Optimal Bayesian distributed methods for nonparametric regression

Author:
Tammo Hesselink

Supervisor:
dr. B.T. Szabo
prof. dr. J. H. van Zanten

Examination date:
October 30, 2018
Abstract

Bayesian methods in nonparametric regression using Gaussian process priors have been widely used in practice. In this thesis we will consider Gaussian processes making use of both the Matérn kernel and the squared exponential kernel. However neither of these Gaussian process priors lead to scalable Bayesian methods and therefore they are highly impractical for large data sets. To solve this, we turn to the distributed setting where we divide the data of size $n$ over $m$ machines, which collectively help to compute a global posterior. For the distributed setting, we will consider multiple methods, which consist of changing the local posterior and different aggregation methods for the local posteriors. These methods will be studied via simulation. We will pose theoretical results for some of the methods. Finally, we will perform simulation studies using the squared exponential kernel, which will show to perform similarly to the Matérn kernel.
Contents

Introduction 5

1. Nonparametric Bayesian regression methods using Gaussian process priors 8
   1.1. Nonparametric Gaussian process regression 8

2. Distributed Bayesian methods using Gaussian process priors 16
   2.1. Naive averaging of posteriors 16
   2.2. Aggregation by product of experts 19
   2.3. Adjusted local likelihoods 21
   2.4. Adjusted priors 22
   2.5. Adjusted local likelihoods and Wasserstein barycenters 24
   2.6. Generalised Product of Experts 26
   2.7. The Bayesian Committee Machine 27
   2.8. Summary of distributed methods 29

3. Distributed Bayesian methods using the squared exponential kernel 31
   3.1. Simulation study 31
   3.2. Distributed adaptive methods 36
   3.3. Summary 38

4. Proofs 39
   4.1. Summary of notations 39
   4.2. The RKHS framework 40
   4.3. Lemmas used in the proofs 44
   4.4. Proof of Lemma 1 52
   4.5. Proof of Theorem 1 53
   4.6. Proof of Theorem 2 58
   4.7. Proof of Theorem 3 62
   4.8. Proof of Theorem 4 63
   4.9. Proof of Theorem 5 64
   4.10. Proof of Theorem 6 66

Conclusion 72

Appendices 75
Introduction

Recent times have seen the importance of data analysis grow immensely. Due to digital techniques it is easier to keep track of data, which allows for various ways of analysis. Terms such as machine learning and big data are being thrown around frequently. Most applications make use of historically kept data to fine tune their application to the users.

When data sets get sufficiently large, the computation can take up an impractical amount of time, or even be unattainable. In case of matrix multiplications, which are also made of great use in the nonparametric regression model, there are methods for sparse matrices available which speed up the process, see [11, 22, 20, 26].

Besides improving the computation time, it is also possible to distribute the data over machines or cores to decrease computation time. After the data is distributed over a cluster of either machines or cores, it is then processed locally after which a central machine aggregates these computations into a global estimation. Next to computational reasons, distributed methods are also used for convenience. For instance, in settings where data is first obtained in different locations, the local machines can already do computations on that data before communicating it to a central machine. Moreover it can be desired to not manage data in a central area for privacy or safety reasons, where distributing data offers a solution.

The nonparametric regression model is widely used for problems in which there is little a priori knowledge of the shape of the curve to be fitted. As opposed to the parametric regression setting where the “true” function $f_0$ is assumed to be in a certain family of functions with a certain amount of parameters to be inferred from the data, the nonparametric setting considers all methods that compute an estimate for the true function without any a priori assumption on the shape of this function.

Bayesian inference on the model offers more flexibility to deal with the uncertainty [15]. Bayesian statistics, as opposed to frequentist statistics, does not look to infer a single true parameter, but rather a distribution of the outcome. This method thus has built in uncertainty quantification. The model requires an assumed prior distribution of the truth which can be chosen in many ways, which offers options for many different settings. The choice of the prior however does add human bias to the model, which for more flexibility can be partially dealt with by endowing the parameters of the prior with hyper priors.

Gaussian processes are frequently used as prior distributions to estimate functions due to their flexibility for unknown functional parameters [10, 21]. They are determined by a mean function that states the expectation at time $t$ and a covariance function determining the covariance between multiple points in time. These functions can be chosen freely, and thus offer a great amount of options to the final shape of the Gaussian process prior.
The Bayesian nonparametric approach using Gaussian process priors has already been well studied [31, 2, 30]. In theory, this method performs very well for estimating curves with little information known a priori. The main problem arises in the scalability of this method. Applying the method to large data sets can take impractical computation time or even become unattainable. This leaves for a solution to make the method scalable, which is possible by using distribution of data.

The theoretical properties of distributed Bayesian methods have been thoroughly studied in the signal in white noise model [25], but not for the more convenient nonparametric regression model. From a methodological view it has been extensively studied, see for example [18, 17, 13]. The aggregation of the local posteriors can be performed using different methods with varying results in contraction rate and credible set coverage.

Methods to adaptively estimate the hyper parameters which have shown optimal results in the nondistributed setting fail to hold in the distributed case, with different methods attempted. Our aim is to find a method for which the distance between the estimator and the true function goes to zero at a similar rate to that of the nondistributed setting, with pointwise posterior credible set coverage for this estimator. For nondistributed setting optimal results have been shown in [33], which will form results to strive for in the distributed setting. These results include $L_\infty$ distances between the posterior mean and the truth, and pointwise posterior coverage of credible bounds.

We do so by applying the methods analysed in [25] and observe that the nonparametric regression model acts similar to the signal in white noise model for the different methods of aggregation of distribution. These include different modifications of the local posteriors and different aggregation methods. They show that the naive method of simply taking the average of local posteriors gives suboptimal results both in terms of recovery and uncertainty quantification. Furthermore, they investigate a variety of other distributed methods which show optimal recovery, uncertainty quantification or both.

We first apply the nondistributed method on the Bayesian nonparametric regression model using the Matérn kernel to see how the method performs in the nondistributed setting as a benchmark method. We recall the $L_\infty$ posterior contraction rate results and pointwise posterior credible set statements derived in [33] for the Matern kernel. These results are used in later chapters to compare the performance of the different distributed methods to.

In the second chapter we perform simulation studies for several distributed methods to see their performance in terms of bias of the posterior mean and coverage of credible bounds. This is done by applying different modifications of the local posteriors, where we will consider raising the power of either the likelihood or the prior, and considering different aggregation methods of the modified posteriors, which include naive averaging of posteriors, product of experts, Wasserstein Barycenter and the Bayesian committee machine [27]. We will prove sup-norm upper bounds for the bias of the posterior mean as well frequentist coverage of the pointwise posterior credible interval confirming the simulation studies.

In Chapter 3 we look at the squared exponential kernel as used in [18]. The distributed methods used for the Matérn kernel are applied to the same setting using the squared
exponential kernel for simulation studies on the posterior mean and credible set coverage. We also discuss the complications of proving sup-norm upper bounds for the posterior mean for the distributed methods using the squared exponential kernel.

Finally, we turn to a more realistic setting in which the optimal parameters of the squared exponential kernel are unknown beforehand. We consider adaptive methods where the parameters of the squared exponential kernel are being tuned solely from the data. We compare the performance of different methods through simulation studies.

Mathematical proofs of the stated results are collected in the last chapter.
1. Nonparametric Bayesian regression methods using Gaussian process priors

In the regression model we consider $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, the noise term, and the set of design points $X_n := \{x_1, \ldots, x_n\}$ is distributed uniformly on the interval [0,1] [33]. The aim is to infer the function $f$ from training data of size $n$, the predictors $X_n$ and dependent variables $Y_n := \{f(x_1) + \epsilon_1, \ldots, f(x_n) + \epsilon_n\}$. This inference can be done in multiple manners, which can be roughly spread into parametric and nonparametric regression.

In the parametric regression model, we assume the function $f$ lies in some family of functions $\mathcal{F}$, defined by a finite amount of parameters. The goal is to estimate these parameters from the data set $X_n, Y_n$.

Nonparametric regression simply encloses all methods that estimate $f$ without assuming it is of a certain form defined by a finite set of parameters. This generally makes for a method that deals better in situations where there is very little knowledge about the function a priori. This is frequently done by using kernel functions, which are used as a weighting function in the process of computing a weighted average. These kernels can take on many different forms, determined by a certain bandwidth parameter [21].

1.1. Nonparametric Gaussian process regression

In the regression setting, we apply the Bayesian approach by placing a prior $\Pi$ of choice on the function of importance $f$, which is then combined with the likelihood of the data $p_f(X_n, Y_n)$ to form the posterior distribution through Bayes’ rule [10],

$$d\Pi(f \mid X_n, Y_n) = \frac{p_f(X_n, Y_n)d\Pi(f)}{\int p_f(X_n, Y_n)d\Pi(f)}.$$

In this setting, it can be favourable to make use of Gaussian processes as prior distributions to estimate functions due to their flexibility for unknown functional parameters [10, 21].

A Gaussian process $W$ is defined as a stochastic process of which all finite dimensional distributions $(W_{t_1}, W_{t_2}, \ldots, W_{t_n})$ are multivariate normal random variables. These Gaussian processes are widely used as priors, with many different applications [18, 33, 25, 21, 16]. The multivariate normal distribution is defined by its mean function $\mu(t) = \mathbb{E}X_t$ and its covariance function $r(s, t) = \text{Cov}(X_s, X_t)$. In our analysis we first consider the
centered Gaussian process prior (having mean function \( \mu(t) = 0 \)) with the Matérn kernel function [21],
\[
k(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu|x - y|})^\nu B_\nu(\sqrt{2\nu|x - y|}),
\]
where \( \nu > 1 \) is the to be chosen smoothness parameter and \( B_\nu \) the Bessel function of the second kind. In the specific case where \( \nu = \frac{3}{2} \), this gives the absolute exponential kernel [21]:
\[
k(x, y) = e^{-|x-y|}.
\]
We want the variance of the prior to be decreasing in \( n \), thus we will scale the prior by \( \sigma^2(\frac{n\lambda}{n\lambda - 1}) \) as proposed in [33], where \( \sigma^2 \) is the error of the measured data, which we will assume to be known. Another useful kernel, which we will study in Chapter 3, is the squared exponential kernel which is defined as
\[
k(x_i, x_j) = e^{-\frac{1}{2}(x_i - x_j)^2}.
\]
This gives us as a prior distribution
\[
f(\cdot) \sim GP(0, \sigma^2(\frac{n\lambda}{n\lambda - 1}) K),
\]
where \( K := k(X_n, X_n) \) or \( \Pi(\cdot) \sim N_n(0, \Sigma) \) where \( \Sigma_{ij} = k(x_i, x_j) \). To compute the posterior distribution, first note that
\[
\text{Cov}(f(x_i), f(x_j)) = k(x_i, x_j),
\]
thus
\[
\text{Cov}(f(x_i) + \epsilon_i, f(x_j) + \epsilon_j) = \text{Cov}(f(x_i), f(x_j)) + \text{Cov}(\epsilon_i, \epsilon_j),
\]
as \( f(x_i) \) and \( \epsilon_j \) are independent for all \( i \) and \( j \). We thus obtain
\[
\text{Cov}(y_i, y_j) = k(x_i, x_j) + \sigma^2 \delta(x_i - x_j),
\]
as \( \text{Cov}(\epsilon_i, \epsilon_j) = 0 \) for \( i \neq j \). Finally, we assume the likelihood \( p_{f_0, \sigma} \) for this model is given as
\[
p_{f_0, \sigma}(Y_n) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \sum_{i=1}^{n}(y_i - f_0(x_i))^2}. \tag{1.3}
\]
The posterior distribution then takes the following form, based on [8], which we will prove for completeness.

**Lemma 1.** Let us have as a prior distribution on \( f \) of the form (1.2) and a likelihood of the form (1.3). Then we have as a posterior distribution for \( y^* := f(x^*) \mid Y_n \sim N(\mu(x^*), s^2(x^*)) \) for a single test input \( x^* \) where
\[
\mu(x^*) = k^T(K + n\lambda I)^{-1}Y_n, \tag{1.4}
\]
\[
s^2(x^*) = \sigma^2(n\lambda)^{-1}(k_{xx} - k^T(K + n\lambda I)^{-1}k), \tag{1.5}
\]
where \( K := k(X_n, X_n) \) is the covariance matrix of the data, \( k_* := k(x^*, X_n) \) the covariance of the test input \( x^* \) with the data and \( k_{xx} = k(x^*, x^*) \).
Proof. The proof to this theorem is given in Section 4.4.

In view of [33], the posterior covariance kernel takes the form

\[ \tilde{C}_n(x, x') = \sigma^2 (n\lambda)^{-1} (k(x, x') - k(x, X_n)(K + n\lambda I)^{-1}k(X_n, x')), \]  

(1.6)

which will be used in later computations regarding the frequentist coverage of the pointwise posterior credible sets. We use the posterior distribution given the data to draw samples for the true function \( f_0 \). When computing the posterior distribution, the matrix operations increase the computation time required of the process significantly, with for example the inversion of a matrix scaling cubically for the data [14]. To prevent unnecessary computations, we will store the matrices that are equal at \( f(x^*) \mid X_n, Y_n \) for every \( x^* \) in a memory cache which can be accessed by the computer. In this case that the matrix \( (K + n\lambda I)^{-1} \) will be calculated on beforehand and applied to the computation of \( f(x^*) \mid X_n, Y_n \) for all \( x^* \). Using the following algorithm we will compute posterior distributions based on data with corresponding credible sets.

**Algorithm 1** Calculating \( f(x^*) \) in the nondistributed setting on the interval \([0,1]\)

1: Given \( X_n, Y_n, \sigma^2, \nu \).
2: Compute \( K \).
3: Compute the inverse of \( K + n\lambda I \) and store in cache.
4: for \( x^* \in [0,1] \) evenly spaced do
5: Compute \( k_*, k_{**} \).
6: Use these to compute \( \mu(x^*) \) and \( \sigma^2(x^*) \).
7: Draw samples for the posterior distribution and compute the \((100 - \beta / 2)\)% and \( \beta / 2 \)% percentiles which form the credible interval for \( \hat{f}(x^*) \)
8: end for

The covariance kernel depends only on the parameter \( \lambda \). Our aim is to find the value of \( \lambda \) for which the \( L_\infty \) distance between the posterior mean and the truth is as small as possible, with high probability. The expected distance is bounded by a bias term and a variance term. The parameter \( \lambda \) for which the bias and the variance are of equal order is considered the value which returns the minimax optimal rate.

Let \( \tilde{f}_n = \mathbb{E}[f \mid X_n, Y_n] \) denote the posterior mean based on data of size \( n \). From [33] we find that the upper bound for \( \|\tilde{f}_n - f_0\|_{\infty} \) is given up to a constant by

\[ \|\mathbb{E}\hat{f} - f_0\|_{\infty} + \sigma \sqrt{\frac{\log n}{nh}}, \]  

(1.7)

with probability at least \( 1 - n^{-10} \), where we define \( h := \lambda^{1/(2\alpha)} \) and \( \|\mathbb{E}\hat{f} - f_0\|_{\infty} \) the bias which we will look deeper into in Section 4.2. Here the first term corresponds to the bias and the second term the variance. To find an upper bound for the bias, we need to define regularity classes in which the true function \( f_0 \) should lie. There are two frequently used classes namely, \( \Theta^\alpha_S(B) \) the set of \( \alpha \)-smooth Sobolev functions and \( \Theta^\alpha_H(B) \)
the set of \( \alpha \)-smooth functions in the H"older class bounded. If we have an orthonormal basis \( \{ \phi_i \}_{i \in \mathbb{N}} \) for \( L^2([0,1]) \) the sets are for \( \alpha > \frac{1}{2} \) and \( B > 0 \) defined as follows

\[
\Theta^{\alpha+1/2}_S(B) = \left\{ f = \sum_{i=1}^{\infty} f_i \phi_i \in L^2([0,1]) : \sum_{i=1}^{\infty} f_i^2 i^{2(\alpha+1/2)} \leq B^2 \right\},
\]

and

\[
\Theta^\alpha_H(B) = \left\{ f = \sum_{i=1}^{\infty} f_i \phi_i \in L^2([0,1]) : \sum_{i=1}^{\infty} |f_i|^\alpha \leq B \right\}. \tag{1.8}
\]

When the smoothness \( \alpha \) of the to be inferred function \( f_0 \) is known, we choose the tuning parameter of the Matérn kernel \( \nu = \alpha + \frac{1}{2} \).

In the optimal situation, we have our bias and variance terms of equal order as in the case where either one is larger, we have a mean square error of greater order, which means the posterior mean will converge to the truth slower. Thus, finding the parameter \( h \) which balances out the bias and variance terms means that the choice of the hyper parameter is optimal. Setting the bias equal to the variance, for a sample size of \( n \) we obtain a hyper parameter which retains the best possible rate

\[
h = \left( \frac{B^2 n}{\sigma^2 \log n} \right)^{-\frac{1}{\alpha+1}}, \tag{1.9}
\]

for \( f \in \Theta^{\alpha+1/2}_S(B) \) or \( f \in \Theta^\alpha_H(B) \) \cite{33}. We use this to set our parameter \( \lambda = h^{\frac{1}{2\nu}} \).

We will use Algorithm 1 for a data set \( X_n, Y_n \), where \( y_i = f_0(x_i) + \epsilon, x_i \sim U(0,1) \) and \( \epsilon \sim N(0, \frac{1}{10}) \) of size \( n = 500 \), for the following function

\[
f_0(x) = \sum_{i=1}^{\infty} \sin(i) \sqrt{2} \cos \left( \pi \left( i - \frac{1}{2} \right) x \right) i^{-\frac{3}{2}}, \tag{1.10}
\]

where we will truncate the infinite sum at \( i = 1000 \) for computational reasons. Let \( || \cdot ||_N \) denote the empirical mean satisfying \( ||f||^2_N = \frac{1}{N} \sum_{i=1}^{N} f(x_i)^2 \), where \( \{ x_i \}_{i=1}^{N} \) are the \( N \) evenly spaced test points on \([0,1]\). Note that these test points are not equal to the training data points \( x_i \). We will use a test input of size \( N = 200 \). Lastly, we will compute the point-wise posterior 95\% credible intervals \( CI(x^*) \) around the posterior mean

\[
\Pi(f(x^*) \in CI(x^*) \mid X_n, Y_n) \geq .95.
\]

This is done by drawing 2000 samples of \( f(x^*) \) at every test coordinate \( x^* \) and calculating the 2.5\% and 97.5\% quantiles at these coordinates. Together, these credible intervals form a set \( CI = \{ CI(x_1), CI(x_2), \ldots, CI(x_{\text{test}}) \} \), which is plotted around the posterior mean. Furthermore, we demonstrate the behaviour of the prior and the corresponding posterior for comparison.
Figure 1.1.: The data \((X_n, Y_n)\) of size \(n = 500\) where 
\(x_i \sim U(0, 1)\) and 
\(y_i \sim f_0(x_i) + \epsilon\) where 
\(\epsilon \sim N(0, \frac{1}{2})\).

Figure 1.2.: \(\hat{f}_n\) on \([0,1]\) with the 95% credible set for tuning hyper parameter \(\nu = \frac{3}{7}\).

Figure 1.3.: \(\hat{f}_n\) on \([0,1]\) with the 95% credible set for \(\nu = \frac{11}{10}\).

Figure 1.4.: \(\hat{f}_n\) on \([0,1]\) with the 95% credible sets for \(\nu = \frac{5}{2}\).

Here one can observe the influence of \(\nu\) on the posterior mean clearly. The bandwidth and smoothness of the kernel grows for larger values of \(\nu\), while it shrinks for smaller values of \(\nu\). From this, it is apparent that both too small and too large values of \(\nu\) are not desired, as they respectively undersmooth and oversmooth the function, resulting in suboptimal estimations. Also one can notice the bias-variance trade-off, where a smoother estimation comes paired with smaller credible sets, and vice versa. This is also visible in the following figure which shows the values of the empirical \(L_2\) distance between the posterior mean and the truth under different values of \(\nu\).
Figure 1.5.: $||\hat{f}_n - f_0||^2_N$ on $[0,1]$ for different values of $\nu$ on an equal data set.

Figure 1.6.: Two draws from the prior with tuning hyper parameter $\nu = \frac{3}{2}$.

Figure 1.7.: Two draws from the prior with $\nu = \frac{11}{10}$. 
From this, we can conclude that the tuning parameter $\nu$ changes the bandwidth and the smoothness of the regression kernel, which is seen in the prior distribution. As seen in the posterior estimations, for the considered $f_0$ we have an undersmooth prior for values of $\nu$ smaller than the smoothness of the truth, and an oversmooth prior for values of $\nu$ greater than the smoothness of the truth.

Regarding the inference of the posterior mean, we are interested in its consistency and contraction rates. If a posterior distribution is consistent, its posterior mean will converge to the true function $f_0$. Let $\mathcal{F}$ be the function space of interest, then the posterior distribution $\Pi_n(\cdot \mid X_n)$ is said to be weakly consistent at $f_0 \in \mathcal{F}$ if $\Pi_n(f : d(f, f_0) > \epsilon \mid X_n) \to 0$ in $\mathbb{P}_{f_0}$ probability as $n \to \infty$ for all $\epsilon > 0$, with strong convergence holding if the convergence is in the almost sure sense [10].

Next to the convergence towards the truth, we are interested in the rate at which this convergence happens. The contraction rate gives an upper bound for this rate. The posterior distribution $\Pi_n(\cdot \mid X_n)$ is said to contract at rate $\epsilon_n \to 0$ at $f_0 \in \mathcal{F}$ if $\Pi_n(f : d(f, f_0) > M_n \epsilon_n \mid X_n) \to 0$ in $\mathbb{P}_{f_0}$ as $n \to \infty$ for every $M_n \to \infty$. This can be interpreted as that the posterior distribution concentrates in balls of radius $\epsilon_n$ around the true function $f_0$ [10].

A method to derive contraction rates using rescaled Gaussian processes was shown in [29], based on a general method to prove contraction rates widely used from [30]. Unfortunately, this method does not pose a way to analogously prove the contraction rate when using the distributed method. However, [33] uses direct computations to find upper bounds for the bias and the variance of the posterior mean, which we will apply to the distributed method.

From Corollary 2.1 of [33] follows that when using the Matérn covariance kernel with probability at least $1 - n^{-10}$ for $h$ equal to (1.9), we have

$$
\left\| \hat{f}_n - f_0 \right\|_{\infty} \lesssim h^{\alpha} + \sigma \sqrt{\frac{\log n}{nh}} = B^{\frac{1}{2\alpha+1}} \left( \frac{\sigma^2 \log n}{n} \right)^{\frac{\alpha}{2\alpha+1}},
$$

for $f \in \Theta_{S}^{\alpha+\frac{1}{2}}(B)$ or $f \in \Theta_{H}^{\alpha}(B)$. The minimax rate is $n^{-\frac{\alpha}{2\alpha+1}}$ [25], so we achieve the minimax rate up to a logarithmic term. Here $a_n \lesssim b_n$ denotes $a_n \leq Cb_n$ for $C$ a constant.
that is universal or fixed in the context. We will use this and throughout the paper, where if we have both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$, we will write $a_n \asymp b_n$ to say $a_n$ is bound both from above and below, up to constants, by $b_n$.

Let us define the half length $l_n(x; \beta)$ by

$$l_n(x; \beta) := z(1+\beta)/2 \sqrt{\hat{C}_n(x, x)},$$

dependent on the posterior covariance kernel (1.6). Then $CI_n(x; \beta)$ denotes the pointwise posterior credible set of probability $\beta$ at the point $x$ defined as

$$CI_n(x; \beta) := (\hat{f}_n(x) - l_n(x; \beta), \hat{f}_n(x) + l_n(x; \beta)).$$

By Corollary 3.1 of [33] the pointwise posterior credible sets in the nondistributed setting that for a correct smoothness of the kernel $\alpha$, for every $\beta \in (0, 1)$ there exists a sequence of functions $\{f_{0,n} \}_{n \in \mathbb{N}} \in \Theta_H^\alpha(B)$ such that

$$P(f_{0,n}(x) \in CI_n(x; \beta)) \to \tilde{\beta}, \quad \text{as } n \to \infty,$$

for $z(1+\beta)/2$ chosen such that

$$\Pi(f(x) \in CI_n(x; \beta)|X_n, Y_n) = \tilde{\beta}.$$
2. Distributed Bayesian methods using Gaussian process priors

In the previous chapter we recalled that the computational complexity of both the posterior mean as the posterior covariance is of order $O(n^3)$, which explodes for larger values of $n$. For the matrix operations which require large amounts of computations, there are in many cases approximations available [11, 22]. In our analysis we turn to distributed methods to decrease the computation time. This is performed by dividing the data of size $n$ over $m$ (where we will, for simplicity assume $n \mod m = 0$) machines and calculate $m$ local posteriors which we will aggregate into a global posterior in different ways.

We will start with the same problem where we have a data set $\{X_n, Y_n\}$ which we divide over $m$ subsets $\{X_{n/m}^{(1)}, Y_{n/m}^{(1)}\}, \{X_{n/m}^{(2)}, Y_{n/m}^{(2)}\}, \ldots, \{X_{n/m}^{(m)}, Y_{n/m}^{(m)}\}$ and send these to $m$ machines. These $m$ machines will use the Bayesian method with Gaussian process priors in the nonparametric regression model on their local data to compute a local posterior $\Pi_L^{(j)}(\cdot | X_{n/m}^{(j)}, Y_{n/m}^{(j)})$. Means and variances are being computed of these local posteriors using Lemma 1. Afterwards, a parent machine computes the global posterior by combining all these local posteriors. This process can be applied recursively up to arbitrary many levels, but we will focus on the case where we have one parent machine and $m$ sub-machines. All local posteriors are trained parallelly and share the same parameters $\nu, \lambda$ where $\nu$ is taken equal to the smoothness of the truth.

2.1. Naive averaging of posteriors

As a starting point which forms our baseline case to compare the other methods to, we analyse a naive averaging distributed approach in which in every local problem we simply use the prior $\Pi(\cdot)$ defined in (1.2). Every local machine then computes its local posterior $\hat{f}_j^{(j)}(x^*) | X_{n/m}, Y_{n/m} \sim N(\mu_j(x^*), s_j^2(x^*))$ where

\begin{align*}
\mu_j(x^*) &= (k_s^{(j)})^T (K^{(j)} + nm^{-1} \lambda I)^{-1} Y_{n/m}^{(j)}, \\
s_j^2(x^*) &= \sigma^2 (nm^{-1} \lambda I)^{-1} (k_s^{(j)})^T (K^{(j)} + nm^{-1} \lambda I)^{-1} k_s^{(j)},
\end{align*}

where $f_{0, \sigma}$ for this model is given as a multiple of $e^{-\frac{1}{2} \sum_{i=1}^n (y_i^{(j)} - f_0(x_i))^2}$.

Locally considering the same prior and likelihood as in Section 1.1 we will have local posteriors $\hat{f}_{n/m}^{(j)}(x^*) | X_n, Y_n, k \sim N(\mu_j(x^*), s_j^2(x^*))$ where

\begin{align*}
\mu_j(x^*) &= (k_s^{(j)})^T (K^{(j)} + nm^{-1} \lambda I)^{-1} Y_{n/m}^{(j)}, \\
s_j^2(x^*) &= \sigma^2 (nm^{-1} \lambda I)^{-1} (k_s^{(j)})^T (K^{(j)} + nm^{-1} \lambda I)^{-1} k_s^{(j)},
\end{align*}
where $K^{(j)}$ and $k^{(j)}$ are similar to the corresponding values in Lemma 1, but based on
the data subset \( \{ X_{n/m}^{(j)}, Y_{n/m}^{(j)} \} \). To have our local posteriors satisfying the upper bound
in a minimax sense, we take

\[
h = \left( \frac{B^2 n/m}{\sigma^2 \log(n/m)} \right)^{-\frac{1}{2n+1}},
\]

for $f \in \Theta^{\alpha+\frac{1}{2}}_S(B)$ or $f \in \Theta^{\alpha}_H(B)$.

In this method, we take a draw from $f_{n/m}^{(j)}(x^*)$ for every local posterior $\Pi_{\cdot|X_{n/m}^{(j)}, Y_{n/m}^{(j)}}$.
Afterwards one takes the average of these $m$ draws to obtain global posterior predictive
distribution

\[
f_{n,m}(x^*) = \frac{1}{m} \sum_{j=1}^{m} f_{n,m}^{(j)}(x^*).
\]

Note that for $f_{n,m}^{(j)}(x^*) \sim N(\mu_j(x^*), s_j^2(x^*))$ we have

\[
\frac{1}{m} \sum_{j=1}^{m} f_{n,m}^{(j)}(x^*) \sim N \left( \frac{1}{m} \sum_{j=1}^{m} \mu_j(x^*), \frac{1}{m^2} \sum_{j=1}^{m} s_j^2(x^*) \right).
\]

Applying this to our Bayesian nonparametric regression method, we gain

\[
f_{n,m}(x^*) | X_n, Y_n \sim N \left( \frac{1}{m} \sum_{j=1}^{m} \mu_j(x^*), \frac{1}{m^2} \sum_{j=1}^{m} s_j^2(x^*) \right).
\]

As we consider multiple different distributed methods, which we will denote the current
method by method $I$, for which we obtain the posterior covariance function

\[
\tilde{C}_{n,m}^{(I)}(x, x') = \sigma^2(n\lambda)^{-1} \sum_{j=1}^{m} \left( k(x, x') - k(x, X_{n/m}^{(j)})(K^{(j)} + (n/m)\lambda I)^{-1}k(X_{n/m}^{(j)}, x') \right),
\]

and posterior mean

\[
\hat{f}_{n,m}^{(I)} = \mathbb{E}[f_{n,m}^{(I)} | X_n, Y_n].
\]

Now, using this for a data set of size $n = 2000$ with parameters $m = 100$, $\sigma = \frac{1}{2}$ and
$\nu = \frac{3}{2}$, we will compare the distributed method to the nondistributed method for the
same data set.

17
From the simulation study, we conclude that this method does not perform well either for posterior estimation or pointwise posterior uncertainty quantification. The posterior mean is smoother than in the nondistributed case, which tells us that the optimal tuning parameters for the local problems might return good results for the local problems, but provide an overly smooth estimator globally. We also see that the credible sets provide worse uncertainty quantification, which is apparent from (2.4) and confirms the suboptimal bias-variance trade-off.

**Theorem 1 (Naive averaging bounds).** Consider a covariance matrix $K$ corresponding to the Matérn kernel for a certain smoothness parameter $\alpha$. Let $h$ be equal to the value (2.3). Then there exists a function in $f_0 \in \Theta^\alpha_B(B)$ such that as $(n/m)^{1-2\alpha \alpha} m^{1/2} \to 0$

$$\left\| \hat{f}_{n,m} - f_0 \right\|_\infty \gtrsim \left( \frac{\log(n/m)}{n/m} \right)^{\frac{1}{2\alpha+1}} / \log \left( \left( \frac{n/m}{\log(n/m)} \right) \right),$$

(2.6)

with probability at least $1 - (2m + 2)(n/m)^{-10}$ with respect to the randomness of the data $X_n, Y_n$, and $f_0 \in \Theta^{\alpha + \frac{1}{2}}_S(B)$ such that as $(n/m)^{1-2\alpha \alpha} m^{1/2} \to 0$

$$\left\| \hat{f}_{n,m} - f_0 \right\|_\infty \gtrsim \left( \frac{\log(n/m)}{n/m} \right)^{\frac{1}{2\alpha+1}} / \log \left( \left( \frac{n/m}{\log(n/m)} \right) \right),$$

(2.7)

with probability at least $1 - (2m + 2)(n/m)^{-10}$ with respect to the randomness of the data $X_n, Y_n$.

**Proof.** The proof of this theorem is given in Section 4.5.

From this theorem, we can conclude that the $L_\infty$ loss of the posterior mean for certain truths is suboptimally large with high probability. This confirms the suboptimal estimation as seen in the simulation experiments.
Theorem 2 (Naive averaging pointwise credible set coverage). Let $h$ be equal to the value (2.3). Then, for any $x \in [0, 1]$ there always exists a function $f_{bad} \in \Theta_{S+1}^2(B)$ such that as $n/m \rightarrow \infty$ and $m/\log^2(n) \rightarrow \infty$ we have
\[ P(f_{bad}(x) \in CI_{n,m}(x)) \rightarrow 0. \]

Proof. The proof to this theorem is given in Section 4.6.

These two theorems now prove the conclusions from the simulation study. The distributed method using naive averaging does not provide an optimal rate of convergence towards the truth, nor does it provide reliable uncertainty quantification.

2.2. Aggregation by product of experts

We saw in the last section that simply taking draws from all separate posteriors and taking an average returns suboptimal results. There are now two options to change the estimation procedure, by either changing the local posteriors, or changing the method of aggregation of these local posteriors. This section will focus on the latter approach.

We can aggregate the posteriors differently by looking at the product of all local posteriors $\prod_{j=1}^{m} d\Pi_{E}^{(j)}(f(x^*) | X_{n/m}^{(j)}, Y_{n/m}^{(j)})$ and normalising this product, which then once more returns a probability measure as proposed in [18]. To do so, we will first look at the product of two Gaussian densities with parameters $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$, respectively.

This is given, up to a constant, by
\[
\exp \left[ -\frac{1}{2} \frac{(x - \mu_1)^2}{\sigma_1^2} \right] \exp \left[ -\frac{1}{2} \frac{(x - \mu_2)^2}{\sigma_2^2} \right] \propto \exp \left[ -\frac{1}{2} \frac{((x^2 - 2x\mu_1)\sigma_2^2 + (x^2 - 2x\mu_2)\sigma_1^2)}{\sigma_1^2 \sigma_2^2} \right] \\
\propto \exp \left[ -\frac{1}{2} \frac{(x^2 - 2x\mu_1)\sigma_2^2 + (x^2 - 2x\mu_2)\sigma_1^2}{\sigma_1^2 \sigma_2^2} \right] \\
\propto \exp \left[ -\frac{1}{2} \frac{x^2 - 2x\mu_1 \sigma_2^2 + \mu_2 \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right] \\
\propto \exp \left[ -\frac{1}{2} \frac{x^2 - 2x\mu_1 \sigma_2^2 + \mu_2 \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right].
\]

Also, we have
\[
\exp \left[ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right] \propto \exp \left[ -\frac{1}{2} \frac{x^2 - 2x\mu}{\sigma^2} \right],
\]
therefore we can conclude that equation (2.8) is proportional to a Gaussian density with
Thus, normalising the product of two Gaussians gives us a Gaussian with these parameters. Generalising this procedure for a product of $m$ Gaussians we find that the normalised product is again Gaussian [3], with parameters

$$
\mu = \frac{\mu_1 \sigma_2^2 + \mu_2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2} = \sigma^2 \left( \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2} \right),
$$

$$
\sigma^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.
$$

Using the same parameters and data set as for the naive averaging method, we now obtain the following simulation results.

**Figure 2.3.** Posterior mean $\hat{f}_{n,m}$ for $n = 2000$, $m = 100$, $\sigma = \frac{1}{2}$ and $\nu = \frac{3}{2}$.

**Figure 2.4.** Nondistributed $\hat{f}_{n,m}$ posterior using equal data.

Here we see the result still being much smoother than in the nondistributed case. Thus the present procedure does not give an optimal result. This suggests that using these aggregation methods the local data does not contain sufficient information for the local machines to compute a good estimation. From now on we will refer to $Na$ and $Np$ for method $N$ using aggregation by naive averaging and product of experts respectively.
2.3. Adjusted local likelihoods

To compensate for the underrepresentation of data in the local problems, we can raise the power of the local likelihoods. Recall that our local likelihoods are of the form

\[ p_{f_0, \sigma}(Y_{n/m}^j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\sum_{i=1}^{n/m} (y_{n/m}^j(i) - f_0(x_{n/m}^j(i)))^2}{2\sigma^2}}. \]

Thus, raising this to the power of \( m \) as proposed in [25], we obtain the modified local likelihood of

\[ (p_{f_0, \sigma}(Y_{n/m}^j))^m \propto e^{-\frac{m\sum_{i=1}^{n/m} (y_{n/m}^j(i) - f_0(x_{n/m}^j))}{2\sigma^2}}. \]

Computing the posterior in this case now easily follows from Lemma 1, but with \( \sigma^2 \) replaced by \( m^{-1}\sigma^2 \). Thus we now have our local posteriors of the form

\[ (f_{II}^{(j)}(x^*) \mid X_{n/m}^{(j)}, Y_{n/m}^{(j)} \sim N(\mu_j(x^*), \sigma_j^2(x^*)). \]

where

\[ \mu_j(x^*) = (k_{s*}^{(j)})^T(K^{(j)} + nm^{-1}\lambda I)^{-1}Y_{n/m}^{(j)}, \] (2.11)

\[ \sigma_j^2(x^*) = \sigma^2(n\lambda)^{-1}(k_{s*}^{(j)})^T(K^{(j)} + nm^{-1}\lambda I)^{-1}k_{s*}^{(j)}). \] (2.12)

from which follows here

\[ h = \left( \frac{B^2n}{\sigma^2 \log(n/m)} \right)^{-\frac{1}{2\nu+1}}, \] (2.13)

for \( f \in \Theta_S^{\alpha+\frac{1}{2}}(B) \) and \( f \in \Theta_H^{\alpha}(B) \), as both the denominator and the divisor contain a term of \( m^{-1} \) in this setting which cancel each other out.

Using this method, for the same parameters as in earlier sections, \( n = 2000, m = 100, \nu = \frac{3}{2} \) and \( \sigma^2 = \frac{1}{2} \), we obtain the following results.

![Figure 2.5: \( \hat{f}_{IIa}^{(j)} \) using the adjusted local likelihoods method, aggregating by averaging the posteriors.](image)

![Figure 2.6: \( \hat{f}_{IIp}^{(j)} \) using the adjusted local likelihoods method, aggregating by product of experts.](image)
One can observe that this method indeed improves the earlier naive methods, as the estimation is more accurate. However, the credible sets in this case are very small and barely visible, thus do not cover \( f_0 \) well, which means there is still room for improvement. Note that using this different setup, we do see difference between the naive averaging and product of experts methods, with the product of experts method providing a better estimation than the naive averaging method.

**Theorem 3** (Adjusted local likelihoods and averaging bounds). Consider a covariance matrix \( K \) corresponding to the Matérn kernel for a certain smoothness parameter \( \alpha \). Let \( h \) be equal to the value (2.11). Then, for every function \( f_0 \in \Theta_\alpha(B) \) or \( f_0 \in \Theta_H(B) \) we obtain as \( n^{1-2\alpha} m^{\frac{1}{2\alpha} - \frac{1-2\alpha}{\alpha}} \to 0 \)

\[
\|\hat{f}_{n,m} - f_0\|_{\infty} \lesssim \left(\frac{\log(n/m)}{n}\right)^{\frac{\alpha}{2\alpha+1}}.
\]  

with probability at least \( 1 - (2m+2)(n/m)^{-10} \) with respect to the randomness of the data \( X_n, Y_n \).

**Proof.** The proof to this theorem is given in Section 4.7.

This shows us the posterior mean converges to the truth of similar order up to a logarithmic term to the nondistributed case, which gives us an optimal estimation in the minimax sense. However, the coverage of the pointwise posterior credible set is suboptimal as shown in the following theorem.

**Theorem 4** (Adjusted local likelihoods and averaging pointwise credible set coverage). Suppose \( h \) chosen minimax optimal as in (2.3). Then, for every \( x \in [0,1] \) there always exists a function \( f_{bad} \in \Theta_\alpha(B) \) such that as \( n/m \to \infty \) and \( m/\log^2(n) \to \infty \) we have

\[
P(f_{bad}(x) \in CI_{n,m}(x)) \to 0.
\]

**Proof.** The proof to this theorem is given in Section 4.8.

This shows us that the pointwise posterior credible intervals are sub-optimal and there is still room for improvement to this method. The simulation studies suggest that aggregation by product of experts does not improve the procedure.

### 2.4. Adjusted priors

Raising the likelihood to the \( m \)’th power poses a solution to the rate of convergence of the posterior as seen in last section. However, using this setup, we saw there was still unreliable uncertainty quantification using the aggregation methods considered. This leaves us to find a procedure giving us reliable uncertainty quantification as well as optimal recovery. Similar to the method used before where the likelihood was raised to the power of \( m \), we can consider raising the prior density to the \( 1/m \)’th power as proposed in [23].
We have seen that raising a Gaussian density to the $1/p$th power is equal to a multiplying the variance by a factor $p$. This also holds for our prior density, where in this case raising the power of the prior density to $1/m$ corresponds to the covariance matrix being multiplied by the factor $m$:

$$
\left(\left|\frac{1}{2\pi\Sigma}\right|^{-\frac{1}{2}}\right)^{\frac{1}{m}} e^{-\frac{1}{2m}(x-\mu)^T\Sigma^{-1}(x-\mu)} \propto e^{-\frac{1}{2}(x-\mu)^T(m\Sigma)^{-1}(x-\mu)}.
$$

Thus in this setting, raising the power of the prior density to $1/m$ is equal to changing the Gaussian process prior to $GP(0,\sigma^2(nm^{-2}\lambda)^{-1}K)$. This is equivalent to rescaling the function $k$ as $k^{I\!I\!I} = mk$.

Computing the posterior predictive and its 95% credible sets using the adjusted prior, we obtain the following results.

Figure 2.7.: $\hat{f}_{n,m}^{I\!I\!I}a$ using the adjusted priors method, aggregating by averaging the posteriors for $n = 2000$, $m = 100$, $\sigma = \frac{1}{2}$ and $\nu = \frac{3}{2}$.

Figure 2.8.: $\hat{f}_{n,m}^{I\!I\!I}p$ using the adjusted local likelihoods method, aggregating by product of experts.

Figure 2.9.: $||\hat{f} - f_0||^2_N$ on $[0,1]$ for different values of $\sigma^2$. 

From the simulations we see that rescaling the prior improves the estimation as well as the coverage of the pointwise posterior credible set. Once more, the aggregation of product experts seems to give a slightly better result than the naive averaging method, the same as perceived in last section. The posterior spread specifically seems to differ between the two aggregation methods. This can be explained because we have

$$\sum_{j=1}^{m} \left( \frac{\sigma_j^2}{m} \right)^{-1} \leq \frac{1}{m} \sum_{j=1}^{m} \sigma_j^2,$$

by the harmonic-arithmetic mean inequality. Multiplying both sides by $1/m$ results in a left hand side equal to the variance in the product of experts setting, and a right hand side equal to the variance in the averaging setting. This confirms the observations.

**Theorem 5** (Adjusted priors and averaging bounds). Consider a covariance matrix $K$ corresponding to the Matérn kernel for a certain smoothness parameter $\alpha$. Let $h$ be equal to the value (2.3). Then, for every function $f_0 \in \Theta_{\alpha+\frac{1}{2}}^{\frac{1}{2}}(B)$ we obtain as $n^{-\frac{1}{\alpha}} m^{\frac{2-\alpha}{2\alpha}} + 2^{\alpha+1} \rightarrow 0$

$$\|\hat{f}^{IIa}_{n,m} - f_0\|_{\infty} \lesssim \left( \frac{\log(n/m)}{n} \right)^{\frac{n}{\alpha+1}},$$

with probability at least $1 - (2m+2)(n/m)^{-10}$ with respect to the randomness of the data $X_n, Y_n$.

**Proof.** The proof to this theorem is given in Section 4.9.

As we have seen in the simulation results, this method does not only offer an optimal sup-norm upper bound, but also point-wise credible set coverage of the truth. This is proven in the next theorem.

**Theorem 6** (Adjusted priors and averaging bounds pointwise credible set coverage). For any $\tilde{\beta} \in (0,1)$ and $x \in [0,1]$, there is some sufficiently large constant $B > 0$ and a sequence of functions $\{f_n\}_{n \in \mathbb{N}}$ where every $f_n \in \Theta_{\alpha+\frac{1}{2}}^{\frac{1}{2}}(B)$, such that

$$\mathbb{P}(f_n(x) \in CI^{IIa}_{n,m}(x;b)) \rightarrow \tilde{\beta},$$

as $m, n/(m^4) \rightarrow \infty$.

**Proof.** The proof to this theorem is given in Section 4.10.

### 2.5. Adjusted local likelihoods and Wasserstein barycenters

As we have seen, adjusting the local likelihoods improves recovery, but results in questionable uncertainty quantification. A way to improve this could be to try another
method of aggregating posteriors, for instance computing their Wasserstein barycenters [25].

To do so, we look at the 2-Wasserstein distance between two probability measures $\mu$ and $\nu$ which is defined as

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

where the infimum is defined over all measures $\gamma$ on $[0, 1] \times [0, 1]$ with marginals $\mu$ and $\nu$. Furthermore, the 2-Wasserstein barycenter between $m$ probability measures $\mu_1, \mu_2, \ldots, \mu_m$ is defined as

$$\bar{\mu} = \arg\min_{\mu} \sum_{i=1}^{m} W_2^2(\mu, \mu_i).$$

From [25] we know that the Wasserstein barycenter of $m$ Gaussians is again Gaussian, with a mean equal to the average mean of the $m$ Gaussians, and a variance parameter equal to the average variance of the Gaussians. This gives a similar result to averaging the posteriors, but has a variance term that is $m$ times as large:

$$\hat{f}_{IV}^{\mu_{n,m}}(x^*) \mid X_n, Y_n \sim N \left( \frac{1}{m} \sum_{j=1}^{m} \mu_j(x^*), \frac{1}{m} \sum_{j=1}^{m} s_j^2(x^*) \right). \quad (2.16)$$

As this variance term is $m$ times as large as the naive averaging variance, we see that applying this aggregation method after raising the power of the likelihood compensates for the small posterior spread of using that method and aggregating by taking the average. Therefore we expect to see optimal results for both the bias and the coverage of the pointwise posterior credible sets.

Figure 2.10.: $\hat{f}_{IV}^{\mu_{n,m}}$ on $[0,1]$ using the adjusted local likelihoods with aggregation by Wasserstein barycenter for $n = 2000, m = 100, \sigma = \frac{1}{2}, \nu = \frac{3}{2}$. 25
2.6. Generalised Product of Experts

Another procedure we can apply to the problem is the generalised product of experts method [5]. This method is based around the same principle of aggregation by product of experts with a slight adjustment. By raising the local posteriors to the power $a_j$, we again get a Gaussian posterior, but with new parameters:

$$
\mu(x^*) = \sigma^2(x^*) \sum_{j=1}^{m} \frac{a_j \mu_j(x^*)}{\sigma_j^2(x^*)},
$$

$$
\sigma^2(x^*) = \left( \sum_{j=1}^{m} a_j \sigma_j^{-2}(x^*) \right)^{-1}.
$$

Because this allows for infinitely many combinations, we will look at the case where $a_j = a$ for all $j$. We will look at the estimation $\hat{f}_{V,n,m}$ for different values of $a$ for the same data set $X_n, Y_n$, with a suggested value of $a = 1/m$ [7].

Figure 2.11.: $||\hat{f} - f_0||^2_N$ per value of $a$.

Here we notice that the $a$ does not seem to have an obvious relation for the contraction of the estimation, but do seem to have an influence on the credible set coverage.

Figure 2.12.: $\hat{f}_{V,n,m}$ on $[0,1]$ for $n = 2000$, $m = 100$, $a = 1/m$, $\sigma = \frac{1}{2}$ and $\nu = \frac{3}{2}$.
From here, we can also conclude that the value of $a$ does not significantly change the estimation of $f_0$, but only the size of the credible sets. Similar to the product of experts aggregation method from Section 2.2 this method also returns suboptimal recovery.

### 2.7. The Bayesian Committee Machine

The Bayesian Committee Machine proposes another option to aggregate posteriors [28]. For this aggregation method we once more divide $X_n, Y_n$ into $m$ subsets which we denote in this setting as $\{D_j\}_{j=1}^m \equiv \{\{X_{n/m}^{(j)}, Y_{n/m}^{(j)}\}\}_{j=1}^m$. Moreover, we define $D_i = \{D_1, \ldots, D_i\}$. Now we are interested in $d\Pi(f^q | D_{i-1}, D_i)$, where $f^q$ is the vector of unknown target measurements corresponding to a given set of test points $x_{q_1}, \ldots, x_{q_N}$.

From Bayes’ rule we know that this is proportional to

$$
d\Pi(f^q) p_{f_0, \sigma}(D_{i-1} | f^q) p_{f_0, \sigma}(D_i | D_{i-1}, f^q).
$$

(2.19)

In this equation, we will now make the approximation

$$
p_{f_0, \sigma}(D_i | D_{i-1}, f^q) \approx p_{f_0, \sigma}(D_i | f^q).
$$

Only conditioned on the whole function, when $f^q \equiv f$, we have that $D_i$ is independent of $D_{i-1}$. The approximation might still be reasonable for large $N$, or when the correlation between $D_i$ and $D_{i-1}$ is very small. By [28] we obtain

$$
d\Pi(f^q | D_{i-1}, D_i) \propto d\Pi(f^q) p_{f_0, \sigma}(D_{i-1} | f^q) p(D_i | D_{i-1}, f^q)$$

$$
\approx d\Pi(f^q) p_{f_0, \sigma}(D_{i-1} | f^q) p_{f_0, \sigma}(D_i | f^q)$$

$$
\propto d\Pi(f^q) \frac{p_{f_0, \sigma}(D_{i-1} | f^q) p_{f_0, \sigma}(D_i | f^q)}{d\Pi(f^q)}$$

$$
\propto \frac{d\Pi(f^q | D_{i-1}) d\Pi(f^q | D_i)}{d\Pi(f^q)}.
$$
Using the fact that $D_{n-1} \cup D_n = D$, we obtain

$$d\Pi(f^q \mid D) \prod_{i=1}^{n-1} d\Pi(f^q \mid D_{i-1}, D_i) = \prod_{i=1}^{n} d\Pi(f^q \mid D_{i-1}, D_i)$$

$$\approx C \prod_{i=1}^{n} \left( \frac{d\Pi(f^q \mid D_{i-1}) d\Pi(f^q \mid D_i)}{d\Pi(f^q)} \right)$$

$$= C \prod_{i=1}^{n} \frac{d\Pi(f^q \mid D_i)}{d\Pi(f^q)^{M-1}}$$

from which follows

$$d\Pi(f^q \mid D) \propto \frac{\prod_{i=1}^{m} d\Pi(f^q \mid D_i)}{d\Pi(f^q)^{m-1}} . \quad (2.20)$$

We know that $d\Pi(f^q \mid D_i)$ and $d\Pi(f^q)$ are both Gaussian in our regression setting, thus we need to know the product of multivariate Gaussian densities and the fraction of multivariate Gaussian densities to explicitly compute the approximate posterior. The product of two multivariate distributions with parameters $\mu_1, \Sigma_1$ and $\mu_2, \Sigma_2$ is proportional to a multivariate Gaussian with parameters

$$\mu = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2),$$

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} .$$

For the fraction of two multivariate Gaussian densities we obtain a multivariate Gaussian density with parameters [27]

$$\mu = (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \mu_1 - \Sigma_2^{-1} \mu_2),$$

$$\Sigma = (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1} .$$

As observed earlier, raising a Gaussian density to the $(m - 1)'th$ power results in a Gaussian density with a variance parameter multiplied by $m - 1$. For a prior with parameters $0, K$, this results in the divisor equal to a density with parameters $0, (m-1)K$. Furthermore, $\prod_{i=1}^{m} d\Pi(f^q \mid D_i)$ is proportional to a multivariate Gaussian distribution with parameters

$$\mu = \Sigma \sum_{i=1}^{m} \Sigma_i^{-1} \mu_i,$$

$$\Sigma = \left( \sum_{i=1}^{m} \Sigma_i^{-1} \right)^{-1} .$$

Combining these two, we see that $\prod_{i=1}^{m} \frac{d\Pi(f^q \mid D_i)}{d\Pi(f^q)^{m-1}}$ is proportional to a multivariate Gauss-
sian distribution with parameters

\[ \mu = \sum_{i=1}^{m} \Sigma_i^{-1} \mu_i, \quad (2.21) \]

\[ \Sigma = \left( \left( \sum_{i=1}^{m} \Sigma_i^{-1} \right)^{-1} - (m - 1)K \right)^{-1}. \quad (2.22) \]

Applying this to the earlier data, we obtain the following result.

![Figure 2.15.: \( \hat{f}_{VI}^{n,m} \) on [0,1] using the Bayesian committee machine for \( n = 2000, m = 100, \sigma = \frac{1}{2} \) and \( \nu = \frac{3}{2} \).](image)

From this we can conclude that this aggregation method has both optimal contraction as well as coverage of the pointwise posterior credible sets. One point to note is that there is a large probability of encountering numerical errors in the calculation resulting in \( \Sigma \) possibly not positive semi-definite, meaning there are some eigenvalues close to 0 with imaginary parts. This problem occurs regularly when creating \( K \) using the Matérn kernel, as well as when computing \( \sigma_i \) for a local machine \( i \) (this problem has not showed up yet, because we evaluated the function at different values of \( x^* \) separately, instead of drawing the whole function from a Gaussian distribution). For a degenerate matrix \( A \) which in theory should be positive semi-definite, one can use \( A + cI \) for a small enough \( c \), as this causes all eigenvalues to be bounded away from 0.

### 2.8. Summary of distributed methods

We have seen that there are many methods available to improve the naive averaging of posteriors. Various methods were introduced to improve performance of the distributed method by changing the computation of the local posteriors, changing the aggregation method of the local posteriors or a combination of both with mixed results. The naive
averaging and product of experts, do not show a significant difference in terms of recovery as well as uncertainty quantification. The only obvious difference between the methods are the slightly rougher paths from the product of experts method, which is also apparent in the theory by the arithmetic-harmonic mean inequality. For methods I, IIa, IIIa we have proven the sup-norm bounds for the $L_\infty$ distance between the estimator and the truth and frequentist coverage of the pointwise posterior credible sets. For all other methods, we have seen simulation studies which suggest likely results for the distance between the truth and the estimator, and frequentist coverage of the pointwise posterior credible sets. We summarise the result of the simulation studies in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Optimal recovery</th>
<th>Uncertainty quantification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method Ia: Naive averaging</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Method Ip: Product of experts</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Method IIa: Adjusted likelihoods + NA</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Method IIp: Adjusted likelihoods + PoE</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Method IIIa: Adjusted priors + NA</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Method IIIp: Adjusted priors + PoE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Method IV: Adjusted likelihoods + Wasserstein barycenter</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Method V: Generalised product of experts</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Method VI: Bayesian committee machine</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2.1.: Performance of the different estimation methods using the Matérn kernel.

From the simulation studies, one can see that methods IIIa, IIIp, IV and VI show results comparable to the nondistributed method. Methods Ia, Ip and V show suboptimal estimation of the truth, but V does show good frequentist coverage of the pointwise posterior credible sets. Methods IIa and IIp show optimal estimation, but too small pointwise posterior credible sets, giving suboptimal frequentist coverage of the pointwise posterior credible sets.

One does have to keep in mind that all these estimations are being made with complete knowledge of the smoothness $\alpha$. For an unknown value of $\alpha$, adaptive methods are needed to estimate the smoothness from the raw data. Methods to adaptively estimate this smoothness in a nondistributed setting have been investigated in [18, 25]. We will see that this results in suboptimal estimations for all methods in the distributed case in Section 3.2.
3. Distributed Bayesian methods using the squared exponential kernel

An initial goal was to improve the distributed methods from [18], which work similar to the distributed methods we have considered in Chapter 2, but using the squared exponential kernel instead. As [31] states that the optimal value of the bandwidth parameter is of order $n^{-1/(2\alpha+1)}$, this leads us to define the squared exponential kernel as

$$k(x_i, x_j) = e^{-\frac{1}{2}(x_i - x_j)^2 \tau^{-1} n^{-1/(2\alpha+1)}},$$

where $\tau$ is a hyper parameter influencing the bandwidth of the kernel.

However, when trying to follow the methods used in [33] to compute upper bounds for the posterior mean analogously for this kernel, a couple of complications occur. The main requirements for the theorems are polynomially decaying eigenvalues and uniformly bounded, Lipschitz continuous eigenfunctions. The eigenvalues of this covariance kernel are not polynomially decaying, but exponential, of the form $\mu_i \asymp \alpha e^{-i\alpha/2}$ [21], which does not satisfy the eigenvalue properties used in the proof for nondistributed setting in [33]. This implies that for the squared exponential kernel, theoretical results in the nondistributed setting are first required before proving them in the distributed setting.

Furthermore, the eigenfunctions of the kernel are of the form

$$\mu_k = C \sqrt{\frac{2\alpha}{A}} B^k,$$

$$\phi_k(x) = e^{-(c-a)x^2} H_k(\sqrt{2cx}),$$

where $H_k$ denotes the $k$’th order Hermite polynomial [12]. We only investigated the squared exponential kernel in a simulation study.

3.1. Simulation study

It is possible to apply the distributed methods investigated in the previous chapter using the squared exponential kernel. We expect to see similar behaviour for these methods in the new setting. We will start by computing the posterior mean using the nondistributed method to have a baseline case to compare the distributed methods to and see the influence of the parameter $\tau$. 

31
Figure 3.1.: The data \((X_n, Y_n)\) of size \(n = 500\) where 
\[x_i \sim U(0, 1)\] and 
\[y_i \sim f_0(x_i) + \epsilon\] where \(\epsilon \sim N(0, \frac{1}{2})\).

Figure 3.2.: \(\hat{f}_n\) on \([0,1]\) with the 95\% pointwise credible set for the optimal value of \(\tau\) in this setting which we denote by \(\tau^\ast\).

Figure 3.3.: \(\hat{f}_n\) on \([0,1]\) with the 95\% credible set for the value of \(10\tau^\ast\).

Figure 3.4.: \(\hat{f}_n\) on \([0,1]\) with the 95\% credible set for the value of \(\frac{1}{10}\tau^\ast\).

One can observe the influence of \(\tau\) on the posterior clearly. The bandwidth of the kernel shrinks for larger values of \(\tau\), whilst it grows for smaller values of \(\tau\). From this, it is apparent that both too large and too small values of \(\tau\) are not desired, as they respectively undersmooth and oversmooth the prior, resulting in sub-optimal estimators. Moreover the bias-variance tradeoff is apparent, where a smaller bias comes paired with higher variance.
Figure 3.5: $||\hat{f} - f_0||^2_{\mathcal{X}}$ on $[0,1]$ for different values of $\tau$ on the same data.

Figure 3.6: Two draws from the prior with $\tau = \tau^\ast$.

Figure 3.7: Two draws from the prior with $\tau = 10\tau^\ast$. 

33
We now consider the same distributed methods as used for the Matérn kernel in Chapter 2 for the squared exponential kernel. We do so by taking data of size $n = 2000$ with $\sigma_\varepsilon = 0.5$ and distribute the data over $m = 50$ machines.
From these simulations we can conclude that the methods perform similar for both covariance kernels. This is unsurprising as the same results were also perceived for the signal-in-white noise model from [25]. We conclude that similar to Matérn covariance kernel, we gain the following results for the squared exponential kernel.
<table>
<thead>
<tr>
<th>Method</th>
<th>Optimal recovery</th>
<th>Uncertainty quantification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method Ia: Naive averaging</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Method Ip: Product of experts</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Method IIa: Adjusted likelihoods + NA</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Method IIp: Adjusted likelihoods + PoE</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Method IIIa: Adjusted priors + NA</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Method IIIp: Adjusted priors + PoE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Method IV: Adjusted likelihoods +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wasserstein barycenter</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Method V: Generalised product of experts</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Method VI: Bayesian committee machine</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3.1.: Performance of the different estimation methods using the squared exponential kernel.

3.2. Distributed adaptive methods

Up until this point, we have assumed that the smoothness parameter $\alpha$ is known. In practice however, this might not be the case. There is thus need for estimation methods which perform without prior knowledge about $\alpha$. Several adaptive methods exist to estimate the parameter from data, of which we will consider by maximising the marginal likelihood [25, 18]. These methods have shown that in the nondistributed case there exists methods that tune the parameter $\alpha$ optimally using only knowledge from the data.

The marginal likelihood in this setting is of the form

$$|2\pi(\sigma^2 K + \sigma^2 I)|^{-\frac{1}{2}} e^{-\frac{1}{2} y^T (\sigma^2 K + \sigma^2 I)^{-1} y}.$$

The logarithm of the marginal likelihood is proportional to

$$-\frac{1}{2} (y^T (\sigma^2 K + \sigma^2 I)^{-1} y + \log |\sigma^2 K + \sigma^2 I|). \quad (3.1)$$

The nondistributed setting uses the value of $\tau$ which maximises this quantity. Numerically, we compute the marginal likelihood for many values of $\tau \in (0, C]$ for a sufficiently large constant $C$ and use the value which returns the largest marginal likelihood. Similarly, for the Matérn kernel, we compute the marginal likelihood for many values $\alpha \in (0.5, C]$ for some large enough constant $C$. In the nondistributed setting, this adaptive method performs well for large enough sample size $n$.

However, in the distributed setting there is no central machine containing all the information. In the distributed setting it was proposed that one should use the average of the local marginal likelihoods to derive an estimator for the regularity hyper parameter $\alpha$, see [18]. Then in [25] it was shown that this method performs sub-optimally, since
the local machines contain insufficient information. For the nonparametric regression model, an adaptive method is proposed in [33] by maximising the sum of the local marginal likelihoods

\[
\sum_{j=1}^{m} \left( -\frac{1}{2} \left( (Y_{n/m}^{(j)})^T (\sigma^2 K^{(j)} + \sigma^2 I)^{-1} Y_{n/m}^{(j)} + \log |\sigma^2 K^{(j)} + \sigma^2 I| \right) \right) .
\] (3.2)

Figure 3.17.: Adaptive estimation for the squared exponential kernel using Method I: Average of local estimated parameters

Figure 3.18.: Adaptive estimation for the squared exponential kernel using Method II: Median of local estimated parameters.

Figure 3.19.: Adaptive estimation for the squared exponential kernel using Method III: Maximising the sum of local likelihoods.

Figure 3.20.: Adaptive estimation for the Matérn kernel using Method I: Average of local estimated parameters.
One can conclude that distributed adaptive methods do not solve the local tuning of the parameters, which is also an open problem for the signal in white noise model [25]. Furthermore, the adaptive methods have shown numerical errors for too large sample sizes, in this setting from as small as $n = 600$, which forms an issue. Subsequent work could consider finding distributed adaptive methods and solving the numerical errors.

3.3. Summary

We have seen that the different distributed methods used in Chapter 2 show equal behaviour when applied with the squared exponential kernel. The theoretical properties of the estimations using this kernel seem to be harder to compute, as the eigenvalues decay at a different rate than the eigenvalues of the Matérn kernel. Moreover, the eigenfunctions of the squared exponential kernel do not satisfy the assumptions used in the proofs for the Matérn kernel.

We have performed a simulation study on methods that estimate the hyper parameters of the kernel from a set of data, instead of assuming these values on beforehand.
4. Proofs

In this chapter we provide the proofs of the theorems and lemmas stated in the previous chapters. Let us first begin with an outline of the notations, in which we will mostly follow the notation from Section 5.1 of [33].

4.1. Summary of notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_n/Y_n$</td>
<td>$x/y$ coordinate data of size $n$, ${x_i/y_i}_{i=1}^n$</td>
</tr>
<tr>
<td>$f_0$</td>
<td>true function</td>
</tr>
<tr>
<td>$w_i$</td>
<td>Gaussian error, $w_i = y_i - f_0(x_i) \sim N(0,\sigma^2)$</td>
</tr>
<tr>
<td>$\hat{f}_n$</td>
<td>posterior mean, $E(f \mid x_n, y_n)$</td>
</tr>
<tr>
<td>$C_n$</td>
<td>posterior covariance function, $\hat{C}_n(x, x') = \text{Cov}(f(x), f(x') \mid X_n, Y_n)$</td>
</tr>
<tr>
<td>${\nu_j}_{j \in \mathbb{N}}$</td>
<td>eigenvalues of the equivalent kernel, $\nu_j = \mu_j / (\mu_j + \lambda)$</td>
</tr>
<tr>
<td>$h$</td>
<td>tuning parameter determined defined as $h = \lambda^{2\alpha}$</td>
</tr>
<tr>
<td>$\hat{K}_N$</td>
<td>equivalent kernel, $\hat{K}<em>N(s, t) = \sum</em>{j=1}^{\infty} \nu_j^{N} \phi_j(s) \phi_j(t)$</td>
</tr>
<tr>
<td>$F_N$</td>
<td>convolution with equivalent kernel, $F_N g(t) = \int g(s) \hat{K}_N(s, t) ds$</td>
</tr>
<tr>
<td>$P_N$</td>
<td>$P_N f = f - F_N f$</td>
</tr>
<tr>
<td>$\hat{f}_{n,\lambda}$</td>
<td>KRR solution, $\arg\min_{f \in H} (n^{-1} \sum_{i=1}^{n} (y_i - f(x_i)) \hat{K}<em>{x_i} - P</em>{\lambda} f)$</td>
</tr>
<tr>
<td>$\hat{C}_n$</td>
<td>approximate covariance function, $\hat{C}_n = \sigma^2 h \hat{K}$</td>
</tr>
<tr>
<td>$A^N_m$</td>
<td>aggregation function over $m$ functions as used in method $N$</td>
</tr>
<tr>
<td>$\hat{f}^{(j)}_{n,m,\lambda}$</td>
<td>KRR solution of distributed method $N$, $A^N_m({\hat{f}^{(j)}<em>{n/m,\lambda}}</em>{j=1}^m)$</td>
</tr>
<tr>
<td>$S_{n,\lambda}$</td>
<td>sample score function, $S_{n,\lambda} f = n^{-1} \sum_{i=1}^{n} (y_i - f(x_i)) \hat{K}<em>{x_i} - P</em>{\lambda} f$</td>
</tr>
<tr>
<td>$\bar{S}_{n,m,\lambda}$</td>
<td>aggregated sample score function $A^N_m({S_{n/m,\lambda}}_{j=1}^m)$</td>
</tr>
<tr>
<td>$S_{\lambda}$</td>
<td>population score function $S_{\lambda} f = F_{\lambda} f_0 - f$</td>
</tr>
<tr>
<td>$\tilde{C}_{n,m}$</td>
<td>global posterior covariance function using method $N$</td>
</tr>
<tr>
<td>$\gamma_n$</td>
<td>$\max{1, n^{-1+1/(2\alpha)} h^{-1/(2\alpha)} \sqrt{\log n} } \sqrt{(nh)^{-1} \log n}$</td>
</tr>
</tbody>
</table>

Table 4.1.: Notation reference

We are using roman numbers to denote our different distributed methods. For any operator $M$, we use $M^N$ to denote operator $M$ in the setting of distributed method $N$, where if the roman number is left out, we refer to the nondistributed setting. For example, this means that $\hat{K}_N$ is only different to $\hat{K}$ in a setting where the eigenvalues
or eigenfunctions are changed. Furthermore $M^{(j)}$ denotes $M$ for the $j$’th local machine, for every symbol.

### 4.2. The RKHS framework

In many of the proofs we will use results on reproducing kernel Hilbert spaces (RKHS), of which further details and proofs can be found in Chapter 1 of [33].

A (real) RKHS $H$ is a Hilbert space of real-valued functions on a general index set $X$, in our setting $X = [0, 1]$, such that for any $t \in X$, the evaluation function $L_t : H \rightarrow \mathbb{R}; f \mapsto f(t)$, is a bounded linear functional. The function $L_t$ is a bounded linear functional if there exists a constant $C_t > 0$ such that

$$|L_t f| = |f(t)| \leq C_t \|f\|_H, \quad f \in H,$$

where $\|f\|_H = \sqrt{(f, f)_H}$ is the norm associated to the Hilbert space $H$. We call a symmetric function $K : [0, 1]^2 \rightarrow \mathbb{R}$ positive definite if for any $n \in \mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}$, and $t_1, \ldots, t_n \in [0, 1],$

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j K(t_i, t_j) > 0.$$

As some evaluation maps are bounded linear operators for $s, t \in [0, 1]$, Riesz’ representation theorem states that for every $t \in [0, 1]$, there exists an element $K_t \in H$ such that

$$f(t) = \langle f, K_t \rangle_H. \quad \text{This element } K_t \text{ is called the representer of evaluation at } t,$$

and the kernel $K(s, t) = K_t(s)$ is positive definite. Furthermore, given any positive definite function $K$, a unique RKHS on $[0, 1]$ with $K$ as its reproducing kernel, the kernel function $K$ for which $f(t) = \langle f, K_t \rangle_H$ holds, can be constructed. We will hence use the notation $K_t(\cdot) = K(t, \cdot)$ throughout for any given kernel $K$. Let $L^2([0, 1])$ denote the space of square integrable functions $f : [0, 1] \rightarrow \mathbb{R}$ which satisfy $\int_0^1 f^2(x)dx < \infty$. Note that for the uniform random variable $X$ on $[0, 1]$ we have $E[f(X)] = \int_0^1 f(x)dx$. The $L^2$ inner product on $[0, 1]$ of two functions is denoted as $(f, g)_{L^2} = \int_0^1 f(x)g(x)dx$. This lets us use Mercer’s theorem [6], which tells us that for a continuous positive definite kernel $K(s, t)$ satisfying $\int_0^1 \int_0^1 K(s, t)dsdt < \infty$, there exists an orthonormal sequence of continuous eigenfunctions denoted by $\{\phi_j\}_{j \in \mathbb{N}}$ in $L^2([0, 1])$ with eigenvalues $\mu_1 \geq \mu_2 \ldots \geq 0$, and

$$\int K(s, t)\phi_j(s)dt = \mu_j \phi_j(s), \quad (4.1)$$

$$K(s, t) = \sum_{j=1}^\infty \mu_j \phi_j(s)\phi_j(t). \quad (4.2)$$

Therefore the RKHS $H$ determined by $K$ is the set

$$\{f \in L^2([0, 1]) : \sum_{j=1}^\infty f_j^2/\mu_j < \infty\},$$
Let us define the kernel ridge regression (KRR) estimator

\[
\hat{f}_{n,\lambda} := \operatorname{argmin}_{f \in \mathcal{H}} \ell_{n,\lambda}(f), \quad \ell_{n,\lambda}(f) := \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}} \right],
\]

(4.3)

which coincides with the posterior mean, \( \hat{f}_n = \mathbb{E}[f | X_n, Y_n] \), in the nonparametric regression setting used [33] (see Subsection 6.2.2 from [21]). In our analysis we consider the KRR estimator instead of the posterior mean for convenience.

Let us define a new inner product on \( \mathcal{H} \) for a fixed \( \lambda > 0 \) as

\[
\langle f, g \rangle_{\lambda} := \langle f, g \rangle_{L^2} + \lambda \langle f, g \rangle_{\mathcal{H}},
\]

and define the norm as

\[
\|f\|_{\lambda} = \langle f, f \rangle_{\lambda}.
\]

Note that this is again a RKHS as we have \( |f(x)| \leq C_x |f|_{\mathcal{H}} \leq C'_x |f|_{\lambda} \) for some constants \( C_x, C'_x \) where the first inequality holds because \( \mathcal{H} \) is an RKHS which means \( |f(x)| \leq C_x |f|_{\mathcal{H}} \) for a certain constant \( C_x > 0 \).

For two elements \( f = \sum_{j=1}^{\infty} f_j \phi_j \) and \( g = \sum_{j=1}^{\infty} g_j \phi_j \) in \( L^2([0,1]) \), we obtain

\[
\langle f, g \rangle_{\lambda} = \langle f, g \rangle_{L^2} + \lambda \langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} f_j g_j + \lambda \sum_{j=1}^{\infty} \frac{f_j g_j}{\nu_j} = \sum_{j=1}^{\infty} \frac{f_j g_j}{\nu_j},
\]

(4.4)

for

\[
\nu_j = \frac{1}{1 + \frac{\lambda}{\mu_j}} = \frac{\mu_j}{\lambda + \mu_j}, \quad j \in \mathbb{N}.
\]

This means the RKHS \( (\mathcal{H}, \langle \cdot, \cdot \rangle_{\lambda}) \) consists of

\[
\{ f = \sum_{j=1}^{\infty} f_j \phi_j \in L^2([0,1]) : \sum_{j=1}^{\infty} \frac{f_j^2}{\nu_j} < \infty \}
\]

with

\[
\langle f, g \rangle_{\lambda} = \sum_{j=1}^{\infty} \frac{f_j g_j}{\nu_j}.
\]

We will denote the reproducing kernel of this new RKHS by

\[
\tilde{K}(s, t) := \sum_{j=1}^{\infty} \nu_j \phi_j(s)\phi_j(t),
\]

(4.5)

which we will call the equivalent kernel of \( k \) defined in (1.1). Let us denote the representer of the evaluation map by \( \tilde{K}_s(\cdot) = \tilde{K}(s, \cdot) \) which implies \( g(s) = \langle g, \tilde{K}_s \rangle_{\lambda} \) for any \( g \in \mathcal{H} \).
The eigenvalues of the Matérn kernel used in this paper are polynomially decaying, of the form
\[ \mu_j \approx j^{-2\alpha} \tag{4.6} \]
for \( \alpha \) the smoothness of the kernel [33], and the eigenfunctions are of the form
\[ \phi_{2j-1}(x) = \sin(\pi jx), \quad \phi_{2j} = \cos(\pi jx). \tag{4.7} \]
One can observe there exist global constants \( C_{\phi}, L_{\phi} > 0 \) such that the eigenfunctions \( \{\phi_j\}_{j \in \mathbb{N}} \) of \( K \) satisfy \( |\phi_j(t)| \leq C_{\phi} \) for all \( j \geq 1, t \in [0, 1] \), and \( |\phi_j(t) - \phi_j(s)| \leq L_{\phi}|t - s| \) for all \( t, s \in [0, 1] \) and \( j \geq 1 \) [33].

We will frequently make use of the linear operator \( F_\lambda : \mathcal{H} \to \mathcal{H} \) defined as
\[ F_\lambda g(t) = \int g(s)\tilde{K}(s,t)ds, \]
which forms the convolution of \( g(s) \) with the equivalent kernel \( \tilde{K} \). In the current setting where \( X \sim U(0,1) \) this can also be seen as \( F_\lambda g(t) = \mathbb{E}_X[g(t)|\tilde{K}_X] \). For \( g = \sum_{j=1}^{\infty} g_j \phi_j \), in view of (4.1)
\[ F_\lambda g(t) = \int \sum_{j=1}^{\infty} g_j \phi_j(t)\tilde{K}(s,t)ds = \sum_{j=1}^{\infty} \nu_j g_j \phi_j(t). \tag{4.8} \]
Next to this, we define another linear operator \( P_\lambda : \mathcal{H} \to \mathcal{H} \) given by
\[ P_\lambda f = f - F_\lambda f, \tag{4.9} \]
where for \( g = \sum_{j=1}^{\infty} g_j \phi_j \) we obtain
\[ P_\lambda g(t) = \sum_{j=1}^{\infty} (1 - \nu_j) g_j \phi_j(t). \tag{4.10} \]

We will use both \( F_\lambda^N \) and \( P_\lambda^N \) for the linear operators in the setting of distributed method \( N \), which differ from the nondistributed operations only if the eigenvalues of the equivalent kernel are changed. For example, using method \( II \) only alters the likelihood, leaving the eigenvalues intact, but method \( III \) considers rescaling the prior, which changes the covariance kernel, and hence the eigenvalues of the equivalent covariance kernel. For the distributed methods, we will define the KRR estimator using distributed method \( N \) as
\[ \hat{f}_{n,m,\lambda}^N = A_m^N \left( \{\hat{f}_{n/m,\lambda}^{(j)}\}_{j=1}^{m} \right) \tag{4.11} \]
where \( A_m^N \) is the posterior aggregation operator of \( m \) random variables as used in distributed method \( N \) and \( \{\hat{f}_{n/m,\lambda}^{(j)}\}_{j=1}^{m} \) is the set of local KRR estimators for the \( m \) separate machines. Recall that for an RKHS \( \mathcal{H} \) corresponding to the prior covariance kernel \( K \),
it is known that the KRR estimator coincides with the corresponding posterior mean. We will denote the sample score function

$$S_{n,\lambda}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i)) \tilde{K}_{x_i} - P_{\lambda}f,$$

which follows from (14) of [33], and the population score function

$$S_{\lambda}(f) = F_{\lambda}f_0 - f. \quad (4.13)$$

Note that we have

$$S_{n,\lambda}(\hat{f}_{n,\lambda}) = 0,$$

as \( \hat{f}_{n,\lambda} \) is the solution to the KRR problem [33]. Furthermore, we obtain

$$S_{\lambda}(F_{\lambda}f_0) = F_{\lambda}f_0 - F_{\lambda}f_0 = 0$$

as well. Note that by definition

$$S_{\lambda}(\hat{f}_{n,\lambda}) = F_{\lambda}f_0 - \hat{f}_{n,\lambda}, \quad (4.14)$$

an identity which will be used frequently later on. For distributed methods, we will again slightly alter the notation and define

$$S_{n,m,\lambda}^N f = A_m^N(\{S_{n/m,\lambda}^{(j)}\}_{j=1}^{m}), \quad (4.15)$$

in the setting of distributed method \( N \), where \( \{S_{n/m,\lambda}^{(j)}\}_{j=1}^{m} \) is the set of local score functions for the separate \( m \) machines. On these score functions, we state a couple of identities. If we define \( \Delta f_n := \hat{f}_{n,\lambda} - F_{\lambda}f_0 \), we obtain

$$\Delta f_n = \hat{f}_{n,\lambda} - F_{\lambda}f_0 = -(F_{\lambda}f_0 - \hat{f}_{n,\lambda}) = -S_{\lambda}(\hat{f}_{n,\lambda}).$$

First note that because \( P_\lambda \) is a linear operator and \( S_{n,\lambda}(\hat{f}_{n,\lambda}) = 0 \), we obtain

$$-S_{n,\lambda}(F_{\lambda}f_0) = S_{n,\lambda}(\hat{f}_{n,\lambda}) - S_{n,\lambda}(F_{\lambda}f_0)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_{n,\lambda}(x_i)) K_{x_i} - P_{\lambda}\hat{f}_{n,\lambda} - \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - F_{\lambda}f_0(x_i)) K_{x_i} - P_{\lambda}(F_{\lambda}f_0) \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \Delta f_n(x_i) \tilde{K}_{x_i} - P_{\lambda}(\Delta f_n)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \Delta f_n(x_i) \tilde{K}_{x_i} - \Delta f_n + F_{\lambda}\Delta f_n.$$

Note that as we are dealing with \( X \sim U(0, 1) \), we have

$$F_{\lambda}f(\cdot) = \int f(\cdot) \tilde{K}(s, \cdot) ds = \mathbb{E}[f \tilde{K}_X]. \quad (4.16)$$
To summarise, we end up with the following identities
\[
-S_{\lambda}(\hat{f}_n,\lambda) - S_{n,\lambda}(F_\lambda f_0) = \Delta f_n - S_{n,\lambda}(F_\lambda f_0),
\]
\[
[S_{n,\lambda}(\hat{f}_n,\lambda) - S_{\lambda}(\hat{f}_n,\lambda)] - [S_{n,\lambda}(F_\lambda f_0) - S_{\lambda}(F_\lambda f_0)] = \frac{1}{n} \sum_{i=1}^{n} \Delta f_n(x_i) \tilde{K}_{x_i} - E[\Delta f_n(X) \tilde{K}_X],
\]
(4.17)
where the last equation holds as we have \( S_{n,\lambda}(\hat{f}_n,\lambda) = 0 = S_{\lambda}(F_\lambda f_0) \).

4.3. Lemmas used in the proofs

In the proofs of this section, we will encounter similar problems, hence it is useful to state some technical results in forms of lemmas.

The first lemma gives a general sup-norm bound for the linear operator \( P_\lambda \) for both the Sobolev and Hölder classes. Furthermore, we show that in both sets there exist functions for which we have matching upper and lower bounds, up to some constants.

**Lemma 2** (Sup-norm bounds for \( P_\lambda \)). Let \( P_\lambda \) be as defined in (4.9) for the Matérn covariance kernel. Then, for \( \lambda = h^{2\alpha} \) the following upper bound holds
\[
\| P_\lambda f \|_\infty \lesssim h^\alpha,
\]
for \( f \in \Theta^3 \frac{1}{2} (B) \) or \( f \in \Theta^0_\lambda (B) \). Moreover, there exist \( f \) in both \( \Theta^3 \frac{1}{2} (B) \) and \( \Theta^0_\lambda (B) \) for which we also have that
\[
\| P_\lambda f \|_\infty \gtrsim h^\alpha / \log(h^{-1}), \quad \| P_\lambda f \|_\infty \gtrsim h^\alpha / \log^2(h^{-1}),
\]
respectively.

**Proof.** Let us first assume that \( f \in \Theta^3 \frac{1}{2} (B) \). By (4.10)
\[
\| P_\lambda f \|_\infty \leq \| P_\lambda f \|_\chi \sup_{x \in [0,1]} \| \tilde{K}_x \|_\chi,
\]
(4.19)
where \( \| \cdot \|_\chi \) denotes the RKHS norm defined in Section 4.2. By the fact that the eigenfunctions are uniformly bounded, we see \( \| \tilde{K}_x \|_\chi^2 = | \sum_{j=1}^{\infty} \nu_j^2 \phi_j^2(x) \nu_j^{-1} | \lesssim \sum_{i=1}^{\infty} \nu_j \).

We have \( \mu_j \approx j^{-2\alpha} \) by (4.6) which implies
\[
\sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \times \sum_{j=1}^{\infty} \frac{1}{1 + \lambda j^{2\alpha}} = \lambda^{-\frac{1}{2\alpha}} \sum_{j=1}^{\infty} \frac{\lambda^\frac{1}{2\alpha}}{1 + \left( \lambda^\frac{1}{2\alpha} j \right)^{2\alpha}}.
\]
For the \( h \) equal to (2.3), \( \lim_{n \to \infty} h \to 0 \), which implies \( \lambda \to 0 \) as well. Notice that this infinite sum can be seen as a Riemann sum with step size \( \lambda^{1/(2\alpha)} \), meaning that we asymptotically have
\[
\sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \approx \lambda^{-\frac{1}{2\alpha}} \int_0^\infty \frac{1}{1 + x^{2\alpha}} dx \approx \lambda^{-\frac{1}{2\alpha}},
\]
(4.20)
as the integral converges for $\alpha > \frac{1}{2}$ which is satisfied by assumption. Next,

$$
\|P_\lambda f\|_\lambda^2 = \sum_{j=1}^{\infty} \frac{(1 - \nu_j)^2 f_j^2}{\nu_j}
= \sum_{j=1}^{\infty} \left( \frac{\lambda}{\lambda + \mu_j} \right)^2 \frac{\lambda + \mu_j}{\mu_j} f_j^2
= \lambda \sum_{j=1}^{\infty} \frac{\lambda}{\lambda + \mu_j} \frac{f_j^2}{\mu_j}
\lesssim \lambda \sum_{j=1}^{\infty} j^{2\alpha} f_j^2
\leq \lambda B^2,
$$

by definition of $\Theta_S^\alpha(B)$. This implies $\|P_\lambda f\|_\lambda \lesssim \lambda^{\frac{1}{2}}$. Combining this with (4.20), we obtain $\|P_\lambda f\|_\infty \lesssim h^{\alpha^{-\frac{1}{2}}}$. Let us consider the function

$$
f_0(x) = a \sum_{j=1}^{\infty} \frac{(2j)^{-\alpha - \frac{1}{2}}}{\log(2j)} \phi_{2j}(x),
$$

for some constant $a$, which is in $\Theta_S^\alpha(B)$ as

$$
a^2 \sum_{j=1}^{\infty} (2j)^{2\alpha} \left( \frac{(2j)^{-\alpha - \frac{1}{2}}}{\log(2j)} \right)^2 = a^2 \sum_{j=1}^{\infty} \frac{1}{2j \log^2(2j)} \leq B^2,
$$

for $a$ appropriately chosen, as the infinite sum is convergent. Let us define $J := \lceil \lambda^{-\frac{1}{2\alpha}} \rceil$,
such that for \( j < J \) we obtain \( \mu_{2j} \lesssim \lambda \). Then we obtain

\[
\|P_\lambda f_0\|_\infty = \left\| \sum_{j=1}^{\infty} \frac{\lambda}{\mu_{2j} + \lambda}(f_0)_{2j} \phi_{2j} \right\|_\infty
\]

\[
= \sum_{j=1}^{\infty} \frac{\lambda}{\mu_{2j} + \lambda}(f_0)_{2j}
\]

\[
= \sum_{j=1}^{\infty} \frac{\lambda}{1 + \lambda \mu_{2j}}(f_0)_{2j}
\]

\[
\lesssim \sum_{j=1}^{\infty} \frac{\lambda(2j)^{2\alpha}}{1 + \lambda(2j)^{2\alpha}}(f_0)_{2j}
\]

\[
= a \sum_{j=1}^{\infty} \frac{\lambda(2j)^{\alpha - \frac{1}{2}}}{\log(2j)(1 + \lambda(2j)^{2\alpha})}
\]

\[
\geq a \sum_{j=1}^{J} \frac{\lambda(2j)^{\alpha - \frac{1}{2}}}{\log(2j)(1 + \lambda(2j)^{2\alpha})}
\]

\[
\geq \frac{1}{\log(2J)} \sum_{j=1}^{J} \frac{\lambda(2j)^{\alpha - \frac{1}{2}}}{1 + \lambda(2j)^{2\alpha}}
\]

\[
\asymp \frac{\lambda^{\frac{1}{2} - \frac{1}{4\alpha}}}{\log(\lambda^{\frac{1}{2\alpha}})} \sum_{j=1}^{J} \frac{\lambda^{\frac{1}{2} + \frac{1}{4\alpha}}(2j)^{\alpha - \frac{1}{2}}}{1 + \lambda(2j)^{2\alpha}},
\]

where (*) holds because we have \( \phi_{2j}(x) = \cos(\pi j x) \) and the supremum of the function is taken at \( x = 0 \).

By repeating the Riemann sum argument we obtain

\[
\sum_{j=1}^{J} \frac{\lambda^{\frac{1}{2} + \frac{1}{4\alpha}}(2j)^{\alpha - \frac{1}{2}}}{1 + \lambda(2j)^{2\alpha}} \asymp \int_{0}^{1} x^{\alpha - \frac{1}{2}} \frac{1}{1 + x^{2\alpha}} = C,
\]

as we have \( \lambda(2j)^{2\alpha} = 1 \), and as \( \lambda \to 0, J \to \infty \). We conclude that for \( \alpha > \frac{1}{2} \)

\[
\|P_\lambda f_0\|_\infty \gtrsim \frac{\lambda^{\frac{1}{2} - \frac{1}{4\alpha}}}{\log(\lambda^{\frac{1}{2\alpha}})} = h^{\alpha - \frac{1}{2}} / \log(h^{-1}).
\]

By looking at \( \alpha' = \alpha + 1/2 \), the result follows by noting \( \Theta_{S}^{a'} = \Theta_{S}^{a + \frac{1}{2}} \), for which \( \|P_\lambda f_0\|_\infty \gtrsim h^{\alpha'} = h^{\alpha + \frac{1}{2}} \).

Let us now assume that \( f \in \Theta_{S}^{a'}(B) \). From Corollary 2.1 in [33] follows \( \|P_\lambda f\|_\infty \lesssim h^{\alpha} \) by showing

\[
\|P_\lambda f\|_\infty \lesssim \sum_{j=1}^{\infty} \frac{\lambda}{\mu_j + \lambda} f_{0,j} = \sqrt{\lambda} \sum_{j=1}^{\infty} \frac{\sqrt{\mu_j}}{\mu_j + \lambda} f_j \lesssim \sqrt{\lambda} \sum_{j=1}^{\infty} j^{\alpha} f_{0,j} \lesssim \sqrt{\lambda} = h^{\alpha},
\]

(4.22)
where (*) follows because \( \mu_j \approx j^{-2\alpha} \) and \( \sqrt{\lambda \mu_j} \leq (\mu_j + \lambda)/2 \) by the arithmetic-geometric mean inequality and the convergence of the infinite sum follows from the definition of \( \Theta_H^{\alpha}(B) \). Now, let us consider the function

\[
f_0(x) = \sum_{j=1}^{\infty} a(2j)^{-\alpha-1} \frac{\phi_{2j}(x)}{(\log(2j))^2},
\]

for a constant \( a \) such that

\[
\sum_{j=1}^{\infty} \frac{a}{(2j)(\log(2j))^2} \leq B,
\]

which confirms \( f_0 \in \Theta_H^{\alpha}(B) \) as the infinite sum is convergent. By similar reasoning as before, we obtain

\[
\|P_\lambda f_0\|_\infty \geq \sum_{j=1}^{\infty} \frac{\lambda(2j)^{\alpha-1}}{\log^2(2j)(1 + \lambda(2j)^{2\alpha})} \gtrsim \sum_{j=1}^{\infty} \frac{\lambda^2(2j)^{\alpha-1}}{\log^2(\lambda^{-\frac{1}{2\alpha}})} \sum_{j=1}^{\infty} \frac{1}{1 + \lambda(2j)^{2\alpha}} \gtrsim \frac{\lambda^2}{\log^2(\lambda^{-\frac{1}{2\alpha}})} \sum_{j=1}^{\infty} \frac{1}{1 + \lambda(2j)^{2\alpha}},
\]

where the final statement lets us once more use Lemma A.1, now for \( s = 2\alpha, \ t = 1, \ r = 1 - \alpha \) and \( \nu = \lambda^{-1/(2\alpha)} \). Analogously to the earlier case, Lemma A.1(i) now implies that for \( \alpha > \frac{1}{2} \)

\[
\|P_\lambda f_0\|_\infty \geq \lambda^2 \frac{\lambda^2}{\log^2(\lambda^{-\frac{1}{2\alpha}})} \sum_{j=1}^{\infty} \frac{1}{1 + \lambda(2j)^{2\alpha}} \gtrsim \frac{\lambda^2}{\log^2(\lambda^{-\frac{1}{2\alpha}})} = \frac{h^\alpha}{\log^2(h^{-1})}.
\]
We conclude that there exists a function $f_0 \in \Theta_H^0(B)$ such that $\|P_\lambda f_0\|_\infty \gtrsim h^\alpha / \log^2(h^{-1})$.

In this chapter, we will often make use of Lemma 5.2 from [33], stated as follows.

**Lemma 3.** Let

$$G_n = \{ g \in \mathcal{H} : \|g\|_\mathcal{H} < A_n, \|g\|_\infty < B_n \}$$

for $A_n, B_n > 0$. For any $\alpha > 1/2$, it holds for some constant $C$ independent of $(n, h, A_n)$ that for $x_1, x_2, \ldots, x_n$ i.i.d.

$$\mathbb{P} \left[ \sup_{t \in T, g \in G_n} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) \tilde{K}(x_i, t) - \mathbb{E}[g(x_1) \tilde{K}(x_1, t)] \right| \right]
\geq (nh)^{-\frac{1}{2}} \left( \sqrt{\log n} + \sqrt{x} + \sqrt{\log \frac{B_n}{\|g\|_\infty} + A_n^{-\frac{1}{n}}} \right) \|g\|_\infty 
+ (nh)^{-1} \left( \log n + x + \log \frac{B_n}{\|g\|_\infty} + A_n^{-\frac{1}{n}} \max(1, (n/h)^{\frac{1}{2n} - \frac{1}{2}}) \right) \|g\|_\infty \right] \leq e^{-x}, \quad x > 0.$$

Now let us define

$$\tilde{A}_n := 2\lambda^{-\frac{1}{2}} \left( \|P_\lambda f_0\|_\lambda + D \sigma(nh)^{-\frac{1}{2}} \log n \right), \quad (4.23)$$

for $D$ some constant larger than 0. As $\| \cdot \|_\mathcal{H} \leq \lambda^{-\frac{1}{2}} \| \cdot \|_\lambda$ by (4.4), $\|\Delta f_n\|_\mathcal{H} \leq \tilde{A}_n$ holds for $D$ and $n$ sufficiently large by (47) from [33] which states

$$\|\Delta f_n\|_\lambda \leq \|\Delta f_n - S_n,\lambda(F_\lambda f_0)\|_\lambda + ||S_n,\lambda(F_\lambda f_0)\|_\lambda \lesssim n^{-\frac{\beta}{2}} \|P_\lambda f_0\|_\lambda + \sigma(nh)^{-\frac{1}{2}} \log n, \quad (4.24)$$

for $\beta = (2\alpha - 1)^2/(4\alpha(2\alpha + 1))$ which is strictly positive. Note that

$$\|f\|_\infty^2 = \sup_{x \in [0,1]} \sum_{j=1}^\infty |f_j(x)|^2 \lesssim \sup_{x \in [0,1]} \left( \sum_{j=1}^\infty \nu_j \phi_j^2(x) \right) \sum_{j=1}^\infty \frac{f_j^2}{\nu_j} \lesssim h^{-1} \|f\|_\lambda^2,$$

where (*) holds by Cauchy-Schwarz and (**) holds by the fact that the eigenfunctions are uniformly bounded. By once more using (4.24), we obtain $||\Delta f_n||_\infty \lesssim h^{-\frac{\beta}{2}} ||\Delta f_n||_\lambda \lesssim n$.

Note that for $A_n = \tilde{A}_n$, $B_n = n$ and $x = c \log n$ for some positive constant $c$, we have

$$(nh)^{-\frac{1}{2}} \left( \sqrt{\log n} + \sqrt{x} + \sqrt{\log \frac{B_n}{\|g\|_\infty} + A_n^{-\frac{1}{n}}} \right) \|g\|_\infty 
\leq (nh)^{-\frac{1}{2}} \left( 2 + c \sqrt{\log \frac{n}{\|g\|_\infty} + A_n^{-\frac{1}{n}}} \right) \|g\|_\infty.$$

48
and
\[
(nh)^{-1} \left( \log n + x + \log \frac{B_n}{\|g\|_{\infty}} + \tilde{A}_n^{\frac{1}{2}} \max(1, (n/h)^{\frac{1}{2\alpha} - \frac{1}{2}}) \right) \|g\|_{\infty}
\leq (nh)^{-1} \left( (2 + c) \log \frac{n}{\min(1, \|g\|_{\infty})} + \tilde{A}_n^{\frac{1}{2}} \max(1, (n/h)^{\frac{1}{2\alpha} - \frac{1}{2}}) \right) \|g\|_{\infty}.
\]

Recall that \(\|P_{\lambda} f_0\|_\lambda \lesssim \sqrt{\lambda}\) by (4.21). For \(h\) equal to (1.9), in view of \(\lambda = h^{2\alpha}\)
\[
h^{-\alpha} (nh)^{-\frac{1}{2}} \log n = h^{-\alpha - \frac{1}{2}} n^{-\frac{1}{2}} \log n \lesssim (n/(\log n))^{\frac{\alpha + \frac{1}{2}}{2\alpha + 1}} n^{-\frac{1}{2}} \log n = (\log n)^{\frac{1}{2}},
\]
which implies \(\tilde{A}_n \lesssim \sqrt{\log n}\), from which follows \(\tilde{A}_n^{\frac{1}{2}} \lesssim \sqrt{\log n}\), as \(\alpha > 1/2\). This implies
\[
(nh)^{-\frac{1}{2}} \left( (2 + c) \sqrt{\log \frac{n}{\min(1, \|g\|_{\infty})}} + \tilde{A}_n^{\frac{1}{2}} \right) \|g\|_{\infty}
\lesssim (nh)^{-\frac{1}{2}} \left( \sqrt{\log \frac{n}{\min(1, \|g\|_{\infty})}} \right) \|g\|_{\infty},
\]
and
\[
(nh)^{-1} \left( (2 + c) \log \frac{n}{\min(1, \|g\|_{\infty})} + \tilde{A}_n^{\frac{1}{2}} \max(1, (n/h)^{\frac{1}{2\alpha} - \frac{1}{2}}) \right) \|g\|_{\infty}
\lesssim (nh)^{-1} \left( \max(1, (n/h)^{\frac{1}{2\alpha} - \frac{1}{2}}) \left( \log \frac{n}{\min(1, \|g\|_{\infty})} \right) \right) \|g\|_{\infty}.
\]

By taking the sum of (4.25) and (4.26) we obtain that for \(g\) satisfying \(\|g\|_{\infty} \leq n\) and \(\|g\|_h \leq \tilde{A}_n\) and \(c, D\) and \(n\) sufficiently large, we obtain with probability at least \(1 - n^{-10}\)
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} g(x_i) \tilde{K}_{x_i} - \mathbb{E}[g(X) \tilde{K}_X] \right\|_{\infty}
\lesssim \left( \sqrt{\log \frac{n}{\min(1, \|g\|_{\infty})}} + (nh)^{-\frac{1}{2}} \max(1, (n/h)^{\frac{1}{2\alpha} - \frac{1}{2}}) \left( \log \frac{n}{\min(1, \|g\|_{\infty})} \right) \right) \|g\|_{\infty} \sqrt{\frac{n}{nh}}.
\]

By (4.27) we now obtain
\[
\left\| \frac{1}{n/m} \sum_{i=1}^{n/m} g(x_i) \tilde{K}_{x_i} - \mathbb{E}[g(X) \tilde{K}_X] \right\|_{\infty}
\lesssim \max(1, (n/m)^{-1 + \frac{1}{2\alpha} h^{-\frac{1}{2\alpha}}} \sqrt{\log \frac{n/m}{\min(1, \|g\|_{\infty})}}) \sqrt{\log \frac{n/m}{\min(1, \|g\|_{\infty})}} \|g\|_{\infty} \sqrt{1/mh}.
\]
for any $\alpha > 1/2$, $g$ satisfying $||g||_{L^1} \lesssim \tilde{A}_{n/m}$ and $||g||_{L^\infty} \leq n/m$, and $\lambda = h^{2\alpha}$, with probability at least $1 - (n/m)^{-10}$ for $c$ sufficiently large. Finally, note that (4.24) also implies $||S_{n/m,\lambda}(F_{\lambda} f_0)||_{L^\infty} \lesssim n/m$ and $||S_{n/m,\lambda}(F_{\lambda} f_0)||_{L^1} \leq \tilde{A}_{n/m}$, which will be used later on.

The following lemma follows from the sup-norm bounds for $\hat{f}_{n,\lambda} - f_0$ as proved in [33]. During the procedure of computing bounds for the $L^\infty$ distance between the estimator and the true function, the triangle inequality is frequently used, where we encounter a term $||\hat{f}^N_{n/m,\lambda} - F_{\lambda}^N f_0||_{L^\infty}$. The following lemma gives us an upper bound for this term.

**Lemma 4.** Let $f_0 \in \Theta^0_H(B)$ or $f_0 \in \Theta^0_S(B)$ and $h$ equal to (2.3). We define $\Delta f^N_{n,m} = \hat{f}^N_{n,m,\lambda} - F_{\lambda}^N f_0$ and $\Delta f^{(j)}_{n/m} = \hat{f}^{(j)}_{n/m,\lambda} - F_{\lambda} f_0$. Let us define

$$\gamma_{n/m} := \max\left(1, (n/m)^{-1 + \frac{1}{2n}} h^{-\frac{1}{2n}} \sqrt{\log(n/m)}\right) \sqrt{\log(n/m)} \frac{1}{\sqrt{nm^{-1}h}}. \quad (4.29)$$

Now if the aggregation function $A^N_{n,m}$ takes the average over $m$ values, we obtain

$$||\Delta f^N_{n,m}||_{L^\infty} \lesssim \max\left((n/m)^{-10} h^{-1}, \gamma_{n/m} \right) ||S^{(1)}_{n/m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty} + ||S^N_{n/m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty},$$

with probability at least $1 - m(n/m)^{-10}$.

**Proof.** The proof is analogous to equation (49) of [33]. Let us begin by using the triangle inequality to obtain

$$||\hat{f}^N_{n,m,\lambda} - F_{\lambda}^N f_0||_{L^\infty} \leq ||\hat{f}^N_{n,m,\lambda} - F_{\lambda}^N f_0 - S^N_{n,m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty} + ||S^N_{n,m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty}.$$

By the identities (4.17) and (4.18) we obtain

$$||\hat{f}^N_{n,m,\lambda} - F_{\lambda}^N f_0 - S^N_{n,m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty} = \frac{1}{m} \sum_{j=1}^{m} (\hat{f}^{(j)}_{n/m,\lambda} - F_{\lambda}^{(j)} f_0 - S^{(j)}_{n/m,\lambda}(F_{\lambda}^N f_0)) \leq \frac{1}{m} \sum_{j=1}^{m} ||\Delta f^{(j)}_{n/m} - S^{(j)}_{n/m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty},$$

which follows by repeatedly using the triangle inequality. By the identities (4.17) and (4.18) combined with (4.28), we obtain for every $j$

$$||\Delta f^{(j)}_{n/m} - S^{(j)}_{n/m,\lambda}(F_{\lambda}^N f_0)||_{L^\infty} = \frac{1}{n/m} \sum_{i=1}^{n/m} \Delta f^{(j)}_{n/m}(x_i^{(j)}) \tilde{K}_{x_i^{(j)}} - E(\Delta f^{(j)}_{n/m}(X) \tilde{K}_{X}) \lesssim \max\left(1, (n/m)^{-1 + \frac{1}{2n}} h^{-\frac{1}{2n}} \sqrt{\log(1/||\Delta f^{(j)}_{n/m}||_{L^\infty})} \right) \frac{\log(n/m)}{\min(1, ||\Delta f^{(j)}_{n/m}||_{L^\infty})} \frac{\log(n/m)}{\min(1, ||\Delta f^{(j)}_{n/m}||_{L^\infty})} \frac{1}{\sqrt{nm^{-1}h}}. \quad (4.30)$$
with probability at least $1 - (n/m)^{-10}$ on an event we will denote by $A_j$. Taking the intersection $\bigcap_{j=1}^m A_j$ tells us that with probability $1 - m(1/m)^{-10}$ all events hold simultaneously. Let us now distinguish between two cases. First assume $\|\Delta f_{n/m}^{(j)}\|_\infty \lesssim (n/m)^{-a}$, for $a$ large enough. We will continue to use $a = 10$ throughout. Then

$$\left\| \frac{1}{n/m} \sum_{i=1}^{n/m} \Delta f_{n/m}^{(j)}(x_i^{(j)} \phi_j(x_i)) - \mathbb{E}(\Delta f_{n/m}^{(j)}(X) \phi_j(X)) \right\|_\infty \lesssim (n/m)^{-10} \left\| \frac{1}{n/m} \sum_{i=1}^{n/m} \phi_j(x_i) - \mathbb{E} \phi_j(x) \right\|_\infty \lesssim (n/m)^{-10} h^{-1},$$

as $\sup_{x,x'} |\phi_j(x) - \phi_j(x')| = \sup_i |\sum_j \nu_j \phi_j(x_i)| \lesssim h^{-1}$ by (4.20). For $h$ equal to (2.3), this is of order at most $(n/m)^{-(9/2)} \log(n/m)^{-\frac{1}{2n+1}}$. We conclude

$$\| f_{n,m,\lambda}^N - F_{\lambda}^N f_0 \|_\infty \lesssim (n/m)^{-(9/2)} \log(n/m)^{-\frac{1}{2n+1}} + \| S_{n,m,\lambda}^N(F_{\lambda}^N f_0) \|_\infty.$$

Now, if $\|\Delta f_{n/m}^{(j)}\|_\infty \geq (n/m)^{-10}$, we obtain

$$\sqrt{\log \frac{(n/m)}{\min(1, \|\Delta f_{n/m}^{(j)}\|_\infty)}} \lesssim \sqrt{\log((n/m)^{11})} = \sqrt{11 \log(n/m)} \lesssim \sqrt{\log(n/m)}.$$

This implies that by (4.30), we obtain for every $j$ that

$$\| \Delta f_{n/m}^{(j)} - S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0) \|_\infty \lesssim \gamma_{n/m} \| \Delta f_{n/m}^{(j)} \|_\infty,$$

with probability at least $1 - (n/m)^{-10}$. This implies that

$$\| \Delta f_{n,m}^N - S_{n,m,\lambda}^N(F_{\lambda}^N f_0) \|_\infty \lesssim \frac{1}{m} \sum_{j=1}^m \| \Delta f_{n/m}^{(j)} - S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0) \|_\infty \lesssim \gamma_{n/m} \max_j \| \Delta f_{n/m}^{(j)} \|_\infty,$$

with probability at least $1 - m(n/m)^{-10}$ on the intersection of the events on which $\|\Delta f_{n/m}^{(j)} - S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0)\|_\infty \lesssim \gamma_{n/m} \|\Delta f_{n/m}^{(j)}\|_\infty$ for all $j$. By using the triangle inequality and the last inequality, we obtain

$$\| \Delta f_{n/m}^{(j)} \|_\infty \leq \| \Delta f_{n/m}^{(j)} - S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0) \|_\infty + \| S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0) \|_\infty$$

$$\Rightarrow \| \Delta f_{n/m}^{(j)} \|_\infty \leq \gamma_{n/m} \| \Delta f_{n/m}^{(j)} \|_\infty + \| S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0) \|_\infty$$

$$\Rightarrow \| \Delta f_{n/m}^{(j)} \|_\infty \lesssim (1 - \gamma_{n/m})^{-1} \| S_{n,m,\lambda}^{(j)}(F_{\lambda}^N f_0) \|_\infty.$$

Let us turn to the term $\gamma_{n/m}$. If $\max(1,(n/m)^{-1} h^{-\frac{1}{2n}} \sqrt{\log(n/m)}) = 1$, we obtain $\gamma_{n/m} \lesssim \sigma^{-1/2(2n+1)}(\log(n/m)/(n/m))^{a/(2n+1)}$ which goes to 0 for $n/m \to \infty$. If

51
\[
\max(1, (n/m)^{-1 + \frac{1}{2\alpha} h^{-\frac{1}{2\alpha}} \sqrt{\log(n/m)}}) = (n/m)^{-1 + \frac{1}{2\alpha} h^{-\frac{1}{2\alpha}} \sqrt{\log(n/m)}}
\]
we obtain for \(h\) equal to (2.3) that
\[
(n/m)^{-1 + \frac{1}{2\alpha} h^{-\frac{1}{2\alpha}} \log(n/m)} = C \sigma^{-\frac{1}{2\alpha}} (n/m)^{-1 + \frac{1}{2\alpha} (n/m)^{\frac{1}{4\alpha^2 + 2\alpha}} \log(n/m)}^{\frac{1}{4\alpha^2 + 2\alpha}}.
\]
Moreover, for \(h\) equal to (2.3), we obtain
\[
\sqrt{\log(n/m)/(nm^{-1}h)} = C \sigma^{-\frac{1}{2\alpha + 1}} (\log(n/m)/(nm))^{\frac{1}{2\alpha + 1}},
\]
which has a positive power of \(n\) for \(\alpha > 1/2\), meaning \(nm^{-1}h \to \infty\) as \(n \to \infty\). This implies that at most
\[
\gamma_{n/m} = O \left( \sigma^{-\frac{2\alpha + 1}{2\alpha^2 + \alpha}} (n/m)^{1 - \frac{2\alpha}{\alpha} (\log(n/m))^{\frac{6\alpha^2 + 2\alpha - 1}{4\alpha^2 + 2\alpha}}} \right),
\]
as
\[
-1 + \frac{1}{2\alpha} + \frac{1}{4\alpha^2 + 2\alpha} - \frac{2\alpha}{2\alpha + 1} = 1 - \frac{2\alpha}{\alpha}.
\]
Combining these two parts, we can conclude that \(\gamma_{n/m} \to 0\) as \(n/m \to \infty\) as \(n/m\) dominates \(\log(n/m)\). Combining this with (4.31) conclude that
\[
\left\| f_{n,m,\lambda}^N - F_\lambda^N f_0 \right\|_\infty \leq \left\| f_{n,m,\lambda}^N - F_\lambda^N f_0 - S_{n,m,\lambda}^N (F_\lambda^N f_0) \right\|_\infty + \left\| S_{n,m,\lambda}^N (F_\lambda^N f_0) \right\|_\infty
\]
\[
\lesssim \gamma_{n/m} \max_j \left\| \Delta f_{n/m}^{(j)} \right\|_\infty + \left\| S_{n,m,\lambda}^N (F_\lambda^N f_0) \right\|_\infty
\]
\[
\lesssim \gamma_{n/m} (1 - \gamma_{n/m})^{-1} \max_j \left\| S_{n,m,\lambda}^{(j)} (F_\lambda^N f_0) \right\|_\infty + \left\| S_{n,m,\lambda}^N (F_\lambda^N f_0) \right\|_\infty
\]
\[
\lesssim \gamma_{n/m} \left\| S_{n,m,\lambda}^{(1)} (F_\lambda^N f_0) \right\|_\infty + \left\| S_{n,m,\lambda}^N (F_\lambda^N f_0) \right\|_\infty,
\]
with probability at least \(1 - m(n/m)^{-10}\). The statement now follows by taking the maximum over the two cases.

Following this lemma, we have all the tools to prove the sup-norm bounds for the posterior mean using the naive averaging method, shown in the following theorem.

### 4.4. Proof of Lemma 1

**Proof.** Note that if we have a \((n + m)\)-dimensional multivariate normal distribution
\[
\begin{bmatrix} x \\ y \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),
\]

52
then we have \( x \mid Y_n \sim \mathcal{N}(\mu', \Sigma') \) where
\[
\mu' = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_n - \mu_2), \tag{4.32}
\]
\[
\Sigma' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{4.33}
\]

In this case, we are looking for \( y^* := \hat{f}(x^*) \mid X_n, Y_n, k \) where \( x^* \) is the test input and \( X_n, Y_n \) the training data. First note that if we draw two functions \( f \) and \( g \) from a Gaussian process with mean function 0 and covariance kernel \( k \) defined in (1.1) marginal to input data sets \( X_n \) and \( X'_n \), we have
\[
\begin{bmatrix} f(X) \\ g(X') \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X_n, X_n) & K(X_n, X'_n) \\ K(X'_n, X_n) & K(X'_n, X'_n) \end{bmatrix} \right),
\]
which in our setting, if we consider a single one-dimensional test input \( x^* \), means:
\[
\begin{bmatrix} y^* \\ Y_n \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2(n\lambda)^{-1}k(x^*, x^*) & \sigma^2(n\lambda)^{-1}k(x^*, X_n) \\ \sigma^2(n\lambda)^{-1}k(X_n, x^*) & \sigma^2(n\lambda)^{-1}k(X_n, X_n) + \sigma^2I \end{bmatrix} \right).
\]
Combining this with (4.32) and (4.33), we obtain
\[
y^* \mid X_n, Y_n, k \sim \mathcal{N}(k^T_f(K + n\lambda I)^{-1}Y_n, \sigma^2(n\lambda)^{-1}(k_{**} - k^T_f(K + n\lambda I)^{-1}k_*)).
\]

\[\square\]

4.5. Proof of Theorem 1

We start by giving a sketch of the proof. To prove the lower bound for the \( L_\infty \) distance between the KRR estimator and the truth, we will use the triangle inequality to split the distance in the difference of two terms, \( ||P_{\lambda}f_0||_\infty \) and \( ||f_{n,m,\lambda}^I - F_{\lambda}^I f_0||_\infty \). As the eigenvalues of the equivalent kernel are equal to the nondistributed setting, it will follow that there exists a function for which \( ||P_{\lambda}f_0||_\infty = ||P_{\lambda}f_0||_\infty \geq h^a / \log^2(h^{-1}) \) by Lemma 2. The upper bound for \( ||f_{n,m,\lambda}^I - F_{\lambda}^I f_0||_\infty \) is different from the nondistributed setting. Analogous to [33], by the triangle inequality
\[
||f_{n,m,\lambda}^I - F_{\lambda}^I f_0||_\infty \leq ||f_{n,m,\lambda}^I - F_{\lambda}^I f_0 - S_{n,m,\lambda}^I(f_{\lambda} f_0)||_\infty + ||S_{n,m,\lambda}^I(f_{\lambda} f_0)||_\infty.
\]
Computing the bounds, we will see that with high probability
\[
||f_{n,m,\lambda}^I - F_{\lambda}^I f_0||_\infty \leq \max \left( (n/m)^{-0.5} \log(n/m)^{-\frac{1}{\alpha+1}}, \gamma \right),
\]
by using Lemma 4. The term \( ||S_{n,m,\lambda}^I(f_{\lambda} f_0)||_\infty \) will be bounded from above by again using the triangle inequality. It follows that \( ||P_{\lambda}f_0||_\infty \) is of higher order than \( ||f_{n,m,\lambda}^I - F_{\lambda}^I f_0||_\infty \). The dominating term is larger than the dominating term in the minimax optimal rate, letting us conclude that there exists a function for which the method is suboptimal. We will now begin the proof of Theorem 1.
Proof. Let us assume that $f_0 \in \Theta_{S}^{\alpha+\frac{1}{2}}(B)$ such that $\|P_{f_0}^{I}\| = \|P_{\lambda f_0}\| \gtrsim h^{\alpha}/\log(h^{-1})$.
We begin by noting

$$\left\| f_{n,m,\lambda}^{I} - f_{0} \right\|_{\infty} = \left\| f_{n,m,\lambda}^{I} - f_{0} + f_{n,m,\lambda}^{I} - f_{n,m,\lambda}^{I} \right\|_{\infty} \geq \left\| P_{f_0}^{I} - f_{n,m,\lambda}^{I} \right\|_{\infty}.$$ 

$$\geq \left\| P_{f_0}^{I} \right\|_{\infty} - \left\| f_{n,m,\lambda}^{I} \right\|_{\infty} \geq \left(\left\| f_{n,m,\lambda}^{I} - f_{n,m,\lambda}^{I} \right\|_{\infty} + \left\| S_{\lambda}(f_{n,m,\lambda}^{I} f_0) \right\|_{\infty}\right).$$

We derived in Lemma 2 bounds for $\|P_{f_0}^{I}\|_{\infty}$, the nondistributed setting. Note that in this distributed setting the equivalent kernel $K$ retains eigenvalues of equal form, which implies both $P_{\lambda} = P_{\lambda}^{I}$ and $F_{\lambda} = F_{\lambda}^{I}$. We continue to denote $I$ for consistency. Thus let us turn to the second term. By Lemma 4 we obtain that there exists an event $A_{n,m}$ with $\mathbb{P}(A_{n,m}) \geq 1 - m(n/m)^{-10}$ on which

$$\left\| \Delta f_{n,m}^{I} \right\|_{\infty} \lesssim \max\left((n/m)^{-(\alpha+\frac{1}{2})} \frac{1}{\sqrt{\log(n/m)}} \right), \gamma_{n/m} \left\| S_{n,m,\lambda}^{(1)}(F_{\lambda}^{I} f_0) \right\|_{\infty} + \left\| S_{n,m,\lambda}(f_{n,m,\lambda}^{I} f_0) \right\|_{\infty}.$$ 

Recall by (1.11) that for a sample size of $n/m$ and $h$ equal to (2.3), we obtain

$$\left\| S_{n,m,\lambda}(F_{\lambda}^{I} f_0) \right\|_{\infty} \lesssim \sigma^{2} \left(\log(n/m) / (n/m)\right)^{\alpha+\frac{1}{2}},$$

with probability at least $1 - (n/m)^{-10}$, on an event we will denote by $B_{n,m}$. Combining this with (4.31), we observe

$$\gamma_{n/m} \left\| S_{n,m,\lambda}^{(1)}(F_{\lambda}^{I} f_0) \right\|_{\infty} \lesssim \sigma^{2} \left(\log(n/m) / (n/m)\right)^{\alpha+\frac{1}{2}}.$$ 

Note that

$$(n/m)^{-(\alpha+\frac{1}{2})} \left(\log(n/m)\right)^{\frac{1}{2}} \lesssim (n/m)^{\frac{1}{2}} \left(\log(n/m)\right)^{\frac{2\alpha^2+2\alpha-1}{4\alpha^2+2\alpha}},$$

from which follows that on $A_{n,m} \cap B_{n,m}$ we obtain

$$\left\| \Delta f_{n,m}^{I} \right\|_{\infty} \lesssim (n/m)^{\frac{1-2\alpha}{\alpha^2+1}} \left(\log(n/m)\right)^{\frac{2\alpha^2+2\alpha-1}{4\alpha^2+2\alpha}} + \left\| S_{n,m,\lambda}(F_{\lambda}^{I} f_0) \right\|_{\infty},$$

with $\mathbb{P}(A_{n,m} \cap B_{n,m}) \geq 1 - (m+1)(n/m)^{-10}$ as in the current setting $\sigma$ does not depend on $n$ or $m$.

Now let us turn to $\left\| S_{n,m,\lambda}(F_{\lambda}^{I} f_0) \right\|_{\infty}$. As we have $A_{n,m}^{I}(\sum_{j=1}^{m} f_{n,m,\lambda}^{(j)}) = m^{-1} \sum_{j=1}^{m} f_{j},$
we find
\[
\left\| S_{n,m,\lambda}^I(F_{\lambda}^I f_0) \right\|_\infty = \left\| \frac{1}{m} \sum_{j=1}^{m} \left( S_{n/m,\lambda}^{(j)}(F_{\lambda}^I f_0) \right) \right\|_\infty \\
= \left\| \frac{1}{m} \sum_{j=1}^{m} \left( \frac{1}{n/m} \sum_{i=1}^{n/m} (y_i^{(j)} - (F_{\lambda}^I f_0)(x_i^{(j)})) \tilde{K}_{x_i^{(j)}} - P_{\lambda}^I(F_{\lambda}^I f_0) \right) \right\|_\infty \\
= \left\| \frac{1}{m} \sum_{j=1}^{m} \left( \frac{1}{n/m} \sum_{i=1}^{n/m} (y_i^{(j)} - f_0(x_i^{(j)})) + f_0(x_i^{(j)}) - (F_{\lambda}^I f_0)(x_i^{(j)}) \right) \right\|_\infty \\
\leq \left\| \frac{1}{m} \sum_{j=1}^{m} \left( \frac{1}{n/m} \sum_{i=1}^{n/m} P_{\lambda}^I f_0(x_i^{(j)}) \tilde{K}_{x_i^{(j)}} - \mathbb{E}(P_{\lambda}^I f_0(X) \tilde{K}_X) \right) \right\|_\infty + \\
\left\| \frac{1}{m} \sum_{j=1}^{m} \left( \frac{1}{n/m} \sum_{i=1}^{n/m} (y_i^{(j)} - f_0(x_i^{(j)})) \tilde{K}_{x_i^{(j)}} \right) \right\|_\infty \\
\leq \frac{1}{m} \sum_{j=1}^{m} \left\| \frac{1}{n/m} \sum_{i=1}^{n/m} P_{\lambda}^I f_0(x_i^{(j)}) \tilde{K}_{x_i^{(j)}} - \mathbb{E}(P_{\lambda}^I f_0(X) \tilde{K}_X) \right\|_\infty + \\
\left\| \frac{1}{m} \sum_{j=1}^{m} \left( \frac{1}{n/m} \sum_{i=1}^{n/m} w_i^{(j)} \tilde{K}_{x_i^{(j)}} \right) \right\|_\infty \\
:= \frac{1}{m} \sum_{j=1}^{m} T_1^j + T_2,
\]
where \( w_i^{(j)} := y_i^{(j)} - f_0(x_i^{(j)}) \) is the \( N(0, \sigma^2) \) distributed error term for the local data \( X_{n/m}^{(j)}, Y_{n/m}^{(j)} \), which form a partition of \( X_n, Y_n \). Inequality (*) holds in view of the triangle
inequality combined with

\[ P_\lambda^l f_0 = P_\lambda^l \left( \sum_{k=1}^{\infty} \nu_k f_{0,k} \phi_k \right) \]

\[ = \sum_{k=1}^{\infty} (1 - \nu_k) f_{0,k} \phi_k \]

\[ = F_\lambda \left( \sum_{k=1}^{\infty} (1 - \nu_k) f_{0,k} \phi_k \right) \]

\[ = F_\lambda^l P_\lambda^l f_0 \]

\[ = \mathbb{E}[P_\lambda^l f_0(X) \hat{K}_X], \]

where last line follows by (4.16). For \( \hat{A}_n \) defined as in (4.23) \( \|P_\lambda^l f_0\|_\lambda = \|P_\lambda f_0\|_\lambda \lesssim \sqrt{\lambda} \lesssim \hat{A}_{n/m} \) by (4.21). Furthermore \( \|P_\lambda^l f_0\|_\infty \lesssim h^n \lesssim n \) for the chosen value of \( h \). This tells us that \( P_\lambda^l f_0 \) satisfies all assumptions for Lemma 3 for \( A_n = \hat{A}_n \) and \( B_n = n \). By inequality (4.28) we know that there exists an event \( \mathcal{C}_{n,m} \) with \( \mathbb{P}(\mathcal{C}_{n,m}) \geq 1 - m(n/m)^{-10} \) on which we have

\[
T_1^{(j)} \lesssim \max \left( 1, \frac{(n/m)^{10+\frac{1}{\alpha}} h^{-\frac{1}{\alpha}} \sqrt{\log \left( \frac{(n/m)}{\min(1, \|P_\lambda^l f_0\|_\infty)} \right)} \sqrt{\log \left( \frac{(n/m)}{\min(1, \|P_\lambda^l f_0\|_\infty)} \right)} }{\|P_\lambda^l f_0\|_\infty} \right)
\]

for all \( j \) simultaneously. Let us now distinguish between two cases. First, if \( \|P_\lambda^l f_0\|_\infty \leq (n/m)^{-10} \), then

\[
\left\| \frac{1}{n/m} \sum_{i=1}^{n/m} P_\lambda^l f_0(x_i^{(j)}) \hat{K}_{x_i^{(j)}} - \mathbb{E}(P_\lambda^l f_0(X) \hat{K}_X) \right\|_\infty \leq \left( \frac{(n/m)^{-10}}{n/m} \sum_{i=1}^{n/m} \hat{K}_{x_i^{(j)}} - \mathbb{E}(\hat{K}_X) \right) \lesssim (n/m)^{-10} h^{-1},
\]

as \( \sup_{x,x'} |\hat{K}(x,x')|_\infty \lesssim h^{-1} \). For \( h \) equal to (2.3), this is of order at most \( (n/m)^{-(9+\frac{1}{\alpha})} \log(n/m)^{-\frac{1}{2\alpha+1}} \). Now, if \( \|P_\lambda^l f_0\|_\infty \geq (n/m)^{-10} \), we obtain

\[
\sqrt{\log \left( \frac{(n/m)}{\min(1, \|P_\lambda^l f_0\|_\infty)} \right)} \lesssim \sqrt{\log(n/m)^{10} = 11 \log(n/m)} \lesssim \sqrt{\log(n/m)}.
\]

This implies \( T_1^{(j)} \lesssim \gamma_{n/m} \|P_\lambda^l f_0\|_\infty \). Now, using (4.31) we obtain

\[
T_1^{(j)} \lesssim \sigma^{-\frac{1}{\alpha}} \frac{(n/m)^{1-2\alpha}}{\alpha (\log(n/m))^{\frac{6n^2+2\alpha-1}{4\alpha+2\alpha}}} \|P_\lambda^l f_0\|_\infty,
\]

which is of larger order than \( (n/m)^{-(9+\frac{1}{\alpha})} \log(n/m)^{-\frac{1}{2\alpha+1}} \). All the pieces combined now give us that on the event \( \mathcal{C}_{n,m} \) with \( \mathbb{P}(\mathcal{C}_{n,m}) \geq 1 - m(n/m)^{-10} \), we have

\[
T_1 := \frac{1}{m} \sum_{j=1}^{m} T_1^{(j)} \lesssim \sigma^{-\frac{\alpha+1}{2\alpha+\alpha}} \frac{(n/m)^{1-2\alpha}}{\alpha (\log(n/m))^{\frac{6n^2+2\alpha-1}{4\alpha+2\alpha}}} \|P_\lambda^l f_0\|_\infty,
\]

(4.36)
which is of smaller order than \( \|P_{x}^{f}f_0\|_{\infty} \) as the power of \((n/m)\) is less than 0 for \( \alpha > \frac{1}{2} \). In this setting the \( \sigma \) term is irrelevant, but it will be needed in later proofs.

To bound \( T_2 \), note that we can write

\[
\left\| \frac{1}{m} \sum_{j=1}^{m} w^{(j)} K_{x_i} \right\|_{\infty} = \left\| \frac{1}{m} \sum_{j=1}^{m} w^{(j)} \left( \sum_{k=1}^{\infty} \phi_{k}(x^{(j)}_{i}) \phi_{k}(\cdot) \nu_{k} \right) \right\|_{\infty}
\]

\[
= \left\| \sum_{k=1}^{\infty} (w^{i,k}_{\cdot} \phi_{k}(\cdot) \nu_{k}) \right\|_{\infty}
\]

where \( w^{i,k}_{\cdot} := m^{-1} \sum_{j=1}^{m} w^{(j)} \phi_{k}(x^{(j)}_{i}) \), which is distributed \( N(0,(\sigma^2 \sum_{j=1}^{m} (\phi_{k}(x^{(j)}_{i}))^2)/m^2) \). Furthermore, denote \( w^{i,k} = w_{i} \phi_{k}(x_{i}) \sim N(0,\sigma^2(\phi_{k}(x_{i}))^2) \) for the nondistributed setting.

Let us state an adaptation of Lemma 5.3 from [33].

**Lemma 5.** There exists some constant \( C \) such that for any \( x > 0 \),

\[
P \left( \left\| \frac{1}{n/m} \sum_{i=1}^{n/m} \sum_{k=1}^{\infty} w^{i,k}_{\cdot} \phi_{k}(\cdot) \nu_{k} \right\|_{\infty} \leq C \sigma \sqrt{\frac{\log(n/m)}{nm}} \right) \geq 1 - e^{-x}.
\]  

(4.37)

The proof to this Lemma can be found in Appendix A.2. For \( x = c \log n \) we obtain

\[
T_2 = \left\| \frac{1}{n/m} \sum_{i=1}^{n/m} \left( \sum_{k=1}^{\infty} w^{i,k}_{\cdot} \phi_{k}(\cdot) \nu_{k} \right) \right\|_{\infty} \leq C\sigma \sqrt{\frac{\log(n/m)}{nh}}.
\]

(4.38)

for some \( C > 0 \), with probability at least \( 1 - (n/m)^{-10} \) for sufficiently large \( c \) on an event which we will denote by event \( D_{n,m} \). Combining this with (4.36) we obtain that on the event \( A_{n,m} \cap B_{n,m} \cap C_{n,m} \cap D_{n,m} \)

\[
\left\| \hat{f}_{n,\lambda} - P_{\lambda}^{f}f_0 \right\|_{\infty} \lesssim (n/m)^{1-2\alpha \over \alpha} \left( \log(n/m) \right)^{8\alpha^2+2\alpha-1 \over 4\alpha^2+2\alpha} 
\]

\[
+ (n/m)^{1-2\alpha \over \alpha} \left( \log(n/m) \right)^{6\alpha^2+2\alpha-1 \over 4\alpha^2+2\alpha} \left\| P_{\lambda}^{f}f_0 \right\|_{\infty} + \sigma \sqrt{\frac{\log(n/m)}{nh}}
\]

\[
\lesssim (n/m)^{1-2\alpha \over \alpha} \left( \log(n/m) \right)^{8\alpha^2+2\alpha-1 \over 4\alpha^2+2\alpha} + \sigma \sqrt{\frac{\log(n/m)}{nh}}.
\]

(4.39)

For \( h \) equal to (2.3), we obtain

\[
\sigma \sqrt{\frac{\log(n/m)}{nh}} \lesssim \sqrt{\frac{\log(n/m)}{n}} \left( \frac{n/m}{\log(n/m)} \right)^{1-\alpha \over 2 \alpha + \alpha} = m^{-1 \over 2} \left( \frac{\log(n/m)}{n/m} \right)^{{\alpha \over 2 \alpha + \alpha}}.
\]

(4.40)

Note that

\[
(n/m)^{1-2\alpha \over \alpha} \left( \log(n/m) \right)^{8\alpha^2+2\alpha-1 \over 4\alpha^2+2\alpha} 
\]

\[
= (n/m)^{1-2\alpha \over \alpha} \left( \log(n/m) \right)^{6\alpha^2+2\alpha-1 \over 4\alpha^2+2\alpha} \left( \log(n/m) \right)^{\alpha \over 2 \alpha + \alpha}
\]

57
which implies that for $\alpha > \frac{1}{2}$ and $(n/m)^{1-2\alpha} m^{\frac{1}{2}} \to 0$

\[
m^{-\frac{1}{2}} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2\alpha+1}} \gtrsim (n/m)^{1-2\alpha} \frac{\alpha}{2\alpha+1} \left( \log(n/m) \right)^{\frac{8\alpha^2+2\alpha-1}{8\alpha^2+2\alpha}}. \tag{4.41}
\]

From Lemma 2 we know that there exists a function $f_0 \in \Theta_2^\alpha(B)$ such that $\|P^I_\lambda f_0\|_\infty \gtrsim h^\alpha / \log^2(h^{-1})$. This lets us conclude that for $f_0$ on $A_{n,m} \cap B_{n,m} \cap C_{n,m} \cap D_{n,m}$ we have

\[
\| \hat{f}_{n,m} - f_0 \|_\infty \geq \| f_0 - F^I_\lambda f_0 \|_\infty - \| f_{n,\lambda} - F^I_\lambda f_0 \|_\infty \\
\geq \| P^I_\lambda f_0 \|_\infty - m^{-\frac{1}{2}} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2\alpha+1}} \\
\geq \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2\alpha+1}} / \log^2 \left( \left( \frac{n/m}{\log(n/m)} \right)^{\frac{1}{2\alpha+1}} \right) - m^{-\frac{1}{2}} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2\alpha+1}},
\]

with $\mathbb{P}(A_{n,m} \cap B_{n,m} \cap C_{n,m} \cap D_{n,m}) \geq 1 - (2m+2)(n/m)^{-10}$. Here the negative terms are of smaller order than the positive term, as we have $n/m \to \infty$ and $m \to \infty$. This tells us the dominating term in the lower bound is $(n/m)^{1-\alpha}$. We have seen that the minimax sup-norm upper bound in the nondistributed case has a dominating term $n^{-\alpha} / \log^{\alpha}(n/m)$, which is smaller than the distributed case. This lets us conclude

\[
\| \hat{f}_{n,m} - f_0 \|_\infty \gtrsim \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2\alpha+1}} / \log^2 \left( \left( \frac{n/m}{\log(n/m)} \right)^{\frac{1}{2\alpha+1}} \right),
\]

with probability at least $1 - (2m+2)(n/m)^{-10}$.

This result now trivially extends to the function $f_0 \in \Theta_2^{\alpha+\frac{1}{2}}(B)$ for which $\|P^I_\lambda f_0\|_\infty \gtrsim h^\alpha / \log(h^{-1})$.

4.6. Proof of Theorem 2

We start by giving a sketch of the proof. We will prove that there exists a function $f_{bad} \in \Theta_2^{\alpha+\frac{1}{2}}(B)$ for which we have asymptotic frequentist coverage of the pointwise credible sets of probability 0. This is done by showing the squared bias dominates the variance.

We will show the possibility of approximating the covariance kernel by some fixed function, independent of the data, in the nondistributed setting. This will show to be useful in proving the pointwise credible set coverage. This theory focusses on the nondistributed setting, which relates to the distributed setting by applying the aggregation operator on the local covariance kernels. By the definition of the covariance kernel (1.6), we can write

\[
\sigma^{-2} n \lambda \hat{C}_n(x,x') = k_x(x') - \hat{K}_x(x'),
\]

58
where $k_x(\cdot) = k(x, \cdot)$ the covariance kernel as defined before, and

$$
\tilde{K}_x(x^*) = k(x^*, X_n)^T (K(X_n, X_n) + n\lambda I)^{-1} k(X_n, x).
$$

If we compare this to (4.3), we obtain that $\tilde{K}_x$ can also be written as the solution of the following KRR problem with noiseless observations of $K_x$ at the random design points $X_n$,

$$
\hat{K}_x = \operatorname{argmin}_{g \in \mathcal{H}} \left[ \frac{1}{n} \sum_{i=1}^{n} (Z_x^i - g(x_i))^2 + \lambda \| g \|_H^2 \right],
$$

(4.42)

where $Z_x^i = k(x, x_i)$, which is noiseless (see Section 2.2 of [33]). Corollary 2.2 from [33] states that for a noiseless version, where the measured random variables do not contain a noise term of a certain variance, of the KRR problem, we obtain both

$$
\| \hat{f}_{n, \lambda} - f_0 \|_{\infty} \leq 2 \| P_\lambda f_0 \|_{\infty},
$$

and

$$
\| f_0 - \hat{f}_{n, \lambda} - P_\lambda f_0 \|_{\infty} \leq C\gamma_n \| P_\lambda f_0 \|_{\infty},
$$

(4.43)

for some constant $C > 0$, with probability at least $1 - n^{-10}$. Notice that for the covariance kernel, the first statement implies

$$
\| \sigma^{-2} n\lambda \hat{C}_n(x, \cdot) \|_{\infty} = \| k_x - \tilde{K}_x \|_{\infty} \leq 2 \| P_\lambda k_x \|_{\infty}.
$$

Note that by (4.2) and the definition of $P_\lambda$, we observe that

$$
P_\lambda k_x = \sum_{j=1}^{\infty} (1 - \nu_j) \mu_j \phi_j(x) \phi_j(\cdot) = \sum_{j=1}^{\infty} \frac{\mu_j \lambda}{\mu_j + \lambda} \phi_j(x) \phi_j(\cdot) = \lambda \tilde{K}_x,
$$

for $\tilde{K}_x(\cdot) = \sum_{j=1}^{\infty} \nu_j \phi_j(x) \phi_j(\cdot)$. By the fact that $\{\phi_j\}_{j \in \mathbb{N}}$ is uniformly bounded, we obtain

$$
\left| \sum_{j=1}^{\infty} \nu_j \phi_j(x) \phi_j(x') \right| \lesssim h^{-1},
$$

from the Riemann sum (4.20). Combining this with the higher order expansion (4.43)

$$
\sigma^{-2} n\lambda \hat{C}_n(x, \cdot) - P_\lambda k_x \|_{\infty} \leq C\gamma_n \| P_\lambda k_x \|_{\infty} \quad \text{implies} \quad \sigma^{-2} n\lambda \hat{C}_n(x, \cdot) - \tilde{K}_x \|_{\infty} \leq C\gamma_n \| \tilde{K}_x \|_{\infty}.
$$

(4.44)

As we have $\gamma_n \to 0$ for $n \to \infty$, this opens an opportunity to numerically approximate $\sigma^{-2} n\lambda \hat{C}_n$ in a way such that there is no need to compute the eigendecomposition and which does not involve the random design points.
Let us now define \( \hat{C}_n(x, x') := \sigma^2 h \tilde{K}(x, x') \). Using (4.44), we now obtain
\[
\sup_{x, x'} |nh \hat{C}_n(x, x') - \hat{C}_n(x, x')| \leq C \sigma^2 h \gamma_n \sup_{x, x'} |\tilde{K}(x, x')| \lesssim \gamma_n,
\]
(4.45)
since \( \sup_{x, x'} |\tilde{K}(x, x')| \lesssim h^{-1} \) by (4.20). As we have seen earlier, for \( h \) equal to (1.9) this goes to 0 for \( n \to \infty \), giving us a covariance kernel to approximate our original covariance kernel \( \hat{C}_n \) with, which does not depend on the random design points but is a fixed function. This will prove to be useful in the computation of pointwise credible set coverage.

With this approximation explained, let us turn to the proof of Theorem 2.

**Proof.** Consider a function \( f_{bad} \in \Theta^{\alpha + \frac{1}{2}}_\alpha(B) \) such that
\[
\|P_\lambda f_{bad}\|_\infty \gtrsim h^{\alpha / \log 2(h^{-1})},
\]
which we know exists by Lemma 2. By the definition of \( \hat{C}_n \) we observe that
\[
\sup_{x, x' \in [0, 1]} |\hat{C}_n(x, x')| = \sup_{x, x' \in [0, 1]} |\sum_{k=1}^\infty \phi_k(x) \phi_k(x') \nu_k| \lesssim hh^{-1} = 1
\]
thus \( \hat{C}_n \) is of constant order. Combining this with (4.45), we conclude that \( \hat{C}_n \) is of order \( (nh)^{-1} \). Recall that
\[
CI_n(x; \beta) = \left[ \hat{f}_n(x) - z_{(1+\beta)/2} \sqrt{\hat{C}_n(x, x)}, \hat{f}_n(x) + z_{(1+\beta)/2} \sqrt{\hat{C}_n(x, x)} \right],
\]
we see that the size of credible set is determined by \( \sqrt{\hat{C}_n(x, x)} \). Now let us turn to the distributed setting, by (2.4) we have
\[
\sqrt{\hat{C}_{n,m}(x, x)} = \sqrt{m^{-2} \sum_{j=1}^m (\hat{C}_{n/m}(x, x))^{(j)}},
\]
which implies that the size of
\[
CI_{n,m}^I(x; \beta) := \left[ \hat{f}_{n,m}(x) - z_{(1+\beta)/2} \sqrt{\hat{C}_{n,m}(x, x)}, \hat{f}_{n,m}(x) + z_{(1+\beta)/2} \sqrt{\hat{C}_{n,m}(x, x)} \right]
\]
is of order \( m^{-\frac{1}{2}}(nm^{-1}h)^{-\frac{1}{2}} = (nh)^{-\frac{1}{2}} \) again for the distributed setting. Let us define
\[ \gamma = (1 + \beta)/2. \] From this follows, analogously to (33) from [33] that:

\[ \mathbb{P}(f_{bad}(x) \in CI_{n,m}(x; \beta)] \]

\[ = \mathbb{P}\left( \sqrt{n}h(\tilde{f}_{n,m}^I(x) - f_{bad}(x)) \right) \leq z_{\gamma} \sqrt{nh\tilde{C}_{n,m}^I(x,x)} \]

\[ \approx \mathbb{P}\left( \sqrt{n}h(\tilde{f}_{n,m}^I(x) - f_{bad}(x)) \right) \leq z_{\gamma} \sqrt{nhm^{-2} \sum_{m=1}^m \tilde{C}_{n,m}^{(j)}(x,x)} \]

\[ \approx \mathbb{P}\left( \sqrt{n}h(\tilde{f}_{n,m}^I(x) - f_{bad}(x)) \right) \leq z_{\gamma} \sqrt{m^{-1} \sum_{j=1}^m \tilde{C}_{n,m}^{(j)}(x,x)} \]

\[ = \mathbb{P}\left( \sqrt{n}h(\tilde{f}_{n,m}^I(x) - F_{\lambda}^I f_{bad}(x)) \right) \in \left[ \sqrt{nhP_{\lambda}^I f_{bad}(x)} \pm z_{\gamma} \sqrt{\tilde{C}_{n,m}^{(1)}(x,x)} \right], \]

where (*) holds because we know the variance of distributed method \( I \) is distributed \( m^{-1} \) times the variance of the nondistributed method for a sample size of \( n/m \), and the approximation follows from (4.45). We have \( m^{-1} \sum_{j=1}^m \tilde{C}_{n,m}^{(j)} = \tilde{C}_{n,m}^{(1)} = \sigma^2 h K \) as all \( \tilde{C}_{n,m}^{(j)} \) are equal. This is of constant order as \( \|K\|_\infty \lesssim \sum_{k=1}^\infty \nu_k \lesssim h^{-1} \). First note that for \( h \) equal to the optimal value (2.3) by (4.39) we obtain with probability at least \( 1 - (n/m)^{-10} \)

\[ \sqrt{n}h(\tilde{f}_{n,m}^I(x) - F_{\lambda}^I f_{bad}(x)) \lesssim (nh)^{1/2} m^{-1/2} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2n+1}} \]

\[ = (n/m)^{1/2} \frac{h}{\sqrt{m}} (n/m)^{-1/2} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2n+1}} \]

\[ \lesssim \left( \frac{n/m}{\log(n/m)} \right)^{-\frac{1}{2n+2}} \left( \frac{n/m}{\log(n/m)} \right)^{\frac{1}{2n+1}} \]

\[ = \sqrt{\log(n/m)}, \]

which goes to infinity as \( n/m, m \to \infty \) as \( \alpha \) is positive. Furthermore, using Lemma 2 for \( h \) equal to 2.3 we have

\[ \sqrt{n}hP_{\lambda}^I f_{bad} \gtrsim (nh)^{1/2} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2n+1}} / \log \left( \left( \frac{n/m}{\log(n/m)} \right)^{\frac{1}{2n+1}} \right) \]

\[ \gtrsim \left( \sqrt{m \log(n/m)} \right) / \log \left( \left( \frac{n/m}{\log(n/m)} \right)^{\frac{1}{2n+1}} \right), \]

by analogous reasoning. This approaches infinity faster than \( \sqrt{n}h(\hat{f}_n - F_{\lambda} f_{bad}(x)) \). Combined, this implies that

\[ \mathbb{P}[f_{bad} \in CI_{n,m}(x; \beta)] \]

\[ \approx \mathbb{P}\left[ \sqrt{n}h(\hat{f}_n^I(x) - F_{\lambda}^I f_{bad}(x)) \right] \in \left[ \sqrt{nhP_{\lambda}^I f_{bad}(x)} \pm z_{\gamma} \sqrt{\tilde{C}_{n,m}^{(1)}(x,x)} \right] \to 0, \]

61
4.7. Proof of Theorem 3

The proof to the upper bound for the $L_\infty$ distance between the KRR estimator and the true function $f_0$ follows by combining the results on the adjusted likelihood method and Theorem 1.

Proof. Let us assume that $f_0 \in \Theta_H^\alpha(B)$. By the triangle inequality we obtain

$$
\|\hat{f}_{n,m}^{IIa} - f_0\|_{\infty} \leq \|P_{\lambda}^{IIa} f_0\|_{\infty} + \|\hat{f}_{n,m,\lambda}^{IIa} - F_{\lambda}^{IIa} f_0\|_{\infty},
$$

where we will bound these two terms similarly to Section 4.5. From Section 2.3 we know the variance of the noise is $\sigma' = \sigma/\sqrt{m}$ in this setting. By Lemma 2 then follows

$$
\|P_{\lambda}^{IIa} f_0\|_{\infty} \lesssim h^\alpha \left( \frac{B^2 n/m}{(\sigma')^2 \log(n/m)} \right)^{-\frac{\alpha}{2r+1}} = \left( \frac{B^2 n/m}{\sigma^2/m \log(n/m)} \right)^{-\frac{\alpha}{2r+2}} \lesssim \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2r+1}}.
$$

First, note that

$$
\gamma_{n/m} \left\| S_{n/m,\lambda}^{(1)}(F_{\lambda}^{IIa} f_0) \right\|_{\infty} \lesssim m^{\frac{1}{2r} - \frac{\alpha}{2r+1}} (n/m)^{1 - \frac{2\alpha}{\alpha} - \frac{\alpha}{2r+1} (\log(n/m))^{\frac{8r^2 + 2r - 1}{4r^2 + 2r}}}.
$$

thus

$$
\max \left( (n/m)^{-10} h^{-1}, \gamma_{n/m} \left\| S_{n/m,\lambda}^{(1)}(F_{\lambda}^{IIa} f_0) \right\|_{\infty} \right)
\lesssim m^{\frac{1}{2r} - \frac{\alpha}{2r+1}} (n/m)^{1 - \frac{2\alpha}{\alpha} - \frac{\alpha}{2r+1} (\log(n/m))^{\frac{8r^2 + 2r - 1}{4r^2 + 2r}}}.
$$

Following this, we combine (4.35), (4.36) (where in this case the value of $\sigma$ is relevant) and (4.39) to obtain

$$
\left\| \hat{f}_{n,m,\lambda}^{IIa} - F_{\lambda}^{IIa} f_0 \right\|_{\infty}
\lesssim m^{\frac{1}{2r} - \frac{\alpha}{2r+1}} (n/m)^{1 - \frac{2\alpha}{\alpha} - \frac{\alpha}{2r+1} (\log(n/m))^{\frac{8r^2 + 2r - 1}{4r^2 + 2r}}} + m^{\frac{1}{2r}} (n/m)^{\frac{6r^2 + 2r - 1}{4r^2 + 2r}} \left\| P_{\lambda}^{IIa} f_0 \right\|_{\infty} + \sigma' \sqrt{\frac{\log(n/m)}{nh}},
$$

as $n/m, m/\log^2 n \to \infty$. □
with probability at least $1 - (2m + 2)(n/m)^{-10}$. Note that
\[ m^{1/2\alpha} (n/m)^{1-2\alpha} (\log(n/m))^{\frac{6\alpha^2+2\alpha-1}{4\alpha^2+2\alpha}} \| P_\lambda f_0 \|_\infty \]
\[ \leq m^{1/2\alpha} (n/m)^{1-2\alpha} (\log(n/m))^{\frac{8\alpha^2+2\alpha-1}{4\alpha^2+2\alpha}}, \]
where the power of $(n/m)$ is negative for $\alpha > \frac{1}{2}$. Now as $n \frac{1-2\alpha}{\alpha} m \frac{1-2\alpha}{\alpha} \to 0$, we have
\[ m^{1/2-\alpha} (n/m)^{1-2\alpha} (\log(n/m))^{\frac{8\alpha^2+2\alpha-1}{4\alpha^2+2\alpha}} \]

Furthermore, we obtain
\[ \sigma' \sqrt{\log(n/m) \frac{n}{h}}^{1/2} = \sigma \sqrt{\log(n/m) \frac{n}{h}}^{1/2} \leq m^{1/2}(\log(n/m) \frac{n}{h})^{\alpha/2\alpha}. \]
This lets us conclude
\[ \| \hat{f}_{IIa} - f_0 \|_\infty \leq \left( \frac{\log(n/m)}{n} \right)^{\alpha/2\alpha}, \]
with probability at least $1 - (2m + 2)(n/m)^{-10}$. Here the dominant term is $n^{-\alpha/2\alpha}$, which is of same order as the dominant term in the minimax rate for the nondistributed method from (23) from [33]. Adjusting the local likelihood thus gives an optimal upper bound for $f_0 \in \Theta_{H,B}^\alpha$.

This result now trivially extends to the function $f_0 \in \Theta_{S,B}^{\alpha+\frac{1}{2}}$.

\[ \Box \]

### 4.8. Proof of Theorem 4

This proof is done similarly to Section 4.6, except for $\sigma' = \sigma/\sqrt{m}$ instead of $\sigma$.

**Proof.** Let us consider $f_{bad} \in \Theta_{S,B}^{\alpha+\frac{1}{2}}$ such that $\| P_\lambda f_{bad} \|_\infty \geq h^\alpha / \log(h^{-1})$. We have seen that in the general distributed setting using the average of the posteriors that $CI_{n,m}(x; \beta)$ is of order $(nh)^{-\frac{1}{2}}$. As the adjusted likelihood setting deals with $\sigma/\sqrt{m}$ as opposed to $\sigma$, the size of $CI_{n,m}(x; \beta)$ can be computed by the definition of $\tilde{C}_n$

\[ \sqrt{\tilde{C}_{n,m}^I(x, x)} = \sqrt{m^{-1}(\tilde{C}_{n,m}^I)(x, x)} \asymp (nmh)^{-\frac{1}{2}}. \]
Implementing this, we get
\[ P[f_{bad} \in CI_{n,m}^{IIa}(x; \beta)] \]
\[ = P \left[ \sqrt{nmh(f_{n,m}^{IIa}(x) - f_{bad}(x))} \leq z_\gamma \sqrt{nmh \hat{C}_{n,m}^{IIa}(x,x)} \right] \]
\[ \approx P \left[ \sqrt{nmh(f_{n,m}^{IIa}(x) - f_{bad}(x))} \leq z_\gamma \sqrt{nmh \sum_{j=1}^{m} \hat{C}_{n,m}^{(j)}(x,x)} \right] \]
\[ = P \left[ \sqrt{nmh(f_{n,m}^{IIa}(x) - F_{\lambda}^{IIa}f_{bad}(x))} \in \left[ \sqrt{nmhP_{\lambda}^{IIa}f_{bad}(x) + z_\gamma \sqrt{\hat{C}_{n/m}(x,x)}} \right] \right] , \]

where \((*)\) holds as \(\hat{C}_{n,m}^{IIa}(x,x) = \frac{1}{m^2} \sum_{j=1}^{m} \hat{C}_{n/m}^{(j)}(x,x)\) and the approximation follows from (4.45) combined with the fact that \(\hat{C}_{n/m}^{IIa} = m^{-1}\hat{C}_{n/m}\) by the adjusted likelihoods.

Using similar computations as in Theorem 2, we obtain
\[ \sqrt{nmh(f_{n,m}^{IIa}(x) - F_{\lambda}^{IIa}f_{bad}(x))} \lesssim \sqrt{nmh \cdot m^{-\frac{1}{2}} \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{\alpha+1}}} = \sqrt{\log(n/m)}. \]

Furthermore, as we have
\[ P_{\lambda}^{IIa}f_{bad}(x) \gtrsim \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{\alpha+1}} / \log \left( \left( \frac{n}{\log(n/m)} \right)^{\frac{n}{\alpha+1}} \right) , \]
in this setting, we obtain
\[ \sqrt{nmhP_{\lambda}^{IIa}f_{bad}(x)} \gtrsim \left( \sqrt{m \log(n/m)} \right) / \log \left( \left( \frac{n}{\log(n/m)} \right)^{\frac{n}{\alpha+1}} \right) . \]

As we have \(n/m \to \infty\) and \(m/\log^2(n) \to \infty\), this implies that \(\sqrt{nmhP_{\lambda}^{IIa}f_{bad}} \to \infty\) faster than \(\sqrt{nmh(f_{n,m}^{IIa} - F_{\lambda}^{IIa}f_{bad}(x))}\). As \(\hat{C}_{n/m}\) is fixed and of constant order, we conclude
\[ P[f_{bad} \in CI_{n,m}^{IIa}(x; \beta)] \to 0, \]
as \(m, n/m \to \infty\).

4.9. Proof of Theorem 5

As the prior is adjusted, the eigenvalues of the kernel get changed by a factor of \(m\). This changes the eigenvalues of the equivalent kernel and changes the variance of the posterior. In the proof below we obtain an upper bound for \(||F_{\lambda}^{IIa}f_0||_{\infty}\) for the new prior, as well as local variance terms for the new prior. We will combine this analogously to the proof of Theorem 1.
Proof. Let $f_0 \in \Theta_S^{\alpha + \frac{1}{2}}(B)$. In this case, we have $K_{\lambda} := mK'$. As the eigenvalues are defined by $Ax = \mu x$, we obtain for $(cA)x = (c\mu)x$, which tells us in the current setting that the eigenvalues of $K_{\lambda}$, $\mu_j$, satisfy $\mu_j^2 = C\lambda m^{-2\alpha}$ for some constant $C$. From this follows that $\|P_{\lambda}^{\lambda}f_0\|_{\infty}$ is different from $\|P_{\lambda}f_0\|_{\infty}$. Note that the rescaling can be seen as a new parameter $\lambda := \lambda/m$. This implies that we have a different value of $h$ as well $h_{\lambda} := (\lambda/m)^{3/2} = hm^{-1/2\alpha}$. Having defined the new value of $h$ for this setting, we can use the results proved for the nondistributed setting in [33] with $h$ replaced by $h_{\lambda}$.

Once more, we use the triangle inequality to observe that

$$\|\hat{J}_{n,m,\lambda}^{\lambda} - f_0\|_{\infty} \leq \|P_{\lambda}^{\lambda}f\|_{\infty} + \|\hat{J}_{n,m,\lambda}^{\lambda} - P_{\lambda}^{\lambda}f_0\|_{\infty}.$$  

To bound the first term, let us assume $f_0 \in \Theta_S^{\alpha}(B)$ for $\alpha' = \alpha + \frac{1}{2}$. Recall that by (4.19) we have $\|P_{\lambda}^{\lambda}f_0\|_{\infty} \leq \|P_{\lambda}^{\lambda}f_0\|_{\lambda}\sup_{x \in [0,1]}||\tilde{K}_{\lambda}^{\lambda}f||_{\lambda}$. Where $\|P_{\lambda}^{\lambda}f_0\|_{\lambda}$ can be bounded analogously to (4.21)

$$\|P_{\lambda}^{\lambda}f_0\|_{\lambda}^2 = \sum_{k=1}^{\infty} \frac{\lambda/m}{\lambda/m + \mu_j} \frac{f_j^2}{m\mu_j} \leq \frac{\lambda}{m} \sum_{j=1}^{\infty} j^{2\alpha'} f_j^2 \leq \frac{\lambda}{m} B^2,$$

as $(\lambda/m)/(\lambda/m + \mu_j) \leq 1$ and $\mu_j \propto j^{-2\alpha'}$. As $\sup_{x \in [0,1]}||\tilde{K}_{\lambda}^{\lambda}f||_{\lambda} \leq \sum_{j=1}^{\infty} \nu_j \alpha' \leq m^{1/(2\alpha')} h^{-1},$ this implies that $\|P_{\lambda}^{\lambda}f_0\|_{\infty} \leq m^{-\frac{\alpha'}{2} + \frac{1}{4\alpha'} h_{\lambda}^{-1}},$ which for $\alpha' = \alpha + \frac{1}{2}$ implies $\|P_{\lambda}^{\lambda}f_0\|_{\infty} \leq m^{-\frac{\alpha}{2} + \frac{1}{4\alpha'} h_{\lambda}^\alpha}$.

For $h$ equal to (2.3) we obtain

$$\|P_{\lambda}^{\lambda}f_0\|_{\infty} \lesssim m^{-\frac{\alpha}{2} + \frac{1}{4\alpha'} h_{\lambda}^\alpha} \lesssim \left(\frac{(n/m)}{n}\right)^{\frac{\alpha}{2\alpha + 1}},$$

(4.46)

which is of equal order to the minimax rate up to a logarithmic term (1.11). First, note that by (1.7) and taking $\gamma_{n/m}$ with $h_{\lambda}$ we have

$\gamma_{n/m} \|S_{n/m,\lambda}^{(1)}(F_{\lambda}^{\lambda} f_0)\|_{\infty} \lesssim \gamma_{n/m} \sqrt{\frac{\log(n/m)}{nm^{-1} h_{\lambda}}} \lesssim n^{-\frac{1-2\alpha}{2} + \frac{2-3\alpha}{4\alpha^2} + \frac{\alpha+1}{4\alpha^2} (\log(n/m))^{\frac{\alpha}{4\alpha^2 + 2\alpha}} \left(\frac{(n/m)}{n}\right)^{\frac{\alpha}{2\alpha + 1}},$  

where we see

$$m^{\frac{1}{2\alpha}} (n/m)^{-\frac{3}{2\alpha}} \log(n/m)^{-\frac{1}{2\alpha + 1}} \lesssim n^{-\frac{1-2\alpha}{2} + \frac{2-3\alpha}{4\alpha^2} + \frac{\alpha+1}{4\alpha^2} (\log(n/m))^{\frac{\alpha}{4\alpha^2 + 2\alpha}} \left(\frac{(n/m)}{n}\right)^{\frac{\alpha}{2\alpha + 1}}.$$  

Now, by repeating all the steps of the proof of Theorem 1 up to (4.39) for the modified equivalent kernel where we replace $h$ by $h_{\lambda}$, we obtain

$$\|\hat{J}_{n,m,\lambda}^{\lambda} - F_{\lambda}^{\lambda} f_0\|_{\infty} \lesssim \gamma_{n/m} \left(\|S_{n/m,\lambda}^{(1)}(F_{\lambda}^{\lambda} f_0)\|_{\infty} + \|P_{\lambda}^{\lambda} f_0\|_{\infty}\right) + m^{-\frac{1}{4}} \sqrt{\frac{\log(n/m)}{nm^{-1} h_{\lambda}}}.$$
with probability at least $1 - (2m + 2)(n/m)^{-10}$. By (1.7) we have

$$\left\| S_{n/m, \lambda}^{(1)}(F_{\lambda}^{\text{IIIa}} f_0) \right\|_{\infty} \lesssim \sqrt{\frac{\log(n/m)}{nm^{-1} h^{\text{IIIa}}}} \lesssim m^{\frac{1}{2}} \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2\alpha + 1}},$$

with probability at least $1 - (2m + 2)(n/m)^{-10}$. Combining this with (4.46) we obtain

$$\gamma_{n/m} \left( \left\| S_{n/m, \lambda}^{(1)}(F_{\lambda}^{\text{IIIa}} f_0) \right\|_{\infty} + \left\| P_{\lambda}^{\text{IIIa}} f_0 \right\|_{\infty} \right) \lesssim \gamma_{n/m} \left( m^{\frac{1}{2}} \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2\alpha + 1}} \right)$$

$$\lesssim n^{\frac{1 - 2\alpha}{\alpha} - \frac{2 - 3\alpha}{2\alpha + 1} + \frac{\alpha + 1}{4\alpha + 2}} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{6\alpha^2 + 2\alpha - 1}{4\alpha^2 + 2\alpha}},$$

$$\lesssim \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2\alpha + 1}},$$

for $n^{\frac{1 - 2\alpha}{\alpha} - \frac{2 - 3\alpha}{2\alpha + 1} + \frac{\alpha + 1}{4\alpha + 2}} \to 0$. Furthermore we see that

$$m^{-\frac{1}{2}} \sqrt{\frac{\log(n/m)}{nm^{-1} h^{\text{IIIa}}}} \lesssim m^{-\frac{1}{2}} \left( \frac{\log(n/m)}{n/m} \right)^{\frac{\alpha}{2\alpha + 1}} = \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2\alpha + 1}},$$

implying with probability at least $1 - (2m + 2)(n/m)^{-10}$

$$\left\| \hat{f}_{n/m, \lambda}^{\text{IIIa}} - F_{\lambda}^{\text{IIIa}} f_0 \right\|_{\infty} \lesssim \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2\alpha + 1}}. \quad (4.47)$$

Combining this with (4.46), we conclude that

$$\left\| \hat{f}_{n/m, \lambda}^{\text{IIIa}} - f_0 \right\| \lesssim \left( \frac{\log(n/m)}{n} \right)^{\frac{\alpha}{2\alpha + 1}},$$

with probability at least $1 - (2m + 2)(n/m)^{-10}$. The order of this term is equal to the order nondistributed minimax rate up to a logarithmic term.

\[ \square \]

### 4.10. Proof of Theorem 6

**Proof.** To prove the pointwise credible set coverage, note that the process from (4.6) can be repeated, but by replacing $\tilde{K}$ by $\tilde{K}^{\text{IIIa}}$, the equivalent kernel of the rescaled covariance kernel. If we now define

$$\hat{C}_{n}^{\text{IIIa}} := \sigma^2 h^{\text{IIIa}} \tilde{K}^{\text{IIIa}}(x, x'),$$

we can approximate

$$(n/m) h^{\text{IIIa}} \hat{C}_{n/m}^{\text{IIIa}}(x, x) \approx \hat{C}_{n/m}^{\text{IIIa}},$$

66

\[ \square \]
by (4.45).

Let us define the process \( U_n(\cdot) = \sqrt{h^{IIIa}/n} \sum_{i=1}^{n} w_i \tilde{K}_{x_i}^{IIIa}(\cdot) \) where we define the covariance function \( C'_n \) which matches the covariance function of the process \( U_n(\cdot) \) on \([0, 1]\) as follows

\[
\tilde{C}'_n(x, x') := \mathbb{E}[U_n(x)U_n(x')]
= \sigma^2 h^{IIIa} \mathbb{E}[\tilde{K}^{IIIa}(X_n, x)\tilde{K}^{IIIa}(X_n, x')]
= \sigma^2 h^{IIIa} \sum_{j=1}^{\infty} \frac{\phi_j(x)\phi_j(x')}{(1 + \frac{h^{IIIa}}{\mu'_j})^2}.
\]

Analogous to (4.20), we obtain as \( h^{IIIa} \to \infty \):

\[
\sum_{j=1}^{\infty} \frac{1}{(1 + h^{IIIa}/\mu'_j)^2} \asymp \sum_{j=1}^{\infty} \frac{h^{IIIa}}{(1 + (h^{IIIa})^{2\alpha})^2} \asymp (h^{IIIa})^{-1} \int_{0}^{\infty} \frac{1}{(1 + x^{2\alpha})^2} \asymp (h^{IIIa})^{-1}.
\]

Combined with the fact that \( \{\phi_j\}_{j \in \mathbb{N}} \) is uniformly bounded, we obtain that \( \tilde{C}'_n(x, x') \) is of constant order. For this process we obtain

\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P}\left( \tilde{C}'_n(x, x) - \frac{1}{2} \sqrt{h^{IIIa}/n} \sum_{i=1}^{n} w_i \tilde{K}_{x_i}^{IIIa}(x) \leq u \right) - \Phi(u) \right| \lesssim (nh^{IIIa})^{-\frac{1}{2}}, \tag{4.48}
\]

by using the Berry-Esseen theorem as in Section 5.5 from [33]. The second display of (27) in [33] tells us that in the nondistributed setting

\[
\left\| \hat{f}_n - F_\lambda f_0 - \frac{1}{n} \sum_{i=1}^{n} w_i \tilde{K}_{x_i}^{IIIa} \right\|_\infty \lesssim \gamma_n \left( \|P_\lambda f_0\|_\infty + \sigma \sqrt{\frac{\log n}{nh}} \right) =: \delta_n, \tag{4.49}
\]

with probability at least \( 1 - n^{-10} \). For the current distributed setting, we obtain by using triangle inequality that

\[
\left\| \hat{f}_{n,m} - F_\lambda^{IIIa} f_0 - \frac{1}{n} \sum_{i=1}^{n} w_i \tilde{K}_{x_i}^{IIIa} \right\|_\infty \leq \frac{1}{m} \sum_{j=1}^{m} \left( \hat{f}_{n/m}^{(j)} - F_\lambda f_0 - \frac{1}{n/m} \sum_{i=1}^{n/m} w_i^{(j)} \tilde{K}_{x_{i}^{(j)}}^{IIIa} \right) \right\|_\infty
\]

\[
\leq \frac{1}{m} \sum_{j=1}^{m} \left\| \hat{f}_{n/m}^{(j)} - F_\lambda f_0 - \frac{1}{n/m} \sum_{i=1}^{n/m} w_i^{(j)} \tilde{K}_{x_{i}^{(j)}}^{IIIa} \right\|_\infty
\]

\[
\lesssim \delta_{n/m}, \tag{4.50}
\]
with probability at least $1 - m(n/m)^{-10}$ on the intersection of the events on which we have
\[
\left\| \tilde{f}_{n/m}^{(j)} - F_\lambda f_0 - \frac{1}{n/m} \sum_{i=1}^{n/m} w_i(i) \tilde{K}_{x_i}^{(j)} \right\|_\infty \lesssim \delta_{n/m},
\]
for all $j$, where $h = h_{IIIa}$ by applying (4.49) to all separate terms. Let us now denote
\[
a := a_{n,m}(x) = \frac{1}{n} \sum_{i=1}^{n} w_i \tilde{K}_{x_i}^{IIIa}(x), \tag{4.51}
\]
\[
b := b_{n,m}(x) = \tilde{f}_{n,m}^{IIIa}(x) - F_\lambda f_0(\hat{x}), \tag{4.52}
\]
\[
c := c_{n,m}(x) = \sqrt{nh_{IIIa}}(\hat{C}_{n}^{\prime}(x, x))^{-\frac{1}{2}}, \tag{4.53}
\]
where we have by (4.48) $\sup_{u \in \mathbb{R}} |\mathbb{P}(ca \leq u) - \Phi(u)| \lesssim (nh_{IIIa})^{-\frac{1}{2}}$ and by (4.50)
\[
||a - b||_\infty \lesssim \delta_{n/m},
\]
which also implies $||c(a - b)||_\infty \lesssim \sqrt{nh_{IIIa}} \delta_{n/m}$. This now implies that
\[
\sup_{u \in \mathbb{R}} |\mathbb{P}(cb \leq u) - \Phi(u)| = \sup_{u \in \mathbb{R}} |\mathbb{P}(cb \leq u - c(a - b)) - \Phi(u - c(a - b))|
\leq \sup_{u \in \mathbb{R}} |\mathbb{P}(cb \leq u - c(a - b)) - \Phi(u) + Dc(a - b)|
\leq \sup_{u \in \mathbb{R}} |\mathbb{P}(cb \leq u - c(a - b)) - \Phi(u)| + |Dc(a - b)|
\lesssim \sup_{u \in \mathbb{R}} |\mathbb{P}(ca \leq u) - \Phi(u)| + \sqrt{nh_{IIIa}} \delta_{n/m}
\lesssim (nh_{IIIa})^{-\frac{1}{2}} + \sqrt{nh_{IIIa}} \delta_{n/m},
\]
with probability at least $1 - (n/m)^{-10}$, where (*) holds for some constant $D$ because $\Phi$ has a bounded derivative \(^1\). We conclude
\[
\sup_{u \in \mathbb{R}} |\mathbb{P}(\hat{C}_{n}^{\prime}(x, x) - \frac{1}{2} \sqrt{nh_{IIIa}}(\tilde{f}_{n,m}^{IIIa} - F_\lambda f_0(\hat{x})) \leq u) - \Phi(u)| \lesssim (nh_{IIIa})^{-\frac{1}{2}} + \sqrt{nh_{IIIa}} \delta_{n/m}.
\tag{4.54}
\]
Note that
\[
\sqrt{nh_{IIIa}} \delta_{n/m} = g_{n/m} \left(\sqrt{nh_{IIIa}} \|P_{xy}f_0\|_\infty + \sigma \sqrt{m \log(n/m)}\right)
\lesssim g_{n/m} \sqrt{m \log(n/m)}
\lesssim (nm^{-1}h_{IIIa})^{-\frac{1}{2}} \sqrt{m \log(n/m)}
\lesssim m \left(\frac{\log(n/m)}{n}\right)^{\frac{\alpha}{2n+1}},
\]
\(^1\)This appears to be a mistake in the proof of Theorem 3.2 in [33], where $\delta_n$ is stated without a $\sqrt{nh}$ factor.
by (4.47). This goes to 0 for \( n/(m^4) \to \infty \) as the power of \( n \) is at most 1/2. By using the triangle inequality in combination with (4.45), we obtain

\[
\sup_{x,x'} |nh^{IIIa} \hat{C}_{n,m}^{IIIa}(x,x') - \hat{C}_{n,m}^{IIIa}(x,x')| = \sup_{x,x'} |nh^{IIIa} \frac{1}{m^2} \sum_{j=1}^{m} \hat{C}_{n,m}^{j}(x,x') - \hat{C}_{n,m}^{IIIa}(x,x')| \\
\leq \frac{1}{m} \sum_{j=1}^{m} \sup_{x,x'} |nm^{-1}h^{IIIa} \hat{C}_{n,m}^{j}(x,x') - \hat{C}_{n,m}^{IIIa}(x,x')| \\
\lesssim \gamma_{n/m}.
\]

(4.55)

Now denote

\[
a := a_{n,m}(x) = f(x) - \hat{f}_{n,m}^{IIIa}(x)
\]

(4.56)

\[
c_1 := (c_1)_{n,m}(x) = \sqrt{\hat{C}_{n,m}^{IIIa}(x,x)}
\]

(4.57)

\[
c_2 := (c_2)_{n,m}(x) = \sqrt{(nh^{IIIa})^{-1} \hat{C}_{n,m}^{IIIa}(x,x)}
\]

(4.58)

from which follows \( a \mid X_n, Y_n \sim N(0, c_1) \). First note that

\[nh^{IIIa}|(c_1 - c_2)(c_1 + c_2)| = |nh^{IIIa}(c_1^2 - c_2^2)| \lesssim \gamma_{n/m},\]

by (4.55). As \((c_1 + c_2) \lesssim (nh^{IIIa})^{-\frac{1}{2}}\) because as \( \hat{C}_{n,m}^{IIIa} \) is of order 1, both terms are of order \((nh^{IIIa})^{-\frac{1}{2}}\), we obtain

\[|c_1 - c_2| = \left| \sqrt{\hat{C}_{n,m}^{IIIa}(x,x)} - \sqrt{(nh^{IIIa})^{-1} \hat{C}_{n,m}^{IIIa}(x,x)} \right| \lesssim (nh^{IIIa})^{-\frac{1}{2}} \gamma_{n/m},\]

(4.59)

by dividing both sides by \((nh^{IIIa})^{\frac{1}{2}}\). Let us denote the half length of the credible interval using distributed method IIIa as

\[l^{IIIa}(x; \beta) = z_{(1+\beta)/2} \sqrt{\hat{C}_{n,m}^{IIIa}(x,x)}\]
Using this, we obtain

\[
\begin{align*}
|\Phi(\sqrt{nh^{IIIa}}(\hat{C}'_n)^{-\frac{1}{2}}(x, x)l_{n,m}^{IIIa}(x; \beta))| \\
= |\Phi \left( \sqrt{nh^{IIIa}}(\hat{C}'_n)^{-\frac{1}{2}}(x, x) \left( l_{n,m}^{IIIa}(x; \beta) - \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{nh^{IIIa}}} z(1+\beta)/2 + \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{nh^{IIIa}}} z(1+\beta)/2 \right) \right) | \\
= \Phi \left( z(1+\beta)/2 \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{C_n(x, x)}} + \sqrt{nh^{IIIa}}(\hat{C}'_n)^{-\frac{1}{2}}(x, x) \left( l_{n,m}^{IIIa}(x; \beta) - \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{nh^{IIIa}}} z(1+\beta)/2 \right) \right) \\
\leq \Phi \left( z(1+\beta)/2 \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{C_n(x, x)}} + D \sqrt{nh^{IIIa}}(\hat{C}'_n)^{-\frac{1}{2}}(x, x) \left( l_{n,m}^{IIIa}(x; \beta) - \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{nh^{IIIa}}} z(1+\beta)/2 \right) \right) \\
\leq \Phi \left( z(1+\beta)/2 \sqrt{\frac{\hat{C}^{IIIa}(x, x)}{C_n(x, x)}} + \gamma_{n/m}, \right)
\end{align*}
\]

where the first inequality follows by the fact that \( \Phi \) has a bounded derivative and the second inequality follows by (4.59). Furthermore, note that

\[
\begin{align*}
\mathbb{P}[f_0(x) \in C_{n,m}^{IIIa}(x; \beta)] &= \Phi(b) - \Phi(a) \quad \text{for} \quad Z \sim N(0, 1), \quad \text{using (4.54) combined with (4.59)}.
\end{align*}
\]

As we can write \( \mathbb{P}(Z \in [a, b]) = \Phi(b) - \Phi(a) \) for \( Z \sim N(0, 1) \), using (4.54) combined with
(4.60), we obtain
\[
|\mathbb{P}[f_0(x) \in C_{I_{IIIa}}(x; \beta)] - \Phi\left( (\hat{C}'_n(x,x))^{-\frac{1}{2}} (\sqrt{nI_{IIIa}} I_{IIIa}^\beta f_0(x) + I_{IIIa}^\beta n,m) \right) |
\]
\[
+ \Phi\left( (\hat{C}'_n(x,x))^{-\frac{1}{2}} (\sqrt{nI_{IIIa}} I_{IIIa}^\beta f_0(x) - I_{IIIa}^\beta n,m) \right) |
\]
\[
\lesssim |\mathbb{P}[f_0(x) \in C_{I_{IIIa}}(x; \beta)] - \Phi\left( \frac{\hat{C}_{I_{IIIa}}(x,x)}{C'(x,x)} z(1+\beta)/2 + \hat{C}'_n(x,x) (1+\beta/2) \right) |
\]
\[
\Phi\left( -\frac{\hat{C}_{I_{IIIa}}(x,x)}{C'(x,x)} z(1+\beta)/2 + \hat{C}'_n(x,x) (1+\beta/2) \right) | + \gamma_n/m
\]
\[
\lesssim (nI_{IIIa})^{-\frac{1}{2}} + \sqrt{nI_{IIIa}} \delta_{n,m} + \gamma_n/m,
\]
where we will denote
\[
u_{I_{IIIa}}(x; \beta) := \sqrt{\frac{\hat{C}_{I_{IIIa}}(x,x)}{C'(x,x)} z(1+\beta)/2}, \quad b_{I_{IIIa}} = (\hat{C}'_n(x,x))^{-\frac{1}{2}} \sqrt{nI_{IIIa}} I_{IIIa}^\beta f_0(x).
\]

(4.63)

Note that all separate terms in (4.63) go to 0 for \( n/(m^4) \to \infty \).

Having shown this, we can analogously to the smooth-match case of Section 5.6 in [33] show the required result. Let \((f_n)\) be such that \(f_n = \lambda \beta^{-\alpha} \), for \(j = j_n = \lfloor (\lambda/m)^{-1/2\alpha} \rfloor \) and \( (f_n)\) is 0 for \(j \neq j_n\). The constant \(C_\beta\) is a parameter which will be determined later. For \(B \geq C_\beta\) trivially follows that \(f_n \in \Theta_{S}^{\alpha+\frac{1}{2}}(B)\). This means we obtain
\[
\sqrt{nI_{IIIa}} I_{IIIa}^\beta (f_n)(x) = \left( \frac{\lambda}{m} \right)^{\frac{1}{2}} \sum_{j=1}^{\infty} \frac{1}{\lambda + j-2\alpha} (f_n)^j \phi_j(x)
\]
\[
= C_\beta \left( \frac{\lambda}{m} \right)^{\frac{1}{2}} \frac{(\lambda/m)^{j-\lambda}}{\lambda/m + (\lambda/m)^{-\frac{1}{2\alpha}}} \phi_j(x)
\]
\[
= c_{\lambda,m} \phi_{j_n}(x),
\]
where \(c_{\lambda,m} \in [\frac{1}{2}, 1]\) as for \( (\lambda/m)^{-\frac{1}{2\alpha}} = (\lambda/m)^{-\frac{1}{2\alpha}} \) we obtain \(c_{\lambda,m} = 1/2\), and furthermore
\[
\left( \frac{\lambda}{m} \right)^{\frac{1}{2}} \frac{(\lambda/m)^{j-\lambda}}{\lambda/m + (\lambda/m)^{-\frac{1}{2\alpha}}} \leq \left( \frac{\lambda}{m} \right)^{\frac{1}{2}} \frac{(\lambda/m)^{j-\lambda}}{\lambda/m + (\lambda/m)^{-\frac{1}{2\alpha}}} \leq 1.
\]

Now let \(\tilde{b}_n\) be the solution of the equation
\[
\Phi(u_{I_{IIIa}}(x; \beta) + \tilde{b}_n) - \Phi((u_{I_{IIIa}}(x; \beta) - \tilde{b}_n) = \tilde{\beta}.
\]
If $C_\beta$ is chosen such that $\tilde{b}_n = C_{\lambda, m}(x, x)^{-\frac{1}{2}} c_{\lambda, m} \phi_{j_n}(x)$, by (4.63) we obtain

$$P[f_n(x) \in C_{\lambda, m}(x; \beta)] \to \tilde{\beta},$$

for $n/m \to \infty$. □
Conclusion

In this thesis we have studied distributed Bayesian methods in the nonparametric regression model. We have briefly studied the characteristics of the nonparametric regression model, for which we showed the motivation behind applying Bayesian methods using Gaussian process priors. The Bayesian methods have built-in uncertainty quantification as opposed to a frequentist setting. We saw that Gaussian processes offer flexible options as priors through the possibilities of covariance kernels available. We have studied two kernels, the Matérn kernel and the squared exponential kernel, of which we saw the eigendecomposition used in the RKHS setting. From the eigendecomposition, the Matérn kernel showed to provide easier computations. We have seen that the required posterior function should not only satisfy a mean close to the true function, but also quantify the uncertainty well by containing the true function inside a band in which with 95% probability the posterior function based on the data lies.

We started by performing simulation studies for the Matérn kernel in a nondistributed setting. For these simulation results, we have seen the theoretical results presented by [33] which form an optimal estimation to aim for in the distributed setting. This paper presents results which we have briefly studied, regarding both the posterior bias and frequentist coverage of pointwise credible sets. The functions for which we computed theoretical results have been in either the Sobolev ball or the Hölder class, which are sets which contain functions satisfying certain smoothness conditions for a smoothness parameter $\alpha$.

Afterwards, we studied the performance of this model in a distributed setting where we considered the smoothness parameter to be known. In the distributed setting, we divided the data of size $n$ over $m$ machines to separately compute local posteriors to be aggregated into a global posterior by various aggregation methods. The first option we considered was naively taking draws from all $m$ local posteriors and taking the average of these draws to form one global posterior. The simulation study shows that it turns out that this method results in both a suboptimal recovery and uncertainty quantification. Afterwards we considered multiple different distributed methods which mostly followed from [25] which studied distributed Bayesian methods in the signal in white noise model. These methods consist of changing the computation of the local posteriors, different aggregation methods or combinations of both. The methods showed to perform similar to the results in the signal in white noise model.

For three of the methods from the simulation studies, we computed theoretical results for both the $L_\infty$ distance between the estimator and the truth, and frequentist coverage of pointwise posterior credible sets. The aim was to compare these results to the optimal results shown for the nondistributed setting. To prove these theoretical results, we considered an RKHS framework as proposed in [33]. The use of this framework
made it possible to compute theoretical results for the model without the use of matrix operations. The theoretical results were computed analogously to the results for the nondistributed setting. These results confirmed the observations from the simulation studies, where the studied methods performed similar to these methods in the signal in white noise model.

After the theoretical results for the distributed methods, we performed simulation studies using the squared exponential kernel instead of the Matérn kernel. For this kernel, the distributed methods performed similar as on the Matérn kernel. Finally, we looked into an adaptive method which tunes the parameters of the kernel from the data by maximising the likelihood, as opposed to it being known for all earlier cases. The adaptive methods have been shown to work in a nondistributed setting, but fail to produce good results for the distributed setting. Both the Matérn kernel and the squared exponential kernel show suboptimal recovery and uncertainty quantification when estimating their hyper parameters in distributed setting.

The research can be continued in various directions. The $L_\infty$ bounds proven can be extended to contraction rates. Moreover, there are still distributed methods for which no $L_\infty$ bounds have been proven. The results on the frequentist coverage of pointwise posterior credible sets can be extended to sup-norm credible bounds around the estimator $\hat{f}_{n,m}$. In particular, all of these results can be extended to results for the squared exponential kernel, which might require more work because of the different structure of the eigenvalues and eigenfunctions. The adaptive methods have been studied via simulation, but require theoretical results as well. Moreover, there has not been any distributed method which performs good results in the simulation studies. It rests to find a solution for this, similar to the open problem from [25].
Appendices
A. Appendix

A.1. Proof of Lemma 5

The proof to this lemma follows analogous to the proof of Lemma 5.3 in [33]. We include this proof for self containedness and completeness.

Proof. To begin our proof, we recall a version of the classical Bernstein inequality and a resulting expectation bound.

**Theorem 7** (Bernstein’s inequality). Let $X_1, \ldots, X_n$ be independent random variables. Let $\nu$ and $c$ be positive numbers such that $\sum_{i=1}^{n} \mathbb{E}X_i^2 \leq \nu$, and $\sum_{i=1}^{n} \mathbb{E}|X_i|^q \leq 2^{-1} q ! \nu c^{q-2}$ for $q \geq 3$. Let $S = \sum_{i=1}^{n} (X_i - \mathbb{E}X_i)$. Then, for any $\lambda \in (0, 1/c)$,

$$
\mathbb{E}e^{\lambda S} \leq \exp\left(\frac{\nu \lambda^2}{2(1 - c\lambda)}\right),
$$

which in particular implies, for any $x > 0$,

$$
\mathbb{P}\left(|S| \geq \sqrt{2\nu x + cx}\right) \leq 2e^{-x}.
$$

(A.1)

Further, if $S$ a random variable satisfying (A.1), then for any $q \geq 1$,

$$
\mathbb{E}[S^{2q}] \leq q!(8\nu)^q + (2q!)^2(4c)^2q.
$$

In particular, $\mathbb{E}[S^2] \lesssim C_1 \nu + C_2 c^2$ where $C_1, C_2$ are nonnegative constants.

This proof makes use of the following tail bound for supremum of empirical processes with sub-exponential increments from [1].

**Theorem 8** (Bernstein-type inequality for suprema of random processes (Theorem 2.1 from [1])). Let $(X_t)_{t \in T}$ be a centered family with $T \subset \mathbb{R}^D$ for some finite $D$. Fix some $t_0 \in T$ and let $Z = \sup_{t} |X_t - X_{t_0}|$. Consider norms $d(s, t) = d(s - t)$ and $\delta(s, t) = \delta(s - t)$ on $\mathbb{R}^D$, and assume there exist $v, b > 0, c \geq 0$ such that

$$
T \subset \{t \in \mathbb{R}^D : d(t, t_0) \leq v, \delta(t, t_0) \leq b\}.
$$

Moreover, assume that for all $s, t, s \neq t$,

$$
\mathbb{E}[e^{\lambda (X_s - X_t)}] \leq \exp\left(\frac{\lambda^2 d^2(t, s)}{2(1 - \lambda c \delta(t, s))}\right), \forall \lambda \in \left[0, \frac{1}{c \delta(t, s)}\right].
$$

(A.2)

Then, with $C = 18$,

$$
\mathbb{P}\left(Z \geq C \left(\sqrt{v^2(1 + x) + b(1 + x)}\right)\right) \leq 2e^{-x}, \forall x > 0.
$$

(A.3)
Without loss of generality, we assume \( \sigma = 1 \). Let
\[
U_t = \frac{n}{m} \sum_{i=1}^{\infty} \nu_j \sqrt{m} \left( \sum_{j=1}^{m} \phi_k^2(X_i^{(j)}) \phi_k(t) \right),
\]
where \( z_i \sim N(0,1/m) \), \( X_i \sim U(0,1) \), and \( X_i \) is \( z_j \) for all \( 0 \leq i,j \leq n \). It suffices to find a tail bound for \( ||U||_\infty = \sup_{t \in [0,1]} |U_t| \). For \( t,s \in [0,1] \), write
\[
U_t = \frac{n}{m} \sum_{i=1}^{n/m} z_i \left( \sum_{k=1}^{\infty} \nu_j \phi_k(t) \right) \sqrt{m} \left( \sum_{j=1}^{m} \phi_k^2(X_i^{(j)}) \right),
\]
From this follows
\[
U_t - U_s = \sum_{i=1}^{n/m} z_i \Gamma_{i,s} = \sum_{i=1}^{n/m} z_i \left( \sum_{k=1}^{\infty} \nu_j (\phi_k(t) - \phi_k(s)) \right) \sqrt{m} \left( \sum_{j=1}^{m} \phi_k^2(X_i^{(j)}) \right),
\]
where we denote
\[
\Gamma_{i,s} := \left( \sum_{k=1}^{\infty} \nu_j (\phi_k(t) - \phi_k(s)) \right) \sqrt{m} \left( \sum_{j=1}^{m} \phi_k^2(X_i^{(j)}) \right),
\]
where we will suppress the dependence on \( s \) and \( t \), and write \( \Gamma_i \) from now on. Notice that
\[
\frac{1}{\sqrt{m}} \sqrt{mC_{\phi}^2} \leq 1 \sqrt{mC_{\phi}^2} = C_{\phi},
\]
for \( C_{\phi} \) the uniform bound for \( \{|\phi_j\}_{j \in \mathbb{N}} \).

We now proceed to the proof by defining the two norms \( d \) and \( \delta \). Let \( d(t,s) = (\mathbb{E}[(U_t - U_s)^2])^{\frac{1}{2}} \) and for any \( \kappa \in (0, \alpha \wedge 1) \), let \( \delta(t,s) = |t - s|^\kappa \). Using these norms, we estimate the quantities \( v, b \) and \( c \) in \((A.3)\). First, note that
\[
d^2(t,s) = \mathbb{E} \left[ \sum_{i=1}^{n/m} (z_i \Gamma_i) \right] = \sum_{i=1}^{n/m} m^{-1} \mathbb{E} \Gamma_i^2.
\]
As the eigenfunctions are uniformly bound and orthonormal, we obtain
\[
\mathbb{E} \Gamma_i^2 = \sum_{k=1}^{\infty} \nu_k |\phi_k(t) - \phi_k(s)|^2 \leq 2C_{\phi}^2 \sum_{k=1}^{\infty} \nu_k^2 \leq Ch^{-1}.
\]
The two previous displays combined now imply
\[
\sup_{t,s \in [0,1]} d(t,s) \leq \sqrt{nm^{-2}h^{-1}}.
\]
In order to satisfy condition (A.2) we require to bounds on $E \sum_{i=1}^{n/m} |z_i \Gamma_i|^q$ for $q \geq 3$. We start by bounding $|\Gamma_i|$. For any $i$ and fixed $\kappa \in (0, 1)$, we use the uniform boundness and Lipschitz continuity of the eigenfunctions to obtain

$$|\Gamma_i| \leq \sum_{k=1}^{\infty} \nu_k |\phi_k(t) - \phi_k(s)|$$

$$\leq C_{\phi} \sum_{k=1}^{\infty} \nu_k |\phi_k(t) - \phi_k(s)|^{2C_{\phi}^{1-\kappa}}$$

$$\leq 2C_{\phi}^{2-\kappa} L_{\phi}^\kappa |t-s|^\kappa \sum_{k=1}^{\infty} k^\kappa \nu_k$$

$$\leq Ch^{-1+\kappa} \delta(t, s).$$

This implies that for any $q \geq 3$,

$$E \left[ \sum_{i=1}^{n/m} |z_i \Gamma_i|^q \right] \leq \sum_{i=1}^{n/m} \frac{m^{-n/2} q/2}{E[|\Gamma_i|^{q-2} \Gamma_i^2]} \leq 2^{-1} q! d^2(t, s) \left( h^{-1+\kappa} \delta(t, s) \right)^{q-2},$$

where we used both the fact that $E[|z_i|^q] \lesssim m^{-n/2} q/2$ and $q! \geq (q/e)^q$. For the $|\Gamma_i|^{q-2}$ term, we used the global bound on $\Gamma_i$ and the fact that $\sum_{i=1}^{n/m} m^{-1} \mathbb{E}[\Gamma_i^2] = d^2(t, s)$.

Now let $c = h^{-1+\kappa}$. Then by Theorem 7, we obtain

$$\mathbb{E} e^{\lambda(U_t - U_s)} \leq \exp \left( \frac{\lambda^2 d^2(t, s)}{2(1-\lambda c \delta(t, s))} \right), \quad \forall \lambda \in \left[ 0, \frac{1}{c \delta(t, s)} \right].$$

Thus condition (A.2) is satisfied, which implies quantities $v^2$ and $b$ in (A.3) are given by $nm^{-2} h^{-1}$ and $h^{-1+\kappa}$, respectively. Theorem 8 now states

$$\mathbb{P} \left( \|U\|_\infty \geq 18 \left( \sqrt{nm^{-2} h^{-1}(1+x)} + h^{-1+\kappa}(1+x) \right) \right) \leq 2e^{-x}, \forall x > 0.$$ 

Dividing both sides in the probability measure by $n/m$ proofs the stated result.

\[\square\]

**A.2. Technical lemmas**

**Lemma A.1.** For $r, s, t \geq 0$ with $st > 1$ consider $f(x) = x^{-r}(x^s + 1)^t$ and set

$$f_\nu = \nu^{-1} \sum_{k=1}^{\infty} f(k/\nu).$$

Then as $\nu \to \infty$,

(i) if $r < 1$, then $f_\nu \asymp \int_0^\infty f(x) dx$. 

78
(ii) if $r = 1$, then $f_\nu \asymp \log \nu$.

(iii) if $r > 1$, then $f_\nu \asymp \nu^{r-1} \sum_{k=1}^{\infty} k^{-r}$.

Proof. A proof to this lemma can be found in [24].
Popular summary

In recent times, digital techniques have made it possible to easily keep track of large sets of data. Due to the availability of data, the importance of data analysis has grown immensely. Topics such as machine learning and big data are mentioned frequently. Many modern applications use the historical data of its users to fine tune the application to the user. As the amount of available data grows, this can result in very long computations.

These computations can for example be sped up by approximating the calculations performed, which should return a similar result in lesser time. Another way to speed up the process is to have multiple machines work on the same computation at a time. These methods are called distributed methods. In distributed methods, we assume we have a single set of data which we will denote by $D$. These sets are divided over multiple machines, which perform the computations that would otherwise be done on just one computer separately on their part of the data. After doing these local computations, we will combine them into a single global outcome by a method which we call an aggregation method.

We are interested in performing these methods on the nonparametric regression model. Regression models in general have some kind of data which consists of multiple variables, of which one variables wants to be predicted by all of the other variables. A simple example of this can be to estimate someones weight based on someones height. To do so, we usually consider different functions for which we calculate the Euclidian distance between this prediction and the true value. The function which turns out to have the smallest average distance between the predictions and the true values is than chosen as the optimal choice based on this data set.

As the potential outcomes are limitless, there is usually assumed for the function to be of a certain shape. An example in this case is linear regression, where we look for the straight line that fits the cloud of data the best. In many cases this can work, but for more complex and unknown functions this can give terrible results!

In nonparametric regression we do not assume that the function belongs to some specific parametric family, hence introducing more flexibility into the model. Where parametric regression already assumes the outcome of a function to be of a certain shape, nonparametric regression takes into account this uncertainty and bases its form more on the data.
As in this case we have the prior knowledge that the actual function on the $x$ data points is linear, we know that the linear estimation will generally perform optimally. As we see, the nonparametric estimator is more likely to shape to a coincidental clustering of points. This can be seen as for example when we have three people of 1.82 who are all somehow relatively light for their length, the estimation of 1.82 will also be lower than in the linear case, which is not optimal. We call this phenomenon overfitting to the data, which will in most real life cases not be a problem as a larger sample size makes for less overfitting.

We want to perform the nonparametric regression in a Bayesian way. The world of statistics can roughly be divided in two schools of thought: the frequentist, and the Bayesian. When estimating from a set of data, the frequentist assumes there is always a true value to be found, while the Bayesian says that the reality is more uncertain, and looks for a distribution of possible outcomes. The Bayesian does so by assuming a distribution for a parameter $\theta$ which will likely be of the shape of the outcome, the so called prior distribution. This is combined with the likelihood of the given data $D$, to find a distribution of $\theta$ given $D$.

As we have seen, this is done by first assuming a prior distribution for the function $f$. We will consider so-called Gaussian processes as prior distributions. Gaussian processes are random functions which are defined by a mean function which tells the expected value at a point in time, and a covariance function, which calculates the covariance between two points in time. Some examples of Gaussian processes can be found below.

Figure 4.1.: Linear regression. 
Figure 4.2.: Nonparametric regression.
As we do not want to add too much bias to our outcome, we will first assume the mean function to be 0 for every point in time, where hence the only characteristics of the Gaussian process prior distribution is determined by the function computing the covariance between the time points.

As visible in Figure 4.5, some Gaussian processes are rougher than others. The covariance kernel we choose for our prior distribution, the Matérn kernel, is only dependent of one parameter $\alpha$, which determines the smoothness of the function. There is no single optimal value of the smoothness, this one depends on the smoothness of the to be estimated function $f$ itself.

Combining this with the likelihood of the data, we now obtain a posterior distribution for $f$ given the data.

Here the black line is the original function $f$, the thick blue line is the expected posterior function, and the dashed blue lines around the expected posterior function show the range in which with 95% certainty the posterior function will be in. We conclude that this method performs a good estimation of the original function, and quantifies the uncertainty well.
However, as data sets grow bigger, the estimation process takes up more time. For a data set of $k$ times larger, the process takes up around $k^3$ times as long! That is 1000 times as long for only 10 times as much data. As this takes unacceptably long for large sets of data, we will use the distributed methods as shown earlier to improve the speed at which the estimation is made.

This is done by dividing the data of size $n$ in $m$ parts which we send to $m$ separate computers. These computers will then compute the posterior distribution based on their local data of size $n/m$ which we will aggregate into a single global posterior distribution afterwards. The most obvious way of aggregating these posterior distributions is by taking the average of the local posterior distributions. This means that when we want to estimate $f(x)$ given the data, we take the average of $f_j(x)$, where $f_j(x)$ is the computed estimate of $f(x)$ by computer $j$, using only its own local data.

![Figure 4.6.: Average of local posterior means with corresponding pointwise credible sets.](image)

From this we can conclude that the average of the estimations gives a too smooth estimation. The local posteriors do not have sufficient data to compute a good estimation for the true function. The low influence of the data can be compensated by raising the power of the likelihood to $m$ in this process, and taking the average of the local posteriors afterwards.
Figure 4.7.: Average of local posterior means with corresponding pointwise credible sets when using adjusted local likelihoods.

One can see that this results in a better estimation of the true function $f$. However, in this case, the sets in which 95% of the posterior distributions are contained, are way smaller than before, and do not seem to cover the true function at many coordinates. This leaves room for improvement. As opposed to raising the power of the likelihoods to the power of $m$, we can also raise the power of the prior distribution to $1/m$ in the process of locally computing the posteriors.

Figure 4.8.: Average of local posterior means with corresponding pointwise credible sets when using adjusted priors.

Here we see that the resulting global posterior does not only estimate the true function well, also does the true function lie in the 95% credible sets around every point. Hence there exist distributed methods which perform as well as the nondistributed setting.

We have seen that the Bayesian nonparametric regression methods form good esti-
mation methods for functions with very little prior knowledge. The model allows for much flexibility in the way the prior distribution is chosen, where we saw that Gaussian processes perform well with the Matérn covariance kernel in the nondistributed setting. However, when distributing the data, we have seen that there are many traps to fall in. The local posteriors are computed with an insufficient amount of data, which will result in an oversmoothed global posterior when naively averaging the local posteriors. There are ways to bypass this problem. One is to raise the power of the likelihood of the data to $m$, which as we saw gives a better estimation, but does not quantify the uncertainty well. A better solution is to raise the power of the priors to $1/m$, which does not only perform similar to the nondistributed setting in terms of estimation, but also quantifies the uncertainty well.
Bibliography


