–Supplementary material–
Fast generalization error bound of deep learning from a kernel perspective

Taiji Suzuki
taiji@mist.i.u-tokyo.ac.jp
Graduate School of Information Science and Technology, The University of Tokyo
PRESTO, Japan Science and Technology Agency, Japan
Center for Advanced Integrated Intelligence Research, RIKEN, Tokyo, Japan

A Approximation error bound and L_∞-norm bound of the finite dimensional model

A.1 Approximation error bound

To derive the approximation error bound, we utilize the following proposition that was proven by Bach (2017).

Proposition 1. *For $\lambda > 0$, there exists a probability density $q_\ell(\tau)$ with respect to the measure Q_ℓ such that, for any $\delta \in (0,1)$, i.i.d. sample $v_1, \dots, v_m$ from q_ℓ satisfies that*

$$\sup_{\|f\|_{\mathcal{H}_\ell} \le 1} \inf_{\beta \in \mathbb{R}^m : \|\beta\|_2^2 \le \frac{4}{m}} \left\| f - \sum_{j=1}^m \beta_j q_\ell(v_j)^{-1/2} \eta(F_{\ell-1}(\cdot, v_j)) \right\|_{L_2(P(X))}^2 \le 4\lambda,$$

with probability $1-\delta$, if

$$m \ge 5N_\ell(\lambda) \log(16N_\ell(\lambda)/\delta).$$

By the scale invaliance of η, $\eta(ax) = a\eta(x)$ $(a > 0)$, we have the following proposition based on Proposition 1.

Lemma 1. *For $\lambda > 0$, and any $1/2 > \delta > 0$, if*

$$m \ge 5N_\ell(\lambda) \log(16N_\ell(\lambda)/\delta),$$

then there exist $v_1, \dots v_m \in \mathcal{T}_\ell$, $w_1, \dots, w_m > 0$ such that

$$\sup_{\|f\|_{\mathcal{H}_\ell} \le R} \inf_{\beta \in \mathbb{R}^m : \|\beta\|_2^2 \le \frac{4R^2}{m}} \left\| f - \sum_{j=1}^m \beta_j \eta(w_j F_{\ell-1}(\cdot, v_j)) \right\|_{L_2(P(X))}^2 \le 4\lambda R^2,$$

and

$$\frac{1}{m} \sum_{j=1}^m w_j^2 \le (1-2\delta)^{-1}.$$

Proof. Notice that $\mathrm{E}[\frac{1}{m}\sum_{j=1}^m q_\ell(v_j)^{-1}] = \mathrm{E}[q_\ell(v)^{-1}] = \int_{\mathcal{T}_\ell} q_\ell(v)^{-1} q_\ell(v) \mathrm{d}Q_\ell(v) = \int_{\mathcal{T}_\ell} 1 \mathrm{d}Q_\ell(v) = 1$, thus an i.i.d. sequence $\{v_1, \dots, v_m\}$ satisfies $\frac{1}{m}\sum_{j=1}^m q_\ell(v_j)^{-1} \le 1/(1-2\delta)$ with probability 2δ by the Markov's inequality. Combining this with Proposition 1, the i.i.d. sequence $\{v_1, \dots, v_m\}$ and $w_j = q_\ell(v_j)^{-1/2}$ satisfies the condition in the statement with probability $1 - (\delta + 1 - 2\delta) = \delta > 0$. This ensures the existence of sequences $\{v_j\}_{j=1}^m$ and $\{w_j\}_{j=1}^m$ that satisfy the assertion. $\square$

From now on, we define

$$c_0 = 4,\ c_1 = 4,\ c_\delta = (1-2\delta)^{-1}.$$

Based on the proposition, we approximate $f^{\rm o}$ given by the integral form (2) by a finite dimensional model f^* given as follows: let m_ℓ be the number of nodes in the ℓ-th internal layer (we set the dimensions of the output and input layers to $m_{L+1}=1$ and $m_1 = d_{\rm x}$) and consider a model

$$\begin{aligned}
&f^*_\ell(g) = W^{(\ell)}\eta(g) + b^{(\ell)} \quad (g \in \mathbb{R}^{m_\ell},\ \ell = 2,\dots,L),\\
&f^*_1(x) = W^{(1)}x + b^{(1)},\\
&f^*(x) = f^*_L \circ f^*_{L-1} \circ \cdots \circ f^*_1(x),
\end{aligned}$$

where $W^{(\ell)} \in \mathbb{R}^{m_{\ell+1}\times m_\ell}$ and $b^{(\ell)} \in \mathbb{R}^{m_{\ell+1}}$.

The next lemma gives an approximation error bound between $f^{\rm o}$ and f^*. The L_∞-norm bounds of $f^{\rm o}$ and f^* are given later in Lemma 3. Substituting $\delta \leftarrow \delta/2$ into the statement in the following Lemma 2 and letting $\hat{c}_\delta = c_1 c_{\delta/2}$, we derive the approximation error $\hat{\delta}_{1,n}$ in Theorem 1 in the main body.

Lemma 2 (Approximation error bound of the nonparametric model)**.** *For any $1/2 > \delta > 0$ and given $\lambda_\ell > 0$, let $m_\ell \geq 5N_\ell(\lambda_\ell)\log(16N_\ell(\lambda_\ell)/\delta)$. Then there exist $W^{(\ell)} \in \mathbb{R}^{m_{\ell+1}\times m_\ell}$ and $b^{(\ell)} \in \mathbb{R}^{m_{\ell+1}}$ ($\ell = 1,\dots,L$) where $m_{L+1} = 1$ and $m_1 = d_{\rm x}$ such that*

$$\begin{aligned}
&\|W^{(\ell)}\|_{\rm F}^2 \leq c_1 c_\delta R^2, \quad \|b^{(\ell)}\|_2 \leq \sqrt{c_\delta} R_b \quad (\ell = 1,\dots,L-1),\\
&\|W^{(L)}\|_{\rm F}^2 \leq c_1 R^2, \quad \|b^{(L)}\|_2 \leq R_b,
\end{aligned}$$

and

$$\|f^{\rm o} - f^*\|_{L_2(P(X))} \leq \sum_{\ell=2}^{L} \sqrt{(c_1 c_\delta)^{L-\ell} c_0} R^{L-\ell+1}\sqrt{\lambda_\ell}.$$

Proof. We construct the asserted finite dimensional network recursively from $\ell = L$ to $\ell = 1$. Let $\{v_j^{(\ell)}\}_{j=1}^{m_\ell}$ and $\{w_j^{(\ell)}\}_{j=1}^{m_\ell}$ be the sequences given in Proposition 1. Let $\widehat{\mathcal{T}}_\ell = \{v_j^{(\ell)}\}_{j=1}^{m_\ell}$. With slight abuse of notation, we identify $f^*_\ell : \mathbb{R}^{m_\ell} \to \mathbb{R}^{m_{\ell+1}}$ to a function $f^*_\ell : \widehat{\mathcal{T}}_\ell \to \widehat{\mathcal{T}}_{\ell+1}$ in a canonical way. For a function $F : \mathbb{R}^{d_{\rm x}} \times \widehat{\mathcal{T}}_\ell \to \mathbb{R}$, we denote by $f^*_\ell[F](x, v_i^{(\ell+1)})$ to express $f^*_\ell[F(x,\cdot)](v_i^{(\ell+1)}) = \sum_{j=1}^{m_\ell} W_{i,j}^{(\ell)} F(x, v_j^{(\ell)}) + b_i^{(\ell)}$ for $v_i^{(\ell+1)} \in \widehat{\mathcal{T}}_{\ell+1}$. When we write $f^*_\ell[F]$ for $F : \mathbb{R}^{d_{\rm x}} \times \mathcal{T}_\ell \to \mathbb{R}$ ($(x,v) \mapsto F(x,v)$), we deal with F as a restriction of F on $\mathbb{R}^{d_{\rm x}} \times \widehat{\mathcal{T}}_\ell$. We define the output from the ℓ-th layer of the approximated network f^* as $F^*_\ell(x,v)$ for $v \in \widehat{\mathcal{T}}_\ell$ and $x \in \mathbb{R}^{d_{\rm x}}$. More precisely, it is recursively defined as $F^*_\ell(x,v) = f^*_\ell[F^*_{\ell-1}](x,v)$.

We use an analogous notation for other networks such as $f^{\rm o}_\ell$. That is, $F^{\rm o}_\ell(x,v) = (f^{\rm o}_\ell \circ \cdots \circ f^{\rm o}_1(x))(v)$ for $v \in \mathcal{T}_\ell$ and $x \in \mathbb{R}^{d_{\rm x}}$, and $F^{\rm o}_\ell(x,v) = f^{\rm o}_\ell[F^{\rm o}_{\ell-1}](x,v)$.

Step 1 (the last layer, $\ell = L$).

We consider the following approximation of the L-th layer (the last layer): Remember that $m_{L+1} = 1$ and thus the output from the L-th layer is just one dimensional. We denote by $\mathcal{T}_{L+1} = \{1\}$ which is the index set of the output (which is just a singleton consisting of an element 1). As a candidate of a good approximation to the true L-th layer, define

$$\tilde{f}^*_L[F_{L-1}](x,1) = \sum_{j=1}^{m_L} \sqrt{m_L}\beta_j^{(L)} \eta\left(\frac{1}{\sqrt{m_L}} w_j^{(L)} F_{L-1}(x, v_j^{(L)})\right) + b_L \tag{S-1}$$

by $\beta^{(L)} \in \mathbb{R}^{m_L}$ and $w^{(L)} \in \mathbb{R}^{m_L}$ satisfying $\|\beta^{(L)}\|_2^2 \leq \frac{1}{m_L} c_1 R^2$ and $\|w^{(L)}\|_2^2 \leq m_L c_\delta$. Here, define that

$$W^{(L)}_{1,:} = \sqrt{m_L}{\beta^{(L)}}^\top, \quad b^{(L)} = (b^{\rm o}_L(1)).$$

Note that the model (S-1) can be rewritten as

$$\tilde{f}^*_L[F_{L-1}](x,1) = \sum_{j=1}^{m_L} W_{1,j}^{(L)} \eta({\sqrt{m_L}}^{-1} w_j^{(L)} F_{L-1}(x, v_j^{(L)})) + b_1^{(L)}.$$

Because of Proposition 1 and Assumption 1, the norms of the weight $W^{(L)}$ and the bias $b^{(L)}$ are bounded as

$$\|W^{(L)}\|_F = \|W^{(L)}_{1,:}\|_2 \le \sqrt{c_1}R, \quad \|b^{(L)}\|_2 = |b_L| \le R_b. \tag{S-2}$$

By the Cauchy-Schwartz inequality and the Lipschitz continuity of η, we have that

$$\begin{aligned}
&|\tilde{f}^*_L[F_{L-1}](x,1) - \tilde{f}^*_L[F'_{L-1}](x,1)| \\
&\le |\sum_{j=1}^{m_L} W^{(L)}_{1,j}(\eta(\sqrt{m_L}^{-1} w^{(L)}_j F_{L-1}(x, v^{(L)}_j)) - \eta(\sqrt{m_L}^{-1} w^{(L)}_j F'_{L-1}(x, v^{(L)}_j)))| \\
&\le \|W^{(L)}_{1,:}\|_2 \sqrt{m_L}^{-1} \|(w^{(L)}_j (F_{L-1}(x, v^{(L)}_j) - F'_{L-1}(x, v^{(L)}_j)))_{j=1}^{m_L}\|_2 \\
&\le \|W^{(L)}_{1,:}\|_2 \sqrt{m_L}^{-1} \|w^{(L)}\|_2 \|(F_{L-1}(x, v^{(L)}_j) - F'_{L-1}(x, v^{(L)}_j))_{j=1}^{m_L}\|_{\max} \\
&\le \sqrt{c_1 R^2}\sqrt{c_\delta m_L/m_L} \|(F_{L-1}(x, v^{(L)}_j) - F'_{L-1}(x, v^{(L)}_j))_{j=1}^{m_L}\|_{\max} \\
&= \sqrt{c_1 c_\delta} R \|(F_{L-1}(x, v^{(L)}_j) - F'_{L-1}(x, v^{(L)}_j))_{j=1}^{m_L}\|_{\max},
\end{aligned}$$

for $F_{L-1}, F'_{L-1} : \widehat{\mathcal{T}}_L \times \mathbb{R}^{d_x} \to \mathbb{R}$. Moreover, Proposition 1 ensures that $\beta^{(L)}$ and $w^{(L)}$ can be taken so that

$$\|\tilde{f}^*_L[F^{\mathrm{o}}_{L-1}](\cdot,1) - f^{\mathrm{o}}_L[F^{\mathrm{o}}_{L-1}](\cdot,1)\|^2_{L_2(P(X))} \le c_0 \lambda_L R^2.$$

Hereinafter, we fix $\beta^{(L)}$ and $w^{(L)}$ so that this inequality and the norm bound (S-2) are satisfied.

Step 2 (internal layers for $\ell = 2, \ldots, L-1$). As for the ℓ-th internal layer, we consider the following approximation:

$$\tilde{f}^*_\ell[g](v^{(\ell+1)}_i) = \sum_{j=1}^{m_\ell} \sqrt{m_\ell} \beta^{(\ell)}_{i,j} \eta(\sqrt{m_\ell}^{-1} w^{(\ell)}_j g(v^{(\ell)}_j)) + b^{\mathrm{o}}_\ell(v^{(\ell+1)}_i),$$

for $g : \widehat{\mathcal{T}}_\ell \to \mathbb{R}$ with $\beta^{(\ell)} \in \mathbb{R}^{m_{\ell+1}\times m_\ell}$ and $w^{(\ell)} \in \mathbb{R}^{m_\ell}$ satisfying $\|\beta^{(\ell)}_{j,:}\|_2^2 \le \frac{1}{m_\ell} c_1 R^2$ ($\forall j = 1, \ldots, m_{\ell+1}$) and $\|w^{(\ell)}\|_2^2 \le m_\ell c_\delta$. Then, the Lipschitz continuity of $\tilde{f}^*_\ell$ can be shown as

$$\begin{aligned}
&|\tilde{f}^*_\ell[F_{\ell-1}](x, v^{(\ell+1)}_i) - \tilde{f}^*_\ell[F'_{\ell-1}](x, v^{(\ell+1)}_i)| \\
&\le \left|\sum_{j=1}^{m_\ell} \sqrt{m_\ell}\beta^{(\ell)}_{i,j}(\eta(\sqrt{m_\ell}^{-1} w^{(\ell)}_j F_{\ell-1}(x, v^{(\ell)}_j)) - \eta(\sqrt{m_\ell}^{-1} w^{(\ell)}_j F'_{\ell-1}(x, v^{(L)}_j)))\right| \\
&\le \|\beta^{(\ell)}_{i,:}\|_2 \|w^{(\ell)}\|_2 \|(F_{\ell-1}(x, v^{(\ell)}_j) - F'_{\ell-1}(x, v^{(\ell)}_j))_{j=1}^{m_\ell}\|_{\max} \\
&\le \sqrt{\frac{c_1}{m_\ell}} R \sqrt{c_\delta m_\ell} \|(F_{\ell-1}(x, v^{(\ell)}_j) - F'_{\ell-1}(x, v^{(\ell)}_j))_{j=1}^{m_\ell}\|_{\max} \\
&= \sqrt{c_1 c_\delta} R \|(F_{\ell-1}(x, v^{(\ell)}_j) - F'_{\ell-1}(x, v^{(\ell)}_j))_{j=1}^{m_L}\|_{\max},
\end{aligned}$$

for any $v^{(\ell+1)}_i \in \widehat{\mathcal{T}}_{(\ell+1)}$. Proposition 1 asserts that there exit $\beta^{(\ell)}$ and $w^{(\ell)}$ that give an upper bound of the approximation error of the ℓ-th layer as

$$\max_{j=1,\ldots,m_\ell} \|\tilde{f}^*_\ell[F^{\mathrm{o}}_{\ell-1}](\cdot, v^{\ell+1}_j) - f^{\mathrm{o}}_\ell[F^{\mathrm{o}}_{\ell-1}](\cdot, v^{\ell+1}_j)\|^2_{L_2(P(X))} \le c_0 \lambda_\ell R^2.$$

Finally, let

$$W^{(\ell)}_{ij} = \sqrt{\frac{m_\ell}{m_{\ell+1}}} \beta^{(\ell)}_{ij} w^{(\ell+1)}_i, \quad b^{(\ell)} = \frac{1}{\sqrt{m_{\ell+1}}}(w^{(\ell+1)}_1 b^{\mathrm{o}}_\ell(v^{(\ell+1)}_1), \ldots, w^{(\ell+1)}_{m_{\ell+1}} b^{\mathrm{o}}_\ell(v^{(\ell+1)}_{m_{\ell+1}}))^\top,$$

then, by Assumption 1 and Proposition 1, the norms of these quantities can be bounded as

$$\begin{aligned}
\|W^{(\ell)}\|_{\mathrm{F}}^2 &= \frac{m_\ell}{m_{\ell+1}} \sum_{i=1}^{m_{\ell+1}} \sum_{j=1}^{m_\ell} \beta^{(\ell)^2}_{ij} w^{(\ell+1)2}_i \\
&\le \frac{m_\ell}{m_{\ell+1}} \sum_{i=1}^{m_{\ell+1}} w^{(\ell+1)2}_i \frac{c_1 R^2}{m_\ell} \le c_1 c_\delta R^2,
\end{aligned}$$

and

$$\|b^{(\ell)}\|_2^2 \leq \frac{1}{m_{\ell+1}} \sum_{j=1}^{m_{\ell+1}} {w^{(\ell+1)}}^2 R_b^2 \leq c_\delta R_b^2.$$

Step 3 (the first layer, $\ell = 1$).

For the first layer, let

$$\tilde{f}^*(x, v_i^{(2)}) = \sum_{j=1}^{d_x} h_1^{\circ}(v_i^{(2)}, j) Q_1(j) x_j + b_1^{\circ}(v_i^{(2)})$$

for $v_i^{(2)} \in \widehat{\mathcal{T}}_2$. By the definition of f°, it holds that

$$\tilde{f}^*(x, v_i^{(2)}) = f^{\circ}(x, v_i^{(2)}).$$

Let $W^{(1)} = \frac{1}{\sqrt{m_2}} (Q_1(j) w_i^{(2)} h_1^{\circ}(v_i^{(2)}, j))_{i,j} \in \mathbb{R}^{m_2 \times d_x}$ and $b^{(1)} = \frac{1}{\sqrt{m_2}} (w_1^{(2)} b_1^{\circ}(1), \ldots, w_{m_2}^{(2)} b_1^{\circ}(m_2))^\top \in \mathbb{R}^{m_2}$. Then, by Assumption 1 and Proposition 1, it holds that

$$\begin{aligned}
\|W^{(1)}\|_{\mathrm{F}}^2 =& \sum_{i=1}^{m_2} \sum_{j=1}^{d_x} \frac{1}{m_2} {w_i^{(2)}}^2 h_1^{\circ}(v_i^{(2)}, j)^2 Q_1(j)^2 \\
\leq& \left(\sum_{i=1}^{m_2} \frac{1}{m_2} {w_i^{(2)}}^2 \right) \max_{1 \leq i \leq m_2} \left(\sum_{j=1}^{d_x} h_1^{\circ}(v_i^{(2)}, j)^2 Q_1(j)^2 \right) \\
\leq& c_\delta \max_{1 \leq i \leq m_2} \left(\sum_{j=1}^{d_x} h_1^{\circ}(v_i^{(2)}, j)^2 Q_1(j) \right) \leq c_\delta R^2,
\end{aligned}$$

and

$$\|b^{(1)}\|_2^2 \leq \frac{1}{m_1} \sum_{i=1}^{m_2} {w_i^{(2)}}^2 R_b^2 \leq c_\delta R_b^2.$$

Step 4.

Finally, we combine the results we have obtained above. Note that

$$\begin{aligned}
& \|f_L^{\circ} \circ f_{L-1}^{\circ} \circ \cdots \circ f_1^{\circ} - \tilde{f}_L^* \circ \tilde{f}_{L-1}^* \circ \cdots \circ \tilde{f}_1^*\|_{L_2(P(X))} \\
=& \|f_L^{\circ} \circ f_{L-1}^{\circ} \circ \cdots \circ f_1^{\circ} - \tilde{f}_L^* \circ f_{L-1}^{\circ} \circ \cdots \circ f_1^{\circ} \\
& \qquad \vdots \\
& + \tilde{f}_L^* \circ \cdots \circ \tilde{f}_{\ell+1}^* \circ f_\ell^{\circ} \circ f_{\ell-1}^{\circ} \circ \cdots \circ f_1^{\circ} - \tilde{f}_L^* \circ \cdots \circ \tilde{f}_{\ell+1}^* \circ \tilde{f}_\ell^* \circ f_{\ell-1}^{\circ} \circ \cdots \circ f_1^{\circ} \\
& \qquad \vdots \\
& + \tilde{f}_L^* \circ \cdots \tilde{f}_2^* \circ f_1^{\circ} - \tilde{f}_L^* \circ \cdots \tilde{f}_2^* \circ \tilde{f}_1^*\|_{L_2(P(X))} \\
\leq& \sum_{\ell=1}^{L} \|\tilde{f}_L^* \circ \cdots \circ \tilde{f}_{\ell+1}^* \circ f_\ell^{\circ} \circ f_{\ell-1}^{\circ} \circ \cdots \circ f_1^{\circ} - \tilde{f}_L^* \circ \cdots \circ \tilde{f}_{\ell+1}^* \circ \tilde{f}_\ell^* \circ f_{\ell-1}^{\circ} \circ \cdots \circ f_1^{\circ}\|_{L_2(P(X))}.
\end{aligned}$$

Then combining the argument given above, we have

$$\begin{aligned}
& \|\tilde{f}_L^* \circ \cdots \circ \tilde{f}_{\ell+1}^* \circ f_\ell^{\circ} \circ f_{\ell-1}^{\circ} \circ \cdots \circ f_1^{\circ} - \tilde{f}_L^* \circ \cdots \circ \tilde{f}_{\ell+1}^* \circ \tilde{f}_\ell^* \circ f_{\ell-1}^{\circ} \circ \cdots \circ f_1^{\circ}\|_{L_2(P(X))} \\
\leq& (\sqrt{c_1 c_\delta} R)^{L-\ell} (\sqrt{c_0 \lambda_\ell} R) = \sqrt{(c_1 c_\delta)^{L-\ell} c_0} R^{L-\ell+1} \sqrt{\lambda_\ell},
\end{aligned}$$

for $\ell = 2, \dots, L$. And the right hand side is 0 for $\ell = 1$. This yields that

$$\|f^{\mathrm{o}} - \tilde{f}^*\|_{L_2(P(X))} \leq \sum_{\ell=2}^{L} R^{L-\ell+1} \sqrt{(c_1 c_\delta)^{L-\ell} c_0} \sqrt{\lambda_\ell}.$$

By substituting $W^{(\ell)}$ and $b^{(\ell)}$ for $\ell = 1, \dots, L$ defined above into the definition of f^*, then it is easy to see that

$$f^* = \tilde{f}^*$$

as a function. Then, we obtain the assertion.

□

A.2 Bounding the L_∞-norm

The next lemma shows the L_∞-norm of the true function f^{o} and that of $f \in \mathcal{F}$.

Lemma 3. *Under Assumptions 1, 2 and 3, the L_∞-norms of f^{o} and that of $f \in \mathcal{F}$ are bounded as*

$$\|f^{\mathrm{o}}\|_\infty \leq R^L D_x + \sum_{\ell=1}^{L} R^{L-\ell} R_b,$$
$$\|f\|_\infty \leq (\sqrt{c_1 c_\delta})^L R^L D_x + \sum_{\ell=1}^{L} (\sqrt{c_1 c_\delta} R)^{L-\ell} \bar{R}_b.$$

Proof. Suppose that

$$\|F^{\mathrm{o}}_{\ell-1}(x, \cdot)\|_{L_2(Q_\ell)} \leq G.$$

Then, F^{o}_ℓ can be bounded inductively: for all $\tau \in \mathcal{T}_{\ell+1}$

$$\begin{aligned} |F^{\mathrm{o}}_\ell(x, \tau)| &= \left| \int_{\mathcal{T}_\ell} h^{\circ}_\ell(\tau, w) \eta(F^{\mathrm{o}}_{\ell-1}(x, w)) \mathrm{d}Q_\ell(w) + b^{\circ}_\ell(\tau) \right| \\ &\leq \|h^{\circ}_\ell(\tau, \cdot)\|_{L_2(Q_\ell)} \|F^{\mathrm{o}}_{\ell-1}(x, \cdot)\|_{L_2(Q_\ell)} + |b^{\circ}_\ell(\tau)| \\ &\leq RG + R_b, \end{aligned}$$

by Assumption 1. Similarly, as for $\ell = 1$, it holds that, for all $\tau \in \mathcal{T}_2$ and $x \in \mathbb{R}^{d_{\mathrm{x}}}$,

$$\begin{aligned} |f^{\mathrm{o}}_1(x, \tau)| &= |\sum_{i=1}^{d_{\mathrm{x}}} h^{\circ}_1(\tau, i) x_i Q_1(i) + b^{\circ}_1(\tau)| \\ &\leq |\sum_{i=1}^{d_{\mathrm{x}}} h^{\circ}_1(\tau, i) x_i Q_1(i)| + |b^{\circ}_1(\tau)| \\ &\leq \|h^{\circ}_1(\tau, \cdot)\|_{L_2(Q_1)} \|x\|_{L_2(Q_1)} + R_b \\ &\leq R D_x + R_b. \end{aligned}$$

Applying the same argument recursively, we have

$$\|f^{\mathrm{o}}\|_\infty \leq R^L D_x + \sum_{\ell=1}^{L} R^{L-\ell} R_b.$$

We can bound the L_∞-norm of any $f \in \mathcal{F}$ through a similar argument. Note that $W^{(\ell)}$ satisfies $\|W^{(\ell)}\|_{\mathrm{F}} \leq \sqrt{c_1 c_\delta} R$ for $\ell = 1, \dots, L-1$, $W^{(L)}$ satisfies $\|W^{(L)}\|_{\mathrm{F}} \leq \sqrt{c_1} R$, and $b^{(\ell)}$ satisfies $\|b^{(\ell)}\|_2 \leq \sqrt{c_\delta} R_b$ by its construction. Therefore, though a similar argument to the bound for f^{o}, we have that

$$\begin{aligned}
\|f\|_\infty \le & \sqrt{c_1}R\left[\prod_{\ell=2}^{L-1}(\sqrt{c_1 c_\delta}R)\right]\sqrt{c_\delta}RD_x \\
& + \left(\sum_{\ell=1}^{L-2}\sqrt{c_1}R\left[\prod_{\ell'=\ell+1}^{L-1}(\sqrt{c_1 c_\delta}R)\right]\sqrt{c_\delta}R_b + \sqrt{c_1}R\sqrt{c_\delta}R_b + \sqrt{c_\delta}R_b\right) \\
& \le (c_1 c_\delta)^{L/2}R^L D_x + \sum_{\ell=1}^{L}(\sqrt{c_1 c_\delta}R)^{L-\ell}\bar{R}_b.
\end{aligned}$$

□

B Bounding the posterior contraction rate (proof of Theorem 2)

In this section, we prove Theorem 2. The proof is divided into two parts: posterior contraction rate with respect to the in-sample error (i.e., the empirical L_2-norm $\|f\|_n = \sqrt{\sum_{i=1}^n f(x_i)^2/n}$) and that with respect to the out-of-sample error (i.e., the population L_2-norm $\|f\|_{L_2(P_X)} = \sqrt{\int f(X)^2 \mathrm{d}P(X)}$).

Here, let

$$\epsilon_n = \hat{\delta}_{1,n} + \sigma\hat{\delta}_{2,n}, \quad \tilde{\epsilon}_n = \hat{\delta}_{1,n} + \hat{\delta}_{2,n}.$$

B.1 In-sample error

Here we show the in-sample error bound. Let $X_n = (x_1, \dots, x_n)$, $Y_n = (y_1, \dots, y_n)$ and $D_n = (X_n, Y_n)$. For given X_n, the probability distribution of Y_n associated with a function f (i.e., $y_i = f(x_i) + \epsilon_i$) is denoted by $P_{n,f}$. The expectation of a function h of Y_n with respect to $P_{n,f}$ is denoted by $P_{n,f}(h)$. The density function of $P_{n,f}$ with respect to Y_n is denoted by $p_{n,f}$.

For $\tilde{r} \ge 1$, let $\mathcal{A}_{\tilde{r}}$ be the event such that

$$\int \frac{p_{n,f}(Y_n)}{p_{n,f^{\mathrm{o}}}(Y_n)}\Pi(\mathrm{d}f) \ge \exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2)\Pi(f : \|f - f^*\|_\infty \le \hat{\delta}_{2,n}\tilde{r}).$$

The probability of this event is bounded by Lemma 4.

Using a test function ϕ_n defined later (here, a test function is a measurable function of D_n that takes its value in $[0,1]$), we decompose the expected posterior mass as

$$\begin{aligned}
& \mathrm{E}\left[\Pi(\|f - f^{\mathrm{o}}\|_n \ge \sqrt{2}\epsilon_n r | D_n)\right] \\
\le & \mathrm{E}\left[\phi_n\right] + P(\mathcal{A}_{\tilde{r}}^c) \\
& + \mathrm{E}[(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\Pi(f \in \mathcal{F}^c | D_n)] \\
& + \mathrm{E}[(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\Pi(f \in \mathcal{F} : \|f - f^{\mathrm{o}}\|_n^2 \ge 2\epsilon r^2 | D_n)] \\
=: & A_n + B_n + C_n + D_n,
\end{aligned} \tag{S-3}$$

for $\epsilon_n > 0$ where the expectation is taken with respect to $D_n = (X_n, Y_n)$ distributed from the true distribution. We give an upper bound of A_n, B_n, C_n and D_n in the following.

Step 1.

For arbitrary $r' > 0$, define $C_{r'} = \{f \in \mathcal{F} \mid r' \le \sqrt{n}\|f - f^{\mathrm{o}}\|_n/\sigma\}$. We construct a maximum cardinality set $\Theta_{r'} \subset C_{r'}$ such that each $f, f' \in \Theta_{r'}$ satisfies $\sqrt{n}\|f - f'\|_n/\sigma \ge r'/2$. Here we denote by $D(\epsilon, \mathcal{F}, \|\cdot\|)$ the ϵ-packing number of a normed space $\mathcal{F}$ attached with a norm $\|\cdot\|$. Then, the cardinality of $\Theta_{r'}$ is equal to

$D(r'/2, C_{r'}, \sqrt{n}\|\cdot\|_n/\sigma)$. Then, following Lemma 13 of van der Vaart and van Zanten (2011), one can construct a test $\tilde{\phi}_{r'}$ such that

$$P_{n,f^\circ}\tilde{\phi}_{r'} \le 9D(r'/2, C_{r'}, \sqrt{n}\|\cdot\|_n/\sigma)e^{-\frac{1}{8}r'^2} \le 9D(r'/2, \mathcal{F}, \sqrt{n}\|\cdot\|_n/\sigma)e^{-\frac{1}{8}r'^2},$$
$$\sup_{f \in C_{r'}} P_{n,f}(1-\tilde{\phi}_{r'}) \le e^{-\frac{1}{8}r'^2},$$

for any $r' > 0$.

Substituting $\sqrt{2}\sqrt{n}\epsilon_n r/\sigma$ into r' and denoting $\phi_n = \tilde{\phi}_{\sqrt{2}\sqrt{n}\epsilon_n r/\sigma}$, we obtain

$$P_{n,f^\circ}\phi_n \le 9e^{-\frac{1}{4\sigma^2}n\epsilon_n^2 r^2 + \log(D(r'/2,\mathcal{F},\sqrt{n}\|\cdot\|_n/\sigma))} \tag{S-4}$$
$$\sup_{f\in C_{2\sqrt{2}\sqrt{n}\epsilon_n r}} P_{n,f}(1-\phi_n) \le e^{-\frac{1}{4\sigma^2}n\epsilon_n^2 r^2}. \tag{S-5}$$

Hence, we just need to evaluate the (log-)packing number $\log(D(r'/2, \mathcal{F}, \sqrt{n}\|\cdot\|_n/\sigma))$ where $r' = \sqrt{2n}\epsilon_n r/\sigma$. It is known that the packing number is bounded from above by the internal covering number[1], and the packing number of unit ball in d-dimensional Euclidean space and that of the covering number is bounded as

$$D(\epsilon, \mathcal{B}_d(1), \|\cdot\|) \le N(\epsilon, \mathcal{B}_d(1), \|\cdot\|) \le \left(\frac{4+\epsilon}{\epsilon}\right)^d.$$

Based on this we evaluate the packing number of $\mathcal{F}$.

Let $f, f' \in \mathcal{F}$ be two functions corresponding to parameters $(W^{(\ell)}, b^{(\ell)})_{\ell=1}^L$ and $(W'^{(\ell)}, b'^{(\ell)})_{\ell=1}^L$. Notice that if $\|W^{(\ell)} - W'^{(\ell)}\|_{\mathrm{F}} \le \epsilon$ and $\|b^{(\ell)} - b'^{(\ell)}\| \le \epsilon$, then

$$\|f - f'\|_\infty \le L\epsilon\bar{R}^{L-1}D_x + \sum_{\ell=1}^L \epsilon\bar{R}^{L-\ell} = \epsilon(L\bar{R}^{L-1}D_x + \sum_{\ell=1}^L \bar{R}^{L-\ell}). \tag{S-6}$$

Therefore, if $\epsilon \le \delta/\hat{G}$ where

$$\hat{G} = (L\bar{R}^{L-1}D_x + \sum_{\ell=1}^L \bar{R}^{L-\ell}),$$

then $\|f - f'\|_\infty \le \delta$. Hence, the packing number of the function space $\mathcal{F}$ can be bounded by using that of the parameter space as

$$\begin{aligned}
&\log(D(r'/2, \mathcal{F}, \sqrt{n}\|\cdot\|_n/\sigma)) = \log(D(r'/2, \mathcal{F}, \sqrt{n}\|\cdot\|_n/\sigma)) \le \log(D(\sigma r'/(2\sqrt{n}), \mathcal{F}, \|\cdot\|_\infty)) \\
&\le \log(N(\sigma r'/(2\sqrt{n}), \mathcal{F}, \|\cdot\|_\infty)) \\
&\le \sum_{\ell=1}^L \log(N(\sigma r'/(2\sqrt{n}\hat{G}), \mathcal{B}_{m_{\ell+1}\times m_\ell}(\bar{R}), \|\cdot\|)) + \sum_{\ell=1}^L \log(N(\sigma r'/(2\sqrt{n}\hat{G}), \mathcal{B}_{m_\ell}(\bar{R}_b), \|\cdot\|)) \\
&\le \sum_{\ell=1}^L m_{\ell+1}m_\ell \log\left(\frac{4 + \frac{\sigma r'}{2\sqrt{n}\hat{G}\bar{R}}}{\frac{\sigma r'}{2\sqrt{n}\hat{G}\bar{R}}}\right) + \sum_{\ell=1}^L m_\ell \log\left(\frac{4 + \frac{\sigma r'}{2\sqrt{n}\hat{G}\bar{R}_b}}{\frac{\sigma r'}{2\sqrt{n}\hat{G}\bar{R}_b}}\right) \\
&= \sum_{\ell=1}^L m_{\ell+1}m_\ell \log\left(1 + \frac{4\sqrt{2}\hat{G}\bar{R}}{\epsilon_n r}\right) + \sum_{\ell=1}^L m_\ell \log\left(1 + \frac{4\sqrt{2}\hat{G}\bar{R}_b}{\epsilon_n r}\right).
\end{aligned} \tag{S-7}$$

Therefore, by Eq. (S-4), we have that

$$A_n \le 9\exp\left[-\frac{1}{4\sigma^2}n\epsilon_n^2 r^2 + \sum_{\ell=1}^L m_{\ell+1}m_\ell \log\left(1 + \frac{4\sqrt{2}\hat{G}\bar{R}}{\epsilon_n r}\right) + \sum_{\ell=1}^L m_\ell \log\left(1 + \frac{4\sqrt{2}\hat{G}\bar{R}_b}{\epsilon_n r}\right)\right].$$

[1] The ϵ-internal covering number of a (semi)-metric space (T, d) is the minimum cardinality of a finite set such that every element in T is in distance ϵ from the finite set with respect to the metric d. We denote by $N(\epsilon, T, d)$ the ϵ-internal covering number of (T, d).

Step 2. Here, we evaluate B_n. It can be evaluated by Lemma 4 as

$$B_n \leq \exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)) + \exp(-n\hat{\delta}_{1,n}^2(\tilde{r}^2-1)^2/(11\hat{R}_\infty^2)).$$

Step 3. Since $\mathcal{F}$ is the support of the prior distribution, it is obvious that $C_n = 0$.

Step 4. Here, we evaluate D_n. Remind that D_n is defined as

$$D_n = \mathrm{E}_{X_n}\left[P_{n,f^\circ}[\Pi(f \in \mathcal{F} : \|f - f^\circ\|_n > \sqrt{2}\epsilon r | Y_n)(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}]\right].$$

Define

$$\Xi_n(\tilde{r}) := -\log(\Pi(f : \|f - f^*\|_\infty \leq \hat{\delta}_{2,n}\tilde{r}))$$

for $\tilde{r} > 0$. Then, D_n can be bounded as

$$\begin{aligned}
D_n &= \mathrm{E}_{X_n}\left\{P_{n,f^\circ}\left[\frac{\int_{\mathcal{F}} \mathbf{1}\{f : \|f-f^\circ\|_n > \sqrt{2}\epsilon r\} p_{n,f}\mathrm{d}\Pi(f)}{\int_{\mathcal{F}} p_{n,f}\mathrm{d}\Pi(f)}(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\right]\right\} \\
&= \mathrm{E}_{X_n}\left\{P_{n,f^\circ}\left[\frac{\int_{\mathcal{F}} \mathbf{1}\{f : \|f-f^\circ\|_n > \sqrt{2}\epsilon r\} \frac{p_{n,f}}{p_{n,f^\circ}}\mathrm{d}\Pi(f)}{\int_{\mathcal{F}} \frac{p_{n,f}}{p_{n,f^\circ}}\mathrm{d}\Pi(f)}(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\right]\right\} \\
&\leq \mathrm{E}_{X_n}\left\{P_{n,f^\circ}\left[\int_{f\in\mathcal{F}:\|f-f^\circ\|_n > \sqrt{2}\epsilon r} p_{n,f}/p_{n,f^\circ}\mathrm{d}\Pi(f)\exp(n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2 + \Xi_n(\tilde{r}))(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\right]\right\} \\
&= \mathrm{E}_{X_n}\left\{\int_{f\in\mathcal{F}:\|f-f^\circ\|_n > \sqrt{2}\epsilon r} P_{n,f}[(1-\phi_n)\mathbf{1}_{\mathcal{A}_{\tilde{r}}}]\exp(n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2 + \Xi_n(\tilde{r}))\mathrm{d}\Pi(f)\right\} \\
&\leq \exp\left(\frac{n\tilde{\epsilon}_n^2\tilde{r}^2}{\sigma^2} + \Xi_n(\tilde{r}) - \frac{n\epsilon_n^2 r^2}{4\sigma^2}\right).
\end{aligned}$$

By using the relation (S-6), the prior mass $\Xi_n(\tilde{r})$ can be bounded as

$$\begin{aligned}
\Xi_n(\tilde{r}) &= -\log(\Pi(f : \|f - f^*\|_\infty \leq \hat{\delta}_{2,n}\tilde{r})) \\
&\leq -\log(\Pi(f : \|f - f^*\|_\infty \leq \hat{\delta}_{2,n})) \\
&\leq -\sum_{\ell=1}^{L}\log(\Pi(W^{(\ell)} : \|W^{(\ell)} - W^{*(\ell)}\|_{\mathrm{F}} \leq \hat{\delta}_{2,n}/\hat{G})) \\
&\quad\ -\sum_{\ell=1}^{L}\log(\Pi(b^{(\ell)} : \|b^{(\ell)} - b^{*(\ell)}\|_2 \leq \hat{\delta}_{2,n}/\hat{G})) \\
&\leq \sum_{\ell=1}^{L} m_\ell m_{\ell+1}\log(\bar{R}\hat{G}/(\hat{\delta}_{2,n}/2)) + \sum_{\ell=1}^{L} m_\ell \log(\bar{R}_b\hat{G}/(\hat{\delta}_{2,n}/2)).
\end{aligned} \tag{S-8}$$

Step 5. Finally, we combine the results obtained above.

$$
\begin{aligned}
&\mathrm{E}\left[\Pi(\|f-f^{\mathrm{o}}\|_n \geq \sqrt{2}\epsilon_n r|Y_n)\right] \\
\leq& 9\exp\left[-\frac{1}{4\sigma^2}n\epsilon_n^2 r^2 + \sum_{\ell=1}^{L} m_{\ell+1}m_\ell \log\left(1+\frac{4\sqrt{2}\hat{G}\bar{R}}{\epsilon_n r}\right) + \sum_{\ell=1}^{L} m_\ell \log\left(1+\frac{4\sqrt{2}\hat{G}\bar{R}_b}{\epsilon_n r}\right)\right] \\
&+\exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)) + \exp(-n\hat{\delta}_{1,n}^2(\tilde{r}^2-1)^2/(11\hat{R}_\infty^2)) \\
&+\exp\left(\frac{n}{\sigma^2}\tilde{\epsilon}_n^2\tilde{r}^2 + \Xi_n(\tilde{r}) - \frac{n\epsilon_n^2 r^2}{4\sigma^2}\right). \qquad \text{(S-9)}
\end{aligned}
$$

Now, let $1 \leq \tilde{r} \leq r$. Then, since $\epsilon_n \geq \hat{\delta}_{2,n}$ and $r \geq 1$, we have that

$$
\max\left\{\log\left(\frac{2\hat{G}R'}{\hat{\delta}_{2,n}}\right), \log\left(1+\frac{4\sqrt{2}\hat{G}R'}{\epsilon_n r}\right)\right\} \leq \log\left(1+\frac{4\sqrt{2}\hat{G}R'}{\hat{\delta}_{2,n}}\right),
$$

for all $R' > 0$. Now, we set $\hat{\delta}_{2,n}$ to satisfy

$$
\frac{n\hat{\delta}_{2,n}^2}{\sigma^2} \geq \sum_{\ell=1}^{L} m_\ell m_{\ell+1}\log\left(1+\frac{4\sqrt{2}\hat{G}\bar{R}}{\hat{\delta}_{2,n}}\right) + \sum_{\ell=1}^{L} m_\ell \log\left(1+\frac{4\sqrt{2}\hat{G}\bar{R}_b}{\hat{\delta}_{2,n}}\right) (\geq \Xi_n(\tilde{r})), \qquad \text{(S-10)}
$$

which can be satisfied by

$$
\hat{\delta}_{2,n}^2 = \frac{2\sigma^2}{n}\sum_{\ell=1}^{L} m_\ell m_{\ell+1}\log_+\left(1+\frac{4\sqrt{2}\hat{G}\max\{\bar{R},\bar{R}_b\}\sqrt{n}}{\sigma\sqrt{\sum_{\ell=1}^{L} m_\ell m_{\ell+1}}}\right).
$$

Then, by noticing $n\hat{\delta}_{2,n}^2 \leq n\tilde{\epsilon}_n^2$ and Eq. (S-8), the RHS of Eq. (S-9) is upper bounded by

$$
\exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)) + \exp(-n\hat{\delta}_{1,n}^2(\tilde{r}^2-1)^2/(11\hat{R}_\infty^2)) + 10\exp\left[2\frac{n}{\sigma^2}\tilde{\epsilon}_n^2\tilde{r}^2 - \frac{n\epsilon_n^2 r^2}{4\sigma^2}\right].
$$

Here, by setting $r^2 = 12\tilde{r}^2 \geq 12$, then the RHS is further bounded as

$$
\begin{aligned}
&\exp(-n\hat{\delta}_{1,n}^2(\tilde{r}^2-1)^2/(11\hat{R}_\infty^2)) + \exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)) + 10\exp(-n\epsilon_n^2\tilde{r}^2/\sigma^2) \\
\leq& \exp\left[-n\hat{\delta}_{1,n}^2(\tilde{r}^2-1)^2/(11\hat{R}_\infty^2)\right] + 11\exp(-n\epsilon_n^2\tilde{r}^2/(8\sigma^2)).
\end{aligned}
$$

Lemma 4. *Then, for any $\tilde{r} > 1$, it holds that*

$$
\begin{aligned}
&P_{D_n}\left(\int \frac{p_{n,f}(Y_n)}{p_{n,f^{\mathrm{o}}}(Y_n)}\Pi(\mathrm{d}f) \geq \exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2)\Pi(f:\|f-f^*\|_\infty \leq \hat{\delta}_{2,n}\tilde{r})\right) \\
&\geq 1-\exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)) - \exp(-n\hat{\delta}_{1,n}^2\min\{(\tilde{r}^2-1)^2, \tilde{r}^2-1\}/(11\hat{R}_\infty^2)).
\end{aligned}
$$

Proof. Note that Lemma 14 of van der Vaart and van Zanten (2011) showed that

$$
P_{Y_n|X_n}\left(\int \frac{p_{n,f}(Y_n)}{p_{n,f^{\mathrm{o}}}(Y_n)}\Pi(\mathrm{d}f) \geq \exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2)\Pi(f:\|f-f^{\mathrm{o}}\|_n \leq \tilde{\epsilon}_n\tilde{r})\right) \geq 1-\exp(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)).
$$

where $P_{Y_n|X_n}$ represents the conditional distribution of $Y_n = (y_i)_{i=1}^n$ conditioned by $X_n = (x_i)_{i=1}^n$. Therefore the proof is reduced to show $\|f-f^{\mathrm{o}}\|_n \leq \hat{\delta}_{1,n}\tilde{r} + \|f-f^*\|_\infty$ with high probability. Note that

$$
\|f-f^{\mathrm{o}}\|_n \leq \|f-f^*\|_n + \|f^*-f^{\mathrm{o}}\|_n \leq \|f-f^*\|_\infty + \|f^*-f^{\mathrm{o}}\|_n.
$$

Hence, we just need to show $\|f^*-f^{\mathrm{o}}\|_n^2 \leq \hat{\delta}_{1,n}^2 + \tilde{r}'\|f^*-f^{\mathrm{o}}\|_{L_2(P_X)}^2 (\leq (1+\tilde{r}')\hat{\delta}_{1,n}^2)$ with high probability for appropriately chosen $\tilde{r}'$. This can be shown by Bernstein's inequality:

$$
P\left(\|f^*-f^{\mathrm{o}}\|_{L_2(P_X)}^2 + \tilde{r}'\hat{\delta}_{1,n}^2 \leq \|f^*-f^{\mathrm{o}}\|_n^2\right) \leq \exp\left(-\frac{n\tilde{r}'^2\hat{\delta}_{1,n}^4}{2(v+\tilde{r}'\|f^*-f^{\mathrm{o}}\|_\infty^2\hat{\delta}_{1,n}^2/3)}\right),
$$

where $v = \mathrm{E}_X[((f^*(X) - f^{\mathrm{o}}(X))^2 - \|f^* - f^{\mathrm{o}}\|_{L_2(P_X)}^2)^2]$. Now $v \leq \mathrm{E}_X[(f^*(X) - f^{\mathrm{o}}(X))^4] \leq \|f^* - f^{\mathrm{o}}\|_\infty^2 \|f^* - f^{\mathrm{o}}\|_{L_2(P_X)}^2 \leq \|f^* - f^{\mathrm{o}}\|_\infty^2 \hat{\delta}_{1,n}^2$. This yields that

$$P\left(\|f^* - f^{\mathrm{o}}\|_{L_2(P_X)}^2 + \tilde{r}'\hat{\delta}_{1,n}^2 \leq \|f^* - f^{\mathrm{o}}\|_n^2\right) \leq \exp\left[-\frac{3n\min\{\tilde{r}'^2, \tilde{r}'\}}{8}\frac{\hat{\delta}_{1,n}^2}{\|f^* - f^{\mathrm{o}}\|_\infty^2}\right]. \tag{S-11}$$

Since $\|f^* - f^{\mathrm{o}}\|_\infty \leq 2\hat{R}_\infty$, the RHS is further bounded by $\exp\left(-\frac{3n\min\{\tilde{r}'^2,\tilde{r}'\}\hat{\delta}_{1,n}^2}{32\hat{R}_\infty^2}\right)$.

Therefore, with probability $1 - \exp\left(-\frac{3n\hat{\delta}_{1,n}^2\min\{\tilde{r}'^2,\tilde{r}'\}}{32\hat{R}_\infty^2}\right)$, it holds that

$$\|f - f^{\mathrm{o}}\|_n \leq \|f - f^*\|_\infty + \sqrt{\|f^* - f^{\mathrm{o}}\|_{L_2(P_X)}^2 + \tilde{r}'\hat{\delta}_{1,n}^2} \leq \|f - f^*\|_\infty + \sqrt{1 + \tilde{r}'}\hat{\delta}_{1,n}$$

for all f such that $\|f\|_\infty < \infty$. Thus by setting $\tilde{r}'$ so that $\tilde{r} = \sqrt{1 + \tilde{r}'}$, we obtain the assertion. □

B.2 Out of sample error

Now, we are going to show the posterior contraction rate with respect to the out-of-sample predictive error:

$$\mathrm{E}_{D_n}\left[\Pi(f : \|f - f^{\mathrm{o}}\|_{L_2(P_X)} \geq \epsilon_n r | D_n)\right], \tag{S-12}$$

for sufficiently large $r \geq 1$.

To bound the posterior tail, we divide that into four parts:

$$\begin{aligned}
&\mathrm{I} = \mathrm{E}_{D_n}\left[\mathbf{1}_{\mathcal{A}_{\tilde{r}}^c}\right], \\
&\mathrm{II} = \mathrm{E}_{D_n}\left[\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\Pi(f : \sqrt{2}\|f - f^{\mathrm{o}}\|_n > \epsilon_n r,\ \|f\|_\infty \leq \hat{R}_\infty \mid D_n)\right], \\
&\mathrm{III} = \mathrm{E}_{D_n}\left[\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\Pi(f : \|f - f^{\mathrm{o}}\|_{L_2(P_X)} > \epsilon_n r \geq \sqrt{2}\|f - f^{\mathrm{o}}\|_n,\ \|f\|_\infty \leq \hat{R}_\infty \mid D_n)\right], \\
&\mathrm{IV} = \mathrm{E}_{D_n}\left[\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\Pi(f : \|f\|_\infty > \hat{R}_\infty \mid D_n)\right].
\end{aligned}$$

The term I and II are already evaluated in Section B.1, that is, $\mathrm{I} + \mathrm{II}$ is bounded by the right hand side of Eq. (S-3) which is what we have upper bounded in Section B.1.

The term III is bounded as follows. To bound this, we need to evaluate the difference between the empirical norm $\|f - f^{\mathrm{o}}\|_n$ and the expected norm $\|f - f^{\mathrm{o}}\|_{L_2(P_X)}$, which can be done by Bernstein's inequality. Following the same argument to derive Eq. (S-11), it holds that

$$P\left(\|f - f^{\mathrm{o}}\|_{L_2(P_X)} \geq \sqrt{2}\|f - f^{\mathrm{o}}\|_n\right) \leq \exp\left(-\frac{n\|f - f^{\mathrm{o}}\|_{L_2(P_X)}^2}{11\hat{R}_\infty^2}\right).$$

Therefore, we arrive at the following bound of III:

$$\begin{aligned}
\mathrm{III} &\leq \mathrm{E}_{X_n}\left[P_{n,f^{\mathrm{o}}}\left[\int_{f\in\mathcal{F}:\|f-f^{\mathrm{o}}\|_{L_2(P_X)}>\epsilon_n r\geq\sqrt{2}\|f-f^{\mathrm{o}}\|_n} p_{n,f}/p_{n,f^{\mathrm{o}}}\mathrm{d}\Pi(f)\right]\exp(n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2 + \Xi_n(\tilde{r}))\mathbf{1}_{\mathcal{A}_{\tilde{r}}}\right] \\
&\leq \exp(n\tilde{\epsilon}_n^2\tilde{r}^2/\sigma^2 + \Xi_n(\tilde{r}))\int_{f\in\mathcal{F}:\|f-f^{\mathrm{o}}\|_{L_2(P_X)}>\epsilon_n r} P(\|f - f^{\mathrm{o}}\|_{L_2(P_X)} \geq \sqrt{2}\|f - f^{\mathrm{o}}\|_n)\mathrm{d}\Pi(f) \\
&\leq \exp\left(\frac{n\tilde{\epsilon}_n^2\tilde{r}^2}{\sigma^2} + \Xi_n(\tilde{r}) - \frac{n\epsilon_n^2 r^2}{11\hat{R}_\infty^2}\right) \\
&\leq \exp\left(\frac{2n\tilde{\epsilon}_n^2\tilde{r}^2}{\sigma^2} - \frac{n\epsilon_n^2 r^2}{11\hat{R}_\infty^2}\right).
\end{aligned}$$

Finally, since all $f \in \mathcal{F}$ satisfies $\|f\|_\infty \leq \hat{R}_\infty$, $\mathrm{IV} = 0$.

Combining the results we arrive at

$$\mathrm{E}_{D_n}\left[\Pi(f : \|f - f^{\mathrm{o}}\|_{L_2(P_X)} \geq \epsilon_n r | D_n)\right] \leq \exp\left[-\frac{n\hat{\delta}_{1,n}^2 \min\{(\tilde{r}^2-1)^2, \tilde{r}^2-1\}}{11\hat{R}_\infty^2}\right] + 12\exp\left(-n\tilde{\epsilon}_n^2\tilde{r}^2/(8\sigma^2)\right),$$

for all $\tilde{r} \geq 1$ and $r \geq \max\{12, 33\hat{R}_\infty^2/\sigma^2\}\tilde{r}^2$. This concludes the proof of Theorem 2.

C Convergence rate for the empirical risk minimizer (proof of Theorem 1)

In this section, we give the proof of Theorem 1 in the main text. To show that, we prepare some lemmas.

Proposition 2 (Gaussian concentration inequality (Theorem 2.5.8 in Giné and Nickl (2015)))**.** *Let $(\xi_i)_{i=1}^n$ be i.i.d. Gaussian sequence with mean 0 and variance σ^2, and $(x_i)_{i=1}^n \subset \mathcal{X}$ be a given set of input variables. Then, for a set $\tilde{\mathcal{F}}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ which is separable with respect to L_∞-norm and $\sup_{f\in\tilde{\mathcal{F}}}\left|\sum_{i=1}^n \frac{1}{n}\xi_i f(x_i)\right| < \infty$ almost surely, it holds that for every $r > 0$,*

$$P\left(\sup_{f\in\tilde{\mathcal{F}}}\left|\sum_{i=1}^n \frac{1}{n}\xi_i f(x_i)\right| \geq \mathrm{E}\left[\sup_{f\in\tilde{\mathcal{F}}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right|\right] + r\right) \leq \exp[-nr^2/2(\sigma\|\tilde{\mathcal{F}}\|_n)^2]$$

where $\|\tilde{\mathcal{F}}\|_n^2 = \sup_{f\in\tilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^n f(x_i)^2$. Here the probability is taken with respect to $(\xi_i)_{i=1}^n$.

Remind that every $f \in \mathcal{F}$ satisfies $\|f\|_n \leq \|f\|_\infty \leq \hat{R}_\infty$. Hence $\|\mathcal{F}\|_n \leq \hat{R}_\infty$. For an observation $(x_i)_{i=1}^n$, let $\mathcal{G}_\delta = \{f - f^* \mid \|f - f^*\|_n \leq \delta, f \in \mathcal{F}\}$. It is obvious that $\mathcal{G}_\delta$ is separable with respect to L_∞-norm. Then, by the Gaussian concentration inequality, we have that

$$P\left(\sup_{f\in\mathcal{G}_\delta}\left|\sum_{i=1}^n \frac{1}{n}\xi_i f(x_i)\right| \geq \mathrm{E}\left[\sup_{f\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right|\right] + r\right) \leq \exp[-nr^2/2(\sigma\delta)^2]$$

for every $r > 0$. By applying this inequality for $\delta_j = 2^{j-1}\sigma/\sqrt{n}$ for $j = 1, \dots, \lceil\log_2(\hat{R}_\infty\sqrt{n}/\sigma)\rceil$ and using the uniform bound, we can show that, for every $r > 0$, with probability $\lceil\log_2(\hat{R}_\infty\sqrt{n}/\sigma)\rceil\exp[-nr^2/2\sigma^2]$, it holds that

$$\left|\frac{1}{n}\sum_{i=1}^n \xi_i(f(x_i) - f^*(x_i))\right| \geq \mathrm{E}\left[\left|\sup_{f\in\mathcal{G}_{2\delta}}\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right|\right] + 2\delta r$$

uniformly for all $f \in \mathcal{G}_\delta$ where δ is any positive real satisfying $\delta \geq \sigma/\sqrt{n}$.

Lemma 5. *There exists a universal constant C such that for any δ it holds that*

$$\mathrm{E}\left[\left|\sup_{f\in\mathcal{G}_{2\delta}}\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right|\right] \leq C\sigma\delta\sqrt{\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log_+\left(1 + \frac{4\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\delta}\right)}.$$

Proof. Since $f \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(x_i)$ is a sub-Gaussian process relative to the metric $\|\cdot\|_n$. By the chaining argument (see, for example, Theorem 2.3.6 of Giné and Nickl (2015)), it holds that

$$\mathrm{E}\left[\left|\sup_{f\in\mathcal{G}_{2\delta}}\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right|\right] \leq 4\sqrt{2}\frac{\sigma}{\sqrt{n}}\int_0^{2\delta}\sqrt{\log(2N(\epsilon, \mathcal{G}_{2\delta}, \|\cdot\|_n))}\mathrm{d}\epsilon.$$

Since $\log N(\epsilon, \mathcal{G}_{2\delta}, \|\cdot\|_n) \leq \log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq 2\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log\left(1 + \frac{4\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\epsilon}\right)$, the right hand side is bounded by

$$\begin{aligned}\int_0^{2\delta}\sqrt{\log(2N(\epsilon, \mathcal{F}, \|\cdot\|_n))}\mathrm{d}\epsilon &\leq \int_0^{2\delta}\sqrt{\log(2) + 2\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log\left(1 + \frac{4\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\epsilon}\right)}\mathrm{d}\epsilon \\ &\leq C\delta\sqrt{\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log_+\left(1 + \frac{4\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\delta}\right)},\end{aligned}$$

where C is a universal constant. This gives the assertion.

$\square$

Therefore, by substituting $\delta \leftarrow \left(\|f - f^*\|_n \vee \sigma\sqrt{\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}}\right)$ and $r \leftarrow \sigma r/\sqrt{n}$, the following inequality holds:

$$
\begin{aligned}
& -\frac{1}{n}\sum_{i=1}^n \xi_i (f(x_i) - f^*(x_i)) \\
& \leq C\sigma \left(\|f - f^*\|_n \vee \sqrt{\frac{\sigma^2 \sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}}\right) \sqrt{\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n} \log_+ \left(1 + \frac{4\sqrt{n}\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\sigma\sqrt{\sum_{\ell=1}^L m_\ell m_{\ell+1}}}\right)} \\
& + 2\left(\|f - f^*\|_n \vee \sqrt{\frac{\sigma^2 \sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}}\right) \sigma \frac{r}{\sqrt{n}} \\
& \leq \frac{1}{4}\left(\|f - f^*\|_n \vee \sqrt{\frac{\sigma^2 \sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}}\right)^2 \\
& + 2C^2\sigma^2 \left(\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n} \log_+ \left(1 + \frac{4\sqrt{n}\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\sigma}\right) + 4\frac{r^2}{n}\right),
\end{aligned}
$$

uniformly for all $f \in \mathcal{F}$ with probability $1 - \lceil \log_2(\hat{R}_\infty\sqrt{n}/\sigma)\rceil \exp[-r^2/2]$. Here let

$$
\Psi_{r,n} := 2C^2\sigma^2 \left(\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n} \log_+ \left(1 + \frac{4\sqrt{n}\hat{G}\max\{\bar{R}, \bar{R}_b\}}{\sigma\sqrt{\sum_{\ell=1}^L m_\ell m_{\ell+1}}}\right) + 4\frac{r^2}{n}\right).
$$

Remind that the empirical risk minimizer in the model $\mathcal{F}$ is denoted by $\widehat{f}$:

$$
\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n (y_i - f(x_i))^2.
$$

Since $\widehat{f}$ minimizes the empirical risk, it holds that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 \leq \frac{1}{n}\sum_{i=1}^n (y_i - f^*(x_i))^2 \\
\Rightarrow \quad & \frac{2}{n}\sum_{i=1}^n y_i (f^*(x_i) - \widehat{f}(x_i)) + \|\widehat{f}\|_n^2 - \|f^*\|_n^2 \leq 0 \\
\Rightarrow \quad & \frac{2}{n}\sum_{i=1}^n (\xi_i + f^{\mathrm{o}}(x_i))(f^*(x_i) - \widehat{f}(x_i)) + \|\widehat{f}\|_n^2 - \|f^*\|_n^2 \leq 0 \\
\Rightarrow \quad & \frac{2}{n}\sum_{i=1}^n \xi_i (f^*(x_i) - \widehat{f}(x_i)) + \frac{2}{n}\sum_{i=1}^n f^{\mathrm{o}}(x_i)(f^*(x_i) - \widehat{f}(x_i)) + \|\widehat{f}\|_n^2 - \|f^*\|_n^2 \leq 0 \\
\Rightarrow \quad & \frac{2}{n}\sum_{i=1}^n \xi_i (f^*(x_i) - \widehat{f}(x_i)) + \|\widehat{f} - f^{\mathrm{o}}\|_n^2 \leq \|f^* - f^{\mathrm{o}}\|_n^2.
\end{aligned}
$$

Therefore, we have

$$
-\frac{1}{4}\left(\|\widehat{f} - f^*\|_n \vee \sqrt{\frac{\sigma^2 \sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}}\right)^2 - \Psi_{r,n} + \|\widehat{f} - f^{\mathrm{o}}\|_n^2 \leq \|f^* - f^{\mathrm{o}}\|_n^2. \tag{S-13}
$$

Let us assume $\|\widehat{f}-f^*\|_n^2 \geq \frac{\sigma^2 \sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}$. Then, by Eq. (S-13), we have

$$\begin{aligned}
& -\frac{1}{4}\|\widehat{f}-f^*\|_n^2 - \Psi_{r,n} + \|\widehat{f}-f^{\mathrm{o}}\|_n^2 \leq \|f^*-f^{\mathrm{o}}\|_n^2 \\
\Rightarrow\quad & -\frac{1}{4}\|\widehat{f}-f^*\|_n^2 - \Psi_{r,n} + \frac{1}{2}\|\widehat{f}-f^*\|_n^2 - \|f^*-f^{\mathrm{o}}\|_n^2 \leq \|f^*-f^{\mathrm{o}}\|_n^2 \\
\Rightarrow\quad & \frac{1}{4}\|\widehat{f}-f^*\|_n^2 \leq 2\|f^*-f^{\mathrm{o}}\|_n^2 + \Psi_{r,n}. \qquad \text{(S-14)}
\end{aligned}$$

Otherwise, we trivially have $\|\widehat{f}-f^*\|_n^2 < \frac{\sigma^2 \sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}$.

Combining the inequalities, it holds that

$$\|\widehat{f}-f^*\|_n^2 \leq 8\|f^*-f^{\mathrm{o}}\|_n^2 + 4\Psi_{r,n} + \frac{\sigma^2 \sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}. \qquad \text{(S-15)}$$

Based on this inequality, we derive a bound for $\|\widehat{f}-f^*\|_{L_2(P_X)}$ instead of the empirical L_2-norm $\|\widehat{f}-f^*\|_n$.

Proposition 3 (Talagrand's concentration inequality (Talagrand, 1996; Bousquet, 2002))**.** *Let $(x_i)_{i=1}^n$ be an i.i.d. sequence of input variables in $\mathcal{X}$. Then, for a set $\tilde{\mathcal{F}}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ which is separable with respect to L_∞-norm and $\|f\|_\infty \leq \tilde{R}$ for all $f \in \tilde{\mathcal{F}}$, it holds that for every $r > 0$,*

$$\begin{aligned}
& P\left(\sup_{f\in\tilde{\mathcal{F}}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)^2 - \mathrm{E}[f^2]\right| \geq C\left\{\mathrm{E}\left[\sup_{f\in\tilde{\mathcal{F}}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)^2 - \mathrm{E}[f^2]\right|\right] + \sqrt{\frac{\|\tilde{\mathcal{F}}^2\|_{L_2(P_X)}^2 r}{n}} + \frac{r\tilde{R}^2}{n}\right\}\right) \\
& \leq \exp(-r)
\end{aligned}$$

where $\|\tilde{\mathcal{F}}^2\|_{L_2(P_X)}^2 = \sup_{f\in\tilde{\mathcal{F}}} \mathrm{E}[f(X)^4]$.

Let $\mathcal{G}'_\delta = \{f - f^* \mid \|f-f^*\|_{L_2(P_X)} \leq \delta, f \in \mathcal{F}\}$. By the bound $\|f\|_\infty \leq \hat{R}_\infty$ for all $f \in \mathcal{F}$ (Lemma 3), $\|g\|_\infty \leq 2\hat{R}_\infty$ for all $g \in \mathcal{G}'_\delta$. Therefore, we have $\|\mathcal{G}'^2_\delta\|_{L_2(P_X)}^2 \leq 4\hat{R}_\infty^2\delta^2$. Hence, Talagrand's concentration inequality yields that

$$\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)^2 - \mathrm{E}[f^2]\right| \geq C_1\left\{\mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)^2 - \mathrm{E}[f^2]\right|\right] + \sqrt{\frac{\delta^2\hat{R}_\infty^2 r}{n}} + \frac{r\tilde{R}^2}{n}\right\} \qquad \text{(S-16)}$$

with probability $1-\exp(-r)$ where C_1 is a universal constant.

Lemma 6. *There exists a universal constant $C > 0$ such that, for all $\delta > 0$,*

$$\begin{aligned}
& \mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)^2 - \mathrm{E}[f^2]\right|\right] \\
& \leq C\Bigg[\delta\hat{R}_\infty\sqrt{\frac{\sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}\log_+\left(1+\frac{4\hat{G}\max\{\bar{R},\bar{R}_b\}}{\delta}\right)} \\
& \quad \vee \hat{R}_\infty^2\frac{\sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}\log_+\left(1+\frac{4\hat{G}\max\{\bar{R},\bar{R}_b\}}{\delta}\right)\Bigg].
\end{aligned}$$

Proof. Let $(\epsilon_i)_{i=1}^n$ be i.i.d. Rademacher sequence. Then, by the standard argument of Rademacher complexity, we have

$$\mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)^2 - \mathrm{E}[f^2]\right|\right] \leq 2\mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)^2\right|\right]$$

(see, for example, Lemma 2.3.1 in van der Vaart and Wellner (1996)). Since $\|f\|_\infty \leq 2\hat{R}_\infty$ for all $f \in \mathcal{G}'_\delta$, the contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12) gives an upper bound of the RHS as

$$2\mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)^2\right|\right] \leq 4(2\hat{R}_\infty)\mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right|\right].$$

We further bound the RHS. By Theorem 3.1 in Giné and Koltchinskii (2006) or Lemma 2.3 of Mendelson (2002) with the covering number bound (S-7), there exists a universal constant C' such that

$$\begin{aligned}
&\mathrm{E}\left[\sup_{f\in\mathcal{G}'_\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right|\right] \\
&\leq C'\left[\delta\sqrt{\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log_+\left(1+\frac{4\hat{G}\max\{\bar{R},\bar{R}_b\}}{\delta}\right)}\right. \\
&\left.\vee\, \hat{R}_\infty\frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log_+\left(1+\frac{4\hat{G}\max\{\bar{R},\bar{R}_b\}}{\delta}\right)\right].
\end{aligned}$$

This concludes the proof. □

Let $\Phi_n := \frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n}\log_+\left(1+\frac{4\sqrt{n}\hat{G}\max\{\bar{R},\bar{R}_b\}}{\hat{R}_\infty\sqrt{\sum_{\ell=1}^L m_\ell m_{\ell+1}}}\right)$. Then, applying the inequality (S-16) for $\delta = 2^{j-1}\hat{R}_\infty/\sqrt{n}$ for $j = 1,\dots,\lceil\log_2(\sqrt{n})\rceil$, it is shown that there exists an event with probability $1-\lceil\log_2(\sqrt{n})\rceil\exp(-r)$ such that, uniformly for all $f\in\mathcal{F}$, it holds that

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n (f(x_i)-f^*(x_i))^2 - \mathrm{E}[(f-f^*)^2]\right| &\leq C_1\left[C(2\delta\hat{R}_\infty\sqrt{\Phi_n})\vee(\hat{R}_\infty^2\Phi_n)+\delta\sqrt{\frac{\hat{R}_\infty^2 r}{n}}+\frac{r\hat{R}_\infty^2}{n}\right] \\
&\leq \frac{\delta^2}{2}+2C_1^2(2C^2+1)\hat{R}_\infty^2\Phi_n+(C_1^2+C_1)\frac{\hat{R}_\infty^2 r}{n},
\end{aligned}$$

where δ is any positive real such that $\delta^2 \geq \mathrm{E}[(f-f^*)^2]$ and $\delta^2 \geq \hat{R}_\infty^2\sum_{\ell=1}^L m_\ell m_{\ell+1}/n$. The right hand side can be further bounded by

$$\frac{\delta^2}{2}+C_2\hat{R}_\infty^2\left(\Phi_n+\frac{r}{n}\right)$$

for an appropriately defined universal constant C_2. Applying this inequality for $f=\widehat{f}$ to Eq. (S-15) gives that

$$\frac{1}{2}\|\widehat{f}-f^*\|_{L_2(P_X)}^2 \leq C_2\hat{R}_\infty^2\left(\Phi_n+\frac{r}{n}\right)+8\|f^*-f^{\mathrm{o}}\|_n^2+4\Psi_{r,n}+\left(\frac{\sigma^2+\hat{R}_\infty^2}{n}\right)\sum_{\ell=1}^L m_\ell m_{\ell+1}.$$

Finally, by the Bernstein's inequality (S-11), the term $\|f^*-f^{\mathrm{o}}\|_n^2$ is bounded as

$$\|f^*-f^{\mathrm{o}}\|_n^2 \leq (1+\tilde{r}')\|f^*-f^{\mathrm{o}}\|_{L_2(P_X)}^2 \leq (1+\tilde{r}')\hat{\delta}_{1,n}^2$$

with probability $1-\exp\left(-\frac{3n\hat{\delta}_{1,n}^2\tilde{r}'^2}{32\hat{R}_\infty^2}\right)$ for every $\tilde{r}'>0$.

Combining all inequalities, we obtain that

$$\|\widehat{f}-f^*\|_{L_2(P_X)}^2 \leq 2C_2\hat{R}_\infty^2\left(\Phi_n+\frac{r}{n}\right)+16(1+\tilde{r}')\hat{\delta}_{1,n}^2+4\Psi_{r,n}+\frac{2(\sigma^2+\hat{R}_\infty^2)}{n}\sum_{\ell=1}^L m_\ell m_{\ell+1}.$$

This gives a bound for the distance between $\widehat{f}$ and f^*. However, what we want is a bound on the distance from the true function f^{o} to $\widehat{f}$. This can be accomplished by noticing that $\|\widehat{f}-f^{\mathrm{o}}\|_{L_2(P_X)}^2 \leq 2(\|\widehat{f}-f^*\|_{L_2(P_X)}^2+\|f^{\mathrm{o}}-$

$f^*\|^2_{L_2(P_X)}) \le 2\|\widehat{f} - f^*\|^2_{L_2(P_X)} + 2\hat{\delta}^2_{1,n}$, and conclude that

$$\|\widehat{f} - f^{\mathrm{o}}\|^2_{L_2(P_X)} \le 4C_2\hat{R}^2_\infty \left(\Phi_n + \frac{r}{n}\right) + (34 + 32\tilde{r}')\hat{\delta}^2_{1,n} + 8\Psi_{r,n} + \frac{4(\sigma^2 + \hat{R}^2_\infty)}{n} \sum_{\ell=1}^{L} m_\ell m_{\ell+1}.$$

More concisely, letting

$$\alpha(U) := U^2 \frac{\sum_{\ell=1}^L m_\ell m_{\ell+1}}{n} \log_+ \left(1 + \frac{4\sqrt{n}\hat{G}\max\{\bar{R}, \bar{R}_b\}}{U\sqrt{\sum_{\ell=1}^L m_\ell m_{\ell+1}}}\right),$$

the right side is further upper bounded as

$$\|\widehat{f} - f^{\mathrm{o}}\|^2_{L_2(P_X)} \le C_3 \left\{\alpha(\hat{R}_\infty) + \alpha(\sigma) + \frac{(\hat{R}^2_\infty + \sigma^2)}{n}\left[\log_+\left(\frac{\sqrt{n}}{\min\{\sigma/\hat{R}_\infty, 1\}}\right) + r\right] + (1 + \tilde{r}')\hat{\delta}^2_{1,n}\right\}$$

with probability $1 - \exp\left(-\frac{3n\hat{\delta}^2_{1,n}\tilde{r}'^2}{32\hat{R}^2_\infty}\right) - 2\exp(-r)$ for every $r > 0$ and $\tilde{r}' > 0$.

References

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. *Comptes Rendus de l'Académie des Sciences -Series I- Mathematics*, 334:495–500, 2002.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces. Isoperimetry and Processes*. Springer Berlin Heidelberg, 1991.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996.

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.