JMLR: Workshop and Conference Proceedings vol 23 (2012) 8.1–8.20 25th Annual Conference on Learning Theory

PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model

Taiji Suzuki S-TAIJI@STAT.T.U-TOKYO.AC.JP
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We develop a PAC-Bayesian bound for the convergence rate of a Bayesian variant of Multiple Kernel Learning (MKL) that is an estimation method for the sparse additive model. Standard analyses for MKL require a strong condition on the design analogous to the restricted eigenvalue condition for the analysis of Lasso and Dantzig selector. In this paper, we apply PAC-Bayesian technique to show that the Bayesian variant of MKL achieves the optimal convergence rate *without* such strong conditions on the design. Basically our approach is a combination of PAC-Bayes and recently developed theories of non-parametric Gaussian process regressions. Our bound is developed in a fixed design situation. Our analysis includes the existing result of Gaussian process as a special case and the proof is much simpler by virtue of PAC-Bayesian technique. We also give the convergence rate of the Bayesian variant of Group Lasso as a finite dimensional special case.

Keywords: PAC-Bayes, Multiple Kernel Learning, Group Lasso, Gaussian Process, Sparse Learning, Additive Model

1. Introduction

Sparse additive modeling is a powerful technique for nonparametric regression in high dimensional data (Ravikumar et al., 2009; Raskutti et al., 2012; Hastie and Tibshirani, 1999). In the past decade, a great amount of studies have been devoted to sparse statistical models. Sparsity gives a nice interpretation of the estimated results and enables statisticians to develop methodologies that yield reasonable performances even for high dimensional data. Although a linear high dimensional modeling has attracted much attentions, there has been also attempts to develop a nonparametric method to achieve more flexible data analysis in high dimensional data. One possible way is to just fit a nonparametric function $f(x)$ to the full input space, but that suffers the curse of dimensionality. To avoid this problem, sparse additive model splits the input data x into M subsets $(x^{(1)}, \dots, x^{(M)})$ and fits the sum of functions $f_m(x^{(m)})$ to the data, $y = \sum_{m=1}^{M} f_m(x^{(m)}) + \xi$, and imposes a sparsity on the set of functions $\{f_m\}_{m=1}^{M}$, that is, only a few components $\{f_m\}_{m \in I_0}$ are meaningful and other components are zero or negligibly small. This is more restrictive than the direct nonparametric fitting using the full input space, but the result is more interpretable and, more importantly, over-fitting can be avoided. One sophisticated approach to estimate the sparse additive model is *Multiple Kernel Learning* (MKL, Lanckriet et al. (2004)). MKL was first developed as a method to "learn a kernel", but afterward Bach et al. (2004) pointed out that MKL can be interpreted as a method to learn a sparse additive model. MKL approximates each component f_m by an element of *Reproducing Kernel Hilbert Space* (RKHS), and imposes L_1-mixed-norm regularization to yield sparsity.

Our main interest in this paper is to theoretically investigate a Bayesian variant of MKL that is a mixture of Bayesian sparse learning and *Gaussian process* estimation. The Gaussian process modeling is a Bayesian alternation of the kernel-based learning (Gibbs, 1997; Seeger, 2004; Rasmussen and Williams, 2006). That has shown nice performances as a non-parametric regression and classification method. It is a natural strategy to apply the Gaussian process modeling to sparse additive model where each component f_m is estimated by the Gaussian process method. Indeed, Gaussian process formulations of the multiple kernel learning framework have been proposed by some authors (Archambeau and Bach, 2010; Tomioka and Suzuki, 2010). In this paper, we analyze a rather different method from those existing ones.

Our theoretical framework is based on the *PAC-Bayesian* technique (McAllester, 1998, 1999; Catoni, 2004). The first PAC-Bayesian bound proposed by McAllester (1998, 1999) was a data-dependent empirical inequality for Bayesian estimators. Afterward Catoni (2004) proposed to utilize the PAC-Bayesian technique to establish sharp oracle inequalities. Recently it has been shown that the PAC-Bayesian technique is quite useful to investigate the statistical convergence rates of Bayesian sparse learning methods. One remarkable insights obtained by PAC-Bayesian bounds for Bayesian sparse learning methods is that no assumption on the condition of design is needed (Dalalyan and Tsybakov, 2008; Alquier and Lounici, 2011; Rigollet and Tsybakov, 2011b). In the theoretical analysis of regularized empirical risk minimization methods such as Lasso and Dantzig selector, we usually assume a strict condition on the design such as restricted eigenvalue condition (see Bickel et al. (2009) and the references therein). On the other hand, through the PAC-Bayesian technique, it has been shown that Bayesian sparse estimation methods achieve the optimal learning rate *without* such a strong condition.

As for theories of Gaussian process modeling, substantial developments have been made recently (van der Vaart and van Zanten, 2008a,b, 2011). van der Vaart and van Zanten (2011) investigated the convergence rate of Gaussian process estimators, and discussed how the estimator behaves according to the geometric relation between the true function and the RKHS corresponding to the Gaussian process prior. Our concern is that they investigated only restricted situations such as Sobolev and Hölder classes.

In this paper, we theoretically investigate a Bayesian variant of MKL, called *Bayesian-MKL*, where each component f_m is modeled by a Gaussian process prior. Our contributions are (i) to develop a PAC-Bayesian bound for Gaussian process regressions, and (ii) to derive the convergence rate of Bayesian-MKL in sparse additive model. More detailed description of our contribution is as follows.

(i) We develop a new PAC-Bayesian oracle inequality for Gaussian process regressions in fixed design situations. Thanks to the PAC-Bayesian technique, we obtain a simple proof of the convergence rate. In our analysis, we relax the normality on the noise unlike the existing researches. Moreover our PAC-Bayesian technique enables us to analyze general classes of model spaces utilizing the notion of interpolation spaces and the metric entropy, while the existing researches are based on the properties specialized to Sobolev and Hölder classes. Moreover, we show that, by putting a prior on the scale of Gaussian process, the estimator possesses adaptivity for the smoothness of the true function in a similar spirit to van der Vaart and van Zanten (2009).

(ii) The convergence rate of Bayesian-MKL is established. Thanks to PAC-Bayesian technique, our convergence analysis does not require any conditions on the design analogous to the re-

stricted eigenvalue condition, while conventional convergence analyses of MKL required that kind of strong assumptions those are sometimes unrealistic (Meier et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Suzuki and Sugiyama, 2012). Moreover our analysis covers the situations where the true function is not contained in the corresponding RKHS.

2. Preliminary

Here we formulate the problem setting and introduce the Bayesian variant of MKL.

2.1. Problem Settings

Suppose we are given n sample input-output pairs $\{(x_i, y_i)\}_{i=1}^n$ generated from the following regression model:

$$y_i = f^{\mathrm{o}}(x_i) + \xi_i, \quad (i = 1, \ldots, n),$$

where $\{x_i\}_{i=1}^n$ are given non-random elements[1] of a set $\mathcal{X}$, $\{\xi_i\}_{i=1}^n$ are i.i.d. zero-mean random variables, and f^{o} is the unknown true function satisfying $f^{\mathrm{o}}(X) = \mathrm{E}[Y|X]$.

In this article, we consider the situation where $\mathcal{X}$ is decomposed into M spaces $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$ and f^{o} is well approximated by a function f^* that can be decomposed into M functions each of which is defined on $\mathcal{X}_m$ $(m = 1, \ldots, M)$, i.e., $f^*(x) = \sum_{m=1}^M f_m^*(x^{(m)})$ where $f_m^* : \mathcal{X}_m \to \mathbb{R}$ and $x = (x^{(1)}, \ldots, x^{(M)}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$. Basically we suppose that f^* is "sparse" in a sense that the number of non-zero components $I_0 := \{m \mid f_m^* \neq 0\}$ is small compared with M. We want to estimate the function f^{o} so that the empirical L_2-norm is minimized:

$$\|f - f^{\mathrm{o}}\|_n^2 := \frac{1}{n}\sum_{i=1}^n (f(x_i) - f^{\mathrm{o}}(x_i))^2.$$

We also define the inner product with respect to the empirical L_2-norm as $\langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i)$. Our strategy is a Bayesian approach where a Gaussian process prior is employed for each component f_m^*. To estimate a sparse model, we put a prior of exponential weight on the number of components to be used. Let $f = (f_1, \ldots, f_M)$ be a concatenation of continuous functions $f_1, \ldots, f_M$ each of which is defined on $\mathcal{X}_m$, then we consider the following prior distribution on the product space $\mathrm{d}f = (\mathrm{d}f_1, \ldots, \mathrm{d}f_M)$:

$$\Pi(\mathrm{d}f) = \sum_{J \in \mathcal{P}(\{1,\ldots,M\})} \pi_J \cdot \prod_{m \in J} \int_{\lambda_m \in \mathbb{R}_+} \mathrm{GP}_m(\mathrm{d}f_m|\lambda_m)\mathcal{G}(\mathrm{d}\lambda_m) \cdot \prod_{m \notin J} \delta_0(\mathrm{d}f_m), \tag{1}$$

where $\mathcal{P}(\{1, \ldots, M\})$ is the set of all subsets of $\{1, \ldots, M\}$ and $\delta_0(\mathrm{d}f_m)$ is the Dirac measure having all its mass at $f_m = 0$; $\{\pi_J\}_{J \in \mathcal{P}(\{1,\ldots,M\})}$ is the exponential weight prior on the model that is given as, for a fixed $\zeta \in (0, 1)$,

$$\pi_J = \frac{\zeta^{|J|}}{\sum_{j=0}^M \zeta^j} \binom{M}{|J|}^{-1},$$

for all $J \in \mathcal{P}(\{1, \ldots, M\})$ (this choice of π_J is suggested by Alquier and Lounici (2011)); $\mathcal{G}(\mathrm{d}\lambda_m)$ is the exponential distribution, $\mathcal{G}(\mathrm{d}\lambda_m) = \exp(-\lambda_m)\mathrm{d}\lambda_m$, that is a conjugate prior for the scale of Gaussian process priors; $\mathrm{GP}_m(\mathrm{d}f|\lambda_m)$ is the Gaussian process prior with scale λ_m that will be defined in the successive subsection.

1. In this paper, we deal with a fixed design situation, i.e., $\{x_i\}_{i=1}^n$ are fixed and non-random.

2.2. Gaussian Process Prior and Corresponding RKHS

We put a zero-mean Gaussian process prior GP_m with a kernel k_m to estimate the function f_m^* on the m-th space $\mathcal{X}_m$. A zero-mean Gaussian process $W = (W_x : x \in \mathcal{X}_m)$ on the input space $\mathcal{X}_m$ is a set of random variable W_x indexed by $\mathcal{X}_m$ and defined on a common probability space $(\Omega_m, \mathcal{U}_m, P_m)$ such that each finite subset $(W_{x_1}, \ldots, W_{x_j})$ $(j = 1, 2, \ldots)$ possesses a zero-mean multivariate normal distribution. We assume that every sample path is bounded $\sup_{x \in \mathcal{X}_m} |W_x| < \infty$, which induces a map $W : \Omega_m \to L_\infty(\mathcal{X}_m)$. Moreover we assume that the map $W : \Omega_m \to L_\infty(\mathcal{X}_m)$ is tight and Borel measurable, that is true if there exits a semi-metric ρ_m on $\mathcal{X}_m$ such that $(\mathcal{X}_m, \rho_m)$ is totally bounded and almost all paths $x \mapsto W_x$ are uniformly ρ-continuous (see Section 1.5 of van der Vaart and Wellner (1996) for the characterization of measurability and tightness). The kernel function $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ corresponding to GP_m is the covariance function defined by

$$k_m(x, x') := \mathrm{E}[W_x W_{x'}].$$

The kernel function completely determines the finite dimensional distribution of the process. Corresponding to the kernel function k_m, we can define the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_m$ as a completion of the linear space spanned by all functions

$$z \mapsto \textstyle\sum_{i=1}^I \alpha_i k_m(z_i, z), \quad (\alpha_1, \ldots, \alpha_I \in \mathbb{R},\ z_1, \ldots, z_I \in \mathcal{X}_m,\ I \in \mathbb{N}),$$

relative to the RKHS norm $\|\cdot\|_{\mathcal{H}_m}$ induced by the inner product

$$\left\langle \textstyle\sum_{i=1}^I \alpha_i k_m(z_i, \cdot), \sum_{j=1}^J \alpha'_j k_m(z'_j, \cdot) \right\rangle_{\mathcal{H}_m} = \textstyle\sum_{i=1}^I \sum_{j=1}^J \alpha_i \alpha'_j k_m(z_i, z'_j). \tag{2}$$

For each element f of $\mathcal{H}_m$, the "function value" at the point $x \in \mathcal{X}_m$ can be recovered by the following reproducing formula:

$$f(x) = \langle f, k_m(\cdot, x) \rangle_{\mathcal{H}_m}.$$

One can show that this reproducing formula is well defined through the completion operation, and compatible with the definition of the inner product Eq. (2). More detailed discussions about the definition of the RKHS attached with the Gaussian process can be found in van der Vaart and van Zanten (2008b).

It is known that the RKHS $\mathcal{H}_m$ is usually much "smaller" than the support of the Gaussian process in an infinite dimensional setting. In fact, typically the prior has probability mass 0 on the infinite dimensional RKHS $\mathcal{H}_m$. That leads to the fact that, under the assumption $f_m^* \in \mathcal{H}_m$, estimating the function f_m^* through the standard Bayesian procedure with Gaussian process prior never achieves the optimal rate in some important examples (van der Vaart and van Zanten, 2011). To overcome this issue, we scale the process by the factor of λ_m and make the estimator close to the small space $\mathcal{H}_m$. The Gaussian process prior $\mathrm{GP}_m(\cdot|\lambda_m)$ with the scale parameter λ_m is the process with the kernel function $\tilde{k}_{m,\lambda_m} = k_m/\lambda_m$. Let $\mathcal{H}_{m,\lambda_m}$ be the RKHS corresponding to $\tilde{k}_{m,\lambda_m}$. Then $f \in \mathcal{H}_m$ can be embedded in $\mathcal{H}_{m,\lambda_m}$, and we have

$$\sqrt{\lambda_m} \|f\|_{\mathcal{H}_m} = \|f\|_{\mathcal{H}_{m,\lambda_m}}.$$

This indicates that with large λ_m the prior $\mathrm{GP}_m(\cdot|\lambda_m)$ imposes a strong regularization, and hence the Bayesian estimator associated with $\mathrm{GP}_m(\cdot|\lambda_m)$ is forced to be concentrated around $\mathcal{H}_m$. To choose the scale parameter λ_m optimally, we put a prior distribution of the exponential distribution $\mathcal{G}(\mathrm{d}\lambda_m)$ for λ_m that is conjugate for the scale of Gaussian process priors.

Example 1 (Matérn Priors) *An important class of Gaussian process priors for smooth functions, such as elements in Sobolev class, is the Matérn priors. Suppose that $\mathcal{X}_m = [0,1]^d$. The Matérn priors on $\mathcal{X}_m$ correspond to the kernel function defined as*

$$k_m(z, z') = \int_{\mathbb{R}^d} e^{is^\top (z-z')} \psi(s) \mathrm{d}s,$$

where $\psi(s)$ is the spectral density given by $\psi(s) = (1+\|s\|^2)^{-(\alpha+d/2)}$, for a smoothness parameter $\alpha > 0$. It is known that the RKHS $\mathcal{H}_m$ corresponding to the Matérn prior is contained in the Sobolev space ($W^{\alpha+d/2}[0,1]^d$) of order $\alpha + d/2$. Moreover, the Bayesian estimator with the Matérn prior yields the optimal rates $n^{-\frac{2\alpha}{2\alpha+d}}$ to estimate a function f_m^* in $C^\alpha[0,1]^d \cap W^\alpha[0,1]^d$ of smoothness order α (van der Vaart and van Zanten, 2011)[2]. Note that, although $f_m^* \in C^\alpha[0,1]^d \cap W^\alpha[0,1]^d$ is not necessarily contained in $W^{\alpha+d/2}[0,1]^d$ (thus is not contained in $\mathcal{H}_m$), the optimal rate is achieved. That means the support of the Matérn prior is much larger than $\mathcal{H}_m$. On the other hand, if $f_m^* \in \mathcal{H}_m$, the optimal rate is never achieved with fixed scale λ_m (van der Vaart and van Zanten, 2011).

2.3. Bayesian Multiple Kernel Learning

Based on the prior introduced in Eq. (1), we construct the "posterior distribution" and the corresponding Bayesian estimator. Let $D_n := (y_1, \ldots, y_n)$. For some constant $\beta > 0$, the posterior probability measure is given as

$$\Pi(\mathrm{d}f|D_n) := \frac{\exp(-\sum_{i=1}^n (y_i - \sum_{m=1}^M f_m(x_i))^2/\beta)}{\int \exp(-\sum_{i=1}^n (y_i - \sum_{m=1}^M \tilde{f}_m(x_i))^2/\beta)\Pi(\mathrm{d}\tilde{f})} \Pi(\mathrm{d}f),$$

for $f = (f_1, \ldots, f_M)$. Corresponding to the posterior, we have the Bayesian estimator $\hat{f}$, say Bayesian-MKL estimator, as the expectation of the posterior:

$$\hat{f} = \int \sum_{m=1}^M f_m \Pi(\mathrm{d}f|D_n).$$

In this paper, we do not pursue the computational aspects of Bayesian-MKL. The Bayesian-MKL estimator is quite computation demanding because it requires summation over all subsets of the index set. However one can utilize an efficient MCMC type method (Marin and Robert, 2007) for this kind of mixture models. In fact, Green (1995) suggested Reversible Jump MCMC method to compute the posterior distribution that possesses mass on several models of different dimensions, and, in the PAC-Bayesian contexts, Dalalyan and Tsybakov (2011) and Alquier and Biau (2011) investigated practical implementations of MCMC for sparse estimation problems.

3. Noise Assumption and PAC-Bayesian Bound

Here we give an assumption on the noise ξ_i to obtain a PAC-Bayesian bound. There are a lot of choices of noise conditions to establish PAC-Bayesian bounds. Here we employ a condition with

2. $C^\alpha[0,1]^d$ denotes the Hölder space of smoothness order α (see Section 2.7.1 of van der Vaart and Wellner (1996) for the definition).

which we can utilize an extension of Stein's identity. Now define a function

$$m_\xi(z) := -\mathrm{E}[\xi_1 \mathbf{1}\{\xi_1 \le z\}] = -\int_{-\infty}^{z} y \mathrm{d}F_\xi(y) = \int_z^\infty y \mathrm{d}F_\xi(y),$$

where $F_\xi(z) = P(\xi_1 \le z)$ is the cumulative distribution function of the noise, and $\mathbf{1}\{\cdot\}$ is the indicator function. Since $\mathrm{E}[\xi_1] = 0$, one can check that $m_\xi(z)$ is non-negative and achieves its maximum at 0: $\max_{z\in\mathbb{R}} m_\xi(z) = m_\xi(0) = \mathrm{E}[|\xi_1|]/2$. Then we impose the following assumption on the noise ξ.

Assumption 1 $\mathrm{E}[\xi_1^2] < \infty$ *and the measure* $m_\xi(z)\mathrm{d}z$ *is absolutely continuous with respect to the density function* $\mathrm{d}F_\xi(z)$ *with a bounded Radon-Nikodym derivative, i.e., there exists a bounded function* $g_\xi : \mathbb{R} \to \mathbb{R}_+$ *such that*

$$\int_a^b m_\xi(z)\mathrm{d}z = \int_a^b g_\xi(z)\mathrm{d}F_\xi(z), \quad \forall a, b \in \mathbb{R}.$$

This characterization of noise gives an extension of the Gaussian noise. Indeed the following examples satisfy the assumption:

- If ξ_1 obeys the Gaussian $\mathcal{N}(0, \sigma^2)$, then $g_\xi(z) = \sigma^2$,
- If ξ_1 obeys the uniform distribution on $[-a, a]$, then $g_\xi(z) = \max(a^2 - z^2, 0)/2$.

Under Assumption 1, Theorem 1 of Dalalyan and Tsybakov (2008) gives the following PAC-Bayesian bound. For a probability measure ρ that is absolutely continuous with respect to Π, let $\mathcal{K}(\rho, \Pi)$ be the KL-divergence between ρ and Π, $\mathcal{K}(\rho, \Pi) := \int \log(\frac{\mathrm{d}\rho}{\mathrm{d}\Pi}(f))\mathrm{d}\rho(f)$.

Theorem 1 *Suppose Assumption 1 is satisfied and* $\beta \ge 4\|g_\xi\|_\infty$. *Then for all probability measure* ρ *that is absolutely continuous with respect to* Π, *we have*

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^{\mathrm{o}}\|_n^2\right] \le \int \|f - f^{\mathrm{o}}\|_n^2 \mathrm{d}\rho(f) + \frac{\beta\mathcal{K}(\rho, \Pi)}{n}. \tag{3}$$

In the following, we assume that β is chosen so that $\beta \ge 4\|g_\xi\|_\infty$ is satisfied.

Remark 2 *If we restrict ourselves to Gaussian noise settings, we obtain a different type of bound such that*

$$P\left[\int \|f - f^{\mathrm{o}}\|_n^2 \mathrm{d}\Pi(f|Y_{1:n}) \ge C\left(\int \|f - f^{\mathrm{o}}\|_n^2 \mathrm{d}\rho(f) + \frac{\beta(\mathcal{K}(\rho, \Pi) + \log(\epsilon^{-1}))}{n}\right)\right] \ge 1 - \epsilon,$$

where exponential tail probability is given and the posterior expectation in the quantity $\int \|f - f^{\mathrm{o}}\|_n^2 \mathrm{d}\Pi(f|Y_{1:n})$ *is taken outside the* L_2*-norm* $\|\cdot - f^{\mathrm{o}}\|_n^2$ *instead of "plugging-in" the estimator as* $\|\hat{f} - f^{\mathrm{o}}\|_n^2$. *However we don't go to this direction. Instead, we deal with a more general class of noise.*

4. Main Results

In this section, we give our main results. The convergence rate of Gaussian process estimators is determined by how the prior distribution concentrates around the true function. The quantitative evaluation of the mass around the true function is given by the following *concentration function* (van der Vaart and van Zanten, 2011, 2008a):

$$\phi_{f_m^*}^{(m)}(\epsilon, \lambda_m) := \inf_{h \in \mathcal{H}_m : \|h - f_m^*\|_\infty \le \epsilon} \left(\|h\|_{\mathcal{H}_{m,\lambda_m}}^2 \vee 1 \right) - \log \mathrm{GP}_m(\{f : \|f\|_\infty \le \epsilon\} | \lambda_m), \quad (4)$$

where $a \vee b := \max(a, b)$. It can be shown that $\phi_{f_m^*}^{(m)}(\epsilon, \lambda_m)$ equals $-\log \mathrm{GP}_m(\{f : \|f_m^* - f\|_\infty \le \epsilon\} | \lambda_m)$ up to constants (van der Vaart and van Zanten, 2008b). The second term $-\log \mathrm{GP}_m(\{f : \|f\|_\infty \le \epsilon\} | \lambda_m)$ measures the small ball probability around the origin. There are large amount of studies for the small probability of Gaussian process measures; see, for example, Kuelbs and Li (1993) and Li and Shao (2001). The first term measures how the small ball probability decreases by shifting the center of the small ball away from the origin.

4.1. General Results

Let $\check{I}_0 := \{m \in I_0 \mid f_m^* \notin \mathcal{H}_m\}$, and $\kappa := \zeta(1 - \zeta)$. The following theorem gives the general theoretical tool to derive the convergence rate of Bayesian-MKL.

Theorem 3 (Convergence rate of Bayesian-MKL) *There exists a constant C_1 depending on only β such that the convergence rate of Bayesian-MKL is bounded as*

$$\begin{aligned} &\mathrm{E}_{Y_{1:n}|x_{1:n}} \left[\|\hat{f} - f^{\mathrm{o}}\|_n^2 \right] \le 2\|f^{\mathrm{o}} - f^*\|_n^2 \\ &+ C_1 \inf_{\epsilon_m, \lambda_m > 0} \left\{ \sum_{m \in I_0} \left(\epsilon_m^2 + \frac{1}{n} \phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) + \frac{\lambda_m}{n} - \frac{\log(\lambda_m)}{n} \right) + \sum_{\substack{m, m' \in \check{I}_0 : \\ m \neq m'}} \epsilon_m \epsilon_{m'} \right\} \\ &+ \frac{\beta |I_0|}{n} \log \left(\frac{Me}{\kappa |I_0|} \right). \end{aligned} \quad (5)$$

The complete proof is placed in Appendix A. Because of the term $\sum_{m, m' \in \check{I}_0 : m \neq m'} \epsilon_m \epsilon_{m'}$, the qualitative behavior of the convergence rate differs depending on how large $\check{I}_0$ is. To see this, we consider the following two extreme situations:

- (Correctly specified situation) $f_m^* \in \mathcal{H}_m$ $(\forall m = 1, \ldots, M)$, i.e., $\check{I}_0 = \emptyset$,
- (Misspecified situation) $f_m^* \notin \mathcal{H}_m$ $(\forall m = 1, \ldots, M)$, i.e., $\check{I}_0 = I_0$.

Roughly speaking, the term $\inf_{\epsilon_m, \lambda_m > 0} \left(\epsilon_m^2 + \frac{1}{n} \phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) + \frac{\lambda_m}{n} - \frac{\log(\lambda_m)}{n} \right)$ gives the convergence rate of Gaussian process estimators for the *single kernel learning*, say $\hat{\epsilon}_m^2$. For simplicity, suppose $\hat{\epsilon}_m^2$ is independent of m (denote it by $\hat{\epsilon}^2$), and assume $f^{\mathrm{o}} = f^*$. Then, in the correctly specified situation, the convergence rate can be evaluated as

$$\mathrm{E}_{Y_{1:n}|x_{1:n}} \left[\|\hat{f} - f^{\mathrm{o}}\|_n^2 \right] = O \left[|I_0| \hat{\epsilon}^2 + \frac{|I_0|}{n} \log \left(\frac{Me}{\kappa |I_0|} \right) \right].$$

This formulation is identical to well-known minimax optimal learning rate (Raskutti et al., 2012), that is, if $\hat{\epsilon}^2$ yields the minimax optimal rate for the single kernel learning (that is typically true), then Bayesian-MKL is also minimax optimal in the MKL setting. Importantly, the theorem does not require any condition on the design such as the restricted eigenvalue condition (Koltchinskii and Yuan, 2010) or the incoherence assumption (Meier et al., 2009). On the other hand, in the misspecified situation, the rate becomes

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f}-f^{\mathrm{o}}\|_n^2\right] = O\left[|I_0|^2\hat{\epsilon}^2 + \frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)\right].$$

Note that dependency of the rate on $|I_0|$ differs according to the situation. This discrepancy is induced by the fact that the cross terms $\langle f_m^* - \hat{f}_m, f_{m'}^* - \hat{f}_{m'}\rangle_n$ in the expansion $\|\sum_{m\in I_0}(f_m^* - \hat{f}_m)\|_n^2 = \sum_{m\in I_0}\|f_m^* - \hat{f}_m\|_n^2 + \sum_{m,m'\in I_0: m\neq m'}\langle f_m^* - \hat{f}_m, f_{m'}^* - \hat{f}_{m'}\rangle_n$ are not negligible because of the bias ($f_m^* \notin \mathcal{H}_m$). If the "design" is well-conditioned ($\|\sum_{m\in I_0}(f_m - f_m^*)\|_n^2 \leq C\sum_{m\in I_0}\|f_m - f_m^*\|_n^2$ for all f_m on the support of the prior), then the cross terms can be omitted and the first term $|I_0|^2\hat{\epsilon}^2$ in the bound is replaced with $|I_0|\hat{\epsilon}^2$.

Note that the second term $\frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)$ is better by an amount of $\frac{|I_0|}{n}\log(|I_0|)$ than that of the ever shown rate of the risk minimization type MKL where the corresponding term is $\frac{|I_0|}{n}\log(M)$.

4.2. Convergence Rates on Several Classes

Here we give convergence rates of Bayesian-MKL on several important examples.

4.2.1. MATÉRN PRIORS

Suppose that $\mathcal{X}_m = [0,1]^{d_m}$, and the kernel function associated with GP_m is the Matérn prior with the smoothness parameter α_m: The spectral density for k_m is given as $\psi(s) = \frac{1}{(1+\|s\|^2)^{\alpha_m + d_m/2}}$. Then the Gaussian process GP_m takes its value in $C^{\alpha'_m}[0,1]^{d_m}$ for any $\alpha'_m < \alpha_m$ while the RKHS $\mathcal{H}_m$ is contained in a Sobolev space $W^{\alpha_m + d_m/2}[0,1]^{d_m}$ with the smoothness $\alpha_m + d_m/2$ (van der Vaart and van Zanten, 2011).

Correctly specified situation Here suppose that $f_m^* \in \mathcal{H}_m$ for all $m \in I_0$, and $\max_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m} \leq R$. Then we obtain the following convergence rate.

Theorem 4 (Matérn prior, correctly specified) *If $f_m^* \in \mathcal{H}_m$ and $\max_{m\in I_0}\|f_m^*\|_{m\in I_0} \leq R$ for a constant R, then there exists a constant C_1' depending on $\{d_m, \alpha_m\}_{m\in I_0}, R, \beta$ such that*

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f}-f^{\mathrm{o}}\|_n^2\right] \leq 2\|f^{\mathrm{o}} - f^*\|_n^2 + C_1'\left\{\sum_{m\in I_0} n^{-\frac{1}{1+d_m/(2\alpha_m+d_m)}} + \frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)\right\}.$$

Note that $n^{-\frac{1}{1+d_m/(2\alpha_m+d_m)}}$ is the optimal rate to estimate $f_m^* \in W^{\alpha_m+d_m/2}[0,1]^{d_m}$ in single kernel learning settings ($M = |I_0| = 1$). If we don't put the exponential prior on the scale λ_m (inverse gamma prior on the scale), the Gaussian process estimation never attains the optimal rate on $\mathcal{H}_m$ (van der Vaart and van Zanten, 2011). However our result achieves the optimal rate. This is because we employed a mixture of Gaussian process priors with various scales that enables the Bayesian estimator to adaptively fit the appropriate scale.

Our convergence rate consists of the sum of the optimal learning rates in single kernel settings and the additional term $\frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)$. For the situation where all α_m, d_ms are same, $\exists\alpha, d$ such that $\alpha_m = \alpha$ and $d_m = d$ $(\forall m)$, it has been shown that this rate is optimal (Raskutti et al., 2012) .

Misspecified situation In the above, we have assumed that f_m^* possesses the smoothness $\alpha_m + d_m/2$. However, one might want to estimate a less smooth function. Here we assume that $f_m^* \in C^{\beta_m}[0,1]^{d_m} \cap W^{\beta_m}[0,1]^{d_m}$ where $\beta_m < \alpha_m + d_m/2$ for all $m \in I_0$. Note that, since $\beta_m < \alpha_m + d_m/2$, f_m^* is not necessarily contained in $\mathcal{H}_m$. Here we denote by $\|f_m\|_{\beta_m|\infty}$ the Besov norm of regularity β_m measured by L_∞-L_∞ norm (see Section 7.32 of Adams and Fournier (2003) for the definition). Then we obtain the following bound.

Theorem 5 (Matérn prior, misspecified) *If* $\max_{m\in I_0}\|f_m^*\|_{\beta_m|\infty} \le R$ *with some constant* R*, then there exists a constant* C_1' *depending on* $\{\alpha_m, \beta_m, d_m\}_{m\in I_0}, \beta, R$ *such that*

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^{\mathrm{o}}\|_n^2\right] \le 2\|f^{\mathrm{o}} - f^*\|_n^2 + C_1'\left\{\left(\sum_{m\in I_0} n^{-\frac{\beta_m}{2\beta_m+d_m}}\right)^2 + \frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)\right\}.$$

This result improves that of van der Vaart and van Zanten (2011) in the following three points:

- The Gaussianity is not assumed,
- The situation where $M > 1$ is covered,
- When $M = 1$, our rate achieves the optimal rate $n^{-\frac{2\beta_m}{2\beta_m+d_m}}$ for all $\beta_m < \alpha_m + d_m/2$ while the rate in van der Vaart and van Zanten (2011) achieves the optimal rate only when $\alpha_m = \beta_m$.

The third point is due to the adaptivity induced by the scale mixture prior. Without the scale mixture prior, the optimal rate can not be achieved whenever $\alpha_m \neq \beta_m$ (Castillo, 2008). An interesting observation here is that the choice of α_m has no influence on the learning rate. In other word, any fine tuning of parameters is not needed to achieve the optimal rate. We just need to choose α_m sufficiently large so that $\beta_m \le \alpha_m + d_m/2$, then the Gaussian process with scale mixture automatically yields the optimal rate. This kind of adaptivity for the smoothness is also pointed out in the context of regularized risk minimization procedures in kernel learning (Steinwart et al., 2009).

4.2.2. KERNELS WITH METRIC ENTROPY OF POLYNOMIAL COMPLEXITY

Here we derive general convergence rate results that are applicable to a general kernel class. We assume that the kernel is attached with an RKHS the unit ball of which possesses a metric entropy of polynomial order complexity. More precisely, there exists a real value $0 < s_m < 1$ such that

$$\log N(\mathcal{B}_{\mathcal{H}_m}, \epsilon, \|\cdot\|_\infty) = O(\epsilon^{-2s_m}), \tag{6}$$

where $N(B, \epsilon, d)$ is the ϵ-covering number of the space B with respect to the metric d (van der Vaart and Wellner, 1996), and $\mathcal{B}_{\mathcal{H}_m}$ is the unit ball of the RKHS $\mathcal{H}_m$. It is known that $-\log(\mathrm{GP}_m(\{f : \|f\|_\infty \le \epsilon\})) = O(\epsilon^{-\frac{2s_m}{1-s_m}})$ under the metric entropy condition (6) (Kuelbs and Li, 1993; Li and Shao, 2001). Thus, if we can evaluate the bias $\inf_{h\in\mathcal{H}_m:\|h-f_m^*\|_\infty\le\epsilon}\|h\|_{\mathcal{H}_{m,\lambda_m}}^2$ in addition

to the evaluation of the small ball probability, we obtain a convergence rate also for misspecified situations $f_m^* \notin \mathcal{H}_m$. Here we consider two situations; (i) $f_m^* \in \mathcal{H}_m$ and (ii) $f_m^* \notin \mathcal{H}_m$ as in previous sections. To derive a convergence rate on an arbitrary augmented space $\mathcal{H}_m (\supset \mathcal{H}_m)$ is a tough problem. However *real interpolation* of spaces (Bennett and Sharpley, 1988) gives a clear characterization of the convergence rate. Suppose that we have a couple of Banach spaces X_0 and X_1 such that $X_0 \supset X_1$ and X_1 is continuously embedded in X_0 (denoted by $X_1 \hookrightarrow X_0$). We define the *K-functional* as

$$K(f,t) = \inf_{f_1 \in X_1} \{\|f - f_1\|_{X_0} + t\|f_1\|_{X_1}\},$$

for all $t > 0$ and $f \in X_0$. Then the real interpolation space $[X_0, X_1]_{\theta,r}$ with $0 < \theta < 1, 1 \leq r < \infty$ or $0 \leq \theta \leq 1, r = \infty$ is a space consisting of all functions $f \in X_0$ that possess the finite norm $\|f\|_{\theta,r}$:

$$\|f\|_{\theta,r} = \|f\|_{\theta,r,[X_0,X_1]} = \begin{cases} \left[\int_0^\infty (t^{-\theta}K(f,t))^r \frac{\mathrm{d}t}{t}\right]^{1/r}, & (0 < \theta < 1,\ 1 \leq r < \infty), \\ \sup_{t>0} t^{-\theta}K(f,t), & (0 \leq \theta \leq 1,\ r = \infty). \end{cases} \tag{7}$$

The real interpolation space $[X_0, X_1]_{\theta,r}$ is an intermediate space between X_0 and X_1, i.e., $X_1 \hookrightarrow [X_0, X_1]_{\theta,r} \hookrightarrow X_0$. One can check that, in extreme cases, we have $[X_0, X_1]_{0,\infty} = X_0$ and $[X_0, X_1]_{1,\infty} = X_1$. In particular, we are interested in the space $[L_\infty(\mathcal{X}_m), \mathcal{H}_m]_{\theta,\infty}$ for which we can give the convergence rate of Bayesian-MKL. To give a concrete example, suppose $\mathcal{H}_m = W^{\alpha_m}(\mathcal{X}_m)$, then Theorem 1.12 of Bennett and Sharpley (1988) gives

$$[L_\infty(\mathcal{X}_m), \mathcal{H}_m]_{\theta,\infty} = [L_\infty(\mathcal{X}_m), W^{\alpha_m}(\mathcal{X}_m)]_{\theta,\infty} \hookrightarrow B_{2,\infty}^{\theta\alpha_m}(\mathcal{X}_m),$$

where $B_{2,\infty}^{\theta\alpha_m}(\mathcal{X}_m)$ denotes a Besov space of regularity $\theta\alpha_m$ with L_2-L_∞ norm[3] (see Adams and Fournier (2003) for the definition). In addition, if $\mathcal{X}_m = [0,1]^{d_m}$, then it is known that $s_m = \frac{d_m}{2\alpha_m}$ satisfies the entropy condition (6) for $\mathcal{H}_m = W^{\alpha_m}(\mathcal{X}_m)$. Now we denote by $\|f_m\|_{\theta,r}^{(m)} := \|f_m\|_{\theta,r,[L_\infty(\mathcal{X}_m),\mathcal{H}_m]}$. Finally we assume that the constant hidden in the small ball probability upper bound is bounded uniformly for all $m = 1, \ldots, M$ for simplicity: $\exists C_0 > 0$ such that

$$-\log(\mathrm{GP}_m(\{f : \|f\|_\infty \leq \epsilon\})) \leq C_0(\epsilon^{-\frac{2s_m}{1-s_m}}) \quad (\forall m = 1, \ldots, M).$$

Then we obtain the following theorem.

Theorem 6 (RKHS with metric entropy condition) *If $f_m^* \in \mathcal{H}_m$ for all $m \in I_0$ and $\max_{m \in I_0} \|f_m^*\|_{\mathcal{H}_m} \leq R$, then there exists a constant C_1' depending on $\{s_m\}_{m \in I_0}, C_0, R, \beta$ such that*

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^{\mathrm{o}}\|_n^2\right] \leq 2\|f^{\mathrm{o}} - f^*\|_n^2 + C_1'\left\{\sum_{m \in I_0} n^{-\frac{1}{1+s_m}} + \frac{|I_0|}{n}\log\left(\frac{Me}{\kappa|I_0|}\right)\right\}.$$

3. $[L_2(\mathcal{X}_m), W^{\alpha_m}(\mathcal{X}_m)]_{\theta,\infty} = B_{2,\infty}^{\theta\alpha_m}(\mathcal{X}_m)$ by the definition.

If $f_m^* \in [L_\infty(\mathcal{X}_m), \mathcal{H}_m]_{\theta,\infty}$ *with* $0 < \theta \leq 1$ *for all* $m \in I_0$ *and* $\max_{m \in I_0} \|f_m^*\|_{\theta,\infty}^{(m)} \leq R$ *with a constant* R*, then there exists a constant* C_1' *depending on* $\{s_m\}_{m \in I_0}, \theta, C_0, R, \beta$ *such that*

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^{\mathrm{o}}\|_n^2\right] \leq 2\|f^{\mathrm{o}} - f^*\|_n^2 + C_1' \left\{ \left(\sum_{m \in I_0} n^{-\frac{1}{2(1+s_m/\theta)}} \right)^2 + \frac{|I_0|}{n} \log\left(\frac{Me}{\kappa |I_0|}\right) \right\}.$$

The proof can be found in Appendix B. Under the metric entropy condition (6), the convergence rate $n^{-\frac{1}{1+s_m}}$ is minimax optimal in typical situations. Moreover, when $\mathcal{X}_m = [0,1]^{d_m}$, since $B_{\infty,\infty}^{\theta\alpha_m}(\mathcal{X}_m) \hookrightarrow [L_\infty(\mathcal{X}_m), W^{\alpha_m}(\mathcal{X}_m)]_{\theta,\infty} \hookrightarrow B_{2,\infty}^{\theta\alpha_m}(\mathcal{X}_m)$, the metric entropy of $[L_\infty(\mathcal{X}_m), W^{\alpha_m}(\mathcal{X}_m)]_{\theta,\infty}$ satisfies (6) where s_m is replaced with $s_m' = \frac{d_m}{2\alpha_m\theta} = \frac{s_m}{\theta}$, and that is tight (see Theorem 2 of Edmunds and Triebel (1996) and A.5.6 of Steinwart (2008)). Thus the convergence rate $n^{-\frac{1}{1+s_m/\theta}}$ is minimax optimal on $[L_\infty(\mathcal{X}_m), W^{\alpha_m}(\mathcal{X}_m)]_{\theta,\infty}$ as long as $s_m/\theta < 1$. In that sense, Theorem 6 states that Bayesian-MKL achieves the optimal rate (as for the misspecified situation, it is true at least when $M = 1$). Here we again observe that the Gaussian process with scale mixture adaptively achieves the optimal rate for all θ such that $s_m < \theta \leq 1$. Thus the convergence rate is *not* influenced by oversmooth specification.

Note that Theorem 6 includes the analysis of the Matérn prior as a special case. Because the RKHS $\mathcal{H}_m$ corresponding to the Matérn prior is continuously embedded in the Sobolev space $W^{\alpha_m + d_m/2}[0,1]^{d_m}$ so that the metric entropy condition (6) is satisfied with $s_m = d_m/(2\alpha_m + d_m)$. Moreover the proof of Lemma 4 of van der Vaart and van Zanten (2011) yields that functions $f_m^* \in C^{\beta_m}[0,1]^{d_m} \cap W^{\beta_m}[0,1]^{d_m}$ with $\|f_m^*\|_{\beta_m|\infty} \leq R$ are included in a ball of the interpolation space $[L_\infty(\mathcal{X}_m), \mathcal{H}_m]_{\theta,\infty}$ with $\theta = \beta_m/(\alpha_m + d_m/2) \leq 1$. Thus Theorems 4 and 5 are recovered by Theorem 6 with the parameter setting $s_m = \frac{d_m}{2\alpha_m + d_m}$ and $\theta = \frac{\beta_m}{\alpha_m + d_m/2}$.

Group Lasso Finally we investigate the situation where each $\mathcal{H}_m$ is finite dimensional. This situation corresponds to Group Lasso (Yuan and Lin, 2006). Suppose $\mathcal{X}_m$ is a compact subset of $\mathbb{R}^{d_m}$ and the Gaussian process prior GP_m is as follows:

$$f_m(x) = \mu^\top x, \ \mu \sim \mathcal{N}(0, I_{d_m}),$$

where I_{d_m} is the $d_m \times d_m$ identity matrix. Then the corresponding kernel function is $k_m(x, x') = x^\top x'$. In this setting, the convergence rate of the Bayesian-MKL is given by the following theorem.

Theorem 7 (Group Lasso) *Suppose that* $f_m^*(x) = \mu_m^\top x$ *for some* $\mu_m \in \mathbb{R}^{d_m}$ *and* $\max_{m \in I_0} \|f_m\|_{\mathcal{H}_m} = \max_{m \in I_0} \|\mu_m\| \leq R$, $\sup_{x^{(m)} \in \mathcal{X}_m} \|x^{(m)}\| \leq R$ *for some constant* R*, then there exits a constant* C_1' *depending on* β, R *such that,*

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^{\mathrm{o}}\|_n^2\right] \leq 2\|f^{\mathrm{o}} - f^*\|_n^2 + C_1' \left\{ \frac{\sum_{m \in I_0} d_m \log(n)}{n} + \frac{|I_0|}{n} \log\left(\frac{Me}{\kappa |I_0|}\right) \right\}.$$

The proof can be found in Appendix C. This is rate optimal up to $\log(n)$ order because the optimal rate of the estimation problem on $\sum_{m \in I_0} d_m$ dimensional parameter space $(\mu_m)_{m \in I_0}$ is $\frac{\sum_{m \in I_0} d_m}{n}$, and $\frac{|I_0|}{n} \log\left(\frac{Me}{|I_0|\kappa}\right)$ is the optimal rate for sparse linear regression with $|I_0|$ non-zeros components (Rigollet and Tsybakov, 2011a).

5. Conclusion and Discussion

In this paper, we developed a PAC-Bayesian bound for Gaussian process model and generalized it to sparse additive model. Important notion was that the optimal rate is achieved *without* any conditions on the design. Interpolations of spaces gave a nice characterization of the convergence rate on the misspecified situation. We have observed that Gaussian processes with scale mixture adaptively achieve the minimax optimal rate on both correctly-specified and misspecified situations.

We bounded the empirical L_2-norm $\|\cdot\|_n$ in this paper. However, the evaluation of the population L_2-norm, $\|f\|^2_{L_2(P_X)} = \int f(X)^2 \mathrm{d}P_X$, between the estimator and the true function is also of interest from the view point of generalization error. For the analysis of the population L_2-norm, the L_∞-norm in the metric entropy condition (6) and the definition (4) of $\phi^{(m)}_{f^*_m}$ could be replaced with the population L_2-morm $\|\cdot\|_{L_2(P_X)}$. To bound the population L_2-norm, we would need to impose some smoothness condition on the prior (see Theorem 2 and the following discussions in van der Vaart and van Zanten (2011)). Our future work includes developing a PAC-Bayesian bound that is also applicable to the population L_2-norm.

Another interesting topic is to compare Bayesian-MKL with a model selection type method that minimizes a penalized risk like the BIC estimator. Rigollet and Tsybakov (2011a) discussed benefits of a model averaging type estimator comparing to a BIC type estimator in a finite dimensional linear model. It is interesting to argue an analogous thing also in a nonparametric regression situation.

Acknowledgments

We would like to thank Alexandre B. Tsybakov and Pierre Alquier for their suggestive advices. TS was partially supported by MEXT Kakenhi 22700289, Global COE Program "The Research and Training Center for New Development in Mathematics," and the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

References

R. A. Adams and J. J. Fournier. *Sobolev Spaces*. Academic Press, New York, 2003. second edition.

P. Alquier and G. Biau. Sparse single-index model. Technical report, 2011. arXiv:1101.3229.

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

C. Archambeau and F. Bach. Multiple Gaussian process models. In *NIPS 2010 Workshop on New Directions in Multiple Kernel Learning*, Whistler, 2010.

F. R. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *the 21st International Conference on Machine Learning*, pages 41–48, 2004.

C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

H. J. Brascamp and E. H. Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorem, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Lecture Notes in Mathematics. Springer, 2004. Saint-Flour Summer School on Probability Theory 2001.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, in press, 2011.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1997.

P. J. Green. Reversible jump markov chain monte carlo computation. *Biometrika*, 82(4):711–732, 1995.

L. Gross. Measurable functions on Hilbert space. *Transactions of the American Mathematical Society*, 105(3):372–390, 1962.

G. Hargé. A particular case of correlation inequality for the gaussian measure. *The Annals of Probability*, 27(4):1939–1951, 1999.

G. Hargé. A convex/log-concave correlation inequality for gaussian measure and an application to abstract wiener spaces. *Probability Theory and Related Fields*, 130(3):415–440, 2004.

T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall Ltd, 1999.

V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6): 3660–3695, 2010.

J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.

G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods*, 19:533–597, 2001.

J.-M. Marin and C. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, 2007.

D. McAllester. Some PAC-Bayesian theorems. In *the Anual Conference on Computational Learning Theory*, pages 230–234, 1998.

D. McAllester. PAC-Bayesian model averaging. In *the Anual Conference on Computational Learning Theory*, pages 164–170, 1999.

L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:389–427, 2012.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71(5):1009–1030, 2009.

P. Rigollet and A. B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011a.

P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. Technical report, 2011b. arXiv:1108.5116.

M. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14 (2), 2004.

I. Steinwart. *Support Vector Machines*. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pages 79–93, 2009.

T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. In *JMLR Workshop and Conference Proceedings 22*, pages 1152–1183, 2012. Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS2012).

R. Tomioka and T. Suzuki. Regularization strategies and empirical bayesian learning for mkl. In *NIPS 2010 Workshop: New Directions in Multiple Kernel Learning*, Whistler, 2010.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008a.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3: 200–222, 2008b. IMS Collections.

A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675, 2009.

A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

Appendix A. Proof of Theorem 3

Fix $\epsilon_m, \lambda_m > 0$. To prove the theorem, we substitute some "dummy" posterior distribution into ρ in Eq. (3) of Theorem 1 (the PAC-Bayes bound). If $f_m^* \in \mathcal{H}_m$, then we take $\tilde{h}_m$ as $\tilde{h}_m = f_m^*$. Otherwise, we take $\tilde{h}_m \in \mathcal{H}_{m,\lambda_m}$ such that

$$\|\tilde{h}_m\|_{\mathcal{H}_m,\lambda_m}^2 \leq 2 \inf_{h \in \mathcal{H}_m : \|h - f_m^*\|_\infty \leq \epsilon_m} \|h\|_{\mathcal{H}_{m,\lambda_m}}^2 .$$

The process $(W_x + \tilde{h}_m(x) : x \in \mathcal{X}_m)$ induces the "shifted" Gaussian process $\mathrm{GP}_m^{W+\tilde{h}_m}(\mathrm{d}f_m|\tilde{\lambda}_m)$ such that $\mathrm{GP}_m^{W+\tilde{h}_m}(A|\tilde{\lambda}_m) := \mathrm{GP}_m(A - \tilde{h}_m|\tilde{\lambda}_m)$ for a measurable set A. Now our choice of ρ is given as follows:

$$\rho(\mathrm{d}f) = \prod_{m \in I_0} \frac{\displaystyle\int_{\frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m} \frac{\mathrm{GP}_m^{W+\tilde{h}_m}(\mathrm{d}f_m|\tilde{\lambda}_m)\mathbf{1}\{\|f_m - \tilde{h}_m\|_\infty \leq \epsilon_m\}}{\mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\tilde{\lambda}_m)} \mathcal{G}(\mathrm{d}\tilde{\lambda}_m)}{\mathcal{G}(\{\tilde{\lambda}_m : \frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m\})} \cdot \prod_{m \notin I_0} \delta_0(\mathrm{d}f_m),$$

We can show that ρ is absolutely continuous with respect to the prior Π as follows. First notice that

$$\begin{aligned}
&\Pi(\mathrm{d}f) \\
\geq &\pi_{I_0} \cdot \prod_{m \in I_0} \int_{\tilde{\lambda}_m \in \mathbb{R}_+} \mathrm{GP}_m(\mathrm{d}f_m|\tilde{\lambda}_m)\mathcal{G}(\mathrm{d}\tilde{\lambda}_m) \cdot \prod_{m \notin I_0} \delta_0(\mathrm{d}f_m) \\
\geq &\pi_{I_0} \cdot \prod_{m \in I_0} \int_{\frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m} \mathrm{GP}_m(\mathrm{d}f_m|\tilde{\lambda}_m)\mathbf{1}\{\|f_m - \tilde{h}_m\|_\infty \leq \epsilon_m\}\mathcal{G}(\mathrm{d}\tilde{\lambda}_m) \cdot \prod_{m \notin I_0} \delta_0(\mathrm{d}f_m). \qquad (8)
\end{aligned}$$

Here we define a linear map $U_{f_m}^{(\tilde{\lambda}_m)} : \mathcal{H}_{m,\tilde{\lambda}_m} \to \mathbb{R}$ by setting $U_{f_m}^{(\tilde{\lambda}_m)} \tilde{k}_{m,\tilde{\lambda}_m}(x, \cdot) = f_m(x)$ and extending linearly and continuously to an arbitrary $h \in \mathcal{H}_m$. This induces an isometry $U_\cdot^{(\tilde{\lambda}_m)} : \mathcal{H}_{m,\tilde{\lambda}_m} \to L_2(\mathrm{GP}_m(\cdot|\tilde{\lambda}_m))$ because $\int [U_{f_m}^{(\tilde{\lambda}_m)}(\sum_{j=1}^J \alpha_j \tilde{k}_{m,\tilde{\lambda}_m}(z_j, \cdot))]^2 \mathrm{GP}_m(\mathrm{d}f_m|\tilde{\lambda}_m) = \sum_{j=1}^J \sum_{j'=1}^J \alpha_j \alpha_{j'} \int f_m(z_j) f_m(z_{j'}) \mathrm{GP}_m(\mathrm{d}f_m|\tilde{\lambda}_m) = \sum_{j=1}^J \sum_{j'=1}^J \alpha_j \alpha_{j'} \tilde{k}_{m,\tilde{\lambda}_m}(z_j, z_{j'})$. According to Lemma 3.1 of van der Vaart and van Zanten (2008a), $\mathrm{GP}_m(\cdot|\tilde{\lambda}_m)$ and $\mathrm{GP}_m^{W+\tilde{h}_m}(\cdot|\tilde{\lambda}_m)$ are equivalent, and moreover, for f_m such that $\|f_m - \tilde{h}_m\|_\infty \leq \epsilon_m$, we have

$$\begin{aligned}
&\frac{\displaystyle\int_{\frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m} \frac{\mathrm{GP}_m^{W+\tilde{h}_m}(\mathrm{d}f_m|\tilde{\lambda}_m)}{\mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\tilde{\lambda}_m)} \mathcal{G}(\mathrm{d}\tilde{\lambda}_m)}{\int_{\frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m} \mathrm{GP}_m(\mathrm{d}f_m|\tilde{\lambda}_m)\mathcal{G}(\mathrm{d}\tilde{\lambda}_m)} \\
\leq &\sup_{\tilde{\lambda}_m : \frac{\lambda_m}{2} \leq \tilde{\lambda}_m \leq \lambda_m} \frac{\mathrm{GP}_m^{W+\tilde{h}_m}(\mathrm{d}f_m|\tilde{\lambda}_m)}{\mathrm{GP}_m(\mathrm{d}f_m|\tilde{\lambda}_m) \cdot \mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\tilde{\lambda}_m)}
\end{aligned}$$

$$\leq \sup_{\tilde{\lambda}_m:\frac{\lambda_m}{2}\leq\tilde{\lambda}_m\leq\lambda_m} \exp\left(U_{f_m}^{(\tilde{\lambda}_m)}\tilde{h}_m - \frac{1}{2}\|\tilde{h}_m\|_{\mathcal{H}_m,\tilde{\lambda}_m}^2\right) \frac{1}{\mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\tilde{\lambda}_m)}$$

($\because$ Lemma 3.1 of van der Vaart and van Zanten (2008a))

$$\leq \sup_{\tilde{\lambda}_m:\frac{\lambda_m}{2}\leq\tilde{\lambda}_m\leq\lambda_m} \exp\left[(U_{f_m-\tilde{h}_m}^{(\tilde{\lambda}_m)} + U_{\tilde{h}_m}^{(\tilde{\lambda}_m)})\tilde{h}_m - \frac{1}{2}\|\tilde{h}_m\|_{\mathcal{H}_{m,\tilde{\lambda}_m}}^2\right] \frac{1}{\mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\tilde{\lambda}_m)}$$

$$\leq \exp\left[|U_{f_m-\tilde{h}_m}^{(\lambda_m)}\tilde{h}_m| + \frac{1}{2}\|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}}^2\right] \frac{1}{\mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\lambda_m/2)} \tag{9}$$

$$< \infty, \quad \text{(a.s.)}. \tag{10}$$

Therefore combining Eq. (8) and Eq. (10), we have that ρ is absolutely continuous with respect to Π. Using the bound (9), we obtain that $\mathcal{K}(\rho,\Pi)$ is bounded from above as

$$\begin{aligned}
&\mathcal{K}(\rho,\Pi)\\
&\leq \int \log\left\{\frac{1}{\pi_{I_0}}\prod_{m\in I_0}\frac{\exp\left[|U_{f_m-\tilde{h}_m}^{(\lambda_m)}\tilde{h}_m| + \frac{1}{2}\|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}}^2\right]}{\mathrm{GP}_m(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}|\lambda_m/2)\mathcal{G}(\{\tilde{\lambda}_m : \frac{\lambda_m}{2}\leq\tilde{\lambda}_m\leq\lambda_m\})}\right\}\rho(\mathrm{d}f)\\
&= \int \sum_{m\in I_0}\left(|U_{f_m-\tilde{h}_m}^{(\lambda_m)}\tilde{h}_m| + \frac{1}{2}\|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}}^2\right)\rho(\mathrm{d}f)\\
&\quad - \sum_{m=1}^{M}\log\left(\mathrm{GP}_m\left(\{\Delta f_m : \|\Delta f_m\|_\infty \leq \epsilon_m\}\,|\frac{\lambda_m}{2}\right)\right)\\
&\quad - \sum_{m=1}^{M}\log\left(\mathcal{G}\left(\{\tilde{\lambda}_m : \frac{\lambda_m}{2}\leq\tilde{\lambda}_m\leq\lambda_m\}\right)\right) - \log(\pi_{I_0}).
\end{aligned} \tag{11}$$

Here we have the following bounds for each term. By Lemma 8, the first term is bounded as

$$\begin{aligned}
&\int \sum_{m\in I_0}\left(|U_{f_m-\tilde{h}_m}^{(\lambda_m)}\tilde{h}_m| + \frac{1}{2}\|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}}^2\right)\rho(\mathrm{d}f)\\
\leq & C\sum_{m\in I_0}\left(\|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}} + \|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}}^2\right) \leq 2C\sum_{m\in I_0}\left(\|\tilde{h}_m\|_{\mathcal{H}_{m,\lambda_m}}^2 \vee 1\right),
\end{aligned} \tag{12}$$

where C is a universal constant. The third term is bounded as

$$\begin{aligned}
&-\log\left(\mathcal{G}\left(\{\tilde{\lambda}_m : \frac{\lambda_m}{2}\leq\tilde{\lambda}_m\leq\lambda_m\}\right)\right) = -\log\left(\int_{\tilde{\lambda}_m:\frac{\lambda_m}{2}\leq\tilde{\lambda}_m\leq\lambda_m}\exp(-\tilde{\lambda}_m)\mathrm{d}\tilde{\lambda}_m\right)\\
\leq & -\log\left(\frac{\lambda_m}{2}\exp(-\lambda_m)\right) = -\log\left(\frac{\lambda_m}{2}\right) + \lambda_m.
\end{aligned} \tag{13}$$

The fourth term is bounded as

$$\begin{aligned}
&-\log(\pi_{I_0}) = -\log\left(\frac{\zeta^{|I_0|}}{\sum_{j=0}^{M}\zeta^j}\binom{M}{|I_0|}^{-1}\right)\\
\leq & |I_0|\log\left(\frac{1}{\zeta}\right) + \log\left(\frac{1}{1-\zeta}\right) + |I_0|\log\left(\frac{Me}{|I_0|}\right)
\end{aligned}$$

$$\leq |I_0| \log\left(\frac{Me}{|I_0|\zeta(1-\zeta)}\right). \tag{14}$$

Substituting Eqs. (12),(13),(14) into Eq. (11), the KL-divergence between the "dummy" posterior ρ and the prior distribution Π is bounded as

$$\begin{aligned}&\frac{1}{n}\mathcal{K}(\rho,\Pi)\\ &\leq C_1' \sum_{m\in I_0}\left(\frac{1}{n}\phi_{f_m^*}^{(m)}(\epsilon_m,\lambda_m/2)+\frac{1}{n}\lambda_m-\frac{1}{n}\log\left(\frac{\lambda_m}{2}\right)\right)+\frac{|I_0|}{n}\log\left(\frac{Me}{|I_0|\zeta(1-\zeta)}\right),\end{aligned} \tag{15}$$

where C_1' is a universal constant.

Finally we bound $\int \|f-f^{\mathrm{o}}\|_n^2 \mathrm{d}\rho(f)$. Notice that $\int \|f-f^{\mathrm{o}}\|_n^2 \mathrm{d}\rho(f) \leq 2\|f^{\mathrm{o}}-f^*\|_n^2 + 2\int \|f-f^*\|_n^2 \mathrm{d}\rho(f)$. Thus we only need to bound $\int \|f-f^*\|_n^2 \mathrm{d}\rho(f)$. By the definition of ρ, we have that

$$\begin{aligned}&\int \|f-f^*\|_n^2 \mathrm{d}\rho(f) = \int \left\|\sum_{m\in I_0}(f_m-f_m^*)\right\|_n^2 \mathrm{d}\rho(f)\\ &= \int \sum_{m\in I_0}\|f_m-f_m^*\|_n^2 \mathrm{d}\rho(f) + \int \sum_{m\neq m'\in I_0}\langle f_m-f_m^*, f_{m'}-f_{m'}^*\rangle_n \mathrm{d}\rho(f).\end{aligned} \tag{16}$$

Since the mean of f_m with respect to ρ is $\tilde{h}_m$, $\|f_m-\tilde{h}_m\|_\infty$ is bounded by ϵ_m on the support of ρ and $\|\tilde{h}_m-f_m^*\|_\infty \leq \epsilon_m$ by the definition, we have

$$\int \|f_m-f_m^*\|_n^2 \mathrm{d}\rho(f) \leq 2\int \|f_m-\tilde{h}_m\|_n^2 \mathrm{d}\rho(f) + 2\int \|\tilde{h}_m-f_m^*\|_n^2 \mathrm{d}\rho(f) \leq 4\epsilon_m^2,$$

and

$$\int \langle f_m-f_m^*, f_{m'}-f_{m'}^*\rangle_n \mathrm{d}\rho(f) = \langle \tilde{h}_m-f_m^*, \tilde{h}_{m'}-f_{m'}^*\rangle_n \begin{cases}\leq \epsilon_m\epsilon_{m'}, & (m,m'\in \check{I}_0),\\ =0, & (\text{otherwise}).\end{cases}$$

These bounds and Eq. (15) give the assertion by resetting $\lambda_m \leftarrow \lambda_m/2$.

Appendix B. Proof of Theorem 6

We show only the second assertion where $f_m^* \notin \mathcal{H}_m$. The first assertion can be shown in the same line. We utilize Theorem 3.

By the definition, we have $\|f_m^*\|_{\theta,\infty}^{(m)} = \sup_{t>0}\inf_{h_m\in\mathcal{H}_m}\{t^{-\theta}\|f_m^*-h_m\|_\infty + t^{1-\theta}\|h_m\|_{\mathcal{H}_m}\}$. If $\inf_{h_m\in\mathcal{H}_m}\|f_m^*-h_m\|_\infty > 0$, then the term $t^{-\theta}\|f_m^*-h_m\|_\infty$ can be arbitrary large. Therefore the assumption $R \geq \|f_m^*\|_{\theta,\infty}^{(m)}$ ensures that there exists $h_m \in \mathcal{H}_m$ such that $\|f_m^*-h_m\|_\infty \leq \epsilon$ for all $\epsilon > 0$. Now we evaluate the quantity $\inf_{h\in\mathcal{H}_m:\|h-f_m^*\|_\infty\leq\epsilon_m}\|h\|_{\mathcal{H}_m}^2$ by the assumption $\|f_m^*\|_{\theta,\infty}^{(m)} < \infty$. For all $t>0$, there exists $h_m^{(t)} \in \mathcal{H}_m$ such that $2\|f_m^*\|_{\theta,\infty}^{(m)} \geq t^{-\theta}\|f_m^*-h_m^{(t)}\|_\infty + t^{1-\theta}\|h_m^{(t)}\|_{\mathcal{H}_m}$. This gives $2\|f_m^*\|_{\theta,\infty}^{(m)} \geq t^{-\theta}\|f_m^*-h_m^{(t)}\|_\infty$ so that we have $t \geq 2^{-\frac{1}{\theta}}\|f_m^*\|_{\theta,\infty}^{(m)\,-\frac{1}{\theta}}\|f_m^*-h_m^{(t)}\|_\infty^{\frac{1}{\theta}}$, and hence $2\|f_m^*\|_{\theta,\infty}^{(m)} \geq t^{1-\theta}\|h_m^{(t)}\|_{\mathcal{H}_m}$ yields

$$\|h_m^{(t)}\|_{\mathcal{H}_m} \leq t^{-(1-\theta)}2\|f_m^*\|_{\theta,\infty}^{(m)} \leq 2^{\frac{1}{\theta}}\|f_m^*\|_{\theta,\infty}^{(m)\,\frac{1}{\theta}}\|f_m^*-h_m^{(t)}\|_\infty^{-\frac{1-\theta}{\theta}}.$$

Therefore we have that

$$\inf_{h\in\mathcal{H}_m:\|h-f_m^*\|_\infty\le\epsilon_m}\|h\|_{\mathcal{H}_m}^2 \le 2^{\frac{2}{\theta}}\|f_m^*\|_{\theta,\infty}^{(m)\,\frac{2}{\theta}}\epsilon_m^{-\frac{2(1-\theta)}{\theta}} \le (2R)^{\frac{2}{\theta}}\epsilon_m^{-\frac{2(1-\theta)}{\theta}},$$

because for all $\epsilon>0$ there exists t such that $\|f_m^*-h_m^{(t)}\|_\infty\le\epsilon$. This and the evaluation

$$-\log(\mathrm{GP}_m(\{f:\|f\|_\infty\le\epsilon\}))\le C_0\epsilon^{-\frac{2s_m}{1-s_m}}$$

gives that

$$\phi_{f_m^*}^{(m)}(\epsilon_m,\lambda_m)\le(2R)^{\frac{2}{\theta}}\lambda_m\epsilon_m^{-\frac{2(1-\theta)}{\theta}}+C_0(\sqrt{\lambda_m}\epsilon_m)^{-\frac{2s_m}{1-s_m}}, \tag{17}$$

where we used $-\log(\mathrm{GP}_m(\{f:\|f\|_\infty\le\epsilon\}|\lambda_m))=-\log(\mathrm{GP}_m(\{f:\|f\|_\infty\le\sqrt{\lambda_m}\epsilon\}))$. Now $\lambda_m=\epsilon_m^{\frac{1-\theta-s_m}{2\theta}}$ balances the two terms in the right hand side of the above display up to constants. With this λ_m, we have that

$$\begin{aligned}&\epsilon_m^2+\frac{1}{n}\phi_{f_m^*}^{(m)}(\epsilon_m,\lambda_m)+\frac{\lambda_m}{n}-\frac{\log(\lambda_m)}{n}\\ \le&\epsilon_m^2+\frac{((2R)^{\frac{2}{\theta}}+C_0)}{n}\epsilon_m^{-\frac{2s_m}{\theta}}+\frac{\epsilon_m^{\frac{1-\theta-s_m}{\theta}}}{n}-\frac{\log(\epsilon_m^{\frac{1-\theta-s_m}{\theta}})}{n}.\end{aligned} \tag{18}$$

Here we take $\epsilon_m=n^{-\frac{\theta}{2(\theta+s_m)}}$ that balances the first two terms of the RHS of the above display (up to constants). Then the RHS of Eq. (18) is further bounded by

$$\begin{aligned}&[1+(2R)^{\frac{2}{\theta}}+C_0]n^{-\frac{1}{1+s_m/\theta}}+n^{-\frac{1+\theta+s_m}{2(\theta+s_m)}}+\frac{1-\theta-s_m}{2(\theta+s_m)}\frac{\log(n)}{n}\\ \le&Cn^{-\frac{1}{1+s_m/\theta}},\end{aligned}$$

where C is a constant depending on s_m,R,C_0,θ. Substituting this bound into Theorem 3, we obtain the assertion.

Appendix C. Proof of Theorem 7

We utilize Theorem 3. Since $\{f:\|f\|_\infty\le\epsilon\}\supseteq\{f(x)=\beta^\top x:\|\beta\|\le\epsilon/R\}$, $-\log\mathrm{GP}_m(\{f:\|f\|_\infty\le\epsilon\}|\lambda_m)$ is bounded as

$$\begin{aligned}&-\log\mathrm{GP}_m(\{f:\|f\|_\infty\le\epsilon\}|\lambda_m)\le-\log\mathcal{N}(\{\beta\in\mathbb{R}^{d_m}:\|\beta\|\le\epsilon/R\}\mid 0,I_{d_m}/\lambda_m)\\ \le&-\log\left[\frac{\exp(-\frac{(\sqrt{\lambda_m}\epsilon/R)^2}{2})}{(2\pi\lambda_m^{-1})^{d_m/2}}\frac{\pi^{d_m/2}}{\Gamma(d_m/2+1)}(\epsilon/R)^{d_m}\right]\\ \le&\frac{(\sqrt{\lambda_m}\epsilon)^2}{2R^2}+\left(\frac{d_m}{2}+1\right)\log(2)-\frac{d_m}{2}\log(\lambda_m)+\frac{d_m}{2}\log\left(\frac{d_m}{2}\right)-d_m\log\left(\frac{\epsilon}{R}\right),\end{aligned}$$

where we used $\Gamma(d_m/2+1)\le2\left(\frac{d_m}{2}\right)^{\frac{d_m}{2}}$. Here set $\lambda_m=1$ and $\epsilon_m=\sqrt{\frac{d_m}{n}}$, then we have

$$-\log\mathrm{GP}_m(\{f:\|f\|_\infty\le\epsilon_m\}|\lambda_m)$$

$$\leq \frac{d_m}{2R^2 n} + \left(\frac{d_m}{2} + 1\right)\log(2) + \frac{d_m}{2}\log\left(\frac{d_m}{2}\right) + d_m \log(R) + \frac{d_m}{2}\log\left(\frac{n}{d_m}\right)$$
$$\leq \left(\frac{1}{2R^2} + 2\log(2) + \log(R)\right) d_m + \frac{d_m}{2}\log(n) \leq C d_m \log(n),$$

where C is a constant depending on R. This gives the following evaluation of $\phi_{f_m^*}^{(m)}$:

$$\phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) \leq C' d_m \log(n),$$

where C' is a constant depending on R. Therefore there exits a constant C'' depending on R such that

$$\epsilon_m^2 + \frac{1}{n}\phi_{f_m^*}^{(m)}(\epsilon_m, \lambda_m) + \frac{\lambda_m}{n} - \frac{\log(\lambda_m)}{n} \leq C'' \frac{d_m}{n}\log(n), \tag{19}$$

which gives the assertion.

Appendix D. Auxiliary Lemma

Lemma 8 *We have that*

$$\frac{\int |U_f^{(\lambda_m)} \tilde{h}_m| \mathbf{1}\{f : \|f\|_\infty \leq \epsilon\} \mathrm{GP}_m(\mathrm{d}f|\lambda_m)}{\mathrm{GP}_m(\{f : \|f\|_\infty \leq \epsilon\}|\lambda_m)} \leq \|\tilde{h}_m\|_{\mathcal{H}_m, \lambda_m}.$$

Proof Since the Gaussian process $W : \Omega \to L_\infty(\mathcal{X}_m)$ with the law $\mathrm{GP}_m(\cdot|\lambda_m)$ is measurable, the norm $\|\cdot\|_\infty$ is a measurable function, that is also true in the sense of Definition 3.1 of Hargé (2004) due to Corollaries 4.5 and 5.2 of Gross (1962). Here we utilize Theorem 3.4 of Hargé (2004) that is a particular infinite dimensional extension of Brascamp-Lieb inequality (Brascamp and Lieb, 1976). That gives

$$\frac{\int |U_f^{(\lambda_m)} \tilde{h}_m| \mathbf{1}\{f : \|f\|_\infty \leq \epsilon\} \mathrm{GP}_m(\mathrm{d}f|\lambda_m)}{\mathrm{GP}_m(\{f : \|f\|_\infty \leq \epsilon\}|\lambda_m)} \leq \int |U_f^{(\lambda_m)} \tilde{h}_m| \mathrm{GP}_m(\mathrm{d}f|\lambda_m).$$

The RHS is further bounded by

$$\sqrt{\int |U_f^{(\lambda_m)} \tilde{h}_m|^2 \mathrm{GP}_m(\mathrm{d}f|\lambda_m)} = \|\tilde{h}_m\|_{\mathcal{H}_m, \lambda_m}, \tag{20}$$

because $U_\cdot^{(\lambda_m)}$ is an isometry from $\mathcal{H}_{m,\lambda_m}$ to $L_2(\mathrm{GP}_m(\cdot|\lambda_m))$. ■

Remark 9 *The key proposition in the proof of Lemma 8 is Theorem 3.4 of Hargé (2004). As we have mentioned, the theorem is an infinite dimensional extension of Brascamp-Lieb inequality (Theorem 5.1 of Brascamp and Lieb (1976)). One big motivation of this line of researches is to prove the Gaussian correlation conjecture:*

$$\mu(A \cap B) \geq \mu(A)\mu(B)$$

where μ is any centered Gaussian measure on a separable Banach space and A and B are any two symmetric convex sets. There is a long history about this conjecture. Brascamp-Lieb inequality can be seen as an application of a particular case of the Gaussian correlation conjecture (see Hargé (1999)). See the survey by Li and Shao (2001) for details of the Gaussian correlation conjecture.