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

Heishiro Kanagawa†
Taiji Suzuki†,‡
† Tokyo Institute of Technology, Tokyo 152-8552, JAPAN
‡ PRESTO, Japan Science and Technological Agency (JST), JAPAN

Hayato Kobayashi⋆
Nobuyuki Shimizu⋆
Yukihiro Tagami⋆
⋆ 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:

\[ y_i = f^o(x_i) + \epsilon_i, \quad (S-1) \]

where \( f^o : \mathcal{X} \rightarrow \mathbb{R} \) 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 \( f^o \simeq f^* \)). When \( f^o = f^* \), 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 \( f^o = f^* \).

A. Noise Assumption and PAC-Bayesian Bound

Here we remind our assumption on the noise \( \epsilon_i \) (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

\[ m_\epsilon(z) := -E[\epsilon_1 1\{\epsilon_1 \leq z\}] = -\int_{-\infty}^{\infty} ydF_\epsilon(y) = \int_{z}^{\infty} ydF_\epsilon(y), \]

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

Then we impose the following assumption on the noise \( \xi \).

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

\[ \int_a^b m_\epsilon(z)dz = \int_a^b g_\epsilon(z)dF_\epsilon(z), \quad \forall a, b \in \mathbb{R}. \]

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

- If \( \epsilon_1 \) obeys the Gaussian \( \mathcal{N}(0, \sigma^2) \), then \( g_\epsilon(z) = \sigma^2 \).
- If \( \epsilon_1 \) 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 \( \rho \) that is absolutely continuous with respect to \( \Pi \), let \( \mathcal{K}(\rho, \Pi) \) be the KL-divergence between \( \rho \) and \( \Pi \), \( \mathcal{K}(\rho, \Pi) := \int \log \left( \frac{d\rho}{d\Pi} (f) \right) d\rho(f) \).

**Theorem A.1.** Suppose Assumption 1 is satisfied and \( \beta \geq 4\|g_\xi\|_\infty \). Then for all probability measure \( \rho \) that is absolutely continuous with respect to \( \Pi \), we have

\[
E_{Y_1:n|X_1:n} \left[ \| \hat{f} - f^0 \|_n^2 \right] \leq \int \| f - f^0 \|_n^2 d\rho(f) + \frac{\beta \mathcal{K}(\rho, \Pi)}{n}. \tag{S-2}
\]

In the following, we assume that \( \beta \) 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 \( \mathcal{H}_{(r,k),\lambda} \) is the RKHS associated with the kernel \( k_{m,\lambda}^{(r)} = k_{r,k}/\lambda_{(k)} \).

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

\[
\phi_{f^*(k)}^{(r,k)}(\epsilon, \lambda, \alpha) := \inf_{h \in \mathcal{H}_{(r,k)} : \| h - f^*(k) \|_\lambda \leq \epsilon} \left( \| h \|_{\mathcal{H}_{(r,k),\lambda}}^2 + 1 \right) - \log \mathcal{G}_{(r,k)}(\{ f : \| f \|_\lambda \leq \epsilon/\sqrt{2} \}) - \log \mathcal{G}_{(r,k)}(\{ f : \| f \|_\infty \leq L/\sqrt{2} \}), \tag{S-3}
\]

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

**B.1. Generalized upper bound**

Define

\[
\tilde{S} := \{ r \mid 1 \leq r \leq d_{\text{max}} \text{, } \exists k \text{ s.t. } f^*_{r,k} \notin \mathcal{H}_{(r,k)} \}.
\]

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

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

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

\[
E_{Y_1:n|X_1:n} \left[ \| \hat{f} - f^0 \|_n^2 \right] \leq 2\| f^0 - f^* \|_n^2 + C_1 \inf_{\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}>0 \epsilon_{(r,k)} \leq L_{(r,k)}} \left\{ \sum_{r=1,\ldots,d^*} c_r \sum_{k=1}^K \left( \hat{R}_{K,\text{max}} \epsilon_{(r,k)}^2 + \frac{1}{n} \phi_{f^*(k)}^{(r,k)}(\epsilon_{(r,k)}, L_{(r,k)}, \lambda_{(r,k)}) \right) + \frac{1}{n} \log \left( \frac{\lambda_{(r,k)}}{n} \right) \right\} + \hat{R}_{K,\text{max}} \left( \sum_{r \in \tilde{S}} \sqrt{\sum_{k=1}^K \epsilon_{(r,k)}^2} \right)^2 + \frac{d^*}{n} \log \left( \frac{1}{\zeta(1 - \zeta)} \right). \tag{S-4}
\]

**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 \( r = 1, \ldots, d_{\text{max}} \) and \( k = 1, \ldots, K \). The typical approach to prove the theorem is that we substitute some “dummy” posterior distribution into \( \rho \) 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 \( S_{(r,k)} \) of a function as

\[
S_{(r,k)} := \{ f : \mathcal{X} \rightarrow \mathbb{R} \mid \| f \|_n \leq \epsilon_{(r,k)}, \| f \|_\infty \leq L_{(r,k)} \}.
\]
We define \( \hat{h}_{r,k} \in \mathcal{H}_{r,k} \) 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 \( \hat{h}_{r,k} \) as \( \hat{h}_{r,k} = f_r^{*}(k) \). Otherwise, we take \( \hat{h}_{r,k} \in \mathcal{H}_{r,k} \) such that

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

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

The process \( (W_x + \hat{h}_{r,k}(x) : x \in \lambda_k) \) induces the “shifted” Gaussian process \( \text{GP}^{W + \hat{h}_{r,k}}(r,k) (d f_r^{*}(k) | \lambda) \) such that \( \text{GP}^{W + \hat{h}_{r,k}}(r,k) (A | \lambda) := \text{GP}(r,k) (A - \hat{h}_{r,k} | \lambda) \) for a measurable set \( A \). Now our choice of \( \rho \) is given as follows:

\[
\rho(\text{df}) = \prod_{r=1}^{d^*} \prod_{k=1}^{K} \int_{\lambda_{r,k}(r) \leq \lambda_{r,k} \leq \lambda_{r,k}} \frac{\text{GP}^{W + \hat{h}_{r,k}}(r,k) (d f_r^{*}(k) | \lambda_{r,k}) \mathbf{1}\{f_r^{*}(k) - \hat{h}_{r,k} \in \mathcal{S}_{r,k}\}}{\text{GP}(r,k) (\mathcal{S}_{r,k} | \lambda_{r,k})} \mathcal{G}(d \hat{\lambda}_{r,k}) \\
\times \prod_{r > d^*} \prod_{k=1}^{K} \delta_0 (d f_r^{*}(k)).
\]

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

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

\[
\| \hat{h}_{r,k} \|_{\infty} \leq \epsilon_{r,k} + R.
\]

Similarly, for all \( f_r^{*}(k) \) in the support of \( \rho \), we have that \( \| f_r^{*}(k) - \hat{h}_{r,k} \|_{\infty} \leq L_{r,k} \) 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 \( f = \sum_{r=1}^{d_{\text{max}}} \prod_{k=1}^{K} f_r^{(k)} \) and \( f^* = \sum_{r=1}^{d_{\text{max}}} \prod_{k=1}^{K} f_r^{*}(k) \), it holds that

\[
\int \| f - f^* \|_n^2 \rho(f) \leq 2 \int \| f - f^* \|_n^2 \rho(f) + 2 \int \| f^* - f^0 \|_n^2 \rho(f).
\]

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

\[
\int \| f - f^* \|_n^2 \rho(f)
\]

\[
= \int \left\| \sum_{r=1}^{d_{\text{max}}} \left( \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} f_r^{*}(k) \right) \right\|_n^2 \rho(f)
\]

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

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

\[
- 2 \sum_{r \neq r': 1 \leq r, r' \leq d^*} \int \left( \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} f_r^{*}(k) \prod_{k=1}^{K} f_{r'}^{(k)} - \prod_{k=1}^{K} f_{r'}^{*}(k) \right) \rho(f).
\]

The first term of the RHS is evaluated by

\[
\sum_{r=1}^{d^*} \int \left\| \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} f_r^{*}(k) \right\|_n^2 \rho(f)
\]
If $k \neq \tilde{k}$ and $r \not\in \tilde{S}$, then the summand of the RHS is 0, otherwise the summand is bounded by
\[
\frac{1}{2} \sum_{k'=k}^K \| f^{(k')}_{r,k'} - f_{r,k'}^{*} \|^2_n \| f_{r,k'}^{*} \|^2_n \leq (R + 2L_{(r,k')})^2 \int \| f_r^{(k')} - f_r^{*} \|^2_n \| f_r^{*} \|^2_n \rho(f) \]
\[
\leq (R + 2 \max_{r,k} L_{(r,k)})^2 \sum_{k'=k}^K \| f_{r,k'}^{*} - f_{r,k'}^{*} \|^2_n \| f_{r,k'}^{*} \|^2_n \rho(f) \]
\[
\leq 4 \left( R + 2 \max_{r,k} L_{(r,k)} \right)^2 (K-1) \sum_{k'=k}^K \| f^{(k')}_{r,k'} - f_{r,k'}^{*} \|^2_n \| f_{r,k'}^{*} \|^2_n \rho(f) \]
\[
\leq 4 \left( R + 2 \max_{r,k} L_{(r,k)} \right)^2 (K-1) \sum_{r=1}^{d^*} c_r \sum_{k=1}^{K} \epsilon_{(r,k)}^2. \tag{S-8} \]

On the other hand, using Lemma B.1 again, an analogous reasoning gives a bound of the second term as
\[
\left| \int \left( \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} f_r^{*} \right) \left( \prod_{k'=1}^{K} f_r^{*} \right) \rho(f) \right| = \left| \int \left( \prod_{k=1}^{K} \tilde{h}_{(r,k)} - \prod_{k=1}^{K} \tilde{h}_{(r,k')} \right) \left( \prod_{k=1}^{K} f_r^{*} \right) \rho(f) \right| \]
\[
\leq \begin{cases} 
0, 
\left( R + 2 \max_{r,k} L_{(r,k)} \right)^{2(K-1)} \sqrt{\sum_{k=1}^{K} \epsilon_{(r,k)}^2} \sqrt{\sum_{k'=1}^{K} \epsilon_{(r,k')}^2}, & (r \not\in \tilde{S} \text{ or } r' \not\in \tilde{S}), \\
(\text{otherwise}). 
\end{cases} \tag{S-9} \]

Now define
\[
\tilde{f}_{r,k}^{*} \equiv f_{r,k}^{*} + \epsilon_{(r,k)} L_{(r,k)} \quad \text{and} \quad \tilde{h}_{(r,k)} \equiv h - f_{r,k}^{*}. 
\]

Then, along with the proof of Theorem 3 in Suzuki (2012), the KL-divergence between the “posterior” $\rho$ and the prior $\Pi$ is bounded as
\[
\frac{1}{n} \mathcal{K}(\rho, \Pi) 
\]
Proof. for all $k$ for all symmetric, convex sets $A$. In particular, $\lambda(\cdot) = \frac{1}{2} \log \left( \frac{\| \cdot \|^{2} \| \cdot \|^{2}}{2} \right)$. Using the same reasoning, we obtain the second assertion by noticing $\lambda(\cdot)$ 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

$$\geq - \log GP_{r,k}(S_{r,k}) \leq - \log GP_{r,k}(\{ f : \| f \|_{n} \leq \epsilon(r,k)/\sqrt{2} \} \lambda) - \log GP_{r,k}(\{ f : \| f \|_{\infty} \leq L(r,k)/\sqrt{2} \} \lambda).$$

Thus

\[ \phi^{r,k}_{f_{z}(r,k)}(\epsilon(r,k), L(r,k), \lambda(r,k)) \leq \phi^{r,k}_{f_{z}(r,k)}(\epsilon(r,k), L(r,k), \lambda(r,k)). \]

(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 $f(x) = \prod_{k=1}^{K} f_{k}(x) : X \mapsto \mathbb{R}$ 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 $f'(x) = \prod_{k=1}^{K} f'_{k}(x) : X \mapsto \mathbb{R}$ 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.**

$$\| f \|_{n}^{2} = \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{K} f_{k}(x_{i})^{2} \leq \frac{1}{n} \sum_{i=1}^{n} f_{k}(x_{i})^{2} \prod_{k' \neq k} \max_{i=1, \ldots, n} \{ f_{k'}(x_{i})^{2} \} \leq R^{K-1} \frac{1}{n} \sum_{i=1}^{n} f_{k}(x_{i})^{2}. $$

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

\[ \square \]

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

**Proposition B.2.** Let $\rho$ 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 $r, k$:

$$\hat{R}_{K, \max}^{2}(r, k) + \frac{1}{n} \phi^{r,k}_{f_{z}(r,k)}(\epsilon(r,k), L(r,k), \lambda(r,k)) + \frac{1}{n} \lambda(r,k) - \frac{1}{n} \log \left( \frac{\lambda(r,k)}{2} \right),$$

by choosing an appropriate $\epsilon(r,k), L(r,k), \lambda(r,k)$ such that $\epsilon(r,k) \leq L(r,k)$.
With a slight abuse of notation, we denote by \( ||f_r^{(k)}||_{\theta, \infty, \mathcal{H}(r, k)} = \sup_{t > 0} \inf_{h_r^{(k)} \in \mathcal{H}(r, k)} \{ t^{-\theta} ||f_r^{(k)} - h_r^{(k)}||_{\infty} + t^{1-\theta} ||h_r^{(k)}||_{\mathcal{H}(r, k)} \} \).  

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: \( \inf_{h_r^{(k)} \in \mathcal{H}(r, k)} ||h_r^{(k)}||_{\mathcal{H}(r, k)} \) for all \( t > 0 \), there exists \( h_r^{(k)} \in \mathcal{H}(r, k) \) such that

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

This gives \( 2||f_r^{(k)}||_{\theta, \infty} \geq t^{-\theta} ||f_r^{(k)} - h_r^{(k)}||_{\infty} \) so that we have \( t \geq 2^{-\frac{1}{\theta}} ||f_r^{(k)}||_{\theta, \infty} - \frac{1}{\theta} ||f_r^{(k)} - h_r^{(k)}||_{\infty} \) and hence \( 2||f_r^{(k)}||_{\theta, \infty} \geq 2^{\frac{1}{\theta}} ||f_r^{(k)} - h_r^{(k)}||_{\infty} \). Therefore we have that

\[
\inf_{h_r^{(k)} \in \mathcal{H}(r, k): ||h_r^{(k)}||_{\mathcal{H}(r, k)} \leq \epsilon} ||h_r^{(k)}||_{\mathcal{H}(r, k)} \leq 2^{\frac{1}{\theta}} ||f_r^{(k)}||_{\theta, \infty} \leq \epsilon \leq \epsilon(r, k)
\]

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

Setting (ii): 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

\[
-\log(\mathcal{G}_r,k(\{ f: \| f \|_n \leq \epsilon \})) \leq C'_0 \epsilon^{-\frac{2s(r, k)}{1-s(r, k)}}
\]

(Kuelbs & Li, 1993; Li & Shao, 2001). Similarly, Assumption 6 gives that there exists \( C'_1 \) such that

\[
-\log(\mathcal{G}_r,k(\{ f: \| f \|_\infty \leq L \})) \leq C'_{1} L^{\frac{2s(r, k)}{1-s(r, k)}}
\]

This and Eq. (S-13) give that

\[
\phi_{f_r^{(k)}}^{(r, k)} (\epsilon(r, k), \lambda(r, k)) \leq \left( 2R \right)^{\frac{1}{\theta}} \lambda \left( \frac{\sqrt{\lambda(r, k) \epsilon(r, k)}}{\sqrt{2}} \right)^{-\frac{2s(r, k)}{1-s(r, k)}} + C'_0 \left( \frac{\sqrt{\lambda(r, k) L(r, k)}}{\sqrt{2}} \right)^{-\frac{2s(r, k)}{1-s(r, k)}} + C'_1 \left( \frac{\sqrt{\lambda(r, k) L(r, k)}}{\sqrt{2}} \right)^{-\frac{2s(r, k)}{1-s(r, k)}}
\]

where we used

\[
\|f\|_{\mathcal{H}(r, k)}^2 = \lambda \|f\|_{\mathcal{H}(r, k)}^2,
\]

\[
-\log(\mathcal{G}_r,k(\{ f: \| f \|_n \leq \epsilon(r, k) \}|\lambda(r, k))) = -\log(\mathcal{G}_r,k(\{ f: \| f \|_n \leq \sqrt{\lambda(r, k) \epsilon(r, k)} \})
\]

\[
-\log(\mathcal{G}_r,k(\{ f: \| f \|_\infty \leq L(r, k) \}|\lambda(r, k))) = -\log(\mathcal{G}_r,k(\{ f: \| f \|_\infty \leq \sqrt{\lambda(r, k) L(r, k)} \})
\]

Now \( \lambda(r, k) = (R \vee 1) \frac{2(1-s(r, k))}{\theta} \epsilon(r, k) \) 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 \( L(r, k) = (R \vee 1)^{\frac{1-s(r, k)}{s(r, k)}} \). With this \( \lambda(r, k) \) and \( L(r, k) \), the RHS of Eq. (S-12) is bounded as

\[
\hat{R}_{k, \max} \lambda^2(r, k) + \frac{1}{n} \phi_{f_r^{(k)}}^{(r, k)} (\epsilon(r, k), L(r, k), \lambda(r, k)) + \lambda(r, k) \frac{\log(\lambda(r, k))}{n} \leq \hat{R}_{k, \max} \epsilon^2(r, k) + \left( \frac{2^{\frac{1}{\theta}} + C'_0 \frac{2s(r, k)}{1-s(r, k)}}{n} \right) (R \vee 1)^{\frac{2s(r, k)}{1-s(r, k)}} \epsilon(r, k)
Thus by taking the number of functions \(\delta\) (Theorem 4) the RHS of Eq. (S-15), the RHS of Eq. (S-15) is bounded by

\[
C \left( \hat{R}_{K, \text{max}} \vee (R \vee 1) \frac{2s(r,k)}{\epsilon(r,k)} \right) n^{-\frac{1}{\bar{\tau}(r,k)/\tau}}.
\]

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

\[
C \left( \hat{R}_{K, \text{max}} \vee (R \vee 1) \frac{2s(r,k)}{\epsilon(r,k)} \right) n^{-\frac{1}{\bar{\tau}(r,k)/\tau}}.
\]

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

\[-\log(GP_{r,k}(\{f : \|f\|_\infty \leq \sqrt{\lambda(r,k)L(r,k)}) \leq -\log(GP_{r,k}(\{f : \|f\|_\infty \leq 1\})) \leq -\log(c_1).\]

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 \(\hat{S} = \emptyset, \theta = 1\) and \((R \vee 1) \frac{2s(r,k)}{\epsilon(r,k)} \leq (R \vee 1)^2\) because of \(s(r,k) < 1\).

**C. Proof of minimax lower bound (Theorem 4)**

**Proof.** (Theorem 4) The \(\delta\)-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 \(\{f_1, \ldots, f_M\} \subseteq \mathcal{G}\) such that \(\|f_i - f_j\| \geq \delta\) for all \(i \neq j\). Generally, it holds that

\[
N(\mathcal{G}, \delta/2, \| \cdot \|) \leq M(\mathcal{G}, \delta, \| \cdot \|) \leq N(\mathcal{G}, \delta, \| \cdot \|).
\]

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

\[
\inf_{\hat{f}} \sup_{f \in \mathcal{H}(d^*, K)(R)} \mathbb{E}[\|\hat{f} - f\|_{L_2(P_X)}^2] \geq \inf_{\hat{f}} \sup_{f \in \mathcal{H}(d^*, K)(R)} \frac{\delta^2}{2} P[\|\hat{f} - f\|_{L_2(P_X)}^2 \geq \delta^2/2] \\
\geq \frac{\delta_n^2}{2} \left( 1 - \frac{\log(N) + \frac{\pi}{\sqrt{2}} \varepsilon_n^2 + \log(2)}{\log(Q)} \right).
\]

Thus by taking \(\delta_n\) and \(\varepsilon_n\) to satisfy

\[
\frac{n}{2\delta_n^2} \varepsilon_n^2 \leq \log(N), \quad S - 8 \log(N) \leq \log(Q), \quad S - 4 \log(2) \leq \log(Q),
\]

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 \( \delta_n \) and \( \varepsilon_n \). For all \( f, f' \in \mathcal{H}_{d,k}^r(R) \), it holds that

\[
\| f - f' \|_{L^2(P_X)}^2 = \left\| \frac{d'}{\sum_{r=1}^{d'} \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} f_r'^{(k)}} \right\|_{L^2(P_X)}^2 \leq \sum_{r=1}^{d'} \left\| \prod_{k=1}^{K} f_r^{(k)} - \prod_{k=1}^{K} f_r'^{(k)} \right\|_{L^2(P_X)}^2
\]

by the construction of \( L_2(P_X) \) and the assumption that \( \mathbb{E}[f_r^{(k)}(X)] = 0 \) 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 \( f_{(r)k} \) is continuous, Mercer’s theorem gives the orthogonal decomposition of the kernel function \( f_{(r)k} \) as

\[
f_{(r)k}(x, x') = \sum_{i=1}^{\infty} \mu_{(r)k,i} \phi_{(r)k,i}(x) \phi_{(r)k,i}(x'), \quad \text{(S-20)}
\]

where the convergence is absolute and uniform, \( \{ \phi_{(r)k,i} \}_{i=1}^{\infty} \) 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 \( \tilde{f}_r^{(k)} \in B_{\mathcal{H}_{r,k}} \) such that \( \| \tilde{f}_r^{(k)} \|_{L^2(P_{X_k})} \geq c_1 \). Without loss of generality, we may assume that \( \tilde{f}_r^{(k)} = \sqrt{\mu_{(r)k,1}} \phi_{(r)k,1} \) because

\[
\sqrt{\mu_{(r)k,1}} \phi_{(r)k,1} = \arg\max_{f \in B_{\mathcal{H}_{r,k}}} \| \tilde{f}_r \|_{L^2(P_X)}.
\]

This can be seen by the relation \( \| f \|^2_{L^2(P_X)} = \sum_{i=1}^{n} \int \int f(x)f(x') \phi_{(r)k,i}(x) \phi_{(r)k,i}(x') / \mu_{(r)k,i} dP_X(x) dP_X(x') \). Now, we consider a subspace which is perpendicular to \( \tilde{f}_r^{(k)} \). Let \( \mathcal{H}_{\perp, r,k} := \{ f \in \mathcal{H}_{r,k} \mid \langle f, \tilde{f}_r^{(k)} \rangle_{L^2(P_X)} = 0 \} \). 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

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

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

\[
k_{\perp, r,k}(x, x') = \sum_{i=2}^{\infty} \mu_{(r)k,i} \phi_{(r)k,i}(x) \phi_{(r)k,i}(x'),
\]

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 \( \mu_{(r)k,i} \sim i^{-1/2} \). Thus, we again use Proposition C.3 to obtain that

\[
\log N(\mathcal{B}_{\mathcal{H}_{\perp, r,k}}, \epsilon, L^2(P_X)) \sim \epsilon^{-2s_{(r,k)}}.
\]

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

\[
\tilde{g}_{[j]} = \frac{(1 - \| g_{[j]} \|_{L^2(P_X)}^2)}{\| \tilde{f}_r^{(k)} \|_{L^2(P_X)}^2} \tilde{f}_r^{(k)} + g_{[j]}.
\]

By the construction of \( g_{[j]} \), we have \( \langle g_{[j]}, \tilde{f}_r^{(k)} \rangle_{L^2(P_X)} = 0 \) and thus

\[
\| \tilde{g}_{[j]} \|_{L^2(P_X)} = (1 - \| g_{[j]} \|_{L^2(P_X)}^2) + \| g_{[j]} \|_{L^2(P_X)}^2 = 1.
\]

(S-21)
Moreover, since Assumption 7 gives \( \| \tilde{f}_r^{(k)} \|_{L^2(P_\alpha)} \geq c_1 \), the RKHS norm of \( \tilde{g}_{[j]} \) is bounded by

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

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

\[
\| \tilde{g}_{[j]} - \tilde{g}_{[j']} \|_{L^2(P_\alpha)} \geq \| g_{[j]} - g_{[j']} \|_{L^2(P_\alpha)} \geq \epsilon
\]

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

\[ M_{(r,k)} \leq M(\mathcal{B}_{\mathcal{H}(r,k)}, \epsilon, L_2(P_{X_{(r,k)}})). \]

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

We construct a packing set of \( \mathcal{H}(d^*, K)(R) \) 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

\[
\| g - g' \|^2_{L^2(P_\alpha)} \geq \sum_{r=1}^{d^*} \min \left\{ \frac{1}{K}, \frac{1}{2} \sum_{k=1}^{K} \| g_r^{(k)} - g_r'^{(k)} \|^2_{L^2(P_\alpha)} \right\}.
\]

(\text{S-22})

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

\[
\| g - g' \|^2_{L^2(P_\alpha)} \geq \frac{d^* K}{2} \epsilon^2
\]

(\text{S-23})

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

\[
|\tilde{\mathcal{G}}| \geq \frac{1}{2} \left( \frac{d^* K}{(d^* K/2) (M + 1)^{d^* K/2}} \right) M_{d^* K},
\]

\[ g, g' \in \tilde{\mathcal{G}}, g \neq g', \implies |(r, k) \mid g_r^{(k)} \neq g_r'^{(k)} | \geq \frac{d^* K}{2}. \]

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

\[
\log |\tilde{\mathcal{G}}| \geq \frac{d^* K}{4} \log(M) - \frac{d^* K}{2} \log(2) \geq d^* K \log(M)
\]

for \( M \geq 5 \). Therefore,

\[
\log(Q) \geq d^* K \log(M) \geq d^* K \epsilon^{-2s}.
\]

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

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

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

then we obtain the assertion.

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

\[
\left\| g - g' \right\|_{L_2(P_X)}^2 = \left\| \sum_{r=1}^{d^*} \left( \prod_{k=1}^{K} g_{r}^{(k)} - \prod_{k=1}^{K} g_{r}^{(k)} \right) \right\|_{L_2(P_X)}^2
\]

Next, we lower bound the summand as follows:

\[
\left\| \prod_{k=1}^{K} g_{r}^{(k)} - \prod_{k=1}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2 = \left\| g_{r}^{(1)} \prod_{k=2}^{K} g_{r}^{(k)} - g_{r}^{(1)} \prod_{k=2}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2
\]

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

\[
\left\| g_{r}^{(1)} \prod_{k=2}^{K} g_{r}^{(k)} - g_{r}^{(1)} \prod_{k=2}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2 \leq \frac{1}{2} \left\| g_{r}^{(1)} - g_{r}^{(1)} \right\|_{L_2(P_X)}^2 \times \left\| \prod_{k=2}^{K} g_{r}^{(k)} - \prod_{k=2}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2
\]

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

\[
\left\| g_{r}^{(1)} \prod_{k=2}^{K} g_{r}^{(k)} - g_{r}^{(1)} \prod_{k=2}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2 \leq \frac{1}{2} \left\| g_{r}^{(1)} - g_{r}^{(1)} \right\|_{L_2(P_X)}^2 \times \left\| \prod_{k=2}^{K} g_{r}^{(k)} - \prod_{k=2}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2 + \left\| \prod_{k=2}^{K} g_{r}^{(k)} - \prod_{k=2}^{K} g_{r}^{(k)} \right\|_{L_2(P_X)}^2
\]
Lemma C.1. Let $\Omega = \{1, \ldots, M\}^s$, and define the Hamming distance in $\Omega$ as $d(x, y) = \sum_{i=1}^s 1[x_i \neq y_i]$. Then, there is a subset $A \subseteq \Omega$ such that every pair $x, y \in A$ s.t. $x \neq x'$ satisfies

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

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

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

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

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

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

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

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

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

in $L_2(\mathcal{X})$ where $\{\phi_i\} \subseteq \mathcal{H}$ 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

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


References


