer's theorem (Mercer, 1909), we can see that the covariance

function of a Gaussian process can be represented as follows

$$K(s,t) = \sum_{j=1}^{+\infty} \lambda_j \phi_j(s)\phi_j(t),$$

because the covariance is a symmetric non-negative definite kernel. Here, the ϕ_j's are eigenfunctions representing an orthonormal basis of $L^2(T)$, the λ_j's are the corresponding non-negative eigenvalues and the convergence of the series is absolute and uniform when the kernel is continuous and T is compact. In addition to Mercer's theorem, the Karhunen-Loève theorem, (Karhunen, 1947) and (Loève, 1978), provides a nice representation of the random variable W_t for any $t \in T$:

$$W_t = \sum_{j=1}^{+\infty} \sqrt{\lambda_j} Z_j \phi_j(t),$$

where the Z_j's are independent standard normal random variables and the convergence is almost sure.

In this thesis, we will focus on the following specific Gaussian processes (GP):

- the GP with a **Matérn covariance kernel**

  $$K(s,t) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} \Big(\frac{|s-t|\sqrt{2\alpha}}{a} \Big)^{\alpha} K_\alpha \Big(\frac{|s-t|\sqrt{2\alpha}}{a} \Big),$$

  where Γ is the gamma function, K_α is the modified Bessel function, a is a rescaling parameter and α is the smoothness parameter. The sample paths $t \mapsto W_t$ of a GP with Matérn covariance function are $\lfloor \alpha \rfloor$ times differentiable.

- the GP with a **squared-exponential covariance kernel**

  $$K(s,t) = \exp\Big(\frac{(s-t)^2}{2a^2} \Big),$$

  where a is a scaling parameter. The squared-exponential covariance kernel can be seen as the limit of Matérn covariance kernels with α going to infinity. A GP with squared-exponential covariance function has infinitely differentiable sample paths.

- the **Brownian motion and its primitives**, the Brownian motion has covariance kernel

  $$K(s,t) = \min(s,t).$$

  Since the sample paths of a Brownian motion are Lipschitz continuous of any order $\alpha < 1/2$ but not differentiable anywhere, it is common to integrate this process k times in order to have a GP with a smoothness order $k + 1/2$. Moreover, the k-fold integrated Brownian motion has vanishing derivatives at zero; hence, it is common to "release" this process by adding a polynomial process $t \mapsto \sum_{j=0}^{k} Z_j t^j / j!$ where the Z_j's are independent standard normal random variables. Thus, the k-fold integrated Brownian motion can be written as

  $$G_t = \sum_{j=0}^{k} \frac{Z_j t^j}{j!} + I^k W_t,$$

where $I^k = I^{k-1}I$ and $If(t) = \int_0^t f(s)ds$.

Both the GP with Matérn and with squared-exponential covariance kernel are stationary; $\mathrm{Cov}(W_t, W_s)$ only depends on the distance between the two points t and s, and not on their position. On the other hand the Brownian motion, while not stationary, has independent and stationary increments, which means the consecutive increments of this process over the same distance are iid random variables.

Gaussian processes appear to be good priors of choice in the statistical models studied in this thesis. For more detailed investigations on these processes, we give a non-exhaustive list of work and references therein: (Tokdar and Ghosal, 2005), (Choudhuri et al., 2007), (van der Vaart and van Zanten, 2007), (van der Vaart and van Zanten, 2008) and (van der Vaart and van Zanten, 2009b).

§1.5 Models

The main non-parametric models investigated in this thesis are the signal in Gaussian white noise model and the non-parametric regression model.

§1.5.1 Signal-in-white-noise model

Let the random function $Y^{(n)}(t)$ be defined as follows

$$Y^{(n)}(t) := \int_0^t \theta_0(s)ds + \frac{1}{\sqrt{n}}W_t, \qquad t \in [0,1],$$

where $\theta_0 \in L^2[0,1]$ is the parameter of interest and W_t is the Brownian motion. This problem can also be interpreted as observing n independent realizations of a random variable defined as

$$Y_j(t) := \int_0^t \theta_0(s)ds + W_{j,t}, \qquad j \in \{1, ..., n\}, \qquad t \in [0,1],$$

where the $W_{j,t}$ are n independent Brownian motions.

It is usual for practicality to convert the problem into the spectral domain using an orthonormal basis $(\phi_i)_{i\geq 1}$ of $L^2[0,1]$. The random function $Y^{(n)}(t)$ becomes a sequence $(Y_i)_{i\geq 1}$ defined as

$$Y_i = \theta_{0,i} + \frac{1}{\sqrt{n}}Z_i, \qquad i \in \mathbb{N},$$

where $(\theta_{0,i})_{i\geq 1} \in \ell^2$ is an infinite sequence and $(Z_i)_{i\geq 1}$ are independent standard normal random variables. The relative simplicity of the model and its relation to non-parametric regression (Brown and Low, 1996) and (Nussbaum, 1996) makes it an important benchmark model in the literature; see for instance (Donoho, 1994), (Tsybakov, 2009) and (Giné and Nickl, 2016).

In the case $\theta_0 \in S^\beta$ with regularity parameter $\beta > 0$, the minimax convergence rate to estimate the parameter is proportional to $n^{-\beta/(1+2\beta)}$ with respect to the L_2-loss; see (Tsybakov, 2009) and reference therein. One way to achieve this rate is to

simply estimate the Fourier coefficients of θ_0 with the first $n^{1/(1+2\beta)}$ coefficients of $(Y_i)_{i\geq 1}$. Furthermore, it has also been shown in (Knapik et al., 2011) and (Ghosal and Van der Vaart, 2017) that using an appropriately smooth GP prior (for instance a squared-exponential GP with rescaling $a_n := n^{1/(1+2\beta)}/\log^{2/(1+2\beta)} n$) leads to optimal posterior contraction rates (up to a multiplicative logarithmic factor).

§1.5.2 Non-parametric regression

Let $(X_i, Y_i)_{i=1}^n$ be n iid pairs of random variables such that

$$Y_i = \theta_0(X_i) + Z_i, \quad i \in \{1, ..., n\}$$

where the X_i's are in a set $\mathcal{X}$ (e.g. $[0,1]$), the Z_i's are iid centered random variables (generally taken as standard normal) and $\theta_0 \in L^2$ is the functional parameter of interest. When X_i is deterministic, we talk about regression with fixed design, for instance the regular grid $X_i = i/n$ on $\mathcal{X} = [0,1]$, while in the case of random X_i, we talk about regression with random design. Generally, θ_0 is assumed to belong to a certain class of smoothness with regularity $\beta > 0$, for instance a Sobolev or Hölder class. The corresponding minimax rate for both classes is, up to a constant multiplier, $n^{-\beta/(1+2\beta)}$ with respect to the empirical L_2-norm in the fixed design case and the usual L_2-norm in the uniformly random design case.

Among the estimators achieving the minimax convergence rate, we can present the Nadaraya-Watson estimator introduced in (Nadaraya, 1964) and (Watson, 1964)

$$\hat{\theta}_n(x) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^n K(\frac{x-X_i}{h})},$$

which uses a kernel K, a non-negative integrable function, as a weight function and $h > 0$ is a smoothing parameter called the bandwidth. In (Bickel and Doksum, 1977), we can see that if the bandwidth is well chosen, namely $h \asymp n^{-1/(2\beta+1)}$, then both the point-wise and integrated mean-square errors are proportional to the optimal convergence rate $n^{-\beta/(1+2\beta)}$. In this model as well, (van der Vaart and van Zanten, 2008), (Knapik et al., 2011), (Ghosal and Van der Vaart, 2017) and (Bhattacharya et al., 2017) have shown that using a GP prior can result in a posterior density which contracts around the true parameter at an optimal rate in L_2-norm up to a possible logarithmic factor.

§1.6 Adaptation

In the previous section, we have seen that some estimators are minimax for the discussed models. However, these estimators require the knowledge of the exact regularity of the true parameter. In practice, this might not be realistic since the true regularity is seldom known in advance; thus, it is desirable to have procedures which do not rely on the true smoothness hyper-parameters, but adapt to them. In mathematical words, we can consider that the parameter of interest θ_0 belongs to a parameter space Θ seen as a collection of subspaces Θ^β indexed by a hyper-parameter

β. We would like to find an estimator $\hat{\theta}_n \in \Theta$ which attains the minimax rate $R_{n,\beta}$ corresponding to Θ^β whatever the true smoothness β is

$$\sup_{\theta \in \Theta^\beta} E_\theta L(\hat{\theta}_n, \theta) \leq C_\beta R_{n,\beta},$$

where $C_\beta > 0$ is a positive constant depending only on β and $R_{n,\beta}$ is defined as

$$R_{n,\beta} = \inf_{\hat{\theta}_n \in \Theta^\beta} \sup_{\theta \in \Theta^\beta} E_\theta L(\hat{\theta}_n, \theta).$$

Although the theory of adaptation is fairly developed in frequentist statistics; see for instance (Bickel, 1982), (Lepski and Spokoiny, 1997), and (Goldenshluger and Nemirovski, 1997), its Bayesian counterpart is relatively recent, and has been getting more attention, see the monograph (Ghosal and Van der Vaart, 2017) and reference therein.

The idea is to construct a prior distribution reaching the optimal contraction rate for a set of possible regularities. In the non-parametric case, the oracle choice, the tuning parameter of the prior associated with the true smoothness, is impossible to select in practice. This encourages us in most cases to make a data-driven selection of this parameter. The two adaptive Bayesian techniques discussed in this thesis are the empirical Bayes and the hierarchical Bayes methods.

§1.6.1 Empirical Bayes

One possible way to proceed is to estimate the tuning parameter from the data. Although not fully Bayesian, since the parameter of interest is estimated using frequentist techniques, this method can be computationally convenient in some statistical settings. This approach is known as the empirical Bayes method in the literature. Basically, if the prior distribution is tuned by a parameter α such that we can denote it by Π_α, we first need to estimate this parameter by maximizing the marginal likelihood $\int_\Theta P_\theta(X) d\Pi_\alpha(\theta)$ seen as a function of α to obtain an estimator $\hat{\alpha}$ with X being our observation, then plug $\hat{\alpha}$ into the posterior used in our inference. In other words, the maximum marginal likelihood estimator (MMLE) is

$$\hat{\alpha} = \arg\max_\alpha \int_\Theta P_\theta(X) d\Pi_\alpha(\theta),$$

and the corresponding posterior would be $\Pi_{\hat{\alpha}}(\cdot|X) = \Pi_\alpha(\cdot|X)\Big|_{\alpha=\hat{\alpha}}$.

We consider the following toy-example to demonstrate the idea. Let $X^{(n)}$ be a sample of n iid observation from a Bernoulli distribution with unknown mean θ. We endow θ with a Beta prior with parameters $a, b > 0$, which are also unknown. We will first estimate the hyper-parameter a and b by maximizing the marginal likelihood

$$(\hat{a}_n, \hat{b}_n) = \arg\max_{a,b>0} \frac{1}{B(a,b)} \theta^{a+\sum_i X_i - 1} (1-\theta)^{b+n-\sum_i X_i - 1},$$

where $B(a,b)$ is the normalizing Beta function. Then, we plug $\hat{a}_n$ and $\hat{b}_n$ in the posterior distribution of the parameter θ

$$\Pi_{\hat{a}_n, \hat{b}_n}(\cdot|X^{(n)}) = \Pi_{a,b}(\cdot|X^{(n)})\Big|_{a=\hat{a}_n, b=\hat{b}_n} \sim \text{Beta}\big(\hat{a}_n + \sum_i X_i, \hat{b}_n + n - \sum_i X_i\big).$$

The approach benefits from the fact that once the estimators of the hyper-parameters are computed, the posterior distribution generally becomes simple to compute. Furthermore, the choice of the hyper-parameters makes sense intuitively for the frequentist community, since it removes part of the subjectivity tied to the prior. For a more comprehensive overview of the use of Empirical Bayes in practice and theoretical properties thereof, see (Johnstone and Silverman, 2005), (Belitser and Enikeeva, 2008), (Jiang and Zhang, 2009) (Szabo et al., 2013) and (Rousseau and Szabo, 2017).

§1.6.2 Hierarchical Bayes

On the other hand, the hierarchical Bayes method is more appealing to Bayesian statisticians. In this approach we treat the hyper-parameter α as a random variable, similarly to the parameter θ, and we endow it with a hyper-prior distribution. Formally,

$$X|\theta \sim P_\theta, \quad \theta|\alpha \sim \Pi_\alpha, \quad \alpha \sim \Lambda,$$

where Λ is a hyper-prior distribution, Π_α is the prior distribution of the parameter of interest θ conditionally on α and P_θ is the distribution of our data conditionally on the parameter θ. This creates a multilevel, hierarchical Bayesian procedure. We are then interested in the marginal posterior distribution $\Pi(\cdot|X) = \int \Pi_\alpha(\cdot|X)d\Lambda(\alpha)$. This fully Bayesian method has been studied in different models, and it has been shown in, among other papers, (Huang, 2004), (Lember and van der Vaart, 2007), (de Jonge and van Zanten, 2009), (van der Vaart and van Zanten, 2009a) and (Arbel et al., 2013) that if the hyper-prior is chosen appropriately, the posterior distribution contracts at an optimal rate adaptively.

Considering our previous toy-example, in a hierarchical Bayes procedure, we would put a hyper-prior on $a, b > 0$, for instance two independent Gamma distributions with parameters $k, \gamma > 0$. Formally, we would have

$$X^{(n)}|\theta \sim \mathrm{Ber}(\theta), \quad \theta|a,b \sim B(a,b), \quad a, b \overset{\text{ind}}{\sim} \Gamma(k,\gamma).$$

In some cases, it is possible to compute the marginal posterior $\theta|X^{(n)}$ straightforwardly; however, in most cases, only approximation techniques (like MCMC methods) can be applied.

One of the reasons hierarchical Bayes procedures are popular is that they ensue directly from the Bayesian philosophy. Indeed, once a suitable sampling method is found for the parameters of interest, most questions about the data-generating process can be answered from a Bayesian point of view.

§1.7 Distributed computation

An asymptotic assessment of a statistical procedure requires the size n of the data to be large; paradoxically, the optimal procedures suffer from having larger and larger data set, consequently increasing computation time. On one hand, a larger sample size is appreciated because it leads to more precise statistical statements, and on the other hand, increasing the number of observations results in more computational burden. Among the statistical procedures showcasing this phenomenon, we can give

the GP non-parametric regression as an example. In this statistical model, we endow the functional parameter θ with a GP prior $\theta \sim \mathrm{GP}(0, K)$ where K is a covariance kernel. When the noise is also Gaussian with $(Z_i)_{i=1}^n \overset{\text{ind}}{\sim} N(0, \sigma^2)$, the corresponding posterior becomes easy to compute and we have $\theta|(X_i, Y_i)_{i=1}^n \sim \mathrm{GP}(\hat{\theta}, \hat{K})$, where the parameters $\hat{\theta}$ and $\hat{K}$ have closed forms:

$$\hat{\theta}(x) = K(x, \mathbf{X})^T (\mathbf{K} + \sigma^2 I_n)^{-1} \mathbf{Y}, \tag{1.7.1}$$

$$\hat{K}(x, x') = K(x, x') - K(x, \mathbf{X})^T (\mathbf{K} + \sigma^2 I_n)^{-1} K(x', \mathbf{X}), \tag{1.7.2}$$

where $\mathbf{Y} = (Y_i)_{i=1}^n$, $K(x, \mathbf{X}) = (K(x, X_i))_{i=1}^n$ and $\mathbf{K} = (K(X_i, X_j))_{i,j=1}^n$ for all $x, x' \in \mathcal{X}$ where $\mathcal{X}$ is a compact space (generally $[0,1]^d$). Even though it is possible to compute the parameters of the posterior directly, they require the inversion of a $n \times n$ matrix, which scales the computational complexity to $O(n^3)$ and requires a memory of order $O(n^2)$.

Against the background of the high computational complexity, methods based on distributed computation have emerged. *Distributed methods* partition the data over several machines called experts, see for instance (Jordan and Jacobs, 1994), (Minsker et al., 2014), (Ng and Deisenroth, 2014), (Cao and Fleet, 2014), (Srivastava et al., 2015) and (Scott et al., 2016). The experts process their share of the data locally, solving smaller versions of the problem. Then, all the local results are aggregated on a central machine to produce a final outcome of the statistical analysis. To formalize these ideas, let us partition the data $(X_i, Y_i)_{i=1}^n$ into m batches $(X_i^j, Y_i^j)_{i=1}^{n_j}$ with $j \in \{1, ..., m\}$, such that $\sum_j n_j = n$. So as to keep the simplicity of the GP non-parametric regression problem, we assume that each machine endows the parameter with a GP prior. Once the local posterior distributions $\theta|(X_i^j, Y_i^j)_{i=1}^{n_j}$ are computed, we can sample from the global posterior by aggregating each sample from the local posterior distributions. There exist different types of distributed computation methods based on the manner the data is partitioned, the way the local posterior or its modification is computed and the aggregation technique used in the process. We provide here a short description of these methods, but more details will be given in Chapters 3 and 4.

§1.7.1 Uniformly random partitioning

The data can be partitioned uniformly randomly among the machines. For simplicity, we assume that $n \equiv 0 \bmod m$. Each machine will receive n/m data points chosen randomly among $(X_i, Y_i)_{i=1}^n$ such that no data point simultaneously belongs to two machines or more.

Although in the classical Bayesian approach, the posterior is merely proportional to the likelihood multiplied by the prior, it is sometimes beneficial in distributed computation to modify the local posteriors. For instance, in order to tone down the effect of the prior, it can be useful to raise the prior to a power decreasing in m, the number of machines, for instance $1/m$. Another technique would be to increase the effect of the likelihood by raising it to a higher power, for instance m.

Furthermore, the aggregation of the local posteriors also affects the quality of the global posterior. In the case the data has been partitioned uniformly randomly, one

can either simply average the local posterior distributions or compute their Wasserstein barycenter, which is based on the Wasserstein distance between probability measures.

§1.7.2 Spatial partitioning

It is also possible to partition the data spatially. In this case, we partition the design space $\mathcal{X}$ into m sub-regions called $\mathcal{D}_j$ with $j \in \{1, ..., m\}$, and the jth machine will deal with the data points $\{(X_i, Y_i)_{i=1}^n, X_i \in \mathcal{D}_j\}$.

In this scenario, we will see that no modification of the local posteriors is needed. Moreover, even though each machine only receives observations with X_i in $\mathcal{D}_j$, it can produce a posterior mean and covariance for all $x \in \mathcal{X}$. Then, it is possible to draw from the local posterior and compute a weighted average of these draw, subsequently constructing a global posterior. Mathematically, if θ_j is a sample from the local posterior Π_j, then

$$\theta = \sum_{j=1}^{m} \omega_j \theta_j,$$

is a draw from the global posterior Π^*. The functions ω_j are defined on $\mathcal{X}$ such that $\sum_{j=1}^m \omega_j(x) = 1$ for all $x \in \mathcal{X}$. The most naive approach would be to take $\omega_j = \mathbf{1}_{\mathcal{D}_j}$, "gluing" in a sense the local GP posteriors. However, this approach leads to discontinuous samples from the global posterior. In order to avoid discontinuities, continuous data-driven weight functions concentrating around $\mathcal{D}_j$ are favored.

§1.8 Overview

This thesis focuses on frequentist properties of Bayesian techniques. The first chapter examines how adaptation affects uncertainty quantification using exponentially decaying covariance kernel. The two following chapters (Chapter 3-4) focuses on distributed computation in the Bayesian non-parametric regression model. In Chapter 3, we derive contraction rates and coverage for uniformly randomly partitioned distributed methods when the smoothness of the true function is known. On the other hand, Chapter 4 deals with adaptive optimal recovery for spatially partitioned distributed methods. Finally, the last chapter is an extensive comparative simulation study of the aforementioned distributed methods.

§1.9 Notations

For two positive sequences a_n, b_n we use the notation $a_n \lesssim b_n$ if there exists an universal positive constant C such that $a_n \leq C b_n$. Along the lines $a_n \asymp b_n$ denotes that $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold, simultaneously. For $\theta \in L_2[0,1]$ we denote the standard L_2-norm as $\|\theta\|_2^2 = \int_0^1 \theta(x)^2 dx$ and let $\mathrm{diam}(S)$ denote the ℓ_2-diameter of the set $S \subset \ell_2$. Throughout the thesis, c and C denote global constants whose value may change one line to another. The dependence of the constants c, C on the model parameters we denote by sub-indexes, e.g. c_β, $C_{\beta,m,M}$.

CHAPTER 2

Adaptive credible sets with a squared-exponential GP prior

This chapter has been published as: A. Hadji, B. Szabó, "Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel?" (2021) in SIAM/ASA Journal on Uncertainty Quantification, (9), 1, 185-230

Abstract. We investigate the frequentist coverage properties of credible sets resulting in from Gaussian process priors with squared exponential covariance kernel. First we show that by selecting the scaling hyper-parameter using the maximum marginal likelihood estimator in the (slightly modified) squared exponential covariance kernel the corresponding credible sets will provide overconfident, misleading uncertainty statements for a large, representative subclass of the functional parameters in context of the Gaussian white noise model. Then we show that by either blowing up the credible sets with a logarithmic factor or modifying the maximum marginal likelihood estimator with a logarithmic term one can get reliable uncertainty statement and adaptive size of the credible sets under some additional restriction. Finally we demonstrate on a numerical study that the derived negative and positive results extend beyond the Gaussian white noise model to the non-parametric regression and classification models for small sample sizes as well. The performance of the squared exponential covariance kernel is also compared to the Matérn covariance kernel.

§2.1 Main results

§2.1.1 Model description

We consider the Gaussian white noise model

$$Y(t) = \int_0^t \theta_0(s)ds + \frac{1}{\sqrt{n}}W_t, \quad t \in [0,1], \tag{2.1.1}$$

where $\theta_0 \in L_2[0,1]$ is the unknown function of interest and W_t denotes the Brownian motion. Let P_0, E_0, and V_0 denote the corresponding probability measure, expected value, and variance, respectively. In the Bayesian approach we endow the unknown function of interest θ_0 with a prior distribution representing our initial belief. In our

work we investigate the popular Gaussian process prior with rescaled squared exponential kernel. Let us consider the sequence representation of the Gaussian white noise model. For an orthonormal basis ψ_i, $i = 1, 2, ...$ (e.g. the Fourier basis) let us denote the sequence decomposition of the functions $\theta_0(t), Y(t)$, and W_t by $Y_i = \langle Y(t), \psi_i(t)\rangle_2$, $\theta_{0,i} = \langle \theta_0, \psi_i(t)\rangle_2$, and $Z_i = \langle W_t, \psi_i(t)\rangle_2 \overset{iid}{\sim} N(0,1)$, $i = 1, 2, ...$, respectively. Then the equivalent sequence model can be given in the form

$$Y_i = \theta_{0,i} + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, ...$$

Slightly abusing our notations we denote by θ_0 both the functional parameter in the Gaussian white noise model and the sequential parameter $\theta_0 = (\theta_{0,1}, \theta_{0,2}, ...)$ in the sequence model. It is common to assume that the true function θ_0 belongs to a hyper-rectangle regularity class, i.e.

$$\theta_0 \in \Theta^\beta(M) = \left\{ \theta \in \ell_2 : \sup_{i \geq 1} \theta_i^2 i^{2\beta+1} \leq M \right\},$$

for some (typically unknown) $\beta, M > 0$. The class $\Theta^\beta(M)$ is closely related to Sobolev type of regularity classes $S^\beta(M) = \{\theta \in \ell_2 : \sum_{i \geq 1} \theta_i^2 i^{2\beta} \leq M\}$ and the derived results can (typically) easily be extended to them, see for instance (Szabo et al., 2015). We note that the minimax estimation rate for the above hyper-rectangle regularity class is $n^{-\beta/(2\beta+1)}$, i.e. there exists $C_\beta > 0$ such that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta^\beta(M)} E_0 \|\theta - \hat{\theta}\|_2 \geq C_\beta n^{-\beta/(2\beta+1)},$$

where the infimum is taken over all estimators, see for instance (Donoho, 1994).

In view of Mercel's theorem we can represent the Gaussian process prior with squared exponential kernel as

$$G_t = \sum_{i=1}^{\infty} \lambda_i \xi_i \psi_i(t),$$

where λ_i, ψ_i, $i = 1, 2, ...$ are the eigenvalues and eigenfunctions of the squared exponential covariance kernel, and ξ_i are iid standard normal random variables, see for instance Chapter 4.3 of (Rasmussen, 2004). The corresponding coefficients λ_i can be approximated as $\lambda_i^2 \approx a^{-1} e^{-i/a}$ for $\mathcal{X} = \mathbb{R}$ and with respect to the Gaussian dominated measure, see for instance (Rasmussen, 2004). In the rest of the chapter we (mainly) work with the prior

$$\theta | a \sim \bigotimes_{i=1}^{\infty} N(0, a^{-1} e^{-i/a}) \tag{2.1.2}$$

in the sequence model, for convenience. Note that in view of $Y|\theta_0 \sim \bigotimes_{i=1}^{\infty} N(\theta_{0,i}, n^{-1})$ and the choice of the prior $\Pi_a(\cdot)$ in (2.1.2), the corresponding posterior $\Pi_a(\cdot|Y)$ takes the form

$$\theta | a, Y \sim \bigotimes_{i=1}^{\infty} \mathcal{N}\left(\frac{nY_i}{ae^{i/a} + n}, \frac{1}{ae^{i/a} + n} \right). \tag{2.1.3}$$

The behavior of the posterior distribution is very sensitive on the choice of the hyper-parameter a. Since the optimal choice of a depends on the (typically) unknown regularity parameter β of the underlying functional parameter of interest θ_0 in practice one uses data driven procedures for selecting a. The two most commonly applied Bayesian techniques for selecting the hyper-parameter are the hierarchical Bayes and the marginal likelihood empirical Bayes methods. In the hierarchical Bayes method the hyper-parameter a is endowed with a prior distribution π (also called hyper-prior distribution), resulting in a two-level, hierarchical prior distribution

$$\Pi(\cdot) = \int_0^\infty \Pi_a(\cdot)\pi(a)da.$$

For technical reasons, we introduce the following assumptions on the hyper-prior density function $\pi(\cdot)$ supported on $[1, A_n]$.

Assumption 2.1.1. *Let us assume that for some $c_1 > 0$ there exist $c_2, c_6 \geq 0$ and $c_3, c_4, c_5 > 0$ such that*

$$c_4^{-1} a^{-c_3} \exp(-c_2 a) \leq \pi(a) \leq c_4 a^{-c_5} \exp(-c_6 a), \tag{2.1.4}$$

for all $c_1 \leq a \leq A_n$.

Note that among others the exponential, the gamma, and the inverse gamma distributions (restricted to $[1, A_n]$) satisfy Assumption 2.1.1.

In contrast to this in the empirical Bayes approach we take the maximum marginal likelihood estimator (MMLE), i.e.

$$\hat{a}_n := \arg \max_{a \in [1, A_n]} \ell_n(a), \tag{2.1.5}$$

where the marginal log-likelihood function (with respect to the measure $\bigotimes_{i=1}^\infty N(0,1)$) is

$$\ell_n(a) = -\frac{1}{2} \sum_{i=1}^\infty \left(\log\left(1 + \frac{n}{ae^{i/a}}\right) - \frac{n^2 Y_i^2}{ae^{i/a} + n} \right)$$

and the parameter $A_n = o(n)$ restricts the parameter space to a compact interval, which is advantageous both from practical and analytical perspective. Then the estimator $\hat{a}_n$ is plugged in into the posterior distribution (2.1.3), resulting in the empirical Bayes posterior $\Pi_{\hat{a}_n}(\cdot|Y)$.

We show in Section 2.4.2 that both of these methods result in optimal recovery for the functional parameter of interest θ_0. These results are of interest on their own right, but our main focus lies on the reliability of Bayesian uncertainty quantification resulting both from the hierarchical and the empirical Bayes procedures, hence we have deferred the contraction rate results to the Section 2.4.2.

§2.1.2 Uncertainty quantification

In our work we investigate the reliability of the built-in uncertainty quantification of the above data-driven posterior distributions. For convenience let $\Pi_n(\cdot|Y)$ denote both the hierarchical and the empirical Bayes posterior distributions in the following.

In Bayesian methods the remaining uncertainty of the procedure is visualized by the credible set. We consider ℓ_2-credible balls centered around the posterior mean, i.e. we analyze credible sets in the form

$$\hat{C}_{n,\alpha} = \left\{\theta \in \ell_2 : \|\theta - \hat{\theta}\|_2 \leq r_\alpha\right\} \tag{2.1.6}$$

where $\hat{\theta}$ is the posterior mean and the radius r_α is chosen such that $\Pi(\theta \in \hat{C}_{n,\alpha}|Y) = 1-\alpha$, for some prescribed significance level $\alpha > 0$.

We are interested in the frequentist properties of ℓ_2-credible balls resulting from the data driven credible balls. Then let us denote by r_α the radius of the ℓ_2-ball centered around the posterior mean $\hat{\theta}$ and accumulating $1-\alpha$ fraction of the posterior mass, i.e.

$$\Pi(\theta : \|\theta - \hat{\theta}\|_2 \leq r_\alpha|Y) = 1-\alpha.$$

In our analysis we introduce some additional flexibility by considering inflated credible balls, i.e.

$$\hat{C}_n(L_n) = \left\{\theta \,\|\theta - \hat{\theta}\|_2 \leq L_n r_\alpha\right\}, \tag{2.1.7}$$

for some blown up factor $L_n \geq 1$, possibly depending on n. As a first step we note that the size of the credible set for both the empirical and hierarchical Bayes procedures adapts to the minimax rate (actually the diameter of the set is even a logarithmic factor faster than the minimax rate in case of the empirical Bayes procedure).

Corollary 2.1.2. *Both the hierarchical and the empirical Bayes credible sets defined in* (2.1.7) *have rate adaptive size, i.e. for every $\beta_0 > 0$ and $M > 0$*

$$\sup_{\beta \geq \beta_0} \sup_{\theta \in \Theta^\beta(M)} P_0\left(diam(\hat{C}_n(1)) \geq M_n n^{-\beta/(2\beta+1)} (\log n)^{-1/(2\beta+1)}\right) \to 0,$$

where the sequence M_n goes to infinity arbitrary slowly in case of the empirical Bayes method and $M_n \gg \log n$ in case of the hierarchical Bayes method.

Proof. The proof is given in Sections 2.7 and 2.10 respectively. □

2.1.2.1 Coverage of credible sets - negative results

Next we investigate how much we can trust the above derived data-driven Bayesian uncertainty quantification from a frequentist perspective. We would like to know whether the true function θ_0 is included in the (blown up) credible set, i.e. if any fixed $\beta_0 > 0$

$$\inf_{\theta_0 \in \cup_{\beta \geq \beta_0} \Theta^\beta(M)} P_0(\theta_0 \in \hat{C}_n(L_n)) \geq 1-\alpha$$

holds for some sufficiently large choice of L_n. Since it is impossible to construct honest confidence sets with rate adaptive size and in view of the adaptive size of the credible sets, see Corollary 2.1.2, they must have poor frequentist coverage properties at least for certain functional parameters θ_0. Actually the radius of the credible sets decays even faster than the minimax rate, which already implies impossibility of coverage. Nevertheless it is of interest to quantify the set of functions for which the Bayesian uncertainty quantification is truth-worthy.

First we note that a representative subset of the hyper-rectangle $\Theta^\beta(M)$ is the set

$$\Theta_s^\beta(m, M) = \left\{ \theta \in \Theta^\beta(M) : \min_{i \geq 1} i^{1+2\beta} \theta_i^2 \geq m \right\}, \tag{2.1.8}$$

for some parameters $0 < m \leq M$. Let us refer to this subclass of sequential parameters as self-similar signals following the similar terminology of (Giné and Nickl, 2010) and (Szabo et al., 2015). It was shown in the later paper that the minimax rate over $\Theta_s^\beta(m, M)$ is the same as over $\Theta^\beta(M)$. The next theorem shows that both of the hierarchical and the empirical Bayes procedures provide unreliable uncertainty quantification over this representative sub-class of functions unless it is blown up with at least a logarithmic factor.

Theorem 2.1.3. *Let us take arbitrary $L_n = o(\sqrt{\log n})$. Then the empirical and hierarchical Bayes credible sets blown up by L_n have frequentist coverage tending to zero for every self-similar signal, i.e. for every $0 < m \leq M$,*

$$\sup_{\theta_0 \in \Theta_s^\beta(m,M)} P_{\theta_0}(\theta_0 \in \hat{C}_n(L_n)) \to 0.$$

Proof. See Section 2.7. □

This negative result draws a dark picture as it tells us that one can not trust Bayesian uncertainty quantification resulting from the investigated prior, even if one allows certain amount of adjustment (i.e. by blowing up the set with a sequence tending to infinity, not too fast). Since the prior used is very closely related to the Gaussian process with squared exponential covariance kernel this gives the intuition that one has to be very cautious working with squared exponential kernel as the corresponding Bayesian uncertainty statement are (typically) unreliable. In the next subsection we will be touching the corners by deriving some positive results on the coverage properties of the credible sets. First we show that for analytic functions the (slightly inflated) credible sets provide reliable uncertainty quantification and second we show that by blowing up the credible sets by a logarithmic factor or by slightly adjusting the maximum marginal likelihood estimator, one gets reliable uncertainty statements for a large subclass of functions, including the self-similar functions.

2.1.2.2 Coverage of credible sets - positive results

Let us consider the set of analytic-type functions defined as

$$\theta_0 \in A^\gamma(M) = \left\{ \theta \in \ell_2 : \sum_{i=1}^{\infty} \theta_i^2 e^{2i\gamma} \leq M \right\},$$

for some $\gamma > 0$. Note that the investigated prior (2.1.2) is more suitable for this class of functions due to the exponential decay of the variances. We show below that, indeed, for the class $A^\gamma(M)$ both the empirical and the hierarchical Bayes procedures provide reliable uncertainty quantification. Note, however, that the present class of functions is substantially smaller than $\Theta^\beta(M)$, for any $\beta > 0$.

Theorem 2.1.4. *The inflated empirical and hierarchical Bayes credible sets $\hat{C}_n(L)$ have frequentist coverage tending to one over the class $\theta_0 \in A^\gamma(M)$ for any $\gamma \geq 1/2$ and sufficiently large constant $L > 0$, i.e.*

$$\inf_{\theta_0 \in A^\gamma(M)} P_0(\theta_0 \in \hat{C}_n(L)) \to 1.$$

Furthermore, the size of the credible set is (nearly) optimal, i.e. for some sufficiently large constant $C > 0$,

$$\inf_{\theta_0 \in A^\gamma(M)} P_0\left(diam(\hat{C}_n(1)) \leq Cn^{-1/2} \log n \right) \to 1.$$

Proof. See Section 2.6. □

Next we investigate the behavior of the credible sets by allowing a logarithmic inflating factor. Since the size of the inflated credible sets are still nearly minimax, the credible sets fail to cover all functional parameter θ_0 of interest, in view of the non-existence result of adaptive confidence sets seen in (Cai and Low, 2004) and (Robins and van der Vaart, 2006). Therefore we restrict the investigated class of functions to the so called polished tail class, introduced in (Szabo et al., 2015) and (Rousseau and Szabo, 2020). We say that a sequential parameter $\theta \in \ell_2(M)$ belongs to the class of polished tail signals denoted by $\Theta_{pt}(L_0, N_0, \rho)$, for some $L_0, \rho, N_0 > 0$ if

$$\sum_{i=N}^{\infty} \theta_i^2 \leq L_0 \sum_{i=N}^{\rho N} \theta_i^2, \qquad \text{for all } N \geq N_0.$$

The above assumption basically requires that knowing the sequential parameter θ up to a certain coordinate enables us to draw conclusion about the tail of the sequence. We require that the energy (sum of the squared coefficients) of the tail is dominated by the energy of a finitely large block of coefficients. This condition makes also sense intuitively as in the stochastic model the signal can be observed only up to some limit, the fluctuation in the later coordinates can equally likely be caused by the noise. Therefore to make reliable uncertainty statement we have to assume that the tail behavior of the signal hidden by the noise is not substantial and can be extrapolated by information available at given signal-to-noise ratio. In (Szabo et al., 2015) it was shown that the above assumption is mild from statistical, topological and Bayesian point of view as well.

The next theorem states that when the sequential parameter θ_0 is restricted to polished tail sequences, then both the empirical and the hierarchical Bayes credible balls blown up by a $\log n$ factor (i.e $\hat{C}_n(L \log^{3/2} n)$) are honest frequentist confidence set, for large enough L.

Theorem 2.1.5. *For any $L_0, N_0, \rho \geq 1$ there exists a constant L such that*

$$\inf_{\theta_0 \in \Theta_{pt}(L_0, N_0, \rho)} P_0(\theta_0 \in \hat{C}_n(L(\log n)^{3/2})) \to 1,$$

where $\hat{C}_n$ denotes either the empirical or the hierarchical Bayes credible sets under Assumption 2.1.1.

Proof. See Sections 2.5 and 2.10.3 respectively. □

Hence one can achieve reliable uncertainty quantification on an arguably large subset of the function space by blowing up the standard credible set with a slowly varying term. This, however, is not very appealing as a practitioner would righteously hesitate introducing the artificial logarithmic blow up. Therefore, we propose another method, where one does not have to introduce a logarithmic blow up factor, but instead adjust the maximum marginal likelihood estimator. Investigating the proof of Theorem 2.1.3 one can see that the MMLE $\hat{a}_n$, given in (2.1.5), is too small, the empirical Bayes procedure is basically oversmoothing. One can compensate for this by undersmoothing the procedure. We propose to adjust the MMLE by a multiplicative logarithmic factor

$$\tilde{a}_n = \log(n)\hat{a}_n. \tag{2.1.9}$$

Then the corresponding empirical Bayes credible set (blown up by a sufficiently large constant $L > 0$) results in reliable uncertainty quantification for self-similar functions $\Theta_s^\beta(m, M)$.

Theorem 2.1.6. *For any $0 < m \leq M$ there exists a constant $L > 0$ such that*

$$\inf_{\theta_0 \in \Theta_s^\beta(m,M)} P_0(\theta_0 \in \tilde{C}_n(L)) \to 1,$$

where $\tilde{C}_n(1)$ denotes the credible set resulting from the empirical Bayes posterior with hyper-parameter $\tilde{a}_n$.

Proof. The proof of the theorem is deferred to Section 2.9. □

§2.2 Numerical analysis

In this section we investigate the numerical properties of the Gaussian process prior with (approximately) squared exponential covariance kernel. First we consider the Gaussian white noise model and the prior (2.1.2). We show that the corresponding Bayesian uncertainty quantification is misleading for various regularly behaving functions. We also demonstrate that a different choice of the covariance kernel or a modified version of the empirical Bayes procedure results in more accurate uncertainty statements. Then we consider the (from practical point of view) more relevant non-parametric regression and classification models, where we also demonstrate the sub-optimal behavior of the (standard) empirical Bayes method with squared exponential covariance kernel and show that the proposed modification results in superior performance compared to it. We also consider GP priors with Matérn covariance kernels and show that although they poses good recovery and uncertainty quantification properties, their run times are substantially slower than using squared exponential kernels.

§2.2.1 Gaussian white noise model

First we demonstrate the sub-optimal performance of the Gaussian process with (approximately) squared exponential covariance kernel (2.1.2) compared to modified versions of the empirical Bayes procedure and to the Gaussian process prior with polynomially decaying variances in the series representation, see (Knapik et al., 2016) and (Szabo et al., 2015). Let us consider the function $\theta_1 \in L_2[0,1]$ given by their Fourier coefficients $\theta_{1,i} = i^{-3/2}\sin(i)$, for $i = 1, 2, ...$, respectively, relative to the Fourier eigenbasis $\psi_i(t) = \sqrt{2}\cos(\pi(i-1/2)t)$. Note that although the function lies outside of the self-similar function class (2.1.8), it has essentially the same behavior. In Figure 2.1 we visualize the 95% credible sets (light blue or light red), the posterior mean (blue or red) and the true function (black), by simulating 2000 draws from the empirical Bayes posterior distribution and plotting the closest 95% of them in L_2-norm to the posterior mean. We note that all credible sets were constructed without any inflation factor, i.e. $L_n = 1$ was taken (except of the case where the choice $L_n = \log n$ was pre-specified). The credible sets are drawn for signal-to-noise ratio $n = 100$, 500, 1000 and 5000, respectively. We also plot the same credible sets blown-up by a $\log n$ factor, the credible sets obtained by the modified empirical Bayes procedure (where the MMLE $\hat{a}_n$ of the scaling parameter a was multiplied by $\log n$) and the empirical Bayes credible sets corresponding to the prior $\theta \sim \otimes_{i=1}^{\infty} N(0, i^{-1-2\alpha})$, with hyper-parameter α estimated by the MMLE. One can see that the standard marginal likelihood empirical Bayes method provides too narrow credible sets failing to cover the underlying true function. Also note that both modifications of the empirical Bayes credible sets and using the prior with polynomially decaying variances provide good coverage, but in contrast to the overly conservative approach of inflating the credible sets with a logarithmic factor the modification of the MMLE results in more informative uncertainty statement (i.e. smaller credible sets).

§2.2.2 Non-parametric regression and classification

In this section we demonstrate on a simulation study that the results derived for the Gaussian white noise model generalize to more complicated statistical models as well. We consider the popular non-parametric regression and classification models specifically. The empirical Bayes posteriors, posterior means and credible sets are computed in both cases using the MatLab package "gpml".

In the non-parametric regression model we assume to observe pairs of random variables $(X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n)$, where

$$Y_i = \theta_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad X_i \overset{iid}{\sim} U(0,1), \quad i = 1, ..., n,$$

and the aim is to estimate the unknown non-parametric regression function θ_0. In the Bayesian approach we endow θ_0 with a Gaussian process prior with squared exponential kernel and estimate the tuning parameter using the MMLE.

In this simulation study we take the Fourier coefficients of the underlying true function θ_2 to be $\theta_{2,i} = i^{-3/2}\cos(i)$, $i = 1, 2, ...$ We take $\sigma^2 = 1/2$, but in the procedure it is considered to be unknown and estimated with the MMLE $\hat{\sigma}^2$. We plot the true function (black), the posterior mean (blue), and the posterior point-wise credible

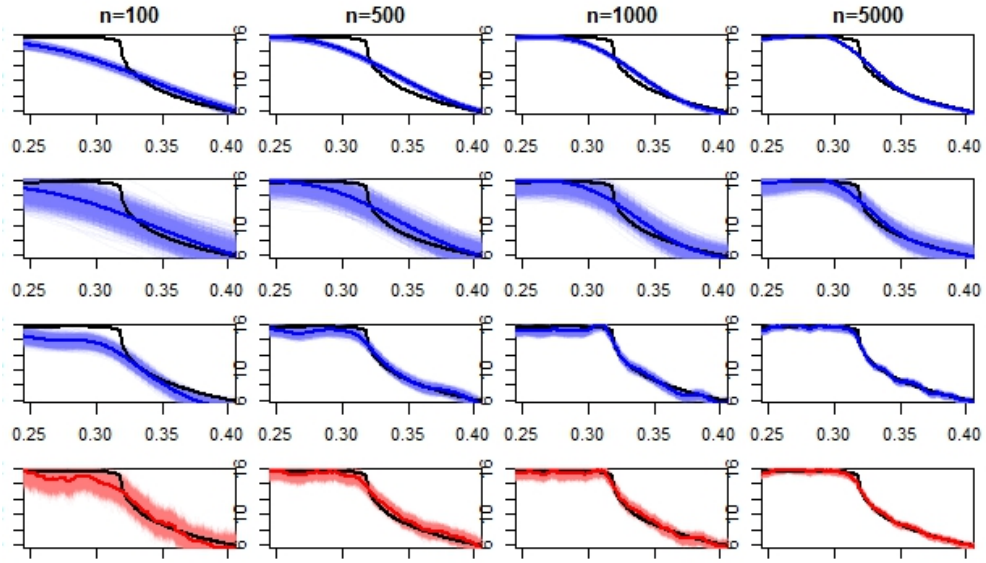

Figure 2.1: Empirical Bayes credible sets for the function θ_1 (drawn in black) zoomed in to the interval $x \in [0.25, 0.4]$. First line: credible set (in light blue) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance. Second line: credible set (in light blue) blown-up by a $\log n$ factor ($L_n = \log n$) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance. Third line: credible set (in light blue) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance and modified empirical Bayes procedure (rescaling factor multiplied by $\log n$). Last line: credible set (in light red) and posterior mean (red curve) corresponding to the prior with polynomially decaying variance. From left to right the signal to noise ratio is $n = 100, 500, 1000, 5000$.

intervals (dashed blue) $[\hat{\theta}(x) - q_{0.025}\sqrt{\hat{c}(x,x)}, \hat{\theta}(x) + q_{0.025}\sqrt{\hat{c}(x,x)}]$, where $\hat{\theta}$ is the posterior mean, q_α the α-th quantile of the standard normal distribution and $\hat{c}(.,.)$ the posterior covariance kernel. We consider the MMLE empirical Bayes method with and without the $\log n$ inflation factor for the credible set, the modified empirical Bayes method (where the MMLE was multiplied by $\log n$), and finally the empirical Bayes method for Matérn covariance kernel with estimating either the regularity or the scale tuning parameter from the data. We take the sample size to be $n = 100, 500, 1000,$ and 2000. Observe in Figure 2.2 that the standard MMLE empirical Bayes method provides unreliable uncertainty quantification in certain points, while the two modified squared exponential credible sets and the empirical Bayes credible sets from the Matérn kernel (with data-driven choice of the regularity hyper-parameter) capture the underlying functional parameter of interest better. Also note that by multiplying the MMLE of the scaling parameter by a $\log n$ factor in the squared exponential kernel case we do not get an overly conservative credible set, unlike in the case when the radius is inflated with a logarithmic factor. Finally, we note that the computation time using the squared exponential kernel is much smaller than working with the Matérn kernel. We note that the computational times corresponding to the Matérn kernel are higher than for the squared exponential kernel. Estimating the regularity hyper-parameter of the kernel is time consuming as the eigenfunctions depend on it. Alternatively, one can consider a rescaled Matérn covariance kernel with fixed regularity. This method is typically faster, however, optimal recovery of the underlying

function is possible only up to the smoothness level $\alpha + d/2$, where α denotes the regularity of the prior, see for instance Szabo et al. (2013). Therefore, we choose α large enough ($\alpha = 10$), which then seemingly slows down the computations. The different running times can be found in Table 2.3. The running time is based on the time spent computing the MMLE, the posterior mean and the point-wise posterior variance using the MatLab package "gpml" in a personal computing environment.

We also investigate empirically the frequentist coverage probabilities of the point-wise credible sets by repeating the experiment 100 times and reporting the frequency that the function at given points (we consider $x = (0.25, 0.3188, 0.75)$ with $0.3188 = \text{argmax}_{x\in[0,1]}\theta_2(x)$) is included in the credible interval, see Table 2.1. Moreover, Table 2.2 shows the average size of the point-wise credible intervals (i.e. $2q_{0.025}\sqrt{\hat{c}(x,x)}$) depending on the sample size n and the procedure used to compute the credible sets. One can observe similar behavior to what we have described above.

Note that Table 2.1 does not quite illustrate the results of Section 2.1.2 since the table shows the point-wise credible intervals whereas most of our theoretical results concern the L_2 credible balls. However, they still give an indication of the reliability of Bayesian uncertainty quantification. The point $x = 0.3188$, at which the maximum of θ_2 is achieved, is seen as one of the clearest way to illustrate our negative results, whereas our positive results could be accepted only if the probability of $\theta_2(x)$ being inside of the corresponding credible interval goes to one for all point $x \in [0, 1]$.

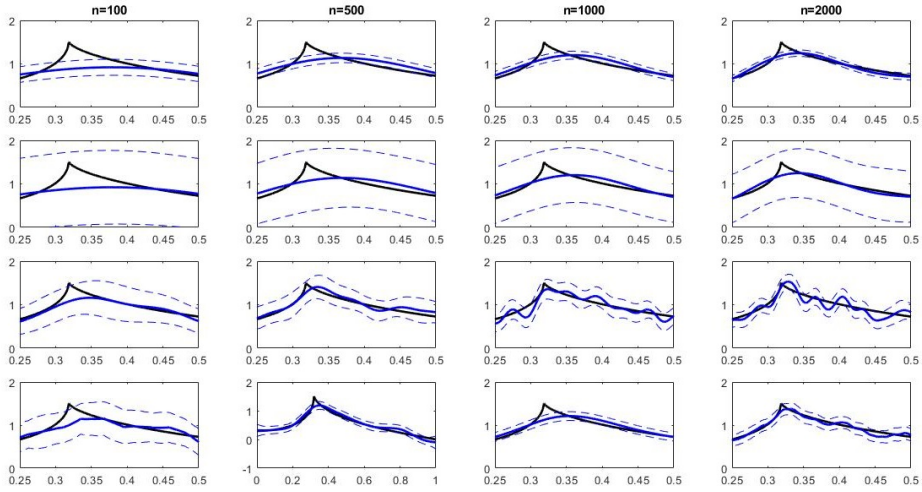

Figure 2.2: Empirical Bayes credible sets for the regression function θ_2 (drawn in black), zoomed in to the interval $x \in [0.25, 0.5]$. The posterior means are drawn by solid blue line, while the 95% point-wise credible sets by dashed blue curves. In the first row we plot the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, the third row the modified MMLE empirical Bayes method using squared exponential Gaussian process prior, while in the fourth row we plot the empirical Byes credible sets using a Matérn kernel with data-driven choice for the regularity hyper-parameter. From left to right the sample size is $n = 100, 500, 1000, 2000$.

Next we consider the non-parametric classification problem. Let us assume that

	$x = 0.25$			$x = 0.3188$			$x = 0.75$		
$n =$	100	500	1000	100	500	1000	100	500	1000
Method 1	0.84	0.69	0.57	0.01	0.01	0.00	0.98	0.92	0.97
Method 2	1.00	1.00	1.00	0.96	0.98	1.00	1.00	1.00	1.00
Method 3	0.98	0.98	0.97	0.35	0.55	0.50	0.99	0.96	0.98
Method 4	0.99	1.00	1.00	0.12	0.35	0.51	1.00	1.00	1.00
Method 5	0.98	1.00	1.00	0.08	0.30	0.47	0.99	1.00	1.00

Table 2.1: Frequencies that $\theta_2(x)$ is inside of the corresponding credible interval for the squared exponential and Matérn Gaussian process prior at given points $x \in \{0.25, 0.3188, 0.75\}$. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$), Method 4: Matérn kernel with smoothness MMLE empirical Bayes, Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

$n =$	100	500	1000
Method 1	0.3956	0.2367	0.1814
Method 2	1.8218	1.4711	1.2533
Method 3	0.7541	0.5279	0.4262
Method 4	0.6346	0.4308	0.3446
Method 5	0.5151	0.3338	0.263

Table 2.2: Average size of the pointwise credible intervals (i.e. $2q_{0.025}\sqrt{\hat{c}(x,x)}$) for $\theta_2(x)$ in the regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$), Method 4: Matérn kernel with smoothness MMLE empirical Bayes, Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

$n =$	100	500	1000	5000	10000	200000
Method 1	0.74 s	2.75 s	10.84 s	3.7 m	25.2 m	1.2 h
Method 4	1.48 s	13.93 s	43.83 s	16.7 m	3.8 h	12.5 h
Method 5	1.37 s	11.15 s	33.5 s	12.3 m	2.8 h	10.5 h

Table 2.3: Average run time of the EB methods for θ_2 in the regression model. Method 1: SE covariance kernel, Method 4: Matérn covariance kernel and MMLE for the regularity hyper-parameter, Method 5: Matérn covariance kernel and MMLE for the scaling hyper-parameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000$

we observe the binary random variables $Y_1, Y_2, ..., Y_n \in \{0, 1\}$, with

$$P(Y_i = 1) = p(X_i), \qquad X_i \overset{iid}{\sim} U(0, 1), \quad i = 1, ..., n,$$

for some non-parametric function $p(x) : [0, 1] \mapsto [0, 1]$. We write $p(x)$ in the form $p(x) = \psi(\theta(x))$, with $\psi(x) = e^x/(1 + e^x)$, for some function $\theta(x) : [0, 1] \mapsto \mathbb{R}$. In the Bayesian approach we endow the functional parameter $\theta(x)$ with a Gaussian process prior with squared exponential or Matérn covariance kernel.

We design similar experiments for the non-parametric classification model as for the non-parametric regression model above, with sample sizes $n = 100, 500, 1000$ and 2000 and the same θ_2 as above. We plot the point-wise credible intervals for θ_2 corresponding to the empirical Bayes procedure, with and without a $\log n$ inflation factor, and to the modified empirical Bayes procedure (where the MMLE is multiplied by

a $\log n$ factor), see Figure 2.3. One can observe that the standard MMLE empirical Bayes procedure produces unreliable uncertainty statements, while by blowing up the credible sets with a logarithmic factor we get overly conservative uncertainty quantification. These problems are resolved by considering the modified empirical Bayes method, which captures the shape of the underlying functional parameter better and provides more reliable uncertainty statements. We also collect the empirical estimation of the frequentist coverage probabilities of the underlying functional parameter $\theta_2(x)$ at points $x = (0.25, 0.3188, 0.75)$ in Table 2.4 and the computation time for different methods in Table 2.6, underlying the conclusions drawn from the figures above. Note that, similarly to Table 5.2, Table 2.4 does not quite illustrate our theoretical results, but is linked to it in a similar fashion as Table 5.2. Moreover, Table 2.5 shows the size of the average credible interval.

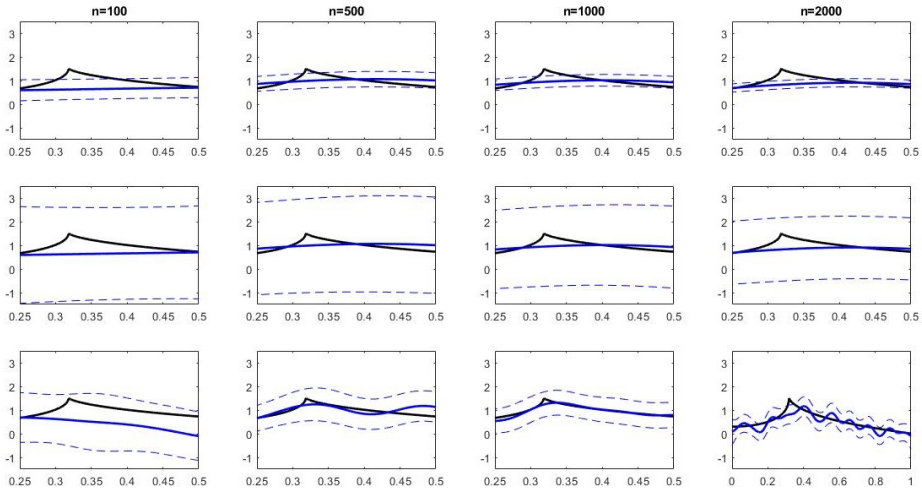

Figure 2.3: Empirical Bayes credible sets using squared exponential Gaussian process priors in the classification model for the function θ_2 (drawn in black). The posterior means are drawn by solid blue line, while the 95% point-wise credible intervals by dashed blue curves. In the first row we plotted the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, while in the the third row the modified MMLE empirical Bayes method. From left to right the sample size is $n = 100, 500, 1000, 2000$.

§2.3 Discussion

We have shown that the MMLE empirical Bayes method for Gaussian process prior with (a slightly modified version of the) squared exponential covariance kernel produces misleading uncertainty statement in context of the Gaussian white noise model. The derived negative results were demonstrated on a simulation study in context of the Gaussian white noise model and extended to the non-parametric regression and classification models as well. Hence we can conclude that one has to be very cautious when applying empirical Bayes methods with squared exponential Gaussian

	$x = 0.25$			$x = 0.3188$			$x = 0.75$		
$n =$	100	200	500	100	200	500	100	200	500
Method 1	0.90	0.90	0.89	0.29	0.16	0.12	0.92	0.88	0.85
Method 2	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00
Method 3	0.91	0.94	0.95	0.42	0.36	0.45	0.94	0.94	0.95
Method 4	0.94	0.96	0.95	0.32	0.27	0.42	0.95	0.96	0.96
Method 5	0.94	0.94	0.96	0.30	0.32	0.35	0.96	0.96	0.96

Table 2.4: Frequencies that $\theta_2(x)$ is inside of the corresponding credible interval for squared exponential and Matérn Gaussian process prior in the logistic regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$), Method 4: Matérn kernel with MMLE for the smoothness, Method 5: Matérn kernel with MMLE for the scaling and taking $\alpha = 10$. From left to right the sample size is $n = 100, 200, 500$.

	n=100	n=200	n=500
Method 1	3.2672	0.8209	0.3485
Method 2	15.0461	4.3495	2.1661
Method 3	3.6777	1.2675	0.7575
Method 4	3.5409	1.1186	0.6212
Method 5	3.4040	0.9698	0.4848

Table 2.5: Average size of the pointwise credible intervals $2q_{0.025}\sqrt{\hat{c}(x,x)}$ for θ_2 in the logistic regression model. The methods and the sample sizes are the same as in Table 2.4.

$n =$	100	500	1000	5000
Method 1	2.23 s	30.81 s	5.9 m	2.8 h
Method 4	4.77 s	3.1 m	23.9 m	11.1 h
Method 5	4.42 s	3 m	15.1 m	8.2 h

Table 2.6: Average run time of the EB methods for θ_2 in the logistic regression model. Method 1: SE covariance kernel, Method 4: Matérn covariance kernel and MMLE for the regularity hyper-parameter, Method 5: Matérn covariance kernel and MMLE for the scaling hyper-parameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000$

processes for uncertainty quantification as typically they provide misleading confidence statements, due to over-smoothing behavior of the MMLE. We note that the bad performance of the prior (2.1.2) is not due to the rescaling factor a^{-1} in the variance, because similar (but easier) computations show that the prior without the a^{-1} factor behaves sub-optimally as well.

One can compensate the haphazard uncertainty statements by blowing up the credible sets with a $\log n$ factor, however, this approach is not appealing from a practical perspective, as demonstrated in our simulation study as well. Instead we propose to modify the MMLE by multiplying it with $\log n$ to compensate for the over-smoothing. This procedure is less conservative than the previous one and hence provides more accurate information about the uncertainty of the method. One can also consider different covariance kernels, with polynomially decaying eigenvalues, like the Matérn kernel, however, these procedures can be computationally less appealing, as demonstrated in the simulation study.

§2.4 Some properties of the MMLE

§2.4.1 Deterministic bounds

As a first step we provide deterministic bounds for the marginal maximum likelihood estimator $\hat{a}_n$ of the rescaling hyper-parameter a. Let us introduce the following functions for $a \in [1, \infty)$:

$$h_n(a, \theta_0) := \frac{1}{\log^2\left(\frac{n}{a}\right)} \sum_{i=1}^{\infty} \frac{n^2 i e^{i/a} \theta_{0,i}^2}{a(ae^{i/a} + n)^2}, \tag{2.4.1}$$

$$g_n(a, \theta_0) := \frac{1}{\log^2\left(\frac{n}{a}\right)} \sum_{i=2a}^{\infty} \frac{n^2 (i-a) e^{i/a} \theta_{0,i}^2}{a(ae^{i/a} + n)^2}. \tag{2.4.2}$$

These functions are derived from the expected value of the score function, see Section Then let us define the deterministic bounds $\underline{a}_n$ and $\overline{a}_n$ for $\hat{a}_n$ with the help of the functions h_n and g_n as

$$\begin{aligned} \underline{a}_n &:= \sup\{a \in [1, A_n] : g_n(a, \theta_0) \geq B \log n\}, \\ \overline{a}_n &:= \sup\{a \in [K_0, A_n] : h_n(a, \theta_0) \geq b\}, \end{aligned} \tag{2.4.3}$$

with some $b, B, K_0 > 0$ to be specified later and $A_n = o(n)$ given in (2.1.5). Then we show that these bounds sandwich $\hat{a}_n$ with high probability.

Theorem 2.4.1. *The MMLE $\hat{a}_n$ satisfies*

$$\inf_{\theta_0 \in \ell_2(M)} P_0(\underline{a}_n \leq \hat{a}_n \leq \overline{a}_n) \to 1, \tag{2.4.4}$$

for $\underline{a}_n, \overline{a}_n$ defined in (2.4.3).

Proof. See Section □

We also derive upper bounds for $\overline{a}_n$, in the case the true function belongs to the hyper-rectangle with regularity hyper-parameter β or or to the analytic function class A^γ and a lower bound for $\underline{a}_n$ in the case of self similar functions $\theta_0 \in \Theta^\beta(m, M)$.

Proposition 2.4.2. *For every $\beta \geq \beta_0$ and $\gamma > 0$ there exist $C_{\beta,b,M}, C_{\gamma,b,M} > 0$ such that*

$$\begin{aligned} \sup_{\theta_0 \in \Theta^\beta(M)} \overline{a}_n &\leq C_{\beta,b,M} n^{1/(2\beta+1)} (\log n)^{-1-1/(2\beta+1)}, \\ \sup_{\theta_0 \in A^\gamma(M)} \overline{a}_n &\leq C_{\gamma,b,M}, \\ \inf_{\theta_0 \in \Theta^\beta(m,M)} \underline{a}_n &\geq C_{\beta,B,m} n^{1/(2\beta+1)} (\log n)^{-1-2/(2\beta+1)}. \end{aligned}$$

Proof. Let us start with the proof of the first inequality. We show that for any $b > 0$ the inequality $h_n(a, \theta_0) < b$ holds for $a \geq C_{\beta,b,M} n^{1/(2\beta+1)} (\log n)^{-1-1/(2\beta+1)}$. Let us introduce the notation $I_a \equiv a \log(n/a)$. Note that by using the inequalities

$ae^{i/a}+n \geq n$ and $ae^{i/a}+n \geq ae^{i/a}$, for all $a \geq 1$, and the sum of geometric series we get

$$\begin{aligned} h_n(a,\theta_0) &\leq \frac{M}{\log^2\left(\frac{n}{a}\right)}\left(\frac{1}{a}\sum_{i=1}^{I_a} e^{i/a} i^{-2\beta} + \frac{n^2}{a^3}\sum_{i>I_a} e^{-i/a} i^{-2\beta}\right) \\ &\leq C_\beta \frac{M}{\log^2\left(\frac{n}{a}\right)}\left(I_a^{-2\beta} e^{I_a/a} + \frac{n^2}{a^2} I_a^{-2\beta} e^{-I_a/a}\right) \\ &\leq 2C_\beta M a^{-1-2\beta} n \left(\log\left(\frac{n}{a}\right)\right)^{-2-2\beta}, \end{aligned}$$

for some constant $C_\beta > 0$ depending only on β. For any $a > 0$ such that $A_n \geq a \geq Kn^{1/(2\beta+1)}(\log n)^{-1-1/(2\beta+1)}$ the preceding display is bounded by a multiple of $2C_\beta M K^{-1-2\beta}$. Then for sufficiently large choice of the constant $K = C_{\beta,b,M}$ (depending only on β, b and M), we get that $h_n(a,\theta_0) < b$ for any a larger than $C_{\beta,b,M} n^{1/(1+2\beta)}(\log n)^{-1-1/(2\beta+1)}$.

The proof of the second inequality of the statement goes similarly, i.e. we prove that for $a \geq C_{\gamma,b,M}$ we have $h_n(a,\theta_0) < b$. Note that by the sum of geometric series we get for every $a \geq 1/\gamma$

$$\begin{aligned} h_n(a,\theta_0) &\leq \frac{M}{\log^2\left(\frac{n}{a}\right)}\left(\frac{1}{a}\sum_{i=1}^{I_a} i e^{i/a} e^{-2\gamma i} + \frac{n^2}{a^3}\sum_{i>I_a} i e^{-i/a} e^{-2\gamma i}\right) \\ &\leq \frac{M}{a\log^2\left(\frac{n}{a}\right)}\sum_{i=1}^{\infty} i e^{-\gamma i} \leq \frac{M}{a(1-e^{-\gamma})^2\log^2\left(\frac{n}{a}\right)}, \end{aligned}$$

which is bounded from above by arbitrarily small b for sufficiently large choice of the constant $C_{\gamma,b,M}$.

Finally we deal with the lower bound for $\underline{a}_n$. Note that by using the inequalities $ae^{i/a}+n \geq n$ and $ae^{i/a}+n \geq ae^{i/a}$, for all $a \geq 1$, and the sum of geometric series we get

$$\begin{aligned} g_n(a,\theta_0) &\geq \frac{m}{4\log^2\left(\frac{n}{a}\right)}\frac{n^2}{a^3}\sum_{i>I_a}(i-a)e^{-i/a} i^{-2\beta} \\ &\geq c_\beta \frac{m}{\log^2\left(\frac{n}{a}\right)}\frac{n^2}{a^2} I_a^{-2\beta} e^{-I_a/a} \\ &= c_\beta m a^{-1-2\beta} n\big(\log(n/a)\big)^{-2-2\beta}, \end{aligned}$$

for some $c_\beta > 0$ depending only on β. For $1 \leq a \leq Kn^{1/(2\beta+1)}(\log n)^{-1-2/(2\beta+1)}$ the preceding display is bounded by a multiple of $mc_\beta K^{-1-2\beta}\log n$. Then for sufficiently small choice of the constant $C_{\beta,B,m}$, we get that $g_n(a,\theta_0) \geq B\log n$ for any $a \leq C_{\beta,B,m} n^{1/(2\beta+1)}(\log n)^{-1-2/(2\beta+1)}$. □

In the next lemma we show that under the polished tail condition the deterministic bounds $\underline{a}_n, \overline{a}_n$ are close to each other.

Lemma 2.4.3. *For every $L_0, \rho, N_0 \geq 1$ we have*

$$\sup_{\theta_0 \in \Theta_{pt}(L_0, N_0, \rho)} \frac{\overline{a}_n \log\left(\frac{n}{\overline{a}_n}\right)}{\underline{a}_n \log\left(\frac{n}{\underline{a}_n}\right)} \leq K \log^2 n,$$

with $K = 8.1e^4\rho^2 L_0 B/b$ for n large enough.

Proof. First of all note that since $\underline{a}_n \leq \overline{a}_n$, there is nothing to prove in the trivial cases $\underline{a}_n = A_n$ or $\overline{a}_n = K_0$. Hence $h_n(\overline{a}_n, \theta_0) \leq b$ and $g_n(a, \theta_0) < B \log n$, for all $a > \underline{a}_n$, hold. Furthermore assume that $\underline{a}_n \leq \rho^{-2}\overline{a}_n$, else the statement is trivial.

Let us divide the interval $[\rho^j, \rho^{j+1})$ into sub-intervals $[\rho^{j+\frac{k}{\lceil \log n \rceil}}, \rho^{j+\frac{k+1}{\lceil \log n \rceil}})$, $k = 0, 1, ...\lceil \log n \rceil - 1$, and introduce the notation

$$k_j = \operatorname{argmax}_{k=0,...,\lceil \log n \rceil - 1} \vartheta_{j,k}, \quad \text{where } \vartheta_{j,k} = \sum_{i=\rho^{j+k/\lceil \log n \rceil}}^{\rho^{j+(k+1)/\lceil \log n \rceil}} \theta_{0,i}^2,$$

with the notational convenience $\sum_{i=a}^{b} c_i = \sum_{i=\lceil a \rceil}^{\lfloor b \rfloor} c_i$, applied later on as well.

Then by the polished tail condition

$$\sum_{i=\rho^j}^{\infty} \theta_{0,i}^2 \leq L_0 \sum_{i=\rho^j}^{\rho^{j+1}} \theta_{0,i}^2 \leq L_0 \log(n) \vartheta_{j,k_j},$$

for $j \geq \log_\rho N_0$. Note that for every $a > 0$ there exists an $\tilde{a} \in (a, \rho^2 a)$ such that

$$I_{\tilde{a}} \equiv \tilde{a} \log(n/\tilde{a}) \in \left[\rho^{j+\frac{k_j}{\lceil \log n \rceil}}, \rho^{j+\frac{k_j+1}{\lceil \log n \rceil}}\right) \tag{2.4.5}$$

for some $j \in \mathbb{N}$ and let us denote this j by $J_{\tilde{a}}$. Then

$$\sum_{i=e^{-1/\log n} I_{\tilde{a}}}^{e^{1/\log n} I_{\tilde{a}}} \theta_{0,i}^2 \geq \vartheta_{J_{\tilde{a}}, k_{J_{\tilde{a}}}}.$$

Let us take any $a_1 \leq \rho^{-2} a_2$ and denote by $\tilde{a}_1 \in (a_1, \rho^2 a_1)$ the value satisfying (2.4.5). Then in view of $\exp\{e^{1/\log n} \log(n/a)\} \leq \exp\{(1+2/\log n)\log(n/a)\} \leq e^2 n/a$, for $n \geq e$, combined with the previous inequalities we get that the ratio $h_n(a_2, \theta_0)/h_n(\tilde{a}_1, \theta_0)$ is bounded from above by

$$\begin{aligned}
&\frac{\tilde{a}_1 \log^2\left(\frac{n}{\tilde{a}_1}\right)}{a_2 \log^2\left(\frac{n}{a_2}\right)} 4e^2 \frac{\sum_{i=1}^{I_{\tilde{a}_1}} i e^{i/a_2} \theta_{0,i}^2 + \sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} \theta_{0,i}^2 + \frac{n^2}{a_2^2} \sum_{i=I_{a_2}}^{\infty} i e^{-i/a_2} \theta_{0,i}^2}{\sum_{i=1}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} \theta_{0,i}^2} \\
&\leq \frac{\tilde{a}_1 \log^2\left(\frac{n}{\tilde{a}_1}\right)}{a_2 \log^2\left(\frac{n}{a_2}\right)} 4e^2 \left(1 + \frac{\sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} \theta_{0,i}^2 + n \log\left(\frac{n}{a_2}\right) \sum_{i=I_{a_2}}^{\infty} \theta_{0,i}^2}{\sum_{i=e^{-1/\log n} I_{\tilde{a}_1}}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} \theta_{0,i}^2}\right).
\end{aligned}$$

Since $ie^{i/\tilde{a}_1} > e^{-2}n\log(n/\tilde{a}_1)$ for $i \geq e^{-1/\log n}I_{\tilde{a}_1}$, and $ie^{i/a_2} \leq n\log(n/a_2)$ for $i \leq I_{a_2}$, we can see that

$$\frac{\sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} ie^{i/a_2}\theta_{0,i}^2 + n\log(\frac{n}{a_2})\sum_{i=I_{a_2}}^{\infty}\theta_{0,i}^2}{\sum_{i=e^{-1/\log n}I_{\tilde{a}_1}}^{e^{1/\log n}I_{\tilde{a}_1}} ie^{i/\tilde{a}_1}\theta_{0,i}^2} \leq e^2\frac{\log\left(\frac{n}{a_2}\right)}{\log\left(\frac{n}{\tilde{a}_1}\right)}\frac{\sum_{i=I_{\tilde{a}_1}}^{\infty}\theta_{0,i}^2}{\sum_{i=e^{-1/\log n}I_{\tilde{a}_1}}^{e^{1/\log n}I_{\tilde{a}_1}}\theta_{0,i}^2}.$$

Moreover, since

$$\sum_{i=I_{\tilde{a}_1}}^{\infty}\theta_{0,i}^2 \leq L_0\log(n)\vartheta_{J_{\tilde{a}_1},k_{J_{\tilde{a}_1}}} \leq L_0\log(n)\sum_{i=e^{-1/\log n}I_{\tilde{a}_1}}^{e^{1/\log n}I_{\tilde{a}_1}}\theta_{0,i}^2,$$

combined with the preceding computations we get that

$$\frac{h_n(a_2,\theta_0)}{h_n(\tilde{a}_1,\theta_0)} \leq 4e\frac{\tilde{a}_1\log^2\left(\frac{n}{\tilde{a}_1}\right)}{a_2\log^2\left(\frac{n}{a_2}\right)}\left(1 + L_0e^2\frac{\log\left(\frac{n}{a_2}\right)}{\log\left(\frac{n}{\tilde{a}_1}\right)}\log n\right). \tag{2.4.6}$$

Furthermore, let us note that for any $\underline{a}_n < a \leq A_n$

$$h_n(a,\theta_0) \leq 2g_n(a,\theta_0) + \frac{2e^2}{\log^2\left(\frac{n}{a}\right)}\sum_{i=1}^{2a}\theta_{0,i}^2 \leq 2B\log(n) + o(1).$$

Then by taking $a_1 = \underline{a}_n$, $\tilde{a}_1 \in (\underline{a}_n, \rho^2\underline{a}_n)$, and $a_2 = \overline{a}_n$ in (2.4.6) we get that

$$\begin{aligned}\frac{b}{2B\log(n)+o(1)} \leq \frac{h_n(\overline{a}_n,\theta_0)}{h_n(\tilde{a}_1,\theta_0)} &\leq 4e^4(1+o(1))L_0\log(n)\frac{\tilde{a}_1\log(\frac{n}{\tilde{a}_1})}{\overline{a}_n\log(\frac{n}{\overline{a}_n})} \\ &\leq 4e^4\rho^2(1+o(1))L_0\log(n)\frac{\underline{a}_n\log(\frac{n}{\underline{a}_n})}{\overline{a}_n\log(\frac{n}{\overline{a}_n})}.\end{aligned}$$

After rearranging the preceding inequality we arrive to our statement. □

§2.4.2 Contraction rates

In this section we provide the contraction rate results both for the empirical and hierarchical Bayes procedures. First we show that the empirical Bayes method achieves the (up to a logarithmic factor) optimal minimax contraction rate around the truth for unknown regularity hyper-parameter $\beta > 0$.

Theorem 2.4.4. *The maximum marginal likelihood empirical Bayes posterior corresponding to the prior* (2.1.2) *achieves the minimax adaptive contraction rate (up to a logarithmic factor), i.e. for given $M, \beta_0 > 0$ we have*

$$\sup_{\beta\geq\beta_0}\sup_{\theta\in\Theta^\beta(M)} E_0\left[\Pi_{\hat{a}_n}(\|\theta-\theta_0\|_2 \geq M_n\left(\frac{n}{\log^2 n}\right)^{-\beta/(2\beta+1)}|Y)\right] \to 0, \tag{2.4.7}$$

for any sequence M_n tending to infinity.

Proof. See Section 2.4.3. □

Using our findings on the empirical Bayes method we can extend the results on the hierarchical Bayes method, derived in (van der Vaart and van Zanten, 2009a) and (Bhattacharya and Pati, 2015) (where typically inverse gamma hyper-prior was considered), by allowing other, more general choices of the hyper-prior distribution as well.

Theorem 2.4.5. *Let us assume that the hyper-prior π satisfies Assumption 2.1.1. Then the corresponding hierarchical Bayes posterior achieves the minimax contraction rate (up to a logarithmic factor), i.e. for given $\beta_0, M > 0$ we have*

$$\sup_{\beta\geq\beta_0}\sup_{\theta\in\Theta^\beta(M)} E_0\left[\Pi(\|\theta-\theta_0\|_2 \geq M_n\left(\frac{n}{\log^2 n}\right)^{-\beta/(2\beta+1)}|Y)\right]\to 0, \tag{2.4.8}$$

for some arbitrary sequence M_n tending to infinity.

Proof. See Section 2.10.1. □

§2.4.3 Proof of Theorem 2.4.4

Let us introduce the shorthand notation

$$\varepsilon_n := n^{-\beta/(2\beta+1)}(\log n)^{2\beta/(2\beta+1)}.$$

In view of Markov's inequality and Theorem 2.4.1, for every $\beta > 0$

$$\sup_{\theta_0\in\Theta^\beta(M)} E_0[\Pi_{\hat{a}_n}(\|\theta-\theta_0\|_2 \geq M_n\varepsilon_n|Y)] \leq \frac{1}{M_n^2\varepsilon_n^2}\sup_{\theta_0\in\Theta^\beta(M)} E_0[\sup_{a\in[\underline{a}_n,\overline{a}_n]} R_n(a)] + o(1), \tag{2.4.9}$$

where

$$R_n(a) = \int \|\theta-\theta_0\|_2^2\Pi_a(d\theta|Y)$$

is the posterior risk. We show below that both

$$\sup_{\theta_0\in\Theta^\beta(M)}\sup_{a\in[\underline{a}_n,\overline{a}_n]} E_0[R_n(a)] = O(\varepsilon_n^2) \qquad \text{and} \tag{2.4.10}$$

$$\sup_{\theta_0\in\Theta^\beta(M)} E_0[\sup_{a\in[\underline{a}_n,\overline{a}_n]} |R_n(a) - E_0(R_n(a))|] = o(\varepsilon_n^2) \tag{2.4.11}$$

hold, which results in that the right-hand side of (2.4.9) vanishes as $n\to\infty$, concluding the proof of Theorem 2.4.4.

2.4.3.1 Bound for the expected posterior risk (2.4.10)

First, note that by elementary computations

$$R_n(a) = \sum_{i=1}^{\infty}(\hat{\theta}_{a,i}-\theta_{0,i})^2 + \sum_{i=1}^{\infty}\frac{1}{ae^{i/a}+n},$$

where $\hat{\theta}_{a,i} = n(ae^{i/a}+n)^{-1}Y_i$ is the ith coefficient of the posterior mean. Therefore the expectation of $R_n(a)$ is given by

$$E_0 R_n(a) = \sum_{i=1}^{\infty} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 + \sum_{i=1}^{\infty} \frac{n}{(ae^{i/a}+n)^2} + \sum_{i=1}^{\infty} \frac{1}{ae^{i/a}+n}. \tag{2.4.12}$$

Note that the second and third terms do not contain θ_0, and that the second term is bounded by the third. By Lemma 2.11.2 (with $r=0$ and $l=1$) and Proposition 2.4.2 the latter is further bounded for $a \le \overline{a}_n$ by a multiple of

$$\frac{a}{n}\log\left(\frac{n}{a}\right) \le \frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right) \le C_{\beta,b,M} n^{-2\beta/(2\beta+1)} (\log n)^{-1/(2\beta+1)},$$

since the function $a \mapsto a\log(n/a)$ is monotone increasing for $a \le n/e$. It remained to deal with the first term on the right hand side of (2.4.12), which we divide into three parts and show that each of the parts have the stated order. First note that for $\theta_0 \in \Theta^\beta(M)$

$$\begin{aligned}\sum_{i=(n/\log^2 n)^{1/(2\beta+1)}}^{\infty} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\le \sum_{i=(n/\log^2 n)^{1/(2\beta+1)}}^{\infty} M i^{-1-2\beta} \\ &\le \frac{M}{2\beta}\left(\frac{n}{\log^2 n}\right)^{-2\beta/(2\beta+1)}.\end{aligned}$$

Next note that for $a \le \overline{a}_n$, in view of Proposition 2.4.2,

$$\begin{aligned}\sum_{i=1}^{2a} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\le \sum_{i=1}^{2a} \frac{a^2 e^{2i/a}}{n^2}\theta_{0,i}^2 \le \frac{a^2 e^4}{n^2}\sum_{i=1}^{2a}\theta_{0,i}^2 \\ &\le e^4 \frac{\overline{a}_n^2}{n^2} \le e^4 M C_{\beta,b,M}^2 n^{-4\beta/(2\beta+1)} (\log n)^{-2-2/(2\beta+1)}.\end{aligned}$$

Furthermore, notice that the maximum of the function $i \mapsto e^{i/a}/(i-a)$ over $[2a, I_a]$ is attained at $i = I_a$, because the function is increasing for $i > 2a$ and $n > 0$. Besides, for $a > \underline{a}_n$ we have $g_n(a, f_0) < B\log n$, hence for any $\underline{a}_n < a \le \bar{a}_n$

$$\begin{aligned}\sum_{i=2a}^{I_a} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\le \frac{a}{n}\frac{\log^2\left(\frac{n}{a}\right)}{\left(\log\left(\frac{n}{a}\right)-1\right)} \sum_{i=2a}^{I_a} \frac{n^2 e^{i/a}(i-a)}{a\log^2\left(\frac{n}{a}\right)(ae^{i/a}+n)^2}\theta_{0,i}^2 \\ &\le 2\overline{a}_n n^{-1}\log\left(\frac{n}{\overline{a}_n}\right) g_n(a,\theta_0) \\ &\le 2\overline{a}_n n^{-1}\log^2 n \le 2C_{\beta,b,M} n^{-2\beta/(2\beta+1)} (\log n)^{2\beta/(2\beta+1)},\end{aligned}$$

where the last inequality follows from Proposition 2.4.2.

It remained to deal with the terms between the term $I_{\underline{a}_n} = \underline{a}_n \log(n/\underline{a}_n)$ and the term $(n/\log^2 n)^{1/(2\beta+1)}$. Let $J = J(n)$ be the smallest integer such that

$$\left(1+\frac{1}{\log n}\right)^J \underline{a}_n \log\left(\frac{n}{\underline{a}_n}\right) \ge \left(\frac{n}{\log^2 n}\right)^{1/(2\beta+1)}$$

and let

$$n_j := \left(1 + \frac{1}{\log n}\right)^j I_{\underline{a}_n}.$$

Note that the sequence n_j is increasing. For notational convenience, we also introduce b_j such that $b_j e^{n_j/b_j} = n$ and $b_j < n_j$. Now we have for any $a \geq 1$

$$\begin{aligned}\sum_{i=I_{\underline{a}_n}}^{(n/\log^2 n)^{1/(2\beta+1)}} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\leq \sum_{j=0}^{J-1}\sum_{i=n_j}^{n_{j+1}} \theta_{0,i}^2 \\ &\leq 4e^2 \sum_{j=0}^{J-1}\sum_{i=n_j}^{n_{j+1}} \frac{nb_j e^{i/b_j}}{(b_j e^{i/b_j}+n)^2}\theta_{0,i}^2. \qquad (2.4.13)\end{aligned}$$

By elementary computations we get that $b_j \asymp n_j/\log n_j$, therefore (2.4.13) is further bounded by constant times

$$\frac{1}{n}\sum_{j=0}^{J-1}\frac{1}{\log n_j}\sum_{i=n_j}^{n_{j+1}}\frac{n^2(i-b_j)e^{i/b_j}}{(b_j e^{i/b_j}+n)^2}\theta_{0,i}^2 \leq \frac{1}{n}\sum_{j=0}^{J-1}\frac{b_j \log^2 n}{\log n_j} g_n(b_j,\theta_0).$$

Since $b_j \geq \underline{a}_n$ we have $g_n(b_j,\theta_0) \leq B\log n$ for all $j \geq 0$. Then by the sum of geometric series we get that

$$\begin{aligned}\frac{1}{n}\sum_{j=0}^{J-1}\frac{n_j}{\log^2 n_j}\log^3 n &\leq 2(1+2\beta)^2\frac{\log n}{n}\frac{I_{\underline{a}_n}\left(1+\frac{1}{\log n}\right)^J}{\frac{1}{\log n}} \\ &\leq 2(1+2\beta)^2 n^{-2\beta/(2\beta+1)}(\log n)^{2-2/(2\beta+1)},\end{aligned}$$

concluding the proof of assertion (2.4.10).

2.4.3.2 Bound for the centered posterior risk (2.4.11)

Note that

$$\begin{aligned}&R_n(a) - E_0 R_n(a) = \mathbb{V}(a)/n - 2\mathbb{W}(a)/\sqrt{n}, \quad \text{where} \\ &\mathbb{V}(a) = n^2\sum_{i=1}^{\infty}\frac{1}{(ae^{i/a}+n)^2}(Z_i^2-1), \text{ and } \mathbb{W}(a) = n\sum_{i=1}^{\infty}\frac{ae^{i/a}\theta_{0,i}}{(ae^{i/a}+n)^2}Z_i.\end{aligned}$$

Therefore it is sufficient to show that there exists a constant $K = K_{\beta,M,b,B} > 0$ such that

$$\begin{aligned}E_0\left(\sup_{a\in[\underline{a}_n,\overline{a}_n]}\frac{|\mathbb{V}(a)|}{n}\right) &\leq Kn^{-2\beta/(2\beta+1)}(\log n)^{-1/(2\beta+1)}, \\ E_0\left(\sup_{a\in[\underline{a}_n,\overline{a}_n]}\frac{|\mathbb{W}(a)|}{\sqrt{n}}\right) &\leq Kn^{-2\beta/(1+2\beta)}.\end{aligned}$$

We deal with the two processes above, separately.

For the process $\mathbb{V}$, Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) implies that

$$E_0\left[\sup_{a\in[\underline{a}_n,\overline{a}_n]}|\mathbb{V}(a)|\right]\lesssim\sup_{a\in[\underline{a}_n,\overline{a}_n]}\sqrt{V_0(\mathbb{V}(a))}+\int_0^{diam_n}\sqrt{N(\varepsilon,[\underline{a}_n,\overline{a}_n],d_n)}d\varepsilon,$$

where $d_n^2(a_1,a_2)=V_0(\mathbb{V}(a_1)-\mathbb{V}(a_2))$, $diam_n$ is the d_n-diameter of $[\underline{a}_n,\overline{a}_n]$ and $N(\varepsilon,B,d_n)$ the covering number of the set B with ε-radius balls relative to the d_n semi-metric. The variance of $\mathbb{V}(a)$ is equal to

$$V_0(\mathbb{V}(a))=2n^4\sum_{i=1}^{\infty}\frac{1}{(ae^{i/a}+n)^4}$$

since $V(Z_i^2)=2$. Using Lemma 2.11.2 (with $r=0$ and $l=4$) we can conclude that the variance of $\mathbb{V}(a)$ is bounded from above by a multiple of $a\log(n/a)$, hence $diam_n\lesssim\sqrt{\overline{a}_n\log n}$. In view of Lemma 2.4.6, the distance $d_n(a_1,a_2)$ is bounded from above by a multiple of $|a_1-a_2|\log^{3/2}n$, hence the interval $[\underline{a}_n,\overline{a}_n]$ can be covered with constant times $\overline{a}_n\varepsilon^{-1}\log^{3/2}n$ amount of ε-balls relative to the d_n semi-metric. In view of the above computation and Proposition 2.4.2

$$E_0\left[\frac{1}{n}\sup_{a\in[\underline{a}_n,\overline{a}_n]}|\mathbb{V}(a)|\right]\lesssim\frac{\overline{a}_n}{n}\log n\le C_{\beta,b,M}n^{-2\beta/(2\beta+1)}(\log n)^{-1/(2\beta+1)}.$$

The process $\mathbb{W}$ can be dealt with similarly to $\mathbb{V}$. The main difference is the bounding of the variance of $\mathbb{W}$, which we describe in details. First note that

$$V_0\left(\frac{\mathbb{W}(a)}{\sqrt{n}}\right)=n\sum_{i=1}^{\infty}\frac{a^2e^{2i/a}}{(ae^{i/a}+n)^4}\theta_{0,i}^2.$$

Let us split the sum at I_a and by applying the inequality $ae^{i/a}+n\ge n$, we get

$$n\sum_{i=1}^{I_a}\frac{a^2e^{2i/a}}{(ae^{i/a}+n)^4}\theta_{0,i}^2\le\frac{1}{n^3}\sum_{i=1}^{I_a}a^2e^{2i/a}\theta_{0,i}^2\le\frac{\|\theta_0\|_2^2}{n}.$$

Then by noting that the function $i\mapsto e^{i/a}/((i-a)(ae^{i/a}+n)^2)$ is decreasing on $[I_a,\infty)$, recalling that $g_n(a,\theta_0)\le B\log n$, for all $a\ge\underline{a}_n$, and in view of Proposition 2.4.2, for $a\le\overline{a}_N$

$$\begin{aligned}n\sum_{i=I_a}^{\infty}\frac{a^2e^{2i/a}}{(ae^{i/a}+n)^4}\theta_{0,i}^2&\le\frac{a\log^2\left(\frac{n}{a}\right)}{4n^2\left(\log\left(\frac{n}{a}\right)-1\right)}\sum_{i=I_a}^{\infty}\frac{n^2(i-a)e^{i/a}}{a\log^2\left(\frac{n}{a}\right)(ae^{i/a}+n)^2}\theta_{0,i}^2\\&\le an^{-2}\log\left(\frac{n}{a}\right)g_n(a,\theta_0)\le B\overline{a}_nn^{-2}\log^2n\\&\le 2BC_{\beta,b,M}n^{\frac{-(4\beta+1)}{2\beta+1}}(\log n)^{2\beta/(2\beta+1)},\end{aligned}$$

hence $\mathrm{diam}_n=O(n^{-\frac{1/2+2\beta}{1+2\beta}}(\log n)^{\beta/(1+2\beta)})$. Then in view of Lemma 2.4.6 the covering number of the interval $[\underline{a}_n,\overline{a}_n]$ is bounded by $C_M\varepsilon^{-1}(\overline{a}_n/\sqrt{n})\log n$ with respect to the semi-metric $d_n(a_1,a_2)=V_0\big(\mathbb{W}(a_1)/\sqrt{n}-\mathbb{W}(a_2)/\sqrt{n}\big)$ and the rest of the proof goes as above.

2.4.3.3 Bounds for the semi-metrics associated to $\mathbb{V}$ and $\mathbb{W}$

Lemma 2.4.6. *For any $1 \le a_1 \le a_2$ and $f_0 \in \ell_2(M)$ we have*

$$V_0(\mathbb{V}(a_1) - \mathbb{V}(a_2)) \lesssim (a_1 - a_2)^2 \log^3 n,$$
$$V_0(\mathbb{W}(a_1) - \mathbb{W}(a_2)) \lesssim (a_1 - a_2)^2 \log^2 n,$$

with constants only depending on M.

Proof. The left-hand side of the first inequality is equal to

$$n^4 \sum_{i=1}^{\infty} (\phi_i(a_1) - \phi_i(a_2))^2 V(Z_i^2),$$

where $\phi_i(a) = (ae^{i/a} + n)^{-2}$. The square of the derivative of ϕ_i is given by $\phi_i'(a)^2 = 4\phi_i(a)^3 e^{2i/a}(i-a)^2/a^2$, hence in view of Lemma 2.11.3 the preceding display is bounded above by a multiple of

$$(a_1 - a_2)^2 \sup_{a \in [a_1, a_2]} \sum_{i=1}^{\infty} \frac{n^4 e^{2i/a}(i-a)^2}{a^2(ae^{i/a} + n)^6} \le (a_1 - a_2)^2 n^4 \sup_{a \in [a_1, a_2]} \sum_{i=1}^{\infty} \frac{e^{2i/a}(i^2 + a^2)}{a^2(ae^{i/a} + n)^6}$$
$$\lesssim (a_1 - a_2)^2 \sup_{a \in [a_1, a_2]} \frac{\log\left(\frac{n}{a}\right)}{a} \left(1 + \log^2\left(\frac{n}{a}\right)\right)$$

with the help of Lemma 2.11.1 (first with $m = 2$ and then with $m = 0$), and Lemma 2.11.2 (with $r = 1$ and $l = 4$).

We next consider the process $\mathbb{W}(a)$. The left-hand side of the second inequality in the statement of the lemma is equal to

$$n^2 \sum_{i=1}^{\infty} (\phi_i(a_1) - \phi_i(a_2))^2 \theta_{0,i}^2 V_0(Z_i),$$

with $\phi_i(a) = ae^{i/a}/(ae^{i/a} + n)^2$. Note that $|\phi_i'(a)| \le (i + a)a^{-2}\phi_i(a)$, hence in view of Lemma 2.11.1 (first with $m = 2$ and then with $m = 0$) and Lemma 2.11.3 the preceding display is bounded by

$$4(a_1 - a_2)^2 n^2 \sup_{a \in [a_1, a_2]} \frac{1}{a^2} \sum_{i=1}^{\infty} \frac{e^{2i/a}(\frac{i^2}{a^2} + 1)}{(ae^{i/a} + n)^4} \theta_{0,i}^2$$
$$\le 4(a_1 - a_2)^2 \sup_{a \in [a_1, a_2]} \frac{1}{a^2} \left(\log^2\left(\frac{n}{a}\right) + 1\right) \|\theta_0\|_2^2,$$

concluding the proof of the lemma. □

§2.5 Proof of the empirical Bayes part of Theorem 2.1.5

First note that we get the empirical Bayes credible set by plugging in the estimator $\hat{a}_n$ into the credible ball $\hat{C}_{a,\alpha}$ defined as

$$\hat{C}_{a,\alpha} = \{\theta \in L_2 : \|\theta - \hat{\theta}_a\|_2 \le L r_\alpha\}$$

satisfying that

$$\Pi_a(\hat{C}_{a,\alpha}|Y) = 1-\alpha,$$

where $\hat{\theta}_a$ is the posterior mean for fixed hyper-parameter $a > 0$.

The proof of the statement is then based on the deterministic bounds for the MMLE $\hat{a}_n$ derived in Theorem 2.4.1 and their distance investigated in Lemma 2.4.3.

Note that $\theta_0 \in \hat{C}_n(L\log^{3/2} n)$ if and only if $\|\theta_0 - \hat{\theta}_{\hat{a}_n}\|_2 \le L(\log n)^{3/2} r_\alpha$. Therefore by triangle inequality it is sufficient to verify that

$$\|W(\hat{a}_n)\|_2 \le L(\log n)^{3/2} r_\alpha(\hat{a}_n) - \|B(\hat{a}_n,\theta_0)\|_2 \tag{2.5.1}$$

holds with high probability, where $W(a) = \hat{\theta}_a - E_0\hat{\theta}_a$ and $B(a,\theta_0) = E_0\hat{\theta}_a - \theta_0$ are the centered posterior mean and the bias of the posterior mean for fixed hyper-parameter $a > 0$, respectively. Note that the ith coefficient of these vectors take the form

$$W_i(a) = \frac{n(Y_i - \theta_{0,i})}{ae^{i/a}+n}, \quad \text{and} \quad B_i(a,\theta_0) = \frac{ae^{i/a}\theta_{0,i}}{ae^{i/a}+n}.$$

We prove below that there exist constants $C_1, C_2 > 0$ depending on ρ, L_0, B and b such that for large enough n,

$$\inf_{\underline{a}_n \le a \le \overline{a}_n} r_\alpha^2(a) \ge \frac{\underline{a}_n}{3n}\log\left(\frac{n}{\underline{a}_n}\right), \tag{2.5.2}$$

$$\inf_{\theta_0 \in \Theta_{pt}(L_0,N_0,\rho)} P_0\left(\sup_{\underline{a}_n \le a \le \overline{a}_n} \|W(a)\|_2^2 \le C_1 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\log^2 n\right) \to 1, \tag{2.5.3}$$

$$\sup_{\theta_0 \in \Theta_{pt}(L_0,N_0,\rho)} \sup_{\underline{a}_n \le a \le \underline{a}_n} \|B(a,\theta_0)\|_2^2 \le C_2 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\log^3 n. \tag{2.5.4}$$

Hence in view of Theorem 2.4.1 assertion (2.5.1) holds with probability tending to one for large enough choice of L, under the polished tail assumption.

Proof of (2.5.2): The radius $r_\alpha(a)$, given in (2.1.6), is defined as $P(U_n(a) < r_\alpha^2(a)) = 1-\alpha$ with $U_n(a) := \sum_{i=1}^{\infty} \frac{1}{ae^{i/a}+n} Z_i^2$, where Z_i's are iid $N(0,1)$. We show below that

$$\liminf_{n\to\infty} \inf_{a\in[\underline{a}_n,\overline{a}_n]} E\left[\frac{nU_n(a)}{a\log\left(\frac{n}{a}\right)}\right] > \frac{1}{2}, \tag{2.5.5}$$

$$E\left[\sup_{a\in[\underline{a}_n,\overline{a}_n]} \frac{n|U_n(a) - E[U_n(a)]|}{a\log\left(\frac{n}{a}\right)}\right] \to 0. \tag{2.5.6}$$

Then by Markov's inequality with probability tending to one we have

$$\inf_{a\in[\underline{a}_n,\overline{a}_n]} \frac{nU_n(a)}{a\log\left(\frac{n}{a}\right)} > 1/3,$$

hence (2.5.2) follows from the definition of $r_\alpha(a)$.

Assertion (2.5.5) follows as

$$E[U_n(a)] \geq \sum_{i=1}^{I_a} \frac{1}{ae^{i/a}+n} \geq \frac{I_a}{2n} \geq \frac{a}{2n} \log\left(\frac{n}{a}\right).$$

To verify (2.5.6), it suffices by Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) (applied with $\psi(x) = x^2$) to show that there exist $K_1, K_2 > 0$ such that for any $a \in [\underline{a}_n, \overline{a}_n]$

$$V\left(\frac{nU_n(a)}{a\log\left(\frac{n}{a}\right)}\right) \leq K_1 \frac{1}{a\log\left(\frac{n}{a}\right)}, \tag{2.5.7}$$

$$\int_0^{diam_n} \sqrt{N(\varepsilon, [\underline{a}_n, \overline{a}_n], d_n)}d\varepsilon \leq \sqrt{A_n/n} = o(1), \tag{2.5.8}$$

where d_n is the semi-metric defined by $d_n^2(a_1, a_2) := V\left(\frac{nU_n(a_1)}{a_1\log(n/a_1)} - \frac{nU_n(a_2)}{a_2\log(n/a_2)}\right)$, $diam_n$ is the diameter of the interval $[\underline{a}_n, \overline{a}_n]$ relative to d_n and $N(\varepsilon, S, d_n)$ is the minimal number of d_n-balls of radius ε needed to cover the set S.

First note that in view of Lemma 2.11.2 (with $r = 0$ and $l = 2$) we have

$$V\left(\frac{nU_n(a)}{a\log\left(\frac{n}{a}\right)}\right) = \frac{2n^2}{a^2\log^2\left(\frac{n}{a}\right)} \sum_{i=1}^{\infty} \frac{1}{(ae^{i/a}+n)^2} \lesssim \frac{1}{a\log\left(\frac{n}{a}\right)}.$$

As a consequence one can see that $diam_n \lesssim (\underline{a}_n \log(n/\underline{a}_n))^{-1/2}$. By Lemma 2.5.1, $d_n(a_1, a_2) \lesssim a_1^{-3/2}\log^{1/2}(n/a_1)n^{-1}|a_1 - a_2|$, hence

$$N(\varepsilon, [\underline{a}_n, \overline{a}_n], d_n) \lesssim \varepsilon^{-1}\log^{1/2}\left(\frac{n}{\underline{a}_n}\right)\underline{a}_n^{-3/2}\frac{\overline{a}_n}{n}.$$

Therefore one can conclude that

$$\int_0^{diam_n} \sqrt{N(\varepsilon, [\underline{a}_n, \overline{a}_n], d_n)}d\varepsilon = \frac{\overline{a}_n^{1/2}\log^{1/4}\left(\frac{n}{\underline{a}_n}\right)}{\underline{a}_n^{3/4}n^{1/2}} \int_0^{C(\underline{a}_n\log(n/\underline{a}_n))^{-1/2}} \varepsilon^{-1/2}d\varepsilon \lesssim \sqrt{A_n/n}.$$

Proof of (2.5.3): The variable $\|W(a)\|_2^2$ is distributed as $\sum_{i=1}^{\infty} \frac{n}{(ae^{i/a}+n)^2} Z_i^2$, with $Z_i \overset{iid}{\sim} N(0,1)$. Observe that

$$E_0[\|W(a)\|_2^2] = \sum_{i=1}^{\infty} \frac{n}{(ae^{i/a}+n)^2}, \text{ and } V_0(\|W(a)\|_2^2) = 2\sum_{i=1}^{\infty} \frac{n^2}{(ae^{i/a}+n)^4}.$$

Furthermore note that by applying Lemma 2.11.2 (with $r = 0$ and $l = 2$) we get

$$\frac{a}{n}\log\left(\frac{n}{a}\right) \leq \frac{4I_a n}{(ae^{I_a/a}+n)^2} \leq \sum_{i=1}^{I_a} \frac{4n}{(ae^{i/a}+n)^2} \leq \sum_{i=1}^{\infty} \frac{4n}{(ae^{i/a}+n)^2} \leq C\frac{a}{n}\log\left(\frac{n}{a}\right),$$

for some universal constant $C > 0$, while by applying the same lemma (with $r = 0$ and $l = 4$) the variance is bounded above by a multiple of $an^{-2}\log(n/a)$. Then similar reasoning to the previous proof results in that

$$\inf_{\theta_0 \in \ell_2(M)} \left(\sup_{\underline{a}_n \le a \le \overline{a}_n} \|W(a)\|_2^2 \le C_2 \frac{\overline{a}_n}{n} \log\left(\frac{n}{\overline{a}_n}\right) \right) \xrightarrow{P_0} 1. \tag{2.5.9}$$

Then in view of Lemma 2.4.3, the right hand side of the inequality in the preceding probability statement is further bounded from above by constant multiplier of $(\underline{a}_n/n)\log(n/\underline{a}_n)\log^2 n$.

Proof of (2.5.4): First note that

$$\|B(a,\theta_0)\|_2^2 \le \sum_{i=1}^{I_a} n^{-2}a^2 e^{2i/a}\theta_{0,i}^2 + \sum_{i=I_a}^{\infty} \theta_{0,i}^2.$$

To bound the first term on the right hand side, we use the inequalities $a/n \le \log(n/a)$ for $a \le A_n$ and $\sum_{i=1}^{\infty}\theta_{0,i}^2 < \infty$, and furthermore note the function $i \mapsto e^{i/a}/(i-a)$ is monotone increasing on the interval $[2a, I_a]$ hence it takes its maximum at I_a. Therefore in view of Lemma 2.4.3 the first part of the bias for functions satisfying the polished tail condition is bounded by

$$\begin{aligned}
\sup_{\underline{a}_n \le a \le \overline{a}_n} \sum_{i=1}^{I_a} \frac{a^2 e^{2i/a}\theta_{0,i}^2}{n^2} &\le \sup_{\underline{a}_n \le a \le \overline{a}_n} \sum_{i=1}^{2a} \frac{a^2 e^{2i/a}\theta_{0,i}^2}{n^2} + \sup_{\underline{a}_n \le a \le \overline{a}_n} \frac{a}{n} \frac{\log^2\left(\frac{n}{a}\right)}{\left(\log\left(\frac{n}{a}\right)-1\right)} g_n(a,\theta_0) \\
&\le \frac{e^4 \overline{a}_n^2}{n^2} \sum_{i=1}^{2a} \theta_{0,i}^2 + (B+o(1))\frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right)\log n \\
&\le (B+o(1))\frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right)\log n \\
&\le K_{\rho,L_0,B,b}\frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\log^3 n,
\end{aligned}$$

for some constant $K_{\rho,L_0,B,b}$ depending on ρ, L_0, B, and b. Furthermore in view of the polished tail assumption we have

$$\sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2 \le L_0 \sum_{i=I_{\underline{a}_n}}^{\rho I_{\underline{a}_n}} \theta_{0,i}^2 \le \log\left(\frac{n}{\underline{a}_n}\right) \sum_{i=I_{\tilde{a}_n}}^{I_{\tilde{a}_n}+\rho\tilde{a}_n} \theta_{0,i}^2,$$

for some $\tilde{a}_n \in [\underline{a}_n, \rho\underline{a}_n]$. Therefore, by using Lemma 2.4.3,

$$\begin{aligned}
\sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2 &\lesssim \log\left(\frac{n}{\underline{a}_n}\right) \sum_{i=I_{\tilde{a}_n}}^{I_{\tilde{a}_n}+\rho\tilde{a}_n} \frac{n^2(i-\tilde{a}_n)e^{i/\tilde{a}_n}}{\tilde{a}_n \log^2\left(\frac{n}{\tilde{a}_n}\right)(\tilde{a}_n e^{i/\tilde{a}_n}+n)^2}\theta_{0,i}^2 \frac{\tilde{a}_n}{n}\log\left(\frac{n}{\tilde{a}_n}\right), \\
&\le \log\left(\frac{n}{\underline{a}_n}\right) g_n(\tilde{a}_n,\theta_0)\frac{\tilde{a}_n}{n}\log\left(\frac{n}{\tilde{a}_n}\right) \le K_{\rho,L_0,B,b}\log^2\left(\frac{n}{\underline{a}_n}\right)\frac{\underline{a}_n}{n}\log n,
\end{aligned}$$

for some large enough constant $K_{\rho,L_0,B,b} > 0$. Combining the two bounds, we see that (2.5.4) holds.

Lemma 2.5.1. *There exists a $K > 0$ such that for any $1 < a_1 < a_2$*

$$V\left(\frac{U_n(a_1)}{a_1 \log\left(\frac{n}{a_1}\right)} - \frac{U_n(a_2)}{a_2 \log\left(\frac{n}{a_2}\right)}\right) \leq K(a_1 - a_2)^2 \frac{\log\left(\frac{n}{a_1}\right)}{a_1^3 n^2}. \tag{2.5.10}$$

Proof. First note that

$$V\left(\frac{U_n(a_1)}{a_1 \log\left(\frac{n}{a_1}\right)} - \frac{U_n(a_2)}{a_2 \log\left(\frac{n}{a_2}\right)}\right) = 2\sum_{i=1}^{\infty}(\phi_i(a_1) - \phi_i(a_2))^2 \tag{2.5.11}$$

with $\phi_i(a) := \frac{1}{a\log(n/a)(ae^{i/a}+n)}$. The derivative of $\phi_i(a)$ is given as

$$\phi_i'(a) = \phi_i(a)\Big(\frac{2(i-a)e^{i/a}}{a(ae^{i/a}+n)} + \frac{1}{a\log(n/a)} - \frac{1}{a}\Big),$$

so we can see that $|\phi_i'(a)| \lesssim \big(\frac{(i+a)e^{i/a}}{a(ae^{i/a}+n)} \vee \frac{1}{a}\big)\phi_i(a)$. Thus in view of Lemma 2.11.3 the right hand side of (2.5.11) is bounded by a multiple of

$$(a_1 - a_2)^2 \sup_{a\in[a_1,a_2]} \sum_{i=1}^{\infty} \Big(\frac{(i^2+a^2)e^{2i/a}}{a^2(ae^{i/a}+n)^2} \vee \frac{1}{a^2}\Big)\phi_i(a)^2.$$

Then in view of Lemma 2.11.1 (first with $m = 2$ and then with $m = 0$) and Lemma 2.11.2 (first with $r = 1$ and $l = 2$ and second with $r = 0$ and $l = 2$) the preceding display is further bounded by the right hand side of (2.5.10), finishing the proof of the statement. □

§2.6 Proof of Theorem 2.1.4

We use the notations introduced in Section 2.5.

First recall that $\theta_0 \in \hat{C}_n(L)$ if and only if $\|\theta_0 - \hat{\theta}\|_2 \leq Lr_\alpha(\hat{a}_n)$. We show below that

$$\inf_{\theta_0\in A^\gamma(M)} P_0\left(\sup_{\underline{a}_n\leq a\leq \overline{a}_n} \|W(a)\|_2^2 \leq C_1 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\right) \to 1, \tag{2.6.1}$$

$$\sup_{\theta_0\in A^\gamma(M)} \sup_{\underline{a}_n\leq a\leq \underline{a}_n} \|B(a, f_0)\|_2^2 \leq C_2 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right), \tag{2.6.2}$$

for some constants $C_1, C_2 > 0$ depending only on M, which together with (2.5.2) and Theorem 2.4.1 results in the statement.

The proof of assertion (2.6.1) follows by combining (2.5.9) and the second inequality of Proposition 2.4.2. Next note that similarly to the proof of (2.5.4), we get that

$$\|B(a,\theta_0)\|_2^2 \leq \sum_{i=1}^{I_a} \frac{a^2 e^{2i/a}\theta_{0,i}^2}{n^2} + \sum_{i=I_a}^{\infty} \theta_{0,i}^2 \lesssim \frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right) + \sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2.$$

Furthermore

$$\sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2 = \sum_{i=I_{\underline{a}_n}}^{\infty} e^{-2i\gamma} e^{2i\gamma} \theta_{0,i}^2 \le M e^{-2I_{\underline{a}_n}\gamma} = M \left(\frac{\underline{a}_n}{n}\right)^{2\underline{a}_n\gamma} \le M \frac{\underline{a}_n}{n} \log\left(\frac{n}{\underline{a}_n}\right)$$

for $\gamma \ge 1/2$, finishing the proof of (2.6.2) and concluding the proof of the theorem.

§2.7 Proof of Theorem 2.1.3 and the empirical Bayes part of Corollary 2.1.2

In the proof we use again the notations introduced in Section 2.5.

First note that $\theta_0 \in \hat{C}_n(L_n)$ implies that $\|B(\hat{a}_n, \theta_0)\|_2 \le L_n r_\alpha(\hat{a}_n) + \|W(\hat{a}_n)\|_2$, which combined with Theorem 2.4.1 provides the upper bound

$$P_0(\theta_0 \in \hat{C}_n(L_n)) \le P_0\left(\inf_{a \le \overline{a}_n} \|B(a,\theta_0)\|_2 \le L_n \sup_{a \le \overline{a}_n} r_\alpha(a) + \sup_{a \le \overline{a}_n} \|W(a)\|_2\right) + o(1). \tag{2.7.1}$$

The proof of assertion (2.5.2) also shows that there exists constants $C_1 > 0$ such that

$$\sup_{a \le \overline{a}_n} r_\alpha^2(a) \le C_1 \frac{\overline{a}_n}{n} \log\left(\frac{n}{\overline{a}_n}\right). \tag{2.7.2}$$

Then in view of assertion (2.5.9) and Proposition 2.4.2, both the squared radius $r_\alpha(a)^2$ and the variance term $\|W(a)\|_2^2$ are bounded by$C_{\beta,b,M} n^{-2\beta/(2\beta+1)} (\log n)^{-1/(2\beta+1)}$, for some $C_{\beta,b,M} > 0$.

Since for $\theta_0 \in \Theta_s^\beta(m,M)$ we have $\|B(a,\theta_0)\|_2^2 = \sum\limits_{i=1}^{\infty} \frac{a^2 e^{2i/a} \theta_{0,i}^2}{(ae^{i/a}+n)^2}$ the bias is bounded from below by

$$\|B(a,\theta_0)\|_2^2 \ge m \sum_{i=I_a}^{\infty} i^{-1-2\beta} > \frac{m}{2\beta} I_a^{-2\beta} \ge \frac{m}{2\beta} a^{-2\beta} \log^{-2\beta}\left(\frac{n}{a}\right).$$

As the function $a \mapsto a^{-2\beta} \log^{-2\beta}\big(n/a\big)$ is monotone decreasing for $a \le A_n$, we see that $\inf\limits_{a \le \overline{a}_n} \|B(a,\theta_0)\|_2^2 \ge (m/(2\beta)) \overline{a}_n^{-2\beta} \log^{-2\beta}(n/\overline{a}_n)$. Hence in view of Proposition 2.4.2 the bias is bounded from below by $c_{m,\beta,b,B,M} n^{-2\beta/(2\beta+1)} (\log n)^{2\beta/(2\beta+1)}$, for some $c_{m,\beta,b,B,M} > 0$. Thus, the above inequalities imply that for arbitrary $\theta_0 \in \Theta_s^\beta(m,M)$ the right hand side of (2.7.1) is further bounded by

$$\sup_{\theta_0 \in \ell_2(M)} P_0\left(n^{-\beta/(2\beta+1)} (\log n)^{\beta/(2\beta+1)} \le L_n C n^{-\beta/(2\beta+1)} (\log n)^{-(1/2)/(2\beta+1)}\right) + o(1),$$

which goes to 0 for arbitrary $L_n = o(\sqrt{\log n})$ and C depending on m, β, b, B and M, concluding the proof of the theorem.

§2.8 Proof of Theorem 2.4.1

First note that the derivative of the marginal likelihood function $\ell_n(a)$ is

$$\mathbb{M}_n(a) = \frac{1}{2}\left(\sum_{i=1}^{\infty} \frac{n^2 Y_i^2 e^{i/a}(i-a)}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\right), \tag{2.8.1}$$

with expected value

$$E_0[\mathbb{M}_n(a)] = \frac{1}{2}\left(\sum_{i=1}^{\infty} \frac{n^2(i-a)e^{i/a}\theta_{0,i}^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\right). \tag{2.8.2}$$

In the following subsections we show with the help of the score function $\mathbb{M}_n(a)$ that the marginal likelihood function $\ell_n(a)$ with probability tending to one has its global maximum outside of the set $[1, \underline{a}_n) \cup (\overline{a}_n, A_n]$.

§2.8.1 $\mathbb{M}_n(a)$ on $[1, \underline{a}_n)$

In this subsection we derive that the process $\mathbb{M}_n(a)$ is bounded from below by $-C_B \log^2(n/a)$ on $[1, \underline{a}_n]$, for some $C_B > 0$, and is bigger than $e^{-5/2} B \log^3(n/\underline{a}_n)$, on the interval

$$\mathcal{I}_n \equiv \left[\frac{\log\left(\frac{n}{\underline{a}_n}\right)}{1+\log\left(\frac{n}{\underline{a}_n}\right)}\underline{a}_n, \underline{a}_n\right] \tag{2.8.3}$$

with probability going to one, where B is the parameter in the definition of $\underline{a}_n$. Hence with probability tending to one for every $a \in [1, \underline{a}_n]/\mathcal{I}_n$

$$\begin{aligned}
\ell_n(\underline{a}_n) - \ell_n(a) &\geq \int_a^{\frac{\log(n/\underline{a}_n)}{1+\log(n/\underline{a}_n)}\underline{a}_n} \mathbb{M}_n(\tilde{a})d\tilde{a} + \int_{\mathcal{I}_n} \mathbb{M}_n(\tilde{a})d\tilde{a} \\
&\geq -(\underline{a}_n - a)C\log^2\left(\frac{n}{\underline{a}_n}\right) + \frac{\tilde{c}_0 B \underline{a}_n \log^3\left(\frac{n}{\underline{a}_n}\right)}{\log\left(\frac{n}{\underline{a}_n}\right)} \\
&\geq (B\tilde{c}_0/2)\underline{a}_n \log^2\left(\frac{n}{\underline{a}_n}\right),
\end{aligned}$$

for $B > 2\tilde{c}_0^{-1}C$. Therefore the global maximum of $\ell_n(a)$ lies outside of the interval $[1, \underline{a}_n)$ with probability tending to one. It remained to show the stated lower bounds for $\mathbb{M}_n(a)$.

By leaving the non-negative stochastic part out we get the lower bound

$$\mathbb{M}_n(a) \geq \frac{1}{2}\left(\sum_{i=1}^{a} \frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\right). \tag{2.8.4}$$

In view of Lemma 2.8.1 the deterministic part in (2.8.4) is bounded from below by a negative constant times $\log^2(n/a)$. The stochastic part is bounded from below by

$-C\sum_{i=1}^{a} Y_i^2$ and since $E_0\sum_{i=1}^{a} Y_i^2 = \sum_{i=1}^{a}\theta_{0,i}^2 + an^{-1}$ and $V_0\big(\sum_{i=1}^{a} Y_i^2\big) = 2n^{-1}\sum_{i=1}^{a}\theta_{0,i}^2 + an^{-2} \to 0$ for all $a \le A_n$ it follows from Chebyshev's inequality that the sum $\sum_{i=1}^{a} Y_i^2$ is bounded with probability going to 1, for all $\theta_0 \in \ell_2(M)$.

Next we deal with the lower bound on the interval $a \in \mathcal{I}_n$. First note that $Y_i^2 \ge \theta_{0,i}^2 + 2\theta_{0,i}Z_i/\sqrt{n}$ implying

$$\mathbb{M}_n(a) \ge \frac{1}{2}\left(\sum_{i=1}^{a}\frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} + \log^2\left(\frac{n}{a}\right)g_n(a,\theta_0) + \mathbb{H}_n(a) - \sum_{i=1}^{\infty}\frac{n(i-a)}{a^2(ae^{i/a}+n)}\right),$$

with the centered Gaussian process

$$\mathbb{H}_n(a) = \sum_{i=2a}^{\infty}\frac{n^{3/2}(i-a)e^{i/a}\theta_{0,i}Z_i}{a(ae^{i/a}+n)^2}. \tag{2.8.5}$$

Note that

$$\begin{aligned} V_0\left(\frac{\mathbb{H}_n(a)}{\log^2\left(\frac{n}{a}\right)}\right) &= \frac{1}{\log^4\left(\frac{n}{a}\right)}\sum_{i=2a}^{\infty}\frac{n^3(i-a)^2e^{2i/a}\theta_{0,i}^2}{a^2(ae^{i/a}+n)^4}V_0(Z_i) \\ &\le \frac{ng_n(a,\theta_0)}{a\log^2\left(\frac{n}{a}\right)}\max_{i\ge 2a}\frac{(i-a)e^{i/a}}{(ae^{i/a}+n)^2} \ge \frac{g_n(a,\theta_0)}{a\log\left(\frac{n}{a}\right)}, \end{aligned}$$

hence the diameter of the interval $\mathcal{I}_n$ with respect to the metric

$$d_n^2(a_1,a_2) = V_0\left(\frac{\mathbb{H}_n(a_1)}{\log^2\left(\frac{n}{a_1}\right)} - \frac{\mathbb{H}_n(a_2)}{\log^2\left(\frac{n}{a_2}\right)}\right)$$

is bounded by a multiple of $\sup_{a\in\mathcal{I}_n} g_n(a,\theta_0)^{1/2}(a\log(n/a))^{-1/2}$.

Next we give an upper bound for the covering number of the interval $\mathcal{I}_n$. Let us take ε-balls centered at $a \in \mathcal{I}_n$, with $2a \in \mathbb{N}$. To cover the remaining part of the interval $\mathcal{I}_n$ it is sufficient to cover all intervals of the form $(a, a+1/2)$, $2a \in \mathbb{N}\cap 2\mathcal{I}_n$. Note that on these intervals for every $a_1, a_2 \in (a, a+1/2)$ we have $\lfloor 2a_1\rfloor - \lfloor 2a_2\rfloor = 0$. Hence in view of Lemma 2.8.2 we have $d_n(a_1,a_2) \lesssim |a_1-a_2|\sup_{a\in\mathcal{I}_n}\sqrt{\log(n/a)g_n(a,\theta_0)/a^3}$. Thus the covering number of the interval $(a, a+1/2)$ relative to d_n is bounded from above by a multiple of $\varepsilon^{-1}\sup_{a\in\mathcal{I}_n}\sqrt{\log(n/a)g_n(a,\theta_0)/a^3}$, which implies that the covering number of the whole interval $\mathcal{I}_n$ is bounded from above by constant times $\varepsilon^{-1}\sup_{a\in\mathcal{I}_n}\sqrt{\log^{-1}(n/a)g_n(a,\theta_0)/a + \underline{a}_n/\log(n/\underline{a}_n)}$.

We show below that for any $c_0 > 2$

$$e^{-2c_0}B\log n + o(1) \le g_n(a,\theta_0) \le e^{c_0}B\log n + o(1), \qquad \text{for } a \in \mathcal{I}_n, \tag{2.8.6}$$

hold. Therefore the covering number of $\mathcal{I}_n$ is bounded from above by a multiple of $\underline{a}_n + \varepsilon^{-1}\sqrt{\log^{-1}(n/\underline{a}_n)\log(n)/\underline{a}_n}$.

By Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) (applied with $\psi(x) = e^{x^2} - 1$) it follows that

$$E_0\left[\sup_{a\in\mathcal{I}_n}\left|\frac{\mathbb{H}_n(a)}{\log^2\left(\frac{n}{a}\right)} - \frac{\mathbb{H}_n(\underline{a}_n)}{\log^2\left(\frac{n}{\underline{a}_n}\right)}\right|\right]$$
$$\lesssim \int_0^{C\log^{1/2}(n)I_{\underline{a}_n}^{-1/2}} \sqrt{\log\left(\underline{a}_n + \varepsilon^{-1}\sqrt{\frac{\log(n)}{\underline{a}_n} log\left(\frac{n}{\underline{a}_n}\right)}\right)}d\varepsilon$$
$$\lesssim \int_0^{C\log^{1/2}(n)I_{\underline{a}_n}^{-1/2}} \sqrt{\log \underline{a}_n}d\varepsilon + \int_0^1 \log(1/\varepsilon)d\varepsilon = O(1).$$

Therefore the process $\mathbb{M}_n(a)$ can be bounded from below on $a \in \mathcal{I}_n$ by

$$\mathbb{M}_n(a) \geq 2^{-1}\inf_{a\in\mathcal{I}_n}\left\{\log^2\left(\frac{n}{a}\right)\left(Be^{-5}\log n - C\right)\right.$$
$$\left. + \sum_{i=1}^{a}\frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty}\frac{n(i-a)}{a^2(ae^{i/a}+n)}\right\}$$

with probability going to one. In view of (2.8.6) and since the third and fourth terms on the right hand side of the preceding display are bounded from below by a fixed negative constant, we get that with probability tending to one $\mathbb{M}_n(a) \geq e^{-5/2}B\log^3(n/\underline{a}_n)$.

It remained to verify assertion (2.8.6). First note that

$$\frac{n^2}{\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=c_0 I_{\underline{a}_n}}^{\infty}\frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2}\theta_{0,i}^2 \leq \frac{n^2}{\underline{a}_n^3\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=c_0 I_{\underline{a}_n}}^{\infty} ie^{-i/\underline{a}_n}\theta_{0,i}^2$$
$$\lesssim \frac{A_n^{c_0-2}}{n^{c_0-2}\log\left(\frac{n}{\underline{a}_n}\right)}\|\theta_0\|_2^2 = o(1).$$

Furthermore, in view of the inequality $c_0 I_{\underline{a}_n}(a^{-1} - \underline{a}_n^{-1}) \leq c_0$, for $a \in \mathcal{I}_n$, we have that

$$g_n(a,\theta_0) \geq \frac{n^2}{\log^2\left(\frac{n}{a}\right)}\sum_{i=2\underline{a}_n}^{c_0 I_{\underline{a}_n}}\frac{(i-a)e^{i/a}}{a(ae^{i/a}+n)^2}\theta_{0,i}^2$$
$$\geq \frac{n^2}{e^{2c_0}\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=2\underline{a}_n}^{c_0 I_{\underline{a}_n}}\frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2}\theta_{0,i}^2.$$

By combining the preceding two displays we get that

$$g_n(a,\theta_0) \geq e^{-2c_0}g_n(\underline{a}_n,\theta_0) - \frac{e^{-2c_0}n^2}{\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=c_0 I_{\underline{a}_n}}^{\infty}\frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2}\theta_{0,i}^2$$
$$\geq e^{-2c_0}B\log n + o(1),$$

finishing the proof of the first inequality in (2.8.6). The proof of the second inequality goes accordingly.

Lemma 2.8.1. *There exists a constant $K > 0$ such that for any $a \in [1, n)$*

$$\sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)} \le K \log^2(n/a)$$

Proof. Note that

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)} &\le \sum_{i=1}^{\infty} \frac{ni}{a^2(ae^{i/a}+n)} \le \sum_{i=1}^{I_a} \frac{i}{a^2} + \sum_{i=I_a}^{\infty} \frac{nie^{-i/a}}{a^3} \\ &\lesssim \log^2(n/a) + \frac{\log(n/a)}{a} \lesssim \log^2(n/a).\end{aligned}$$

□

Lemma 2.8.2. *There exists a constant $K > 0$ such that for any positive a_1 and a_2 such that $a_1 < a_2$, $\lfloor 2a_2 \rfloor - \lfloor 2a_1 \rfloor = 0$*

$$V_0\Big(\frac{\mathbb{H}_n(a_1)}{\log(n/a_1)^2} - \frac{\mathbb{H}_n(a_2)}{\log(n/a_2)^2}\Big) \le K(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \frac{\log(n/a) g_n(a,\theta_0)}{a^3}.$$

Proof. Recall that the left hand side of the display in the lemma was denoted by $d_n^2(a_1, a_2)$ and note that

$$d_n^2(a_1,a_2) = \sum_{i=2a_2}^{\infty} \big(\phi_i(a_1) - \phi_i(a_2)\big)^2 n^3 \theta_{0,i}^2 \tag{2.8.7}$$

with $\phi_i(a) := \frac{(i-a)e^{i/a}}{\log(n/a)^2 a(ae^{i/a}+n)^2}$. Then by elementary, but cumbersome computations we get that $|\phi_i'(a)| \lesssim ia^{-2}\phi_i(a)$. Thus, in view of Lemma 2.11.3, the right hand side of (2.8.7) is bounded by

$$\begin{aligned}&n^3(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \sum_{i=2a}^{\infty} \frac{i^2}{a^4}\phi_i(a)^2\theta_{0,i}^2 \\ &\qquad\lesssim (a_1-a_2)^2 \sup_{a\in[a_1,a_2]} g_n(a,\theta_0) \sup_{i\in\mathbb{N}} \frac{ni^3 e^{i/a}}{a^5 \log^2(n/a)(ae^{i/a}+n)^2}.\end{aligned}$$

Then the statement of the lemma follows by applying Lemma 2.11.1 (with $m = 3$). □

§2.8.2 $\mathbb{M}_n(a)$ on $[\overline{a}_n, A_n]$

We prove that for sufficiently large choice of $K_0 > 0$ in the definition of $\overline{a}_n$

$$\limsup_n \sup_{\theta_0\in\ell_2(M)} \sup_{a\in[\overline{a}_n,A_n]} E_0\Big[\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big] < -2^{-5}, \tag{2.8.8}$$

$$\limsup_n \sup_{\theta_0\in\ell_2(M)} E_0\left[\sup_{a\in[\overline{a}_n,A_n]} \frac{|\mathbb{M}_n(a) - E_0[\mathbb{M}_n(a)]|}{\log^2(n/a)}\right] \le 2^{-6}. \tag{2.8.9}$$

These imply that with probability tending to one $\mathbb{M}_n(a) < -2^{-6}\log^2(n/a)$, for every $a \in [\bar{a}_n, A_n]$, hence the marginal likelihood function $\ell_n(a)$ is monotone decreasing and does not attain its global (or local) maximum on the interval $[\overline{a}_n, A_n]$, i.e.

$$\inf_{\theta_0 \in \ell_2(M)} P_0(\hat{a}_n \leq \overline{a}_n) \to 1. \tag{2.8.10}$$

Proof of assertion (2.8.8): In view of $h_n(a, \theta_0) \leq b$ for all $a \in [\overline{a}_n, A_n]$ (assuming that $\overline{a}_n > K_0$), we get that

$$\begin{aligned} E_0\Big[\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big] &= \frac{1}{2}\Big(h_n(a,\theta_0) - \frac{1}{\log^2(n/a)}\sum_{i=1}^{\infty}\frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big) \\ &\leq \frac{1}{2}\Big(b - \frac{1}{\log^2(n/a)}\sum_{i=1}^{\infty}\frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big). \end{aligned}$$

In view of Lemma 2.11.2 (with $r = 0$ and $l = 2$), we have $\sum_{i=1}^{\infty}\frac{n^2}{a(ae^{i/a}+n)^2} \lesssim \log(n/a)$. Furthermore,

$$\sum_{i=1}^{\infty}\frac{in^2}{a^2(ae^{i/a}+n)^2} \geq \sum_{i=1}^{I_a}\frac{i}{4a^2} = \frac{I_a(I_a+1)}{8a^2} \geq 2^{-3}\log^2\big(\frac{n}{a}\big),$$

which implies that

$$E_0\big[\mathbb{M}_n(a)/\log^2(n/a)\big] \leq (b - 2^{-3} + o(1))/2,$$

concluding the proof of assertion (2.8.8), for small enough choice of b ($b < 2^{-4}$ is small enough).

Proof of assertion (2.8.9): In view of Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) (applied with $\psi(x) = x^2$) it is sufficient to show that there exist universal constants $K_1, K_2 > 0$ such that for any $a \in [\overline{a}_n, A_n]$

$$V_0\big(\mathbb{M}_n(a)/\log^2(n/a)\big) \leq K_1/\log(n/a), \tag{2.8.11}$$

$$\int_0^{diam_n}\sqrt{N(\varepsilon, [\overline{a}_n, A_n], d_n)}d\varepsilon \leq K_2/K_0^{1/4}, \tag{2.8.12}$$

where d_n is the semi-metric defined by $d_n^2(a_1,a_2) := V_0\Big(\frac{\mathbb{M}_n(a_1)}{\log^2(n/a_1)} - \frac{\mathbb{M}_n(a_2)}{\log^2(n/a_2)}\Big)$, $diam_n$ is the diameter of $[\overline{a}_n, A_n]$ relative to d_n and $N(\varepsilon, S, d_n)$ is the minimal number of d_n-balls of radius ε needed to cover the set S, since by sufficiently large choice of K_0 ($K_0 \geq (2^6K_2)^4$ is sufficiently large) assertion (2.8.9) holds.

Note that Lemma 2.8.3 immediately implies assertion (2.8.11) and

$$diam_n \lesssim \sup_{a\in[\bar{a}_n, A_n]}(a\log(n/a))^{-1/2} \lesssim \log^{-1/2} n.$$

Then let us introduce the cover

$$[\overline{a}_n, A_n] \subset \bigcup_{k=0}^{K_n-1}[2^k\overline{a}_n, 2^{k+1}\overline{a}_n]$$

with $K_n = \lceil \log(A_n/\overline{a}_n) \rceil$. In view of Lemma 2.8.4, when a_1 and a_2 are on the interval $[2^k\overline{a}_n, 2^{k+1}\overline{a}_n]$

$$d_n(a_1,a_2) \lesssim \big(2^k\overline{a}_n\big)^{-3/2} \log^{1/2}(n)|a_1-a_2|,$$

hence

$$N(\varepsilon,[\overline{a}_n,A_n],d_n) \lesssim \sum_{k=0}^{K_n-1} \frac{\log^{1/2}(n)}{\varepsilon\big(2^k\overline{a}_n\big)^{1/2}} \lesssim \frac{\log^{1/2}(n)}{\varepsilon\overline{a}_n^{1/2}}.$$

This results in

$$\int_0^{diam_n} \sqrt{N(\varepsilon,[\overline{a}_n,A_n],d_n)}d\varepsilon \le K_2/\overline{a}_n^{1/4} \le K_2/K_0^{1/4}.$$

Lemma 2.8.3. *For all $a \in [\overline{a}_n, A_n]$, we have $V_0\big(\mathbb{M}_n(a)/\log^2(n/a)\big) \lesssim (a\log(n/a))^{-1}$.*

Proof. We know that the Y_is are independent and $V_0(Y_i^2) = 2/n^2 + 4\theta_{0,i}^2/n$, so the variance is equal to

$$\begin{aligned} V_0\Big(\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big) &= \frac{1}{4}\sum_{i=1}^{\infty} \frac{n^4V_0(Y_i^2)e^{2i/a}(i-a)^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4} \\ &= \frac{1}{2}\sum_{i=1}^{\infty} \frac{n^2e^{2i/a}(i-a)^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty} \frac{n^3e^{2i/a}(i-a)^2\theta_{0,i}^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4}. \end{aligned} \tag{2.8.13}$$

In view of $(i-a)^2 \le a^2+i^2$, for any $a,i>0$, and by applying Lemma 2.11.1 (with $m=2$) and Lemma 2.11.2 (first with $r=2$ and $l=4$ and then with $r=1$ and $l=2$) the first sum in (2.8.13) is bounded from above by a multiple of

$$\begin{aligned} &\sum_{i=1}^{\infty} \frac{n^2e^{2i/a}}{\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty} \frac{ne^{i/a}}{a\log^2(n/a)(ae^{i/a}+n)^2} \\ &\qquad \lesssim \frac{1}{a\log^3(n/a)} + \frac{1}{a\log(n/a)} \lesssim \frac{1}{a\log(n/a)}. \end{aligned}$$

Similarly, following from Lemma 2.11.1 (with $m=1$ and $m=-1$) and $h_n(a,\theta_0) \le b$ for $a \ge \overline{a}_n$, the second sum in (2.8.13) is bounded by a multiple of

$$\begin{aligned} &\Big(\max_{i\in\mathbb{N}} \frac{ane^{i/a}}{i\log^2(n/a)(ae^{i/a}+n)^2} + \max_{i\in\mathbb{N}} \frac{ine^{i/a}}{a\log^2(n/a)(ae^{i/a}+n)^2}\Big)h_n(a,\theta_0) \\ &\qquad \lesssim \Big(\frac{1}{a\log^3(n/a)} + \frac{1}{a\log(n/a)}\Big) \lesssim \frac{1}{a\log(n/a)}, \end{aligned}$$

concluding the proof of the lemma. □

Lemma 2.8.4. *For all $1 \le a_1 < a_2 < A_n$, we have*

$$d_n^2(a_1,a_2) \le C_0(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \frac{\log(n/a)}{a^3}(1+h_n(a,\theta_0)),$$

for some universal constant $C_0 > 0$.

Proof. Note that

$$d_n^2(a_1,a_2) = n^4 \sum_{i=1}^{\infty} (\phi_i(a_1) - \phi_i(a_2))^2 V_0(Y_i^2),$$

with $\phi_i(a) = \frac{e^{i/a}(i-a)}{2a\log^2(n/a)(ae^{i/a}+n)^2}$. By elementary computations one can see that $|\phi_i(a)'|^2 \lesssim (i^2a^{-4} + a^{-2})\phi_i^2(a)$, hence in view of Lemma 2.11.3,

$$d_n^2(a_1,a_2) \lesssim (a_1-a_2)^2 n^4 \sup_{a\in[a_1,a_2]} \sum_{i=1}^{\infty} \frac{e^{2i/a}(i^4+a^4)}{a^6\log^4(n/a)(ae^{i/a}+n)^4} V_0(Y_i^2).$$

Since $V_0(Y_i^2) = 2/n^2 + 4\theta_{0,i}^2/n$ the preceding sum is bounded by

$$\sum_{i=1}^{\infty} \frac{2e^{2i/a}(i^4+a^4)}{a^6n^2\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty} \frac{4e^{2i/a}(i^4+a^4)}{a^6n\log^4(n/a)(ae^{i/a}+n)^4}\theta_{0,i}^2. \tag{2.8.14}$$

Then in view of Lemma 2.11.1 (applied with $m=4$ and $m=0$) and Lemma 2.11.2 (applied with $r=1$ and $l=2$) the first term of (2.8.14) is bounded from above by a multiple of

$$\sum_{i=1}^{\infty} \frac{e^{i/a}}{a^3n^3(ae^{i/a}+n)^2} \lesssim \frac{\log(n/a)}{a^3n^4}.$$

Similarly in view of Lemma 2.11.1 (with $m=3$ and $m=-1$) the second term of (2.8.14) is bounded by

$$\begin{aligned}\max_{i\in\mathbb{N}} &\frac{\big((i/a)^3 + (i/a)^{-1}\big)e^{i/a}}{a^2n^3\log^2(n/a)(ae^{i/a}+n)^2} h_n(a,\theta_0) \\ &\lesssim \Big(\frac{\log(n/a)}{a^3n^4} + \frac{1}{n^5a^2}\Big)h_n(a,\theta_0) \lesssim \frac{\log(n/a)}{a^3n^4} h_n(a,\theta_0),\end{aligned}$$

concluding the proof of the lemma. □

§2.9 Proof of Theorem 2.1.6

Similarly to the previous sections we use the notations introduced in Section 2.5. We show below that there exists a constant $c>0$ depending only on m, M and β_0 such that

$$\inf_{\beta\geq\beta_0} \inf_{\theta_0\in\Theta_s^\beta(m,M)} P_0(\hat{a}_n \geq c(n/\log n)^{1/(1+2\beta)}/\log n) \to 1, \tag{2.9.1}$$

which combined with Proposition 2.4.2 and Theorem 2.4.1 results in

$$\inf_{\beta\geq\beta_0} \inf_{\theta_0\in\Theta_s^\beta(m,M)} P_0(c(n/\log n)^{1/(1+2\beta)} \leq \tilde{a}_n \leq C(n/\log n)^{1/(1+2\beta)}) \to 1,$$

for some positive constants c, C depending on b, B, m, M and β. Let us introduce then the notation

$$\tilde{\mathcal{I}}_n = [c(n/\log n)^{\frac{1}{1+2\beta}}, C(n/\log n)^{\frac{1}{1+2\beta}}].$$

As before, note that $\theta_0 \in \hat{C}_n(L)$ is equivalent to $\|\theta_0 - \hat{\theta}\|_2 \le Lr_\alpha(\tilde{a}_n)$, hence by proving that

$$\inf_{a\in\tilde{\mathcal{I}}_n} r_\alpha^2(a) \ge C_1(n/\log n)^{-2\beta/(1+2\beta)},$$

$$\inf_{\beta\ge\beta_0}\inf_{\theta_0\in\Theta_s^\beta(m,M)} P_0\big(\inf_{a\in\tilde{\mathcal{I}}_n}\|W(a)\|_2^2 \le C_2(n/\log n)^{-2\beta/(1+2\beta)}\big) \to 1,$$

$$\sup_{\beta\ge\beta_0}\sup_{\theta_0\in\Theta_s^\beta(m,M)}\sup_{a\in\tilde{\mathcal{I}}_n}\|B(a,\theta_0)\|_2^2 \le C_3(n/\log n)^{-2\beta/(1+2\beta)},$$

hold for some constants $C_1, C_2, C_3 > 0$, the statement of the theorem follows immediately. The proof of the first two inequalities follow from (2.5.2) and (2.5.9) (with $\underline{a}_n$ and $\overline{a}_n$ replaced by a multiple of $(n/\log n)^{1/(1+2\beta)}$), respectively. To prove the last inequality we note that for $\theta_0 \in \Theta_s^\beta(m, M)$, $a \in \tilde{\mathcal{I}}_n$, and $\beta \ge \beta_0$ we have that

$$\begin{aligned}\|B(a,\theta_0)\|_2^2 &\lesssim \sum_{i=1}^{I_a/2} a^2 e^{2i/a} n^{-2} i^{-1-2\beta} + \sum_{i=I_a/2}^{\infty} i^{-1-2\beta} \lesssim a/n + I_a^{-2\beta}\\ &= o\big((n/\log n)^{-2\beta/(1+2\beta)}\big).\end{aligned}$$

It remained to prove assertion (2.9.1). Let us introduce the slightly modified version of $\underline{a}_n$ as

$$\underline{a}'_n := \sup\{a \in [1, A_n] : g_n(a, \theta_0) \ge B\},$$

for some sufficiently large constant $B > 0$ to be specified later. Then we show below that

$$P_0(\hat{a}_n \ge \underline{a}'_n) \to 1, \quad \text{and} \quad \underline{a}'_n \ge c(n/\log n)^{1/(1+2\beta)}/\log n, \tag{2.9.2}$$

for some sufficiently small constant $c > 0$.

For the second statement note that

$$g_n(a,\theta_0) \ge \frac{m}{\log^2(n/a)} n^2 \sum_{i=I_a}^{\infty} e^{-i/a} i^{-2\beta} \gtrsim mna^{-1-2\beta}\log^{-2-2\beta}(n/a), \tag{2.9.3}$$

hence for any fixed $B > 0$ there exists a small enough $c > 0$ such that the right hand side of the preceding display with $a = c(n/\log n)^{1/(1+2\beta)}/\log n$ is bigger than B. It remained to deal with the first part of (2.9.2). We show below that with probability tending to one $\inf_{a\in[\underline{a}'_n/2,\underline{a}'_n]}\mathbb{M}_n(a) \ge cB\log^2(n/a)$, for some small enough constant $c > 0$, not depending on B. Then with probability tending to one for any $a \in [1, \underline{a}'_n/2]$ we have

$$\begin{aligned}\ell_n(\underline{a}_n) - \ell_n(a) &\ge \int_a^{\underline{a}'_n/2}\mathbb{M}_n(\tilde{a})d\tilde{a} + \int_{[\underline{a}'_n/2,\underline{a}'_n]}\mathbb{M}_n(\tilde{a})d\tilde{a}\\ &\ge -(\underline{a}'_n/2 - a)C\log^2(n/\underline{a}_n) + cB(\underline{a}'_n/2)\log^2(n/\underline{a}'_n)\\ &\ge (c/4)B\underline{a}'_n\log^2(n/\underline{a}'_n),\end{aligned}$$

for large enough choice of $B > 0$, hence the global maximum of $\ell_n(a)$ lies outside of the interval $[1, \underline{a}'_n]$.

It remained to verify the lower bound for $M_n(a)$. First note that for $a \le A_n = o(n)$

$$g_n(a,\theta_0) \le \frac{M}{\log^2(n/a)}\Big(\frac{1}{a}\sum_{i=2a}^{I_a} e^{i/a} i^{-2\beta} + \frac{n^2}{a^3}\sum_{i=I_a}^{\infty} e^{-i/a} i^{-2\beta}\Big)$$
$$\le c_{M,\beta} n a^{-1-2\beta} (\log n)^{-2-2\beta},$$

hence $\underline{a}'_n \le c'_{M,\beta} B^{-1/(1+2\beta)} (n/\log n)^{1/(1+2\beta)}/\log n$. Therefore in view of (2.9.3) for every $a \ge \underline{a}'_n/2$ we have $g_n(a,\theta_0) \ge c_{M,\beta,m} B$, for some positive constant $c_{M,\beta,m} > 0$ not depending on B. Similarly we can show that $g_n(a,\theta_0) \le c'_{M,\beta,m} B$, for every $a \ge \underline{a}'_n/2$, for some $c'_{M,\beta,m} > 0$ not depending on B. Then following the same line of reasoning as in Section 2.8.1, with the only main difference that instead of the interval given in (2.8.3) we are working with the interval $[\underline{a}'_n/2, \underline{a}'_n]$ we get that with probability going to one

$$\inf_{a\in[\underline{a}'_n/2,\underline{a}'_n]} \mathbb{M}_n(a) \ge 2^{-1} \inf_{a\in[\underline{a}'_n/2,\underline{a}'_n]} \Big\{ \log^2(n/a)\Big(c_{M,\beta,m} B - \sqrt{c'_{M,\beta,m} B}\Big)$$
$$+ \sum_{i=1}^{a} \frac{n^2 (i-a) e^{i/a} Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\Big\}$$
$$\gtrsim B \log^2(n/\underline{a}'_n),$$

for large enough choice of $M > 0$, finishing the proof of the theorem.

§2.10 Proofs for the Hierarchical Bayes procedure

In this section we prove the results on the hierarchical Bayes procedure (i.e. Theorems 2.4.5 and 2.1.5 and Corollary 2.1.2) based on the results derived for the empirical Bayes procedure. First we state that under the conditions of Theorem 2.4.5 the hyper-posterior distribution on the hyper-parameter a concentrates most of its mass on the interval $\mathcal{I}_n = [\underline{a}_n \log(n)/(1+\log n), C\overline{a}_n]$, for some large enough constant $C > 0$.

Lemma 2.10.1. *If $a \sim \pi(.)$ such that π verifies Assumption 2.1.1 then for sufficiently large $C > 0$ we have for every $\beta_0 > 0$ that*

$$\inf_{\beta>\beta_0} \inf_{\theta_0\in\Theta^\beta(M)} E_0 \Pi\Big(\underline{a}_n \log(n)/(1+\log n) \le a \le C\overline{a}_n | Y\Big) = 1 + o(1/n).$$

§2.10.1 Proof of Theorem 2.4.5

Take $\varepsilon_n = (n/\log^2 n)^{-\beta/(1+2\beta)}$. Then following from Lemma 2.10.1, we have

$$\sup_{\theta_0\in\Theta^\beta(M)} E_0 \Pi\big(\theta : \|\theta-\theta_0\|_2 > M_n\varepsilon_n | Y\big) \le \sup_{\theta_0\in\Theta^\beta(M)} \Big(E_0\Pi(a \notin \mathcal{I}_n | Y)$$
$$+ E_0 \sup_{a\in\mathcal{I}_n} \Pi_a\big(\theta : \|\theta-\theta_0\|_2 > M_n\varepsilon_n | Y\big)\Big) = o(1),$$

where the last equation follows by similar arguments as given in (2.4.9) and the displays below it (the only difference is that the supremum is taken over the interval $\mathcal{I}_n$ instead of $[\underline{a}_n, \overline{a}_n]$, but it only changes the constant factors which do not play an essential role. This concludes the proof of the theorem.

§2.10.2 Proof of Theorem 2.1.3 - Hierarchical Bayes part

In the proof we use again the notations introduced in Section 2.5.

Let $a' := n^{1/(1+2\beta)}(\log n)^{-1-1/(1+2\beta)} \asymp \overline{a}_n \asymp \underline{a}_n$ with probability going to one thanks to Proposition 2.4.2. One can see that in the hierarchical case,

$$P_0(\theta_0 \in \hat{C}_n(L_n)) \leq P_0\Big(\|B(a', \theta_0)\|_2 \leq L_n r_\alpha + \|W(a')\|_2 + \|\hat{\theta} - \hat{\theta}_{a'}\|_2\Big) + o(1), \tag{2.10.1}$$

which is a slightly modified version of (2.7.1) thanks to the triangle inequality. In order to prove that the right hand-side tends to zero, it is sufficient to show that there exist constants $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 > 0$ such that

$$r_\alpha^2 \leq \tilde{C}_1 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}, \tag{2.10.2}$$

$$P_0(\|W(a')\|_2^2 \leq \tilde{C}_2 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}) \to 1, \tag{2.10.3}$$

$$\|B(a', \theta_0)\|_2^2 \geq \tilde{C}_3 n^{-2\beta/(1+2\beta)} \log(n)^{2\beta/(1+2\beta)}, \tag{2.10.4}$$

$$P_0(\|\hat{\theta} - \hat{\theta}_{a'}\|_2 \leq \tilde{C}_4 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}) \to 1. \tag{2.10.5}$$

The bounds on the variance and the bias are obtained in a similar manner as in Section 2.7 and 2.10.3. Next we deal with assertion (2.10.5).

By Jensen's inequality, Fubini's theorem and triangle inequality one can obtain that

$$\begin{aligned}
\|\hat{\theta} - \hat{\theta}_{a'}\|_2 &= \| \int (\hat{\theta}_a - \hat{\theta}_{a'}) \Pi(da|Y) \|_2^2 \qquad (2.10.6)\\
&\leq \sum_{i=1}^{\infty} \int (\hat{\theta}_{a,i} - \hat{\theta}_{a',i})^2 \Pi(da|Y) \\
&\leq \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - \hat{\theta}_{a',i})^2 \Pi(a_1 \in \mathcal{I}_n|Y) + \sup_{a_1 \notin \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - \hat{\theta}_{a',i})^2 \Pi(a_1 \notin \mathcal{I}_n|Y).
\end{aligned}$$

Starting with the first term, we use the trivial bound 1 for $\Pi(.|Y)$. We have with P_0-probability tending to 1 that

$$\begin{aligned}
\sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - \hat{\theta}_{a',i})^2 \leq & \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (E_0\hat{\theta}_{a_1,i} - E_0\hat{\theta}_{a',i})^2 \qquad (2.10.7)\\
&+ \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - E_0\hat{\theta}_{a_1,i})^2 + \sum_{i=1}^{\infty} (\hat{\theta}_{a',i} - E_0\hat{\theta}_{a',i})^2
\end{aligned}$$

The two last term on the right hand-side are bounded by a constant multiplier of $n^{-2\beta/(1+2\beta)}\log(n)^{-1/(1+2\beta)}$ from (2.5.3). The first term can be written as $\sup_{a_1\in\mathcal{I}_n}\sum_{i=1}^{\infty}(g_i(a_1)-g_i(a'))^2$ for $g_i(a)=n\theta_{0,i}^2/(ae^{i/a}+n)$. The derivative of $g_i(a)$ is $-n\theta_{0,i}^2(a-i)e^{i/a}/(a(ae^{i/a}+n)^2)$. Without loss of generality, when $a_1<a'$ writing the difference as the integral of $g_i'(a)$, applying Cauchy-Schwartz inequality to its squares and then interchanging the sum and the integral, we get that

$$\begin{aligned}\sum_{i=1}^{\infty}(E_0\hat{\theta}_{a_1,i}-E_0\hat{\theta}_{a',i})^2&=\sum_{i=1}^{\infty}\Big(\int_{a_1}^{a'}g_i'(a)da\Big)^2\leq\sum_{i=1}^{\infty}(a'-a_1)\int_{a_1}^{a'}g_i'(a)^2da\\&=(a'-a_1)\int_{a_1}^{a'}\sum_{i=1}^{\infty}g_i'(a)^2da\leq(a'-a_1)^2\sup_{a\in\mathcal{I}_n}\sum_{i=1}^{\infty}g_i'(a)^2da\\&\leq(a'-a_1)^2\sup_{a\in\mathcal{I}_n}\sum_{i=1}^{\infty}\frac{n^2\theta_{0,i}^4(i-a)^2e^{2i/a}}{a^2(ae^{i/a}+n)^4}\end{aligned}$$

For fixed a, the sum in the preceding display is bounded from above by constant times

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{n^2i^{-2-4\beta}(i-a)^2e^{2i/a}}{a^2(ae^{i/a}+n)^4}&\leq\frac{1}{a^2n^2}\sum_{i=1}^{I_a}(i^2+a^2)i^{-2-4\beta}e^{2i/a}\\&\quad+\frac{n^2}{a^6}\sum_{i>I_a}(i^2+a^2)i^{-2-4\beta}e^{-2i/a}\\&\lesssim a^{-3-4\beta}\log\big(\frac{n}{a}\big)^{1-4\beta}.\end{aligned}$$

Therefore, one can see that

$$\begin{aligned}\sup_{a_1\leq\bar{a}_n}\sum_{i=1}^{\infty}(E_0\hat{\theta}_{a_1,i}-E_0\hat{\theta}_{a',i})^2&\lesssim\sup_{a_1\in\mathcal{I}_n}(a'-a_1)^2\underline{a}_n^{-3-4\beta}\log(n)^{1-4\beta}.\\&\lesssim n^{-1-2\beta/(1+2\beta)}\log(n)^{7+1/(1+2\beta)}=o(1/n),\end{aligned}$$

with probability tending to one using Proposition 2.4.2

It is left to deal with the second term on the right hand-side of (2.10.6). Following from (2.10.7), we get with P_0-probability tending to 1 that

$$\begin{aligned}\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(\hat{\theta}_{a_1,i}-\hat{\theta}_{a',i})^2\leq 2\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(E_0\hat{\theta}_{a_1,i})^2+\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(\hat{\theta}_{a_1,i}-E_0\hat{\theta}_{a_1,i})^2\\+2\sum_{i=1}^{\infty}(E_0\hat{\theta}_{a',i})^2+\sum_{i=1}^{\infty}(\hat{\theta}_{a',i}-E_0\hat{\theta}_{a',i})^2,\end{aligned}\tag{2.10.8}$$

where all terms on the right hand side are $O(1)$. Since

$$E_0\Pi(a\notin\mathcal{I}_n|Y)=o(1/n),$$

applying Markov's inequality leads to the second term on the right hand-side of (2.10.6) being of lower order than n^{-1}.

It remained to deal with assertion (2.10.2). We show below that

$$r_\alpha \le \tilde{r} := \sup_{a\in\mathcal{I}_n}\Big(\|\hat{\theta}-\hat{\theta}_a\|_2 + r_{\alpha/2}(a)\Big). \tag{2.10.9}$$

Then in view of the inequality

$$\sup_{a\in\mathcal{I}_n}\|\hat{\theta}-\hat{\theta}_a\|_2 \le \|\hat{\theta}-\hat{\theta}_{a'}\|_2 + \sup_{a\in\mathcal{I}_n}\|\hat{\theta}_a-\hat{\theta}_{a'}\|_2$$

and assertions (2.10.5), (2.10.7), and (2.7.2) we get that with probability tending to one $r_\alpha^2 \le \tilde{C}_1 n^{-2\beta/(1+2\beta)}\log(n)^{-1/(1+2\beta)}$ and since r_α is deterministic the inequality holds almost surely.

Finally we verify assertion (2.10.9). Note that

$$\begin{aligned}\Pi(\theta:\ \|\hat{\theta}-\theta\|_2\le\tilde{r}|Y) &\ge \int_{\mathcal{I}_n}\Pi_a(\theta:\ \|\theta-\hat{\theta}\|_2\le\tilde{r}|Y)\pi(a|Y)da\\ &\ge \int_{\mathcal{I}_n}\Pi_a(\theta:\ \|\theta-\hat{\theta}_a\|_2\le r_{\alpha/2}(a)|Y)\pi(a|Y)da\\ &\ge \int_{\mathcal{I}_n}(1-\alpha/2)\pi(a|Y)da < 1-\alpha,\end{aligned}$$

for large enough n, concluding the proof of our theorem for the Hierarchical Bayes method.

§2.10.3 Proof of Theorem 2.1.5 - Hierarchical Bayes part

Let us introduce the notations $W=\hat{\theta}-E_0\hat{\theta}$ and $B(\theta_0)=E_0\hat{\theta}-\theta_0$, for the centered hierarchical posterior mean and the bias of the posterior mean, respectively. Then $P_0(\theta_0\in\hat{C}(L\log n))$ if and only if

$$\|W\|_2 \le L\log(n)r_\alpha - \|B(\theta_0)\|_2 \tag{2.10.10}$$

holds. Using assertions (2.5.2), (2.5.3), and (2.5.4) we show below that, there exist constants $\tilde{C}_1,\tilde{C}_2,\tilde{C}_3>0$, such that

$$r_\alpha^2 \ge \tilde{C}_1(\underline{a}_n/n)\log(n/\underline{a}_n), \tag{2.10.11}$$

$$\inf_{\theta_0\in\Theta_{pt}(L_0,N_0,\rho)} P_0\big(\|W\|_2^2 \le \tilde{C}_2(\underline{a}_n/n)\log(n/\underline{a}_n)\log^2 n\big)\to 1, \tag{2.10.12}$$

$$\|B(\theta_0)\|_2^2 \le \tilde{C}_3(\underline{a}_n/n)\log^2\big(n/\underline{a}_n\big)\log n, \tag{2.10.13}$$

resulting in (2.10.10) for sufficiently large choice of $L>0$.

Proof of (2.10.11): Let us take any $\alpha'>\alpha$ and note that in view of (2.5.2) we have

$$\inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)^2 \ge C_1(\underline{a}_n/n)\log(n/\underline{a}_n).$$

Next, in view of Lemma 2.10.1 and Anderson's lemma, we get for arbitrary $r \leq \inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)$ that

$$\begin{aligned}\Pi(\theta:\ \|\theta-\hat\theta\|_2\leq r|Y) &= \int_{\mathcal{I}_n}\Pi_a(\theta:\ \|\theta-\hat\theta\|_2\leq r|Y)\pi(a|Y)da+o(1)\\ &\leq \int_{\mathcal{I}_n}\Pi_a(\theta:\ \|\theta-\hat\theta_a\|_2\leq r_{\alpha'}(a)|Y)\pi(a|Y)da+o(1)\\ &\leq 1-\alpha'+o(1),\end{aligned}$$

hence $r_\alpha^2\geq \inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)^2\geq C_1(\underline{a}_n/n)\log(n/\underline{a}_n)$.

Proof of (2.10.12): Note that by triangle inequality, Fubini's theorem, assertion (2.5.3), and Lemma 2.10.1 we get that under the polished tail condition with P_0-probability tending to one

$$\begin{aligned}\|W\|_2 &= \Big\|\int(\hat\theta_a-E_0\hat\theta_a)\pi(a|Y)da\Big\|_2\\ &\leq \sup_{a\in\mathcal{I}_n}\|W(a)\|_2\pi(\mathcal{I}_n|Y)+\sup_{1\leq a\leq A_n}\|W(a)\|_2\pi(\mathcal{I}_n^c|Y)\\ &\leq (C_2\underline{a}_n/n)^{1/2}\log(n/\underline{a}_n)^{1/2}\log n+o(1/n)\end{aligned}$$

where $\pi(\mathcal{I}_n|Y)$ denotes (by slightly abusing our notation) the posterior probability that the hyper-parameter a lies in the interval $\mathcal{I}_n$ and in the last inequality we used in view of the proof of assertion (2.5.3) that $\sup_{1\leq a\leq A_n}\|W(a)\|_2=O(1)$.

Proof of (2.10.13): Similarly to the proof of (2.10.12) we get that

$$\begin{aligned}\|B(\theta_0)\|_2^2 &\lesssim \sup_{a\in\mathcal{I}_n}\|B(a,\theta_0)\|_2^2+o(\sup_{a\in[1,A_n]}\|B(a,\theta_0)\|_2^2/n)\\ &\leq C_3(\underline{a}_n/n)\log^2\big(n/\underline{a}_n\big)\log n+o(1/n),\end{aligned}$$

where the last inequality follows from $\|B(a,\theta_0)\|_2^2\leq\|\theta_0\|_2^2=O(1)$, finishing the proof of the theorem.

§2.10.4 Proof of Corollary 2.1.2

Let $\varepsilon_n=(n/\log^2 n)^{-\beta/(1+2\beta)}$ and first note that in view of assertions (2.10.12) and (2.10.13) combined with triangle inequality and Proposition 2.4.2 we have with P_0-probability tending to one that

$$\|\theta_0-\hat\theta\|_2\leq\|W\|_2+\|B(\theta_0)\|_2\lesssim\sqrt{\underline{a}_n/n}\log(n/\underline{a}_n)\lesssim\varepsilon_n.$$

Then in view of Theorem 2.4.5 and by applying again the triangle inequality we get with probability tending to one that

$$\Pi(\theta:\ \|\theta-\hat\theta\|_2\leq M_n\varepsilon_n|Y)\geq\Pi(\theta:\ \|\theta-\theta_0\|_2\leq M_n\varepsilon_n-\|\theta_0-\hat\theta\|_2|Y)=1-o(1),$$

concluding the proof of the corollary.

§2.10.5 Proof of Lemma 2.10.1

In Section 2.8 it was shown that $\mathbb{M}_n(a) = \frac{\partial \ell_n(a)}{\partial a}$ satisfies, for positive constants K_1, K_2 and K_3,

$$\frac{\mathbb{M}_n(a)}{\log^2(n/a)} \begin{cases} \leq -K_1, & \text{for } a \geq \overline{a}_n \\ \geq K_2 \log(n/\underline{a}_n), & \text{for } a \in [\underline{a}_n^*, \underline{a}_n] \\ \geq -K_3, & \text{for } a \leq \underline{a}_n^*, \end{cases}$$

where $\underline{a}_n^* = \underline{a}_n \log n/(1+\log n)$. Furthermore, the constant K_2 can be chosen arbitrarily large by choosing B large enough, while the constant K_3 is fixed.

For $a \geq C\overline{a}_n$ with $C \geq 3$, we have

$$\ell_n(a) - \ell_n(2\overline{a}_n) \leq -K_1 \log^2(n/\overline{a}_n)(a - 2\overline{a}_n) \leq -K_4 \log^2(n/\overline{a}_n)\overline{a}_n$$

with $K_4 = K_1(C-2)$. Consequently $e^{\ell_n(a)} \leq e^{\ell_n(2\overline{a}_n) - K_4 \log^2(n/\overline{a}_n)\overline{a}_n}$ for $a \geq C\overline{a}_n$. Since also $e^{\ell_n(a)} \geq e^{\ell_n(2\overline{a}_n)}$ for $a \in [\overline{a}_n, 2\overline{a}_n]$, we find

$$\Pi(a \geq C\overline{a}_n | Y) \leq \frac{\int_{C\overline{a}_n}^{\infty} e^{\ell_n(a)}\pi(a)da}{\int_{\overline{a}_n}^{2\overline{a}_n} e^{\ell_n(a)}\pi(a)da} \leq \frac{\Pi\big([C\overline{a}_n, \infty)\big)e^{-K_4 \log^2(n/\overline{a}_n)\overline{a}_n}}{\Pi([\overline{a}_n, 2\overline{a}_n])}. \tag{2.10.14}$$

Note that by Assumption 2.1.1

$$\Pi\big([\overline{a}_n, 2\overline{a}_n]\big) \gtrsim \overline{a}_n^{1-c_3} e^{-c_2\overline{a}_n} \gg e^{-K_4 \log^2(n/\overline{a}_n)\overline{a}_n},$$

hence the right hand side of (2.10.14) tends to zero.

The analysis of the left tail goes similarly. Note that for $a < \underline{a}_n^*/2$ we have $\ell_n(\underline{a}_n^*) - \ell_n(a) \geq -K_3(\underline{a}_n^* - a)\log^2(n/\underline{a}_n)$, hence $e^{\ell_n(a)} \leq e^{\ell_n(\underline{a}_n^*) + K_3\underline{a}_n \log^2(n/\overline{a}_n)}$ and analogously for $(\underline{a}_n + \underline{a}_n^*)/2 < a < \underline{a}_n$ we have $\ell_n(a) - \ell_n(\underline{a}_n^*) \geq K_2(a - \underline{a}_n^*)\log^3(n/\underline{a}_n)$, which implies $e^{\ell_n(a)} \geq e^{\ell_n(\underline{a}_n^*) + K_2(\underline{a}_n/4)\log^2(n/\underline{a}_n)}$. Therefore

$$\Pi(a \leq \underline{a}_n^* | Y) \leq \frac{\int_1^{\underline{a}_n^*} e^{\ell_n(a)}\pi(a)da}{\int_{(\underline{a}_n + \underline{a}_n^*)/2}^{\underline{a}_n} e^{\ell_n(a)}\pi(a)da} \leq \frac{\Pi([1, \underline{a}_n])e^{K_3\underline{a}_n \log^2(n/\underline{a}_n)}}{\Pi([(\underline{a}_n + \underline{a}_n^*)/2, \underline{a}_n])e^{K_2(\underline{a}_n/4)\log^2(n/\underline{a}_n)}}. \tag{2.10.15}$$

Since

$$\Pi([(\underline{a}_n + \underline{a}_n^*)/2, \underline{a}_n])^{-1} \lesssim \log(n)\underline{a}_n^{c_5 - 1} e^{c_6\underline{a}_n} \ll e^{K_2(\underline{a}_n/8)\log^2(n/\underline{a}_n)},$$

for large enough choice of K_2, the right hand side of (2.10.15) tends to zero, finishing the proof of the lemma.

§2.11 Technical Lemmas

Lemma 2.11.1. *Let $i, m \in \mathbb{N}$ and $a \geq 1$, then for any $n/a \geq e^m$*

$$\frac{ne^{i/a}i^m}{a^m(ae^{i/a} + n)^2} \leq \frac{1}{a}\log^m\big(\frac{n}{a}\big) \vee e\frac{a^{-m}}{n}.$$

Proof. Assume first that $i \leq I_a \equiv a\log(n/a)$. Note that the function $f(x) = e^{x/a}(x/a)^m$ is monotone decreasing on $(-\infty, -ma]$ and monotone increasing on $[-ma, \infty]$. Then by the inequality $ae^{i/a} + n \geq n$,

$$\frac{ne^{i/a}i^m}{a^m(ae^{i/a}+n)^2} \leq \frac{e^{i/a}(i/a)^m}{n} \leq \frac{1}{a}\log^m\big(\frac{n}{a}\big) \vee e\frac{a^{-m}}{n}.$$

Next assume that $i > I_a$. Note that the derivative of the function $f(x) = e^{-x/a}x^m$ is $f'(x) = e^{-x/a}x^{m-1}(m - x/a)$, hence the function $f(i)$ is monotone decreasing for $i \geq am$. Thus for $n/a \geq e^m$, $f(i)$ takes its maximum at $i = I_a$, which implies that

$$\frac{ne^{i/a}i^m}{a^m(ae^{i/a}+n)^2} \leq \frac{ne^{-i/a}i^m}{a^{m+2}} \leq \frac{1}{a}\log^m\big(\frac{n}{a}\big).$$

□

Lemma 2.11.2. *Let $l > r \geq 0$, then for $n/a \geq e^{l-r}$*

$$\sum_{i=1}^{\infty} \frac{e^{ir/a}}{(ae^{i/a}+n)^l} \lesssim \frac{n^{r-l}}{a^{r-1}}\log\big(\frac{n}{a}\big).$$

Proof. First note that following from the inequality $ae^{i/a} + n \geq n$ and the sum of geometric series we get

$$\sum_{i=1}^{I_a} \frac{e^{ir/a}}{(ae^{i/a}+n)^l} \leq n^{-1}\sum_{i=1}^{I_a} e^{ir/a} \lesssim \frac{n^{r-l}}{a^{r-1}}\log\big(\frac{n}{a}\big),$$

where $I_a \equiv a\log(n/a)$. Then similarly, using the inequality $ae^{i/a} + n \geq ae^{i/a}$ and the sum of geometric series,

$$\sum_{I_a}^{\infty} \frac{e^{ir/a}}{(ae^{i/a}+n)^l} \leq a^{-l}\sum_{I_a}^{\infty} e^{(r-l)i/a} \leq \frac{n^{r-l}}{a^r}\frac{1}{e^{(l-r)/a}-1} \lesssim \frac{n^{r-l}}{a^{r-1}}\log\big(\frac{n}{a}\big)$$

because $e^{(l-r)/a} - 1 \geq \frac{l-r}{a}$ and $\log\big(\frac{n}{a}\big) \geq l - r$ for $\frac{n}{a} \geq e^{l-r}$. □

Lemma 2.11.3 (Lemma C.11 of (van der Pas et al., 2017))**.** *For any stochastic process $(V_a : a > 0)$ with continuously differentiable sample paths $a \mapsto V_a$, with derivative written as $\dot{V}_a$,*

$$E(V_{a_2} - V_{a_1})^2 \leq (a_2 - a_1)^2 \sup_{a\in[a_1,a_2]} E\dot{V}_a^2.$$

§2.12 Extra simulation study

The purpose of this section is to reinforce the evidence shown in Section 2.2. To this end, we will show graphically and numerically the sub-optimal performance of the Gaussian process with (approximately) squared exponential covariance kernel compared to other methods in the non-parametric regression model specifically. In this

simulation study we take the Fourier coefficients of the underlying true function θ_3 to be $\theta_{3,i} = i^{-3/2}\cos(i)$, $i = 1, 2, ...$ We take $\sigma^2 = 1/2$, but in the procedure it is considered to be unknown and estimated with the MMLE $\hat{\sigma}^2$. We take the sample size to be $n = 500, 1000, 5000$, and 10000. Observe in Figure 2.4 that the standard MMLE empirical Bayes method provides unreliable uncertainty quantification in certain points, especially compared to the three other methods. One can also observe through different running times in Table 2.9, that while the Matérn covariance kernels might provide robust credible sets, they substantially slow down the computations for large n.

We also investigate empirically the frequentist coverage probabilities of the point-wise credible sets by repeating the experiment 100 times and reporting the frequency that the function at given points (we consider $x = (0.25, 0.6474, 0.75)$ with $0.6474 = \text{argmax}_{x\in[0,1]}\theta_3(x)$) is included in the credible interval, see Table 2.7. Moreover, Table 2.8 shows the average size of the point-wise credible intervals (i.e. $2q_{0.025}\sqrt{\hat{c}(x,x)}$) depending on the sample size n and the procedure used to compute the credible sets. One can observe similar behavior to what we have described above.

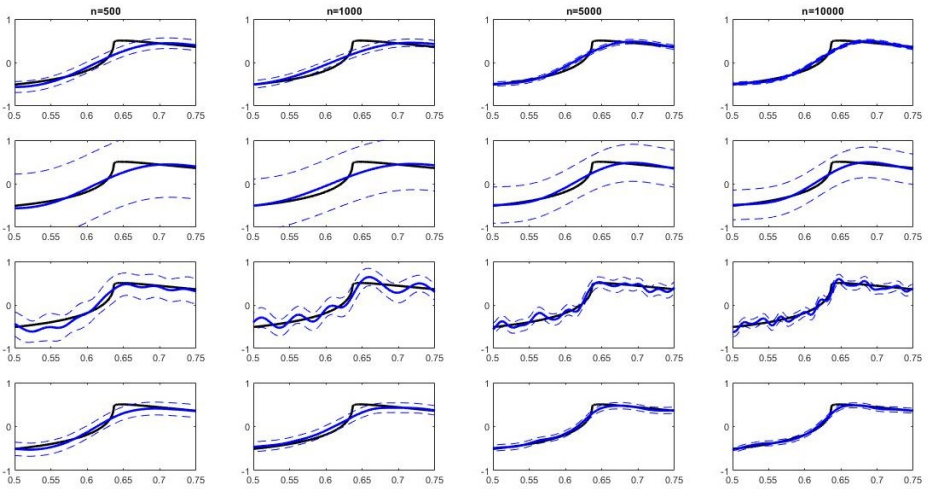

Figure 2.4: Empirical Bayes credible sets for the regression function θ_3 (drawn in black), zoomed in to the interval $x \in [0.5, 0.75]$. The posterior means are drawn by solid blue line, while the 95% point-wise credible sets by dashed blue curves. In the first row we plot the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, the third row the modified MMLE empirical Bayes method using squared exponential Gaussian process prior, while in the fourth row we plot the empirical Byes credible sets using a Matérn kernel with data-driven choice for the regularity hyper-parameter. From left to right the sample size is $n = 500, 1000, 5000, 10000$.

We also consider a multi-variable version of the previous regression with $d = 10$ variables. The Fourier coefficients of the underlying true function θ_4 become $\theta_{4,i} = \prod_{k=1}^{10}(i_k^{-3/2}\cos(i_k))$, $i_k = 1, 2, ...$ for all $k = 1, 2, ..., 10$, relative to the Fourier eigenbasis $\psi_i(t) = 32\prod_{k=1}^{10}\cos(\pi(i_k - 1/2)t)$. We have collected the frequentist coverage probabilities of the point-wise credible sets at given points (we consider $x = (\{0.25\}^{10}, \{0.3188\}^{10}, \{0.75\}^{10})$) in Table 2.10 and note that similar conclusions

	$x = 0.25$			$x = 0.6474$			$x = 0.75$		
$n =$	100	500	1000	100	500	1000	100	500	1000
Method 1	0.00	0.00	0.00	0.86	0.82	0.71	0.00	0.01	0.00
Method 2	0.94	1.00	1.00	0.64	1.00	1.00	1.00	1.00	1.00
Method 3	0.11	0.12	0.17	0.95	0.95	0.96	0.62	0.47	0.75
Method 4	0.00	0.13	0.14	0.86	0.86	0.90	0.24	0.27	0.35
Method 5	0.00	0.08	0.10	0.83	0.84	0.87	0.19	0.19	0.34

Table 2.7: Frequencies that $\theta_3(x)$ is inside of the corresponding credible interval for the squared exponential and Matérn Gaussian process prior at given points $x \in \{0.25, 0.3188, 0.75\}$. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$), Method 4: Matérn kernel with smoothness MMLE empirical Bayes, Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

$n =$	100	500	1000
Method 1	0.4184	0.2335	0.1804
Method 2	1.9270	1.4516	1.2459
Method 3	0.7949	0.5271	0.4267
Method 4	0.6694	0.4292	0.3446
Method 5	0.5439	0.2987	0.2625

Table 2.8: Average size of the pointwise credible intervals (i.e. $2q_{0.025}\sqrt{\hat{c}(x,x)}$) for $\theta_3(x)$ in the regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$), Method 4: Matérn kernel with smoothness MMLE empirical Bayes, Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

$n =$	100	500	1000	5000	10000
Method 1	0.82 s	2.70 s	10.66 s	3.6 m	19.1 m
Method 4	1.61 s	14.52 s	45.77 s	19 m	4.2 h
Method 5	1.37 s	11.45 s	34.29 s	14.1 m	2.4 h

Table 2.9: Average run time of the EB methods for θ_3 in the regression model. Method 1: SE covariance kernel, Method 4: Matérn covariance kernel and MMLE for the regularity hyper-parameter, Method 5: Matérn covariance kernel and MMLE for the scaling hyper-parameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000, 10000$

can be drawn as in the $d = 1$ dimensional case. Surprisingly, the computation times for the posterior in higher dimension is of similar order as their one-dimensional counterpart, hence they are omitted.

	$x = \{0.25\}^{10}$			$x = \{0.3188\}^{10}$			$x = \{0.75\}^{10}$		
$n =$	100	500	1000	100	500	1000	100	500	1000
Method 1	1.00	0.97	0.96	0.85	0.76	0.70	1.00	0.98	0.95
Method 2	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00
Method 3	1.00	1.00	1.00	0.86	0.87	0.89	1.00	1.00	1.00
Method 4	1.00	1.00	1.00	0.96	0.95	0.97	1.00	1.00	1.00
Method 5	1.00	1.00	1.00	0.95	0.95	0.94	1.00	1.00	1.00

Table 2.10: Frequencies that $\theta_4(x)$ is inside of the corresponding credible interval for the squared exponential and Matérn Gaussian process prior in the multivariate ($d = 10$) regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$), Method 4: Matérn kernel with smoothness MMLE empirical Bayes, Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

CHAPTER 3

Optimal recovery and coverage for distributed Bayesian non-parametric regression

Abstract. Gaussian Processes (GP) are widely used for probabilistic modeling and inference for non-parametric regression. However, their computational complexity scales cubicly with the sample size rendering them unfeasible for large data sets. To speed up the computations various distributed methods were proposed in the literature. These methods have, however, limited theoretical underpinning. In our work we derive frequentist theoretical guarantees and limitations for a range of distributed methods for general GP priors in context of the non-parametric regression model, both for recovery and uncertainty quantification. As specific examples we consider covariance kernels both with polynomially and exponentially decaying eigenvalues. We demonstrate the practical performance of the investigated approaches in a numerical study using synthetic data sets.

§3.1 GP regression framework

In our analysis we consider the multivariate random design regression model. Let us assume that we observe (X_i, Y_i), $i = 1, ..., n$, i.i.d pairs of random variables satisfying

$$Y_i = \theta_0(X_i) + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right), \tag{3.1.1}$$

with design points X_i, $i = 1, ..., n$, belonging to some compact set $\mathcal{X} \subset \mathbb{R}^d$, observations $Y_i \in \mathbb{R}$, noise variance $\sigma^2 > 0$, and functional parameter $\theta_0 : \mathcal{X} \to \mathbb{R}$. For simplicity we take $\mathcal{X} = [0, 1]^d$, assume that the design points are uniformly distributed, i.e. $X_i \overset{iid}{\sim} U[0, 1]^d$, and $\sigma^2 \gtrsim 1$ to be known. We use the notation $\mathbb{D}_n = (Y_i, X_i)_{i=1,...,n}$ for the observations and P_0 and E_0 for the probability measure and expected value corresponding to the underlying regression function θ_0.

In order to perform inference on the regression function θ_0, we consider a non-parametric Bayesian approach. We endow θ_0 with a mean-zero Gaussian Process (GP) prior $GP(0, K)$, where $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is a positive definite stationary kernel. For matrices $A \in \mathbb{R}^{d\times n}$ and $B \in \mathbb{R}^{d\times n'}$, let $K(A, B)$, denote the $n \times n'$ matrix of $(K(A_{\cdot i}, B_{\cdot j}))_{1\le i\le n,\, 1\le j\le n'}$.

By conjugacy the posterior distribution of θ is also a Gaussian process and by the same conjugate computation as in Chapter 2 of (Rasmussen and Williams, 2006), $\theta|\mathbb{D}_n \sim \mathrm{GP}(\hat{\theta}_n, \hat{C}_n)$, where for any $x, x' \in [0,1]^d$

$$\hat{\theta}_n(x) = K(x, \mathbb{X})\left(K(\mathbb{X}, \mathbb{X}) + \sigma^2 I_n\right)^{-1} \mathbb{Y}, \tag{3.1.2}$$

$$\hat{C}_n(x, x') = K(x, x') - K(x, \mathbb{X})\left(K(\mathbb{X}, \mathbb{X}) + \sigma^2 I_n\right)^{-1} K(\mathbb{X}, x'), \tag{3.1.3}$$

where $\mathbb{X} \in [0,1]^{d \times n}$, $\mathbb{Y} \in \mathbb{R}^n$ are the collection of design points and observations, respectively, and I_n denotes the $n \times n$ identity matrix.

We assume that the eigenfunctions $\{\psi_j\}_{j \in \mathbb{N}^d}$ of the above covariance kernel K factorize, i.e.

$$\psi_j = \prod_{k=1}^{d} \psi_{j_k}, j \in \mathbb{N}^d, \tag{3.1.4}$$

where $\{\psi_{j_k}\}_{j_k \in \mathbb{N}}$ are the eigenfunctions corresponding to the one dimensional kernel on $[0,1]$. We further assume that the eigenfunctions of the kernel K are bounded.

Assumption 3.1.1. *There exists a global constant $C_\psi > 0$ such that the eigenfunctions $\{\psi_j\}_{j \in \mathbb{N}^d}$ of K satisfy $|\psi_j(t)| \leq C_\psi$ for all $j \in \mathbb{N}^d, t \in \mathcal{X}$.*

The corresponding eigenvalues of K are

$$\mu_j = \prod_{k=1}^{d} \mu_{j_k}, j \in \mathbb{N}^d, \tag{3.1.5}$$

with $\{\mu_{j_k}\}_{j_k \in \mathbb{N}}$ the eigenvalues of the k-th component of the kernel (Berlinet and C. Thomas-Agnan, 2004). Although our results hold more generally, as specific examples we consider polynomially and exponentially decaying eigenvalues

Assumption 3.1.2. *The one dimensional eigenvalues μ_j, $j \in \mathbb{N}$ are either*

- *Polynomially decaying:*

$$C^{-1} j^{-2\alpha/d-1} \leq \mu_j \leq C j^{-2\alpha/d-1}, \tag{3.1.6}$$

  for some $\alpha, C > 0$, or

- *Exponentially decaying:*

$$C^{-1} b e^{-aj} \leq \mu_j \leq C b e^{-aj}, \tag{3.1.7}$$

  for some $a, b, C > 0$.

In non-parametric statistics, it is common to assume that the underlying functional parameter of interest belongs to some regularity class. In our analysis we consider Sobolev-type of regularity classes defined with the basis ψ_j, i.e. for any $\beta > 0$ and $B > 0$, define as in (Bényi and Oh, 2013), (Hunter, 2013) and (Cobos et al., 2015) the function space

$$\Theta^\beta(B) = \left\{ \theta = \sum_{j \in \mathbb{N}^d} \theta_j \psi_j \in L_2\left([0,1]^d\right) : \sum_{j \in \mathbb{N}^d} \left(\sum_{k=1}^{d} j_k \right)^{2\beta} \theta_j^2 \leq B^2 \right\}. \tag{3.1.8}$$

For the Fourier basis or the basis corresponding to the Matérn covariance kernel, $\Theta^\beta(B)$ is equivalent to β-smooth Sobolev balls and are known as *isotropic Sobolev spaces*, see (Cobos et al., 2015).

The frequentist properties of Gaussian process priors for recovery are well understood in the literature. It was shown in various specific examples and choices of priors that for appropriately scaled Gaussian priors the corresponding posterior can recover the underlying functional parameter of interest $\theta_0 \in \Theta^\beta(B)$ with the optimal minimax estimation rate $n^{-\beta/(2\beta+d)}$, see for instance (van der Vaart and van Zanten, 2007), (van der Vaart and van Zanten, 2008) and (van der Vaart and van Zanten, 2011). Another, from a practical perspective very appealing property of Bayesian methods is the built-in uncertainty quantification. Bayesian credible sets accumulate prescribed (typically 95%) posterior mass and can take various forms. In our analysis we consider L_2 credible balls, i.e. we define the credible set as $\hat{B}_n = \{\theta : \|\theta - \hat{\theta}_n\| \leq r_\gamma\}$, satisfying $\Pi(\theta \in \hat{B}_n | \mathbb{D}_n) = 1 - \gamma$, for some $\gamma \in (0, 1)$. Credible sets do not provide automatically valid confidence statements. In recent years the frequentist coverage properties of Bayesian credible sets were widely studied and it was shown for appropriate choices of the prior distribution the corresponding posterior can provide reliable frequentist uncertainty quantification for functions satisfying certain regularity assumptions, see for instance (Szabo et al., 2015), (Belitser, 2017), (Castillo and Nickl, 2014), (Serra and Krivobokova, 2017), (Sniekers and van der Vaart, 2015a), (Yoo and Ghosal, 2016), (Bhattacharya et al., 2017), (Ray, 2017), (Rousseau and Szabo, 2020) and (Hadji and Szabo, 2021). However, our setting wasn't covered by these results yet.

Despite the fact that the mean (3.1.2) and covariance (3.1.3) functions can be explicitly computed, consequently solving the model, their computation requires inverting the matrix $(K(\mathbb{X}, \mathbb{X}) + \sigma^2 I_n)$. The inversion of this $n \times n$ matrix is of $O(n^3)$ computational complexity, which rapidly explodes as n grows. One way to speed up the computations is to consider sparse approximations of the matrices, see for instance (Gibbs et al., 1976), (Saad, 1990), (Quiñonero-Candela and Rasmussen, 2005) and (Titsias, 2009). In this work we focus on a different, distributed approach to decrease computational complexity.

§3.2 Distributed GP regression

In distributed methods, the data are divided among multiple local machines or servers, and the computations are carried out locally, in parallel to each other. Then the outcome of the computations are transmitted to a center machine or server where they are aggregated somehow forming the final outcome of the distributed method. In the random design regression model it means that we divide the data of size n over m machines (we assume for simplicity that $n \bmod m = 0$), i.e. in each machine $k = 1, ..., m$ we observe iid pairs of random variables $(X_i^{(k)}, Y_i^{(k)}) \in [0, 1]^d \times \mathbb{R}$, $i = 1, ..., n/m$, satisfying

$$Y_i^{(k)} = \theta_0\left(X_i^{(k)}\right) + \varepsilon_i^{(k)}, \quad \varepsilon_i^{(k)} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right), \tag{3.2.1}$$

where $\theta_0 : [0, 1]^d \mapsto \mathbb{R}$ is the unknown functional parameter of interest, and $\sigma^2 > 0$ the known variance of the noise. For convenience, let us introduce the notations $\mathbb{D}_n^{(k)} =$

$(X_i^{(k)}, Y_i^{(k)})_{i=1,\dots,\tilde{n}}$, $\mathbb{X}^{(k)} = (X_i^{(k)})_{i=1,\dots,\tilde{n}}$, $\mathbb{Y}^{(k)} = (Y_i^{(k)})_{i=1,\dots,\tilde{n}}$ for the whole data set, the design points, and observations in the k-th local machine, respectively. Similarly to the non-distributed method (with only one local machine $m = 1$), we assume that the true function belongs to some Sobolev-type of regularity class $\theta_0 \in \Theta^\beta(B)$, for given $\beta, B > 0$, see (3.1.8).

We consider distributed Bayesian approaches for recovering θ_0. First, we endow the function θ_0 in each local machine $k = 1, ..., m$ with a Gaussian process prior and compute the corresponding local (adjusted) posterior distribution $\Pi^{(k)}(.|\mathbb{D}_n^{(k)})$. Then, we transmit the m local posteriors into a central machine where we aggregate them somehow into a global (adjusted) posterior $\Pi_{n,m}^\dagger(\cdot|\mathbb{D}_n)$. We further denote by $\hat{\theta}_n^{(k)}$ the local (adjusted) posterior mean, and by $\hat{\theta}_{n,m}$ the global (adjusted) posterior mean. For quantifying the uncertainty of the distributed Bayesian procedure we consider L_2-credible balls resulting in from the aggregated posterior distribution, i.e. let

$$\begin{aligned} &\hat{B}_{n,m,\gamma} = \left\{\theta : \|\theta - \hat{\theta}_{n,m}\|_2 \le r_{n,m,\gamma}\right\}, \quad \text{satisfying} \\ &\Pi_{n,m}^\dagger\left(\theta \in \hat{B}_{n,m,\gamma}|\mathbb{D}_n\right) = 1 - \gamma, \end{aligned} \tag{3.2.2}$$

for some prescribed $\gamma \in (0, 1)$.

Distributed methods vary according to the way the local (adjusted) posterior distributions are computed and aggregated to obtain the global posterior. The behavior of the aggregated posterior crucially depends on the applied techniques. To demonstrate this let us consider a naive method where in each local machine we endow $\theta_0 \in \Theta^\beta(B)$ with a Gaussian process prior and compute the corresponding unadjusted local posterior distribution $\Pi_n^*(\cdot|\mathbb{D}_n^{(k)})$. We consider a centered GP with polynomially decaying eigenvalues as in Assumption 3.1.2 with matching regularity hyper-parameter $\alpha = \beta$. Note that this choice of the hyper-parameter is optimal in the non-distributed case (with only one local machine $m = 1$). Then the local posteriors are aggregated to a global posterior $\Pi_{n,m}^\dagger(\cdot|\mathbb{D}_n)$ in the following way: a draw from the aggregated posterior is taken to be the average of a single draw from each local posteriors. The theorem below shows that such method results in sub-optimal concentration for the posterior mean and contraction rate for the whole posterior distribution.

Theorem 3.2.1. *Take $\beta \ge 2$ and consider the function $\theta_0 \in \Theta_\beta(L)$ of the form $\theta_0(x) = c_L \sum_{j=1}^\infty j^{-1-2\beta}(\log j)^{-2}\psi_j(x)$, $x \in [0, 1]$, for sufficiently small $c_L > 0$. Then for the covariance kernel K with polynomially decaying eigenvalues (3.1.6) with $\alpha = \beta$ and $d = 1$, and $(\log n)^2 \ll m \lesssim n^{1/(1+2\beta)}$ the corresponding naive aggregated posterior mean $\hat{\theta}_{n,m}$ has sub-optimal concentration and the posterior itself achieves sub-optimal contraction rate, i.e.*

$$E_0 \left\|\hat{\theta}_{n,m} - \theta_0\right\|_2^2 \ge c\,(\log n)^{-2}\,(n/m)^{-\beta/(2\beta+1)}, \tag{3.2.3}$$

$$E_0 \Pi_{n,m}^\dagger\left(\theta : \|\theta - \theta_0\|_2^2 \le c\,(\log n)^{-2}\,(n/m)^{-\beta/(2\beta+1)}|\mathbb{D}_n\right) \to 0, \tag{3.2.4}$$

for sufficiently small $c > 0$, where $\hat{\theta}_{n,m}$ is the mean of the global posterior $\Pi_{n,m}^\dagger$ obtained with the naive method.

The proof is given in Section 3.5.4.

§3.2.1 Optimal Distributed Methods

In this paper we consider two methods, for which optimal performance were derived in context of the Gaussian white noise setting, see (Szabó and van Zanten, 2019). We investigate these methods here in the practically more relevant and technically substantially more complex non-parametric regression model. We note that in (Guhaniyogi et al., 2017) in context of the regression model an approach closely related to Method II was derived and its contraction properties were investigated for a rescaled covariance kernel with polynomially decaying eigenvalues. In our work we consider more general kernel structures and in contrast to (Guhaniyogi et al., 2017) do not require that the functional parameter belongs to the Reproducing Kernel Hilbert Space (RKHS) of the Gaussian Process prior. Furthermore, we also derive guarantees and limitations to uncertainty quantification. Therefore, our results are of different nature requiring a different approach.

3.2.1.1 Method I

Rescaling the priors. In the first method, introduced by (Scott et al., 2016) in a parametric setting, we consider raising the prior density to the power $1/m$, which is formally equivalent to multiplying the kernel K by m, i.e. the adjusted kernel takes the form $K^I := mK$. Then the eigenvalues of the kernel K^I are $\{\mu_j^I\}_{j\in\mathbb{N}^d} = \{m\mu_j\}_{j\in\mathbb{N}^d}$. Hence, in view of (3.1.1) the posterior distribution, for each machine $k = 1, ..., m$, is also a Gaussian process $\theta|\mathbb{D}_n^{(k)} \sim \mathrm{GP}(\hat{\theta}_n^{(k)}, \hat{C}_n^{(k)})$ with

$$\hat{\theta}_n^{(k)}(x) = K\left(x, \mathbb{X}^{(k)}\right)\left(K\left(\mathbb{X}^{(k)}, \mathbb{X}^{(k)}\right) + m^{-1}\sigma^2 I_{\tilde{n}}\right)^{-1}\mathbb{Y}^{(k)},$$
$$\hat{C}_n^{(k)}(x, x') = m\left(K(x, x') - K\left(x, \mathbb{X}^{(k)}\right)\left(K\left(\mathbb{X}^{(k)}, \mathbb{X}^{(k)}\right) + m^{-1}\sigma^2 I_{\tilde{n}}\right)^{-1} K(\mathbb{X}^{(k)}, x')\right).$$

Averaging the local draws. A draw from the global posterior is generated by first drawing a single sample from each local posteriors and then taking the averages of these draws over all machines. Since the data sets and the priors in the local machines are independent, the so generated average of the local posteriors is also a Gaussian process with mean $\hat{\theta}_{n,m}^I = m^{-1}\sum_{k=1}^m \hat{\theta}_n^{(k)}$ and covariance kernel $\hat{C}_{n,m}^I = m^{-2}\sum_{k=1}^m \hat{C}_n^{(k)}$, where $\hat{\theta}_n^{(k)}$ and $\hat{C}_n^{(k)}$ denote the posterior mean and covariance functions in the kth local machine.

3.2.1.2 Method II

Rescaling the likelihood. In the second method proposed by (Srivastava et al., 2015), we adjust the local likelihood by raising its power to m in every machine, which is equivalent to rescaling the variance of the observations by a factor m^{-1}. Then, by elementary computations similar to (3.1.1), we obtain that for each machine, the

posterior distribution is $GP(\hat{\theta}_n^{(k)}, \hat{C}_n^{(k)})$, with

$$\hat{\theta}_n^{(k)}(x) = K\left(x, \mathbb{X}^{(k)}\right)\left(K\left(\mathbb{X}^{(k)}, \mathbb{X}^{(k)}\right) + m^{-1}\sigma^2 I_{\tilde{n}}\right)^{-1}\mathbb{Y}^{(k)},$$
$$\hat{C}_n^{(k)}(x, x') = K(x, x') - K\left(x, \mathbb{X}^{(k)}\right)\left(K\left(\mathbb{X}^{(k)}, \mathbb{X}^{(k)}\right) + m^{-1}\sigma^2 I_{\tilde{n}}\right)^{-1} K\left(\mathbb{X}^{(k)}, x'\right).$$

Wasserstein barycenter. This approach consists in aggregating the local posteriors by computing their Wasserstein barycenter. The 2-Wasserstein distance $W_2^2(\mu, \nu)$ between two probability measures μ and ν is defined as

$$W_2^2(\mu, \nu) := \inf_{\gamma} \int\int \|x - y\|_2^2 \gamma(dx, dy),$$

where the infimum is taken over all measures γ with marginals μ and ν. The corresponding 2-Wasserstein barycenter of m probability measures $\mu_1, ..., \mu_m$ is defined by

$$\bar{\mu} = \arg\min_{\mu} \frac{1}{m}\sum_{k=1}^{m} W_2^2(\mu, \mu_k),$$

where the minimum is taken over all probability measures with finite second moments. In view of Theorem 4 in (Mallasto and Feragen, 2017), the global posterior is a Gaussian process with mean $\hat{\theta}_{n,m}^{II}$ and covariance $\hat{C}_{n,m}^{II}$ satisfying

$$\hat{\theta}_{n,m}^{II} = \frac{1}{m}\sum_{k=1}^{m} \hat{\theta}_n^{(k)},$$
$$\hat{C}_{n,m}^{II} = \frac{1}{m}\sum_{k=1}^{m}\left(\left(\hat{C}_{n,m}^{II}\right)^{1/2}\hat{C}_n^{(k)}\left(\hat{C}_{n,m}^{II}\right)^{1/2}\right)^{1/2}.$$

In particular, the posterior variance function is

$$\mathrm{Var}_{n,m}^{II}\left(f(x)|\mathbb{D}_n\right) = \frac{1}{m}\sum_{k=1}^{m}\mathrm{Var}\left(f(x)|\mathbb{D}_n^{(k)}\right)$$

for all $x \in \mathcal{X}$.

§3.2.2 Posterior contraction rate

We show that the above proposed distributed methods (i.e. Methods I- II) provide optimal recovery of the underlying functional parameter of interest. The methods result in different global posteriors which can have different finite sample size behavior, but their asymptotic properties are similar.

Theorem 3.2.2. *Let $\beta, B > 0$, K a kernel with eigenvalues $(\mu_j)_{j\in\mathbb{N}^d}$ satisfying $|\{j \in \mathbb{N}^d : \mu_j n \geq \sigma^2\}| \leq n$ and corresponding eigenfunctions satisfying Assumption 3.1.1. Furthermore, let*

$$\nu_j = \frac{n\mu_j}{\sigma^2 + n\mu_j}, \quad \textit{for all } j \in \mathbb{N}^d, \tag{3.2.5}$$

and $\tilde{P}$ *a linear operator defined as* $\tilde{P}(\theta) := \sum_{j\in\mathbb{N}^d}(1-\nu_j)\theta_j\psi_j$ *for all* $\theta \in L^2(\mathcal{X})$. *Then*

$$E_0\|\hat{\theta}_{n,m} - \theta_0\|_2^2 \lesssim \|\tilde{P}(\theta_0)\|_2^2 + \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j^2 + \delta_n, \tag{3.2.6}$$

$$E_0\Pi_{n,m}^{\dagger}\left(\|\theta-\theta_0\|_2^2 > M_n\Big(\|\tilde{P}(\theta_0)\|_2^2 + \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j + \delta_n\Big)|\mathbb{D}_n\right) \to 0, \tag{3.2.7}$$

for arbitrary sequence M_n *tending to infinity, where* $\hat{\theta}_{n,m}$ *is the mean of the global posterior* $\Pi_{n,m}^{\dagger}(.|\mathbb{D}_n)$ *obtained with either Methods* $I-II$ *and*

$$\delta_n = \inf\Big\{n\sum_{j\in\mathbb{N}^d}\nu_j^2\sum_{\ell\in\mathcal{I}^c}\mu_\ell : \mathcal{I}\subset\mathbb{N}^d, |\mathcal{I}| \le n\big(m^2\log n\sum_{j\in\mathbb{N}^d}\nu_j^2\big)^{-1}\Big\} \tag{3.2.8}$$

is a (typically) negligible technical term.

The proof of the theorem is deferred to Section 3.5.3.

First we note that the condition $|\{j\in\mathbb{N}^d : \mu_j n \ge \sigma^2\}| \le N$ is very mild and is satisfied by the eigenvalues considered in Assumption 3.1.2. The sequence $(\nu_j)_{j\in\mathbb{N}}$ can be thought of as the population eigenvalues of the posterior. Next note that the bound (3.2.6) has two main components. The first term $\|\tilde{P}(\theta_0)\|_2^2$ measures how close θ_0 is (in L_2-norm) to its convolution with the eigenvalues $(\nu_j)_{j\in\mathbb{N}^d}$, hence it accounts for the bias of the estimator. In the meanwhile the second term $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j^2$ can be thought of as the variance term. In a similar fashion, the contraction rate (3.2.7) has also two main components: $\|\tilde{P}(\theta_0)\|_2^2$ and $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$, where the former is the squared bias while the latter is the expected value of the posterior variance under the true parameter. The remaining δ_n term is of technical nature. It bounds the tail behavior of the eigen-decomposition of the variance of the posterior mean. This term is shown to be negligible in our examples. Since all the above terms are related to the kernel K, explicit bounds on the expectation of $\|\hat{\theta}_n - \theta_0\|_2$, as well as explicit posterior contraction rates of the global posterior $\Pi_{n,m}^{\dagger}(.|\mathbb{D}_n)$, can be achieved for specific choices of the kernels.

Corollary 3.2.3. ***(Polynomial)*** *For given* $B > 0$ *and* $\beta \ge 3d/2$, *assume that the covariance kernel* K *satisfies Assumptions 3.1.1 and* (3.1.6) *with* $\alpha = \beta$. *Then for* $m = o(n^{\frac{2\beta-3d}{4\beta}})$ *the aggregated posterior distribution* $\Pi_{n,m}^{\dagger}(.|\mathbb{D}_n)$ *and the corresponding aggregated posterior mean* $\hat{\theta}_{n,m}$ *resulting from either of the Methods* $I-II$ *achieve the minimax convergence rate up to a logarithmic factor, i.e.*

$$\sup_{\theta_0\in\Theta^\beta(B)} E_0\|\hat{\theta}_{n,m} - \theta_0\|_2^2 \lesssim (n/\sigma^2)^{-2\beta/(2\beta+d)}(\log(n/\sigma^2))^{d-1}$$

and for all sequences $M_n \to +\infty$,

$$\sup_{\theta_0\in\Theta^\beta(B)} E_0\Pi_{n,m}^{\dagger}(\theta : \|\theta-\theta_0\|_2 > M_n(n/\sigma^2)^{-\beta/(2\beta+d)}(\log(n/\sigma^2))^{(d-1)/2}|\mathbb{D}_n) \to 0.$$

The proof is given in Section 3.6.1.

Corollary 3.2.4. ***(Exponential)*** *For given $B > 0$ and $\beta \geq d/2$ assume that the covariance kernel K satisfies Assumptions 3.1.1 and* (3.1.7) *with rescaling parameter $a = (\sigma^2/n)^{1/(2\beta+d)} \log(n/\sigma^2)$ and $b = 1$. Then for $m = o(n^{\frac{2\beta-d}{2(2\beta+d)}})$ the aggregated posterior distribution $\Pi^{\dagger}_{n,m}(.|\mathbb{D}_n)$ and the corresponding aggregated posterior mean $\hat{\theta}_{n,m}$ resulting from either of the methods $I - II$ achieve the minimax convergence rate, i.e.*

$$\sup_{\theta_0 \in \Theta^{\beta}(B)} E_0 \left\| \hat{\theta}_{n,m} - \theta_0 \right\|_2^2 \lesssim (n/\sigma^2)^{-2\beta/(2\beta+d)},$$

and for all sequences $M_n \to +\infty$,

$$\sup_{\theta_0 \in \Theta^{\beta}(B)} E_0 \Pi^{\dagger}_{n,m} \left(\theta : \|\theta - \theta_0\|_2 > M_n (n/\sigma^2)^{-\beta/(2\beta+d)} | \mathbb{D}_n \right) \to 0.$$

The proof is given in Section 3.6.2. We note that the conditions on β and m in both corollaries follow from the remaining technical term δ_n. These conditions are not optimized and are of technical nature.

§3.3 Distributed uncertainty quantification

In the following, we study the frequentist coverage properties of the L_2 credible balls defined in (3.2.2) resulting from Method I. For convenience we allow some additional flexibility by allowing the credible balls to be blown up by a constant factor $L > 0$, i.e. we consider balls

$$\hat{B}_{n,m,\gamma}(L) = \left\{ \theta \in L_2(\mathcal{X}) : \left\| \theta - \hat{\theta}_{n,m} \right\| \leq L r_{n,m,\gamma} \right\},$$

where for the choice $L = 1$ we get back our original credible ball (3.2.2). The frequentist validity of $\hat{B}_{n,m,\gamma}(L)$ will be established in two steps: First we approximate the centered posterior measure $\theta - \hat{\theta}_{n,m}|\mathbb{D}_n$ and second we study the asymptotic behavior of the radius, the bias and the variance of the posterior mean corresponding to the approximated posterior.

In the non-distributed case (i.e. $m = 1$), the posterior distribution can be approximated by an auxiliary GP. For the GP posterior $\theta - \hat{\theta}_n|\mathbb{D}_n \sim \mathrm{GP}(0, \hat{C}_n)$, the covariance kernel $\hat{C}_n$ given in (3.1.3) is hard to analyze due to its dependence on $\mathbb{X}$. Against this background, following the idea of (Bhattacharya et al., 2017), we define a population level GP $\tilde{W} \sim \mathrm{GP}(0, \tilde{C}_n)$, where $\tilde{C}_n(x, x') = \sigma^2/n \sum_{j \in \mathbb{N}^d} \nu_j \boldsymbol{\psi}_j(x) \boldsymbol{\psi}_j(x')$, and show that the two kernels are close with respect to the L_2-norm. Then using this result we can provide the following frequentist coverage results for the credible balls.

Theorem 3.3.1. *Let $\beta, B > 0$, K be a kernel with eigenvalues $(\mu_j)_{j \in \mathbb{N}^d}$ satisfying $|\{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}| \leq n$ and corresponding eigenfunctions satisfying Assumption 3.1.1. Furthermore, assume that $n\delta_n / \sum_{j \in \mathbb{N}^d} \nu_j = o(1)$, where the (typically) negligible term δ_n was defined in* (3.2.8)*. Then in case the bias term $\|\tilde{P}(\theta_0)\|_2$ satisfies that*

$$\frac{n}{\sigma^2} \frac{\|\tilde{P}(\theta_0)\|_2^2}{\sum_{j \in \mathbb{N}^d} \nu_j} \leq c \tag{3.3.1}$$

for some $c \geq 0$, the frequentist coverage of the (inflated) credible set resulting from Method I tends to one, i.e. for arbitrary $L_n \to +\infty$

$$P_{\theta_0}\left(\theta_0 \in \hat{B}_{n,m,\gamma}\left(L_n\right)\right) \overset{n\to\infty}{\to} 1.$$

On the other hand, if the bias term $\|\tilde{P}(\theta_0)\|_2$ satisfies that

$$\frac{n}{\sigma^2}\frac{\|\tilde{P}(\theta_0)\|_2^2}{\sum_{j\in\mathbb{N}^d}\nu_j} \overset{n\to\infty}{\to} \infty, \tag{3.3.2}$$

then the aggregated and inflated credible set resulting from Method I has frequentist coverage tending to zero, i.e. for any $L > 0$,

$$P_{\theta_0}\left(\theta_0 \in \hat{B}_{n,m,\gamma}\left(L\right)\right) \overset{n\to\infty}{\to} 0.$$

We briefly discuss the assumptions. Condition (3.3.1) requires that the squared bias term is dominated by the posterior variance, which is a natural and standard assumption for coverage. On the other hand condition (3.3.2) resulting in the lack of coverage assumes that the squared bias dominates the variance which is again natural and standard. The assumption $n\delta_n/\sum_{j\in\mathbb{N}^d}\nu_j = o(1)$ is of technical nature, and is required to deal with the tail of the eigen-decomposition of the posterior. This condition is not optimized but it is already sufficiently general to cover our examples. The blow up constant of the credible sets are again of technical nature, it can be equivalently replaced by slightly under-smoothing the priors, see (Knapik et al., 2011).

Below we consider specific choices of the covariance kernel K, both with polynomially and exponentially decaying eigenvalues. We show below that by not over-smoothing the priors, Method I results in frequentist coverage tending to one in both examples.

Corollary 3.3.2. ***(Polynomial)*** *For given $B > 0$ and $\beta \geq 3d/2$, assume that the covariance kernel K satisfies Assumptions 3.1.1 and (3.1.6) with $\alpha \leq \beta$. Then for $m = o(n^{\frac{2\beta-3d}{4\beta}})$ and L_n tending to infinity arbitrarily slowly the aggregated posterior credible set $\hat{B}_{n,m,\gamma}(L_n)$ attains asymptotic frequentist coverage one, i.e.*

$$\inf_{\theta_0\in\Theta^\beta(B)} P_0\left(\theta_0 \in \hat{B}_{n,m,\gamma}\left(L_n\right)\right) \to 1.$$

The proof is given in Section 3.6.3.

Corollary 3.3.3. ***(Exponential)*** *For given $B > 0$ and $\beta \geq d/2$, let us take $m = o(n^{\frac{2\beta-d}{2(2\beta+d)}})$ and assume that the covariance kernel K satisfies Assumptions 3.1.1 and (3.1.7) with $(m/n)^{1/(2d)}(\log n)^{1-1/(2d)} \lesssim a \lesssim \left(\sigma/n\right)^{1/(2\beta+d)}\log n$ and $b = 1$. Then for L_n tending to infinity arbitrarily slowly the aggregated posterior credible set $\hat{B}_{n,m,\gamma}(L_n)$ obtains asymptotic frequentist coverage one, i.e.*

$$\inf_{\theta_0\in\Theta^\beta(B)} P_0\left(\theta_0 \in \hat{B}_{n,m,\gamma}\left(L_n\right)\right) \to 1.$$

The proof is given in Section 3.6.4. We note that in both examples the conditions on β and m are of technical nature and they were not optimized.

§3.4 Discussion

In this chapter, we have shown that distributed methods can be applied in the context of Gaussian Process regression and give accurate results in terms of recovery and uncertainty quantification. Although a naive averaging of the local posteriors will fail to capture the true functional parameters, there exist techniques obtaining a global posterior distribution which has similar asymptotic behavior as the non-distributed posterior distribution. We demonstrate through various examples (including both polynomially and exponentially decaying eigenvalues for the covariance kernel) that the aggregated posterior distribution can achieve optimal minimax contraction rates and good frequentist coverage.

One of the main contributions of our paper is that we do not need to assume that the true functional parameter belongs to the Reproducing Kernel Hilbert Space (RKHS) corresponding to the considered Gaussian Process prior, which is a typical assumption in the literature. This way our results are less restrictive and can be applied for a larger class of functions and priors. For instance squared exponential covariance kernels contain analytic functions in their RKHS, hence assuming that the truth belongs to that space would substantially reduce the applicability of the method. Also, in case of Matérn kernels by relaxing this assumption we do not have to introduce an (artificial) rescaling factor which is needed otherwise as the regularity of the Matérn kernel can't be chosen to match the regularity of the truth.

The optimal choice of the tuning hyper-parameter in the covariance kernel depends on the regularity of the underlying function, which is typically unknown in practice. In the non-distributed setting various adaptive techniques were proposed to solve this problem, including hierarchical and empirical Bayes methods. However, in the distributed setting standard approaches based on the (marginal) likelihood fail, as it was demonstrated in the context of the Gaussian white noise model, see (Szabó and van Zanten, 2019). An open and interesting line of research is to understand whether adaptation is possible at all in the distributed regression framework (3.1) and if yes to provide method achieving it.

§3.5 Proofs of the main results

§3.5.1 Kernel Ridge Regression in non-distributed setting

Let us first consider the non-distributed case, i.e. take $m = 1$. We introduce some notations and recall standard results for the kernel ridge regression method. The posterior mean $\hat{\theta}_n$ coincides with the kernel ridge regression (KRR) estimator

$$\hat{\theta}_n = \hat{\theta}_{KRR} = \arg\min_{\theta \in \mathcal{H}} \left[-\ell_n(\theta)\right], \quad -\ell_n(\theta) := \sum_{i=1}^{n} \left(Y_i - \theta(X_i)\right)^2 + \sigma^2 \|\theta\|_{\mathcal{H}}^2, \quad (3.5.1)$$

where the RKHS $\mathcal{H}$ corresponds to the prior covariance kernel K, see Chapter 6 in (Rasmussen and Williams, 2006). The objective function of the KRR is composed of the average squared-error loss and an RKHS penalty term. In view of the representer

theorem for RKHSs, the solution to (3.5.1) is a linear combination of kernel functions, which renders it equivalent to a quadratic program.

By the reproducing property, all functions θ in the RKHS $\mathcal{H}$ can be evaluated as $\theta(X_i) = \langle \theta, K_{X_i} \rangle_{\mathcal{H}}$ with $K_{X_i} = K(X_i, \cdot)$, and $\|\theta\|_{\mathcal{H}}^2 = \langle \theta, \theta \rangle_{\mathcal{H}}$. The corresponding log-likelihood function takes the form (up to an additive constant term)

$$-\ell_n(\theta) := \sum_{i=1}^{n} \left(Y_i - \langle \theta, K_{X_i} \rangle_{\mathcal{H}}\right)^2 + \sigma^2 \langle \theta, \theta \rangle_{\mathcal{H}}.$$

Performing a Fréchet derivation on $\ell_n : (\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}) \to \mathbb{R}$ with respect to θ, one can obtain the score function. By multiplying the score function with $1/(2n)$ we arrive at the function $\hat{S}_n : \mathcal{H} \to \mathcal{H}$ given as

$$\hat{S}_n(\theta) = \frac{1}{n}\left[\sum_{i=1}^{n} \left(Y_i - \theta(X_i)\right) K_{X_i} - \sigma^2 \theta\right]. \tag{3.5.2}$$

For simplicity we refer to $\hat{S}_n(\theta)$ as the score function from now on and note that the *KRR* estimate $\hat{\theta}_n = \hat{\theta}_{KRR}$ then verifies

$$\hat{S}_n\left(\hat{\theta}_n\right) = 0.$$

Define also $S_n(\theta) := E_0 \hat{S}_n(\theta)$ to be the population version of the score function, i.e.

$$S_n(\theta) = \int_{\mathcal{X}} \left(\theta_0(x) - \theta(x)\right) K_x dx - \frac{\sigma^2}{n}\theta = F(\theta_0 - \theta) - \frac{\sigma^2}{n}\theta, \tag{3.5.3}$$

where the operator $F : L_2(\mathcal{X}) \to \mathcal{H}$ is a convolution with the kernel K, in other words $F(\theta) = \int \theta(x) K_x dx$. Considering $\theta = \sum_{j \in \mathbb{N}^d} \theta_j \psi_j$, a straightforward calculation yields $F(\theta) = \sum_{j \in \mathbb{N}^d} \mu_j \theta_j \psi_j$. We can then rewrite $S_n(\theta)$ as

$$S_n(\theta) = \sum_{j \in \mathbb{N}^d} \left(\mu_j \theta_{0,j} - \frac{\sigma^2 + n\mu_j}{n}\theta_j\right) \psi_j, \tag{3.5.4}$$

which leads immediately to a solution of $S_n(\theta) = 0$ with $\theta_j = \nu_j \theta_{0,j}$, where $\nu_j = \nu_{n,j} = \frac{n\mu_j}{\sigma^2 + n\mu_j}$.

Let us define another operator $\tilde{F} : L_2(\mathcal{X}) \to \tilde{\mathcal{H}}$, with $\tilde{\mathcal{H}}$ denoting the Hilbert space with inner product $\langle \theta, \theta' \rangle_{\tilde{\mathcal{H}}} = \sum_{j \in \mathbb{N}^d} \nu_j^{-2} \theta_j \theta'_j$, as $\tilde{F}(\theta) = \sum_{j \in \mathbb{N}^d} \nu_j \theta_j \psi_j$ (we omit the dependence on n in the notation). Note that both operators F and $\tilde{F}$ are bijective and linear, which allows us to rewrite (3.5.3) as

$$S_n(\theta) = F(\theta_0) - F \circ \tilde{F}^{-1}(\theta) = F\left(\theta_0 - \tilde{F}^{-1}(\theta)\right).$$

Hence, using the notation $\Delta\hat{\theta}_n = \hat{\theta}_n - \tilde{F}(\theta_0)$ we get

$$\Delta\hat{\theta}_n = -\tilde{F} \circ F^{-1} \circ S_n\left(\hat{\theta}_n\right). \tag{3.5.5}$$

It will also be useful to define the operator $\tilde{P} = \mathrm{id} - \tilde{F}$, where id denotes the identity operator on $L_2(\mathcal{X})$. Also note that $S_n(\tilde{F}(\theta_0)) = 0$.

Table 3.1 provides a summary of the key above notations in order to help the reader find a way in the proofs.

Table 3.1: Notation references

Symbol	Definition
$\mathbb{D}_n$	Data, $\{(Y_i, X_i)_{i=1}^n\}$.
θ_0	True function.
ε_i	Gaussian error, $\varepsilon_i = Y_i - \theta_0(X_i) \sim \mathcal{N}(0, \sigma^2)$.
$\hat{\theta}_n$	posterior mean function, $E_X[\theta \mid \mathbb{D}_n]$, equal to the KRR solution.
	$\hat{\theta}_n = \arg\min_{\theta \in \mathcal{H}} \left[n^{-1} \sum_{i=1}^n (Y_i - \theta(X_i))^2 + n^{-1}\sigma^2 \|\theta\|_{\mathcal{H}}^2 \right]$.
F	Convolution with kernel K, $F(\theta) = \sum_{j \in \mathbb{N}^d} \mu_j \theta_j \psi_j$.
F^{-1}	Inverse of F, $F^{-1}(\theta) = \sum_{j \in \mathbb{N}^d} (\theta_j / \mu_j) \psi_j$.
$\{\nu_j\}_{j \in \mathbb{N}^d}$	Eigenvalues of the equivalent kernel $\nu_j = n\mu_j / (\sigma^2 + n\mu_j)$.
$\tilde{F}$	Convolution with the equivalent kernel $\tilde{F}(\theta) = \sum_{j \in \mathbb{N}^d} \nu_j \theta_j \psi_j$.
$\tilde{F}^{-1}$	Inverse of $\tilde{F}$, $\tilde{F}^{-1}(\theta) = \sum_{j \in \mathbb{N}^d} (\theta_j / \nu_j) \psi_j$.
$\tilde{P}$	$\tilde{P} = \mathrm{id} - \tilde{F}$.
$\hat{S}_n$	Sample score function, $\hat{S}_n(\theta) = n^{-1}[\sum_{i=1}^n (Y_i - \theta(X_i)) K_{X_i} - \sigma^2 \theta]$.
S_n	Population score function, $S_n(\theta) = F(\theta_0 - \tilde{F}^{-1}(\theta))$.

§3.5.2 Kernel Ridge Regression in distributed setting

In the distributed setting (both in Methods I and II), accordingly, the kth local sample and population score functions are given (up to constant multipliers) by

$$\hat{S}_n^{(k)}(\theta) = \frac{1}{n/m} \left[\sum_{i=1}^{n/m} \left(Y_i^{(k)} - \theta\left(X_i^{(k)}\right) \right) K_{X_i^{(k)}} - \frac{\sigma^2}{m}\theta \right],$$

$$S_n^{(k)}(\theta) = \int_{\mathcal{X}} \left(\theta_0(x) - \theta(x)\right) K_x dx - \frac{\sigma^2}{n}\theta = S_n(\theta), \tag{3.5.6}$$

respectively. Analogously to (3.5.2), every local KRR estimate satisfies $\hat{S}_n^{(k)}(\hat{\theta}_n^{(k)}) = 0$. In view of $S_n^{(k)} = S_n$ we have $S_n^{(k)}\big(\tilde{F}(\theta_0)\big) = 0$, hence for each machine, let $\Delta\hat{\theta}_n^{(k)} = \hat{\theta}_n^{(k)} - \tilde{F}(\theta_0)$ denote the difference between the empirical and the population minimizer of the KRR.

§3.5.3 Proof of Theorem 3.2.2

In the proof we use ideas from the proof of Theorem 2.1 of (Bhattacharya et al., 2017). The main differences between their and our results are that we are considering (various) distributed Bayesian methods (not just the standard posterior with $m = 1$) and that we extend the results to general Gaussian process priors (including kernel with polynomially decaying and exponentially decaying eigenvalues), while the proof (Bhattacharya et al., 2017) only covered the rescaled version of the kernel with polynomially decaying eigenvalues, with scaling factor depending on the sample size. More specifically we do not require that the true function belongs to the RKHS of

the GP prior, which substantially extends the applicability of our results. Finally in our analysis we consider the multivariate d-dimensional case, work with L_2-norm and consider Sobolev type of regularity classes rather than L_∞ norm and hyper-rectangles induced by the series decomposition with respect to the eigenbasis $\boldsymbol{\psi}_j$. These extensions and conceptual differences required substantially different proof techniques than in (Bhattacharya et al., 2017).

First note that in view of the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get

$$E_0 \left\| \hat{\theta}_{n,m} - \theta_0 \right\|_2^2 \leq 2 \left\| \theta_0 - \tilde{F}(\theta_0) \right\|_2^2 + 2E_0 \left\| \hat{\theta}_{n,m} - \tilde{F}(\theta_0) \right\|_2^2,$$

where $\hat{\theta}_{n,m}$ is the mean of the global posterior $\Pi^{\dagger}_{n,m}(.|\mathbb{D}_n)$ obtained with either method I or II. Then we show in Section 3.5.3.1 that for $\theta_0 \in \Theta^{\beta}(B)$

$$E_0 \left\| \hat{\theta}_{n,m} - \tilde{F}\left(\theta_0\right) \right\|_2^2 \lesssim \left(\frac{1}{n} \sum_{j \in \mathbb{N}^d} \nu_j^2 \right) \left(\|\tilde{P}(\theta_0)\|_2^2 + \sigma^2 \right) + \delta_n, \tag{3.5.7}$$

where

$$\delta_n = \inf \left\{ n \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell : \mathcal{I} \subset \mathbb{N}^d, |\mathcal{I}| \leq \frac{n}{m^2} \left(\sum_{j \in \mathbb{N}^d} \nu_j^2 \right)^{-1} \right\}$$

concluding the proof of the first statement.

For the contraction rate note that by using Markov's and triangle inequalities we get

$$E_0 \Pi^{\dagger}_{n,m} \left(\theta : \|\theta - \theta_0\|_2 \geq M_n \varepsilon_n | \mathbb{D}_n \right) \leq 2 \frac{E_0 E^{\dagger}_{n,m} \left[\left\| \theta - \hat{\theta}_{n,m} \right\|_2^2 | \mathbb{D}_n \right] + E_0 \left\| \hat{\theta}_{n,m} - \theta_0 \right\|_2^2}{M_n^2 \varepsilon_n^2}.$$

Therefore it is sufficient to show that

$$E_0 E^{\dagger}_{n,m} \left[\left\| \theta - \hat{\theta}_{n,m} \right\|_2^2 | \mathbb{D}_n \right] = O\left(\frac{\sigma^2}{n} \sum_j \nu_j \right).$$

In view of Fubini's theorem the expected squared L_2-norm of the process $\theta - \hat{\theta}_{n,m} | \mathbb{D}_n$ is the integral of the aggregated posterior variance of $\theta(x)$ over $\mathcal{X}$,

$$E^{\dagger}_{n,m} \left[\left\| \theta - \hat{\theta}_{n,m} \right\|_2^2 | \mathbb{D}_n \right] = \int_{\mathcal{X}} \mathrm{Var}^{\dagger}_{n,m} \left(\theta \left(x \right) | \mathbb{D}_n \right) dx.$$

In the non-distributed setting, the posterior variance only depends on the design matrix $\mathbb{X}$. The expectation of this integral is known as the *learning curve* in Chapter 7 of (Rasmussen and Williams, 2006). In Section 3.5.3.2 we prove that

$$E_0 \int_{\mathcal{X}} \mathrm{Var}^{\dagger}_{n,m} \left(\theta \left(x \right) | \mathbb{D}_n \right) dx \asymp \sigma^2 \sum_{j \in \mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j} = \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j, \tag{3.5.8}$$

concluding the proof of the statement.

3.5.3.1 Proof of (3.5.7)

First note, that in view of the inequality $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2^2 \leq 2\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2 + 2\left\|\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2.$$

Then we show below that

$$E_0\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2 \lesssim \frac{1}{m}E_0\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2^2 + \delta_n, \tag{3.5.9}$$

which together with the preceding display implies

$$E_0\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2^2 \leq (2+o(1))\left(E_0\left\|\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2 + C\delta_n\right).$$

By combining the preceding two displays we arrive at

$$\begin{aligned}E_0&\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2\\ &\lesssim \frac{1}{m}E_0\left\|\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2 + \delta_n.\end{aligned}$$

For the aggregated estimator we get that

$$\begin{aligned}\left\|\Delta\hat{\theta}_{n,m}\right\|_2^2 \lesssim{}& \left\|\Delta\hat{\theta}_{n,m} - \frac{1}{m}\sum_{k=1}^m\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2\\ &+ \left\|\frac{1}{m}\sum_{k=1}^m\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2.\end{aligned}$$

Then in view of the preceding display, the independence of the data across machines and $E_0\big(\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\big) = 0$ we get that

$$E_0\left\|\Delta\hat{\theta}_{n,m}\right\|_2^2 \lesssim \frac{1}{m}E_0\left\|\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2 + \delta_n.$$

Finally we verify below that

$$E_0\left\|\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right\|_2^2 \lesssim \left(\frac{1}{n/m}\sum_{j\in\mathbb{N}^d}\nu_j^2\right)\left(\left\|\tilde{P}\left(\theta_0\right)\right\|_2^2 + \sigma^2\right), \tag{3.5.10}$$

which together with $\|\tilde{P}(\theta_0)\|_2^2 \leq \|\theta_0\|_2^2 \leq B^2$ provides us (3.5.7).

Proof of (3.5.9)**:** First note that the identity $\Delta\hat{\theta}_n^{(k)} = -\tilde{F}\circ F^{-1}\circ S_n^{(k)}(\hat{\theta}_n^{(k)})$ follows from assertions (3.5.5) and (3.5.6). This implies together with the properties of $\hat{S}_n^{(k)}$ and $S_n^{(k)}$, that

$$\begin{aligned}\left(\hat{S}_n^{(k)}\left(\hat{\theta}_n^{(k)}\right) - S_n^{(k)}\left(\hat{\theta}_n^{(k)}\right)\right) - \left(\hat{S}_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right) - S_n^{(k)}\left(\tilde{F}\left(\theta_0\right)\right)\right)&\\ = F\circ\tilde{F}^{-1}\left(\Delta\hat{\theta}_n^{(k)}\right) - \hat{S}_n^{(k)}\left(\tilde{F}(\theta_0)\right).&\end{aligned} \tag{3.5.11}$$

On the other hand, in view of (3.5.6),

$$\hat{S}_n^{(k)}(\theta) - S_n^{(k)}(\theta) = \frac{1}{n/m}\sum_{i=1}^{n/m}\left(Y_i^{(k)} - \theta\left(X_i^{(k)}\right)\right)K_{X_i^{(k)}} - \int_{\mathcal{X}}(\theta_0(x) - \theta(x))\,K_x dx$$

for all functions $\theta \in \mathcal{H}$. Therefore, by applying the preceding display twice with $\theta = \hat{\theta}_n^{(k)}$ and $\theta = \tilde{F}(\theta_0)$, we get that

$$\begin{aligned}\left(\hat{S}_n^{(k)}\left(\hat{\theta}_n^{(k)}\right) - S_n^{(k)}\left(\hat{\theta}_n^{(k)}\right)\right) - \left(\hat{S}_n^{(k)}\left(\tilde{F}(\theta_0)\right) - S_n^{(k)}\left(\tilde{F}(\theta_0)\right)\right)\\ = -\frac{1}{n/m}\sum_{i=1}^{n/m}\Delta\hat{\theta}_n^{(k)}(X_i^{(k)})K_{X_i^{(k)}} + \int_{\mathcal{X}}\Delta\hat{\theta}_n^{(k)}(x)K_x dx.\end{aligned}$$

Combining assertion (3.5.11) with the preceding display and then using Lemma 3.7.2 (with $\hat{\vartheta} = \Delta\hat{\theta}_n^{(k)}$, satisfying the boundedness assumption, see Lemma 3.7.9) together with Lemma 3.7.7, we get for arbitrary index set $\mathcal{I} \subset \mathbb{N}^d$ that

$$\begin{aligned}&E_0\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ S_n^{(k)}\left(\tilde{F}(\theta_0)\right)\right\|_2^2\\ &\quad= E_0\left\|\left(\tilde{F}\circ F^{-1}\right)\left(\frac{1}{n/m}\sum_{i=1}^{n/m}\Delta\hat{\theta}_n^{(k)}\left(X_i^{(k)}\right)K_{X_i^{(k)}} - \int_{\mathcal{X}}\Delta\hat{\theta}_n^{(k)}(x)\,K_x dx\right)\right\|_2^2\\ &\quad\lesssim \frac{|\mathcal{I}|\log n}{n/m}\sum_{j\in\mathbb{N}^d}\nu_j^2 E_0\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2^2 + n\sum_{j\in\mathbb{N}^d}\nu_j^2\sum_{\ell\in\mathcal{I}^c}\mu_\ell.\end{aligned}$$

Taking the minimum over $|\mathcal{I}| \le \frac{n}{m^2\log n}(\sum_{j\in\mathbb{N}^d}\nu_j^2)^{-1}$, we get that

$$E_0\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ S_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 \lesssim \frac{1}{m}E_0\|\Delta\hat{\theta}_n^{(k)}\|_2^2 + \delta_n \tag{3.5.12}$$

concluding the proof of (3.5.9).

Proof of (3.5.10). In view of the linearity of the operator $\tilde{F}\circ F^{-1}$, the inequality $\|\theta_1+\theta_2\|_2^2 \le 2\|\theta_1\|_2^2 + 2\|\theta_2\|_2^2$, and

$$\begin{aligned}\hat{S}_n^{(k)}\left(\tilde{F}(\theta_0)\right) &= \frac{1}{n/m}\sum_{i=1}^{n/m}\left(Y_i^{(k)} - \theta_0\left(X_i^{(k)}\right)\right)K_{X_i^{(k)}}\\ &\quad+ \frac{1}{n/m}\sum_{i=1}^{n/m}\tilde{P}(\theta_0)\left(X_i^{(k)}\right)K_{X_i^{(k)}} - \frac{\sigma^2}{n}\tilde{F}(\theta_0),\end{aligned}$$

the left hand side of (3.5.10) can be bounded from above as

$$
\begin{aligned}
&E_0 \left\| \tilde{F} \circ F^{-1} \left(\hat{S}_n^{(k)} \left(\tilde{F} \left(\theta_0 \right) \right) - S_n^{(k)} \left(\tilde{F} \left(\theta_0 \right) \right) \right) \right\|_2^2 \\
&\quad \leq 2E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \tilde{P} \left(\theta_0 \right) \left(X_i^{(k)} \right) K_{X_i^{(k)}} - \int_{\mathcal{X}} \tilde{P} \left(\theta_0 \right) (x) \, K_x dx \right) \right\|_2^2 \\
&\qquad + 2E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \varepsilon_i^{(k)} K_{X_i^{(k)}} \right) \right\|_2^2 \\
&\quad =: (T_1 + T_2).
\end{aligned}
$$

We deal with terms T_1 and T_2 separately. In view of Lemma 3.7.1 (with $\vartheta = \tilde{P}(\theta_0)$)

$$
T_1 \leq \frac{2C}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 \left\| \tilde{P} \left(\theta_0 \right) \right\|_2^2,
$$

for some $C > 0$. Since the operator $\tilde{F} \circ F^{-1}$ is linear, we get that

$$
\begin{aligned}
T_2 &= \frac{2}{(n/m)^2} \sum_{i=1}^{n/m} E_0 \left(\left(\varepsilon_i^{(k)} \right)^2 \left\| \tilde{F} \circ F^{-1} \left(K_{X_i^{(k)}} \right) \right\|_2^2 \right) \\
&\quad + \frac{4}{(n/m)^2} \sum_{1 \leq i < \ell \leq n} E_0 \left(\varepsilon_i^{(k)} \varepsilon_\ell^{(k)} \tilde{F} \circ F^{-1} \left(\left\langle K_{X_i^{(k)}}, K_{X_\ell^{(k)}} \right\rangle_2 \right) \right) \\
&= \frac{2\sigma^2}{n/m} E_0 \left\| \tilde{F} \circ F^{-1} \left(K_{X_1^{(k)}} \right) \right\|_2^2 = \frac{2\sigma^2}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2,
\end{aligned}
$$

because the cross terms are equal to 0 due to independence of the noise $\varepsilon_i^{(k)}$, $i = 1, ..., n/m$, $k = 1, ..., m$.

3.5.3.2 Proof of (3.5.8)

In this section we give upper bounds for the learning curves in case of both distributed methods.

Method I: Let us denote by $\mu_j^I = m\mu_j$ the eigenvalues of the local covariance kernel. Then in view of Lemma 3.7.4, the expectations of the m local posterior variances are all of the same order

$$
E_0 E_X \operatorname{Var} \left(\theta(X) | \mathbb{D}_n^{(k)} \right) \asymp \sigma^2 \sum_{j \in \mathbb{N}^d} \frac{\mu_j^I}{\sigma^2 + (n/m)\mu_j^I} = \sigma^2 \sum_{j \in \mathbb{N}^d} \frac{m\mu_j}{\sigma^2 + n\mu_j} = \frac{\sigma^2}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j.
$$

Since the variance of the global posterior distribution $\Pi_{n,m}^I(.|\mathbb{D}_n)$ satisfies the following equality

$$
\operatorname{Var}_{n,m}^I \left(\theta(x) \right) = m^{-2} \sum_{k=1}^{m} \operatorname{Var} \left(\theta(x) | \mathbb{D}_n^{(k)} \right),
$$

one can see that

$$E_0 E_X \operatorname{Var}^I_{n,m}(\theta(X)) \asymp \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j.$$

Method II: First note that $\mu_j^{II} = \mu_j$ the eigenvalues of the local covariance kernel. Note that the expectations of the m local posterior variances are all of the same order

$$E_0 E_X \operatorname{Var}\left(\theta(X)|\mathbb{D}_n^{(k)}\right) \asymp \frac{\sigma^2}{m} \sum_{j \in \mathbb{N}^d} \frac{\mu_j^{II}}{\sigma^2/m + (n/m)\mu_j^{II}} = \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j,$$

because the variance of the noise is σ^2/m for each machine. The variance of the aggregated posterior distribution $\Pi^{II}_{n,m}(.|\mathbb{D}_n)$ satisfies

$$E_0 E_X \operatorname{Var}^{II}_{n,m}(\theta(X)|\mathbb{D}_n) \asymp \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j$$

because we know that

$$\operatorname{Var}^{II}_{n,m}(\theta(X)|\mathbb{D}_n) = m^{-1} \sum_{k=1}^{m} \operatorname{Var}\left(\theta(x)|\mathbb{D}_n^{(k)}\right)$$

proving assertion (3.5.8).

§3.5.4 Proof of Theorem 3.2.1

The proof follows similar lines of reasoning as Theorem 3.2.2, where we provided general upper bounds for the contraction rate of the distributed posterior.

First we prove (3.2.3). For the naive averaging method the local sample and population score functions coincide to the non-distributed case given in Section 3.5.1 with sample size n/m, i.e.

$$\hat{S}_n^{*(k)}(\theta) = \frac{1}{n/m}\left[\sum_{i=1}^{n/m}\left(Y_i^{(k)} - \theta\left(X_i^{(k)}\right)\right) K_{X_i^{(k)}} - \sigma^2\theta\right],$$

$$S_n^{*(k)}(\theta) = \int_{\mathcal{X}} (\theta_0(x) - \theta(x))\, K_x dx - \frac{\sigma^2}{n/m}\theta = F(\theta_0 - \theta) - \frac{\sigma^2}{n/m}\theta.$$

Note that the solution of the equation $S_n^{*(k)}(\theta) = 0$ is given by the coefficients $\theta_j = \nu_j^* \theta_{0,j}$, with $\nu_j^* = \frac{n\mu_j}{m\sigma^2 + n\mu_j}$, $j \in \mathbb{N}^d$.

Then using the inequality $a^2 \geq (a-b)^2/2 - b^2$ one can obtain that

$$E_0 \left\|\hat{\theta}^*_{n,m} - \theta_0\right\|_2^2 \geq \frac{1}{2}\left\|\theta_0 - \tilde{F}^*(\theta_0)\right\|_2^2 - E_0 \left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2^2,$$

where $\tilde{F}^*(\theta) = \sum_{j \in \mathbb{N}} \nu_j^* \theta_j \psi_j$ and $\hat{\theta}^*_{n,m}$ is the mean of the global posterior $\Pi^{\dagger}_{n,m}(.|\mathbb{D}_n)$ obtained with the naive averaging method.

First note that

$$\left\|\theta_0 - \tilde{F}^*(\theta_0)\right\|_2^2 = \sum_{j=1}^{\infty} \frac{m}{m\sigma^2 + n\mu_j}\theta_{0,j}^2 \geq \frac{c_L}{2} \sum_{(n/(m\sigma^2))^{1/(2\beta+1)} \leq j} j^{-1-2\beta}(\log j)^{-2}$$
$$\geq c_0(n/m)^{-2\beta/(2\beta+1)}(\log(n/m))^{-2}, \tag{3.5.13}$$

for some small enough $c_0 > 0$. We conclude the proof of (3.2.3) by showing below that $E_0\|\hat{\theta}_{n,m} - \tilde{F}(\theta_0)\|_2^2 = o((n/m)^{-2\beta/(2\beta+1)}(\log(n/m))^{-2})$.

Similarly to (3.5.7) we can derive (by replacing $\tilde{F}$ and ν with $\tilde{F}^*$ and ν^*, respectively) that

$$E_0\left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2^2 \lesssim \left(\frac{1}{n}\sum_{j=1}^{\infty}(\nu_j^*)^2\right)\left(\left\|\tilde{P}^*(\theta_0)\right\|_2^2 + \sigma^2\right) + \delta_n^*, \tag{3.5.14}$$

where $\delta_n^* = n\sum_{j=1}^{\infty}(\nu_j^*)^2\sum_{\ell=I}^{\infty}\mu_\ell$, with $I = \frac{n}{m^2\log n}(\sum_{j=1}^{\infty}(\nu_j^*)^2)^{-1}$. Note that $\|\tilde{P}^*(\theta_0)\|_2^2 = O(1)$ and in view of Lemma 3.7.5, $\sum_{j=1}^{\infty}(\nu_j^*)^2 \asymp (n/m)^{1/(2\beta+1)}$; hence

$$I \asymp \frac{(n/m)^{2\beta/(2\beta+1)}}{m\log n}.$$

Therefore the first term on the right hand side of (3.5.14) is $O(n^{-2\beta/(2\beta+1)}m^{-1/(2\beta+1)})$ and

$$\delta_n^* \lesssim n(n/m)^{1/(1+2\beta)}I^{-2\beta} \asymp n^{2-2\beta}m^{-1+4\beta}(\log n)^{2\beta} = o\left((\log(n/m))^{-2}\right),$$

where the last step holds for large enough choice of β and not too large choice of m. For instance taking $\beta > 2$ and $m = o(n^{1/(2+2\beta)})$, we get that

$$\delta_n^*(n/m)^{2\beta/(2\beta+1)} \lesssim n^{-c_0}\log^4 n = o\left((\log(n/m))^{-2}\right),$$

for some $c_1 > 0$.

It remained to deal with (3.2.4). First note that by the computations above combined with Markov's inequality there exists a sequence $\rho_n \to 0$ such that

$$P_0\left(\left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2 \geq \rho_n(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1}\right) \to 0.$$

Then by triangle inequality, (3.5.13) and Markov's inequality we get for $c < c_0$ that

$$E_0\Pi^*_{n,m}\left(\theta : \|\theta - \theta_0\|_2 \leq c(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1}|\mathbb{D}_n\right)$$
$$\leq E_0\Pi^*_{n,m}\Big(\left\|\theta_0 - \tilde{F}^*(\theta_0)\right\|_2 - \left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2$$
$$- c(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1} \leq \left\|\theta - \hat{\theta}^*_{n,m}\right\|_2|\mathbb{D}_n\Big)$$
$$\leq E_0\Pi^*_{n,m}\left((c_0 - c - \rho_n)(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1} \leq \left\|\theta - \hat{\theta}^*_{n,m}\right\|_2|\mathbb{D}_n\right) + o(1)$$
$$\lesssim (n/m)^{2\beta/(2\beta+1)}(\log(n/m))^2E_0E^*_{n,m}\left\|\theta - \hat{\theta}^*_{n,m}\right\|_2^2.$$

We conclude the proof by noting that

$$E_0 E_X \operatorname{Var}\left(\theta(X)|\mathbb{D}_n^{(k)}\right) = \sigma^2 \sum_{j=1}^{\infty} \frac{\mu_j}{\sigma^2 + (n/m)\mu_j} = \frac{\sigma^2}{n/m} \sum_{j=1}^{\infty} \nu_j^*,$$

for all $k \in \{1, ..., m\}$; hence

$$E_0 E_{n,m}^* \left\|\theta - \hat{\theta}_{n,m}^*\right\|_2^2 = \frac{1}{m^2} \sum_{k=1}^{m} E_0 E_X \operatorname{Var}\left(\theta(X)|\mathbb{D}_n^{(k)}\right) = \frac{\sigma^2}{n} \sum_{j=1}^{\infty} \nu_j^* \lesssim \frac{\sigma^2}{m} (n/m)^{-2\beta/(2\beta+1)}.$$

§3.5.5 Proof of Theorem 3.3.1

We first consider the non-distributed case $m = 1$ for clearer presentation and then extend our results to the distributed setting.

3.5.5.1 Non-distributed setting

Connection to KRR Similarly to the posterior mean, the posterior covariance function $\hat{C}_n$ can be given as

$$\hat{C}_n(x, x') = K(x, x') - \hat{K}_n(x, x'),$$

where $\hat{K}_n(x, \cdot) = K(\cdot, \mathbb{X})[K(\mathbb{X}, \mathbb{X}) + \sigma^2 I_n]^{-1} K(\mathbb{X}, x)$, or equivalently

$$\hat{K}_{x,n} = \hat{K}_n(x, \cdot) = \arg\min_{\vartheta \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(K(x, X_i) - \vartheta(X_i)\right)^2 + \frac{\sigma^2}{n} \|\vartheta\|_{\mathcal{H}}^2 \right], \tag{3.5.15}$$

see assertion (8) of (Bhattacharya et al., 2017).

Then by taking the Fréchet derivative of the expression on the right hand side we arrive to the (adjusted) score function and its expected value

$$\hat{S}_{K_x,n}(\vartheta) = \frac{1}{n} \left(\sum_{i=1}^{n} \left(K_x(X_i) - \vartheta(X_i)\right) K_{X_i} - \sigma^2 \vartheta \right),$$
$$S_{K_x,n}(\vartheta) = E\hat{S}_{K_x,n}(\vartheta) = \int_{\mathcal{X}} \left(K_x(z) - \vartheta(z)\right) K_z dz - \frac{\sigma^2}{n} \vartheta.$$

Then similarly to the posterior mean in Section 3.5.1 the following assertions hold

$$S_{K_x,n}(\vartheta) = F(K_x) - F \circ \tilde{F}^{-1}(\vartheta) = F\left(K_x - \tilde{F}^{-1}(\vartheta)\right), \tag{3.5.16}$$

$$\Delta\hat{K}_{x,n} = \hat{K}_{x,n} - \tilde{F}(K_x) = -\tilde{F} \circ F^{-1} \circ S_{K_x,n}\left(\hat{K}_{x,n}\right), \tag{3.5.17}$$

$$\hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \tilde{P}(K_x)(X_i)K_{X_i} - \sigma^2\tilde{F}(K_x)\right), \tag{3.5.18}$$

$$\begin{aligned} F \circ \tilde{F}^{-1}\left(\Delta\hat{K}_{x,n}\right) - \hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right) \\ = -\frac{1}{n}\sum_{i=1}^{n} \Delta\hat{K}_{x,n}(X_i)K_{X_i} + \int_{\mathcal{X}} \Delta\hat{K}_{x,n}(x')K_{x'}dx', \end{aligned} \tag{3.5.19}$$

and note that $\hat{K}_{x,n}$ and $\tilde{F}(K_x)$ are the zero points of the functions $\hat{S}_{K_x,n}$ and $S_{K_x,n}$, respectively.

Under-smoothing Following from the triangle inequality, to obtain frequentist coverage for the credible ball it is sufficient to show that for $L_n \to \infty$

$$P_0\left(\left\|\tilde{P}(\theta_0)\right\|_2 + \left\|\hat{\theta}_n - \tilde{F}(\theta_0)\right\|_2 \leq L_n r_{n,\gamma}\right) \to 1.$$

The preceding display is implied by assumption (3.3.1) and assertions

$$P_0\left(\left\|\Delta\hat{\theta}_n\right\|_2^2 \leq L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \to 1, \tag{3.5.20}$$

$$P_0\left(r_{n,\gamma}^2 \geq \frac{1}{2C_\psi^2}\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \to 1, \tag{3.5.21}$$

where $\Delta\hat{\theta}_n := \hat{\theta}_n - \tilde{F}(\theta_0)$, verified below.

Proof of (3.5.20)**:** In view of assertion (3.5.7) with $m = 1$ and Markov's inequality we get

$$\begin{aligned} P_0\left(\left\|\Delta\hat{\theta}_n\right\|_2^2 \leq L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) &\leq \frac{E_0\left\|\Delta\hat{\theta}_n\right\|_2^2}{L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j} \\ &\lesssim \frac{\left(\frac{1}{n}\sum_{j\in\mathbb{N}^d}\nu_j^2\right) + \delta_n}{L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j} \\ &= O\left(\frac{1}{L_n} + \frac{n\delta_n}{\sum_{j\in\mathbb{N}^d}\nu_j}\right) = o(1). \end{aligned}$$

Proof of (3.5.21)**:** The radius $r_{n,\gamma}$ is defined, conditionally on $\mathbb{X}$, as $P(\|W_n\|_2^2 \leq r_{n,\gamma}^2|\mathbb{X}) = 1-\gamma$, where W_n is a centered GP with covariance kernel $\hat{C}_n$ given in (3.1.3).

In view of Chebyshev's inequality

$$r_{n,\gamma}^2 \geq E\left[\|W_n\|_2^2|\mathbb{X}\right] - (1-\gamma)^{-1/2}\operatorname{Var}\left(\|W_n\|_2^2|\mathbb{X}\right)^{1/2}.$$

Using Fubini's theorem, the first term on the right hand side of the preceding display can be rewritten as

$$E\left[\|W_n\|_2^2|\mathbb{X}\right] = E\left[\|\theta - \hat{\theta}_n\|_2^2|\mathbb{D}_n\right] = \int_{\mathcal{X}} \operatorname{Var}\left(\theta(x)|\mathbb{D}_n\right) dx.$$

The integral on the right-hand side of the display, called the *generalization error*, see Chapter 7 of (Rasmussen and Williams, 2006), is asymptotically bounded from below almost surely by

$$\sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j \sup_{x\in\mathcal{X}} \psi_j^2(x)} \geq \frac{\sigma^2 C_\psi^{-2}}{n} \sum_{j\in\mathbb{N}^d} \nu_j, \tag{3.5.22}$$

in view of assertion (12) of (Opper and Vivarelli, 1999) and Assumption 3.1.1. Furthermore, the variance of $\|W_n\|_2^2$, conditional on the design $\mathbb{X}$, is

$$\operatorname{Var}\left(\|W_n\|_2^2|\mathbb{X}\right) = E\left[\|W_n\|_2^4|\mathbb{X}\right] - E^2\left[\|W_n\|_2^2|\mathbb{X}\right].$$

The first term on the right hand-side satisfies

$$\begin{aligned}
E\left[\|W_n\|_2^4|\mathbb{X}\right] &= E\left[\|\theta - \hat{\theta}_n\|_2^4|\mathbb{D}_n\right] \qquad (3.5.23)\\
&= \int \left(\int_{\mathcal{X}} \left(\theta(x) - \hat{\theta}_n(x)\right)^2 dx \int_{\mathcal{X}} \left(\theta(x') - \hat{\theta}_n(x')\right)^2 dx'\right) \Pi(d\theta|\mathbb{D}_n)\\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int \left(\theta(x) - \hat{\theta}_n(x)\right)^2 \left(\theta(x') - \hat{\theta}_n(x')\right)^2 \Pi(d\theta|\mathbb{D}_n) dx dx'\\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \left(\operatorname{Var}\left(\theta(x)|\mathbb{D}_n\right)\operatorname{Var}\left(\theta(x')|\mathbb{D}_n\right) + 2\hat{C}_n(x,x')^2\right) dx' dx\\
&= \left(\int_{\mathcal{X}} \operatorname{Var}\left(\theta(x)|\mathbb{D}_n\right) dx\right)^2 + 2\int_{\mathcal{X}} \left\|\hat{C}_n(x,.)\right\|_2^2 dx\\
&= E^2\left[\|W_n\|_2^2|\mathbb{X}\right] + 2\int_{\mathcal{X}} \left\|\hat{C}_n(x,.)\right\|_2^2 dx, \qquad (3.5.24)
\end{aligned}$$

using Fubini's theorem and the reduction formula $EX_1^2X_2^2 = Var(X_1)Var(X_2) + 2Cov(X_1,X_2)^2$ for X_1, X_2 centered Gaussian random variables, see for instance page 189 of (Isserlis, 1916). Hence, again in view of Fubini's theorem,

$$E_0 \operatorname{Var}\left(\|W_n\|_2^2|\mathbb{X}\right) = 2\int_{\mathcal{X}} E_0 \left\|\hat{C}_n(x,.)\right\|_2^2 dx. \tag{3.5.25}$$

Recall that the covariance function $\hat{C}_n(x,x') = K(x,x') - \hat{K}_n(x,x')$, where $\hat{K}_{x,n} = \hat{K}_n(x,.)$ is the solution to (3.5.15). We show below that for all $x \in \mathcal{X}$

$$E_0 \left\|\hat{C}_n(x,.)\right\|_2^2 \lesssim \left\|\tilde{P}(K_x)\right\|_2^2 + \tilde{\delta}_n, \tag{3.5.26}$$

for

$$\tilde{\delta}_n = \inf\left\{\left(\sum_{j\in\mathbb{N}^d}\nu_j^2\right)^2 \sum_{\ell\in\mathcal{I}^c}\mu_\ell : \mathcal{I}\subset\mathbb{N}^d, |\mathcal{I}| = o\left(n\left(\sum_{j\in\mathbb{N}^d}\nu_j^2\right)^{-1}\right)\right\}.$$

In view of the definition of the linear operator $\tilde{P}$ and the eigenvalues ν_j, μ_j we get

$$\tilde{P}\left(K(x,x')\right) = \sum_{j\in\mathbb{N}^d}(1-\nu_j)\mu_j\boldsymbol{\psi}_j(x)\boldsymbol{\psi}_j(x') = \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\boldsymbol{\psi}_j(x)\boldsymbol{\psi}_j(x'), \tag{3.5.27}$$

for all $x, x' \in \mathcal{X}$. Then by combining the last three displays

$$\begin{aligned}
E_0\,\mathrm{Var}\left(\|W_n\|_2^2|\mathbb{X}\right) &= 2\int_{\mathcal{X}} E_0\left\|\hat{C}_n(x,.)\right\|_2^2 dx \\
&\lesssim \int_{\mathcal{X}}\left\|\tilde{P}(K(x,.))\right\|_2^2 dx + \tilde{\delta}_n \\
&= \left(\frac{\sigma^2}{n}\right)^2\int_{\mathcal{X}}\sum_{j\in\mathbb{N}^d}\nu_j^2\boldsymbol{\psi}_j(x)^2dx + \delta_n\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{n} \\
&= \left(\frac{\sigma^2}{n}\right)^2\sum_{j\in\mathbb{N}^d}\nu_j^2 + \delta_n\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{n}.
\end{aligned} \tag{3.5.28}$$

Therefore, by Markov's inequality and Lemmas 3.7.5 and 3.7.6,

$$P_0\left(\mathrm{Var}\left(\|W_n\|_2^2|\mathbb{X}\right)^{1/2} \geq t\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \lesssim t^{-2}\left(\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{(\sum_{j\in\mathbb{N}^d}\nu_j)^2} + \frac{n\delta_n\sum_{j\in\mathbb{N}^d}\nu_j^2}{(\sum_{j\in\mathbb{N}^d}\nu_j)^2}\right) \to 0$$

for all $t > 0$. Hence by combining (3.5.22) and the preceding display (with $t = (1-\gamma)^{1/2}C_\psi^{-2}/2$),

$$P_0\left(E\left[\|W_n\|_2^2|\mathbb{X}\right] - (1-\gamma)^{-1/2}\,\mathrm{Var}\left(\|W_n\|_2^2|\mathbb{X}\right)^{1/2} \geq (C_\psi^{-2}/2)\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \to 1.$$

This implies that all the quantiles of $\|W_n\|_2^2$, conditionally on $\mathbb{X}$, are of the order $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$ with P_0-probability going to one, including $r_{n,\gamma}^2$.

Proof of (3.5.26)**:** First note that by the inequality $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\left\|\hat{C}_n(x,.)\right\|_2^2 \leq 2\left\|\tilde{P}(K_x)\right\|_2^2 + 2\left\|\Delta\hat{K}_{x,n}\right\|_2^2,$$

where $\Delta\hat{K}_{x,n} = \hat{K}_{x,n} - \tilde{F}(K_x)$.

Next we give an upper bound for the second term of the preceding display similarly to Section 3.5.3.1. First note that

$$\left\|\Delta\hat{K}_{x,n}\right\|_2^2 \lesssim \left\|\Delta\hat{K}_{x,n} - \tilde{F}\circ F^{-1}\circ\hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2 + \left\|\tilde{F}\circ F^{-1}\circ\hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2.$$

Then by showing below that

$$E_0 \left\| \Delta\hat{K}_{x,n} - \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n}(\tilde{F}(K_x)) \right\|_2^2 \leq o\left(E_0 \left\| \Delta\hat{K}_{x,n} \right\|_2^2 \right) + \tilde{\delta}_n, \qquad (3.5.29)$$

we arrive at

$$E_0 \left\| \Delta\hat{K}_{x,n} \right\|_2^2 \lesssim E_0 \left\| \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 + \tilde{\delta}_n.$$

Next, in view of (3.5.18),

$$\begin{aligned} E_0 \left\| \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 &= E_0 \left\| \tilde{F} \circ F^{-1} \left(\hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) - S_{K_x,n} \left(\tilde{F}(K_x) \right) \right) \right\|_2^2 \\ &= E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n} \sum_{i=1}^n \tilde{P}(K_x)(X_i) K_{X_i} - \int_{\mathcal{X}} \tilde{P}(K_x)(x') K_{x'} dx' \right) \right\|_2^2 \\ &\leq \left(\frac{1}{n} \sum_{j \in \mathbb{N}^d} \nu_j^2 \right) \left\| \tilde{P}(K_x) \right\|_2^2 = o\left(\left\| \tilde{P}(K_x) \right\|_2^2 \right), \end{aligned}$$

where the last line follows from Lemma 3.7.1 with $\vartheta = \tilde{P}(K_x)$ (and $m = 1$), concluding the proof of (3.5.26).

Proof of (3.5.29)**:** Similarly to (3.5.12), by using assertion (3.5.19), Lemma 3.7.2 (with $\hat{\vartheta} = \Delta\hat{K}_{x,n}$, sample size n) and Lemma 3.7.3 (with $m = 1$), we can show that for all $x \in \mathcal{X}$

$$\begin{aligned} &E_0 \left\| \Delta\hat{K}_{x,n} - \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 \\ &\quad = E_0 \left\| (\tilde{F} \circ F^{-1}) \left(\frac{1}{n} \sum_{i=1}^n \Delta\hat{K}_{x,n}(X_i) K_{X_i} - \int_{\mathcal{X}} \Delta\hat{K}_{x,n}(x') K_{x'} dx' \right) \right\|_2^2 \\ &\quad \lesssim \frac{|\mathcal{I}| \log n \sum_{j \in \mathbb{N}^d} \nu_j^2}{n} E_0 \left\| \Delta\hat{K}_{x,n} \right\|_2^2 + \left(\sum_{j \in \mathbb{N}^d} \nu_j^2 \right)^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell. \end{aligned}$$

Taking the infimum over $|\mathcal{I}| = o\big(n/(\log n \sum_{j \in \mathbb{N}^d} \nu_j^2)\big)$, we get that the left hand-side of the preceding display is bounded from above by $o(E_0 \|\Delta\hat{K}_{x,n}\|_2^2) + \tilde{\delta}_n$, concluding the proof of the statement.

Over-smoothing By the definition of credible sets and using the triangle inequality, we get that

$$\begin{aligned} P_0 \left(\theta_0 \in \hat{B}_{n,\gamma}(L) \right) &\leq P_0 \left(\left\| \tilde{P}(\theta_0) \right\|_2 \leq \left\| \hat{\theta}_n - \tilde{F}(\theta_0) \right\|_2 + L r_{n,\gamma} \right) \\ &\leq P_0 \left(\left\| \tilde{P}(\theta_0) \right\|_2 \leq 2 \left\| \hat{\theta}_n - \tilde{F}(\theta_0) \right\|_2 \right) + P_0 \left(\left\| \tilde{P}(\theta_0) \right\|_2 \leq 2 L r_{n,\gamma} \right) \end{aligned}$$

and we show below that both probabilities on the right hand side tend to zero.

The first term disappears in view of (3.5.20) and assumption (3.3.2). For the second term note, that in view of Markov's inequality and $P_0(\|W_n\|_2^2 \geq r_{n,\gamma}^2|\mathbb{X}) = \gamma$, where W_n is a centered GP with covariance kernel $\hat{C}_n$, we have $\gamma r_{n,\gamma}^2 \leq E[\|W_n\|_2^2|\mathbb{X}]$. Then

$$P_0\left(2Lr_{n,\gamma}^2 \geq \left\|\tilde{P}(\theta_0)\right\|_2^2\right) \leq P_0\left(E\left[\|W_n\|_2^2|\mathbb{X}\right] \geq \frac{\gamma}{2L}\left\|\tilde{P}(\theta_0)\right\|_2^2\right) \leq \frac{2LE_0(\int_{\mathcal{X}} \mathrm{Var}\left(\theta(x)|\mathbb{D}_n\right)dx)}{\gamma\left\|\tilde{P}(\theta_0)\right\|_2^2}. \tag{3.5.30}$$

The expectation in the numerator, known as the *learning curve*, is of order $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$ according to Lemma 3.7.4; thus for all $L > 0$ not depending on n the right hand side of the preceding display goes to 0 in view of assumption (3.3.2).

3.5.5.2 Distributed setting

Preliminary results. We start by introducing the distributed version of the notations introduced in Section 3.5.5.1. The aggregated posterior covariance function is $\hat{C}_{n,m}^I(x,x') = m^{-2}\sum_{k=1}^m \hat{C}_n^{I,(k)}(x,x')$, where the local posterior covariance functions can be given as $\hat{C}_n^{I,(k)}(x,x') = K_x^I(x') - \hat{K}_x^{I,(k)}(x')$ with

$$\begin{aligned}\hat{K}_x^{I,(k)}(\cdot) &= K^I\left(\cdot,\mathbb{X}^{(k)}\right)\left[K^I\left(\mathbb{X}^{(k)},\mathbb{X}^{(k)}\right)+\sigma^2 I_{n/m}\right]^{-1}K^I\left(\mathbb{X}^{(k)},x\right)\\ &= mK\left(\cdot,\mathbb{X}^{(k)}\right)\left[K\left(\mathbb{X}^{(k)},\mathbb{X}^{(k)}\right)+m^{-1}\sigma^2 I_{n/m}\right]^{-1}K\left(\mathbb{X}^{(k)},x\right).\end{aligned}$$

Then in view of (3.5.15),

$$m^{-1}\hat{K}_x^{I,(k)} = \arg\min_{\vartheta\in\mathcal{H}}\frac{1}{n/m}\left[\sum_{i=1}^{n/m}\left(K_x(X_i^{(k)}) - \vartheta(X_i^{(k)})\right)^2 + \frac{\sigma^2}{m}\|\vartheta\|_{\mathcal{H}}^2\right].$$

For convenience let us introduce the notation $\tilde{K}_x^{I,(k)} = m^{-1}\hat{K}_x^{I,(k)}$. Then the corresponding score function (up to constant multipliers) is given by

$$\hat{S}_{K_x,n}^{I,(k)}(\vartheta) = \frac{1}{n/m}\left(\sum_{i=1}^{n/m}\left(K_x(X_i^{(k)}) - \vartheta(X_i^{(k)})\right)K_{X_i^{(k)}} - \frac{\sigma^2}{m}\vartheta\right)$$

satisfying $\hat{S}_{K_x,n}^{I,(k)}(\tilde{K}_x^{I,(k)}) = 0$. Furthermore the expected value of the score function is

$$S_{K_x,n}^I(\vartheta) = E\hat{S}_{K_x,n}^{I,(k)}(\vartheta) = \int_{\mathcal{X}}\left(K_x(z) - \vartheta(z)\right)K_z dz - \frac{\sigma^2}{n}\vartheta = S_{K_x,n}(\vartheta),$$

hence $S_{K_x,n}^I(\tilde{F}(K_x)) = 0$.

Then similarly to the posterior mean in Section 3.5.1 the following assertions hold

$$\Delta\tilde{K}_x^{I,(k)} = \tilde{K}_x^{I,(k)} - \tilde{F}(K_x) = -\tilde{F} \circ F^{-1} \circ S^I_{K_x,n}\left(\tilde{K}_x^{I,(k)}\right),$$

$$\hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right) = \frac{1}{n/m}\left(\sum_{i=1}^{n/m} \tilde{P}(K_x)\left(X_i^{(k)}\right) K_{X_i^{(k)}} - \frac{\sigma^2}{m}\tilde{F}(K_x)\right), \tag{3.5.31}$$

$$\begin{aligned} F \circ \tilde{F}^{-1}\left(\Delta\tilde{K}_x^{I,(k)}\right) &- \hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right) \\ &= -\frac{1}{n/m}\sum_{i=1}^{n/m} \Delta\tilde{K}_x^{I,(k)}\left(X_i^{(k)}\right) K_{X_i^{(k)}} + \int_{\mathcal{X}} \Delta\tilde{K}_x^{I,(k)}(x') K_{x'} dx'. \end{aligned} \tag{3.5.32}$$

Main assertions. Similarly to the non-distributed case in Section 3.5.5.1, for the coverage of the credible sets it is sufficient to show that

$$P_0\left(r^2_{n,m}(\gamma) \geq C_2 \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d} \nu_j\right) \to 1, \tag{3.5.33}$$

$$P_0\left(\left\|\hat{\theta}_{n,m} - \tilde{F}(\theta_0)\right\|_2^2 \leq L_n \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d} \nu_j\right) \to 1, \tag{3.5.34}$$

where the radius $r_{n,m}(\gamma)$ is defined as $P(\|W_{n,m}\|_2^2 \leq r^2_{n,m}(\gamma)|\mathbb{X}) = 1-\gamma$ and $W_{n,m}$ is a centered GP with the same covariance kernel as $\Pi^{\dagger}_{n,m}(.|\mathbb{D}_n)$. Furthermore, the lack of coverage under (3.3.2) follows from

$$P_0\left(Lr^2_{n,m}(\gamma) \geq \left\|\tilde{P}(\theta_0)\right\|_2^2\right) \to 0. \tag{3.5.35}$$

We prove below the above assertions.

Proof of (3.5.33)**:** Similarly to the proof of (3.5.21) we get by Chebyshev's inequality that

$$r^2_{n,m}(\gamma) \geq E\left[\|W_{n,m}\|_2^2|\mathbb{X}\right] - (1-\gamma)^{-1/2}\operatorname{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right)^{1/2}.$$

Then in view of

$$\operatorname{Var}^I_{n,m}\left(\theta(x)\right) = m^{-2}\sum_{k=1}^{m}\operatorname{Var}^I\left(\theta(x)|\mathbb{D}_n^{(k)}\right), \qquad \text{for all } x\in\mathcal{X}, \tag{3.5.36}$$

and Lemma 3.7.4 it holds almost surely that

$$E\left[\|W_{n,m}\|_2^2|\mathbb{X}\right] = \int_{x\in\mathcal{X}} \operatorname{Var}^I_{n,m}\left(\theta(x)\right) dx \gtrsim \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j. \tag{3.5.37}$$

Furthermore, as in (3.5.25),

$$\operatorname{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right) = 2\int_{\mathcal{X}}\left\|\hat{C}^I_{n,m}(x,.)\right\|_2^2 dx.$$

Recall that the covariance function $\hat{C}^I_{n,m}(x,x') = m^{-2}\sum_{k=1}^m \hat{C}^{I,(k)}_n(x,x')$. Then in view of $(\sum_{i=1}^m a_i)^2 \le m\sum_{i=1}^m a_i^2$,

$$\left\|\hat{C}^I_{n,m}(x,.)\right\|_2^2 = \left\|m^{-2}\sum_{k=1}^m \hat{C}^{I,(k)}_n(x,.)\right\|_2^2 \le m^{-3}\sum_{k=1}^m \left\|\hat{C}^{I,(k)}_n(x,.)\right\|_2^2.$$

We show below that

$$E_0\left\|\hat{C}^{I,(k)}_n(x,.)\right\|_2^2 \lesssim m^2\left(\left\|\tilde{P}(K_x)\right\|_2^2 + \tilde{\delta}_n\right), \tag{3.5.38}$$

for $\tilde{\delta}_n = \inf\{(\sum_{j\in\mathbb{N}^d}\nu_j^2)^2\sum_{\ell\in\mathcal{I}^c}\mu_\ell : |\mathcal{I}| \le n/(m^2\log n\sum_{j\in\mathbb{N}^d}\nu_j^2)\}$ similarly to the non-distributed case. Then in view of assertion (3.5.27), the variance of $\|W_{n,m}\|_2^2$, similarly to (3.5.28), is bounded from above by

$$\begin{aligned} E_0\,\mathrm{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right) &= 2\int_{\mathcal{X}} E_0\left\|\hat{C}^I_{n,m}(x,.)\right\|_2^2 dx \\ &\lesssim \left(\int_{\mathcal{X}}\left\|\tilde{P}(K^I_x)\right\|_2^2 dx + \tilde{\delta}_n\right) \\ &= \frac{\sigma^4}{n^2}\sum_{j\in\mathbb{N}^d}\nu_j^2 + \tilde{\delta}_n. \end{aligned}$$

Hence for all $t>0$ we get by Markov's inequality and Lemmas 3.7.5 and 3.7.6 that

$$\begin{aligned} &P_0\left(\mathrm{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right) \ge t\frac{\sigma^4}{n^2}\left(\sum_{j\in\mathbb{N}^d}\nu_j\right)^2\right) \\ &\qquad\lesssim t^{-2}\left(\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{\left(\sum_{j\in\mathbb{N}^d}\nu_j\right)^2} + \frac{\tilde{\delta}_n n^2}{\sigma^4\left(\sum_{j\in\mathbb{N}^d}\nu_j\right)^2}\right) = o(1). \end{aligned}$$

Hence with P_0-probability tending to one $E[\|W_{n,m}\|_2^2|\mathbb{X}_n]$ is of higher order than $\mathrm{Var}(\|W_{n,m}\|_2^2)^{1/2}$. Therefore, the quantiles of $\|W_{n,m}\|_2^2$ are of the order $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$ with P_0-probability going to one, including $r^2_{n,m}(\gamma)$.

Proof of (3.5.38)**:** We adapt the proof of (3.5.26) to the distributed setting. First note that

$$\begin{aligned} \left\|\hat{C}^{I,(k)}_n(x,.)\right\|_2^2 \lesssim m^2\Bigg(&\left\|\tilde{P}(K_x)\right\|_2^2 + \left\|\Delta\tilde{K}^{I,(k)}_x - \tilde{F}\circ F^{-1}\circ\hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2 \\ &+ \left\|\tilde{F}\circ F^{-1}\circ\hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2\Bigg), \end{aligned}$$

where $\Delta\tilde{K}_x^{I,(k)} = \hat{K}_x^{I,(k)}/m - \tilde{F}(K_x)$. Then for, in view of (3.5.31), we get that

$$
\begin{aligned}
&E_0 \left\| \tilde{F} \circ F^{-1} \circ \hat{S}^{I,(k)}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 \\
&\quad = E_0 \left\| \tilde{F} \circ F^{-1} \left(\hat{S}^{I,(k)}_{K_x,n} \left(\tilde{F}\big(K_x\big) \right) - S^{I,(k)}_{K_x,n} \left(\tilde{F}\big(K_x\big) \right) \right) \right\|_2^2 \\
&\quad = E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \tilde{P}(K_x) \left(X_i^{(k)} \right) K_{X_i^{(k)}} - \int_{\mathcal{X}} \tilde{P}(K_x)(x') K_{x'} dx' \right) \right\|_2^2 \\
&\quad \leq \left(\frac{1}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 \right) \|\tilde{P}(K_x)\|_2^2 = o\left(\|\tilde{P}(K_x)\|_2^2 \right),
\end{aligned}
$$

where the penultimate inequality follows from Lemma 3.7.1 with $\vartheta = \tilde{P}(K_x)$.

Furthermore, similarly to the proof in Section 3.5.5.1, by using assertion (3.5.32), Lemma 3.7.2 (with $\hat{\vartheta}^{(k)} = \Delta\tilde{K}_x^{I,(k)}$, sample size n/m) and Lemma 3.7.3, we can show that for all $x \in \mathcal{X}$

$$
\begin{aligned}
&E_0 \left\| \Delta\tilde{K}_x^{I,(k)} - \tilde{F} \circ F^{-1} \circ \hat{S}^{I,(k)}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 \\
&\quad = E_0 \left\| (\tilde{F} \circ F^{-1}) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \Delta\tilde{K}_x^{I,(k)} \left(X_i^{(k)} \right) K_{X_i^{(k)}} - \int_{\mathcal{X}} \Delta\tilde{K}_x^{I,(k)}(x') K_{x'} dx' \right) \right\|_2^2 \\
&\quad \lesssim \frac{|\mathcal{I}| \log n \sum_{j \in \mathbb{N}^d} \nu_j^2}{n/m} E_0 \|\Delta\tilde{K}_x^{I,(k)}\|_2^2 + \tilde{\delta}_n.
\end{aligned}
$$

Taking the infimum over $|\mathcal{I}| = o\big(n/(m \log n \sum_{j \in \mathbb{N}^d} \nu_j^2)\big)$ we get that the left hand side of the preceding display is bounded from above by $o(E_0 \|\Delta\tilde{K}^{(k)}_{x,n}\|_2^2) + \tilde{\delta}_n$. We conclude the proof of (3.5.38) by combining the above three displays.

Proof of (3.5.34)**:** Exactly the same as the proof of (3.5.20).

Proof of (3.5.35)**:** Similarly to assertion (3.5.30) we get in view of (3.5.36) and Lemma 3.7.4 in the case where assumption (3.3.2) holds

$$
\begin{aligned}
P_0 \left(L r_{n,m}^2(\gamma) \geq \left\| \tilde{P}(\theta_0) \right\|_2^2 \right) &\leq \frac{2LE_0 \int_{\mathcal{X}} \mathrm{Var}_{n,m}\left(\theta(x)\right) dx.}{\gamma \left\| \tilde{P}(\theta_0) \right\|_2^2} \\
&\lesssim \frac{\sigma^2 \sum_{j \in \mathbb{N}^d} \frac{m\mu_j}{\sigma^2 + n\mu_j}}{m \left\| \tilde{P}(\theta_0) \right\|_2^2} \\
&= \frac{\sigma^2 \sum_{j \in \mathbb{N}^d} \nu_j}{n \left\| \tilde{P}(\theta_0) \right\|_2^2} = o(1).
\end{aligned}
$$

§3.6 Proof of the Corollaries

§3.6.1 Proof of Corollary 3.2.3

First note that for any $\mathcal{N} \subset \mathbb{N}^d$

$$\begin{aligned}\left\|\tilde{P}(\theta_0)\right\|_2^2 &= \sum_{j\in\mathbb{N}^d}(1-\nu_j)^2\theta_{0,j}^2 = \sum_{j\in\mathbb{N}^d}\frac{\sigma^4}{(\sigma^2+n\mu_j)^2}\theta_{0,j}^2 \\ &\leq (n/\sigma^2)^{-2}\sum_{j\in\mathcal{N}}\frac{1}{\mu_j^2}\theta_{0,j}^2 + \sum_{j\in\mathbb{N}^d/\mathcal{N}}\theta_{0,j}^2. \end{aligned} \tag{3.6.1}$$

Consider eigenvalues satisfying (3.1.6) with $\alpha = \beta$, i.e. $\mu_j \asymp \left(\prod_{i=1}^d j_i\right)^{-2\beta/d-1}$. Let us take $\mathcal{N} = \{j \in \mathbb{N}^d : \Pi_{i=1}^d j_i \leq J_\beta\}$ with $J_\beta := (n/\sigma^2)^{d/(d+2\beta)}$ and note that in view of (3.7.6) [with $I = J_\beta$] we have

$$|\mathcal{N}| \lesssim J_\beta \log^{d-1} J_\beta = o(n) \tag{3.6.2}$$

Furthermore, we also get that

$$\begin{aligned}\sup_{\theta_0\in\Theta^\beta(B)}\left\|\tilde{P}(\theta_0)\right\|_2^2 &\lesssim \sup_{\theta_0\in\Theta^\beta(B)}\left[\frac{\sigma^4}{n^2}\max_{j\in\mathcal{N}}\left(\prod_{i=1}^d j_i\right)^{4\beta/d+2}\left(\sum_{i=1}^d j_i\right)^{-2\beta}\sum_{j\in\mathcal{N}}\left(\sum_{i=1}^d j_i\right)^{2\beta}\theta_{0,j}^2\right. \\ &\qquad \left. + \sup_{j\notin\mathcal{N}}\left(\sum_{i=1}^d j_k\right)^{-2\beta}\sum_{j\notin\mathcal{N}}\left(\sum_{i=1}^d j_i\right)^{2\beta}\theta_{0,j}^2\right] \\ &\lesssim (n/\sigma^2)^{-2}J_\beta^{2\beta/d+2}B^2 + J_\beta^{-2\beta/d}B^2 \\ &\lesssim (n/\sigma^2)^{-2\beta/(d+2\beta)},\end{aligned}$$

using Lemmas 3.7.10 [with $r = 4\beta/d+2$, $s = 2\beta$ and $J = J_\beta$] and 3.7.11 [with $s = 2\beta$ and $J = J_\beta$].

Moreover, in view of Lemma 3.7.5 and $\nu_j \leq 1$

$$\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j^2 \lesssim \frac{\sigma^2}{n}J_\beta\log^{d-1}J_\beta = (n/\sigma^2)^{-2\beta/(d+2\beta)}\left(\log\left(\frac{n}{\sigma^2}\right)\right)^{d-1}.$$

Finally we show that the remaining term is $\delta_n = o(n^{-2\beta/(d+2\beta)})$ for the choice

$$\mathcal{I} = \left\{j \in \mathbb{N}^d : \prod_{i=1}^d j_i \leq I\right\} \qquad \text{with } I = \frac{n}{m^2\log^d(n/m)}\Big(\sum_{j\in\mathbb{N}^d}\nu_j^2\Big)^{-1},$$

where $\frac{n}{m^2\log^d(n/m)}\Big(\sum_{j\in\mathbb{N}^d}\nu_j^2\Big)^{-1} \geq 1$ holds because m is small enough. Note that in view of Lemma 3.7.8 the cardinality of $\mathcal{I}$ satisfies $|\mathcal{I}| \lesssim \frac{n}{m^2\log n}\Big(\sum_{j\in\mathbb{N}^d}\nu_j^2\Big)^{-1}$, hence

it satisfies the cardinality assumption on $\mathcal{I}$. Then in view of Lemma 3.7.8 and Lemma 3.7.5

$$\begin{aligned}\delta_n &\lesssim n \sum_{j\in\mathbb{N}^d} \nu_j^2 \sum_{\ell:\, \prod_{i=1}^d \ell_i > I} \mu_\ell \lesssim n \sum_{j\in\mathbb{N}^d} \nu_j^2 I^{-2\beta/d} \log^{d-1} I \\ &\ll n^{1-2\beta/d} m^{4\beta/d} \left(\sum_{j\in\mathbb{N}^d} \nu_j^2\right)^{2\beta/d+1} (\log n)^{2\beta+d-1} \\ &\lesssim n^{2-2\beta/d} m^{4\beta/d} (\log n)^{2\beta+d-1}.\end{aligned}$$

The right hand side is of order $o(n^{-2\beta/(2\beta+d)})$ for all $m = o(n^{1/2-3d/(4\beta)})$ with $\beta > 3d/2$. Combining the above inequality with Theorem 3.2.2 concludes the proof for the polynomially decaying eigenvalues.

§3.6.2 Proof of Corollary 3.2.4

For arbitrary index set $\mathcal{N} \subset \mathbb{N}^d$ we get that

$$\begin{aligned}\sup_{\theta_0\in\Theta^\beta(B)} \left\|\tilde{P}(\theta_0)\right\|_2^2 \leq \sup_{\theta_0\in\Theta^\beta(B)} \Bigg[&\frac{\sigma^4}{n^2} \max_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} e^{2a\sum_{i=1}^d j_i} \sum_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2 \\ &+ \sup_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} \sum_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2 \Bigg]. \qquad (3.6.3)\end{aligned}$$

We deal with the two terms on the right hand side separately. Note that the function $x \mapsto x^{-2\beta} e^{2ax}$ is convex on $[1, J_a]$, for $J_a = a^{-1}\log(n/\sigma^2)$ with $a \leq 1$, and achieves its maximum at one of the end points. Let us take the set $\mathcal{N} = \{j \in \mathbb{N}^d : \sum_{k=1}^d j_k \leq J_a\}$ and note that

$$|\mathcal{N}| \leq a^{-d} \log^d n = o(n), \qquad (3.6.4)$$

, by the lower bound on a. Furthermore, by noting that $(\sum_{i=1}^d j_i)^2 \leq d\sum_{i=1}^d j_i^2$, the maximum of the last display over $\mathcal{N}$ is bounded from above by

$$\max_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} e^{2a\sum_{i=1}^d j_i} \lesssim 1 + J_a^{-2\beta} e^{2aJ_a}.$$

The second term in (3.6.3) is directly bounded from above by $J_a^{-2\beta} B^2$. Therefore, by combining the inequalities above,

$$\left\|\tilde{P}(\theta_0)\right\|_2^2 \lesssim \frac{\sigma^4}{n^2} + \left(a^{-1}\log(n/\sigma^2)\right)^{-2\beta}. \qquad (3.6.5)$$

Moreover, in view of Lemma 3.7.6

$$\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j \asymp \frac{\sigma^2}{n} J_a^d = \frac{\sigma^2}{n} a^{-d} \log^d(n/\sigma^2). \qquad (3.6.6)$$

For $a := (n/\sigma^2)^{-1/(2\beta+d)} \log(n/\sigma^2)$ both of the preceding displays are bounded from above by a multiple of $(n/\sigma^2)^{-2\beta/(2\beta+d)}$.

Finally we show that the remainder term δ_n is of lower order than $(n/\sigma^2)^{-2\beta/(2\beta+d)}$. We take $\mathcal{I} = \{j \in \mathbb{N}^d : \sum_{i=1}^d j_i \leq I\}$, with $I = n^{1/d}\big(m^2 \log n \sum_{j\in\mathbb{N}^d} \nu_j^2\big)^{-1/d}$. Then it is easy to see that $|\mathcal{I}| \leq I^d \leq n\big(m^2 \log n \sum_{j\in\mathbb{N}^d} \nu_j^2\big)^{-1}$ holds. Note that $|\mathcal{I}| \geq 1$ holds because m is small enough. Furthermore, in view of the upper bound $p(j,d) \leq \frac{1}{2}\binom{j-1}{d-1} + 1/2 \leq j^d$ on the d partition of $j \in \mathbb{N}$, we get that

$$\begin{aligned}\delta_n &= n \sum_{j\in\mathbb{N}^d} \nu_j^2 \sum_{\ell\in\mathcal{I}^c} \mu_\ell \leq n \sum_{j\in\mathbb{N}^d} \nu_j^2 \sum_{\ell\geq I} \ell^d e^{-a\ell} \\ &\lesssim n I^d e^{-aI} \sum_{j\in\mathbb{N}^d} \nu_j^2 \lesssim (n/m)^2 e^{-aI} (\log n)^{-1} \end{aligned} \tag{3.6.7}$$

Since $\beta \geq d/2$, we have

$$\begin{aligned} aI &= (n/\sigma^2)^{-1/(2\beta+d)} \log(n/\sigma^2) n^{1/d} m^{-2/d} (\log n)^{-1/2} \left(\sum_{j\in\mathbb{N}^d} \nu_j^2\right)^{-1/d} \\ &\gtrsim n^{\frac{2\beta-d}{d(2\beta+d)}} m^{-2/d} (\log n)^{1-1/d} \geq L \log n. \end{aligned}$$

Hence the right hand side of (3.6.7) is $o(n^{-L})$, for arbitrary $L > 0$, when $m = o(n^{\frac{2\beta-d}{2(2\beta+d)}})$ concluding the proof of the corollary using Theorem 3.2.2.

§3.6.3 Proof of Corollary 3.3.2

We proceed by proving that the conditions of Theorem 3.3.1 hold for this choice of the kernel and the parameters, which directly provides us the statements.

Let us take $\mathcal{N} = \{j \in \mathbb{N}^d : \Pi_{k=1}^d j_k \leq J_\alpha\}$ with $J_\alpha := (n/\sigma^2)^{1/(2\alpha+d)}$ in (3.6.1). The cardinality of this set is $o(n)$, see (3.6.2). Furthermore, in view of $\alpha \leq \beta$,

$$\begin{aligned} \sup_{\theta_0\in\Theta^\beta(B)} \left\|\tilde{P}(\theta_0)\right\|_2^2 &\lesssim \sup_{\theta_0\in\Theta^\beta(B)} \left[\frac{\sigma^4}{n^2} \max_{j\in\mathcal{N}} \left(\prod_{i=1}^d j_i\right)^{4\alpha/d+2} \left(\sum_{i=1}^d j_i\right)^{-2\beta} \sum_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2 \right. \\ &\qquad \left. + \sup_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} \sum_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2\right] \\ &\lesssim (n/\sigma^2)^{-2} J_\alpha^{4\alpha/d-2\beta/d+2} B^2 + J_\alpha^{-2\beta/d} B^2 \lesssim (n/\sigma^2)^{-2\beta/(2\alpha+d)}. \end{aligned}$$

Then, in view of Lemma 3.7.5, $\nu_j \leq 1$ and the preceding display,

$$\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j \asymp \frac{\sigma^2}{n} J_\alpha (\log J_\alpha)^{d-1} = (n/\sigma^2)^{-2\alpha/(2\alpha+d)} \left(\log(n/\sigma^2)\right)^{d-1} \gtrsim \sup_{\theta_0\in\Theta^\beta(B)} \left\|\tilde{P}(\theta_0)\right\|_2^2,$$

when $\alpha \leq \beta$. Finally in view of Corollary 3.2.3 we have that

$$\delta_n = o\left((n/\sigma^2)^{-2\alpha/(2\alpha+d)}\right) = o\left(\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j\right),$$

finishing the proof of the corollary.

§3.6.4 Proof of Corollary 3.3.3

We again prove that the conditions of Theorem 3.3.1 hold in this setting.

In view of assertions (3.6.5) and (3.6.6), we get for $a \lesssim (\sigma^2/n)^{1/(2\beta+d)} \log(n/\sigma^2)$ that

$$\left\|\tilde{P}(\theta_0)\right\|_2^2 \lesssim \frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j.$$

Furthermore, the cardinality of the set $\{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}$ is $o(n)$, see (3.6.4). Finally, in view of Corollary 3.2.4, $\delta_n = o(n^{-c})$, hence the condition $\delta_n = o\Big(\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j\Big)$ of Theorem 3.3.1 also holds, concluding the proof.

§3.7 Technical lemmas

Lemma 3.7.1. *Consider the local regression problem* (3.1.1) *for arbitrary* $k \in \{1, ..., m\}$ *and let* $\vartheta \in L_2(\mathcal{X})$. *Then there exists a universal constant* C *not depending on* ϑ *such that*

$$E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \vartheta(X_i^{(k)}) K_{X_i^{(k)}} - E_X[\vartheta(x) K_x dx] \right) \right\|_2^2 \leq \frac{C}{n/m} \|\vartheta\|_2^2 \sum_{j\in\mathbb{N}^d} \nu_j^2, \tag{3.7.1}$$

where X *is a uniform random variable on* $\mathcal{X}$, *and* ν_j*'s are the eigenvalues of the operator* $\tilde{F}$.

Proof. For simplicity we omit the reference to the local k machine in the proof by writing $X_i = X_i^{(k)}$. Let $\vartheta = \sum_{j\in\mathbb{N}^d} \vartheta_j \psi_j \in L_2(\mathcal{X})$. Since

$$\vartheta(X) K_X = \sum_{j,k\in\mathbb{N}^d} \mu_j \vartheta_k \psi_j(X) \psi_k(X) \psi_j,$$

and $(\psi_j)_{j\in\mathbb{N}^d}$ is an orthonormal basis of $L_2(\mathcal{X})$, we have $E_X[\vartheta(X)K_X] = \sum_{j\in\mathbb{N}^d} \mu_j \vartheta_j \psi_j$. Furthermore, the linearity of the operator $\tilde{F} \circ F^{-1}$ implies that $\tilde{F} \circ F^{-1}(\vartheta(X)K_X) = \sum_{j,k\in\mathbb{N}^d} \nu_j \vartheta_k \psi_j(X) \psi_k(X) \psi_j$, providing

$$\begin{aligned}
\tilde{F} \circ F^{-1} \left(E_X[\vartheta(X)K_X]\right) &= \sum_{j\in\mathbb{N}^d} \nu_j \vartheta_j \psi_j, \\
\tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \vartheta(X_i) K_{X_i} \right) &= \frac{1}{n/m} \sum_{i=1}^{n/m} \tilde{F} \circ F^{-1} \left(\vartheta(X_i) K_{X_i}\right) \\
&= \frac{1}{n/m} \sum_{i=1}^{n/m} \sum_{j,k\in\mathbb{N}^d} \nu_j \vartheta_k \psi_j(X_i) \psi_k(X_i) \psi_j. \qquad (3.7.2)
\end{aligned}$$

Then using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we get

$$
\begin{aligned}
&E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \vartheta(X_i) K_{X_i} - E_X[\vartheta(X) K_X] \right) \right\|_2^2 \\
&\quad = E_0 \left\| \sum_{j,k \in \mathbb{N}^d} \nu_j \vartheta_k \psi_j \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_k(X_i) - \delta_{jk} \right) \right\|_2^2 \\
&\quad = \sum_{j \in \mathbb{N}^d} \frac{\nu_j^2}{n/m} E_0 \left(\vartheta(X_i)\psi_j(X_i) - \vartheta_j\right)^2 \\
&\quad \leq 2 \sum_{j \in \mathbb{N}^d} \frac{\nu_j^2}{n/m} \left(E_0 \vartheta^2(X_i)\psi_j^2(X_i) + \vartheta_j^2 \right) \leq \frac{2(C_\psi^2+1)\|\vartheta\|_2^2}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2,
\end{aligned}
$$

finishing the proof of the statement. □

Lemma 3.7.2. *Consider the local regression problem* (3.2.1) *for arbitrary* $k \in \{1, ..., m\}$. *Then for any finite index set* $\mathcal{I} \subset \mathbb{N}^d$, $|\mathcal{I}| \leq n^C$ *and data dependent function* $\hat{\vartheta}^{(k)} : \mathcal{X}^{n/m} \mapsto \mathbb{R}$, $\|\hat{\vartheta}^{(k)}\|_2 \leq n^C$, *for some* $C > 0$,

$$
\begin{aligned}
&E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \hat{\vartheta}^{(k)}(X_i^{(k)}) K_{X_i^{(k)}} - E_X[\hat{\vartheta}^{(k)}(X) K_X] \right) \right\|_2^2 \\
&\lesssim \quad \frac{|\mathcal{I}| \log n}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 E_0 \left\| \hat{\vartheta}^{(k)} \right\|_2^2 + E_0 \left\| \hat{\vartheta}^{(k)}_{\mathcal{I}^c} \right\|_{\mathcal{H}}^2 \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell + n^{-C_0}, \qquad (3.7.3)
\end{aligned}
$$

where X *is a uniform random variable on* $\mathcal{X}$, ν_j*'s are the eigenvalues of the operator* $\tilde{F}$, C_0 *can be chosen arbitrarily large, and* $\hat{\vartheta}^{(k)}_{\mathcal{I}^c}(\cdot) = \sum_{j \in \mathcal{I}^c} \hat{\vartheta}^{(k)}_j \psi_j(\cdot)$.

Proof. For simplicity we omit the reference to the kth local problem and write $X_i = X_i^{(k)}$ and $\hat{\vartheta} = \hat{\vartheta}^{(k)}$. Let us next define the set $\mathcal{A}_{\mathcal{I},j} \subset \mathcal{X}^{n/m}$ as

$$
\mathcal{A}_{\mathcal{I},j} = \left\{ \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right)^2 \leq \frac{8C_\psi^2 C \log n}{n/m}, \quad \ell \in \mathcal{I} \right\}. \qquad (3.7.4)
$$

Note that by Hoeffding's inequality, for arbitrary $\ell \in \mathcal{I}$,

$$
\begin{aligned}
P\left(\mathcal{A}_{\mathcal{I},j}^c\right) &\leq |\mathcal{I}| P \left(\left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right)^2 > \frac{8C_\psi^2 C \log n}{n/m} \right) \\
&\leq 2|\mathcal{I}| \exp\left\{ -\frac{4C_\psi^2 C \log n}{C_\psi^2} \right\} \leq O(|\mathcal{I}| n^{-3C}).
\end{aligned}
$$

Then using $(a+b)^2 \le 2a^2 + 2b^2$ and Cauchy-Schwarz inequality

$$
\begin{aligned}
&E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \hat{\vartheta}(X_i) K_{X_i} - E_X[\hat{\vartheta}(X) K_X] \right) \right\|_2^2 \\
&\quad = E_0 \left\| \sum_{j \in \mathbb{N}^d} \sum_{\ell \in \mathbb{N}^d} \nu_j \hat{\vartheta}_\ell \psi_j \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right) \right\|_2^2 \\
&\quad \lesssim E_0 \sum_{j \in \mathbb{N}^d} \nu_j^2 \left(\sum_{\ell \in \mathcal{I}} \hat{\vartheta}_\ell \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right) \right)^2 \\
&\qquad + E_0 \sum_{j \in \mathbb{N}^d} \nu_j^2 \left(\sum_{\ell \in \mathcal{I}^c} |\hat{\vartheta}_\ell| (C_\psi^2 + 1) \right)^2 \\
&\quad \lesssim E_0 \sum_{j \in \mathbb{N}^d} \nu_j^2 |\mathcal{I}| \sum_{\ell \in \mathcal{I}} \hat{\vartheta}_\ell^2 \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right)^2 \\
&\qquad + \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell E_0 \sum_{\ell \in \mathcal{I}^c} \hat{\vartheta}_\ell^2 \mu_\ell^{-1} \\
&\quad \le \sum_{j \in \mathbb{N}^d} \nu_j^2 E_0 \|\hat{\vartheta}\|_2^2 \left(\frac{8 C_\psi^2 C |\mathcal{I}| \log n}{n/m} + 1_{A_{j,\mathcal{I}}^c} |\mathcal{I}| \right) + E_0 \|\hat{\vartheta}_{\mathcal{I}^c}\|_{\mathcal{H}}^2 \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell \\
&\quad \lesssim \frac{|\mathcal{I}| \log n}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 E_0 \left\| \hat{\vartheta} \right\|_2^2 + E_0 \left\| \hat{\vartheta}_{\mathcal{I}^c} \right\|_{\mathcal{H}}^2 \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell + O(n^{-C}),
\end{aligned}
$$

where C can be chosen arbitrarily large, concluding the proof of our statement. □

Lemma 3.7.3. *There exists $C > 0$ such that*

$$
E_0 \|\hat{K}_{x,n}^{I,(k)}/m - \tilde{F}(K_x)\|_{\mathcal{H}}^2 \le C \sum_{j \in \mathbb{N}^d} \nu_j^2.
$$

Proof. First note that

$$
\|\hat{K}_{x,n}^{I,(k)}/m - \tilde{F}(K_x)\|_{\mathcal{H}}^2 \le 2m^{-2} \|\hat{K}_{x,n}^{I,(k)}\|_{\mathcal{H}}^2 + 2\|\tilde{F}(K_x)\|_{\mathcal{H}}^2.
$$

The second term on the right hand is bounded by

$$
\|\tilde{F}(K_x)\|_{\mathcal{H}}^2 = \sum_{j \in \mathbb{N}^d} \mu_j^{-1} \nu_j^2 \mu_j^2 \psi_j(x)^2 \le C_\psi^2 \sum_{j \in \mathbb{N}^d} \mu_j \nu_j^2 \lesssim \sum_{j \in \mathbb{N}^d} \nu_j^2.
$$

Since $\hat{K}_{x,n}^{I,(k)}$ is a KRR estimator, we get that

$$\begin{aligned} E_0\sigma^2\|\tilde{K}_{x,n}^{(k)}\|_{\mathcal{H}}^2 &\leq E_0\left(\sum_{i=1}^{n/m}(\tilde{K}_{x,n}^{(k)}(X_i^{(k)}) - K_x(X_i^{(k)}))^2 + \sigma^2\|\tilde{K}_{x,n}^{(k)}\|_{\mathcal{H}}^2\right) \\ &\leq E_0\left(\sum_{i=1}^{n/m}(\tilde{F}(K_x)(X_i^{(k)}) - K_x(X_i^{(k)}))^2 + \sigma^2\|\tilde{F}(K_x)\|_{\mathcal{H}}^2\right) \\ &\leq \sum_{i=1}^{n/m} E_0\tilde{P}(K_x)^2(X_i^{(k)}) + \sigma^2\|\tilde{F}(K_x)\|_{\mathcal{H}}^2 = O(\sum_{j\in\mathbb{N}^d}\nu_j^2), \end{aligned}$$

where the last inequality follows from (3.5.27). □

Lemma 3.7.4. *Assume that the eigenvalues μ_j of the covariance kernel K satisfy $\sum_{j\in\mathbb{N}^d}\mu_j < \infty$, $|\{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}| \leq n$, and $\sigma^2 \geq c > 0$. Then the expectation of the posterior variance is of the following order*

$$E_0E_X \operatorname{Var}(\theta(X)|\mathbb{D}_n) \asymp \sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j},$$

where the expectation E_X corresponds to the random variable $X \sim U[0,1]^d$ and the multiplicative constant depends on $\sum_{j\in\mathbb{N}^d}\mu_j$ and c.

Proof. It is shown in Section 6 of (Opper and Vivarelli, 1999) that the expectation of the posterior variance, named "generalization error", is bounded from below as follows

$$E_0E_X \operatorname{Var}(\theta(X)|\mathbb{D}_n) \geq \sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j E_X\psi_j^2(X)} = \sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j}.$$

In (Ferrari-Trecate et al., 1998), it has been shown that for stationary GPs, for any $\mathcal{J} \subset \mathbb{N}^d$, with $|\mathcal{J}| \leq n$, the learning curve is bounded from above by

$$E_0E_X \operatorname{Var}(\theta(X)|\mathbb{D}_n) \leq \sum_{j\in\mathbb{N}^d}\mu_j - n\sum_{j\in\mathcal{J}}\frac{\mu_j^2}{c_j},$$

where

$$c_j = (n-1)\mu_j + \sigma^2 + \sum_{j\in\mathbb{N}^d}\mu_j.$$

Let us take $\mathcal{J} = \{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}$ and by assumption its cardinality is

bounded by n. Then

$$\begin{aligned}\sum_{j\in\mathbb{N}^d}\mu_j - n\sum_{j\in\mathcal{J}}\frac{\mu_j^2}{c_j} &= \sum_{j\in\mathcal{J}}\mu_j\frac{c_j-n\mu_j}{c_j}+\sum_{j\notin\mathcal{J}}\mu_j\\ &= \sum_{j\in\mathcal{J}}\mu_j\frac{\sum_{j\in\mathbb{N}^d}\mu_j+\sigma^2-\mu_j}{\sum_{j\in\mathbb{N}^d}\mu_j+\sigma^2+(n-1)\mu_j}+\sum_{j\notin\mathcal{J}}\mu_j\\ &\le \sigma^2\sum_{j\in\mathcal{J}}\mu_j\frac{\sum_{j\in\mathbb{N}^d}\mu_j/\sigma^2+1}{\sigma^2+n\mu_j}+2\sigma^2\sum_{j\notin\mathcal{J}}\frac{\mu_j}{\sigma^2+n\mu_j}\\ &\lesssim \sigma^2\sum_{j\in\mathbb{N}^d}\frac{\mu_j}{\sigma^2+n\mu_j},\end{aligned}$$

concluding our proof. □

Lemma 3.7.5. *For ν_j, $j\in\mathbb{N}^d$, defined in* (3.2.5) *with eigenvalues μ_j polynomially decaying according to Assumption 3.1.2 and $k\in\mathbb{N}$,*

$$\sum_{j\in\mathbb{N}^d}\nu_j^k \asymp J_\alpha\log^{d-1}J_\alpha,$$

where $J_\alpha=(n/\sigma^2)^{d/(2\alpha+d)}$.

Proof. Let $\mathcal{N}:=\{j\in\mathbb{N}^d : n\mu_j\ge\sigma^2\}=\{j\in\mathbb{N}^d:\prod_{i=1}^d j_i\le CJ_\alpha\}$ and we apply Lemma 3.7.8 [with $\mathcal{I}$=$\mathcal{N}$, I=CJ_α and $\gamma=k(2\alpha/d+1)-1$]. First, we prove the upper bound,

$$\begin{aligned}\sum_{j\in\mathbb{N}^d}\nu_j^k &= \sum_{j\in\mathbb{N}^d}\frac{(n\mu_j)^k}{(\sigma^2+n\mu_j)^k}\\ &\le \sum_{j\in\mathcal{N}}1+(n/\sigma^2)^k\sum_{j\notin\mathcal{N}}\mu_j^k\\ &\lesssim J_\alpha\log^{d-1}J_\alpha+(n/\sigma^2)^kJ_\alpha^{-k(2\alpha/d+1)+1}\log^{d-1}J_\alpha\\ &\lesssim J_\alpha\log^{d-1}J_\alpha.\end{aligned}$$

The lower bound follows similarly,

$$\sum_{j\in\mathbb{N}^d}\nu_j^k\ge\Big(\frac{n}{2\sigma^2}\Big)^k\sum_{j\notin\mathcal{N}}\mu_j^k\gtrsim(n/\sigma^2)^kJ_\alpha^{-k(2\alpha/d+1)+1}\log^{d-1}J_\alpha\gtrsim J_\alpha\log^{d-1}J_\alpha.$$

□

Lemma 3.7.6. *For ν_j, $j\in\mathbb{N}^d$, defined in* (3.2.5) *with eigenvalues μ_j exponentially decaying according to Assumption 3.1.2 with $b=1$, $a<1$ and $k\in\mathbb{N}$,*

$$\sum_{j\in\mathbb{N}^d}\nu_j^k\asymp J_a^d,$$

where $J_a=a^{-1}\log(n/\sigma^2)$.

Proof. Let $\mathcal{N}_d := \{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\} = \{j \in \mathbb{N}^d : \sum_{i=1}^d j_i \leq J_a + c/a\}$ with $c > 0$ a positive constant. Then it is easy to see that $|\mathcal{N}_d| \leq 2^d J_a^d$. Moreover, we will show by induction on d that

$$\sum_{j \notin \mathcal{N}_d} e^{-ak\sum_{i=1}^d j_i} \lesssim a^{-d}(n/\sigma^2)^{-k} \log^{d-1}(n/\sigma^2).$$

Let us start with the case $d = 1$. We can directly see that

$$\sum_{j > J_a} e^{-akj} \leq C e^{-akJ_a} \frac{e^{ak}}{e^{ak} - 1} \lesssim a^{-1}(n/\sigma^2)^{-k}.$$

Now, assume that our assumption holds for d and consider the case $d + 1$, then

$$\begin{aligned}
\sum_{j \notin \mathcal{N}_{d+1}} e^{-ak\sum_{i=1}^{d+1} j_i} &\lesssim \sum_{j_{1:d} \in \mathbb{N}^d} e^{-ak\sum_{i=1}^d j_i} \sum_{j_{d+1} > \max(J_a - \sum_{i=1}^d j_i, 0)} e^{-akj_{d+1}} \\
&\lesssim \sum_{j_{1:d} \in \mathbb{N}^d} (e^{-ak\sum_{i=1}^d j_i} \wedge e^{-akJ_a}) \frac{e^{ak}}{e^{ak} - 1} \\
&\lesssim \sum_{j_{1:d} \in \mathcal{N}_d} a^{-1} e^{-akJ_a} + \sum_{j_{1:d} \notin \mathcal{N}_d} a^{-1} e^{-ak\sum_{i=1}^d j_i} \\
&\lesssim a^{-1}|\mathcal{N}_d|(n/\sigma^2)^{-k} + a^{-d-1}(n/\sigma^2)^{-k} \log^{d-1}(n/\sigma^2) \\
&\lesssim a^{-d-1}(n/\sigma^2)^{-k} \log^d(n/\sigma^2),
\end{aligned}$$

which concludes the induction proof.

Using these two results, we can easily show that

$$\sum_{j \in \mathbb{N}^d} \nu_j^k \lesssim \sum_{j \in \mathcal{N}_d} 1 + (n/\sigma^2)^k \sum_{j \notin \mathcal{N}_d} e^{-ak\sum_{i=1}^d j_i} \lesssim |\mathcal{N}_d| + a^{-d} \log^{d-1}(n/\sigma^2) \lesssim J_a^d.$$

On the other hand, we can show by induction that for all $J > d$, the cardinality of $\mathcal{N}_d := \{j \in \mathbb{N}^d : \sum_{i=1}^d j_i \leq J\}$ is bounded from below as follows

$$|\mathcal{N}_d| \geq (J - d)^d / d!.$$

Note that it holds trivially for $d = 1$. Now assume it holds for d, then we can write $\mathcal{N}_{d+1}$ as a partition as follows

$$\mathcal{N}_{d+1} = \left\{ j \in \mathbb{N}^{d+1} : \sum_{k=1}^{d+1} j_k \leq J \right\} = \bigcup_{i=1}^{J-d} \left\{ j \in \mathbb{N}^{d+1} : j_{d+1} = i; \sum_{k=1}^{d} j_k \leq J - i \right\}.$$

According to our induction assumption, the cardinality of all these subsets are bounded from below by $(J - d - i)^d / d!$, hence we have

$$|\mathcal{N}_{d+1}| \geq \sum_{i=1}^{J-d} \frac{(J - d - i)^d}{d!} \geq \int_1^{J-d} \frac{(J - d - t)^d}{d!} dt = \frac{(J - d - 1)^{d+1}}{(d+1)!},$$

which concludes our induction proof. Using this result, we can now show that

$$\sum_{j\in\mathbb{N}^d} \nu_j^k \geq \sum_{j\in\mathcal{N}_d} 1 = |\mathcal{N}_d| \gtrsim J_a^d,$$

concluding the proof. □

Lemma 3.7.7. *For arbitrary $\theta_0 \in \ell_2(L)$ we get that*

$$E_0\|\Delta\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \leq Cn,$$

for some universal constant $C > 0$.

Proof. First note that

$$\|\Delta\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \leq 2\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 + 2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2.$$

For $\theta_0 \in \ell_2(L)$ the second term on the right hand is bounded by

$$\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2 = \sum_{j\in\mathbb{N}^d} \mu_j^{-1}\nu_j^2\theta_{0,j}^2 \leq \sum_{j\in\mathbb{N}^d} \frac{n^2\mu_j}{(\sigma^2+n\mu_j)^2}\theta_{0,j}^2 \leq nL^2/\sigma^2.$$

Then by the definition of $\hat{\theta}_n^{(k)}$ we get that

$$\begin{aligned}
\sigma^2\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 &\leq \sum_{i=1}^{n/m}\left(\hat{\theta}_n^{(k)}(X_i^{(k)}) - Y_i^{(k)}\right)^2 + \sigma^2\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \\
&\leq \left(\sum_{i=1}^{n/m}\left(\tilde{F}(\theta_0)(X_i^{(k)}) - \theta_0(X_i^{(k)}) - \varepsilon_i^{(k)}\right)^2 + \sigma^2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2\right) \\
&\leq 2\sum_{i=1}^{n/m}\tilde{P}(\theta_0)^2(X_i^{(k)}) + 2\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 + \sigma^2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2. \qquad (3.7.5)
\end{aligned}$$

We conclude the proof by taking the expectation of both sides

$$\sigma^2 E_0\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \lesssim \sum_{i=1}^{n/m} E_0\tilde{P}(\theta_0)^2(X_i^{(k)}) + \sum_{i=1}^{n/m} E_0(\varepsilon_i^{(k)})^2 + \sigma^2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2 = O(n).$$

□

Lemma 3.7.8. *The cardinality of the set*

$$\mathcal{I}_{I,d} = \left\{ j \in \mathbb{N}_+^d : \prod_{i=1}^d j_i \leq I \right\} \qquad (3.7.6)$$

satisfies that $|\mathcal{I}_{I,d}| \leq 2^d I \log^{d-1} I$. Furthermore,

$$\sum_{j\in\mathcal{I}_{I,d}^c}\prod_{i=1}^d j_i^{-\gamma-1} \asymp I^{-\gamma}\left(\log I\right)^{d-1}, \qquad (3.7.7)$$

for some large enough constant $C_{\gamma,d}$.

Proof. We prove both statements by induction, starting with the first one. For $d = 1$ it is trivial. Let us assume that it holds for d and consider the case $d + 1$. We distinguish cases according the value of j_{d+1}. If $j_{d+1} = 1$, then $\prod_{i=1}^{d} j_i \leq I$ holds, if $j_{d+1} = 2$, then $\prod_{i=1}^{d} j_i \leq I/2$ holds, and so on. Hence we can write that

$$|\mathcal{I}_{I,d+1}| \leq \sum_{j_{d+1}=1}^{I} \left|\mathcal{I}_{I/j_{d+1},d+1}\right| \leq 2^d \sum_{j_{d+1}=1}^{I} \frac{I}{j_{d+1}} \log^{d-1} \frac{I}{j_{d+1}} < 2^{d+1} I \log^d I,$$

where in the last inequality we have used that $\sum_{i=1}^{n} 1/i < 1 + \log n < 2 \log n$.

Note again that for $d = 1$ the second statement holds trivially (using Riemann sums for instance). Then assume that it holds for d and consider the case $d+1$. First we deal with the upper bound, where we note that

$$\begin{aligned}
\sum_{j \in \mathcal{I}^c_{I,d+1}} \prod_{i=1}^{d+1} j_i^{-\gamma-1} &= \sum_{j_{d+1}=1}^{I} j_{d+1}^{-\gamma-1} \sum_{j \in \mathcal{I}^c_{I/j_{d+1},d}} \prod_{i=1}^{d} j_i^{-\gamma-1} \\
&\quad + \sum_{j_{d+1}=I}^{\infty} j_{d+1}^{-\gamma-1} \prod_{k=1}^{d} \sum_{j_k=1}^{\infty} j_k^{-\gamma-1} \\
&\lesssim \sum_{j_{d+1}=1}^{I} \frac{1}{j_{d+1}} I^{-\gamma} \left(\log I / j_{d+1}\right)^{d-1} + \sum_{j_{d+1}=I}^{\infty} j_{d+1}^{-\gamma-1} \\
&\leq I^{-\gamma} \log^{d-1} I \sum_{j_{d+1}=1}^{I} \frac{1}{j_{d+1}} + I^{-\gamma} \leq I^{-\gamma} \log^d I.
\end{aligned}$$

Finally, it remained to deal with the lower bound. First, note that it is sufficient to show the result for $I \geq C$, for some C large enough (depending only on d and γ). Then by noting that for $x \geq e^{d-1}$ the function $x \mapsto x^{-1} \log^{d-1} x$ is monotone decreasing, we get that

$$\begin{aligned}
\sum_{j \in \mathcal{I}^c_{I,d+1}} \prod_{i=1}^{d+1} j_i^{-\gamma-1} &\geq \sum_{j_{d+1}=1}^{I} j_{d+1}^{-\gamma-1} \sum_{j \in \mathcal{I}^c_{I/j_{d+1},d}} \prod_{i=1}^{d} j_i^{-\gamma-1} \\
&\gtrsim I^{-\gamma} \left(\sum_{j_{d+1}=1}^{I} j_{d+1}^{-1} \log^{d-1} I - \sum_{j_{d+1}=1}^{I} j_{d+1}^{-1} \log^{d-1} j_{d+1} \right) \\
&\geq I^{-\gamma} \Big(\log^{d-1} I \int_1^I x^{-1} dx - \sum_{j_{d+1}=1}^{e^{d-1}} j_{d+1}^{-1} \log^{d-1} j_{d+1} \\
&\qquad - \int_{e^{d-1}}^{I} x^{-1} \log^{d-1} x dx \Big) \\
&\geq I^{-\gamma} \left(\log^d I - C_{d,\gamma} - \log^d I/2) \right) \gtrsim I^{-\gamma} \log^d I,
\end{aligned}$$

concluding the proof of our statement. □

Lemma 3.7.9. *There exists an event $A_n^{(k)}$ such that for any $\theta_0 \in L_\infty(L)$ and $n \le (n/m)^{C_1}$, for some $C_1 \ge 1$ there exist constants $C_2, C_3 > 0$ such that*

$$\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2 1_{A_n^{(k)}} \le (n/m)^{C_2},$$
$$E_{\theta_0}\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 1_{(A_n^{(k)})^c} = e^{-C_3 n/m}.$$

Proof. Let us take $A_n^{(k)} = \{\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 \le (n/m)^{C_0}\}$, for arbitrary $C_0 > 1$. Then in view of (3.7.5) we have on the event $A_n^{(k)}$ that

$$\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2 \le \left\|\hat{\theta}_n^{(k)}\right\|_2 + \|\tilde{F}(\theta_0)\|_2 \lesssim n^{1/2} + (n/m)^{C_0} + L \lesssim (n/m)^{C_0\vee C_1/2}.$$

Furthermore, note that

$$\begin{aligned}\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 &\lesssim \left\|\Delta\hat{\theta}_n^{(k)}\right\|_2^2 + \left\|\tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 \\ &\lesssim n + \sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 + n^2\left\|\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2.\end{aligned}$$

Furthermore from the definition of $\hat{S}_n^{(k)}$, the boundedness of $\mathcal{X}$ and $\|K\|_\infty = O(1)$ we get that

$$\begin{aligned}\left\|\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 \le \left\|\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_\infty^2 &\lesssim \left(\frac{1}{n/m}\sum_{i=1}^{n/m}|\varepsilon_i^{(k)}|\right)^2 + \|\theta_0\|_\infty^2 \\ &\lesssim \frac{1}{n/m}\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 + 1.\end{aligned}$$

Since $W_n = (n/m)^{-1}\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 \sim \chi^2_{n/m}$, note that for n/m large enough

$$\begin{aligned}&E_{\theta_0}\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 1_{(A_n^{(k)})^c} \\ &\quad\lesssim E_{\theta_0}\left((n/m)^{2C_1}W_n + (n/m)^{2C_1}\right)1_{W_n > (n/m)^{C_0}} = O(e^{-n/m}),\end{aligned}$$

concluding the proof of the lemma. □

Lemma 3.7.10. *Let $r, s > 0$ such that $r > s/d$ and $f : [1,\infty)^d \to \mathbb{R}$ defined as*

$$f(x) = \left(\prod_{i=1}^d x_i\right)^r \left(\sum_{i=1}^d x_i\right)^{-s}.$$

Then f is bounded from above by $d^{-s}J^{r-s/d}$ on the set $\mathcal{N} := \{x \in [1,\infty)^d : \prod_{i=1}^d x_i \le J\}$ with $J > 1$.

Proof. From the inequality of arithmetic and geometric means, we know that for all $x \in [1, \infty)^d$

$$\sum_{i=1}^{d} x_i \geq d \left(\prod_{i=1}^{d} x_i \right)^{1/d}.$$

Thus, we can bound f from above by

$$f(x) \leq d^{-s} \left(\prod_{i=1}^{d} x_i \right)^{r-s/d} \leq d^{-s} J^{r-s/d},$$

on $\mathcal{N}$ concluding the proof. □

Lemma 3.7.11. *Let $s > 0$ and $f : [1, \infty)^d \to \mathbb{R}$ defined as*

$$f(x) = (\sum_{i=1}^{d} x_i)^{-s}.$$

Then f is bounded from above by $d^{-s} J^{-s/d}$ on the set $\mathcal{N} := \{x \in [1, \infty)^d : \prod_{i=1}^{d} x_i \geq J\}$ with $J > 1$.

Proof. Since f is differentiable on its domain, we can compute its gradient

$$(\nabla f)_\ell = -s(\sum_{k=1}^{d} x_k)^{-s-1} < 0,$$

for all $\ell \in \{1, ..., d\}$. Thus, the function attains its maximum at $\prod_{i=1}^{d} x_i = J$. At the maximum point, in view of the inequality of arithmetic and geometric means, $\sum_{i=1}^{d} x_i \geq d \left(\prod_{i=1}^{d} x_i \right)^{1/d} = dJ^{1/d}$. The statement of the lemma follows by raising both sides to the $-s$ power. □

CHAPTER 4

Optimal recovery for spatially distributed Gaussian process regression

Abstract. This chapter delves into spatially distributed Gaussian process regression. Even though it has been seen that distributed Bayesian non-parametric regression not only allow tractable computation, but can also lead to optimal contraction rate and good coverage of the credible sets when the data is distributed uniformly randomly across the machines when the smoothness of the underlying parameter is known. Nonetheless, the knowledge of said smoothness is often not realistic in practice. Therefore, we look into another way to split the data between the machines. We see that if the data is distributed spatially, in other words, if machines receive data from non-overlapping sub-regions of the design points, then appropriately scaled and aggregated local priors result in adaptive global posterior contraction rates. Furthermore, this result holds true when using a fully Bayesian adaptive procedure without any knowledge of the smoothness of the true parameter.

§4.1 Spatially distributed GP regression

In our analysis we consider the non-parametric regression model, where the observed data $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ satisfy the regression relation

$$Y_i = \theta_0(x_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, ..., n,$$

where $x_1, ..., x_n$ are the fixed covariates and the goal is to estimate the unknown regression function θ_0 belonging to a Hölder regularity class (i.e. $\theta_0 \in C^\beta([0,1])$). For simplicity of the analysis we take $\sigma = 1$ known. Then the data is distributed over m machines spatially in the following way. The kth machine, $k \in \{1, ..., m\}$ received the observations Y_i with design points belonging to the kth equidistant sub-interval, i.e. $x_i \in (\frac{k-1}{m}, \frac{k}{m}]$. For convenience we introduce the shorthand notation $S^{(k)} = (\frac{k-1}{m}, \frac{k}{m}]$, $\mathbf{x}^{(k)} = \{x_i : x_i \in S^{(k)}\}$, and $Y^{(k)} = \{Y_i : x_i \in S^{(k)}\}$.

In the Bayesian approach we endow the functional parameter θ_0 in each machine with the same Gaussian Process prior G_t supported on $t \in [0,1]$. These Gaussian processes are often endowed by scaling factor a which can be adjusted to achieve bigger flexibility. As a next step the corresponding local posteriors are computed

based on the data available at the local machines. Finally the local Gaussian processes are aggregated to a global one by restricting them to the intervals the corresponding data falls and pasting them together, i.e. a draw θ from the aggregated posterior is defined as

$$\theta(x) = \sum_{k=1}^{m} 1_{(\frac{k-1}{m}, \frac{k}{m}]}(x)\theta^{(k)}(x),$$

where $\theta^{(k)}$ is a draw from the kth local posterior. Then the aggregated posterior measure takes the form

$$\Pi_{n,m}(\theta \in B|\mathbf{Y}) = \prod_{k=1}^{m} \Pi^{(k)}(\theta^{(k)} \in B_k|\mathbf{Y}^{(k)}),$$

for any set of functions B, where B_k is the measurable set of functions whose restrictions to $S^{(k)}$ is equal to the corresponding restriction of an element of B, i.e

$$B_k = \left\{\vartheta : \exists \theta \in B \text{ such that } \vartheta(x) = \theta(x), x \in S^{(k)}\right\},$$

and $\Pi^{(k)}(\cdot|\mathbf{Y}^{(k)})$ the kth posterior distribution.

§4.1.1 Posterior contraction for distributed GP

We investigate the contraction rate of the aggregate posterior constructed this way $\Pi_{n,m}(\cdot|\mathbf{Y})$. To this end we introduce the local version of the concentration function originally introduced in (van der Vaart and van Zanten, 2008) for the non-distributed model. Let us formally restrict the prior in the kth machine to the $S^{(k)} = (\frac{k-1}{m}, \frac{k}{m}]$ interval and define

$$\phi_{\theta_0}^{(k)}(\varepsilon) = \inf_{\vartheta \in \mathbb{H}^{(k)} : \|\theta_0 - \vartheta\|_{\infty,k} \leq \varepsilon} \|\vartheta\|_{\mathbb{H}^{(k)}}^2 - \log \Pi\left(\theta : \|\theta\|_{\infty,k} < \varepsilon\right),$$

where $\|\theta\|_{\infty,k} = \sup_{x \in S^{(k)}} |\theta(x)|$ denotes the L_∞-norm restricted to $S^{(k)}$, and $\|.\|_{\mathbb{H}^{(k)}}$ is the norm corresponding to the Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}^{(k)}$ of the Gaussian Process prior G_t on $S^{(k)}$. Then the aggregated posterior contraction rate can be expressed with the help of the local concentration functions.

Theorem 4.1.1. *Let θ_0 be a bounded function and assume that there exists $\varepsilon_n \to 0$, $(n/m^2)\varepsilon_n^2 \to \infty$, such that $\phi_{\theta_0}^{(k)}(\varepsilon_n) \leq \varepsilon_n^2(n/m)$, $k = 1, ..., m$. Then*

$$E_0 \Pi_{n,m}\left(\theta : \|\theta - \theta_0\|_n \geq M_n \varepsilon_n | \, \boldsymbol{Y}\right) \to 0,$$

for arbitrary $M_n \to \infty$, where $\|\cdot\|_n$ denotes the empirical L_2-norm, i.e. $\|\theta\|_n^2 = n^{-1}\sum_{i=1}^{n} \theta(x_i)^2$.

The proof is deferred to Section 4.3.1.

In Section 4.3.3 we apply the above results to provide minimax contraction rates for a properly tuned Gaussian Process priors: the integrated Brownian motion.

A common practice to tune the GP is to insert a time scale parameter a, i.e. consider a GP $t \mapsto G_t^a := G_{at}$ instead of the original process. Even though the qualitative smoothness of the sample paths should not change for any a, a dramatic impact on the posterior contraction rate can be observed when $a = a_n$ goes to infinity or zero with the sample size n. For $a > 1$ this entails shrinking a process on a bigger time set to the time set $[0, 1]$, whereas $a < 1$ corresponds to stretching. Intuitively shrinking makes the sample paths more variable, as the randomness on a bigger time set is packed inside a smaller one, whereas stretching creates a smoother process.

§4.1.2 Adaptation

The time scale parameter a usually depends on the regularity or smoothness class of the underlying function θ_0. However, since the regularity class of θ_0 is typically not available in practice, one would like to provide a method which doesn't rely on this knowledge and uses a data driven tuning of the Gaussian Process. In this chapter we consider a fully Bayesian approach, where the scaling parameter a is taken to be a random variable A, i.e. we consider the hierarchical prior $G^A = \{G_{At} : t \in [0, 1]\}$ restricted to $[0, 1]$ and endow A with another layer of prior. We will denote by g the Lebesgue density of A. In each local model we take an iid random variable A_k, $k = 1, ..., m$ resulting in the independent hierarchical priors

$$\Pi^{A_k}(\cdot) = \int \Pi^{(a)}(\cdot) g(a) da. \tag{4.1.1}$$

Then the aggregated posterior is constructed similarly as in the non-adaptive case, i.e. a draw θ from the aggregated posterior is defined as

$$\theta(x) = \sum_{k=1}^{m} 1_{(\frac{k-1}{m}, \frac{k}{m}]}(x) \theta^{(k)}(x),$$

where $\theta^{(k)}$ is a draw from the kth local hierarchical posterior. Then the aggregated posterior measure takes the form

$$\Pi_{n,m}^{A}(B|\mathbf{Y}) = \prod_{k=1}^{m} \Pi^{A_k}(B_k|\mathbf{Y}^{(k)}), \tag{4.1.2}$$

for any set of the form $B = \bigotimes_{k=1}^{m} B_k$, where B_k is a measurable set of functions restricted to the interval $(\frac{k-1}{m}, \frac{k}{m}]$, and $\Pi^{(k)}(\cdot|\mathbf{Y}^{(k)})$ the kth posterior distribution.

Theorem 4.1.2. *Let θ_0 be a bounded function and assume that there exists a sieve $B_{n,m}^{(k)}$, such that for all local hierarchical prior Π^{A_k} given in* (4.1.1) *it holds that*

$$\Pi^{A_k}(\theta : \theta \notin B_{n,m}^{(k)}) \leq e^{-4(n/m)\varepsilon_n^2} \tag{4.1.3}$$

$$\Pi^{A_k}(\theta : \|\theta - \theta_0\|_{\infty,k} \leq \varepsilon_n) \geq e^{-(n/m)\varepsilon_n^2} \tag{4.1.4}$$

$$\log N(\varepsilon_n, B_{n,m}^{(k)}, \|.\|_{\infty,k}) \leq (n/m)\varepsilon_n^2, \tag{4.1.5}$$

with $\varepsilon_n \to 0$. *Then for* m *satisfying* $(n/m^2)\varepsilon_n^2 \to \infty$ *the aggregated hierarchical posterior* $\Pi^A_{n,m}(\cdot|\,\boldsymbol{Y})$ *achieves the following contraction rate, i.e.*

$$E_0\Pi^A_{n,m}(\|\theta-\theta_0\|_n \geq M_n\varepsilon_n|\,\boldsymbol{Y}) \to 0, \tag{4.1.6}$$

for arbitrary $M_n \to \infty$, *where* $\|.\|_n$ *denotes the empirical* L_2*-norm.*

The proof is deferred to Section 4.3.2.

We will apply this result as well in Section 4.3.3 in order to obtain minimax contraction rates for the integrated Brownian motion with adaptive rescaling.

§4.2 Application to the Integrated Brownian Motion

The "released" ℓ-fold integrated Brownian motion is defined as

$$G_t := B\sum_{j=0}^{\ell}\frac{Z_j t^j}{j!} + I^\ell W_t,$$

with $t \in [0,1]$, random variables $(Z_j)_{j=1}^{\ell} \overset{iid}{\sim} \mathcal{N}(0,1)$ independent from the Brownian motion W_t, and constant $B > 0$. The functional operator I^ℓ is defined recursively as $I^1 = I$ and $I^\ell = I^{\ell-1}I$ for all $\ell \in \mathbb{N}$ with $IW_t = \int_0^t W_s ds$. It is considered as a way of smoothing out the rough Brownian motion W_t. Indeed, the sample paths of this process are ℓ times differentiable and their last derivative is Hölder of order almost $1/2$. It has been seen in Section 11.4 of (Ghosal and Van der Vaart, 2017) that the posterior contraction rate for this particular Gaussian process prior achieves the minimax rate for β-regular functions if and only if $\beta = \ell + 1/2$. Since this condition is restrictive, we introduce some form of rescaling of the prior.

It is known that the integrated Brownian motion is self-similar of order $\ell + 1/2$. Accordingly, a time rescaling is equivalent to a space rescaling with another coefficient. That is why we consider a time rescaling of $I^\ell W_t$ and introduce

$$G^a_t := B\sum_{j=0}^{\ell} Z_j (at)^j + I^\ell W_{at}, \quad t \in [0,1], \tag{4.2.1}$$

where $a > 0$ is a rescaling factor on the time axis. This process has been studied in Section 11.5 of (Ghosal and Van der Vaart, 2017) where they have demonstrated that taking $a := a_n = n^{\frac{\ell+1/2-\beta}{(\ell+1/2)(2\beta+1)}}$ also leads to optimal contraction rate in the minimax sense around $\theta_0 \in C^\beta[0,1]$ for $\beta \leq \ell + 1$, i.e.

$$\Pi^{a_n}\left(f : \|\theta-\theta_0\|_n \gtrsim n^{\frac{-\beta}{2\beta+1}}|\mathbf{Y}^n\right) \to 0.$$

We will show that choosing $a = a_n$ for all local distributions allows us to achieve the same contraction rate for the distributed version of this process.

Corollary 4.2.1. *Let $\beta > 1/2$ and consider the rescaled integrated Brownian motion prior G_t^a (4.2.1) for some $\ell \geq \beta - 1$, $a = n^{\frac{\ell+1/2-\beta}{(\ell+1/2)(2\beta+1)}}$ and $B^2 > n^{\frac{-1}{2\beta+1}} m$. Let $\Pi_{n,m}^a(\cdot|Y)$ the corresponding aggregated posterior; then for all $\theta_0 \in C^\beta([0,1])$, if $m = o(n^{\frac{1/2}{2\beta+1}})$*

$$E_0 \Pi_{n,m}^a \left(\theta : \|\theta - \theta_0\|_n \geq M_n n^{\frac{-\beta}{2\beta+1}} | (Y_i)_{i=1}^n \right) \to 0,$$

for arbitrary $M_n \to \infty$, where $\|.\|_n$ is the empirical L_2-norm on $[0,1]$.

This result shows that the aggregated posterior based on an appropriately rescaled integrated Brownian motion prior contracts around any Hölder smooth truth at an optimal rate provided that the number of machines does not increase at more than a sub-linear polynomial rate with the number of data points. Although the contraction rates are optimal, the procedure necessitates knowledge of the smoothness level ahead of time, considering it appears in the expression of the optimal rescaling factor. This is generally not realistic in practice; the smoothness of the function of interest is typically not known. Fortunately, it will be shown further that if each local rescaling factors A_k are iid random variables independent of W, and follow the hyper-prior distribution $g_{\ell,n,m}$ verifying, for positive constants C_1, D_1, C_2, D_2, non-negative constant p and every $a > 0$,

$$C_1 a^p \exp\{-D_1 \frac{n^{\frac{1}{2(\ell+1)}}}{m} a^{\frac{\ell+1/2}{\ell+1}}\} \leq g_{\ell,n,m}(a) \leq C_2 a^p \exp\{-D_2 \frac{n^{\frac{1}{2(\ell+1)}}}{m} a^{\frac{\ell+1/2}{\ell+1}}\}, \quad (4.2.2)$$

then the optimal contraction rate can be achieved in the distributed case as well.

Corollary 4.2.2. *Consider the hierarchical prior (4.1.2) with the Gaussian Process (4.2.1) taken as the rescaled integrated Brownian motion with some $\ell \geq 1$. Then if $m = o(n^{\frac{1}{4(\ell+1/2)}})$ and $\theta_0 \in C^\beta([0,1])$, for some $1/2 < \beta \leq \ell + 1/2$ the posterior adapts to the optimal minimax contraction rate, i.e.*

$$E_0 \Pi_{n,m}^A \left(\theta : \|\theta - \theta_0\|_n \geq M_n n^{\frac{-\beta}{2\beta+1}} | (Y_i)_{i=1}^n \right) \to 0,$$

for arbitrary $M_n \to \infty$, where $\|.\|_n$ is the empirical L_2-norm on $[0,1]$.

The proof is given in Section 4.3.4.

In the above theorem, we can see that if all the local rescaling parameters follow a suitably chosen hyper-prior, the resulting aggregated posterior will, in some way, adapt to the true Hölder smoothness of the function of interest. Indeed, the contraction rates around the true function will be optimal for all smoothness β in some interval as long as the hyper-prior and the number of experts are chosen appropriately.

§4.3 Proofs of the general results

In this section we have collected the proofs of the abstract contraction rate theorems both for the non-adaptive and adaptive settings.

§4.3.1 Proof of Theorem 4.1.1

In view of Theorem 3.3 of (van der Vaart and van Zanten, 2008) and Theorem 4 of (Ghosal and van der Vaart, 2007) the local contraction functions provides us that

$$E_0\Pi^{(k)}\left(\theta : \|\theta-\theta_0\|_{n,k} \geq M_n\varepsilon_n|\mathbf{Y}\right) \lesssim \frac{1}{(n/m)\varepsilon_n^2},$$

where $\|\theta\|_{n,k}^2 = \frac{1}{n/m}\sum_{i\in S^{(k)}}\theta(x_i)^2$. Note that

$$\|\theta\|_n^2 = \frac{1}{n}\sum_{i=1}^n \theta^2(x_i) = \frac{1}{m}\sum_{k=1}^m \frac{1}{n/m}\sum_{x_i\in S^{(k)}}\theta^2(x_i) = \frac{1}{m}\sum_{k=1}^m \|\theta\|_{n,k}^2.$$

Then by triangle inequality

$$\begin{aligned}
&E_0\Pi_{n,m}\left(\theta : \|\theta-\theta_0\|_n \geq M_n\varepsilon_n|\mathbf{Y}\right)\\
&\quad\leq \sum_{k=1}^m E_0\Pi_{n,m}\left(\theta : \|\theta-\theta_0\|_{n,k} \geq M_n\varepsilon_n|\mathbf{Y}\right)\\
&\quad= \sum_{k=1}^m E_0\Pi^{(k)}\left(\theta : \|\theta-\theta_0\|_{n,k} \geq M_n\varepsilon_n|\mathbf{Y}\right) \lesssim \frac{m^2}{n\varepsilon_n^2} = o(1),
\end{aligned}$$

concluding the proof of the statement.

§4.3.2 Proof of Theorem 4.1.2

The result in the adaptive case is based on a comparison between the comparisons of the Kullback–Leibler divergence and variance, and the norm $\|\cdot\|_n$ to the uniform norm on the Gaussian process as seen in (van der Vaart and van Zanten, 2008) and (Ghosal and van der Vaart, 2007). Indeed, (van der Vaart and van Zanten, 2008) and van der Vaart and van Zanten (2009a) showed that (4.1.3)-(4.1.5) map one-to-one to the conditions of Theorem 1 of (Ghosal and van der Vaart, 2007) over posterior contractions rates.

We make use of the latter theorem in order to bound the expected adaptive local posterior distribution

$$E_0\Pi^{A_k}\left(\theta : \|\theta-\theta_0\|_{n_k} \geq M_n\varepsilon_n|\mathbf{Y}\right) \lesssim \frac{1}{(n/m)\varepsilon_n^2}.$$

Then by triangle inequality

$$\begin{aligned}
&E_0\Pi_{n,m}^A\left(\theta : \|\theta-\theta_0\|_n \geq M_n\varepsilon_n|\mathbf{Y}\right)\\
&\quad\leq \sum_{k=1}^m E_0\Pi_{n,m}^A\left(\theta : \|\theta-\theta_0\|_{n,k} \geq M_n\varepsilon_n|\mathbf{Y}\right)\\
&\quad= \sum_{k=1}^m E_0\Pi^{A_k}\left(\theta : \|\theta-\theta_0\|_{n,k} \geq M_n\varepsilon_n|\mathbf{Y}\right) \lesssim \frac{m^2}{n\varepsilon_n^2} = o(1),
\end{aligned}$$

concluding the proof of the statement.

§4.3.3 Proof of Corollary 4.2.1

We show below that for $\theta_0 \in C^\beta[0,1]$ with $B^2 > a^{2\ell+1} m\varepsilon^{\frac{2(\ell+1-\beta)}{\beta}}$, $a \gtrsim \varepsilon^{\frac{\ell-\beta}{\ell\beta}}$ and $\beta \leq \ell+1$, the following bounds are satisfied for $C_1^*, C_2^* > 0$ not depending on ε, m and n

$$\inf_{h:\|\theta_0-h\|_{\infty,k}\leq\varepsilon} \|h\|^2_{\mathbb{H}^{a,(k)}} \leq C_1^* \frac{a^{-2\ell-1}}{m} \varepsilon^{\frac{-2(\ell+1-\beta)}{\beta}}, \tag{4.3.1}$$

$$-\log \Pi^a\left(\theta : \|\theta\|_{\infty,k} < \varepsilon\right) \leq C_2^* \frac{a}{m} \varepsilon^{\frac{-1}{\ell+1/2}}, \tag{4.3.2}$$

where Π^a is the prior associated with the rescaled GP G_t^a and $\mathbb{H}^{a,(k)}$ is its RKHS on $S^{(k)}$. The local modulus inequality is hence verified for

$$\varepsilon_n \geq K^* \left(a^{\frac{-\beta(\ell+1/2)}{\ell+1}} n^{\frac{-\beta}{2(\ell+1)}} \vee (a/n)^{\frac{\ell+1/2}{2(\ell+1)}} \right),$$

for some large enough constant $K^* > 0$. For $a \asymp n^{\frac{\ell+1/2-\beta}{(\ell+1/2)(2\beta+1)}}$ the above inequality results in $\varepsilon_n \geq K n^{\frac{-\beta}{2\beta+1}}$ for $B^2 > n^{\frac{-1}{2\beta+1}} m$. Therefore, the statement is a direct consequence of Theorem 4.1.1.

Proof of (4.3.1) Let $\theta_0 \in C^\beta([0,1])$ with $\beta \leq \ell+1$. By Whitney's theorem, we can extend $\theta_0 \in C^\beta([0,1])$ to a function $\theta_0 \in C^\beta(\mathbb{R})$ with compact support. We notice, as in the proof of Lemma 11.31 in (Ghosal and Van der Vaart, 2017), that for any ℓth order kernel ψ (an integrable function with $\int \psi(s)ds = 1$, $\int x^l \psi(s)ds = 0$ for all $l \leq \ell$, and $\int |s|^{\ell+1}\psi(s)ds < \infty$), the scaled version $\psi_\sigma(\cdot) := \frac{1}{\sigma}\psi(\frac{\cdot}{\sigma})$ satisfies

$$\sup_{0\leq x\leq 1} |\theta_0 - \theta_0 * \psi_\sigma|(x) \lesssim \sigma^\beta,$$

where $\theta_0 * \psi_\sigma$ is the convolution between θ_0 and ψ_σ. Moreover, for all $\beta \leq l+1$

$$\sup_{0\leq x\leq 1} |(\theta_0 * \psi_\sigma)^{(l+1)}(x)| \lesssim \sigma^{\lfloor\beta\rfloor - l - 1},$$

because $(\theta_0 * \psi_\sigma)^{(l+1)} = \theta_0^{(\lfloor\beta\rfloor)} * \psi_\sigma^{(l+1-\lfloor\beta\rfloor)}$ using the fact that the derivative of the convolution $f * g$ is given as

$$(f * g)' = f' * g = f * g'.$$

Taking $h \in \mathbb{H}^{a,(k)}$ defined as

$$h(t) := I^{\ell+1}\left((\theta_0 * \psi_\sigma)^{(\ell+1)} \mathbf{1}_{[\frac{k-1}{m}, \frac{k}{m}]} \right)(t) + \sum_{j=0}^{\ell} \frac{(t - \frac{k-1}{m})^j}{j!} (\theta_0 * \psi_\sigma)^{(j)}\left(\frac{k-1}{m}\right),$$

where $t \in \left[\frac{k-1}{m}, \frac{k}{m}\right]$, leads to $h^{(\ell+1)} = (\theta_0 * \psi_\sigma)^{(\ell+1)}$ and $h = \theta_0 * \psi_\sigma$ on $[\frac{k-1}{m}, \frac{k}{m}]$. Therefore, h verifies the inequality $\|h - \theta_0\|_{\infty,k} \leq \sigma^\beta$. Moreover, in view of Lemmas 4.3.1 and 4.3.2, the local RKHS norm of h is

$$\|h\|^2_{\mathbb{H}^{a,(k)}} = B^{-2} \sum_{j=0}^{\ell} \frac{h^{(j)}(0)^2}{a^{2j}} + a^{-2\ell-1} \int_{\frac{k-1}{m}}^{\frac{k}{m}} (\theta_0 * \psi_\sigma)^{(\ell+1)}(s)^2 ds.$$

The second part of the right hand-side of this display can be directly bounded from above by $a^{-2\ell-1}\sigma^{-2(\ell+1-\beta)}/m$. On the other hand, the first part is bounded as follows

$$\begin{aligned}\sum_{j=0}^{\ell}\frac{h^{(j)}(0)^2}{a^{2j}} &\leq \sum_{j=0}^{\ell}\frac{1}{a^{2j}}\left(\sum_{r=j}^{\ell}\frac{(-1)^{r-j}(k-1)^{r-j}}{m^{r-j}(r-j)!}(\theta_0*\psi_\sigma)^{(j)}\left(\frac{k-1}{m}\right)\right)^2\\ &\leq \sum_{j=0}^{\ell}\frac{(\ell+1-j)}{a^{2j}}\sum_{r=j}^{\ell}\frac{(k-1)^{2r-2j}}{m^{2r-2j}((r-j)!)^2}(\theta_0*\psi_\sigma)^{(j)}\left(\frac{k-1}{m}\right)^2\\ &\leq \sum_{j=0}^{\ell}\frac{(\ell+1-j)^2}{a^{2j}}\sigma^{2\lfloor\beta\rfloor-2j}\\ &\lesssim (\sigma^{-2(\ell-\beta)}a^{-2\ell}\vee 1).\end{aligned}$$

If we choose $\sigma=\varepsilon^{1/\beta}$, we can deduce that

$$\inf_{h:\|\theta_0-h\|_{\infty,k}\leq\varepsilon}\|h\|^2_{\mathbb{H}^{a,(k)}}\lesssim B^{-2}+\frac{a^{-2\ell-1}}{m}\varepsilon^{\frac{-2(\ell+1-\beta)}{\beta}},$$

when $a\gtrsim\varepsilon^{\frac{\ell-\beta}{\ell\beta}}$. The right-hand side of the display is dominated by $\frac{a^{-2\ell-1}}{m}\varepsilon^{\frac{-2(\ell+1-\beta)}{\beta}}$ when $B^2>a^{2\ell+1}m\varepsilon^{2\frac{(\ell+1-\beta)}{\beta}}$, which concludes the proof.

Proof of (4.3.2) For $\theta\sim B\sum_{j=0}^{\ell}Z_j(at)^j+I^\ell W^a_t$, the centered small ball probability is lower bounded by

$$\begin{aligned}\Pi^a\left(\theta:\|\theta\|_{\infty,k}<\varepsilon\right)&\geq\Pi\left(\|I^\ell W^a\|_{\infty,k}<\varepsilon/2\right)\Pi\left(\|B\sum_{j=0}^{\ell}Z_j(at)^j\|_{\infty,k}<\varepsilon/2\right)\\ &\geq\Pi\left(\|I^\ell W^a\|_{\infty,k}<\varepsilon/2\right)\Pr\left(\sum_{j=0}^{\ell}|Z_j|<\frac{\varepsilon}{2B(a^\ell\vee 1)}\right).\end{aligned}$$

The small ball probability of the ℓ-fold integrated Brownian motion on $S^{(k)}$ can be bounded from below as follows

$$\Pi\left(\sup_{\frac{a(k-1)}{m}\leq t\leq\frac{ak}{m}}|I^\ell W_t|<\varepsilon/2\right)\geq C_k\exp\left\{\frac{-aC}{m\varepsilon^{1/(\ell+1/2)}}\right\}$$

using Lemma 4.3.3 [with $x_0=\frac{a(k-1)}{m}$ and $x_1=\frac{ak}{m}$].

On the other hand, the probability of the sum of absolute values of $(Z_j)_{j=1}^{\ell}$ being smaller than $\varepsilon/(2B(a^\ell\vee 1))$ can be bounded from below by the probability of the absolute value of the $\ell+1$ independent standard normal random variables being

bounded by $\varepsilon/(B(\ell+1)(a^\ell \vee 1))$. Consequently

$$\Pr\left(\sum_{j=0}^{\ell}|Z_j| < \frac{\varepsilon}{2B(a^\ell \vee 1)}\right) \geq 2^{\ell+1}\left(\Phi\left(\frac{\varepsilon}{2B(\ell+1)(a^\ell \vee 1)}\right) - \Phi(0)\right)^{\ell+1}$$
$$\geq 2^{\ell+1}\frac{\exp\left\{\frac{-\varepsilon^2}{8B^2(\ell+1)^2(a^{2\ell}\vee 1)}\right\}}{(2\pi)^{(\ell+1)/2}}\left(\frac{\varepsilon}{B(\ell+1)(a^\ell \vee 1)}\right)^{\ell+1},$$

using the standard lower bound of an integral on a closed set.

Applying the function $x \mapsto -\log(x)$ to the centered small ball probability provides us the following lower bound

$$-\log \Pi^a\left(\theta : \|\theta\|_{\infty,k} < \varepsilon\right) \leq C_1\left(\frac{a}{m}\varepsilon^{\frac{-1}{\ell+1/2}} + \frac{\varepsilon^{\frac{-2(\ell+1)}{\beta}}}{(a^{2\ell}\vee 1)a^{2\ell+1}m} + \log(\varepsilon/a^\ell)\right) + C_2,$$

since $B^2 > a^{2\ell+1}m\varepsilon^{2\frac{(\ell+1-\beta)}{\beta}}$. We note that the upper bound is dominated by the first term when $\varepsilon \geq n^{-L_1}$ and $\varepsilon^{\frac{1}{2\ell+1}-\frac{\ell+1}{\beta}} \leq a^{2\ell+1} \leq n^{L_2}$, for arbitrary $L_1, L_2 > 0$ concludes the proof of the upper bound in (4.3.2).

Lemma 4.3.1. *The RKHS $\mathbb{H}^a$ of the rescaled process G_t^a given in* (4.2.1) *on* $[0,1]$ *is the Sobolev space $S^\ell([0,1])$ with inner product*

$$\langle \vartheta_1, \vartheta_2\rangle_{\mathbb{H}^a} = B^{-2}\sum_{j=0}^{\ell}\frac{\vartheta_1^{(j)}(0)\vartheta_2^{(j)}(0)}{a^{2j}} + a^{-2\ell-1}\int_0^1 \vartheta_1^{(\ell+1)}(s)\vartheta_2^{(\ell+1)}(s)ds.$$

Proof. Lemmas 11.29 and 11.52 of (Ghosal and Van der Vaart, 2017) tell that the RKHS of the integrated Brownian motion part of G_t^a is the subset of the Sobolev space $S^{\ell+1}([0,1])$ of functions ϑ with $\vartheta(0) = ... = \vartheta^{(\ell)}(0) = 0$ under the following inner product

$$\langle \vartheta_1, \vartheta_2\rangle = a^{-2\ell-1}\int_0^1 \vartheta_1^{(\ell+1)}(s)\vartheta_2^{(\ell+1)}(s)ds.$$

In addition, the RKHS of the rescaled polynomial process is the set of ℓth degree polynomials with inner product the Euclidean product of the rescaled coefficients. In other words, for ϑ_1 and ϑ_2 two ℓth degree polynomials in this RKHS, their inner product is

$$\langle \vartheta_1, \vartheta_2\rangle = B^{-2}\sum_{j=0}^{\ell}\frac{\vartheta_1^{(j)}(0)\vartheta_2^{(j)}(0)}{a^{2j}}.$$

The RKHS $\mathbb{H}^a$ of the process G_t^a can then be obtained by applying the general result for the RKHS of a sum of independent Gaussian random elements (Lemma I.18 in (Ghosal and Van der Vaart, 2017)), which concludes the proof. □

Lemma 4.3.2. *Let G_t be a Gaussian process on* $[0,1]$ *with RKHS $\mathbb{H}$. The local RKHS $\mathbb{H}^I$ of the process G_t on $I \subset [0,1]$ is contained in the set of functions $h \in \mathbb{H}$ restricted to I, with the norm*

$$\|h\|_{\mathbb{H}^I} = \inf_{h^* \in \mathbb{H}; h^*(t)=h(t):t\in I}\|h^*\|_{\mathbb{H}}.$$

Proof. Let $z_H \in \mathbb{H}^I$, then according to the definition of the RKHS of a mean zero Gaussian process, there exists a random element $H \in \overline{\text{lin}}(G_t : t \in I)$ such that

$$z_H(t) = EHG_t$$

for all $t \in I$. Since $H \in \overline{\text{lin}}(G_t : t \in I)$, it is also an element of $\overline{\text{lin}}(G_t : t \in [0,1])$ which means there is an element $z_H^* \in \mathbb{H}$ such that

$$z_H^*(t) := EHG_t = z_H(t)$$

for all $t \in I$. Moreover, by the definition of the inner product in the RKHS of a Gaussian process, one can notice that

$$\|z_H\|_{\mathbb{H}^I} = \|z_H^*\|_{\mathbb{H}}.$$

The next step is to show that the norm of any $z_{H^*} \in \mathbb{H}$ such that

$$z_{H^*}(t) = z_H(t)$$

for all $t \in I$ is larger than $\|z_H\|_{\mathbb{H}^I}$. By the definition of the $\mathbb{H}$ norm, we can directly see that

$$\|z_{H^*}\|_{\mathbb{H}} = \sqrt{E(H^*)^2}$$

with $H^* \in \overline{\text{lin}}(G_t : t \in [0,1])$. Since for all $t \in I$, we have $z_{H^*}(t) = z_H(t)$, then

$$E(H^* - H)G_t = 0.$$

This in turn, following from $H \in \overline{\text{lin}}(G_t : t \in I)$, implies that

$$E(H^*)^2 = E(H^* - H)^2 + EH^2 \geq EH^2$$

concluding the proof. □

§4.3.4 Proof of Corollary 4.2.2

In view of Theorem 4.1.2 it is sufficient to construct a sieve $B_{n,m}^{(k)}$ such that the assumptions (4.1.3), (4.1.4) and (4.1.5) hold. The construction of the sieves $B_{n,m}^{(k)}$ for a rescaled integrated Brownian motion is analogous to its counterpart in van der Vaart and van Zanten (2009a) for a squared exponential Gaussian process.

Let us take $K_{n,m}^2 := 8C_0(n/m)\varepsilon_n^2$ and $r_n := n\varepsilon_n^{2+\frac{1}{\ell+1/2}}/D_2$. Define

$$B_{n,m}^{(k)} = \{\theta(x)1_{S^{(k)}}(x) : \theta \in \bigcup_{a<r_n} \left(K_{n,m}\mathbb{H}_1^{a,(k)}\right) + \varepsilon_n \mathbb{B}_1^{(k)}\}, \tag{4.3.3}$$

where $\mathbb{H}_1^{a,(k)}$ is the unit ball in the RKHS of the local rescaled integrated Brownian motion G_{at} on $S^{(k)}$ and $\mathbb{B}_1^{(k)}$ is the unit ball in the Banach space of functions on $S^{(k)}$ equipped with the supremum norm, denoted by $C(S^{(k)})$.

It will be shown below that when $\varepsilon_n \lesssim n^{-1/4}$ and

$$m = o\left(n^{\frac{1}{1+2\beta}} \wedge n^{\frac{\beta}{(\ell+1/2)(1+2\beta)}}\right),$$

which is true for $m = O(n^{\frac{1}{4\ell+2}})$ since $1/2 < \beta \leq \ell + 1/2$, the following bounds are satisfied

$$\log \Pi\left(G^{A_k,(k)} \notin B_{n,m}^{(k)}\right) \leq -K_1(n/m)\varepsilon_n^2, \tag{4.3.4}$$

$$-\log \Pi\left(G^{A_k,(k)} : \|G^{A_k,(k)} - \theta_0\|_{\infty,k} \leq \varepsilon_n\right) \leq K_2 \frac{\varepsilon_n^{-1/\beta}}{m}, \tag{4.3.5}$$

$$\log N\left(\varepsilon_n, B_{n,m}^{(k)}, \|.\|_{\infty,k}\right) \leq K_3(n/m)\varepsilon_n^2, \tag{4.3.6}$$

for $\theta_0 \in C^\beta[0,1]$.

The local conditions on $B_{n,m}^{(k)}$ are hence automatically verified for $\varepsilon_n^{-1/\beta} \leq K^* n\varepsilon_n^2$ for some large enough constant $K^* > 0$. Therefore, the statement is a direct consequence of Theorem 4.1.2.

Proof of (4.3.4) We first consider the local small ball exponent $\phi^{a,(k)}(\varepsilon)$ for the rescaled process G^a

$$\phi^{a,(k)}(\varepsilon) = -\log \Pi^a\left(\theta : \|\theta\|_{\infty,k} < \varepsilon\right),$$

and the corresponding concentration function for θ_0

$$\phi_{\theta_0}^{a,(k)}(\varepsilon) = \inf_{h:\|\theta_0 - h\|_{\infty,k} \leq \varepsilon} \|h\|_{\mathbb{H}^{a,(k)}}^2 - \log \Pi^a\left(\theta : \|\theta\|_{\infty,k} < \varepsilon\right).$$

In view of Borell's inequality,

$$\begin{aligned}\Pi(G^{a,(k)} \notin B_{n,m}^{(k)}) &\leq \Pi(G^{a,(k)} \notin K_{n,m}\mathbb{H}_1^{r_n,(k)} + \varepsilon_n \mathbb{B}_1^{(k)}) \\ &\leq 1 - \Phi(\Phi^{-1}(e^{-\phi^{r_n,(k)}(\varepsilon_n)}) + K_{n,m}).\end{aligned}$$

Let $\varepsilon_n < \rho$ so that $e^{-\phi^{1,(k)}(\rho)} < 1/4$. Seeing as $m = o(\varepsilon_n^{\frac{-1}{\ell+1/2}})$, we can have $\rho \lesssim m^{-\ell-1/2}$ so $\rho \lesssim k^{\ell+1/2-j} m^{-\ell-1/2}$ for all $1 \leq k < m$ and $0 \leq j \leq \ell$; hence in view of Lemma 4.3.3 [with $x_0 = \frac{k-1}{m}$ and $x_1 = \frac{k}{m}$], ρ exists. Consequently, for all $n^{-\frac{(\ell+1/2)}{2(\ell+1)}} < \varepsilon_n < \rho$, we have $r_n > 1$ and thus $e^{-\phi^{r_n,(k)}(\varepsilon_n)} < e^{-\phi^{1,(k)}(\varepsilon_n)} < 1/4$. Furthermore, it is possible to rewrite $K_{n,m}^2$ as follows

$$K_{n,m}^2 = 8C_0 D_2 \frac{r_n \varepsilon_n^{\frac{-1}{\ell+1/2}}}{m}.$$

Similarly to the proof of (4.3.2), we can lower bound for the local centered small ball probability

$$e^{-\phi^{r_n,(k)}(\varepsilon_n)} \geq \Pi\left(\|I^\ell W^{r_n}\|_{\infty,k} < \varepsilon_n/2\right) \Pr\left(\sum_{j=0}^{\ell} |Z_j| < \frac{\varepsilon_n}{2B(r_n^\ell \vee 1)}\right),$$

where the small ball probability of the rescaled ℓ-fold integrated Brownian motion is bounded from below as follows

$$\Pi\left(\sup_{\frac{r_n(k-1)}{m} \leq t \leq \frac{r_n k}{m}} |I^\ell W_t| < \varepsilon_n/2\right) \geq C_k \exp\left\{\frac{-r_n C}{m\varepsilon_n^{1/(\ell+1/2)}}\right\}$$

as long as $m = o(n\varepsilon_n^2)$ and $\varepsilon_n \gtrsim n^{-1/2}$ using Lemma 4.3.3 [with $x_0 = \frac{r_n(k-1)}{m}$ and $x_1 = \frac{r_n k}{m}$], while the probability of the sum of absolute values of $(Z_j)_{j=1}^{\ell}$ being smaller than $\varepsilon_n/(2B(r_n^{\ell} \vee 1))$ is also bounded from below

$$\Pr\left(\sum_{j=0}^{\ell} |Z_j| < \frac{\varepsilon_n}{2B(r_n^{\ell} \vee 1)}\right) \geq 2^{\ell+1} \frac{\exp\left\{\frac{-\varepsilon_n^2}{8B^2(\ell+1)^2(a^{2\ell}\vee 1)}\right\}}{(2\pi)^{(\ell+1)/2}} \left(\frac{\varepsilon_n}{B(\ell+1)(a^{\ell}\vee 1)}\right)^{\ell+1}.$$

Therefore, we can deduce the following

$$\phi^{r_n,(k)}(\varepsilon_n) \leq C_1 \left(\frac{r_n \varepsilon_n^{\frac{-1}{\ell+1/2}}}{m} + \frac{\varepsilon_n^{\frac{-2(\ell+1)}{\beta}}}{(r_n^{2\ell}\vee 1) r_n^{2\ell+1} m} + \log(\varepsilon_n / r_n^{\ell})\right) + C_2,$$

as long as $B^2 > r_n^{2\ell+1} m \varepsilon_n^{2\frac{(\ell+1-\beta)}{\beta}}$. We note that the upper bound is dominated by the first term when $\varepsilon_n \gtrsim n^{-1/2}$ and $\varepsilon_n^{\frac{1}{2\ell+1}-\frac{\ell+1}{\beta}} \leq r_n^{2\ell+1} \leq n^L$, for arbitrary $L > 0$. Hence, we can conclude that

$$K_{n,m}^2 \geq 16\phi^{r_n,(k)}(\varepsilon_n),$$

as long as $m = o(n\varepsilon_n^2)$ and $\varepsilon_n \gtrsim n^{-1/2}$. In accordance to Lemma 4.4.4 [with $u = \exp\{-\phi^{r_n(k)}(\varepsilon_n)\}$], the inequality $K_{n,m}^2 \geq -2\Phi^{-1}(\exp\{-\phi^{r_n,(k)}(\varepsilon_n)\})$ holds; thus,

$$\begin{aligned}\Pi(G^{a,(k)} \notin B_{n,m}^{(k)}) &\leq 1 - \Phi(K_{n,m}/2) \\ &\leq e^{-K_{n,m}^2/8},\end{aligned}$$

using Lemma 4.4.5. The local remaining masses can then be bounded from above as follows,

$$\begin{aligned}\Pi(G^{A_k,(k)} \notin B_{n,m}^{(k)}) &\leq P(A_k > r_n) + \int_0^{r_n} \Pi(G^{a,(k)} \notin B_{n,m}^{(k)}) g(a) da \\ &\leq 2C_2 r_n^{p+\frac{1}{2(\ell+1)}} e^{-D_2 n^{\frac{1}{2(\ell+1)}} r_n^{\frac{\ell+1/2}{\ell+1}}/m} + e^{-K_{n,m}^2/8} \leq e^{-c(n/m)\varepsilon_n^2},\end{aligned}$$

for some $c > 0$ small enough in view of Lemma 4.4.3. Thus, to conclude it suffices that $m = o\left(\varepsilon_n^{\frac{-1}{\ell+1/2}} \wedge n\varepsilon_n^2\right)$, so that the inequality (4.3.4) be satisfied.

Proof of (4.3.6) Lemma 11.52 of (Ghosal and Van der Vaart, 2017) [with $\alpha = \ell + 1/2$] implies that the set $\mathbb{H}_1^a$ can be rewritten as $a^{\ell+1/2}\mathbb{H}_1$; hence $\bigcup_{a<r_n}(K_{n,m}\mathbb{H}_1^{a,(k)}) = r_n^{\ell+1/2} K_{n,m}\mathbb{H}_1^{(k)} = \mathbb{H}_1^{r_n,(k)}$, and we get

$$N\left(2\varepsilon_n, \bigcup_{a<r_n}(K_{n,m}\mathbb{H}_1^{a,(k)} + \varepsilon_n \mathbb{B}_1^{(k)}), \|.\|_{\infty,k}\right) \leq N\left(\varepsilon_n, K_{n,m}\mathbb{H}_1^{r_n,(k)}, \|.\|_{\infty,k}\right).$$

Since the function $x \mapsto \log(x)$ is increasing on its domain, and in view of Lemma 4.4.2

$$\begin{aligned}
\log N\left(2\varepsilon_n, \bigcup_{a<r_n}(K_{n,m}\mathbb{H}_1^{a,(k)} + \varepsilon_n \mathbb{B}_1^{(k)}), \|.\|_{\infty,k}\right) &\lesssim \frac{1}{m}\Big(\frac{r_n^{\ell+1/2}K_{n,m}}{\varepsilon_n}\Big)^{1/(\ell+1)} \\
&= \frac{1}{m}\left(\left(n\varepsilon_n^{2+\frac{1}{\ell+1/2}}\right)^{\ell+1/2}\sqrt{n/m}\right)^{1/(\ell+1)} \\
&= \frac{1}{m}\left(m^{-1/2}n^{\ell+1}\varepsilon_n^{2\ell+2}\right)^{1/(\ell+1)} \\
&= m^{\frac{-2\ell-3}{2(\ell+1)}} n\varepsilon_n^2 \\
&\lesssim (n/m)\varepsilon_n^2,
\end{aligned}$$

since $m = o(n\varepsilon_n^2)$; hence $(n/m)\varepsilon_n^2 \to +\infty$ and $\varepsilon_n/(K_{n,m}r_n^{\ell+1/2}) \downarrow 0$. From this display, we have that the assertion (4.3.6) is verified.

Proof of (4.3.5) Suppose that $\theta_0 \in C^\beta([0,1])$. In view of assertion (4.3.1), there exists $C_1^* > 0$ such that

$$\inf_{h:\|\theta_0-h\|_{\infty,k}\leq\varepsilon}\|h\|^2_{\mathbb{H}^{a,(k)}} \leq C_1^*\frac{a^{-2\ell-1}}{m}\varepsilon^{\frac{-2(\ell+1-\beta)}{\beta}},$$

provided that $B^2 > a^{2\ell+1}m\varepsilon^{\frac{2(\ell+1-\beta)}{\beta}}$ and $a \gtrsim \varepsilon^{\frac{\ell-\beta}{\ell\beta}}$. Additionally, (4.3.2) tells that there exists $C_2^* > 0$ such that

$$\phi^{a,(k)}(\varepsilon) \leq C_2^*\frac{a}{m}\varepsilon^{\frac{-1}{\ell+1/2}}.$$

Therefore,

$$\begin{aligned}
\phi_{\theta_0}^{a,(k)}(\varepsilon) &= \inf_{h:\|\theta_0-h\|_{\infty,k}\leq\varepsilon}\|h\|^2_{\mathbb{H}^{a,(k)}} + \phi^{(k)}(a^{-\ell-1/2}\varepsilon) \\
&\leq \frac{1}{m}\left(C_1^* a^{-2\ell-1}\varepsilon^{\frac{-2(\ell+1-\beta)}{\beta}} + C_2^* a\varepsilon^{\frac{-1}{\ell+1/2}}\right).
\end{aligned}$$

By taking a which verifies

$$\varepsilon^{\frac{1}{\ell+1/2}-\frac{1}{\beta}} \leq a \leq 2\varepsilon^{\frac{1}{\ell+1/2}-\frac{1}{\beta}},$$

one can find a constant $C_0 > 0$ such that the local concentration function is bounded from above as follows

$$\phi_{\theta_0}^{a,(k)}(\varepsilon) \leq C_0\frac{\varepsilon^{-1/\beta}}{m}.$$

Let $\alpha := \frac{1}{\beta} - \frac{1}{\ell+1/2}$ so that there exists a constant $C > 0$ such that,

$$\begin{aligned}
\Pi^A(\|G^A - \theta_0\|_{\infty,k} \leq 2\varepsilon_n) &\geq \int_0^\infty e^{-\phi_{\theta_0}^{a,(k)}(\varepsilon_n)}g(a)da \\
&\geq \int_{\varepsilon_n^{-\alpha}}^{2\varepsilon_n^{-\alpha}} \exp\big\{-c_0\varepsilon_n^{-1/\beta}/m\big\}g(a)da.
\end{aligned}$$

The integral on the right-hand side is larger than the infimum of the integrand multiplied by $\varepsilon_n^{-\alpha}$; hence, if we apply the function $x \mapsto -\log(x)$ to both sides of the display, we obtain the following bound

$$-\log \Pi^A(\|G^A - \theta_0\|_{\infty,k} \leq 2\varepsilon_n) \lesssim \frac{\varepsilon_n^{-1/\beta}}{m} + \frac{n^{\frac{1}{2(\ell+1)}} \varepsilon_n^{-\alpha\frac{(\ell+1/2)}{\ell+1}}}{m} + \log \varepsilon_n.$$

which concludes the proof of (4.3.5) when $\varepsilon_n \lesssim n^{\frac{-\beta}{1+2\beta}}$.

Lemma 4.3.3. *Let $(I^\ell W_t : t \in [x_0, x_1])$ be the ℓ-fold primitive of Brownian motion W_t on $[x_0, x_1]$ with $0 < x_0 < x_1 \leq T$, then the process $(I^\ell W_t : t \in [x_0, x_1])$ has the same distribution as $(I^\ell W_{t+x_0} : t \in [0, x_1 - x_0])$. Moreover,*

$$I^\ell W_{t+x_0} = I^\ell W_t^* + \sum_{j=0}^{\ell} \frac{t^j}{j!} I^{\ell-j} W_{x_0},$$

with W_t^* a Brownian motion independent from W_t, and for any ε verifying

$$\varepsilon \leq \frac{(\ell+1)(x_1 - x_0)^j x_0^{\ell+1/2-j}}{\ell!\sqrt{2\ell+1}},$$

for all $0 \leq j \leq \ell$, we obtain the following bounds

$$\frac{-q_2(x_1 - x_0)}{\varepsilon^{\frac{1}{\ell+1/2}}} \leq \log Pr\left(\sup_{x_0 \leq t \leq x_1} |I^\ell W_t| < \varepsilon\right) \leq \frac{-q_1(x_1 - x_0)}{\varepsilon^{\frac{1}{\ell+1/2}}}$$

when $x_1 - x_0 \leq x_0$, where the constants $q_1, q_2 > 0$ depend only on ℓ.

Proof. For W_t a standard Brownian motion, let $W_t^* := W_{x_0+t} - W_{x_0}$ for a given $x_0 > 0$, then the process $(W_t^*)_{t \geq 0}$ is also a standard Brownian motion independent from $(W_t)_{0 \leq t \leq x_0}$ because Brownian motions are translation invariant.

In order to prove the first part of the lemma, it is possible to first show by induction that

$$I^\ell W_t^* = I^\ell W_{x_0+t} - \sum_{j=0}^{\ell} \frac{t^j}{j!} I^{\ell-j} W_{x_0}.$$

Indeed, the identity is trivial for $\ell = 0$. Moreover, if the claim is true for ℓ, then

$$\begin{aligned} I^{\ell+1} W_t^* &= \int_0^t I^\ell W_s^* ds \\ &= \int_0^t I^\ell W_{x_0+s} ds - \sum_{j=0}^{\ell} \int_0^t \frac{s^j}{j!} I^{\ell-j} W_{x_0} ds \\ &= I^{\ell+1} W_{x_0+t} - I^{\ell+1} W_{x_0} - \sum_{j=0}^{\ell} \frac{t^{j+1}}{(j+1)!} I^{\ell-j} W_{x_0} = I^{\ell+1} W_{x_0+t} - \sum_{j=0}^{\ell+1} \frac{t^j}{j!} I^{\ell+1-j} W_{x_0}, \end{aligned}$$

which confirms the claim for $\ell + 1$.

Using this newfound identity, one can rewrite the small ball probability of $(I^\ell W_t : t \in [x_0, x_1])$ as follows

$$\Pr\left(\sup_{x_0\leq t\leq x_1}|I^\ell W_t|<\varepsilon\right)=\Pr\left(\sup_{0\leq t\leq x_1-x_0}\left|I^\ell W_t^*+\sum_{j=0}^{\ell}\frac{t^j}{j!}I^{\ell-j}W_{x_0}\right|<\varepsilon\right).$$

Then, we can use Anderson's theorem and the self-similarity of $I^\ell W_t$ with index $\ell+1/2$ in order to bound this probability from above

$$\begin{aligned}\Pr\left(\sup_{x_0\leq t\leq x_1}|I^\ell W_t|<\varepsilon\right)&\leq\Pr\left(\sup_{0\leq t\leq 1}|I^\ell W^*_{(x_1-x_0)t}|<\varepsilon\right)\\&=\Pr\left(\sup_{0\leq t\leq 1}|I^\ell W_t^*|<(x_1-x_0)^{-\ell-1/2}\varepsilon\right).\end{aligned}$$

Hence, there exists a $q_1>0$ such that

$$\log\Pr\left(\sup_{x_0\leq t\leq x_1}|I^\ell W_t|<\varepsilon\right)\leq\frac{-q_1(x_1-x_0)}{\varepsilon^{\frac{1}{\ell+1/2}}}.$$

Besides, one can bound the small ball probability of $(I^\ell W_t : t \in [x_0, x_1])$ from below as follows

$$\begin{aligned}\Pr\left(\sup_{x_0\leq t\leq x_1}|I^\ell W_t|<\varepsilon\right)&=\Pr\left(\sup_{0\leq t\leq x_1-x_0}\left|I^\ell W_t^*+\sum_{j=0}^{\ell}\frac{t^j}{j!}I^{\ell-j}W_{x_0}\right|<\varepsilon\right)\\&\geq\Pr\left(\sup_{0\leq t\leq x_1-x_0}|I^\ell W_t^*|<\varepsilon/2\right)\times\\&\quad\Pr\left(\sup_{0\leq t\leq x_1-x_0}\left|\sum_{j=0}^{\ell}\frac{t^j}{j!}I^{\ell-j}W_{x_0}\right|<\varepsilon/2\right).\end{aligned}$$

The logarithm of the first part can be bounded from below by a constant multiplier of $-(x_1-x_0)\varepsilon^{\frac{-1}{\ell+1/2}}$ because $I^\ell W^*$ is self-similar with index $\ell+1/2$. On the other hand, one can bound the probability in the second part as follows

$$\begin{aligned}\Pr\left(\sup_{0\leq t\leq x_1-x_0}\left|\sum_{j=0}^{\ell}\frac{t^j}{j!}I^{\ell-j}W_{x_0}\right|<\varepsilon\right)&\geq\Pr\left(\sum_{j=0}^{\ell}\sup_{0\leq t\leq x_1-x_0}\frac{t^j}{j!}\left|I^{\ell-j}W_{x_0}\right|<\varepsilon\right)\\&\geq\prod_{j=0}^{\ell}\Pr\left(\frac{(x_1-x_0)^j}{j!}\left|I^{\ell-j}W_{x_0}\right|<\varepsilon/(\ell+1)\right).\end{aligned}$$

The variance of the elements $I^{\ell-j}W_{x_0}$ for $0\leq j\leq\ell$ can be derived as follows

$$\begin{aligned}&\operatorname{Var}\left(I^{\ell-j}W_{x_0}\right)=E\left(I^{\ell-j}W_{x_0}\right)^2\\&=\frac{1}{(\ell-j-1)!^2}\int_0^{x_0}\int_0^{x_0}(x_0-s)^{\ell-j-1}(x_0-u)^{\ell-j-1}EW_sW_u\,du\,ds\end{aligned}$$

using the Cauchy formula for repeated integration and Fubini's theorem. The expectation inside the integral is simply the covariance kernel of the Brownian motion $\min(s,u)$. Without loss of generality, let's compute this integral when $u \leq s$

$$
\begin{aligned}
&2\int_0^{x_0}\int_0^{s}(x_0-s)^{\ell-j-1}(x_0-u)^{\ell-j-1}u\,du\,ds \\
&= 2\int_0^{x_0}(x_0-s)^{\ell-j-1}\left(\frac{x_0^{\ell+1-j}}{(\ell+1-j)(\ell-j)}-\frac{x_0(x_0-s)^{\ell-j}}{(\ell+1-j)(\ell-j)}-\frac{s(x_0-s)^{\ell-j}}{\ell+1-j}\right)ds \\
&=: I_1+I_2+I_3.
\end{aligned}
$$

The integrals I_1, I_2 and I_3 can be computed straightforwardly:

$$
\begin{aligned}
I_1 &= \frac{2x_0^{2\ell+1-2j}}{(\ell+1-j)(\ell-j)^2} \\
I_2 &= \frac{-x_0^{2\ell+1-2j}}{(\ell+1-j)(\ell-j)^2} \\
I_3 &= \frac{-x_0^{2\ell+1-2j}}{(2\ell+1-2j)(\ell+1-j)(\ell-j)}.
\end{aligned}
$$

Consequently, the total integral is

$$
2\int_0^{x_0}\int_0^{x_0}(x_0-s)^{\ell-j-1}(x_0-u)^{\ell-j-1}u\,du\,ds = \frac{x_0^{2\ell+1-2j}}{(2\ell+1-2j)(\ell-j)^2},
$$

and the variance of the integrated Brownian motion at point x_0 is

$$
\operatorname{Var}\left(I^{\ell-j}W_{x_0}\right) = \frac{x_0^{2\ell+1-2j}}{(2\ell+1-2j)(\ell-j)!^2}.
$$

Accordingly, it is possible to obtain the following bound for all $0 \leq j \leq \ell$

$$
\Pr\left(\frac{(x_1-x_0)^j}{j!}\left|I^{\ell-j}W_{x_0}\right| < \varepsilon/(\ell+1)\right) \geq 2\Phi\left(\frac{\varepsilon j!(\ell-j)!\sqrt{2\ell+1-2j}}{(\ell+1)(x_1-x_0)^j x_0^{\ell+1/2-j}}\right)-1.
$$

It can be verified that the function given by $x \mapsto \log(2\Phi(x)-1)$ is increasing from $-\infty$ to 0, and that $\log(2\Phi(x)-1) \geq -1-|\log x|$ when x is small enough. This implies that the log-probability of interest is bounded from below by

$$
-1-\left|\log\left(\frac{\varepsilon j!(\ell-j)!\sqrt{2\ell+1-2j}}{(\ell+1)(x_1-x_0)^j x_0^{\ell+1/2-j}}\right)\right|.
$$

Furthermore, seeing as $-1-|\log x| \geq -1/x$ for small values of x and because

$$
\varepsilon \leq \frac{(\ell+1)(x_1-x_0)^j x_0^{\ell+1/2-j}}{\ell!\sqrt{2\ell+1}} \leq \frac{(\ell+1)(x_1-x_0)^j x_0^{\ell+1/2-j}}{j!(\ell-j)!\sqrt{2\ell+1-2j}}
$$

for all $0 \leq j \leq \ell$, we have that

$$\log \Pr\left(\frac{(x_1-x_0)^j}{j!}\left|I^{\ell-j}W_{x_0}\right| < \varepsilon/(\ell+1)\right) \geq -1 - \left|\log\left(\frac{\varepsilon j!(\ell-j)!\sqrt{2\ell+1-2j}}{(\ell+1)(x_1-x_0)^j x_0^{\ell+1/2-j}}\right)\right|.$$
$$\gtrsim \frac{-(x_1-x_0)^j x_0^{\ell+1/2-j}}{\varepsilon}$$
$$\gtrsim \frac{-(x_1-x_0)}{\varepsilon^{\frac{1}{\ell+1/2}}}.$$

The last inequality results from the fact that ε is of smaller order than

$$\varepsilon \lesssim (x_1-x_0)^j x_0^{\ell+1/2-j}$$
$$\lesssim (x_1-x_0)^{\frac{(j-1)(\ell+1/2)}{\ell-1/2}} x_0^{\frac{(\ell+1/2-j)(\ell+1/2)}{(\ell-1/2)}},$$

for all $0 \leq j \leq \ell$ when $x_1 - x_0 \leq x_0$. We conclude the proof by noting that the logarithm of a product is the sum of the logarithm of the terms.

□

§4.4 Auxiliary Lemmas

Lemma 4.4.1 (Proposition 11.19 of (Ghosal and Van der Vaart, 2017))**.** *For any mean zero Gaussian random element G in a separable Banach space, any θ in the closure of its RKHS and any $\varepsilon > 0$*

$$\phi_\theta(\varepsilon) \leq -\log \Pi(\|G-\theta\| < \varepsilon) \leq \phi_\theta(\varepsilon/2),$$

where the norm is taken as the norm of the Banach space.

Lemma 4.4.2 (Proposition C.7 of (Ghosal and Van der Vaart, 2017))**.** *For $M > 0$, $k \geq 0$ and $a > 0$, let $\mathbb{H}_1^a$ the unit ball in the RKHS of G^a on $\mathcal{X}$ where G is defined as in* (4.2.1)*, then there exists a constant K such that*

$$\log N(\varepsilon, M\mathbb{H}_1^a, \|.\|_\infty) \leq K\, Vol(\mathcal{X}) a^{\frac{\ell+1/2}{\ell+1}} \left(\frac{M}{\varepsilon}\right)^{\frac{1}{\ell+1}}.$$

Lemma 4.4.3. *If the random variable A has a density $g_{\ell,n,m}$ that satisfies* (4.2.2)*, then for $a^{\frac{\ell+1/2}{\ell+1}} > 2mn^{\frac{-1}{2(\ell+1)}}|1-2(\ell+1)p|/(D_2(2\ell+1)$ and $a > e$,*

$$P(A > a) \leq 2C_2 a^{p+\frac{1}{2(\ell+1)}} e^{-D_2 n^{\frac{1}{2(\ell+1)}} a^{\frac{\ell+1/2}{\ell+1}}/m}$$

Proof. As this lemma is analogous to Lemma 4.9 of (van der Vaart and van Zanten, 2009b), the proof follows the same steps. Set $j_p(a) = a^p \exp\{-D_2 n^{\frac{1}{2(\ell+1)}} a^{\frac{\ell+1/2}{\ell+1}}/m\}$ and $J_p(a) = \int_a^{+\infty} j_p(t)dt$ with $p \geq 0$. The derivative of the function j_p can, with the help of the chain rule, be expressed as the sum of two terms. By integrating this identity we see that

$$j_p(a) = \frac{D_2 n^{\frac{1}{2(\ell+1)}} \frac{(\ell+1/2)}{\ell+1}}{m} J_{p-\frac{1}{2(\ell+1)}}(a) - pJ_{p-1}(a).$$

The first term on the right is non-negative, while the second is negative if and only if $p > 0$. By the transformation $p - \frac{1}{2(\ell+1)} \to p$ we conclude that

$$\frac{D_2 n^{\frac{1}{2(\ell+1)}} \frac{(\ell+1/2)}{\ell+1}}{m} J_p(a) - \left| p + \frac{1}{2(\ell+1)} \right| J_{p-\frac{\ell+1/2}{\ell+1}}(a) = j_{p+\frac{1}{2(\ell+1)}}(a).$$

Here

$$J_{p-\frac{\ell+1/2}{\ell+1}}(a) = \int_a^{+\infty} t^{\frac{-(\ell+1/2)}{\ell+1}} j_p(t) dt \leq a^{\frac{-(\ell+1/2)}{\ell+1}} J_p(a).$$

By substituting this inequality in the left-hand side and rearranging we obtain the bound

$$\left(\frac{D_2 n^{\frac{1}{2(\ell+1)}} \frac{(\ell+1/2)}{\ell+1}}{m} - a^{\frac{-(\ell+1/2)}{\ell+1}} \left| p - \frac{1}{2(\ell+1)} \right| \right) J_p(a) \leq j_{p+\frac{1}{2(\ell+1)}}(a)$$

on $P(A > a) \leq C_2 J_p(a)$ asserted by the lemma. □

The following lemmas are standard results from the literature.

Lemma 4.4.4. *For Φ the standard normal cumulative distribution function, $\Phi^{-1}(u) \geq -\sqrt{2\log(1/u)}$ for $u \in (0,1)$ and $\Phi^{-1}(u) \leq -1/2\sqrt{\log(1/u)}$ for $u \in (0, 1/2)$.*

Lemma 4.4.5. *For Φ the standard normal cumulative distribution function, $1 - \Phi(u) \geq e^{-u^2/2}$ for $u > 0$.*

CHAPTER 5
Simulation study

Abstract. In this chapter we investigate the numerical properties of the different Gaussian process regression techniques using distributed methods. Distributed methods use a divide-and-conquer strategy: the supposedly large data set is divided among m machines. This strategy helps reducing the computational costs of the typical Bayesian non-parametric regression. It should be noted that there exist multiple ways of partitioning the data among the machines

§5.1 Distributed GP regression

Gaussian process regression is arguably a very useful tool in machine learning since it can elegantly capture complex relationships in data (Rasmussen and Williams, 2006). However, it scales very poorly in computation and memory ($O(n^3)$ and $O(n^2)$ respectively, where n is the number of data points). This limitation inspired different approximation approaches, among which the *divide-and-conquer* strategy where the design is partitioned into m "expert" machines; then, the kth partition, with $k \in \{1, ..., m\}$ of size n_k is modeled by the "expert" to which it was assigned. Different models arose according to the way the data is allotted to the expert machines. A uniformly random partition model (see (Cao and Fleet, 2014) and (Tresp, 2000)) are built by allotting each machine a random subset of the data of size n/m in order to independently compute predictive distributions which will be aggregated. These have been shown to not only be Kolmogorov inconsistent (Samo and Roberts, 2016), but also to have posterior which contracts at a sub-optimal rates (Szabó and van Zanten, 2019).

On the other hand, spatial partition models are based on a division of the design space into non-overlapping region. Each machine is assigned a specific region and inference is made using the data in this region. For instance, the Naive-Local-Experts model (Kim et al., 2005) models each region with an independent GP. Its main drawback is the introduction of discontinuities in the prediction at the border of each region. There exist multiple ways to address the issue. Patched GPs (see (Park and Huang, 2016) and (Park and Apley, 2018)) for instance impose continuity constraints such that two adjacent local GPs are patched to share the nearly identical predictions on the boundary. Two-step Mixtures introduce a latent variable to the model which dynamically selects an expert to draw prediction on a given point (Tresp, 2001), (Rasmussen and Ghahramani, 2002), (Meeds and Osindero, 2006). Recently, hierarchical spatial partitioning models (Ng and Deisenroth, 2014) have been developed. They

typically result in posterior predictive distributions in the form of an average of all the "expert" predictions with a weight supposed to indicate how the confidence level of each prediction.

§5.1.1 Uniformly random

The data can be randomly split among the m machines. Each machine will receive a random sub-sample of the data, and will therefore solve a smaller version of the initial regression by computing a local posterior. The different posteriors will then be aggregated to form a global posterior. As seen in (Szabó and van Zanten, 2019), some adjustments should be made locally in order to obtain theoretical guaranties for this method. In this simulation study, we chose to adjust the local prior by raising it to the power $1/m$ and to average the local posteriors. However, (Szabó and van Zanten, 2019) also shows that despite the modifications on the local prior, adaptation leads to sub-optimal contraction rates and bad coverage for some true functions.

§5.1.2 Spatial

The data can also be split into subsets of the design point set. Each machine will receive data such that the design points belong to a certain sub-region. The sub-regions are not to overlap. The machine can then be seen as local experts; each expert is specialized in one particular sub-region of design points. A draw from the global posterior will thus consist of the local posterior draws restricted to their corresponding intervals and pasted together. Due to the localized structure, there is no need for alterations in the local prior. Moreover, this structure allows the local posterior to adapt to the unknown smoothness as we showed in Chapter 4. Unfortunately, the global posterior obtained by this procedure contains unwanted discontinuities at the border of each regions.

§5.1.3 Weighted-average model

One can note that both global posteriors produce samples in the form of a weighted average of the local samples. In the first scenario, a global posterior draw θ is defined as

$$\theta(x) = \frac{1}{m} \sum_{k=1}^{m} \theta^{(k)}(x),$$

for all $x \in \mathcal{X}$ where the $\theta^{(k)}$'s are local draws. In the second scenario, a global posterior draw can be written as

$$\theta(x) = \sum_{k=1}^{m} 1_{\mathcal{D}_k}(x) \theta^{(k)}(x),$$

where $\mathcal{D}_k$'s are the sub-regions into which the design points are partitioned. This observation explains the discontinuities in the latter case, since the weights are discontinuous themselves. In order to palliate this problem, we propose using data-driven weight functions which are both continuous and close to indicator functions. One can

find in the literature (Ng and Deisenroth, 2014) weights proportional to the inverse variances:

$$\theta(x) = \Sigma(x) \sum_{k=1}^{m} \frac{\theta^{(k)}(x)}{\sigma_k^2(x)},$$

where the σ_j^2's are the local posterior variance and $\Sigma(x) = \left(\sum_{k=1}^{m} \sigma_k^{-2}(x)\right)^{-1}$. Although these weights are data-driven and continuous, we will see that the corresponding global posterior exhibit sub-optimal asymptotic characteristics in an adaptive setting. Indeed, adapting to local smoothnesses may lead to shrinking variances for some machines, which in turn will lead the corresponding weight to be overly large even outside of the expert's domain. That is to say that experts are overly confident about the behavior of the true function in the whole space when this function is particularly smooth in this expert's domain. That is why we propose a modification of the weight functions so that they shrink quickly outside of their corresponding region minimizing the behavior of indicator functions. Namely, we choose

$$\theta(x) = W(x) \sum_{k=1}^{m} w_k(x)\theta^{(k)}(x),$$

where $w_k(x) = e^{-m^2(x-c_k)^2}/\sigma_k^2(x)$ with c_k being the center of gravity of $\mathcal{D}_k$, and $W(x) = \left(\sum_{k=1}^{m} w_k(x)\right)^{-1}$. These weights are both continuous and data-driven.

§5.2 Numerical study

§5.2.1 Simulated Data

First, we consider Gaussian process regression with simulated data that will allow us to compare the different distributed techniques to the non-distributed Gaussian regression, which will act as a benchmark. We see that if the smoothness of the true function is known and the Gaussian process parameters are chosen accordingly, all proposed distributed methods behave similarly. They present comparable L_2 distance between the true function and the posterior mean, and they portray similar coverage for their point-wise credible sets. Moreover, if the number of machines m does not grow too fast, these distributed methods are also similar to our benchmark. On the other hand, we show that in the adaptive setting, the way the data is distributed among the machines affect greatly the performance of the regression.

In this model we assume to observe n independent pairs of random variables $(X_1, Y_1), ...(X_n, Y_n)$, where

$$Y_i = \theta_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad X_i \overset{iid}{\sim} U(0, 1),$$

and the aim is to estimate the unknown non-parametric regression function θ_0. In the Bayesian approach we endow θ_0 with a Gaussian process prior with squared exponential kernel and estimate the tuning parameter using the MMLE. In addition, we wish to reduce the computational time by distributing the n pairs of random variables among m machines such that each machine only deals with the independent

pairs $(X_1^{(k)}, Y_1^{(k)}), ...(X_{n_k}^{(k)}, Y_{n_k}^{(k)})$ which represent a subset of the original sample and $n_k = n/m$ is the number of random variable pairs in the corresponding machine. All the posterior means, credible sets, and the empirical Bayes posteriors in the adaptive setting are all computed using the MatLab package gpml. Let us consider the function $\theta_0 \in L_2[0,1]$ given by the coefficients $\theta_{0,i} = i^{-3/2}\sin(i)$, for $i \geq 3$ and $\theta_{0,i} = 0$ otherwise, relative to the cosine eigenbasis $\psi_i(t) = \sqrt{2}\cos(\pi(i-1/2)t)$. Note that although the function lies outside of the self-similar function class, it has essentially the same behavior.

We take $\sigma^2 = 1$, but in the procedure it is considered to be unknown and estimated with the MMLE $\hat{\sigma}^2$. We plot in figures the true function (black), the posterior mean (colored), and the posterior point-wise credible intervals (shaded area) $[\hat{\theta}(x) - q_{0.025}\sqrt{\hat{c}(x,x)}, \hat{\theta}(x) + q_{0.025}\sqrt{\hat{c}(x,x)}]$, where $\hat{\theta}$ is the posterior mean, q_α the α-th quantile of the standard normal distribution and $\hat{c}(.,.)$ the posterior covariance kernel. We consider the non-distributive method (at the top) along with the four distributed methods proposed. We will compare the methods in different figures depending on the setting (non-adaptive or adaptive), the sample size ($n = 100, 500, 1000$ or 2000) and the number of machines ($m = 10$, $m \approx n^{1/3}$ or $m = n/100$).

We also investigate empirically the rate at which the posterior mean concentrates around the truth and the frequentist coverage probabilities of the point-wise credible sets by repeating the experiment 100 times and reporting the average integrated mean squared error and the frequency that the function at given points (we consider $x = (0.5, 0.41148, 0.31143)$ with $0.41148 = \text{argmin}_{x\in[0,1]}\theta_0(x)$ and $0.31143 = \text{argmax}_{x\in[0,1]}\theta_0(x)$) is included in the credible interval. See Tables 5.1afor the average L_2-norm between the posterior mean and the true function, and see Tables 5.2for the frequentist coverage of the credible sets.

The different methods we study are summed up in this table:

Method	*Description*
1	uniformly random partitioning + adjusting the prior with power $1/m$
2	spatial partitioning
3	spatial partitioning + inverse local post variance weights
4	spatial partitioning + inverse centered squared-exponential weights

Figure 5.1 illustrates that when m increases at sub-linear rate with n, the global posterior means obtained via the different methods are similar in a non-adaptive setting. Besides, these global posteriors means look similar to the traditional posterior mean. Nonetheless, the global posteriors as wholes do not behave similarly. **Method 2**, for instance, produces visible discontinuous predictions on the boundaries of subregions One can also see in Table 5.4 that global posteriors of distributed Bayesian regression take substantially less time to compute. It should also be taken into account that the all computations have been done sequentially; the parameters of all local posteriors have been computed and stored in the same computer. This may explain why adding more experts does not necessarily decreases the computation time. In practice, one can imagine that these running times might be reduced using multiple machines or cores, and that the effective time of the operation would roughly equal the present computation time divided by the number of machines.

Observe in Table 5.1a that the posterior mean in all distributed methods concen-

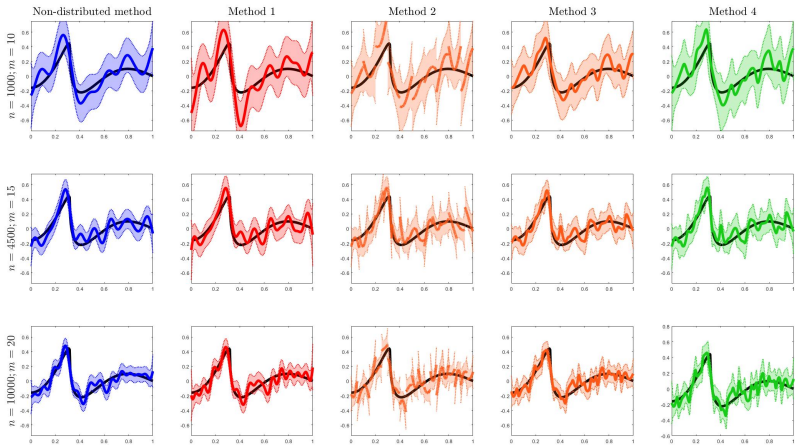

Figure 5.1: Non-adaptive posterior density for the function θ_0 (drawn in black) on $x \in [0,1]$. The posterior means are drawn by solid line, while the 95% point-wise credible sets are shaded between two dotted lines. In the first column, we plot the non-distributed method, in the second column the distributed method with random partitioning, the third column the distributed method with spatial partitioning, while in the fourth and fifth column we plot the distributed method with spatial partitioning with inverse variance weights and exponential weights respectively. From top to bottom the sample size is $n = 1000, 5000, 10000$ and the number of experts is $m = 10, 15, 20$.

	$n =$	1000	4500	10000	1000	4500	10000
	BM	**39.0041**	**23.5635**	**17.0517**	**49.9397**	**28.4795**	**20.3814**
$m = 10$	M 1	48.0590	26.1719	18.6091	64.2480	30.9439	22.2252
	M 2	47.5648	27.4998	19.5424	67.2354	31.2043	23.6621
	M 3	36.5442	24.0340	17.8270	56.6046	28.7473	22.2875
	M 4	42.1584	25.5088	18.3610	55.1067	29.5429	22.5838
$m \approx n^{1/3}$	M 1	48.0590	27.0015	19.2581	64.2480	31.9520	23.1445
	M 2	47.5648	29.4000	22.1400	67.2354	33.4671	26.8086
	M 3	36.5442	24.3732	18.6893	56.6046	29.5770	23.6461
	M 4	42.1584	26.7001	20.0192	55.1067	30.9875	24.7048
$m = \frac{n}{100}$	M 1	48.0590	36.5080	35.7912	64.2480	61.4936	84.4595
	M 2	47.5648	37.7861	32.7527	67.2354	43.1683	39.6620
	M 3	36.5442	24.9314	23.6126	56.6046	30.5618	23.9152
	M 4	42.1584	31.7063	25.9903	55.1067	36.9440	32.3062

(a) Average L_2 distance between θ_0 and the posterior mean for the squared exponential Gaussian process prior in a non-adaptive setting.

(b) Average L_2 credible ball radius for the squared exponential Gaussian process prior in a non-adaptive setting.

Table 5.1: BM: Benchmark, Non-distributed method. M 1: Random partitioning, M 2: Spatial partitioning, M 3: Spatial partitioning with inverse variance weights, M 4: Spatial partitioning with exponential weights. From left to right the sample size is $n = 1000, 4500, 10000$.

trates around the true function at a similar rate in a non-adaptive setting. When $m \lesssim \sqrt[3]{n}$, the contraction rate of the distributed posterior mean is virtually the same as the non-distributed case. However, as soon as the number of machine increases

		$x = 0.5$			$x = 0.41148$			$x = 0.31143$		
	$N =$	1000	4500	10000	1000	4500	10000	1000	4500	10000
$m = 10$	Benchmark	**0.99**	**0.96**	**0.95**	**0.96**	**0.91**	**1.00**	**0.57**	**0.48**	**0.30**
	Method 1	0.98	0.97	0.94	0.93	0.92	0.99	0.73	0.62	0.43
	Method 2	0.97	0.99	0.98	0.97	0.95	1.00	0.91	0.80	0.46
	Method 3	0.99	0.98	0.97	0.99	0.95	1.00	0.69	0.60	0.37
	Method 4	1.00	0.98	0.99	1.00	0.97	1.00	0.94	0.76	0.55
$m \approx \sqrt[3]{n}$	Method 1	0.98	0.95	0.97	0.93	0.96	0.98	0.73	0.66	0.49
	Method 2	0.97	0.97	0.98	0.97	0.95	0.98	0.91	0.67	0.45
	Method 3	0.99	0.99	0.97	0.99	0.95	0.99	0.69	0.47	0.28
	Method 4	1.00	0.97	0.99	1.00	0.97	0.98	0.94	0.63	0.52
$m = \frac{n}{100}$	Method 1	0.98	0.95	1.00	0.93	0.91	1.00	0.73	0.78	0.70
	Method 2	0.97	0.95	1.00	0.97	0.94	0.98	0.91	0.98	0.98
	Method 3	0.99	0.98	0.96	0.99	0.95	0.78	0.69	0.38	0.04
	Method 4	1.00	0.97	1.00	1.00	0.97	1.00	0.94	0.97	0.96

Table 5.2: Frequencies that $\theta_0(x)$ is inside of the corresponding credible interval for the squared exponential Gaussian process prior in a non-adaptive setting at given points $x \in \{0.5, 0.41148, 0.31143\}$. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning, Method 3: Spatial partitioning with inverse variance weights, Method 4: Spatial partitioning with exponential weights. From left to right the sample size is $n = 1000, 4500, 10000$.

	$N =$	1000	4500	10000
$m = 10$	Benchmark	**0.93**	**0.96**	**0.95**
	Method 1	0.96	0.95	0.92
	Method 2	0.95	0.95	0.95
	Method 3	0.99	0.98	0.99
	Method 4	0.99	0.96	0.98
$m \approx \sqrt[3]{n}$	Method 1	0.96	0.93	0.93
	Method 2	0.95	0.98	0.99
	Method 3	0.99	0.99	1.00
	Method 4	0.99	0.97	1.00
$m = \frac{n}{100}$	Method 1	0.96	1.00	1.00
	Method 2	0.95	1.00	1.00
	Method 3	0.99	0.98	0.57
	Method 4	0.99	0.98	1.00

Table 5.3: Frequencies that θ_0 is inside of the credible ball for the squared exponential Gaussian process prior in a non-adaptive setting at given points $x \in \{0.5, 0.41148, 0.31143\}$. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning, Method 3: Spatial partitioning with inverse variance weights, Method 4: Spatial partitioning with exponential weights. From left to right the sample size is $n = 1000, 4500, 10000$.

linearly with the quantity of data available, the posterior means of the distributed methods concentrates at a sub-optimal rate. On the other hand, we can notice on Table 5.1b that the radius of the L_2 credible ball for every methods is bigger than the L_2 distance between the posterior mean and the true function on average. Furthermore, Tables 5.2 and 5.3 corroborate this statement by showcasing that the coverage obtained by distributed methods in a non-adaptive setting is as good as in the non-

	$n =$	1000	4500	10000
$m = 10$	Benchmark	**1.9800 sec**	**16.5773 sec**	**70.5413 sec**
	Random	1.8536 sec	8.2083 sec	18.3162 sec
	Spatial	1.8900 sec	8.1137 sec	18.3614 sec
$m \approx \sqrt[3]{n}$	Random	1.8536 sec	8.8279 sec	18.1707 sec
	Spatial	1.8900 sec	8.6763 sec	18.5524 sec
$m = \frac{n}{100}$	Random	1.8536 sec	8.5768 sec	18.5996 sec
	Spatial	1.8900 sec	8.3238 sec	18.4250 sec

Table 5.4: Average running time for the computation of the posterior for θ_0 for the squared exponential Gaussian process prior in a non-adaptive setting. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning. From left to right the sample size is $n = 1000, 4500, 10000$

distributed case. It is also noted that **Method 2** and **4** may lead to better point-wise coverage than using a non-distributed regression method. This may be due to the capacity of spatially distributed regression to capture localized properties of the "truth".

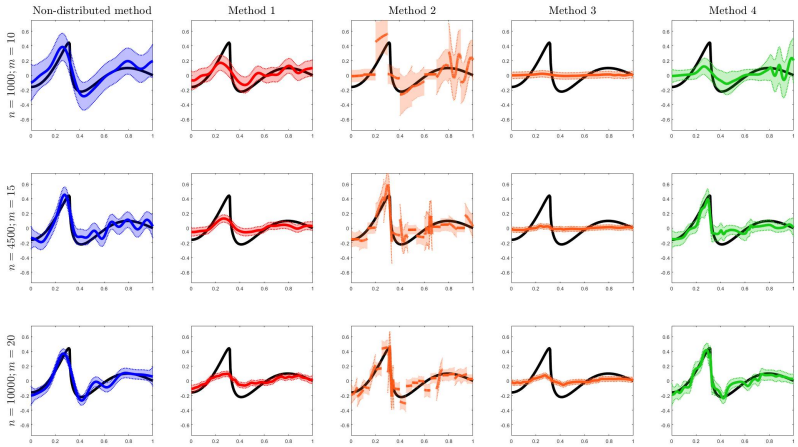

Figure 5.2: Adaptive posterior density for the function θ_0 (drawn in black) on $x \in [0, 1]$. The posterior means are drawn by solid line, while the 95% point-wise credible sets are shaded between two dotted lines. In the first column, we plot the non-distributed method, in the second column the distributed method with random partitioning, the third column the distributed method with spatial partitioning, while in the fourth and fifth column we plot the distributed method with spatial partitioning with inverse variance weights and exponential weights respectively. From top to bottom the sample size is $n = 1000, 5000, 10000$ and the number of experts is $m = 10, 15, 20$.

Although the last-mentioned empirical results draw an optimistic picture of distributed methods, these results are only valid when the exact smoothness β of the function θ_0 is known in advance. In real life applications, this might not be realistic and the smoothness is generally learned before doing any inference on the true function. One can for example estimate the Gaussian process prior parameter from the data using a frequentist technique. We will be using the maximum marginal likelihood estimator (MMLE) in the present study. Despite exhibiting reasonably good contraction rates, the Gaussian process with squared exponential kernel where the

	$n =$	1000	4500	10000
$m = 10$	BM	**32.1979**	**20.2424**	**16.0613**
	M 1	43.3699	34.1139	18.4541
	M 2	41.1122	22.3420	15.6854
	M 3	46.9816	41.1088	33.6787
	M 4	36.6261	21.9179	15.9172
$m \approx n^{1/3}$	M 1	43.3699	38.6536	36.6079
	M 2	41.1122	24.2863	17.9899
	M 3	46.9816	44.6756	43.7188
	M 4	36.6261	22.5384	15.6955
$m = \frac{n}{100}$	M 1	43.3699	45.1984	46.0332
	M 2	41.1122	33.8245	32.4338
	M 3	46.9816	47.9000	48.2887
	M 4	36.6261	27.5078	26.9132

(a) Average L_2 distance between θ_0 and the posterior mean for the squared exponential Gaussian process prior in an adaptive setting.

	$n =$	1000	4500	10000
$m = 10$	BM	**46.6420**	**27.5647**	**20.4758**
	M 1	40.9491	21.6266	17.2895
	M 2	62.1316	39.4716	33.1512
	M 3	23.6971	17.5376	18.1167
	M 4	42.4518	30.6414	27.0113
$m \approx n^{1/3}$	M 1	40.9491	19.4399	14.3160
	M 2	62.1316	43.3218	38.5396
	M 3	23.6971	15.0748	15.0664
	M 4	42.4518	32.0551	28.8076
$m = \frac{n}{100}$	M 1	40.9491	26.2542	27.0996
	M 2	62.1316	55.1480	46.9959
	M 3	23.6971	11.1716	16.3951
	M 4	42.4518	37.2899	38.5320

(b) Average L_2 credible ball radius for the squared exponential Gaussian process prior in an adaptive setting.

Table 5.5: BM: Benchmark, Non-distributed method. M 1: Random partitioning, M 2: Spatial partitioning, M 3: Spatial partitioning with inverse variance weights, M 4: Spatial partitioning with exponential weights. From left to right the sample size is $n = 1000, 4500, 10000$.

		$x = 0.5$			$x = 0.41148$			$x = 0.31143$		
	$n =$	1000	4500	10000	1000	4500	10000	1000	4500	10000
$m = 10$	Benchmark	**0.96**	**0.97**	**0.96**	**0.79**	**0.89**	**0.96**	**0.14**	**0.04**	**0.01**
	Method 1	0.61	0.68	0.89	0.34	0.51	0.79	0.11	0.02	0.00
	Method 2	0.84	0.96	0.93	0.77	0.90	0.98	0.47	0.65	0.42
	Method 3	0.00	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00
	Method 4	0.62	0.74	0.83	0.59	0.85	0.96	0.21	0.34	0.24
$m \approx \sqrt[3]{n}$	Method 1	0.61	0.48	0.23	0.34	0.12	0.03	0.11	0.00	0.00
	Method 2	0.84	0.91	0.94	0.77	0.92	0.95	0.47	0.63	0.67
	Method 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Method 4	0.62	0.73	0.84	0.59	0.83	0.95	0.21	0.17	0.36
$m = \frac{n}{100}$	Method 1	0.61	0.09	0.11	0.34	0.06	0.05	0.11	0.03	0.02
	Method 2	0.84	0.64	0.73	0.77	0.75	0.79	0.47	0.72	0.92
	Method 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Method 4	0.62	0.43	0.54	0.59	0.72	0.64	0.21	0.32	0.32

Table 5.6: Frequencies that $\theta_0(x)$ is inside of the corresponding credible interval for the squared exponential Gaussian process prior in an adaptive setting at given points $x \in \{0.5, 0.41148, 0.31143\}$. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning, Method 3: Spatial partitioning with inverse variance weights, Method 4: Spatial partitioning with exponential weights. From left to right the sample size is $n = 1000, 4500, 10000$.

rescaling parameter is estimated using the MMLE is known to provide unreliable uncertainty coverage since the credible sets based thereon fail to cover the true function, see Chapter 2. We will witness this pattern in the following results, in particular in Table 2.7.

In Figure 5.2, one can observe that as soon as one considers adaptation, the

	$N =$	1000	4500	10000
$m = 10$	Benchmark	**0.87**	**0.79**	**0.60**
	Method 1	0.33	0.30	0.38
	Method 2	0.98	0.99	0.99
	Method 3	0.09	0.00	0.02
	Method 4	0.72	0.88	0.93
$m \approx \sqrt[3]{n}$	Method 1	0.33	0.05	0.01
	Method 2	0.98	0.98	1.00
	Method 3	0.09	0.00	0.02
	Method 4	0.72	0.90	0.98
$m = \frac{n}{100}$	Method 1	0.33	0.13	0.15
	Method 2	0.98	1.00	1.00
	Method 3	0.09	0.00	0.09
	Method 4	0.72	0.95	0.99

Table 5.7: Frequencies that θ_0 is inside of the credible ball for the squared exponential Gaussian process prior in an adaptive setting at given points $x \in \{0.5, 0.41148, 0.31143\}$. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning, Method 3: Spatial partitioning with inverse variance weights, Method 4: Spatial partitioning with exponential weights. From left to right the sample size is $n = 1000, 4500, 10000$.

	$n =$	1000	4500	10000
$m = 10$	Benchmark	**8.0086 sec**	**3.40244 min**	**23.81 min**
	Random	3.7013 sec	20.6797 sec	79.5956 sec
	Spatial	3.6724 sec	21.5390 sec	81.8753 sec
$m \approx \sqrt[3]{n}$	Random	3.7013 sec	17.5368 sec	49.5480 sec
	Spatial	3.6724 sec	16.36173 sec	49.1387 sec
$m = \frac{n}{100}$	Random	3.7013 sec	16.4724 sec	29.9921 sec
	Spatial	3.6724 sec	15.0549 sec	32.0985 sec

Table 5.8: Average running time for the computation of the posterior for θ_0 for the squared exponential Gaussian process prior in an adaptive setting. Benchmark: Non-distributed method. From left to right the sample size is $n = 1000, 4500, 10000$

distributed methods do not behave similarly any longer. Indeed, the way the data is partitioned influences how well the true function can be recovered with the global posteriors. For instance, if we randomly partition the data across the machines and the draw local Bayesian inference using local empirical Bayes estimates, the global posterior will not contract optimally around the true function. This phenomenon has already been studied theoretically in (Szabó and van Zanten, 2019) in the signal-in-white-noise model. Using spatial partitioning could seem more sensible since each machine could adapt to the smoothness locally on a smaller region. This method does result in a global posterior which behaves similarly to the non-distributed method, and yet the discontinuities it inherently produces make it unattractive to practitioners. Fortunately, putting appropriate weights on each draw of the local posteriors, namely $\omega_k(x) = W(x)w_k(x)$, where $w_k(x) = e^{-m^2(x-c_k)^2}/\sigma_k^2(x)$ with c_k being the center of the local data region and $W(x) = \left(\sum_{j=1}^m w_k(x)\right)^{-1}$, alleviate the issue of the "jumps" at the border of the different partitioning regions while preserving the optimal recovery property of the method.

Tables 5.5a, 5.5b, 5.6 and 5.7 support the conclusion drawn from Figures 5.2. They illustrate that if one wants to adapt the priors to the smoothness locally before computing the global posterior, then the distributed method of choice influences greatly the performance of the corresponding posterior. Some methods (**Methods 1** and **3**) result in very poor performances, which can be exacerbated if the number of experts increases with the data. Considering this dysfunction, it may seem that distributed adaptation sometimes leads to a global posterior behaving as bad as the worst local posterior in terms of contraction around the true function and credible set coverage. On the other hand, the other methods (**Methods 2** and **4**) exhibit promising results when that $m \lesssim \sqrt[3]{n}$. Not only are the rates at which the posterior means approach the true function for those methods on par with their non-distributed counterpart, both the point-wise and the L_2 coverage are sometimes slightly improved due to the localized aspect of the former methods. Moreover, we can notice that the coverage is still good even when the number of machines increases linearly with n, which indicates that even when the methods do not achieve optimal recovery, the global posterior does not concentrate too much around the global posterior mean. It should be taken into account that these results are nonetheless only numerical.

While Table 5.4 highlighted the gain in computation times distributed methods offer when the smoothness of the true function is correctly assumed, Table 5.8 emphasizes that this gain is considerable in an adaptive setting. As a matter of fact, the computation of the MMLE is also heavily influenced by the size of the data which explains why the computation of distributed methods takes much less time than the computation of the classical posterior.

Overall, most methods are of interest, especially when the smoothness is assumed to be known, although some of them perform sub-optimally in the adaptive setting. It seems that spatial partitioning with exponentially decreasing weights is the method that generates a global posterior closest to the long-established conventional posterior when the number of experts increase reasonably fast.

§5.2.2 Airline Delays (USA Flight)

Next, we will compare the performance of our different distributive method on a large-scale data set: flight arrival and departure times for every commercial flight in the USA from January 2008 to April 2008. This data set covers more than 5 million flights and contains exhaustive information about the flights, including delays at arrival (in minutes). The average delay in first the quarter of 2008 was about 30 minutes, but one may be interested in estimating this delay more accurately thanks to wealth of data available. However, the usual non-parametric Bayesian regression is discouraged due to the mere size of the data set

This data set has already been studied before by (Hensman et al., 2013), (Gal et al., 2014), (Ng and Deisenroth, 2014) and (Ng and Deisenroth, 2015) using various methods to speed-up the regression. (Hensman et al., 2013) used Stochastic Variational inference (SVI) with inducing points, whereas (Ng and Deisenroth, 2015) compared different distributed methods with random partitioning, among which their robust Bayesian Committee Machine (rBCM) performed the best. We decided to follow the same procedure described in those articles; in order to predict the delay at arrival we select $P = 70K, 2M$ and $5M$ data points to train our models and $100,000$

$P =$	$70K$			$2M$			$5M$		
	CT	RMSE	SE	CT	RMSE	SE	CT	RMSE	SE
SVI	–	33.0	–	–	–	–	–	–	–
rBCM	13 s	27.1	< 0.3	39 s	34.4	< 0.3	90 s	35.5	< 0.3
M 1	24.2 m	29.1	0.52	35.4 m	34.8	0.49	41.3 m	41.5	0.5
M 2	22.5 m	25.0	0.12	31.1 m	27.1	0.14	39.9 m	30.2	0.17
M 3	24.6 m	33.4	0.35	35.0 m	40.4	0.29	40.8 m	45.1	0.28
M 4	23.7 m	26.6	0.15	35.5 m	31.5	0.14	42.1 m	31.8	0.15

Table 5.9: US Flight Data Set. Performance of different method in terms of computation time (CT), root-mean-square error (RMSE) and standard error (SE). SVI and rBCM results are reported from (Hensman et al., 2013) and (Ng and Deisenroth, 2015) respectively. Best and worse performance by training size are highlighted in blue and red, respectively. M 1: Random partitioning, M 2: Spatial partitioning, M 3: Spatial partitioning with inverse variance weights, M 4: Spatial partitioning with exponential weights.

other data points to test them. The dependent variable is of dimension 8 and encompass: the age of the aircraft (number of years since deployment), distance between the two airports (in miles), airtime (in minutes), departure time, arrival time, month, day of the week and day of the month. We conducted 10 experiments with 256, 512 and 1024 machines respectively. The computation times, the root-mean-square errors and the standard errors (i.e. root of the mean of the squares of the deviations within the training set) of the different methods are all reported in Table 5.9, along with the reported performance of the SVI and the rBCM. All the simulation have been made on a single workstation using an Intel Core i7-8700 CPU operating at 3.40 GHz and 16 GB of RAM using sequential computation of the different local GP posteriors (i.e. all the local posteriors have been computed and stored on the same station).

On the table, one can observe a decrease in performances with the number of training data which has already been reported in Ng and Deisenroth (2015). Nonetheless, it is also noticeable that partitioning randomly the data across machines leads to similar RMSE as the one obtaining by an rBCM, which is not surprising. It is noticeable that the spatial partitioning consistently outperforms all the other GP methods. Nonetheless, we should remind the reader that despite those performance, the resulting global posterior contains multiple discontinuity regions. The table draws however a dark picture on the prediction of weighting the local posterior with inverse variance after a spatial partitioning of the data. Luckily, this can be compensated by exponential weights which largely improve the prediction. Indeed, **Method 4** achieves consistently better RMSE than other reported methods in the literature while still providing a global posterior with continuous draws.

Bibliography

Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*, 40(3):549–570.

Balder, E., Gilliland, D., and van Houwelingen, J. (1983). On the essential completeness of Bayes empirical Bayes decision rules. *Statistics & Risk Modeling*, 1(4-5):503–510.

Belitser, E. (2017). On coverage and local radial rates of credible sets. *Ann. Statist.*, 45(3):1124–1151.

Belitser, E. and Enikeeva, F. (2008). Empirical Bayesian test of the smoothness. *Math. Methods Statist.*, 17(1):1–18.

Bényi, A. and Oh, T. (2013). The sobolev inequality on the torus revisited. *Publicationes Mathematicae Debrecen*, 83:359–374.

Berlinet, A. and C. Thomas-Agnan, C. (2004). *RKHS and Stochastic Processes*, pages 55–108. Springer US.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.

Bhattacharya, A. and Pati, D. (2015). Adaptive Bayesian inference in the Gaussian sequence model using exponential-variance priors. *Statist. Prob. Letters*, 103:100–104.

Bhattacharya, A., Pati, D., and Yang, Y. (2017). Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv e-prints*.

Bickel, P. J. (1982). On Adaptive Estimation. *The Annals of Statistics*, 10(3):647 – 671.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day Company, Oakland, California.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brown, L. and Low, M. (1996). A constrained risk inequality with applications to nonparametric function estimation. *Ann. Statist.*, 24:2524–2535.

Cai, T. and Low, M. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.*, 32:1805–1840.

Cao, Y. and Fleet, D. (2014). Generalized product of experts for automaticand principled fusion of gaussian process predictions. *arXiv e-prints*.

Castillo, I. and Nickl, R. (2013). Nonparametric Bernsteinñ-von Mises theorems in gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.

Castillo, I. and Nickl, R. (2014). On the Bernsteinñ-von Mises phenomenon for nonparametric bayes procedures. *Ann. Statist.*, 42(5):1941–1969.

Choudhuri, N., Ghosal, S., and Roy, A. (2007). Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology*, 4:227–243.

Cobos, F., Kühn, T., and Sickel, W. (2015). Optimal approximation of multivariate periodic Sobolev functions in the sup-norm. *Journal of Functional Analysis*, 270.

Cox, D. D. (1993). An analysis of bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923.

de Jonge, R. and van Zanten, J. H. (2009). Adaptive nonparametric bayesian inference using location-scale mixture priors. Technical report.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, 14:1–26.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.*, 22(1):238–270.

Doob, J. L. (1949). Application of the theory of martingales. In *Le calcul des probabilités et ses applications*, number 13 in CNRS International Colloquia, pages 23–27. CNRS, Paris.

Ferrari-Trecate, G., Williams, C., and Opper, M. (1998). Finite-dimensional approximation of gaussian processes. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press.

Freedman, D. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386 – 1403.

Gal, Y., van der Wilk, M., and Rasmussen, C. E. (2014). Distributed variational inference in sparse Gaussian process regression and latent variable models. *arXiv:1402.1389*.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28:500–531.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *Ann. Statist.*, 35(1):192–223.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Gibbs, N., Jr, W. P., and Stockmeyer, P. (1976). An algorithm for reducing the bandwidth and profile of a sparse matrix. *SIAM J. Numer. Anal.*, 13(2):236–250.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge series in statistical and probabilistic mathematics.

Goldenshluger, A. and Nemirovski, A. (1997). On spatial adaptive estimation of nonparametric regression. *Math. Meth. Statistics*, 6:135–170.

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2017). A divide-and-conquer bayesian approach to large-scale kriging. *arXiv preprint arXiv:1712.09767*.

Hadji, A. and Szabo, B. (2021). Can we trust bayesian uncertainty quantification from gaussian process priors with squared exponential covariance kernel? *SIAM/ASA Journal on Uncertainty Quantification*, 9(1):185–230.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. UAI'13, page 282–290, Arlington, Virginia, USA. AUAI Press.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32(4):1556 – 1593.

Hunter, J. (2013). Distributions and sobolev spaces. Lecture Notes: Analysis Prelim Workshop. Department of Mathematics of the University of California Davis.

Isserlis, L. (1916). On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression. *Biometrika*, 11(3):185–190.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *Ann. Statist.*

Johnstone, I. and Silverman, B. (2005). Empirical bayes selection of wavelet thresholds. *Ann. Statist.*, 33:1700–1752.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Comput.*, 6(2):181–214.

Karhunen, K. (1947). über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae Series A. I. Mathematica*, (37):1–79.

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *The Annals of Statistics*, 29(3):666 – 686.

Knapik, B., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657.

Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3):771–813.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation (revised edition)*. Springer-Verlag, New York.

Lember, J. and van der Vaart, A. (2007). On universal bayesian adaptation. *Statistics & Decisions*, 25(2):127–152.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512 – 2546.

Loève, M. (1978). *Probability theory Vol. II*, volume 46 of *Graduate Texts in Mathematics*. Springer-Verlag.

Mallasto, A. and Feragen, A. (2017). Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5660–5670. Curran Associates, Inc.

Meeds, E. and Osindero, S. (2006). An alternative infinite mixture of gaussian process experts. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014). Robus and scalabale Bayes via a media of subset posterior measures. *arXiv preprint*.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Ng, J. and Deisenroth, M. (2014). Hierarchical mixture-of-experts modelfor large-scale gaussian process regression. *arXiv e-prints*.

Ng, J. and Deisenroth, M. (2015). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1481–1490.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and gaussian white noise. *Ann. Statist.*, 24(6):2399–2430.

Opper, M. and Vivarelli, F. (1999). General bounds on bayes errors for regression with gaussian processes. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems II*, pages 302–308. MIT Press.

Park, C. and Apley, D. W. (2018). Patchwork kriging for large-scale gaussian process regression. *J. Mach. Learn. Res.*, 19:7:1–7:43.

Park, C. and Huang, J. Z. (2016). Efficient computation of gaussian process regression for large spatial data sets by patching local gaussian processes. *Journal of Machine Learning Research*, 17(174):1–29.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *J. Machine Learning Research*, 6:1939–1959.

Rasmussen, C. and Ghahramani, Z. (2002). Infinite mixtures of gaussian process experts. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press, Boston.

Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer.

Ray, K. (2017). Adaptive bernstein-von mises theorems in gaussian white noise. *Ann. Statist.*, 45(6):2511–2536.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.*, 34(1):229–253.

Rousseau, J. and Szabo, B. (2017). Asymptotic behaviour of the empirical bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.*, 45:833–865.

Rousseau, J. and Szabo, B. (2020). Asymptotic frequentist coverage properties of bayesian credible sets for sieve priors. *Annals of Statistics*, 48(4):2155–2179.

Saad, Y. (1990). Sparskit: a basic tool kit for sparse matrix computations.

Samo, Y.-L. K. and Roberts, S. J. (2016). String and membrane gaussian processes. *Journal of Machine Learning Research*, 17(131):1–87.

Schwartz, L. (1965). On Bayes procedures. *Z. Warsch. Verw. Gebiete*, 4:10–26.

Scott, S., Blocker, A., Bonassi, F., Chipman, H., George, E., and McCulloch, R. (2016). Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Serra, P. and Krivobokova, T. (2017). Adaptive empirical bayesian smoothing splines. *Bayesian Anal.*, 12(1):219–238.

Sniekers, S. and van der Vaart, A. (2015a). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.*, 9(2):2475–2527.

Sniekers, S. and van der Vaart, A. (2015b). Credible sets in the fixed design model with Brownian motion prior. *Journal of Statistical Planning and Inference*, 166:78–86.

Srivastava, S., Cevher, V., Dinh, Q., and Dunson, D. (2015). WASP: Scalable Bayes via barycenters of subset posteriors. In Lebanon, G. and Vishwanathan, S., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 912–920, San Diego, California, USA. PMLR.

Szabó, B. and van Zanten, H. (2019). An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research*, 20(87):1–30.

Szabo, B. T., van der Pas, S., and van der Vaart, A. W. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221 – 1274.

Szabo, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013). Empirical bayes scaling of gaussian priors in the white noise model. *Electron. J. Statist.*, 7:991–1018.

Szabo, B. T., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Annals of Statistics*, 43(4):1391–1428.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 567–574.

Tokdar, S. and Ghosal, S. (2005). Posterior consistency of Gaussian process priors in density estimation. *J. Statist. Plann. Inference*, 137:34–42.

Tresp, V. (2000). The generalized bayesian committee machine. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 130–139, New York, NY, USA. Association for Computing Machinery.

Tresp, V. (2001). Mixtures of gaussian processes. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats.

van der Pas, S., Szabo, B., and van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Statist.*, 11(2):3196–3225.

van der Vaart, A. and van Zanten, J. H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Statist.*, 1:433–448.

van der Vaart, A. and van Zanten, J. H. (2009a). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37:2655–2675.

van der Vaart, A. and van Zanten, J. H. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2009b). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer-Verlag, Berlin, Heidelberg.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser.*, 26:359–372.

Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.

Summary

This thesis investigates the frequentist properties of Bayesian procedures, specifically the use of Gaussian process priors in Bayesian nonparametric statistics. Typically, in Bayesian statistics, there is no fixed unique parameter of our model, but rather a realization of a random variable which will act as a parameter. To this end, an prior distribution on the parameters is assumed. After observing the data, this leads to a posteriori distribution on the parameters, which is used to compute estimates of the parameters. Although this paradigm differs significantly from frequentist statistics, it is still interesting to see how the posterior distribution behaves as a random measure which depends on the "true" parameter.

One, very appealing advantage of the Bayesian framework is that it readily provides built-in uncertainty quantification. Indeed, since it is possible to sample from the posterior, the construction of credible sets containing a fraction of the posterior mass is relatively simple. In Chapter 2, we study the coverage of those credible sets resulting in from Gaussian process priors with squared exponential covariance kernel. As the sample paths of the process are infinitely smooth, the common practice is to rescale it in order to recover the underlying functional parameter. The optimal scaling depends on the smoothness of this parameter, which is generally unknown. The scaling hyper-parameter is thus generally learnt from the data using either hierarchical Bayes or Empirical Bayes techniques. Unfortunately, both methods initially lead to overconfident, unreliable uncertainty statements for a large class of parameters in the context of the Gaussian white noise model. However, blowing up the credible sets with a logarithmic factor or modifying the estimated hyper-parameter with a logarithmic term can get good frequentist coverage while maintaining a reasonable adaptive size.

The following chapters focus on the scalability of Bayesian methods in the context of Gaussian process nonparametric regression. Coming from the Bayesian paradigm, Gaussian process regression allows to make probabilistic statements the regression function based on the data. Moreover, when the noise in the model is Gaussian, it is possible to make use of conjugacy to obtain a closed form solution for the posterior distribution of the regression function. Nonetheless, this model is extremely greedy as its computational complexity scales cubicly with the number of observations. Distributed methods allow to divide the data across different machines which will all perform a local Bayesian nonparametric regression. The local solutions will then be collected by a global machine and aggregated into a global distribution for the regression function.

The naive approach is to simply perform a Gaussian process nonparametric regression with a random subset of the data in each machine and average all the local distributions into a global one. This approach quickly shows its frequentist limits as the convergence rate of the posterior will depends on the number of machines. Other methods are possible: down-scaling the prior locally and then averaging the results,

or up-scaling the likelihood and find the Wasserstein-barycenter of the resulting local distributions for instance. If the smoothness parameter of the Gaussian process prior matches the true regularity of the regression function, then both methods lead to near-optimal recovery and good uncertainty quantification, provided the number of machines does not increase too fast compared to the number of observations.

Another approach would be to partition the design space of the regression such that each machine performs a regression on one of the non-overlapping resulting sub-regions. Though the final posterior distribution contains discontinuities at the borders of each partitions, it will contract optimally to the "true" regression function when the rescaling hyper-parameter of the prior is well chosen. Furthermore, this approach can also lead to adaptive optimal contraction rates. Even when the "true" regularity is unknown, it is possible to learn the hyper-parameter using a hierarchical framework. This will also lead to optimal recovery of the regression function.

This approach can be seen as an aggregation of the posterior samples obtained by the different machines combined with weight functions. Indeed, the weight functions would be indicator functions of a sub-region of the design space. The discontinuities of the latter are then the result of the discontinuities of the weights. A thorough simulation study suggests that by choosing appropriate, data-driven weights, it is possible to achieve adaptive near-optimal recovery and coverage of the underlying regression function.

Samenvatting

Dit proefschrift onderzoekt de frequentistische eigenschappen van Bayesiaanse procedures, met name het gebruik van Gauss-proces als *a-priori*-verdeling in Bayesiaanse niet-parametrische statistiek. In Bayesiaanse statistiek is er typisch geen vaste unieke parameter van het model, maar eerder een realisatie van een stochastische variabele die als parameter zal fungeren. Daartoe wordt een *a-priori*-verdeling op de parameters aangenomen. Na waarneming van de data leidt dit tot een *a-posteriori*-verdeling op de parameters, welke gebruikt wordt om tot een schatting van de parameters te komen. Hoewel dit paradigma aanzienlijk verschilt van de frequentistische statistiek, is het toch interessant om te zien hoe de *a-posteriori*-verdeling zich gedraagt als een willekeurige maat die afhangt van de "ware" parameter.

Een zeer aantrekkelijk voordeel van het Bayesiaanse framework is dat het gemakkelijk ingebouwde kwantificering van onzekerheid biedt. Aangezien het mogelijk is een steekproef te nemen uit de *a-posteriori*-verdeling, is de constructie van geloofwaardigheidsgebieden die een fractie van de *a-posteriori*-verdelingsmassa bevatten relatief eenvoudig. In hoofdstuk 2 bestuderen we de dekking van die geloofwaardigheidsgebied die het resultaat zijn van Gauss-proces als *a-priori*-verdelingen met gekwadrateerde exponentiële covariance kernel. Omdat de steekproefpaden van het proces oneindig glad zijn, bestaat de gangbare praktijk erin de schaal te wijzigen om de onderliggende functionele parameter terug te vinden. De optimale schaling hangt af van de gladheid van deze parameter, die meestal onbekend is. De hyperparameter voor de schaling wordt dus meestal uit de gegevens geleerd met behulp van hierarchical Bayes of Empirical Bayes technieken. Helaas leiden beide methoden aanvankelijk tot overmoedige, onbetrouwbare onzekerheidsuitspraken voor een grote klasse parameters in de context van het Gaussian white noise model. Door echter de geloofwaardigheidsgebieden op te blazen met een logaritmische factor of de geschatte hyperparameter aan te passen met een logaritmische term kan een goede frequentistische dekking worden verkregen met behoud van een redelijke adaptieve grootte.

De volgende hoofdstukken richten zich op de schaalbaarheid van Bayesiaanse methoden in de context van niet-parametrische regressie op basis van Gauss-proceses. Vanuit het Bayesiaanse paradigma maakt Gauss-procesregressie het mogelijk probabilistische uitspraken te doen over de regressiefunctie op basis van de gegevens. Wanneer de ruis in het model Gaussisch is, is het bovendien mogelijk gebruik te maken van conjugatie om een relatief eenvoudig formule voor de a-posteriori-verdeling van de regressiefunctie. Dit model is echter zeer inhalig, want de rekencomplexiteit ervan schaalt kubiek met het aantal waarnemingen. Gedistribueerde methoden maken het mogelijk de gegevens te verdelen over verschillende machines die allemaal een lokale Bayesiaanse niet-parametrische regressie uitvoeren. De lokale oplossingen worden dan door een globale machine verzameld en samengevoegd tot een globale verdeling van de regressiefunctie.

De naïeve aanpak bestaat erin gewoon een niet-parametrische Gaussische regressie

uit te voeren met een willekeurige subset van de gegevens in elke machine en het gemiddelde van alle lokale verdelingen te nemen als globale verdeling. Deze aanpak toont al snel zijn frequentistische grenzen, doordat de convergentiesnelheid van de *a-posteriori*-verdeling afhangt van het aantal machines. Andere methoden zijn mogelijk: de *a-priori*-verdeling lokaal omlaag schalen en dan de resultaten middelen, of de aannemelijkheid omhoog schalen en bijvoorbeeld het Wasserstein-barycentrum van de resulterende lokale verdelingen vinden. Indien de gladheidsparameter van de prior van het Gaussische proces overeenkomt met de werkelijke regelmaat van de regressiefunctie, dan leiden beide methoden tot een bijna optimaal terugvinding en een goede kwantificering van de onzekerheid, mits het aantal machines niet te snel toeneemt ten opzichte van het aantal waarnemingen.

Een andere aanpak zou zijn om de ontwerpruimte van de regressie zodanig te partitioneren dat elke machine een regressie uitvoert op een van niet-overlappende resulterende subgebieden. Hoewel de uiteindelijke *a-posteriori*-verdeling discontinuïteiten bevat aan de grenzen van elke partitie, zal die optimaal samentrekken tot de "ware" regressiefunctie wanneer de herschalingshyperparameter van de prior goed gekozen is. Bovendien kan deze aanpak ook leiden tot adaptieve optimale contractiesnelheden. Zelfs wanneer de "ware" regelmaat onbekend is, is het mogelijk de hyperparameter te leren met behulp van een hiërarchisch framework. Ook dit leidt tot een optimaal herstel van de regressiefunctie.

Deze aanpak kan worden gezien als een aggregatie van de proefsteken van de *a-posteriori*-verdeling verkregen door de verschillende machines in combinatie gewichtsfuncties. De gewichtsfuncties zouden immers indicatorfuncties zijn van een subregio van de ontwerpruimte. De discontinuïteiten van deze laatste zijn dan het gevolg van de discontinuïteiten van de gewichten. Uit een grondige simulatiestudie blijkt dat het mogelijk is om door het kiezen van passende, gegevensgestuurde gewichten een adaptief bijna-optimaal herstel en dekking van de onderliggende regressiefunctie te bereiken.

Acknowledgements

Here comes the part I have been dreading to write for so long. Not that I do not feel thankful in any way, but I have always feared to forget to mention anyone. Indeed, I consider the completion of this thesis is not only my own work, but the collaboration of all the people in my life.

First of all, I would like to thank the two people who gave me the opportunity to write this piece, my supervisors, Aad van der Vaart and Botond Szabó. What struck me the most from my first day as a PhD candidate was Botond's limitless and contagious enthusiasm. He was always ready to partake in sometimes long and elaborate discussions about our projects. I was never afraid to ask a question or suggest an idea because even when they seemed trivial, he took the time to discuss them with me. During these few years, he trusted me more than I could trust myself, and you encouraged me to the point where I accomplished what I never thought I could. In addition, Aad's insight during this journey has been one of the most valuable in my short academic career. Our meetings, albeit occasional, helped me understood my own work better, and his piercing questions motivated the rest of my work.

I would also like to express my gratitude to the members of the dissertation committee for their time in reading my manuscript.

I thank my colleagues, not only from my own group, but from the whole Mathematical Institute. The numerous seminars, talks, but also drinks and barbecues we had together showed me that there is more to research than writing articles. I thank especially my past and present fellow PhD candidates, who I cannot list here not for a lack of memories, but for a lack of space. I will miss our lunches, our coffee breaks, our PhD colloquiums, our dinners, our pointless discussions, and secretly, I hope you will miss me too.

Finally, I would like to thank my housemates, my friends from all around the world, and my family who supported me and put up with me during these challenging years. However, as I said in the beginning, the completion of this thesis is the mere collaboration of all the people in my life. Thus, if you have the chance, or misfortune, of reading this, I would like to thank you.

Curriculum Vitae

Mohamed Amine Hadji was born on August 15, 1993 in Algiers, Algeria. After attending high school in Algiers, he started his bachelor in Applied mathematics and Operation Research in 2011 at the University of Science and Technology - Houari Boumediene (USTHB). He completed his bachelor degree in 2014 with a bachelor thesis on tiling problems using polyominos, supervised by Kahina Meslem. He pursued a master degree in Applied mathematics and Statistical learning at the university of Paris-Dauphine PSL. He obtained his master degree cum laude in 2016.

In December 2016, Mohamed Amine started his PhD at Leiden University under the supervision of dr. Botond Szabó, with Prof. dr. Aad van der Vaart acting as his promotor. During his PhD, Mohamed Amine acted out as a teaching assistant and lecturer for different statistics courses. He was also a member of various master defense committees. Moreover, Mohamed Amine presented his research on Gaussian processes at numerous international conferences, online and physically.