Scalability and uncertainty of Gaussian processes

Hadji, M.A.

CHAPTER 3

Optimal recovery and coverage for distributed Bayesian non-parametric regression

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

§3.1 GP regression framework

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

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

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

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

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

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

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

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

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

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

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

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

The corresponding eigenvalues of K are

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

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

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

- *Polynomially decaying:*

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

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

- *Exponentially decaying:*

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

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

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

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

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

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

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

§3.2 Distributed GP regression

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

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

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

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

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

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

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

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

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

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

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

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

The proof is given in Section 3.5.4.

§3.2.1 Optimal Distributed Methods

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

3.2.1.1 Method I

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

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

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

3.2.1.2 Method II

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

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

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

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

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

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

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

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

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

In particular, the posterior variance function is

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

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

§3.2.2 Posterior contraction rate

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

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

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

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

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

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

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

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

is a (typically) negligible technical term.

The proof of the theorem is deferred to Section 3.5.3.

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

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

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

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

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

The proof is given in Section 3.6.1.

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

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

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

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

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

§3.3 Distributed uncertainty quantification

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The proof is given in Section 3.6.3.

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

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

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

§3.4 Discussion

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

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

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

§3.5 Proofs of the main results

§3.5.1 Kernel Ridge Regression in non-distributed setting

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 3.1: Notation references

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

§3.5.2 Kernel Ridge Regression in distributed setting

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

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

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

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

§3.5.3 Proof of Theorem 3.2.2

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

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

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

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

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

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

where

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

concluding the proof of the first statement.

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

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

Therefore it is sufficient to show that

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

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

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

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

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

concluding the proof of the statement.

3.5.3.1 Proof of (3.5.7)

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

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

Then we show below that

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

which together with the preceding display implies

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

By combining the preceding two displays we arrive at

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

For the aggregated estimator we get that

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

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

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

Finally we verify below that

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

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

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

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

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

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

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

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

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

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

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

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

concluding the proof of (3.5.9).

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

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

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

$$
\begin{aligned}
&E_0 \left\| \tilde{F} \circ F^{-1} \left(\hat{S}_n^{(k)} \left(\tilde{F} \left(\theta_0 \right) \right) - S_n^{(k)} \left(\tilde{F} \left(\theta_0 \right) \right) \right) \right\|_2^2 \\
&\quad \leq 2E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \tilde{P} \left(\theta_0 \right) \left(X_i^{(k)} \right) K_{X_i^{(k)}} - \int_{\mathcal{X}} \tilde{P} \left(\theta_0 \right) (x) \, K_x dx \right) \right\|_2^2 \\
&\qquad + 2E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \varepsilon_i^{(k)} K_{X_i^{(k)}} \right) \right\|_2^2 \\
&\quad =: (T_1 + T_2).
\end{aligned}
$$

We deal with terms T_1 and T_2 separately. In view of Lemma 3.7.1 (with $\vartheta = \tilde{P}(\theta_0)$)

$$
T_1 \leq \frac{2C}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 \left\| \tilde{P} \left(\theta_0 \right) \right\|_2^2,
$$

for some $C > 0$. Since the operator $\tilde{F} \circ F^{-1}$ is linear, we get that

$$
\begin{aligned}
T_2 &= \frac{2}{(n/m)^2} \sum_{i=1}^{n/m} E_0 \left(\left(\varepsilon_i^{(k)} \right)^2 \left\| \tilde{F} \circ F^{-1} \left(K_{X_i^{(k)}} \right) \right\|_2^2 \right) \\
&\quad + \frac{4}{(n/m)^2} \sum_{1 \leq i < \ell \leq n} E_0 \left(\varepsilon_i^{(k)} \varepsilon_\ell^{(k)} \tilde{F} \circ F^{-1} \left(\left\langle K_{X_i^{(k)}}, K_{X_\ell^{(k)}} \right\rangle_2 \right) \right) \\
&= \frac{2\sigma^2}{n/m} E_0 \left\| \tilde{F} \circ F^{-1} \left(K_{X_1^{(k)}} \right) \right\|_2^2 = \frac{2\sigma^2}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2,
\end{aligned}
$$

because the cross terms are equal to 0 due to independence of the noise $\varepsilon_i^{(k)}$, $i = 1, ..., n/m$, $k = 1, ..., m$.

3.5.3.2 Proof of (3.5.8)

In this section we give upper bounds for the learning curves in case of both distributed methods.

Method I: Let us denote by $\mu_j^I = m\mu_j$ the eigenvalues of the local covariance kernel. Then in view of Lemma 3.7.4, the expectations of the m local posterior variances are all of the same order

$$
E_0 E_X \operatorname{Var} \left(\theta(X) | \mathbb{D}_n^{(k)} \right) \asymp \sigma^2 \sum_{j \in \mathbb{N}^d} \frac{\mu_j^I}{\sigma^2 + (n/m)\mu_j^I} = \sigma^2 \sum_{j \in \mathbb{N}^d} \frac{m\mu_j}{\sigma^2 + n\mu_j} = \frac{\sigma^2}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j.
$$

Since the variance of the global posterior distribution $\Pi_{n,m}^I(.|\mathbb{D}_n)$ satisfies the following equality

$$
\operatorname{Var}_{n,m}^I \left(\theta(x) \right) = m^{-2} \sum_{k=1}^{m} \operatorname{Var} \left(\theta(x) | \mathbb{D}_n^{(k)} \right),
$$

one can see that

$$E_0 E_X \operatorname{Var}^I_{n,m}(\theta(X)) \asymp \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j.$$

Method II: First note that $\mu_j^{II} = \mu_j$ the eigenvalues of the local covariance kernel. Note that the expectations of the m local posterior variances are all of the same order

$$E_0 E_X \operatorname{Var}\left(\theta(X)|\mathbb{D}_n^{(k)}\right) \asymp \frac{\sigma^2}{m} \sum_{j \in \mathbb{N}^d} \frac{\mu_j^{II}}{\sigma^2/m + (n/m)\mu_j^{II}} = \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j,$$

because the variance of the noise is σ^2/m for each machine. The variance of the aggregated posterior distribution $\Pi^{II}_{n,m}(.|\mathbb{D}_n)$ satisfies

$$E_0 E_X \operatorname{Var}^{II}_{n,m}(\theta(X)|\mathbb{D}_n) \asymp \frac{\sigma^2}{n} \sum_{j \in \mathbb{N}^d} \nu_j$$

because we know that

$$\operatorname{Var}^{II}_{n,m}(\theta(X)|\mathbb{D}_n) = m^{-1} \sum_{k=1}^{m} \operatorname{Var}\left(\theta(x)|\mathbb{D}_n^{(k)}\right)$$

proving assertion (3.5.8).

§3.5.4 Proof of Theorem 3.2.1

The proof follows similar lines of reasoning as Theorem 3.2.2, where we provided general upper bounds for the contraction rate of the distributed posterior.

First we prove (3.2.3). For the naive averaging method the local sample and population score functions coincide to the non-distributed case given in Section 3.5.1 with sample size n/m, i.e.

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

$$S_n^{*(k)}(\theta) = \int_{\mathcal{X}} (\theta_0(x) - \theta(x))\, K_x dx - \frac{\sigma^2}{n/m}\theta = F(\theta_0 - \theta) - \frac{\sigma^2}{n/m}\theta.$$

Note that the solution of the equation $S_n^{*(k)}(\theta) = 0$ is given by the coefficients $\theta_j = \nu_j^* \theta_{0,j}$, with $\nu_j^* = \frac{n\mu_j}{m\sigma^2 + n\mu_j}$, $j \in \mathbb{N}^d$.

Then using the inequality $a^2 \geq (a-b)^2/2 - b^2$ one can obtain that

$$E_0 \left\|\hat{\theta}^*_{n,m} - \theta_0\right\|_2^2 \geq \frac{1}{2}\left\|\theta_0 - \tilde{F}^*(\theta_0)\right\|_2^2 - E_0 \left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2^2,$$

where $\tilde{F}^*(\theta) = \sum_{j \in \mathbb{N}} \nu_j^* \theta_j \psi_j$ and $\hat{\theta}^*_{n,m}$ is the mean of the global posterior $\Pi^{\dagger}_{n,m}(.|\mathbb{D}_n)$ obtained with the naive averaging method.

First note that

$$\left\|\theta_0 - \tilde{F}^*(\theta_0)\right\|_2^2 = \sum_{j=1}^{\infty} \frac{m}{m\sigma^2 + n\mu_j}\theta_{0,j}^2 \geq \frac{c_L}{2} \sum_{(n/(m\sigma^2))^{1/(2\beta+1)} \leq j} j^{-1-2\beta}(\log j)^{-2}$$
$$\geq c_0(n/m)^{-2\beta/(2\beta+1)}(\log(n/m))^{-2}, \tag{3.5.13}$$

for some small enough $c_0 > 0$. We conclude the proof of (3.2.3) by showing below that $E_0\|\hat{\theta}_{n,m} - \tilde{F}(\theta_0)\|_2^2 = o((n/m)^{-2\beta/(2\beta+1)}(\log(n/m))^{-2})$.

Similarly to (3.5.7) we can derive (by replacing $\tilde{F}$ and ν with $\tilde{F}^*$ and ν^*, respectively) that

$$E_0\left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2^2 \lesssim \left(\frac{1}{n}\sum_{j=1}^{\infty}(\nu_j^*)^2\right)\left(\left\|\tilde{P}^*(\theta_0)\right\|_2^2 + \sigma^2\right) + \delta_n^*, \tag{3.5.14}$$

where $\delta_n^* = n\sum_{j=1}^{\infty}(\nu_j^*)^2\sum_{\ell=I}^{\infty}\mu_\ell$, with $I = \frac{n}{m^2\log n}(\sum_{j=1}^{\infty}(\nu_j^*)^2)^{-1}$. Note that $\|\tilde{P}^*(\theta_0)\|_2^2 = O(1)$ and in view of Lemma 3.7.5, $\sum_{j=1}^{\infty}(\nu_j^*)^2 \asymp (n/m)^{1/(2\beta+1)}$; hence

$$I \asymp \frac{(n/m)^{2\beta/(2\beta+1)}}{m\log n}.$$

Therefore the first term on the right hand side of (3.5.14) is $O(n^{-2\beta/(2\beta+1)}m^{-1/(2\beta+1)})$ and

$$\delta_n^* \lesssim n(n/m)^{1/(1+2\beta)}I^{-2\beta} \asymp n^{2-2\beta}m^{-1+4\beta}(\log n)^{2\beta} = o\left((\log(n/m))^{-2}\right),$$

where the last step holds for large enough choice of β and not too large choice of m. For instance taking $\beta > 2$ and $m = o(n^{1/(2+2\beta)})$, we get that

$$\delta_n^*(n/m)^{2\beta/(2\beta+1)} \lesssim n^{-c_0}\log^4 n = o\left((\log(n/m))^{-2}\right),$$

for some $c_1 > 0$.

It remained to deal with (3.2.4). First note that by the computations above combined with Markov's inequality there exists a sequence $\rho_n \to 0$ such that

$$P_0\left(\left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2 \geq \rho_n(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1}\right) \to 0.$$

Then by triangle inequality, (3.5.13) and Markov's inequality we get for $c < c_0$ that

$$E_0\Pi^*_{n,m}\left(\theta : \|\theta - \theta_0\|_2 \leq c(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1}|\mathbb{D}_n\right)$$
$$\leq E_0\Pi^*_{n,m}\Big(\left\|\theta_0 - \tilde{F}^*(\theta_0)\right\|_2 - \left\|\hat{\theta}^*_{n,m} - \tilde{F}^*(\theta_0)\right\|_2$$
$$- c(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1} \leq \left\|\theta - \hat{\theta}^*_{n,m}\right\|_2|\mathbb{D}_n\Big)$$
$$\leq E_0\Pi^*_{n,m}\left((c_0 - c - \rho_n)(n/m)^{-\beta/(2\beta+1)}(\log(n/m))^{-1} \leq \left\|\theta - \hat{\theta}^*_{n,m}\right\|_2|\mathbb{D}_n\right) + o(1)$$
$$\lesssim (n/m)^{2\beta/(2\beta+1)}(\log(n/m))^2E_0E^*_{n,m}\left\|\theta - \hat{\theta}^*_{n,m}\right\|_2^2.$$

We conclude the proof by noting that

$$E_0 E_X \operatorname{Var}\left(\theta(X)|\mathbb{D}_n^{(k)}\right) = \sigma^2 \sum_{j=1}^{\infty} \frac{\mu_j}{\sigma^2 + (n/m)\mu_j} = \frac{\sigma^2}{n/m} \sum_{j=1}^{\infty} \nu_j^*,$$

for all $k \in \{1, ..., m\}$; hence

$$E_0 E_{n,m}^* \left\|\theta - \hat{\theta}_{n,m}^*\right\|_2^2 = \frac{1}{m^2} \sum_{k=1}^{m} E_0 E_X \operatorname{Var}\left(\theta(X)|\mathbb{D}_n^{(k)}\right) = \frac{\sigma^2}{n} \sum_{j=1}^{\infty} \nu_j^* \lesssim \frac{\sigma^2}{m} (n/m)^{-2\beta/(2\beta+1)}.$$

§3.5.5 Proof of Theorem 3.3.1

We first consider the non-distributed case $m = 1$ for clearer presentation and then extend our results to the distributed setting.

3.5.5.1 Non-distributed setting

Connection to KRR Similarly to the posterior mean, the posterior covariance function $\hat{C}_n$ can be given as

$$\hat{C}_n(x, x') = K(x, x') - \hat{K}_n(x, x'),$$

where $\hat{K}_n(x, \cdot) = K(\cdot, \mathbb{X})[K(\mathbb{X}, \mathbb{X}) + \sigma^2 I_n]^{-1} K(\mathbb{X}, x)$, or equivalently

$$\hat{K}_{x,n} = \hat{K}_n(x, \cdot) = \arg\min_{\vartheta \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(K(x, X_i) - \vartheta(X_i)\right)^2 + \frac{\sigma^2}{n} \|\vartheta\|_{\mathcal{H}}^2 \right], \tag{3.5.15}$$

see assertion (8) of (Bhattacharya et al., 2017).

Then by taking the Fréchet derivative of the expression on the right hand side we arrive to the (adjusted) score function and its expected value

$$\hat{S}_{K_x,n}(\vartheta) = \frac{1}{n} \left(\sum_{i=1}^{n} \left(K_x(X_i) - \vartheta(X_i)\right) K_{X_i} - \sigma^2 \vartheta \right),$$
$$S_{K_x,n}(\vartheta) = E\hat{S}_{K_x,n}(\vartheta) = \int_{\mathcal{X}} \left(K_x(z) - \vartheta(z)\right) K_z dz - \frac{\sigma^2}{n} \vartheta.$$

Then similarly to the posterior mean in Section 3.5.1 the following assertions hold

$$S_{K_x,n}(\vartheta) = F(K_x) - F \circ \tilde{F}^{-1}(\vartheta) = F\left(K_x - \tilde{F}^{-1}(\vartheta)\right), \tag{3.5.16}$$

$$\Delta\hat{K}_{x,n} = \hat{K}_{x,n} - \tilde{F}(K_x) = -\tilde{F} \circ F^{-1} \circ S_{K_x,n}\left(\hat{K}_{x,n}\right), \tag{3.5.17}$$

$$\hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \tilde{P}(K_x)(X_i)K_{X_i} - \sigma^2\tilde{F}(K_x)\right), \tag{3.5.18}$$

$$\begin{aligned} F \circ \tilde{F}^{-1}\left(\Delta\hat{K}_{x,n}\right) - \hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right) \\ = -\frac{1}{n}\sum_{i=1}^{n} \Delta\hat{K}_{x,n}(X_i)K_{X_i} + \int_{\mathcal{X}} \Delta\hat{K}_{x,n}(x')K_{x'}dx', \end{aligned} \tag{3.5.19}$$

and note that $\hat{K}_{x,n}$ and $\tilde{F}(K_x)$ are the zero points of the functions $\hat{S}_{K_x,n}$ and $S_{K_x,n}$, respectively.

Under-smoothing Following from the triangle inequality, to obtain frequentist coverage for the credible ball it is sufficient to show that for $L_n \to \infty$

$$P_0\left(\left\|\tilde{P}(\theta_0)\right\|_2 + \left\|\hat{\theta}_n - \tilde{F}(\theta_0)\right\|_2 \leq L_n r_{n,\gamma}\right) \to 1.$$

The preceding display is implied by assumption (3.3.1) and assertions

$$P_0\left(\left\|\Delta\hat{\theta}_n\right\|_2^2 \leq L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \to 1, \tag{3.5.20}$$

$$P_0\left(r_{n,\gamma}^2 \geq \frac{1}{2C_\psi^2}\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \to 1, \tag{3.5.21}$$

where $\Delta\hat{\theta}_n := \hat{\theta}_n - \tilde{F}(\theta_0)$, verified below.

Proof of (3.5.20)**:** In view of assertion (3.5.7) with $m = 1$ and Markov's inequality we get

$$\begin{aligned} P_0\left(\left\|\Delta\hat{\theta}_n\right\|_2^2 \leq L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) &\leq \frac{E_0\left\|\Delta\hat{\theta}_n\right\|_2^2}{L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j} \\ &\lesssim \frac{\left(\frac{1}{n}\sum_{j\in\mathbb{N}^d}\nu_j^2\right) + \delta_n}{L_n\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j} \\ &= O\left(\frac{1}{L_n} + \frac{n\delta_n}{\sum_{j\in\mathbb{N}^d}\nu_j}\right) = o(1). \end{aligned}$$

Proof of (3.5.21)**:** The radius $r_{n,\gamma}$ is defined, conditionally on $\mathbb{X}$, as $P(\|W_n\|_2^2 \leq r_{n,\gamma}^2|\mathbb{X}) = 1-\gamma$, where W_n is a centered GP with covariance kernel $\hat{C}_n$ given in (3.1.3).

In view of Chebyshev's inequality

$$r_{n,\gamma}^2 \geq E\left[\|W_n\|_2^2|\mathbb{X}\right] - (1-\gamma)^{-1/2}\operatorname{Var}\left(\|W_n\|_2^2|\mathbb{X}\right)^{1/2}.$$

Using Fubini's theorem, the first term on the right hand side of the preceding display can be rewritten as

$$E\left[\|W_n\|_2^2|\mathbb{X}\right] = E\left[\|\theta - \hat{\theta}_n\|_2^2|\mathbb{D}_n\right] = \int_{\mathcal{X}} \operatorname{Var}\left(\theta(x)|\mathbb{D}_n\right) dx.$$

The integral on the right-hand side of the display, called the *generalization error*, see Chapter 7 of (Rasmussen and Williams, 2006), is asymptotically bounded from below almost surely by

$$\sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j \sup_{x\in\mathcal{X}} \psi_j^2(x)} \geq \frac{\sigma^2 C_\psi^{-2}}{n} \sum_{j\in\mathbb{N}^d} \nu_j, \tag{3.5.22}$$

in view of assertion (12) of (Opper and Vivarelli, 1999) and Assumption 3.1.1. Furthermore, the variance of $\|W_n\|_2^2$, conditional on the design $\mathbb{X}$, is

$$\operatorname{Var}\left(\|W_n\|_2^2|\mathbb{X}\right) = E\left[\|W_n\|_2^4|\mathbb{X}\right] - E^2\left[\|W_n\|_2^2|\mathbb{X}\right].$$

The first term on the right hand-side satisfies

$$\begin{aligned}
E\left[\|W_n\|_2^4|\mathbb{X}\right] &= E\left[\|\theta - \hat{\theta}_n\|_2^4|\mathbb{D}_n\right] \qquad (3.5.23)\\
&= \int \left(\int_{\mathcal{X}} \left(\theta(x) - \hat{\theta}_n(x)\right)^2 dx \int_{\mathcal{X}} \left(\theta(x') - \hat{\theta}_n(x')\right)^2 dx'\right) \Pi(d\theta|\mathbb{D}_n)\\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int \left(\theta(x) - \hat{\theta}_n(x)\right)^2 \left(\theta(x') - \hat{\theta}_n(x')\right)^2 \Pi(d\theta|\mathbb{D}_n) dx dx'\\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \left(\operatorname{Var}\left(\theta(x)|\mathbb{D}_n\right)\operatorname{Var}\left(\theta(x')|\mathbb{D}_n\right) + 2\hat{C}_n(x,x')^2\right) dx' dx\\
&= \left(\int_{\mathcal{X}} \operatorname{Var}\left(\theta(x)|\mathbb{D}_n\right) dx\right)^2 + 2\int_{\mathcal{X}} \left\|\hat{C}_n(x,.)\right\|_2^2 dx\\
&= E^2\left[\|W_n\|_2^2|\mathbb{X}\right] + 2\int_{\mathcal{X}} \left\|\hat{C}_n(x,.)\right\|_2^2 dx, \qquad (3.5.24)
\end{aligned}$$

using Fubini's theorem and the reduction formula $EX_1^2X_2^2 = Var(X_1)Var(X_2) + 2Cov(X_1,X_2)^2$ for X_1, X_2 centered Gaussian random variables, see for instance page 189 of (Isserlis, 1916). Hence, again in view of Fubini's theorem,

$$E_0 \operatorname{Var}\left(\|W_n\|_2^2|\mathbb{X}\right) = 2\int_{\mathcal{X}} E_0 \left\|\hat{C}_n(x,.)\right\|_2^2 dx. \tag{3.5.25}$$

Recall that the covariance function $\hat{C}_n(x,x') = K(x,x') - \hat{K}_n(x,x')$, where $\hat{K}_{x,n} = \hat{K}_n(x,.)$ is the solution to (3.5.15). We show below that for all $x \in \mathcal{X}$

$$E_0 \left\|\hat{C}_n(x,.)\right\|_2^2 \lesssim \left\|\tilde{P}(K_x)\right\|_2^2 + \tilde{\delta}_n, \tag{3.5.26}$$

for

$$\tilde{\delta}_n = \inf\left\{\left(\sum_{j\in\mathbb{N}^d}\nu_j^2\right)^2 \sum_{\ell\in\mathcal{I}^c}\mu_\ell : \mathcal{I}\subset\mathbb{N}^d, |\mathcal{I}| = o\left(n\left(\sum_{j\in\mathbb{N}^d}\nu_j^2\right)^{-1}\right)\right\}.$$

In view of the definition of the linear operator $\tilde{P}$ and the eigenvalues ν_j, μ_j we get

$$\tilde{P}\left(K(x,x')\right) = \sum_{j\in\mathbb{N}^d}(1-\nu_j)\mu_j\boldsymbol{\psi}_j(x)\boldsymbol{\psi}_j(x') = \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\boldsymbol{\psi}_j(x)\boldsymbol{\psi}_j(x'), \tag{3.5.27}$$

for all $x, x' \in \mathcal{X}$. Then by combining the last three displays

$$\begin{aligned}
E_0\,\mathrm{Var}\left(\|W_n\|_2^2|\mathbb{X}\right) &= 2\int_{\mathcal{X}} E_0\left\|\hat{C}_n(x,.)\right\|_2^2 dx \\
&\lesssim \int_{\mathcal{X}}\left\|\tilde{P}(K(x,.))\right\|_2^2 dx + \tilde{\delta}_n \\
&= \left(\frac{\sigma^2}{n}\right)^2\int_{\mathcal{X}}\sum_{j\in\mathbb{N}^d}\nu_j^2\boldsymbol{\psi}_j(x)^2dx + \delta_n\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{n} \\
&= \left(\frac{\sigma^2}{n}\right)^2\sum_{j\in\mathbb{N}^d}\nu_j^2 + \delta_n\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{n}.
\end{aligned} \tag{3.5.28}$$

Therefore, by Markov's inequality and Lemmas 3.7.5 and 3.7.6,

$$P_0\left(\mathrm{Var}\left(\|W_n\|_2^2|\mathbb{X}\right)^{1/2} \geq t\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \lesssim t^{-2}\left(\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{(\sum_{j\in\mathbb{N}^d}\nu_j)^2} + \frac{n\delta_n\sum_{j\in\mathbb{N}^d}\nu_j^2}{(\sum_{j\in\mathbb{N}^d}\nu_j)^2}\right) \to 0$$

for all $t > 0$. Hence by combining (3.5.22) and the preceding display (with $t = (1-\gamma)^{1/2}C_\psi^{-2}/2$),

$$P_0\left(E\left[\|W_n\|_2^2|\mathbb{X}\right] - (1-\gamma)^{-1/2}\,\mathrm{Var}\left(\|W_n\|_2^2|\mathbb{X}\right)^{1/2} \geq (C_\psi^{-2}/2)\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j\right) \to 1.$$

This implies that all the quantiles of $\|W_n\|_2^2$, conditionally on $\mathbb{X}$, are of the order $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$ with P_0-probability going to one, including $r_{n,\gamma}^2$.

Proof of (3.5.26)**:** First note that by the inequality $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\left\|\hat{C}_n(x,.)\right\|_2^2 \leq 2\left\|\tilde{P}(K_x)\right\|_2^2 + 2\left\|\Delta\hat{K}_{x,n}\right\|_2^2,$$

where $\Delta\hat{K}_{x,n} = \hat{K}_{x,n} - \tilde{F}(K_x)$.

Next we give an upper bound for the second term of the preceding display similarly to Section 3.5.3.1. First note that

$$\left\|\Delta\hat{K}_{x,n}\right\|_2^2 \lesssim \left\|\Delta\hat{K}_{x,n} - \tilde{F}\circ F^{-1}\circ\hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2 + \left\|\tilde{F}\circ F^{-1}\circ\hat{S}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2.$$

Then by showing below that

$$E_0 \left\| \Delta\hat{K}_{x,n} - \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n}(\tilde{F}(K_x)) \right\|_2^2 \leq o\left(E_0 \left\| \Delta\hat{K}_{x,n} \right\|_2^2 \right) + \tilde{\delta}_n, \qquad (3.5.29)$$

we arrive at

$$E_0 \left\| \Delta\hat{K}_{x,n} \right\|_2^2 \lesssim E_0 \left\| \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 + \tilde{\delta}_n.$$

Next, in view of (3.5.18),

$$\begin{aligned} E_0 \left\| \tilde{F} \circ F^{-1} \circ \hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 &= E_0 \left\| \tilde{F} \circ F^{-1} \left(\hat{S}_{K_x,n} \left(\tilde{F}(K_x) \right) - S_{K_x,n} \left(\tilde{F}(K_x) \right) \right) \right\|_2^2 \\ &= E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n} \sum_{i=1}^n \tilde{P}(K_x)(X_i) K_{X_i} - \int_{\mathcal{X}} \tilde{P}(K_x)(x') K_{x'} dx' \right) \right\|_2^2 \\ &\leq \left(\frac{1}{n} \sum_{j \in \mathbb{N}^d} \nu_j^2 \right) \left\| \tilde{P}(K_x) \right\|_2^2 = o\left(\left\| \tilde{P}(K_x) \right\|_2^2 \right), \end{aligned}$$

where the last line follows from Lemma 3.7.1 with $\vartheta = \tilde{P}(K_x)$ (and $m = 1$), concluding the proof of (3.5.26).

Proof of (3.5.29)**:** Similarly to (3.5.12), by using assertion (3.5.19), Lemma 3.7.2 (with $\hat{\vartheta} = \Delta\hat{K}_{x,n}$, sample size n) and Lemma 3.7.3 (with $m = 1$), we can show that for all $x \in \mathcal{X}$

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

Taking the infimum over $|\mathcal{I}| = o\big(n/(\log n \sum_{j \in \mathbb{N}^d} \nu_j^2)\big)$, we get that the left hand-side of the preceding display is bounded from above by $o(E_0 \|\Delta\hat{K}_{x,n}\|_2^2) + \tilde{\delta}_n$, concluding the proof of the statement.

Over-smoothing By the definition of credible sets and using the triangle inequality, we get that

$$\begin{aligned} P_0 \left(\theta_0 \in \hat{B}_{n,\gamma}(L) \right) &\leq P_0 \left(\left\| \tilde{P}(\theta_0) \right\|_2 \leq \left\| \hat{\theta}_n - \tilde{F}(\theta_0) \right\|_2 + L r_{n,\gamma} \right) \\ &\leq P_0 \left(\left\| \tilde{P}(\theta_0) \right\|_2 \leq 2 \left\| \hat{\theta}_n - \tilde{F}(\theta_0) \right\|_2 \right) + P_0 \left(\left\| \tilde{P}(\theta_0) \right\|_2 \leq 2 L r_{n,\gamma} \right) \end{aligned}$$

and we show below that both probabilities on the right hand side tend to zero.

The first term disappears in view of (3.5.20) and assumption (3.3.2). For the second term note, that in view of Markov's inequality and $P_0(\|W_n\|_2^2 \geq r_{n,\gamma}^2|\mathbb{X}) = \gamma$, where W_n is a centered GP with covariance kernel $\hat{C}_n$, we have $\gamma r_{n,\gamma}^2 \leq E[\|W_n\|_2^2|\mathbb{X}]$. Then

$$P_0\left(2Lr_{n,\gamma}^2 \geq \left\|\tilde{P}(\theta_0)\right\|_2^2\right) \leq P_0\left(E\left[\|W_n\|_2^2|\mathbb{X}\right] \geq \frac{\gamma}{2L}\left\|\tilde{P}(\theta_0)\right\|_2^2\right) \leq \frac{2LE_0(\int_{\mathcal{X}} \mathrm{Var}\left(\theta(x)|\mathbb{D}_n\right)dx)}{\gamma\left\|\tilde{P}(\theta_0)\right\|_2^2}. \tag{3.5.30}$$

The expectation in the numerator, known as the *learning curve*, is of order $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$ according to Lemma 3.7.4; thus for all $L > 0$ not depending on n the right hand side of the preceding display goes to 0 in view of assumption (3.3.2).

3.5.5.2 Distributed setting

Preliminary results. We start by introducing the distributed version of the notations introduced in Section 3.5.5.1. The aggregated posterior covariance function is $\hat{C}_{n,m}^I(x,x') = m^{-2}\sum_{k=1}^m \hat{C}_n^{I,(k)}(x,x')$, where the local posterior covariance functions can be given as $\hat{C}_n^{I,(k)}(x,x') = K_x^I(x') - \hat{K}_x^{I,(k)}(x')$ with

$$\begin{aligned}\hat{K}_x^{I,(k)}(\cdot) &= K^I\left(\cdot,\mathbb{X}^{(k)}\right)\left[K^I\left(\mathbb{X}^{(k)},\mathbb{X}^{(k)}\right)+\sigma^2 I_{n/m}\right]^{-1}K^I\left(\mathbb{X}^{(k)},x\right)\\ &= mK\left(\cdot,\mathbb{X}^{(k)}\right)\left[K\left(\mathbb{X}^{(k)},\mathbb{X}^{(k)}\right)+m^{-1}\sigma^2 I_{n/m}\right]^{-1}K\left(\mathbb{X}^{(k)},x\right).\end{aligned}$$

Then in view of (3.5.15),

$$m^{-1}\hat{K}_x^{I,(k)} = \arg\min_{\vartheta\in\mathcal{H}}\frac{1}{n/m}\left[\sum_{i=1}^{n/m}\left(K_x(X_i^{(k)}) - \vartheta(X_i^{(k)})\right)^2 + \frac{\sigma^2}{m}\|\vartheta\|_{\mathcal{H}}^2\right].$$

For convenience let us introduce the notation $\tilde{K}_x^{I,(k)} = m^{-1}\hat{K}_x^{I,(k)}$. Then the corresponding score function (up to constant multipliers) is given by

$$\hat{S}_{K_x,n}^{I,(k)}(\vartheta) = \frac{1}{n/m}\left(\sum_{i=1}^{n/m}\left(K_x(X_i^{(k)}) - \vartheta(X_i^{(k)})\right)K_{X_i^{(k)}} - \frac{\sigma^2}{m}\vartheta\right)$$

satisfying $\hat{S}_{K_x,n}^{I,(k)}(\tilde{K}_x^{I,(k)}) = 0$. Furthermore the expected value of the score function is

$$S_{K_x,n}^I(\vartheta) = E\hat{S}_{K_x,n}^{I,(k)}(\vartheta) = \int_{\mathcal{X}}\left(K_x(z) - \vartheta(z)\right)K_z dz - \frac{\sigma^2}{n}\vartheta = S_{K_x,n}(\vartheta),$$

hence $S_{K_x,n}^I(\tilde{F}(K_x)) = 0$.

Then similarly to the posterior mean in Section 3.5.1 the following assertions hold

$$\Delta\tilde{K}_x^{I,(k)} = \tilde{K}_x^{I,(k)} - \tilde{F}(K_x) = -\tilde{F} \circ F^{-1} \circ S^I_{K_x,n}\left(\tilde{K}_x^{I,(k)}\right),$$

$$\hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right) = \frac{1}{n/m}\left(\sum_{i=1}^{n/m} \tilde{P}(K_x)\left(X_i^{(k)}\right) K_{X_i^{(k)}} - \frac{\sigma^2}{m}\tilde{F}(K_x)\right), \tag{3.5.31}$$

$$\begin{aligned} F \circ \tilde{F}^{-1}\left(\Delta\tilde{K}_x^{I,(k)}\right) &- \hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right) \\ &= -\frac{1}{n/m}\sum_{i=1}^{n/m} \Delta\tilde{K}_x^{I,(k)}\left(X_i^{(k)}\right) K_{X_i^{(k)}} + \int_{\mathcal{X}} \Delta\tilde{K}_x^{I,(k)}(x') K_{x'} dx'. \end{aligned} \tag{3.5.32}$$

Main assertions. Similarly to the non-distributed case in Section 3.5.5.1, for the coverage of the credible sets it is sufficient to show that

$$P_0\left(r^2_{n,m}(\gamma) \geq C_2 \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d} \nu_j\right) \to 1, \tag{3.5.33}$$

$$P_0\left(\left\|\hat{\theta}_{n,m} - \tilde{F}(\theta_0)\right\|_2^2 \leq L_n \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d} \nu_j\right) \to 1, \tag{3.5.34}$$

where the radius $r_{n,m}(\gamma)$ is defined as $P(\|W_{n,m}\|_2^2 \leq r^2_{n,m}(\gamma)|\mathbb{X}) = 1-\gamma$ and $W_{n,m}$ is a centered GP with the same covariance kernel as $\Pi^{\dagger}_{n,m}(.|\mathbb{D}_n)$. Furthermore, the lack of coverage under (3.3.2) follows from

$$P_0\left(Lr^2_{n,m}(\gamma) \geq \left\|\tilde{P}(\theta_0)\right\|_2^2\right) \to 0. \tag{3.5.35}$$

We prove below the above assertions.

Proof of (3.5.33)**:** Similarly to the proof of (3.5.21) we get by Chebyshev's inequality that

$$r^2_{n,m}(\gamma) \geq E\left[\|W_{n,m}\|_2^2|\mathbb{X}\right] - (1-\gamma)^{-1/2}\operatorname{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right)^{1/2}.$$

Then in view of

$$\operatorname{Var}^I_{n,m}\left(\theta(x)\right) = m^{-2}\sum_{k=1}^{m}\operatorname{Var}^I\left(\theta(x)|\mathbb{D}_n^{(k)}\right), \qquad \text{for all } x\in\mathcal{X}, \tag{3.5.36}$$

and Lemma 3.7.4 it holds almost surely that

$$E\left[\|W_{n,m}\|_2^2|\mathbb{X}\right] = \int_{x\in\mathcal{X}} \operatorname{Var}^I_{n,m}\left(\theta(x)\right) dx \gtrsim \frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j. \tag{3.5.37}$$

Furthermore, as in (3.5.25),

$$\operatorname{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right) = 2\int_{\mathcal{X}}\left\|\hat{C}^I_{n,m}(x,.)\right\|_2^2 dx.$$

Recall that the covariance function $\hat{C}^I_{n,m}(x,x') = m^{-2}\sum_{k=1}^m \hat{C}^{I,(k)}_n(x,x')$. Then in view of $(\sum_{i=1}^m a_i)^2 \le m\sum_{i=1}^m a_i^2$,

$$\left\|\hat{C}^I_{n,m}(x,.)\right\|_2^2 = \left\|m^{-2}\sum_{k=1}^m \hat{C}^{I,(k)}_n(x,.)\right\|_2^2 \le m^{-3}\sum_{k=1}^m \left\|\hat{C}^{I,(k)}_n(x,.)\right\|_2^2.$$

We show below that

$$E_0\left\|\hat{C}^{I,(k)}_n(x,.)\right\|_2^2 \lesssim m^2\left(\left\|\tilde{P}(K_x)\right\|_2^2 + \tilde{\delta}_n\right), \tag{3.5.38}$$

for $\tilde{\delta}_n = \inf\{(\sum_{j\in\mathbb{N}^d}\nu_j^2)^2\sum_{\ell\in\mathcal{I}^c}\mu_\ell : |\mathcal{I}| \le n/(m^2\log n\sum_{j\in\mathbb{N}^d}\nu_j^2)\}$ similarly to the non-distributed case. Then in view of assertion (3.5.27), the variance of $\|W_{n,m}\|_2^2$, similarly to (3.5.28), is bounded from above by

$$\begin{aligned} E_0\,\mathrm{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right) &= 2\int_{\mathcal{X}} E_0\left\|\hat{C}^I_{n,m}(x,.)\right\|_2^2 dx \\ &\lesssim \left(\int_{\mathcal{X}}\left\|\tilde{P}(K^I_x)\right\|_2^2 dx + \tilde{\delta}_n\right) \\ &= \frac{\sigma^4}{n^2}\sum_{j\in\mathbb{N}^d}\nu_j^2 + \tilde{\delta}_n. \end{aligned}$$

Hence for all $t>0$ we get by Markov's inequality and Lemmas 3.7.5 and 3.7.6 that

$$\begin{aligned} &P_0\left(\mathrm{Var}\left(\|W_{n,m}\|_2^2|\mathbb{X}\right) \ge t\frac{\sigma^4}{n^2}\left(\sum_{j\in\mathbb{N}^d}\nu_j\right)^2\right) \\ &\qquad\lesssim t^{-2}\left(\frac{\sum_{j\in\mathbb{N}^d}\nu_j^2}{\left(\sum_{j\in\mathbb{N}^d}\nu_j\right)^2} + \frac{\tilde{\delta}_n n^2}{\sigma^4\left(\sum_{j\in\mathbb{N}^d}\nu_j\right)^2}\right) = o(1). \end{aligned}$$

Hence with P_0-probability tending to one $E[\|W_{n,m}\|_2^2|\mathbb{X}_n]$ is of higher order than $\mathrm{Var}(\|W_{n,m}\|_2^2)^{1/2}$. Therefore, the quantiles of $\|W_{n,m}\|_2^2$ are of the order $(\sigma^2/n)\sum_{j\in\mathbb{N}^d}\nu_j$ with P_0-probability going to one, including $r^2_{n,m}(\gamma)$.

Proof of (3.5.38)**:** We adapt the proof of (3.5.26) to the distributed setting. First note that

$$\begin{aligned} \left\|\hat{C}^{I,(k)}_n(x,.)\right\|_2^2 \lesssim m^2\Bigg(&\left\|\tilde{P}(K_x)\right\|_2^2 + \left\|\Delta\tilde{K}^{I,(k)}_x - \tilde{F}\circ F^{-1}\circ\hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2 \\ &+ \left\|\tilde{F}\circ F^{-1}\circ\hat{S}^{I,(k)}_{K_x,n}\left(\tilde{F}(K_x)\right)\right\|_2^2\Bigg), \end{aligned}$$

where $\Delta\tilde{K}_x^{I,(k)} = \hat{K}_x^{I,(k)}/m - \tilde{F}(K_x)$. Then for, in view of (3.5.31), we get that

$$
\begin{aligned}
&E_0 \left\| \tilde{F} \circ F^{-1} \circ \hat{S}^{I,(k)}_{K_x,n} \left(\tilde{F}(K_x) \right) \right\|_2^2 \\
&\quad = E_0 \left\| \tilde{F} \circ F^{-1} \left(\hat{S}^{I,(k)}_{K_x,n} \left(\tilde{F}\big(K_x\big) \right) - S^{I,(k)}_{K_x,n} \left(\tilde{F}\big(K_x\big) \right) \right) \right\|_2^2 \\
&\quad = E_0 \left\| \tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \tilde{P}(K_x) \left(X_i^{(k)} \right) K_{X_i^{(k)}} - \int_{\mathcal{X}} \tilde{P}(K_x)(x') K_{x'} dx' \right) \right\|_2^2 \\
&\quad \leq \left(\frac{1}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 \right) \|\tilde{P}(K_x)\|_2^2 = o\left(\|\tilde{P}(K_x)\|_2^2 \right),
\end{aligned}
$$

where the penultimate inequality follows from Lemma 3.7.1 with $\vartheta = \tilde{P}(K_x)$.

Furthermore, similarly to the proof in Section 3.5.5.1, by using assertion (3.5.32), Lemma 3.7.2 (with $\hat{\vartheta}^{(k)} = \Delta\tilde{K}_x^{I,(k)}$, sample size n/m) and Lemma 3.7.3, we can show that for all $x \in \mathcal{X}$

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

Taking the infimum over $|\mathcal{I}| = o\big(n/(m \log n \sum_{j \in \mathbb{N}^d} \nu_j^2)\big)$ we get that the left hand side of the preceding display is bounded from above by $o(E_0 \|\Delta\tilde{K}^{(k)}_{x,n}\|_2^2) + \tilde{\delta}_n$. We conclude the proof of (3.5.38) by combining the above three displays.

Proof of (3.5.34)**:** Exactly the same as the proof of (3.5.20).

Proof of (3.5.35)**:** Similarly to assertion (3.5.30) we get in view of (3.5.36) and Lemma 3.7.4 in the case where assumption (3.3.2) holds

$$
\begin{aligned}
P_0 \left(L r_{n,m}^2(\gamma) \geq \left\| \tilde{P}(\theta_0) \right\|_2^2 \right) &\leq \frac{2LE_0 \int_{\mathcal{X}} \mathrm{Var}_{n,m}\left(\theta(x)\right) dx.}{\gamma \left\| \tilde{P}(\theta_0) \right\|_2^2} \\
&\lesssim \frac{\sigma^2 \sum_{j \in \mathbb{N}^d} \frac{m\mu_j}{\sigma^2 + n\mu_j}}{m \left\| \tilde{P}(\theta_0) \right\|_2^2} \\
&= \frac{\sigma^2 \sum_{j \in \mathbb{N}^d} \nu_j}{n \left\| \tilde{P}(\theta_0) \right\|_2^2} = o(1).
\end{aligned}
$$

§3.6 Proof of the Corollaries

§3.6.1 Proof of Corollary 3.2.3

First note that for any $\mathcal{N} \subset \mathbb{N}^d$

$$\begin{aligned}\left\|\tilde{P}(\theta_0)\right\|_2^2 &= \sum_{j\in\mathbb{N}^d}(1-\nu_j)^2\theta_{0,j}^2 = \sum_{j\in\mathbb{N}^d}\frac{\sigma^4}{(\sigma^2+n\mu_j)^2}\theta_{0,j}^2 \\ &\leq (n/\sigma^2)^{-2}\sum_{j\in\mathcal{N}}\frac{1}{\mu_j^2}\theta_{0,j}^2 + \sum_{j\in\mathbb{N}^d/\mathcal{N}}\theta_{0,j}^2. \end{aligned} \tag{3.6.1}$$

Consider eigenvalues satisfying (3.1.6) with $\alpha = \beta$, i.e. $\mu_j \asymp \left(\prod_{i=1}^d j_i\right)^{-2\beta/d-1}$. Let us take $\mathcal{N} = \{j \in \mathbb{N}^d : \Pi_{i=1}^d j_i \leq J_\beta\}$ with $J_\beta := (n/\sigma^2)^{d/(d+2\beta)}$ and note that in view of (3.7.6) [with $I = J_\beta$] we have

$$|\mathcal{N}| \lesssim J_\beta \log^{d-1} J_\beta = o(n) \tag{3.6.2}$$

Furthermore, we also get that

$$\begin{aligned}\sup_{\theta_0\in\Theta^\beta(B)}\left\|\tilde{P}(\theta_0)\right\|_2^2 &\lesssim \sup_{\theta_0\in\Theta^\beta(B)}\left[\frac{\sigma^4}{n^2}\max_{j\in\mathcal{N}}\left(\prod_{i=1}^d j_i\right)^{4\beta/d+2}\left(\sum_{i=1}^d j_i\right)^{-2\beta}\sum_{j\in\mathcal{N}}\left(\sum_{i=1}^d j_i\right)^{2\beta}\theta_{0,j}^2\right. \\ &\qquad \left. + \sup_{j\notin\mathcal{N}}\left(\sum_{i=1}^d j_k\right)^{-2\beta}\sum_{j\notin\mathcal{N}}\left(\sum_{i=1}^d j_i\right)^{2\beta}\theta_{0,j}^2\right] \\ &\lesssim (n/\sigma^2)^{-2}J_\beta^{2\beta/d+2}B^2 + J_\beta^{-2\beta/d}B^2 \\ &\lesssim (n/\sigma^2)^{-2\beta/(d+2\beta)},\end{aligned}$$

using Lemmas 3.7.10 [with $r = 4\beta/d+2$, $s = 2\beta$ and $J = J_\beta$] and 3.7.11 [with $s = 2\beta$ and $J = J_\beta$].

Moreover, in view of Lemma 3.7.5 and $\nu_j \leq 1$

$$\frac{\sigma^2}{n}\sum_{j\in\mathbb{N}^d}\nu_j^2 \lesssim \frac{\sigma^2}{n}J_\beta\log^{d-1}J_\beta = (n/\sigma^2)^{-2\beta/(d+2\beta)}\left(\log\left(\frac{n}{\sigma^2}\right)\right)^{d-1}.$$

Finally we show that the remaining term is $\delta_n = o(n^{-2\beta/(d+2\beta)})$ for the choice

$$\mathcal{I} = \left\{j \in \mathbb{N}^d : \prod_{i=1}^d j_i \leq I\right\} \qquad \text{with } I = \frac{n}{m^2\log^d(n/m)}\Big(\sum_{j\in\mathbb{N}^d}\nu_j^2\Big)^{-1},$$

where $\frac{n}{m^2\log^d(n/m)}\Big(\sum_{j\in\mathbb{N}^d}\nu_j^2\Big)^{-1} \geq 1$ holds because m is small enough. Note that in view of Lemma 3.7.8 the cardinality of $\mathcal{I}$ satisfies $|\mathcal{I}| \lesssim \frac{n}{m^2\log n}\Big(\sum_{j\in\mathbb{N}^d}\nu_j^2\Big)^{-1}$, hence

it satisfies the cardinality assumption on $\mathcal{I}$. Then in view of Lemma 3.7.8 and Lemma 3.7.5

$$\begin{aligned}\delta_n &\lesssim n \sum_{j\in\mathbb{N}^d} \nu_j^2 \sum_{\ell:\, \prod_{i=1}^d \ell_i > I} \mu_\ell \lesssim n \sum_{j\in\mathbb{N}^d} \nu_j^2 I^{-2\beta/d} \log^{d-1} I \\ &\ll n^{1-2\beta/d} m^{4\beta/d} \left(\sum_{j\in\mathbb{N}^d} \nu_j^2\right)^{2\beta/d+1} (\log n)^{2\beta+d-1} \\ &\lesssim n^{2-2\beta/d} m^{4\beta/d} (\log n)^{2\beta+d-1}.\end{aligned}$$

The right hand side is of order $o(n^{-2\beta/(2\beta+d)})$ for all $m = o(n^{1/2-3d/(4\beta)})$ with $\beta > 3d/2$. Combining the above inequality with Theorem 3.2.2 concludes the proof for the polynomially decaying eigenvalues.

§3.6.2 Proof of Corollary 3.2.4

For arbitrary index set $\mathcal{N} \subset \mathbb{N}^d$ we get that

$$\begin{aligned}\sup_{\theta_0\in\Theta^\beta(B)} \left\|\tilde{P}(\theta_0)\right\|_2^2 \leq \sup_{\theta_0\in\Theta^\beta(B)} \Bigg[&\frac{\sigma^4}{n^2} \max_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} e^{2a\sum_{i=1}^d j_i} \sum_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2 \\ &+ \sup_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} \sum_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2 \Bigg]. \qquad (3.6.3)\end{aligned}$$

We deal with the two terms on the right hand side separately. Note that the function $x \mapsto x^{-2\beta} e^{2ax}$ is convex on $[1, J_a]$, for $J_a = a^{-1}\log(n/\sigma^2)$ with $a \leq 1$, and achieves its maximum at one of the end points. Let us take the set $\mathcal{N} = \{j \in \mathbb{N}^d : \sum_{k=1}^d j_k \leq J_a\}$ and note that

$$|\mathcal{N}| \leq a^{-d} \log^d n = o(n), \qquad (3.6.4)$$

, by the lower bound on a. Furthermore, by noting that $(\sum_{i=1}^d j_i)^2 \leq d\sum_{i=1}^d j_i^2$, the maximum of the last display over $\mathcal{N}$ is bounded from above by

$$\max_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} e^{2a\sum_{i=1}^d j_i} \lesssim 1 + J_a^{-2\beta} e^{2aJ_a}.$$

The second term in (3.6.3) is directly bounded from above by $J_a^{-2\beta} B^2$. Therefore, by combining the inequalities above,

$$\left\|\tilde{P}(\theta_0)\right\|_2^2 \lesssim \frac{\sigma^4}{n^2} + \left(a^{-1}\log(n/\sigma^2)\right)^{-2\beta}. \qquad (3.6.5)$$

Moreover, in view of Lemma 3.7.6

$$\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j \asymp \frac{\sigma^2}{n} J_a^d = \frac{\sigma^2}{n} a^{-d} \log^d(n/\sigma^2). \qquad (3.6.6)$$

For $a := (n/\sigma^2)^{-1/(2\beta+d)} \log(n/\sigma^2)$ both of the preceding displays are bounded from above by a multiple of $(n/\sigma^2)^{-2\beta/(2\beta+d)}$.

Finally we show that the remainder term δ_n is of lower order than $(n/\sigma^2)^{-2\beta/(2\beta+d)}$. We take $\mathcal{I} = \{j \in \mathbb{N}^d : \sum_{i=1}^d j_i \leq I\}$, with $I = n^{1/d}\big(m^2 \log n \sum_{j\in\mathbb{N}^d} \nu_j^2\big)^{-1/d}$. Then it is easy to see that $|\mathcal{I}| \leq I^d \leq n\big(m^2 \log n \sum_{j\in\mathbb{N}^d} \nu_j^2\big)^{-1}$ holds. Note that $|\mathcal{I}| \geq 1$ holds because m is small enough. Furthermore, in view of the upper bound $p(j,d) \leq \frac{1}{2}\binom{j-1}{d-1} + 1/2 \leq j^d$ on the d partition of $j \in \mathbb{N}$, we get that

$$\begin{aligned}\delta_n &= n \sum_{j\in\mathbb{N}^d} \nu_j^2 \sum_{\ell\in\mathcal{I}^c} \mu_\ell \leq n \sum_{j\in\mathbb{N}^d} \nu_j^2 \sum_{\ell\geq I} \ell^d e^{-a\ell} \\ &\lesssim n I^d e^{-aI} \sum_{j\in\mathbb{N}^d} \nu_j^2 \lesssim (n/m)^2 e^{-aI} (\log n)^{-1} \end{aligned} \tag{3.6.7}$$

Since $\beta \geq d/2$, we have

$$\begin{aligned} aI &= (n/\sigma^2)^{-1/(2\beta+d)} \log(n/\sigma^2) n^{1/d} m^{-2/d} (\log n)^{-1/2} \left(\sum_{j\in\mathbb{N}^d} \nu_j^2\right)^{-1/d} \\ &\gtrsim n^{\frac{2\beta-d}{d(2\beta+d)}} m^{-2/d} (\log n)^{1-1/d} \geq L \log n. \end{aligned}$$

Hence the right hand side of (3.6.7) is $o(n^{-L})$, for arbitrary $L > 0$, when $m = o(n^{\frac{2\beta-d}{2(2\beta+d)}})$ concluding the proof of the corollary using Theorem 3.2.2.

§3.6.3 Proof of Corollary 3.3.2

We proceed by proving that the conditions of Theorem 3.3.1 hold for this choice of the kernel and the parameters, which directly provides us the statements.

Let us take $\mathcal{N} = \{j \in \mathbb{N}^d : \Pi_{k=1}^d j_k \leq J_\alpha\}$ with $J_\alpha := (n/\sigma^2)^{1/(2\alpha+d)}$ in (3.6.1). The cardinality of this set is $o(n)$, see (3.6.2). Furthermore, in view of $\alpha \leq \beta$,

$$\begin{aligned} \sup_{\theta_0\in\Theta^\beta(B)} \left\|\tilde{P}(\theta_0)\right\|_2^2 &\lesssim \sup_{\theta_0\in\Theta^\beta(B)} \left[\frac{\sigma^4}{n^2} \max_{j\in\mathcal{N}} \left(\prod_{i=1}^d j_i\right)^{4\alpha/d+2} \left(\sum_{i=1}^d j_i\right)^{-2\beta} \sum_{j\in\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2 \right. \\ &\qquad \left. + \sup_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{-2\beta} \sum_{j\notin\mathcal{N}} \left(\sum_{i=1}^d j_i\right)^{2\beta} \theta_{0,j}^2\right] \\ &\lesssim (n/\sigma^2)^{-2} J_\alpha^{4\alpha/d-2\beta/d+2} B^2 + J_\alpha^{-2\beta/d} B^2 \lesssim (n/\sigma^2)^{-2\beta/(2\alpha+d)}. \end{aligned}$$

Then, in view of Lemma 3.7.5, $\nu_j \leq 1$ and the preceding display,

$$\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j \asymp \frac{\sigma^2}{n} J_\alpha (\log J_\alpha)^{d-1} = (n/\sigma^2)^{-2\alpha/(2\alpha+d)} \left(\log(n/\sigma^2)\right)^{d-1} \gtrsim \sup_{\theta_0\in\Theta^\beta(B)} \left\|\tilde{P}(\theta_0)\right\|_2^2,$$

when $\alpha \leq \beta$. Finally in view of Corollary 3.2.3 we have that

$$\delta_n = o\left((n/\sigma^2)^{-2\alpha/(2\alpha+d)}\right) = o\left(\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j\right),$$

finishing the proof of the corollary.

§3.6.4 Proof of Corollary 3.3.3

We again prove that the conditions of Theorem 3.3.1 hold in this setting.

In view of assertions (3.6.5) and (3.6.6), we get for $a \lesssim (\sigma^2/n)^{1/(2\beta+d)} \log(n/\sigma^2)$ that

$$\left\|\tilde{P}(\theta_0)\right\|_2^2 \lesssim \frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j.$$

Furthermore, the cardinality of the set $\{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}$ is $o(n)$, see (3.6.4). Finally, in view of Corollary 3.2.4, $\delta_n = o(n^{-c})$, hence the condition $\delta_n = o\Big(\frac{\sigma^2}{n} \sum_{j\in\mathbb{N}^d} \nu_j\Big)$ of Theorem 3.3.1 also holds, concluding the proof.

§3.7 Technical lemmas

Lemma 3.7.1. *Consider the local regression problem* (3.1.1) *for arbitrary* $k \in \{1, ..., m\}$ *and let* $\vartheta \in L_2(\mathcal{X})$. *Then there exists a universal constant* C *not depending on* ϑ *such that*

$$E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \vartheta(X_i^{(k)}) K_{X_i^{(k)}} - E_X[\vartheta(x) K_x dx] \right) \right\|_2^2 \leq \frac{C}{n/m} \|\vartheta\|_2^2 \sum_{j\in\mathbb{N}^d} \nu_j^2, \tag{3.7.1}$$

where X *is a uniform random variable on* $\mathcal{X}$, *and* ν_j*'s are the eigenvalues of the operator* $\tilde{F}$.

Proof. For simplicity we omit the reference to the local k machine in the proof by writing $X_i = X_i^{(k)}$. Let $\vartheta = \sum_{j\in\mathbb{N}^d} \vartheta_j \psi_j \in L_2(\mathcal{X})$. Since

$$\vartheta(X) K_X = \sum_{j,k\in\mathbb{N}^d} \mu_j \vartheta_k \psi_j(X) \psi_k(X) \psi_j,$$

and $(\psi_j)_{j\in\mathbb{N}^d}$ is an orthonormal basis of $L_2(\mathcal{X})$, we have $E_X[\vartheta(X)K_X] = \sum_{j\in\mathbb{N}^d} \mu_j \vartheta_j \psi_j$. Furthermore, the linearity of the operator $\tilde{F} \circ F^{-1}$ implies that $\tilde{F} \circ F^{-1}(\vartheta(X)K_X) = \sum_{j,k\in\mathbb{N}^d} \nu_j \vartheta_k \psi_j(X) \psi_k(X) \psi_j$, providing

$$\begin{aligned}
\tilde{F} \circ F^{-1} \left(E_X[\vartheta(X)K_X]\right) &= \sum_{j\in\mathbb{N}^d} \nu_j \vartheta_j \psi_j, \\
\tilde{F} \circ F^{-1} \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \vartheta(X_i) K_{X_i} \right) &= \frac{1}{n/m} \sum_{i=1}^{n/m} \tilde{F} \circ F^{-1} \left(\vartheta(X_i) K_{X_i}\right) \\
&= \frac{1}{n/m} \sum_{i=1}^{n/m} \sum_{j,k\in\mathbb{N}^d} \nu_j \vartheta_k \psi_j(X_i) \psi_k(X_i) \psi_j. \qquad (3.7.2)
\end{aligned}$$

Then using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we get

$$
\begin{aligned}
&E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \vartheta(X_i) K_{X_i} - E_X[\vartheta(X) K_X] \right) \right\|_2^2 \\
&\quad = E_0 \left\| \sum_{j,k \in \mathbb{N}^d} \nu_j \vartheta_k \psi_j \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_k(X_i) - \delta_{jk} \right) \right\|_2^2 \\
&\quad = \sum_{j \in \mathbb{N}^d} \frac{\nu_j^2}{n/m} E_0 \left(\vartheta(X_i)\psi_j(X_i) - \vartheta_j\right)^2 \\
&\quad \leq 2 \sum_{j \in \mathbb{N}^d} \frac{\nu_j^2}{n/m} \left(E_0 \vartheta^2(X_i)\psi_j^2(X_i) + \vartheta_j^2 \right) \leq \frac{2(C_\psi^2+1)\|\vartheta\|_2^2}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2,
\end{aligned}
$$

finishing the proof of the statement. □

Lemma 3.7.2. *Consider the local regression problem* (3.2.1) *for arbitrary* $k \in \{1, ..., m\}$. *Then for any finite index set* $\mathcal{I} \subset \mathbb{N}^d$, $|\mathcal{I}| \leq n^C$ *and data dependent function* $\hat{\vartheta}^{(k)} : \mathcal{X}^{n/m} \mapsto \mathbb{R}$, $\|\hat{\vartheta}^{(k)}\|_2 \leq n^C$, *for some* $C > 0$,

$$
\begin{aligned}
&E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \hat{\vartheta}^{(k)}(X_i^{(k)}) K_{X_i^{(k)}} - E_X[\hat{\vartheta}^{(k)}(X) K_X] \right) \right\|_2^2 \\
&\lesssim \quad \frac{|\mathcal{I}| \log n}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 E_0 \left\| \hat{\vartheta}^{(k)} \right\|_2^2 + E_0 \left\| \hat{\vartheta}^{(k)}_{\mathcal{I}^c} \right\|_{\mathcal{H}}^2 \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell + n^{-C_0}, \qquad (3.7.3)
\end{aligned}
$$

where X *is a uniform random variable on* $\mathcal{X}$, ν_j*'s are the eigenvalues of the operator* $\tilde{F}$, C_0 *can be chosen arbitrarily large, and* $\hat{\vartheta}^{(k)}_{\mathcal{I}^c}(\cdot) = \sum_{j \in \mathcal{I}^c} \hat{\vartheta}^{(k)}_j \psi_j(\cdot)$.

Proof. For simplicity we omit the reference to the kth local problem and write $X_i = X_i^{(k)}$ and $\hat{\vartheta} = \hat{\vartheta}^{(k)}$. Let us next define the set $\mathcal{A}_{\mathcal{I},j} \subset \mathcal{X}^{n/m}$ as

$$
\mathcal{A}_{\mathcal{I},j} = \left\{ \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right)^2 \leq \frac{8C_\psi^2 C \log n}{n/m}, \quad \ell \in \mathcal{I} \right\}. \qquad (3.7.4)
$$

Note that by Hoeffding's inequality, for arbitrary $\ell \in \mathcal{I}$,

$$
\begin{aligned}
P\left(\mathcal{A}_{\mathcal{I},j}^c\right) &\leq |\mathcal{I}| P \left(\left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right)^2 > \frac{8C_\psi^2 C \log n}{n/m} \right) \\
&\leq 2|\mathcal{I}| \exp\left\{ -\frac{4C_\psi^2 C \log n}{C_\psi^2} \right\} \leq O(|\mathcal{I}| n^{-3C}).
\end{aligned}
$$

Then using $(a+b)^2 \le 2a^2 + 2b^2$ and Cauchy-Schwarz inequality

$$
\begin{aligned}
&E_0 \left\| \left(\tilde{F} \circ F^{-1}\right) \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \hat{\vartheta}(X_i) K_{X_i} - E_X[\hat{\vartheta}(X) K_X] \right) \right\|_2^2 \\
&\quad = E_0 \left\| \sum_{j \in \mathbb{N}^d} \sum_{\ell \in \mathbb{N}^d} \nu_j \hat{\vartheta}_\ell \psi_j \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right) \right\|_2^2 \\
&\quad \lesssim E_0 \sum_{j \in \mathbb{N}^d} \nu_j^2 \left(\sum_{\ell \in \mathcal{I}} \hat{\vartheta}_\ell \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right) \right)^2 \\
&\qquad + E_0 \sum_{j \in \mathbb{N}^d} \nu_j^2 \left(\sum_{\ell \in \mathcal{I}^c} |\hat{\vartheta}_\ell| (C_\psi^2 + 1) \right)^2 \\
&\quad \lesssim E_0 \sum_{j \in \mathbb{N}^d} \nu_j^2 |\mathcal{I}| \sum_{\ell \in \mathcal{I}} \hat{\vartheta}_\ell^2 \left(\frac{1}{n/m} \sum_{i=1}^{n/m} \psi_j(X_i)\psi_\ell(X_i) - \delta_{j\ell} \right)^2 \\
&\qquad + \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell E_0 \sum_{\ell \in \mathcal{I}^c} \hat{\vartheta}_\ell^2 \mu_\ell^{-1} \\
&\quad \le \sum_{j \in \mathbb{N}^d} \nu_j^2 E_0 \|\hat{\vartheta}\|_2^2 \left(\frac{8 C_\psi^2 C |\mathcal{I}| \log n}{n/m} + 1_{A_{j,\mathcal{I}}^c} |\mathcal{I}| \right) + E_0 \|\hat{\vartheta}_{\mathcal{I}^c}\|_{\mathcal{H}}^2 \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell \\
&\quad \lesssim \frac{|\mathcal{I}| \log n}{n/m} \sum_{j \in \mathbb{N}^d} \nu_j^2 E_0 \left\| \hat{\vartheta} \right\|_2^2 + E_0 \left\| \hat{\vartheta}_{\mathcal{I}^c} \right\|_{\mathcal{H}}^2 \sum_{j \in \mathbb{N}^d} \nu_j^2 \sum_{\ell \in \mathcal{I}^c} \mu_\ell + O(n^{-C}),
\end{aligned}
$$

where C can be chosen arbitrarily large, concluding the proof of our statement. □

Lemma 3.7.3. *There exists $C > 0$ such that*

$$
E_0 \|\hat{K}_{x,n}^{I,(k)}/m - \tilde{F}(K_x)\|_{\mathcal{H}}^2 \le C \sum_{j \in \mathbb{N}^d} \nu_j^2.
$$

Proof. First note that

$$
\|\hat{K}_{x,n}^{I,(k)}/m - \tilde{F}(K_x)\|_{\mathcal{H}}^2 \le 2m^{-2} \|\hat{K}_{x,n}^{I,(k)}\|_{\mathcal{H}}^2 + 2\|\tilde{F}(K_x)\|_{\mathcal{H}}^2.
$$

The second term on the right hand is bounded by

$$
\|\tilde{F}(K_x)\|_{\mathcal{H}}^2 = \sum_{j \in \mathbb{N}^d} \mu_j^{-1} \nu_j^2 \mu_j^2 \psi_j(x)^2 \le C_\psi^2 \sum_{j \in \mathbb{N}^d} \mu_j \nu_j^2 \lesssim \sum_{j \in \mathbb{N}^d} \nu_j^2.
$$

Since $\hat{K}_{x,n}^{I,(k)}$ is a KRR estimator, we get that

$$\begin{aligned} E_0\sigma^2\|\tilde{K}_{x,n}^{(k)}\|_{\mathcal{H}}^2 &\leq E_0\left(\sum_{i=1}^{n/m}(\tilde{K}_{x,n}^{(k)}(X_i^{(k)}) - K_x(X_i^{(k)}))^2 + \sigma^2\|\tilde{K}_{x,n}^{(k)}\|_{\mathcal{H}}^2\right) \\ &\leq E_0\left(\sum_{i=1}^{n/m}(\tilde{F}(K_x)(X_i^{(k)}) - K_x(X_i^{(k)}))^2 + \sigma^2\|\tilde{F}(K_x)\|_{\mathcal{H}}^2\right) \\ &\leq \sum_{i=1}^{n/m} E_0\tilde{P}(K_x)^2(X_i^{(k)}) + \sigma^2\|\tilde{F}(K_x)\|_{\mathcal{H}}^2 = O(\sum_{j\in\mathbb{N}^d}\nu_j^2), \end{aligned}$$

where the last inequality follows from (3.5.27). □

Lemma 3.7.4. *Assume that the eigenvalues μ_j of the covariance kernel K satisfy $\sum_{j\in\mathbb{N}^d}\mu_j < \infty$, $|\{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}| \leq n$, and $\sigma^2 \geq c > 0$. Then the expectation of the posterior variance is of the following order*

$$E_0E_X \operatorname{Var}(\theta(X)|\mathbb{D}_n) \asymp \sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j},$$

where the expectation E_X corresponds to the random variable $X \sim U[0,1]^d$ and the multiplicative constant depends on $\sum_{j\in\mathbb{N}^d}\mu_j$ and c.

Proof. It is shown in Section 6 of (Opper and Vivarelli, 1999) that the expectation of the posterior variance, named "generalization error", is bounded from below as follows

$$E_0E_X \operatorname{Var}(\theta(X)|\mathbb{D}_n) \geq \sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j E_X\psi_j^2(X)} = \sigma^2 \sum_{j\in\mathbb{N}^d} \frac{\mu_j}{\sigma^2 + n\mu_j}.$$

In (Ferrari-Trecate et al., 1998), it has been shown that for stationary GPs, for any $\mathcal{J} \subset \mathbb{N}^d$, with $|\mathcal{J}| \leq n$, the learning curve is bounded from above by

$$E_0E_X \operatorname{Var}(\theta(X)|\mathbb{D}_n) \leq \sum_{j\in\mathbb{N}^d}\mu_j - n\sum_{j\in\mathcal{J}}\frac{\mu_j^2}{c_j},$$

where

$$c_j = (n-1)\mu_j + \sigma^2 + \sum_{j\in\mathbb{N}^d}\mu_j.$$

Let us take $\mathcal{J} = \{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\}$ and by assumption its cardinality is

bounded by n. Then

$$\begin{aligned}\sum_{j\in\mathbb{N}^d}\mu_j - n\sum_{j\in\mathcal{J}}\frac{\mu_j^2}{c_j} &= \sum_{j\in\mathcal{J}}\mu_j\frac{c_j-n\mu_j}{c_j}+\sum_{j\notin\mathcal{J}}\mu_j\\ &= \sum_{j\in\mathcal{J}}\mu_j\frac{\sum_{j\in\mathbb{N}^d}\mu_j+\sigma^2-\mu_j}{\sum_{j\in\mathbb{N}^d}\mu_j+\sigma^2+(n-1)\mu_j}+\sum_{j\notin\mathcal{J}}\mu_j\\ &\le \sigma^2\sum_{j\in\mathcal{J}}\mu_j\frac{\sum_{j\in\mathbb{N}^d}\mu_j/\sigma^2+1}{\sigma^2+n\mu_j}+2\sigma^2\sum_{j\notin\mathcal{J}}\frac{\mu_j}{\sigma^2+n\mu_j}\\ &\lesssim \sigma^2\sum_{j\in\mathbb{N}^d}\frac{\mu_j}{\sigma^2+n\mu_j},\end{aligned}$$

concluding our proof. □

Lemma 3.7.5. *For ν_j, $j\in\mathbb{N}^d$, defined in* (3.2.5) *with eigenvalues μ_j polynomially decaying according to Assumption 3.1.2 and $k\in\mathbb{N}$,*

$$\sum_{j\in\mathbb{N}^d}\nu_j^k \asymp J_\alpha\log^{d-1}J_\alpha,$$

where $J_\alpha=(n/\sigma^2)^{d/(2\alpha+d)}$.

Proof. Let $\mathcal{N}:=\{j\in\mathbb{N}^d : n\mu_j\ge\sigma^2\}=\{j\in\mathbb{N}^d:\prod_{i=1}^d j_i\le CJ_\alpha\}$ and we apply Lemma 3.7.8 [with $\mathcal{I}$=$\mathcal{N}$, I=CJ_α and $\gamma=k(2\alpha/d+1)-1$]. First, we prove the upper bound,

$$\begin{aligned}\sum_{j\in\mathbb{N}^d}\nu_j^k &= \sum_{j\in\mathbb{N}^d}\frac{(n\mu_j)^k}{(\sigma^2+n\mu_j)^k}\\ &\le \sum_{j\in\mathcal{N}}1+(n/\sigma^2)^k\sum_{j\notin\mathcal{N}}\mu_j^k\\ &\lesssim J_\alpha\log^{d-1}J_\alpha+(n/\sigma^2)^kJ_\alpha^{-k(2\alpha/d+1)+1}\log^{d-1}J_\alpha\\ &\lesssim J_\alpha\log^{d-1}J_\alpha.\end{aligned}$$

The lower bound follows similarly,

$$\sum_{j\in\mathbb{N}^d}\nu_j^k\ge\Big(\frac{n}{2\sigma^2}\Big)^k\sum_{j\notin\mathcal{N}}\mu_j^k\gtrsim(n/\sigma^2)^kJ_\alpha^{-k(2\alpha/d+1)+1}\log^{d-1}J_\alpha\gtrsim J_\alpha\log^{d-1}J_\alpha.$$

□

Lemma 3.7.6. *For ν_j, $j\in\mathbb{N}^d$, defined in* (3.2.5) *with eigenvalues μ_j exponentially decaying according to Assumption 3.1.2 with $b=1$, $a<1$ and $k\in\mathbb{N}$,*

$$\sum_{j\in\mathbb{N}^d}\nu_j^k\asymp J_a^d,$$

where $J_a=a^{-1}\log(n/\sigma^2)$.

Proof. Let $\mathcal{N}_d := \{j \in \mathbb{N}^d : n\mu_j \geq \sigma^2\} = \{j \in \mathbb{N}^d : \sum_{i=1}^d j_i \leq J_a + c/a\}$ with $c > 0$ a positive constant. Then it is easy to see that $|\mathcal{N}_d| \leq 2^d J_a^d$. Moreover, we will show by induction on d that

$$\sum_{j \notin \mathcal{N}_d} e^{-ak\sum_{i=1}^d j_i} \lesssim a^{-d}(n/\sigma^2)^{-k} \log^{d-1}(n/\sigma^2).$$

Let us start with the case $d = 1$. We can directly see that

$$\sum_{j > J_a} e^{-akj} \leq C e^{-akJ_a} \frac{e^{ak}}{e^{ak} - 1} \lesssim a^{-1}(n/\sigma^2)^{-k}.$$

Now, assume that our assumption holds for d and consider the case $d + 1$, then

$$\begin{aligned}
\sum_{j \notin \mathcal{N}_{d+1}} e^{-ak\sum_{i=1}^{d+1} j_i} &\lesssim \sum_{j_{1:d} \in \mathbb{N}^d} e^{-ak\sum_{i=1}^d j_i} \sum_{j_{d+1} > \max(J_a - \sum_{i=1}^d j_i, 0)} e^{-akj_{d+1}} \\
&\lesssim \sum_{j_{1:d} \in \mathbb{N}^d} (e^{-ak\sum_{i=1}^d j_i} \wedge e^{-akJ_a}) \frac{e^{ak}}{e^{ak} - 1} \\
&\lesssim \sum_{j_{1:d} \in \mathcal{N}_d} a^{-1} e^{-akJ_a} + \sum_{j_{1:d} \notin \mathcal{N}_d} a^{-1} e^{-ak\sum_{i=1}^d j_i} \\
&\lesssim a^{-1}|\mathcal{N}_d|(n/\sigma^2)^{-k} + a^{-d-1}(n/\sigma^2)^{-k} \log^{d-1}(n/\sigma^2) \\
&\lesssim a^{-d-1}(n/\sigma^2)^{-k} \log^d(n/\sigma^2),
\end{aligned}$$

which concludes the induction proof.

Using these two results, we can easily show that

$$\sum_{j \in \mathbb{N}^d} \nu_j^k \lesssim \sum_{j \in \mathcal{N}_d} 1 + (n/\sigma^2)^k \sum_{j \notin \mathcal{N}_d} e^{-ak\sum_{i=1}^d j_i} \lesssim |\mathcal{N}_d| + a^{-d} \log^{d-1}(n/\sigma^2) \lesssim J_a^d.$$

On the other hand, we can show by induction that for all $J > d$, the cardinality of $\mathcal{N}_d := \{j \in \mathbb{N}^d : \sum_{i=1}^d j_i \leq J\}$ is bounded from below as follows

$$|\mathcal{N}_d| \geq (J - d)^d / d!.$$

Note that it holds trivially for $d = 1$. Now assume it holds for d, then we can write $\mathcal{N}_{d+1}$ as a partition as follows

$$\mathcal{N}_{d+1} = \left\{ j \in \mathbb{N}^{d+1} : \sum_{k=1}^{d+1} j_k \leq J \right\} = \bigcup_{i=1}^{J-d} \left\{ j \in \mathbb{N}^{d+1} : j_{d+1} = i; \sum_{k=1}^{d} j_k \leq J - i \right\}.$$

According to our induction assumption, the cardinality of all these subsets are bounded from below by $(J - d - i)^d / d!$, hence we have

$$|\mathcal{N}_{d+1}| \geq \sum_{i=1}^{J-d} \frac{(J - d - i)^d}{d!} \geq \int_1^{J-d} \frac{(J - d - t)^d}{d!} dt = \frac{(J - d - 1)^{d+1}}{(d+1)!},$$

which concludes our induction proof. Using this result, we can now show that

$$\sum_{j\in\mathbb{N}^d} \nu_j^k \geq \sum_{j\in\mathcal{N}_d} 1 = |\mathcal{N}_d| \gtrsim J_a^d,$$

concluding the proof. □

Lemma 3.7.7. *For arbitrary $\theta_0 \in \ell_2(L)$ we get that*

$$E_0\|\Delta\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \leq Cn,$$

for some universal constant $C > 0$.

Proof. First note that

$$\|\Delta\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \leq 2\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 + 2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2.$$

For $\theta_0 \in \ell_2(L)$ the second term on the right hand is bounded by

$$\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2 = \sum_{j\in\mathbb{N}^d} \mu_j^{-1}\nu_j^2\theta_{0,j}^2 \leq \sum_{j\in\mathbb{N}^d} \frac{n^2\mu_j}{(\sigma^2+n\mu_j)^2}\theta_{0,j}^2 \leq nL^2/\sigma^2.$$

Then by the definition of $\hat{\theta}_n^{(k)}$ we get that

$$\begin{aligned}
\sigma^2\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 &\leq \sum_{i=1}^{n/m}\left(\hat{\theta}_n^{(k)}(X_i^{(k)}) - Y_i^{(k)}\right)^2 + \sigma^2\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \\
&\leq \left(\sum_{i=1}^{n/m}\left(\tilde{F}(\theta_0)(X_i^{(k)}) - \theta_0(X_i^{(k)}) - \varepsilon_i^{(k)}\right)^2 + \sigma^2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2\right) \\
&\leq 2\sum_{i=1}^{n/m}\tilde{P}(\theta_0)^2(X_i^{(k)}) + 2\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 + \sigma^2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2. \qquad (3.7.5)
\end{aligned}$$

We conclude the proof by taking the expectation of both sides

$$\sigma^2 E_0\|\hat{\theta}_n^{(k)}\|_{\mathcal{H}}^2 \lesssim \sum_{i=1}^{n/m} E_0\tilde{P}(\theta_0)^2(X_i^{(k)}) + \sum_{i=1}^{n/m} E_0(\varepsilon_i^{(k)})^2 + \sigma^2\|\tilde{F}(\theta_0)\|_{\mathcal{H}}^2 = O(n).$$

□

Lemma 3.7.8. *The cardinality of the set*

$$\mathcal{I}_{I,d} = \left\{ j \in \mathbb{N}_+^d : \prod_{i=1}^d j_i \leq I \right\} \qquad (3.7.6)$$

satisfies that $|\mathcal{I}_{I,d}| \leq 2^d I \log^{d-1} I$. Furthermore,

$$\sum_{j\in\mathcal{I}_{I,d}^c}\prod_{i=1}^d j_i^{-\gamma-1} \asymp I^{-\gamma}\left(\log I\right)^{d-1}, \qquad (3.7.7)$$

for some large enough constant $C_{\gamma,d}$.

Proof. We prove both statements by induction, starting with the first one. For $d = 1$ it is trivial. Let us assume that it holds for d and consider the case $d + 1$. We distinguish cases according the value of j_{d+1}. If $j_{d+1} = 1$, then $\prod_{i=1}^{d} j_i \leq I$ holds, if $j_{d+1} = 2$, then $\prod_{i=1}^{d} j_i \leq I/2$ holds, and so on. Hence we can write that

$$|\mathcal{I}_{I,d+1}| \leq \sum_{j_{d+1}=1}^{I} \left|\mathcal{I}_{I/j_{d+1},d+1}\right| \leq 2^d \sum_{j_{d+1}=1}^{I} \frac{I}{j_{d+1}} \log^{d-1} \frac{I}{j_{d+1}} < 2^{d+1} I \log^d I,$$

where in the last inequality we have used that $\sum_{i=1}^{n} 1/i < 1 + \log n < 2 \log n$.

Note again that for $d = 1$ the second statement holds trivially (using Riemann sums for instance). Then assume that it holds for d and consider the case $d+1$. First we deal with the upper bound, where we note that

$$\begin{aligned}
\sum_{j \in \mathcal{I}^c_{I,d+1}} \prod_{i=1}^{d+1} j_i^{-\gamma-1} &= \sum_{j_{d+1}=1}^{I} j_{d+1}^{-\gamma-1} \sum_{j \in \mathcal{I}^c_{I/j_{d+1},d}} \prod_{i=1}^{d} j_i^{-\gamma-1} \\
&\quad + \sum_{j_{d+1}=I}^{\infty} j_{d+1}^{-\gamma-1} \prod_{k=1}^{d} \sum_{j_k=1}^{\infty} j_k^{-\gamma-1} \\
&\lesssim \sum_{j_{d+1}=1}^{I} \frac{1}{j_{d+1}} I^{-\gamma} \left(\log I / j_{d+1}\right)^{d-1} + \sum_{j_{d+1}=I}^{\infty} j_{d+1}^{-\gamma-1} \\
&\leq I^{-\gamma} \log^{d-1} I \sum_{j_{d+1}=1}^{I} \frac{1}{j_{d+1}} + I^{-\gamma} \leq I^{-\gamma} \log^d I.
\end{aligned}$$

Finally, it remained to deal with the lower bound. First, note that it is sufficient to show the result for $I \geq C$, for some C large enough (depending only on d and γ). Then by noting that for $x \geq e^{d-1}$ the function $x \mapsto x^{-1} \log^{d-1} x$ is monotone decreasing, we get that

$$\begin{aligned}
\sum_{j \in \mathcal{I}^c_{I,d+1}} \prod_{i=1}^{d+1} j_i^{-\gamma-1} &\geq \sum_{j_{d+1}=1}^{I} j_{d+1}^{-\gamma-1} \sum_{j \in \mathcal{I}^c_{I/j_{d+1},d}} \prod_{i=1}^{d} j_i^{-\gamma-1} \\
&\gtrsim I^{-\gamma} \left(\sum_{j_{d+1}=1}^{I} j_{d+1}^{-1} \log^{d-1} I - \sum_{j_{d+1}=1}^{I} j_{d+1}^{-1} \log^{d-1} j_{d+1} \right) \\
&\geq I^{-\gamma} \Big(\log^{d-1} I \int_1^I x^{-1} dx - \sum_{j_{d+1}=1}^{e^{d-1}} j_{d+1}^{-1} \log^{d-1} j_{d+1} \\
&\qquad - \int_{e^{d-1}}^{I} x^{-1} \log^{d-1} x dx \Big) \\
&\geq I^{-\gamma} \left(\log^d I - C_{d,\gamma} - \log^d I/2) \right) \gtrsim I^{-\gamma} \log^d I,
\end{aligned}$$

concluding the proof of our statement. □

Lemma 3.7.9. *There exists an event $A_n^{(k)}$ such that for any $\theta_0 \in L_\infty(L)$ and $n \le (n/m)^{C_1}$, for some $C_1 \ge 1$ there exist constants $C_2, C_3 > 0$ such that*

$$\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2 1_{A_n^{(k)}} \le (n/m)^{C_2},$$
$$E_{\theta_0}\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 1_{(A_n^{(k)})^c} = e^{-C_3 n/m}.$$

Proof. Let us take $A_n^{(k)} = \{\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 \le (n/m)^{C_0}\}$, for arbitrary $C_0 > 1$. Then in view of (3.7.5) we have on the event $A_n^{(k)}$ that

$$\left\|\Delta\hat{\theta}_n^{(k)}\right\|_2 \le \left\|\hat{\theta}_n^{(k)}\right\|_2 + \|\tilde{F}(\theta_0)\|_2 \lesssim n^{1/2} + (n/m)^{C_0} + L \lesssim (n/m)^{C_0\vee C_1/2}.$$

Furthermore, note that

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

Furthermore from the definition of $\hat{S}_n^{(k)}$, the boundedness of $\mathcal{X}$ and $\|K\|_\infty = O(1)$ we get that

$$\begin{aligned}\left\|\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 \le \left\|\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_\infty^2 &\lesssim \left(\frac{1}{n/m}\sum_{i=1}^{n/m}|\varepsilon_i^{(k)}|\right)^2 + \|\theta_0\|_\infty^2 \\ &\lesssim \frac{1}{n/m}\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 + 1.\end{aligned}$$

Since $W_n = (n/m)^{-1}\sum_{i=1}^{n/m}(\varepsilon_i^{(k)})^2 \sim \chi^2_{n/m}$, note that for n/m large enough

$$\begin{aligned}&E_{\theta_0}\left\|\Delta\hat{\theta}_n^{(k)} - \tilde{F}\circ F^{-1}\circ\hat{S}_n^{(k)}(\tilde{F}(\theta_0))\right\|_2^2 1_{(A_n^{(k)})^c} \\ &\quad\lesssim E_{\theta_0}\left((n/m)^{2C_1}W_n + (n/m)^{2C_1}\right)1_{W_n > (n/m)^{C_0}} = O(e^{-n/m}),\end{aligned}$$

concluding the proof of the lemma. □

Lemma 3.7.10. *Let $r, s > 0$ such that $r > s/d$ and $f : [1,\infty)^d \to \mathbb{R}$ defined as*

$$f(x) = \left(\prod_{i=1}^d x_i\right)^r \left(\sum_{i=1}^d x_i\right)^{-s}.$$

Then f is bounded from above by $d^{-s}J^{r-s/d}$ on the set $\mathcal{N} := \{x \in [1,\infty)^d : \prod_{i=1}^d x_i \le J\}$ with $J > 1$.

Proof. From the inequality of arithmetic and geometric means, we know that for all $x \in [1, \infty)^d$

$$\sum_{i=1}^{d} x_i \geq d \left(\prod_{i=1}^{d} x_i \right)^{1/d}.$$

Thus, we can bound f from above by

$$f(x) \leq d^{-s} \left(\prod_{i=1}^{d} x_i \right)^{r-s/d} \leq d^{-s} J^{r-s/d},$$

on $\mathcal{N}$ concluding the proof. □

Lemma 3.7.11. *Let $s > 0$ and $f : [1, \infty)^d \to \mathbb{R}$ defined as*

$$f(x) = (\sum_{i=1}^{d} x_i)^{-s}.$$

Then f is bounded from above by $d^{-s} J^{-s/d}$ on the set $\mathcal{N} := \{x \in [1, \infty)^d : \prod_{i=1}^{d} x_i \geq J\}$ with $J > 1$.

Proof. Since f is differentiable on its domain, we can compute its gradient

$$(\nabla f)_\ell = -s(\sum_{k=1}^{d} x_k)^{-s-1} < 0,$$

for all $\ell \in \{1, ..., d\}$. Thus, the function attains its maximum at $\prod_{i=1}^{d} x_i = J$. At the maximum point, in view of the inequality of arithmetic and geometric means, $\sum_{i=1}^{d} x_i \geq d \left(\prod_{i=1}^{d} x_i \right)^{1/d} = dJ^{1/d}$. The statement of the lemma follows by raising both sides to the $-s$ power. □