Supplementary material: Gaussian process nonparametric tensor estimator and its minimax optimality

Heishiro Kanagawa[†] KANAGAWAH.AB@M.TITECH.AC.JP

Taiji Suzuki[†,‡] SUZUKI.T.CT@M.TITECH.AC.JP

[†] Tokyo Institute of Technology, Tokyo 152-8552, JAPAN
[‡] PRESTO, Japan Science and Technological Agency (JST), JAPAN

Hayato Kobayashi[⋆] HAKOBAYA@YAHOO-CORP.JP

Nobuyuki Shimizu[⋆] NOBUSHIM@YAHOO-CORP.JP

Yukihiro Tagami[⋆] YUTAGAMI@YAHOO-CORP.JP

[⋆] Yahoo Japan Corporation, Tokyo 107-6211, JAPAN

In this supplementary material, we give the comprehensive proof and the generalized theorems. We consider a more general regression setting:

$$y_i = f^{\mathrm{o}}(x_i) + \epsilon_i, \tag{S-1}$$

where $f^{\mathrm{o}} : \mathcal{X} \to \mathbb{R}$ is the unknown true function. We suppose that the true function f^{o} is well approximated by $f^* = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f_r^{*(k)}$ (that is $f^{\mathrm{o}} \simeq f^*$). When $f^{\mathrm{o}} = f^*$, this generalized regression problem is equivalent to that in the main body. In that sense, the model (S-1) contains the model in the main body as a special setting $f^{\mathrm{o}} = f^*$.

A. Noise Assumption and PAC-Bayesian Bound

Here we remind our assumption on the noise ϵ_i (Assumption 1). There are a lot of choices of noise conditions to establish PAC-Bayesian bounds. Here we employ a condition with which we can utilize an extension of Stein's identity. Now define a function

$$m_\epsilon(z) := -\mathrm{E}[\epsilon_1 \mathbf{1}\{\epsilon_1 \le z\}] = -\int_{-\infty}^{z} y \mathrm{d}F_\epsilon(y) = \int_{z}^{\infty} y \mathrm{d}F_\epsilon(y),$$

where $F_\epsilon(z) = P(\epsilon_1 \le z)$ is the cumulative distribution function of the noise, and $\mathbf{1}\{\cdot\}$ is the indicator function. Since $\mathrm{E}[\epsilon_1] = 0$, one can check that $m_\epsilon(z)$ is non-negative and achieves its maximum at 0: $\max_{z\in\mathbb{R}} m_\epsilon(z) = m_\epsilon(0) = \mathrm{E}[|\epsilon_1|]/2$. Then we impose the following assumption on the noise ξ.

Assumption A.1. $\mathrm{E}[\epsilon_1^2] < \infty$ and the measure $m_\epsilon(z)\mathrm{d}z$ is absolutely continuous with respect to the density function $\mathrm{d}F_\epsilon(z)$ with a bounded Radon-Nikodym derivative, i.e., there exists a bounded function $g_\xi : \mathbb{R} \to \mathbb{R}_+$ such that

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

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

- If ϵ_1 obeys the Gaussian $\mathcal{N}(0, \sigma^2)$, then $g_\epsilon(z) = \sigma^2$,
- If ϵ_1 obeys the uniform distribution on $[-a, a]$, then $g_\epsilon(z) = \max(a^2 - z^2, 0)/2$.

Under Assumption 1, Theorem 1 of (Dalalyan & 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 A.1. *Suppose Assumption 1 is satisfied and $\beta \geq 4\|g_\epsilon\|_\infty$. Then for all probability measure ρ that is absolutely continuous with respect to Π, we have*

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

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

B. Upper bound analysis

Let $\mathcal{H}_{(r,k),\lambda}$ be the “scaled” version of an RKHS $\mathcal{H}_{(r,k)}$. That is $\mathcal{H}_{(r,k),\lambda}$ it the RKHS associated with the kernel $\tilde{k}_{m,\lambda_r^{(k)}} = k_{r,k}/\lambda_r^{(k)}$.

The quantitative evaluation of the mass around the true function is given by the following *concentration function* (van der Vaart & van Zanten, 2011; 2008a):

$$\begin{aligned}\phi^{(r,k)}_{f_r^{*(k)}}(\epsilon,L,\lambda) := &\inf_{h\in\mathcal{H}_{(r,k)}:\|h-f_r^{*(k)}\|_\infty\leq\epsilon}\left(\|h\|^2_{\mathcal{H}_{(r,k),\lambda}}\vee 1\right) - \log \mathrm{GP}_{(r,k)}(\{f:\|f\|_n\leq\epsilon/\sqrt{2}\}|\lambda),\\ &-\log\mathrm{GP}_{(r,k)}(\{f:\|f\|_\infty\leq L/\sqrt{2}\}|\lambda),\end{aligned} \tag{S-3}$$

where $a\vee b := \max(a,b)$. It can be shown that $\phi^{(r,k)}_{f_r^{*(k)}}(\epsilon,\lambda)$ equals $-\log\mathrm{GP}_{(r,k)}(\{f:\|f_r^{*(k)}-f\|_\infty\leq\epsilon\}|\lambda)$ up to constants (van der Vaart & van Zanten, 2008b).

B.1. Generalized upper bound

Define

$$\check{S} := \{r \mid 1\leq r\leq d_{\max},\ \exists k \text{ s.t. } f_r^{*(k)}\notin\mathcal{H}_{(r,k)}\}.$$

Theorem B.1 (Convergence rate of GP-Tensor)**.** *Let*

$$\hat{R}_{K,\max} := \left(R + 2\max_{r,k} L_{(r,k)}\right)^{2(K-1)},$$

and set $c_r = 1$ if $r\notin\check{S}$, and $c_r = K$ otherwise. Then, there exists a constant C_1 depending on only β such that the convergence rate of Bayesian-MKL is bounded as

$$\begin{aligned}&\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f}-f^{\mathrm{o}}\|_n^2\right] \leq 2\|f^{\mathrm{o}}-f^*\|_n^2\\ &+C_1\inf_{\epsilon_{(r,k)},L_{(r,k)},\lambda_{(r,k)}>0;\epsilon_{(r,k)}\leq L_{(r,k)}}\left\{\sum_{r=1,\dots,d^*}c_r\sum_{k=1}^K\left(\hat{R}_{K,\max}\epsilon^2_{(r,k)}+\frac{1}{n}\phi^{(r,k)}_{f_r^{*(k)}}(\epsilon_{(r,k)},L_{(r,k)},\lambda_{(r,k)})+\frac{\lambda_{(r,k)}}{n}-\frac{\log(\lambda_{(r,k)})}{n}\right)\right.\\ &\left.+\hat{R}_{K,\max}\left(\sum_{r\in\check{S}}\sqrt{\sum_{k=1}^K\epsilon^2_{(r,k)}}\right)^2\right\}+\frac{d^*}{n}\log\left(\frac{1}{\zeta(1-\zeta)}\right).\end{aligned} \tag{S-4}$$

B.2. Proof of Theorem B.1

We go along the same line with (Suzuki, 2012). Fix $\epsilon_{(r,k)},\lambda_{(r,k)},L_{(r,k)}>0$ for $r=1,\dots,d_{\max}$ and $k=1,\dots,K$. The typical approach to prove the theorem is that we substitute some “dummy” posterior distribution into ρ in Eq. (S-2) of Theorem A.1 (the PAC-Bayes bound). For $\epsilon_{(r,k)}>0$ and $L_{(r,k)}>0$, define a set $\mathcal{S}_{(r,k)}$ of a function as

$$\mathcal{S}_{(r,k)} := \{f:\mathcal{X}\to\mathbb{R}\mid \|f\|_n\leq\epsilon_{(r,k)},\|f\|_\infty\leq L_{(r,k)}\}.$$

We define $\tilde{h}_{(r,k)} \in \mathcal{H}_{(r,k)}$ for each r, k so that it is an approximation of $f_r^{*(k)}$ as follows. If $f_r^{*(k)} \in \mathcal{H}_{(r,k)}$, then we take $\tilde{h}_{(r,k)}$ as $\tilde{h}_{(r,k)} = f_r^{*(k)}$. Otherwise, we take $\tilde{h}_{(r,k)} \in \mathcal{H}_{(r,k)}$ such that

$$\|\tilde{h}_{(r,k)}\|^2_{\mathcal{H}_{(r,k)},\lambda_{(r,k)}} \leq 2 \inf_{h \in \mathcal{H}_{(r,k)}: \|h - f_r^{*(k)}\|_\infty \leq \epsilon_{(r,k)}} \|h\|^2_{\mathcal{H}_{(r,k)},\lambda_{(r,k)}}, \tag{S-5}$$

$$\tilde{h}_{(r,k)} - f_r^{*(k)} \in \mathcal{S}_{(r,k)}. \tag{S-6}$$

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

$$\rho(\mathrm{d}f) = \prod_{r=1,\dots,d^*} \prod_{k=1}^K \frac{\int_{\frac{\lambda_{(r,k)}}{2} \leq \tilde{\lambda}_{(r,k)} \leq \lambda_{(r,k)}} \frac{\mathrm{GP}_{(r,k)}^{W+\tilde{h}_{(r,k)}}(\mathrm{d}f_r^{(k)}|\tilde{\lambda}_{(r,k)}) \mathbf{1}\{f_r^{(k)} - \tilde{h}_{(r,k)} \in \mathcal{S}_{(r,k)}\}}{\mathrm{GP}_{(r,k)}(\mathcal{S}_{(r,k)}|\tilde{\lambda}_{(r,k)})} \mathcal{G}(\mathrm{d}\tilde{\lambda}_{(r,k)})}{\mathcal{G}(\{\tilde{\lambda}_{(r,k)} : \frac{\lambda_{(r,k)}}{2} \leq \tilde{\lambda}_{(r,k)} \leq \lambda_{(r,k)}\})}$$
$$\times \prod_{r > d^*} \prod_{k=1}^K \delta_0(\mathrm{d}f_r^{(k)}).$$

According to the proof of Theorem 3 in Suzuki (2012), it is shown that ρ is absolutely continuous with respect to the prior Π. Therefore, we may apply Theorem A.1.

Let $R := \max_{r,k}\{\|f_r^{*(k)}\|_\infty\}$. Then, since $\tilde{h}_{(r,k)}$ satisfies $\|\tilde{h}_{(r,k)} - f_r^{*(k)}\|_\infty \leq L_{(r,k)}$ by the definition (Eq. (S-6)), it holds that

$$\|\tilde{h}_{(r,k)}\|_\infty \leq \epsilon_{(r,k)} + R.$$

Similarly, for all $f_r^{(k)}$ in the support of ρ, we have that $\|f_r^{(k)} - \tilde{h}_{(r,k)}\|_\infty \leq L_{(r,k)}$ and by assuming $\epsilon_{(r,k)} \leq L_{(r,k)}$,

$$\|f_r^{(k)}\|_\infty \leq \epsilon_{(r,k)} + L_{(r,k)} + R \leq 2L_{(r,k)} + R. \tag{S-7}$$

Note that for $f = \sum_{r=1}^{d_{\max}} \prod_{k=1}^K f_r^{(k)}$ and $f^* = \sum_{r=1}^{d_{\max}} \prod_{k=1}^K f_r^{*(k)}$, it holds that

$$\int \|f - f^{\mathrm{o}}\|_n^2 \mathrm{d}\rho(f) \leq 2 \int \|f - f^*\|_n^2 \mathrm{d}\rho(f) + 2 \int \|f^* - f^{\mathrm{o}}\|_n^2 \mathrm{d}\rho(f).$$

Thus we just need to bound the first term of the RHS:

$$\begin{aligned}
&\int \|f - f^*\|_n^2 \mathrm{d}\rho(f) \\
&= \int \left\| \sum_{r=1}^{d_{\max}} \left(\prod_{k=1}^K f_r^{(k)} - \prod_{k=1}^K f_r^{*(k)} \right) \right\|_n^2 \mathrm{d}\rho(f) \\
&= \int \left\| \sum_{r=1}^{d^*} \left(\prod_{k=1}^K f_r^{(k)} - \prod_{k=1}^K f_r^{*(k)} \right) \right\|_n^2 \mathrm{d}\rho(f) \\
&= \sum_{r=1}^{d^*} \int \left\| \prod_{k=1}^K f_r^{(k)} - \prod_{k=1}^K f_r^{*(k)} \right\|_n^2 \mathrm{d}\rho(f) \\
&\quad - 2 \sum_{r \neq r' : 1 \leq r, r' \leq d^*} \int \left\langle \prod_{k=1}^K f_r^{(k)} - \prod_{k=1}^K f_r^{*(k)}, \prod_{k=1}^K f_{r'}^{(k)} - \prod_{k=1}^K f_{r'}^{*(k)} \right\rangle_n \mathrm{d}\rho(f).
\end{aligned}$$

The first term of the RHS is evaluated by

$$\sum_{r=1}^{d^*} \int \left\| \prod_{k=1}^K f_r^{(k)} - \prod_{k=1}^K f_r^{*(k)} \right\|_n^2 \mathrm{d}\rho(f)$$

$$
\begin{aligned}
=&\sum_{r=1}^{d^*}\int\left\|\sum_{k=1}^{K}\left(\prod_{k'=1}^{k-1}f_r^{(k')}\right)\left(f_r^{(k)}-f_r^{*(k)}\right)\left(\prod_{k'=k+1}^{K}f_r^{*(k')}\right)\right\|_n^2\mathrm{d}\rho(f)\\
=&\sum_{r=1}^{d^*}\sum_{k=1}^{K}\sum_{\tilde{k}=1}^{K}\int\Bigg\langle\left(\prod_{k'=1}^{k-1}f_r^{(k')}\right)\left(f_r^{(k)}-f_r^{*(k)}\right)\left(\prod_{k'=k+1}^{K}f_r^{*(k')}\right),\\
&\left(\prod_{k'=1}^{\tilde{k}-1}f_r^{(k')}\right)\left(f_r^{(\tilde{k})}-f_r^{*(\tilde{k})}\right)\left(\prod_{k'=\tilde{k}+1}^{K}f_r^{*(k')}\right)\Bigg\rangle_n\mathrm{d}\rho(f)\\
=&\sum_{r=1}^{d^*}\sum_{k=1}^{K}\sum_{\tilde{k}=1}^{K}\int\Bigg\langle\left(\prod_{k'=1}^{k-1}f_r^{(k')}\right)\left(f_r^{(k)}-f_r^{*(k)}\right)\left(\prod_{k'=k+1}^{K}f_r^{*(k')}\right),\\
&\left(\prod_{k'=1}^{\tilde{k}-1}f_r^{(k')}\right)\left(f_r^{(\tilde{k})}-f_r^{*(\tilde{k})}\right)\left(\prod_{k'=\tilde{k}+1}^{K}f_r^{*(k')}\right)\Bigg\rangle_n\mathrm{d}\rho(f).
\end{aligned}
$$

If $k \neq \tilde{k}$ and $r \notin \check{S}$, then the summand of the RHS is 0, otherwise the summand is bounded by $\frac{1}{2}\sum_{k''=k,\tilde{k}}\int\left\|\left(\prod_{k'=1}^{k''-1}f_r^{(k')}\right)\left(f_r^{(k'')}-f_r^{*(k'')}\right)\left(\prod_{k'=k''+1}^{K}f_r^{*(k')}\right)\right\|_n^2\mathrm{d}\rho(f)$. Hence, by setting $c_r = 1$ if $r \notin \check{S}$, and $c_r = K$ otherwise, then Lemma B.1 and Eq. (S-7) give an upper bound of the RHS as

$$
\begin{aligned}
&\sum_{r=1}^{d^*}c_r\sum_{k=1}^{K}\int\left\|\left(\prod_{k'=1}^{k-1}f_r^{(k')}\right)\left(f_r^{(k)}-f_r^{*(k)}\right)\left(\prod_{k'=k+1}^{K}f_r^{*(k')}\right)\right\|_n^2\mathrm{d}\rho(f)\\
\le&\sum_{r=1}^{d^*}c_r\sum_{k=1}^{K}\prod_{k'\neq k}(R+2L_{(r,k')})^2\int\|f_r^{(k)}-f_r^{*(k)}\|_n^2\mathrm{d}\rho(f)\\
\le&(R+2\max_{r,k}L_{(r,k)})^{2(K-1)}\sum_{r=1}^{d^*}c_r\sum_{k=1}^{K}2\int(\|f_r^{(k)}-\tilde{h}_{(r,k)}\|_n^2+\|\tilde{h}_{(r,k)}-f_r^{*(k)}\|_n^2)\mathrm{d}\rho(f)\\
\le&4\left(R+2\max_{r,k}L_{(r,k)}\right)^{2(K-1)}\sum_{r=1}^{d^*}c_r\sum_{k=1}^{K}\epsilon_{(r,k)}^2.
\end{aligned}
\tag{S-8}
$$

On the other hand, using Lemma B.1 again, an analogous reasoning gives a bound of the second term as

$$
\begin{aligned}
&\left|\int\left\langle\prod_{k=1}^{K}f_r^{(k)}-\prod_{k=1}^{K}f_r^{*(k)},\prod_{k=1}^{K}f_{r'}^{(k)}-\prod_{k=1}^{K}f_{r'}^{*\,(k)}\right\rangle_n\mathrm{d}\rho(f)\right|\\
=&\left|\left\langle\prod_{k=1}^{K}\tilde{h}_{(r,k)}-\prod_{k=1}^{K}f_r^{*(k)},\prod_{k=1}^{K}\tilde{h}_{(r',k)}-\prod_{k=1}^{K}f_{r'}^{*\,(k)}\right\rangle_n\right|\\
\le&\begin{cases}0, & (r\notin\check{S}\text{ or }r'\notin\check{S}),\\ \left(R+2\max_{\tilde{r},\tilde{k}}L_{(\tilde{r},\tilde{k})}\right)^{2(K-1)}\sqrt{\sum_{k=1}^{K}\epsilon_{(r,k)}^2}\sqrt{\sum_{k=1}^{K}\epsilon_{(r',k)}^2}, & (\text{otherwise}).\end{cases}
\end{aligned}
\tag{S-9}
$$

Now define

$$
\widehat{\phi^{(r,k)}}_{f_r^{*(k)}}(\epsilon_{(r,k)},L_{(r,k)},\lambda):=\inf_{h\in\mathcal{H}_{(r,k)}:\|h-f_r^{*(k)}\|_\infty\le\epsilon}\left(\|h\|_{\mathcal{H}_{(r,k),\lambda}}^2\vee 1\right)-\log \mathrm{GP}_{(r,k)}(\mathcal{S}_{(r,k)}|\lambda).
$$

Then, along with the proof of Theorem 3 in Suzuki (2012), the KL-divergence between the "posterior" ρ and the prior Π is bounded as

$$
\frac{1}{n}\mathcal{K}(\rho,\Pi)
$$

$$\leq C_1' \sum_{r=1}^{d^*} \sum_{k=1}^{K} \left(\frac{1}{n} \widehat{\phi^{(r,k)}}_{f_r^{*(k)}}(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}) + \frac{1}{n}\lambda_{(r,k)} - \frac{1}{n}\log\left(\frac{\lambda_{(r,k)}}{2}\right)\right) + \frac{d^*}{n}\log\left(\frac{1}{\zeta(1-\zeta)}\right), \tag{S-10}$$

where C_1' is a universal constant. Here, since both of the sets $\{f : \|f\|_n \leq \epsilon\}$ and $\{f : \|f\|_\infty \leq L\}$ are convex and symmetric, we obtain by Proposition B.2 that

$$-\log \mathrm{GP}_{(r,k)}(\mathcal{S}_{(r,k)}|\lambda) \leq -\log \mathrm{GP}_{(r,k)}(\{f : \|f\|_n \leq \epsilon_{(r,k)}/\sqrt{2}\}|\lambda) - \log \mathrm{GP}_{(r,k)}(\{f : \|f\|_\infty \leq L_{(r,k)}/\sqrt{2}\}|\lambda).$$

Thus

$$\widehat{\phi^{(r,k)}}_{f_r^{*(k)}}(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}) \leq \phi^{(r,k)}{}_{f_r^{*(k)}}(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}). \tag{S-11}$$

Finally, combining Eq. (S-9), Eq. (S-8), and Eq. (S-10) with Eq. (S-11), we obtain the assertion.

Lemma B.1. For $f(x) = \prod_{k=1}^K f_k(x) : \mathcal{X} \mapsto \mathbb{R}$ such that $\|f_k\|_\infty \leq R$ $(\forall k)$, it holds that

$$\|f\|_n^2 \leq R^{K-1}\|f_k\|_n^2,$$

for all $k = 1, \ldots, K$. In addition, for $f'(x) = \prod_{k=1}^K f_k'(x) : \mathcal{X} \mapsto \mathbb{R}$ such that $\|f_k'\|_\infty \leq R$ $(\forall k)$, it holds that

$$\langle f, f' \rangle_n \leq R^{K-1}\|f_k\|_n \|f_k'\|_n,$$

for all $k = 1, \ldots, K$.

Proof.

$$\begin{aligned}\|f\|_n^2 &= \frac{1}{n}\sum_{i=1}^n \prod_{k=1}^K f_k(x_i)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^n f_k(x_i)^2 \prod_{k' \neq k} \max_{i=1,\ldots,n}\{f_{k'}(x_i)^2\} \leq R^{K-1}\frac{1}{n}\sum_{i=1}^n f_k(x_i)^2.\end{aligned}$$

Using the same reasoning, we obtain the second assertion by noticing $\frac{1}{n}\sum_{i=1}^n |f_k(x_i) f_k'(x_i)| \leq \|f_k\|\|f_k'\|$.

□

Schechtman et al. (1998); Li (1999) showed the following theorem.

Proposition B.2. Let ρ be a centered Gaussian measure on a separable Banach space E. Then for any $0 < \lambda < 1$, any symmetric, convex sets A and B in E,

$$\rho(A \cap B)\rho(\lambda^2 A + (1-\lambda^2)B) \geq \rho(\lambda A)\rho((1-\lambda^2)^{1/2}B).$$

In particular,

$$\rho(A \cap B) \geq \rho(\lambda A)\rho((1-\lambda^2)^{1/2}B).$$

Schechtman et al. (1998) probed the above statement for $\lambda = 1/\sqrt{2}$ and $E = \mathbb{R}^n$, and Li (1999) extended the results as above.

B.3. Proof of Theorems 1 and 2 in the main body

We just need to bound the following term for each r, k:

$$\hat{R}_{K,\max}\epsilon_{(r,k)}^2 + \frac{1}{n}\phi^{(r,k)}_{f_r^{*(k)}}(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}) + \frac{1}{n}\lambda_{(r,k)} - \frac{1}{n}\log\left(\frac{\lambda_{(r,k)}}{2}\right), \tag{S-12}$$

by choosing an appropriate $\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}$ such that $\epsilon_{(r,k)} \leq L_{(r,k)}$.

By the definition, we have

$$\|f_r^{*(k)}\|_{\theta,\infty,\mathcal{H}_{(r,k)}} = \sup_{t>0} \inf_{h_r^{(k)} \in \mathcal{H}_{(r,k)}} \{t^{-\theta}\|f_r^{*(k)} - h_r^{(k)}\|_\infty + t^{1-\theta}\|h_r^{(k)}\|_{\mathcal{H}_{(r,k)}}\}.$$

With a slight abuse of notation, we denote by $\|f_r^{*(k)}\|_{\theta,\infty} = \|f_r^{*(k)}\|_{\theta,\infty,\mathcal{H}_{(r,k)}}$. If $\inf_{h_r^{(k)} \in \mathcal{H}_{(r,k)}} \|f_r^{*(k)} - h_r^{(k)}\|_\infty > 0$, then the term $t^{-\theta}\|f_r^{*(k)} - h_r^{(k)}\|_\infty$ can be arbitrary large. Therefore the assumption $R \geq \|f_r^{*(k)}\|_{\theta,\infty}$ ensures that there exists $h_r^{(k)} \in \mathcal{H}_{(r,k)}$ such that $\|f_r^{*(k)} - h_r^{(k)}\|_\infty \leq \epsilon$ for all $\epsilon > 0$. Using this, we evaluate the RKHS norm of the approximator: $\inf_{h \in \mathcal{H}_{(r,k)}:\|h - f_r^{*(k)}\|_\infty \leq \epsilon_{(r,k)}} \|h\|^2_{\mathcal{H}_{(r,k)}}$. For all $t > 0$, there exists $h_r^{(k)}{}_{[t]} \in \mathcal{H}_{(r,k)}$ such that $2\|f_r^{*(k)}\|_{\theta,\infty} \geq t^{-\theta}\|f_r^{*(k)} - h_r^{(k)}{}_{[t]}\|_\infty + t^{1-\theta}\|h_r^{(k)}{}_{[t]}\|_{\mathcal{H}_{(r,k)}}$. This gives $2\|f_r^{*(k)}\|_{\theta,\infty} \geq t^{-\theta}\|f_r^{*(k)} - h_r^{(k)}{}_{[t]}\|_\infty$ so that we have $t \geq 2^{-\frac{1}{\theta}}\|f_r^{*(k)}\|_{\theta,\infty}{}^{-\frac{1}{\theta}}\|f_r^{*(k)} - h_r^{(k)}{}_{[t]}\|_\infty^{\frac{1}{\theta}}$, and hence $2\|f_r^{*(k)}\|_{\theta,\infty} \geq t^{1-\theta}\|h_r^{(k)}{}_{[t]}\|_{\mathcal{H}_{(r,k)}}$ yields

$$\|h_r^{(k)}{}_{[t]}\|_{\mathcal{H}_{(r,k)}} \leq t^{-(1-\theta)}2\|f_r^{*(k)}\|_{\theta,\infty} \leq 2^{\frac{1}{\theta}}\|f_r^{*(k)}\|_{\theta,\infty}{}^{\frac{1}{\theta}}\|f_r^{*(k)} - h_r^{(k)}{}_{[t]}\|_\infty^{-\frac{1-\theta}{\theta}}.$$

Therefore we have that

$$\inf_{h \in \mathcal{H}_{(r,k)}:\|h - f_r^{*(k)}\|_\infty \leq \epsilon_{(r,k)}} \|h\|^2_{\mathcal{H}_{r,k}} \leq 2^{\frac{2}{\theta}}\|f_r^{*(k)}\|_{\theta,\infty}{}^{\frac{2}{\theta}}\epsilon_{(r,k)}^{-\frac{2(1-\theta)}{\theta}} \leq (2R)^{\frac{2}{\theta}}\epsilon_{(r,k)}^{-\frac{2(1-\theta)}{\theta}}, \tag{S-13}$$

because for all $\epsilon > 0$ there exists t such that $\|f_r^{*(k)} - h_r^{(k)}{}_{[t]}\|_\infty \leq \epsilon$.

Setting (i): From now on, we assume that $1 - \theta - s_{(r,k)} \geq 0$. Here, the metric entropy condition (Assumption 2) gives that there exists C_0' such that

$$-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_n \leq \epsilon\})) \leq C_0'\epsilon^{-\frac{2s_{(r,k)}}{1-s_{(r,k)}}}$$

(Kuelbs & Li, 1993; Li & Shao, 2001). Similary, Assumption 6 gives that there exists C_1' such that

$$-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq L\})) \leq C_1'L^{-\frac{2\tilde{s}_{(r,k)}}{1-\tilde{s}_{(r,k)}}}.$$

This and Eq. (S-13) give that

$$\phi^{(r,k)}_{f_r^{*(k)}}(\epsilon_{(r,k)}, \lambda_{(r,k)}) \leq (2R)^{\frac{2}{\theta}}\lambda_{(r,k)}\epsilon_{(r,k)}^{-\frac{2(1-\theta)}{\theta}} + C_0'\left(\frac{\sqrt{\lambda_{(r,k)}}\epsilon_{(r,k)}}{\sqrt{2}}\right)^{-\frac{2s_{(r,k)}}{1-s_{(r,k)}}} + C_1'\left(\frac{\sqrt{\lambda_{(r,k)}}L_{(r,k)}}{\sqrt{2}}\right)^{-\frac{2\tilde{s}_{(r,k)}}{1-\tilde{s}_{(r,k)}}} \tag{S-14}$$

where we used

$$\begin{aligned}
&\|f\|^2_{\mathcal{H}_{(r,k),\lambda}} = \lambda\|f\|^2_{\mathcal{H}_{(r,k)}},\\
&-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_n \leq \epsilon_{(r,k)}\}|\lambda_{(r,k)})) = -\log(\mathrm{GP}_{r,k}(\{f : \|f\|_n \leq \sqrt{\lambda_{(r,k)}}\epsilon_{(r,k)}\})),\\
&-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq L_{(r,k)}\}|\lambda_{(r,k)})) = -\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq \sqrt{\lambda_{(r,k)}}L_{(r,k)}\})).
\end{aligned}$$

Now $\lambda_{(r,k)} = (R \vee 1)^{-\frac{2(1-s_{(r,k)})}{\theta}}\epsilon_{(r,k)}^{\frac{2(1-\theta-s_{(r,k)})}{\theta}}$ balances the first two terms in the right hand side of Eq. (S-14) up to constants. In addition to $\lambda_{(r,k)}$, we set $L_{(r,k)} = (R \vee 1)^{\frac{1-s_{(r,k)}}{\theta}}$. With this $\lambda_{(r,k)}$and $L_{(r,k)}$, the RHS of Eq. (S-12) is bounded as

$$\begin{aligned}
&\hat{R}_{K,\max}\epsilon^2_{(r,k)} + \frac{1}{n}\phi^{(r,k)}_{f_r^{*(k)}}(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}) + \frac{\lambda_{(r,k)}}{n} - \frac{\log(\lambda_{(r,k)})}{n}\\
\leq&\hat{R}_{K,\max}\epsilon^2_{(r,k)} + \frac{(2^{\frac{2}{\theta}} + C_0'2^{\frac{s_{(r,k)}}{1-s_{(r,k)}}})}{n}(R \vee 1)^{\frac{2s_{(r,k)}}{\theta}}\epsilon_{(r,k)}^{-\frac{2s_{(r,k)}}{\theta}}
\end{aligned}$$

$$+\frac{C_1' 2^{\frac{2\tilde{s}_{(r,k)}}{1-\tilde{s}_{(r,k)}}}}{n}\epsilon_{(r,k)}^{-\frac{2(1-\theta-s_{(r,k)})\tilde{s}_{(r,k)}}{\theta(1-\tilde{s}_{(r,k)})}}$$
$$+\frac{\epsilon_{(r,k)}^{\frac{1-\theta-s_{(r,k)}}{\theta}}}{n}-\frac{\log(\epsilon_{(r,k)}^{\frac{1-\theta-s_{(r,k)}}{\theta}})}{n}. \tag{S-15}$$

Here, we set $\epsilon_{(r,k)}^2 = n^{-\frac{1}{1+s_{(r,k)}/\theta}}$. When $1-\theta-s_{(r,k)} \geq 0$, by the assumption that $\tilde{s}_{(r,k)} \leq \frac{s_{(r,k)}}{1-\theta}$,

$$\epsilon_{(r,k)}^{-\frac{2(1-\theta-s_{(r,k)})\tilde{s}_{(r,k)}}{\theta(1-\tilde{s}_{(r,k)})}} \leq \epsilon_{(r,k)}^{-\frac{2s_{(r,k)}}{\theta}}. \tag{S-16}$$

Therefore, by applying Eq. (S-16) to the RHS of Eq. (S-15), the RHS of Eq. (S-15) is bounded by

$$C\left(\hat{R}_{K,\max} \vee (R\vee 1)^{\frac{2s_{(r,k)}}{\theta}}\right) n^{-\frac{1}{1+s_{(r,k)}/\theta}}. \tag{S-17}$$

where C is a constant independent of n, R.

Setting (ii): As for the situation, $1-\theta-s_{(r,k)} \leq 0$, we also use the same setting. Then $\sqrt{\lambda_{(r,k)}}L_{(r,k)} \geq 1$. Thus we have another bound like

$$-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq \sqrt{\lambda_{(r,k)}}L_{(r,k)}\})) \leq -\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq 1\})) \leq -\log(c_1).$$

Then along with the same reasoning as for the situation $1-\theta-s_{(r,k)} \geq 0$, the same upper bound of Eq. (S-12) as Eq. (S-17) with a different constant. This concludes the proof of Theorem 2 by substituting the setting $(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)})$ as described above into Eq. (S-4) in the statement of Theorem B.1.

Theorem 1 is proved by the same reasoning, but it should be noticed that $\check{S} = \emptyset$, $\theta = 1$ and $(R\vee 1)^{\frac{2s_{(r,k)}}{\theta}} \leq (R\vee 1)^2$ because of $s_{(r,k)} < 1$.

C. Proof of minimax lower bound (Theorem 4)

Proof. (Theorem 4) The δ-packing number $M(\mathcal{G}, \delta, \|\cdot\|)$ of a function class $\mathcal{G}$ with respect to a norm $\|\cdot\|$ is the largest number of functions $\{f_1, \dots, f_{\mathcal{M}}\} \subseteq \mathcal{G}$ such that $\|f_i - f_j\| \geq \delta$ for all $i \neq j$. Generally, it holds that

$$N(\mathcal{G}, \delta/2, \|\cdot\|) \leq M(\mathcal{G}, \delta, \|\cdot\|) \leq N(\mathcal{G}, \delta, \|\cdot\|). \tag{S-18}$$

For a given $\delta_n > 0$ and $\varepsilon_n > 0$, let Q be the δ_n packing number $M(\mathscr{H}_{(d^*,K)}(R), \delta_n, L_2(P_\mathcal{X}))$ of $\mathscr{H}_{(d^*,K)}(R)$ and N be the ε_n covering number $N(\mathscr{H}_{(d^*,K)}(R), \varepsilon_n, L_2(P_\mathcal{X}))$ of $\mathscr{H}_{(d^*,K)}(R)$. (Raskutti et al., 2010) utilized the techniques developed by (Yang & Barron, 1999) to show the following inequality in their proof of Theorem 2(b) :

$$\begin{aligned}\inf_{\hat{f}} \sup_{f^*\in\mathscr{H}_{(d^*,K)}(R)} \mathrm{E}[\|\hat{f}-f^*\|^2_{L_2(P_\mathcal{X})}] &\geq \inf_{\hat{f}} \sup_{f^*\in\mathscr{H}_{(d^*,K)}(R)} \frac{\delta_n^2}{2} P[\|\hat{f}-f^*\|^2_{L_2(P_\mathcal{X})} \geq \delta_n^2/2] \\ &\geq \frac{\delta_n^2}{2}\left(1-\frac{\log(N)+\frac{n}{2\sigma^2}\varepsilon_n^2+\log(2)}{\log(Q)}\right).\end{aligned}$$

Thus by taking δ_n and ε_n to satisfy

$$\frac{n}{2\sigma^2}\varepsilon_n^2 \leq \log(N), \tag{S-19a}$$
$$8\log(N) \leq \log(Q), \tag{S-19b}$$
$$4\log(2) \leq \log(Q), \tag{S-19c}$$

the minimax rate is lower bounded by $\frac{\delta_n^2}{4}$.

From now on, we are going to evaluate $\log(N)$ and $\log(Q)$ in terms of δ_n and ε_n. For all $f, f' \in \mathscr{H}_{(d^*,K)}(R)$, it holds that

$$\|f - f'\|^2_{L_2(P_\mathcal{X})} = \left\| \sum_{r=1}^{d^*} (\prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} {f'}_r^{(k)}) \right\|^2_{L_2(P_\mathcal{X})}$$
$$= \sum_{r=1}^{d^*} \left\| \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} {f'}_r^{(k)} \right\|^2_{L_2(P_\mathcal{X})}$$

by the construction of $L_2(P_\mathcal{X})$ and the assumption that $\mathrm{E}[f_r^{(k)}(X)] = 0$ for all $f_r^{(k)} \in \mathcal{H}_{(r,k)}$.

To evaluate the covering number and packing numbers, we construct a packing sets on the "sphere" of each $\mathcal{H}_{(r,k)}$. Since $\mathcal{X}_k$ is a compact metric space and $k_{(r)}k$ is continuous, Mercer's theorem gives the orthogonal decomposition of the kernel function $k_{(r,k)}$ as

$$k_{(r,k)}(x, x') = \sum_{i=1}^{\infty} \mu_{(r,k),i} \psi_{(r,k),i}(x) \psi_{(r,k),i}(x'), \tag{S-20}$$

where the convergence is absolute and uniform, $\{\psi_{(r,k),i}\}_{i=1}^{\infty}$ forms an orthonormal system and $\mu_{(r,k),i} \geq 0$ is the i-th eigen-value (see Theorem 4.49 in Steinwart & Christmann (2008) for example). We assume that $\mu_{(r,k),1} \geq \mu_{(r,k),2} \geq \cdots$. As in Assumption 7, there exists $\widehat{f}_r^{(k)} \in \mathcal{B}_{\mathcal{H}_{(r,k)}}$ such that $\|\widehat{f}_r^{(k)}\|_{L_2(P_{\mathcal{X}_k})} \geq c_1$. Without loss of generality, we may assume that $\widehat{f}_r^{(k)} = \sqrt{\mu_{(r,k),1}} \psi_{(r,k),1}$ because

$$\sqrt{\mu_{(r,k),1}} \psi_{(r,k),1} = \underset{f \in \mathcal{B}_{\mathcal{H}_{(r,k)}}}{\operatorname{argmax}} \|f\|_{L_2(P_\mathcal{X})}.$$

This can be seen by the relation $\|f\|^2_{\mathcal{H}_{(r,k)}} = \sum_{i=1}^n \int\int f(x) f(x') \psi_{(r,k),i}(x) \psi_{(r,k),i}(x') / \mu_{(r,k),i} \mathrm{d}P_\mathcal{X}(x) \mathrm{d}P_\mathcal{X}(x')$. Now, we consider a subspace which is perpendicular to $\widehat{f}_r^{(k)}$. Let $\mathcal{H}_{\perp,(r,k)} := \{f \in \mathcal{H}_{(r,k)} \mid \langle f, \widehat{f}_r^{(k)} \rangle_{L_2(P_\mathcal{X})} = 0\}$. Then, by the orthogonal decomposition (S-20) and the Mercer representation of RKHSs (Theorem 4.51 of Steinwart & Christmann (2008)), the space $\mathcal{H}_{\perp,(r,k)}$ can be represented by

$$\mathcal{H}_{\perp,(r,k)} = \left\{ \sum_{i=2}^{\infty} \alpha_i \psi_{(r,k),i} \mid \sum_{i=2}^{\infty} \alpha_i^2 / \mu_{(r,k),i} < \infty \right\}$$

where $0/0$ is defined as 0. $\mathcal{H}_{\perp,(r,k)}$ is also an RKHS with a kernel function

$$k_{\perp,(r,k)}(x, x') = \sum_{i=2}^{\infty} \mu_{(r,k),i} \psi_{(r,k),i}(x) \psi_{(r,k),i}(x'),$$

and $\|f\|_{\mathcal{H}_{\perp,(r,k)}} = \|f\|_{\mathcal{H}_{(r,k)}}$ for all $f \in \mathcal{H}_{\perp,(r,k)}$. Now, we evaluate the covering number of $\mathcal{H}_{\perp,(r,k)}$. Proposition C.3 with Assumption 7 gives that $\mu_{(r,k),i} \sim i^{-1/s_{(r,k)}}$. Thus, we again use Proposition C.3 to obtain that

$$\log N(\mathcal{B}_{\mathcal{H}_{\perp,(r,k)}}, \epsilon, L_2(P_{\mathcal{X}_{(r,k)}})) \sim \epsilon^{-2s_{(r,k)}}.$$

Let $g_{[j]}$ $(j = 1, \ldots, M_{(r,k)})$ be the packing set that gives the packing number $M_{(r,k)} = M(\mathcal{B}_{\mathcal{H}_{\perp,(r,k)}}, \epsilon, L_2(P_{\mathcal{X}_{(r,k)}}))$. Note that $\log M_{(r,k)} \sim \epsilon^{-2s_{(r,k)}}$. Then, $\|g_{[j]}\|_{\mathcal{H}_{(r,k)}} \leq 1$ and thus $\|g_{[j]}\|_{L_2(P_\mathcal{X})} \leq \|g_{[j]}\|_\infty \leq \sup_x k_{(r,k)}(x,x) \|g_{[j]}\|_{\mathcal{H}_{(r,k)}} \leq 1$. Now let,

$$\tilde{g}_{[j]} = \sqrt{(1 - \|g_{[j]}\|^2_{L_2(P_\mathcal{X})})} \frac{\widehat{f}_r^{(k)}}{\|\widehat{f}_r^{(k)}\|_{L_2(P_\mathcal{X})}} + g_{[j]}.$$

By the construction of $g_{[j]}$, we have $\langle g_{[j]}, \widehat{f}_r^{(k)} \rangle_{L_2(P_\mathcal{X})} = 0$ and thus

$$\|\tilde{g}_{[j]}\|^2_{L_2(P_\mathcal{X})} = (1 - \|g_{[j]}\|^2_{L_2(P_\mathcal{X})}) + \|g_{[j]}\|^2_{L_2(P_\mathcal{X})} = 1. \tag{S-21}$$

Moreover, since Assumption 7 gives $\|\widehat{f}_r^{(k)}\|_{L_2(P_{\mathcal{X}})} \geq c_1$, the RKHS norm of $\tilde{g}_{[j]}$ is bounded by

$$\|\tilde{g}_{[j]}\|_{\mathcal{H}_{(r,k)}} \leq \frac{\sqrt{(1-\|g_{[j]}\|_{L_2(P_{\mathcal{X}})}^2)}}{\|\widehat{f}_r^{(k)}\|_{L_2(P_{\mathcal{X}})}} \|\widehat{f}_r^{(k)}\|_{\mathcal{H}_{(r,k)}} + \|g_{[j]}\|_{\mathcal{H}_{(r,k)}} \leq \frac{1+c_1}{c_1}.$$

Moreover, $\{\tilde{g}_{[j]}\}_j$ satisfies

$$\|\tilde{g}_{[j]} - \tilde{g}_{[j']}\|_{L_2(P_{\mathcal{X}})} \geq \|g_{[j]} - g_{[j']}\|_{L_2(P_{\mathcal{X}})} \geq \epsilon$$

where we used the orthogonality between $\widehat{f}_r^{(k)}$ and $g_{[j]} - g_{[j']}$. Therefore, we have that

$$M_{(r,k)} \leq M(\mathcal{B}_{\mathcal{H}_{(r,k)}}, \epsilon, L_2(P_{\mathcal{X}_{(r,k)}})).$$

We denote by $\mathcal{G}_{(r,k)} := \{\tilde{g}_{[j]} \ (j = 1, \dots, M_{(r,k)})\}$.

We construct a packing set of $\mathscr{H}_{(d^*,K)}(R)$ as follows. Let

$$\mathcal{G} = \left\{ g = \sum_{r=1}^{d^*} \prod_{k=1}^{K} g_r^{(k)} \mid g_r^{(k)} \in \mathcal{G}_{(r,k)} \right\}.$$

Note that

$$|\mathcal{G}| = \prod_{r=1}^{d^*} \prod_{k=1}^{K} M_{(r,k)}.$$

It will be shown later that any $g, g' \in \mathcal{G}$ satisfy

$$\|g - g'\|_{L_2(P_{\mathcal{X}})}^2 \geq \sum_{r=1}^{d^*} \min\left\{ \frac{1}{K}, \frac{1}{2} \sum_{k=1}^{K} \left\| g_r^{(k)} - {g'}_r^{(k)} \right\|_{L_2(P_{\mathcal{X}})}^2 \right\}. \tag{S-22}$$

Thus, if $|\{(r,k) \mid g_r^{(k)} \neq {g'}_r^{(k)}\}| \geq \frac{d^*K}{2}$, then the right hand side of Eq. (S-22) is lower bounded by

$$\|g - g'\|_{L_2(P_{\mathcal{X}})}^2 \geq \frac{d^*K}{2}\epsilon^2 \tag{S-23}$$

for sufficiently small ϵ. Now, by the assumption that $s_{(r,k)} = s$ for all r, k, we may assume that $\exists M$ such that $M_{(r,k)} = M$ for all r, k. By Lemma C.1, we can construct a subset $\widetilde{\mathcal{G}}$ of $\mathcal{G}$ such that

$$|\widetilde{\mathcal{G}}| \geq \frac{1}{2} \frac{M^{d^*K}}{\binom{d^*K}{d^*K/2}(M+1)^{d^*K/2}},$$

$$g, g' \in \widetilde{\mathcal{G}}, \ g \neq g', \quad \Rightarrow \quad |\{(r,k) \mid g_r^{(k)} \neq {g'}_r^{(k)}\}| \geq \frac{d^*K}{2}.$$

Once this is shown, $\widetilde{\mathcal{G}}$ is actually a packing set of $\mathscr{H}_{(d^*,K)}(R)$ with $\epsilon_n = \frac{d^*K}{2}\epsilon^2$, and $Q = |\widetilde{\mathcal{G}}|$ satisfies

$$\log|\widetilde{\mathcal{G}}| \geq \frac{d^*K}{4}\log(M) - \frac{d^*K}{2}\log(2) \gtrsim d^*K\log(M)$$

for $M \geq 5$. Therefore,

$$\log(Q) \gtrsim d^*K\log(M) \gtrsim d^*K\epsilon^{-2s}.$$

By setting δ_n appropriately like $\delta_n = C\epsilon_n$, we have $\log(Q)/2 \leq 8\log(N) \leq \log(Q)$, and let ϵ to satisfy

$$\frac{n}{2\sigma^2}d^*K\epsilon^2 \lesssim d^*K\epsilon^{-2s}$$

then the inequalities (S-19) are satisfied for $\epsilon_n = \frac{d^* K}{2}\epsilon^2$. To satisfy this, we set $\epsilon \simeq n^{-\frac{1}{1+s}}$ and thus

$$\epsilon_n^2 \simeq \sum_{r=1}^{d^*}\sum_{k=1}^{K} n^{-\frac{1}{1+s_{(r,k)}}} = d^* K n^{-\frac{1}{1+s}},$$

then we obtain the assertion.

What remains to be shown is Eq. (S-22). This is shown as follows. First notice that

$$\begin{aligned}\|g-g'\|_{L_2(P_\mathcal{X})}^2 &= \|\sum_{r=1}^{d^*}(\prod_{k=1}^{K} g_r^{(k)} - \prod_{k=1}^{K} {g'}_r^{(k)})\|_{L_2(P_\mathcal{X})}^2 \\ &= \sum_{r=1}^{d^*}\left\|\prod_{k=1}^{K} g_r^{(k)} - \prod_{k=1}^{K} {g'}_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 .\end{aligned}$$

Next, we lower bound the summand as follows:

$$\begin{aligned}&\left\|\prod_{k=1}^{K} g_r^{(k)} - \prod_{k=1}^{K} {g'}_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 \\ =&\left\|g_r^{(1)}\prod_{k=2}^{K} g_r^{(k)} - {g'}_r^{(1)}\prod_{k=2}^{K} {g'}_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 \\ =&\left\|(g_r^{(1)} - {g'}_r^{(1)})\prod_{k=2}^{K} g_r^{(k)} - {g'}_r^{(1)}(\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)})\right\|_{L_2(P_\mathcal{X})}^2 \\ =&\left\|(g_r^{(1)} - {g'}_r^{(1)})\prod_{k=2}^{K} g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 - 2\left\langle (g_r^{(1)} - {g'}_r^{(1)})\prod_{k=2}^{K} g_r^{(k)}, {g'}_r^{(1)}(\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)})\right\rangle_{L_2(P_\mathcal{X})} \\ &+\left\|{g'}_r^{(1)}(\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)})\right\|_{L_2(P_\mathcal{X})}^2 \\ =&\left\|g_r^{(1)} - {g'}_r^{(1)}\right\|_{L_2(P_\mathcal{X})}^2\prod_{k=2}^{K}\left\|g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 \\ &-2\left\langle g_r^{(1)} - {g'}_r^{(1)}, {g'}_r^{(1)}\right\rangle_{L_2(P_\mathcal{X})}\times\left\langle \prod_{k=2}^{K} g_r^{(k)}, \prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)}\right\rangle_{L_2(P_\mathcal{X})} \\ &+\left\|{g'}_r^{(1)}\right\|_{L_2(P_\mathcal{X})}^2\left\|\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 .\end{aligned}$$

Using Lemma C.2 with Eq. (S-21), the RHS is equivalent to

$$\begin{aligned}&\left\|g_r^{(1)} - {g'}_r^{(1)}\right\|_{L_2(P_\mathcal{X})}^2\prod_{k=2}^{K}\left\|g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 - \frac{1}{2}\|g_r^{(1)} - {g'}_r^{(1)}\|_{L_2(P_\mathcal{X})}^2\times\left\|\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 \\ &+\left\|{g'}_r^{(1)}\right\|_{L_2(P_\mathcal{X})}^2\left\|\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2\end{aligned}$$

By using Eq. (S-21), we have that every $g_r^{(k)} \in \mathcal{G}_{(r,k)}$ satisfies $\|g_r^{(k)}\|_{L_2(P_\mathcal{X})} = 1$, and thus the RHS is lower bounded as

$$\left\|g_r^{(1)} - {g'}_r^{(1)}\right\|_{L_2(P_\mathcal{X})}^2 - \frac{1}{2}\|g_r^{(1)} - {g'}_r^{(1)}\|_{L_2(P_\mathcal{X})}^2\times\left\|\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2 + \left\|\prod_{k=2}^{K} {g'}_r^{(k)} - \prod_{k=2}^{K} g_r^{(k)}\right\|_{L_2(P_\mathcal{X})}^2$$

$$
\begin{aligned}
&\geq \left\| g_r^{(1)} - {g'}_r^{(1)} \right\|_{L_2(P_{\mathcal{X}})}^2 + \left(1 - \frac{1}{2}\|g_r^{(1)} - {g'}_r^{(1)}\|_{L_2(P_{\mathcal{X}})}^2\right) \left\| \prod_{k=2}^K {g'}_r^{(k)} - \prod_{k=2}^K g_r^{(k)} \right\|_{L_2(P_{\mathcal{X}})}^2 \\
&\geq \min\left\{ \frac{1}{K}, \left\| g_r^{(1)} - {g'}_r^{(1)} \right\|_{L_2(P_{\mathcal{X}})}^2 + (1 - 1/2K) \left\| \prod_{k=2}^K {g'}_r^{(k)} - \prod_{k=2}^K g_r^{(k)} \right\|_{L_2(P_{\mathcal{X}})}^2 \right\}.
\end{aligned}
$$

Applying the same argument K times, the right hand side is lower bounded by

$$
\begin{aligned}
&\min\left\{ \frac{1}{K}, (1 - 1/2K)^{K-1} \sum_{k=1}^K \left\| g_r^{(k)} - {g'}_r^{(k)} \right\|_{L_2(P_{\mathcal{X}})}^2 \right\} \\
\geq &\min\left\{ \frac{1}{K}, \frac{1}{2} \sum_{k=1}^K \left\| g_r^{(k)} - {g'}_r^{(k)} \right\|_{L_2(P_{\mathcal{X}})}^2 \right\}.
\end{aligned}
$$

This shows Eq. (S-22). Then we complete the proof.

Lemma C.1. Let $\Omega = \{1, \dots, M\}^s$, and define the Hamming distance in Ω as $d(x, y) = \sum_{i=1}^s \mathbf{1}[x_j \neq y_j]$. Then, there is a subset $\mathcal{A} \subseteq \Omega$ such that every pair $x, y \in \mathcal{A}$ s.t. $x \neq x'$ satisfies

$$
d(x, y) \geq s/2
$$

and $|\mathcal{A}| \geq \frac{M^s}{2\binom{s}{s/2}(M+1)^{s/2}}$.

Proof. The proof is given in the proof of Lemma 4 in Raskutti et al. (2012). □

Lemma C.2. Suppose that $\mathcal{H} \subseteq L_2(P_{\mathcal{X}})$ is a Hilbert space and $x, y \in \mathcal{H}$ satisfy $\|x\|_{L_2(P_{\mathcal{X}})} = \|y\|_{L_2(P_{\mathcal{X}})}$, then it holds that

$$
\langle x - y, y \rangle_{L_2(P_{\mathcal{X}})} = -\frac{1}{2} \|x - y\|_{L_2(P_{\mathcal{X}})}^2.
$$

Proof. Since $\|x\|_{L_2(P_{\mathcal{X}})}^2 = \|y\|_{L_2(P_{\mathcal{X}})}^2$, we have that

$$
\begin{aligned}
&\|x\|_{L_2(P_{\mathcal{X}})}^2 = \|x - y + y\|_{L_2(P_{\mathcal{X}})}^2 = \|x - y\|_{L_2(P_{\mathcal{X}})}^2 + 2\langle x - y, y \rangle_{L_2(P_{\mathcal{X}})} + \|y\|_{L_2(P_{\mathcal{X}})}^2 \\
\Rightarrow\ & 0 = \|x - y\|_{L_2(P_{\mathcal{X}})}^2 + 2\langle x - y, y \rangle_{L_2(P_{\mathcal{X}})}.
\end{aligned}
$$

This is equivalent to the assertion. □

Proposition C.3 (Theorem 15 in Steinwart et al. (2009))**.** Let $\mathcal{H}$ be an RKHS associated with a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Suppose that a kernel function k has an expansion such as

$$
k(x, x') = \sum_{i=1}^\infty \mu_i \psi_i(x) \psi_i(x')
$$

in $L_2(P_{\mathcal{X}})$ where $\{\phi_i\}_i \subseteq \mathcal{H}$ is an orthonormal system and $\mu_1 \geq \mu_2 \geq \cdots \geq 0$. Then, given $s > 0$, we have that $\mu_i \sim i^{-1/s}$ if and only if

$$
\mathcal{N}(\mathcal{B}_{\mathcal{H}}, \epsilon, L_2(P_{\mathcal{X}})) \sim \epsilon^{-2s}.
$$

□

References

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

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

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

Li, Wenbo V. A gaussian correlation inequality and its applications to small ball probabilities. *Electronic Communications in Probability*, 4:111–118, 1999.

Raskutti, Garvesh, Wainwright, Martin, and Yu, Bin. Minimax-optimal rates for sparse additive models over kernel classes via convex programming, 2010. arXiv:1008.3654.

Raskutti, Garvesh, Wainwright, Martin J, and Yu, Bin. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, 13(1):389–427, 2012.

Schechtman, Gideon, Schlumprecht, Th, and Zinn, Joel. On the gaussian measure of the intersection. *Annals of probability*, pp. 346–357, 1998.

Steinwart, Ingo and Christmann, Andreas. *Support Vector Machines*. Springer, 2008.

Steinwart, Ingo, Hush, Don, and Scovel, Clint. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pp. 79–93, 2009.

Suzuki, Taiji. Pac-bayesian bound for gaussian process regression and multiple kernel additive model. In *JMLR Workshop and Conference Proceedings*, volume 23, pp. 8.1–8.20, 2012. Conference on Learning Theory (COLT2012).

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

van der Vaart, Aad W. and van Zanten, J. Harry. 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.

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

Yang, Yuhong and Barron, Andrew. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.