A Approximation error bound and $L_{\infty}$-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 $Q_\ell$ such that, for any $\delta \in (0, 1)$, i.i.d. sample $v_1, \ldots, v_m$ from $q_\ell$ satisfies that

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

with probability $1 - \delta$, if

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

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

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

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

then there exist $v_1, \ldots v_m \in T_\ell$, $w_1, \ldots, w_m > 0$ such that

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

and

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

**Proof.** Notice that $E[\frac{1}{m} \sum_{j=1}^m q_\ell(v_j)^{-1}] = E[q_\ell(v)^{-1}] = \int_{T_\ell} q_\ell(v)^{-1} q_\ell(v) dQ_\ell(v) = \int_{T_\ell} 1 dQ_\ell(v) = 1$, thus an i.i.d. sequence $\{v_1, \ldots, v_m\}$ satisfies $\frac{1}{m} \sum_{j=1}^m q_\ell(v_j)^{-1} \leq 1/(1 - 2\delta)$ with probability $2\delta$ by the Markov’s inequality. Combining this with Proposition 1, the i.i.d. sequence $\{v_1, \ldots, v_m\}$ 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
\[ c_0 = 4, \quad c_1 = 4, \quad c_5 = (1 - 2\delta)^{-1}. \]

Based on the proposition, we approximate \( f^* \) 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 \( \ell \)-th internal layer (we set the dimensions of the output and input layers to \( m_{L+1} = 1 \) and \( m_1 = d_x \)) and consider a model
\[
\begin{align*}
f_{\ell}^i(g) &= W^i(\eta)(g) + b^i(\ell = 2, \ldots, L), \\
f_{\ell}^i(x) &= W^i(x) + b^i(1), \\
f^i(x) &= f_L^\ell \circ f_{L-1}^\ell \circ \cdots \circ f_1^\ell(x),
\end{align*}
\]
where \( W^i(\eta) \in \mathbb{R}^{m_{\ell+1} \times m_\ell} \) and \( b^i \in \mathbb{R}^{m_{\ell+1}} \).

The next lemma gives an approximation error bound between \( f^o \) and \( f^* \). The \( L_\infty \)-norm bounds of \( f^o \) and \( f^* \) are given later in Lemma 3. Substituting \( \delta \leftarrow \delta/2 \) into the statement in the following Lemma 2 and letting \( c_5 = c_1/2 \), we derive the approximation error \( \delta_1 \) 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 \( W^i(\eta)||_2 \leq c_1 R^2 \), \( ||b^i||_2 \leq \sqrt{\lambda_\ell} R_b \), \( ||W^i||_2 \leq c_1 R^2 \), \( ||b^i||_2 \leq R_b \), and
\[
\|f^o - f^*\|_{L_2(P(X))} \leq \sum_{\ell=2}^L \sqrt{(c_1 c_5)^L-1 c_0 R^L-1 \sqrt{\lambda_\ell}.}
\]

**Proof.** We construct the asserted finite dimensional network recursively from \( \ell = L \) to \( \ell = 1 \). Let \( \{v_{i(j)}^\ell\}_{j=1}^{m_\ell} \) and \( \{w_{i(j)}^\ell\}_{j=1}^{m_\ell} \) be the sequences given in Proposition 1. Let \( \hat{\theta} = \{v_{i(j)}^\ell\}_{j=1}^{m_\ell} \). With slight abuse of notation, we identify \( f^\ell \circ F \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{m_\ell} \rightarrow \mathbb{R}^{m_{\ell+1}} \) to a function \( f^\ell : \hat{\theta} \rightarrow \hat{\theta}_{\ell+1} \) in a canonical way. For a function \( F : \mathbb{R}^{d_x} \times \hat{\theta} \rightarrow \mathbb{R} \), we denote by \( f^\ell(F(x,v)) \) to express \( f^\ell(F(x,v)) = \sum_{j=1}^{m_\ell} W_{i(j)}^\ell F(x,v_j^\ell) + b^\ell \) for \( v_j^\ell \in \hat{\theta}_{\ell+1} \). When we write \( f^\ell(F) \) for \( F : \mathbb{R}^{d_x} \times \hat{\theta} \rightarrow \mathbb{R} \), we deal with \( F \) as a restriction of \( F \) on \( \mathbb{R}^{d_x} \times \hat{\theta} \). We define the output from the \( \ell \)-th layer of the approximated network \( f^* \) as \( F^\ell(x,v) \) for \( v \in \hat{\theta} \) and \( x \in \mathbb{R}^{d_x} \). 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 \( f^\ell \). That is, \( F^\ell(x,v) = (f^\ell \circ \cdots \circ F^\ell(x))(v) \) for \( v \in \hat{\theta} \) and \( x \in \mathbb{R}^{d_x} \), and \( F^\ell(x,v) = f^\ell[F^\ell-1(x,v)] \).

**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 \( \hat{\theta}_{L+1} = \{1\} \) 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
\[
\hat{f}_L^1[F_{L-1}](x,1) = \sum_{j=1}^{m_L} \frac{1}{\sqrt{m_L}} \eta \left( \frac{1}{\sqrt{m_L}} w_j^L F_{L-1}(x,v_j^L) \right) + b^L \quad (S-1)
\]
by \( \beta^L \in \mathbb{R}^{m_L} \) and \( w^L \in \mathbb{R}^{m_L} \) satisfying \( \|\beta^L\|_2 \leq \frac{1}{m_L} c_1 R^2 \) and \( \|w^L\|_2 \leq m_L c_6 \). Here, define that \( W_{i,j}^L = \sqrt{m_L} \beta^L \) and \( b^L = (b^L_0(1)) \).

Note that the model (S-1) can be rewritten as
\[
\hat{f}_L^1[F_{L-1}](x,1) = \sum_{j=1}^{m_L} W_{i,j}^L \eta(\sqrt{m_L}^{-1} w_j^L F_{L-1}(x,v_j^L)) + b^L_1.
\]
Finally, let then, by Assumption 1 and Proposition 1, the norms of these quantities can be bounded as
\[ \|W^{(L)}\|_F = \|W^{(L)}_{1}\|_2 \leq \sqrt{c_1} R_t, \quad \|b^{(L)}\|_2 = |b_L| \leq R_b. \]  

(S-2)

By the Cauchy-Schwartz inequality and the Lipschitz continuity of \( \eta \), we have that
\[
\left| \hat{f}^* [F_{L-1}] (x, L) - \hat{f}^* [F'_{L-1}] (x, L) \right| \\
\leq \left| \sum_{j=1}^{m_L} W_{1,j} (\eta(\sqrt{m_L} - 1)w_j^{(L)} F_{L-1}(x, v_j^{(L)}) - \eta(\sqrt{m_L} - 1)w_j^{(L)} F'_{L-1}(x, v_j^{(L)}))) \right| \\
\leq \|W_{1,1}\|_2 \sqrt{m_L} \| (w_j^{(L)} F_{L-1}(x, v_j^{(L)}) - F'_{L-1}(x, v_j^{(L)})) \|_{\infty} \\
\leq \|W_{1,1}\|_2 \sqrt{m_L} \| (F_{L-1}(x, v_j^{(L)}) - F'_{L-1}(x, v_j^{(L)})) \|_{\infty} \\
= \sqrt{c_1 R^2} \| (F_{L-1}(x, v_j^{(L)}) - F'_{L-1}(x, v_j^{(L)})) \|_{\infty},
\]
for \( F_{L-1}, F'_{L-1} : \tilde{T} \times \mathbb{R}^{d_x} \to \mathbb{R} \). Moreover, Proposition 1 ensures that \( \beta^{(L)} \) and \( w^{(L)} \) can be taken so that
\[
\|\hat{f}^* [F_{L-1}] (\cdot, L) - \hat{f}^* [F'_{L-1}] (\cdot, L)\|_{L_2(P(X))}^2 \leq c_0 \lambda_L R^2.
\]

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 \( \ell \)-th internal layer, we consider the following approximation:
\[
\hat{f}^* [F_{\ell-1}] (x, v^{(\ell+1)}) = \sum_{j=1}^{m_{\ell}} \sqrt{m_{\ell} \beta^{(\ell)}_{ij} \eta (\sqrt{m_{\ell}} - 1) w_{ij}^{(\ell)} (v^{(\ell)})} + b_{ij}^{(\ell)} (v^{(\ell+1)}),
\]
for \( g : \tilde{T} \to \mathbb{R} \) with \( \beta^{(\ell)} \in \mathbb{R}^{m_{\ell+1} \times m_{\ell}} \) and \( w^{(\ell)} \in \mathbb{R}^{m_{\ell}} \) satisfying \( \|\beta^{(\ell)}\|_2^2 \leq \frac{1}{m_{\ell}} c_1 R^2 \) for \( j = 1, \ldots, m_{\ell+1} \) and \( \|w^{(\ell)}\|_2^2 \leq m_{\ell} c_3 \). Then, the Lipschitz continuity of \( \hat{f}^* [F_{\ell-1}] \) can be shown as
\[
\left| \hat{f}^* [F_{\ell-1}] (x, v^{(\ell+1)}) - \hat{f}^* [F'_{\ell-1}] (x, v^{(\ell+1)}) \right| \\
\leq \sum_{j=1}^{m_{\ell}} \sqrt{m_{\ell} \beta^{(\ell)}_{ij} \eta (\sqrt{m_{\ell}} - 1) w_{ij}^{(\ell)} (F_{\ell-1}(x, v_j^{(\ell)}) - F'_{\ell-1}(x, v_j^{(\ell)}))} \\
\leq \|\beta^{(\ell)}\|_2 \|w^{(\ell)}\|_2 \| (F_{\ell-1}(x, v_j^{(\ell)}) - F'_{\ell-1}(x, v_j^{(\ell)})) \|_{\infty} \\
\leq \sqrt{c_1 R^2 m_{\ell}} \| (F_{\ell-1}(x, v_j^{(\ell)}) - F'_{\ell-1}(x, v_j^{(\ell)})) \|_{\infty} \\
= \sqrt{c_1 R^2} \| (F_{\ell-1}(x, v_j^{(\ell)}) - F'_{\ell-1}(x, v_j^{(\ell)})) \|_{\infty},
\]
for any \( v_j^{(\ell+1)} \in \tilde{T} \). Proposition 1 asserts that there exist \( \beta^{(\ell)} \) and \( w^{(\ell)} \) that give an upper bound of the approximation error of the \( \ell \)-th layer as
\[
\max_{j=1, \ldots, m_{\ell}} \|\hat{f}^* [F_{\ell-1}] (\cdot, v_j^{(\ell+1)}) - \hat{f}^* [F'_{\ell-1}] (\cdot, v_j^{(\ell+1)})\|_{L_2(P(X))}^2 \leq c_0 \lambda_{\ell} R^2.
\]

Finally, let
\[
W_{ij}^{(\ell)} = \sqrt{\frac{m_{\ell}}{m_{\ell+1}}} (w_{ij}^{(\ell+1)} \beta^{(\ell+1)} \eta (\sqrt{m_{\ell}} - 1) w_i^{(\ell+1)})^{\top}, \quad b^{(\ell)} = \frac{1}{\sqrt{m_{\ell+1}}} \left( (w_1^{(\ell+1)} b_1^{(\ell+1)} (v_1^{(\ell+1)}), \ldots, w_{m_{\ell+1}}^{(\ell+1)} b_{m_{\ell+1}}^{(\ell+1)} (v_{m_{\ell+1}}^{(\ell+1)}))^{\top} \right),
\]
then, by Assumption 1 and Proposition 1, the norms of these quantities can be bounded as
\[
\|W^{(\ell)}\|_F = \\frac{m_{\ell}}{m_{\ell+1}} \sum_{i=1}^{m_{\ell+1}} \sum_{j=1}^{m_{\ell}} \beta^{(\ell)}_{ij} w_i^{(\ell+1)} \|w^{(\ell+1)}\|_2^2 \\
\leq \frac{m_{\ell}}{m_{\ell+1}} \sum_{i=1}^{m_{\ell+1}} \sum_{j=1}^{m_{\ell}} w_i^{(\ell+1)} \|w^{(\ell+1)}\|_2^2 \leq c_1 R^2.
\]
Finally, we combine the results we have obtained above. Note that \( \sum_{j=1}^{m_{\ell+1}} w_j^{(\ell+1)} R_b \le c_3 R_b^2 \).

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

For the first layer, let

\[
\hat{f}^*(x, v^{(2)}_i) = \sum_{j=1}^{d_s} h_1^o(v^{(2)}_i, j) Q_1(j) x_j + b_1^o(v^{(2)}_i)
\]

for \( v^{(2)}_i \in \tilde{F}_2 \). By the definition of \( f^o \), it holds that

\[
\hat{f}^*(x, v^{(2)}_i) = f^o(x, v^{(2)}_i).
\]

Let \( W^{(1)} = \frac{1}{\sqrt{m_2}} (Q_1(j) w^{(2)}_i h_1^o(v^{(2)}_i, j))_{i, j} \in \mathbb{R}^{m_2 \times d_s} \) and \( b^{(1)} = \frac{1}{\sqrt{m_2}} w^{(2)}_1 b_1^o(1), \ldots, w^{(2)}_m b_1^o(m_2) \) \( \top \in \mathbb{R}^{m_2} \). Then, by Assumption 1 and Proposition 1, it holds that

\[
||W^{(1)}||_F^2 = \sum_{i=1}^{m_2} \sum_{j=1}^{d_s} \frac{1}{m_2} w^{(2)}_i h_1^o(v^{(2)}_i, j)^2 Q_1(j)^2 \le \left( \sum_{i=1}^{m_2} \frac{1}{m_2} w^{(2)}_i \right)^2 \max_{1 \le i \le m_2} \left( \sum_{j=1}^{d_s} h_1^o(v^{(2)}_i, j)^2 Q_1(j)^2 \right) \le c_3 \max_{1 \le i \le m_2} \left( \sum_{j=1}^{d_s} h_1^o(v^{(2)}_i, j)^2 Q_1(j)^2 \right) \le c_3 R^2,
\]

and

\[
||b^{(1)}||_2^2 \le \frac{1}{m_1} \sum_{i=1}^{m_2} w^{(2)}_i R_b \le c_3 R_b^2.
\]

**Step 4.**

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

\[
\|f^o_L \circ f^o_{L-1} \circ \cdots \circ f^o_1 - \hat{f}^*_L \circ \hat{f}^*_{L-1} \circ \cdots \circ \hat{f}^*_1\|_{L_2(P(X))} = \|f^o_L \circ f^o_{L-1} \circ \cdots \circ f^o_1 - \hat{f}^*_L \circ \hat{f}^*_{L-1} \circ \cdots \circ \hat{f}^*_1\|_{L_2(P(X))} \le \sum_{\ell=1}^{L} \|\hat{f}^*_\ell \circ \cdots \circ \hat{f}^*_{\ell+1} \circ f^o_{\ell+1} \circ \cdots \circ f^o_1 - \hat{f}^*_L \circ \cdots \circ \hat{f}^*_1\|_{L_2(P(X))}.
\]

Then combining the argument given above, we have

\[
\|\hat{f}^*_L \circ \cdots \circ \hat{f}^*_{\ell+1} \circ f^o_{\ell+1} \circ \cdots \circ f^o_1 - \hat{f}^*_L \circ \cdots \circ \hat{f}^*_1\|_{L_2(P(X))} \le (\sqrt{c_1 c_3} R)^{L-\ell} (\sqrt{c_0 c_2} R) = (c_1 c_3)^{L-\ell} c_0 R^{L-\ell+1} \sqrt{\lambda_\ell},
\]
for \( \ell = 2, \ldots, L \). And the right hand side is 0 for \( \ell = 1 \). This yields that
\[
\| f^0 - \hat{f}^* \|_{L_2(P(X))} \leq \sum_{\ell=2}^{L} R^{L-\ell+1} \sqrt{(c_1 c_3) L^{-\ell} c_0} \sqrt{\lambda_\ell}.
\]
By substituting \( W(\ell) \) and \( b^{(\ell)} \) for \( \ell = 1, \ldots, L \) defined above into the definition of \( f^* \), then it is easy to see that
\[
f^* = \hat{f}^*
\]
as a function. Then, we obtain the assertion.

A.2 Bounding the \( L_\infty \)-norm

The next lemma shows the \( L_\infty \)-norm of the true function \( f^0 \) and that of \( f \in \mathcal{F} \).

Lemma 3. Under Assumptions 1, 2 and 3, the \( L_\infty \)-norms of \( f^0 \) and that of \( f \in \mathcal{F} \) are bounded as
\[
\| f^0 \|_\infty \leq R L D_x + \sum_{\ell=1}^{L} R^{L-\ell} R_b,
\]
\[
\| f \|_\infty \leq (\sqrt{c_1 c_3}) L^L D_x + \sum_{\ell=1}^{L} (\sqrt{c_1 c_3}) R^{L-\ell} R_b.
\]

Proof. Suppose that
\[
\| F_{\ell-1}^0 (x, \cdot) \|_{L_2(Q_\ell)} \leq G.
\]
Then, \( F_\ell^0 \) can be bounded inductively: for all \( \tau \in \mathcal{T}_{\ell+1} \)
\[
| F_\ell^0 (x, \tau) | = \left| \int_{\mathcal{T}_\ell} h_\ell^0 (\tau, w) \eta (F_{\ell-1}^0 (x, w)) dQ_\ell (w) + b_\ell^0 (\tau) \right|
\leq \| h_\ell^0 (\tau, \cdot) \|_{L_2(Q_\ell)} \| F_{\ell-1}^0 (x, \cdot) \|_{L_2(Q_\ell)} + | b_\ell^0 (\tau) |
\leq R G + R_b,
\]
by Assumption 1. Similarly, as for \( \ell = 1 \), it holds that, for all \( \tau \in \mathcal{T}_2 \) and \( x \in \mathbb{R}^{d_x} \),
\[
| f_1^0 (x, \tau) | = \left| \sum_{i=1}^{d_x} h_1^0 (\tau, i) x_i Q_1 (i) + b_1^0 (\tau) \right|
\leq \left| \sum_{i=1}^{d_x} h_1^0 (\tau, i) x_i Q_1 (i) \right| + | b_1^0 (\tau) |
\leq \| h_1^0 (\tau, \cdot) \|_{L_2(Q_1)} \| x \|_{L_2(Q_1)} + R_b
\leq R D_x + R_b.
\]
Applying the same argument recursively, we have
\[
\| f^0 \|_\infty \leq R L D_x + \sum_{\ell=1}^{L} R^{L-\ell} R_b.
\]
We can bound the \( L_\infty \)-norm of any \( f \in \mathcal{F} \) through a similar argument. Note that \( W(\ell) \) satisfies \( \| W(\ell) \|_F \leq \sqrt{c_1 c_3} R \) for \( \ell = 1, \ldots, L - 1 \), \( W(L) \) satisfies \( \| W(L) \|_F \leq \sqrt{c_3} R \), and \( b^{(\ell)} \) satisfies \( \| b^{(\ell)} \|_2 \leq \sqrt{c_3} R_b \) by its construc-
tion. Therefore, though a similar argument to the bound for \( f^o \), we have that
\[
\|f\| \leq \sqrt{c_1} R \left[ \prod_{\ell=2}^{L-1} (\sqrt{c_1} R) \right] \sqrt{c_3} RD_x
\]
\[
+ \left( \sum_{\ell=1}^{L-2} \sqrt{c_1} R \left[ \prod_{\ell'=\ell+1}^{L-1} (\sqrt{c_1} R) \right] \sqrt{c_3} R_b + \sqrt{c_1} R \sqrt{c_3} R_b + \sqrt{c_3} R_b \right)
\]
\[
\leq (c_1 c_3)^{L/2} R^L D_x + \sum_{\ell=1}^{L} (\sqrt{c_1} c_3 R)^{L-\ell} R_b. 
\]

\[\square\]

### 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 \( \|f\|_{L_2(P_X)} = \sqrt{\int f(X)^2 dP(X)} \)).

Here, let
\[
\epsilon_n = \delta_1 + \sigma \delta_2, \quad \bar{\epsilon}_n = \delta_1 + \tilde{\delta}_2.
\]

#### B.1 In-sample error

Here we show the in-sample error bound. Let \( X_n = (x_1, \ldots, x_n), Y_n = (y_1, \ldots, y_n) \) and \( D_n = (X_n, Y_n) \). For given \( X_n \), the probability distribution of \( Y_n \) is distributed from the true distribution. 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 \( \tilde{r} \geq 1 \), let \( A_{\tilde{r}} \) be the event such that
\[
\int_{p_n,f} \Pi(df) \geq \exp(-n \epsilon_n^2 \bar{\epsilon}_n^2 / \sigma^2) \Pi(f : \|f - f^o\|_\infty \leq \tilde{\delta}_2 \tilde{r}).
\]

The probability of this event is bounded by Lemma 4.

Using a test function \( \phi_n \) 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
\[
E \left[ \Pi(\|f - f^o\|_n \geq \sqrt{2} \epsilon_n r | D_n) \right] 
\leq E \left[ \phi_n \right] + P(A_{\tilde{r}})
\leq E \left[ (1 - \phi_n) I_{A_{\tilde{r}}} \Pi(f \in \mathcal{F} | D_n) \right]
+ E \left[ (1 - \phi_n) I_{A_{\tilde{r}}} \Pi(f \in \mathcal{F} : \|f - f^o\|_n^2 \geq 2 \epsilon n \bar{\epsilon}_n^2 | D_n) \right]
\leq A_n + B_n + C_n + D_n, \tag{S-3}
\]
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 \( \Theta_{r'} = \{ f \in \mathcal{F} \mid r' \leq \sqrt{n} \|f - f^o\|_n / \sigma \} \). We construct a maximum cardinality set \( \Theta_{r'} \subset C_{r'} \) such that each \( f, f' \in \Theta_{r'} \) satisfies \( \sqrt{n} \|f - f'\|_n / \sigma \geq r' / 2 \). Here we denote by \( D(\epsilon, \mathcal{F}, \| \cdot \|) \) the \( \epsilon \)-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 $\phi_{r'}$ such that

$$P_n, f' \overline{\phi}_{r'} \leq 9D(r'/2, C_{r'}, \sqrt{n} \cdot \|n/\sigma\)e^{-\frac{1}{2}r'^2} \leq 9D(r'/2, \mathcal{F}, \sqrt{n} \cdot \|n/\sigma\)e^{-\frac{1}{2}r'^2},$$

for any $r' > 0$.

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

$$P_n, f' \phi_n \leq 9e^{-\frac{1}{2}n r'^2 + \log(D(r'/2, \mathcal{F}, \sqrt{n} \cdot \|n/\sigma\))} \leq 9e^{-\frac{1}{4}r'^2}.$$ (S-4)

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

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

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

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

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

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

$$\log(D(r'/2, \mathcal{F}, \sqrt{n} \cdot \|n/\sigma\)) \leq \log(D(\sigma r'/2\sqrt{n}, \mathcal{F}, \|\cdot\|_\infty) \leq \log(D(\sigma r'/2\sqrt{n}, \mathcal{F}, \|\cdot\|_\infty)) \leq \log(D(r'/2, \mathcal{F}, \sqrt{n} \cdot \|n/\sigma\)) \leq \log(D(\sigma r'/2\sqrt{n}, \mathcal{F}, \|\cdot\|_\infty))$$

$$\leq \sum_{\ell=1}^L \log(N(\sigma r'/2\sqrt{n} \hat{G}) \cdot B_{m_{\ell+1} \times m_{\ell}}(\hat{R}, \|\cdot\|)) + \sum_{\ell=1}^L \log(N(\sigma r'/2\sqrt{n} \hat{G}) \cdot B_{m_{\ell}}(\hat{R}_b, \|\cdot\|))$$

$$\leq \sum_{\ell=1}^L \frac{m_{\ell+1} m_{\ell}}{4 + \frac{\sigma r'}{2\sqrt{n} \hat{G}R}} + \sum_{\ell=1}^L \frac{m_{\ell}}{4 + \frac{\sigma r'}{2\sqrt{n} \hat{G} R_b}}$$

$$= \sum_{\ell=1}^L m_{\ell+1} m_{\ell} \log \left(1 + 4\sqrt{2\hat{G} \hat{R}}/\epsilon n r'\right) + \sum_{\ell=1}^L m_{\ell} \log \left(1 + 4\sqrt{2\hat{G} \hat{R}_b}/\epsilon n r'\right).$$ (S-7)

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

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

The $\epsilon$-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 $\epsilon$ from the finite set with respect to the metric $d$. We denote by $N(\epsilon, T, d)$ the $\epsilon$-internal covering number of $(T, d)$. 

---

1 The $\epsilon$-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 $\epsilon$ from the finite set with respect to the metric $d$. We denote by $N(\epsilon, T, d)$ the $\epsilon$-internal covering number of $(T, d)$. 

---

Manuscript under review by AISTATS 2018
Step 2. Here, we evaluate $B_n$. It can be evaluated by Lemma 4 as

$$B_n \leq \exp(-n\tilde{c}_n r^2/(8\sigma^2)) + \exp(-n\tilde{d}_{1,n}(r^2 - 1)/(11R_\infty^2)).$$

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

$$D_n = E_{X_n} \left[ P_{n,f^o} \Pi(f \in \mathcal{F} : \|f - f^o\|_n > \sqrt{2}er|Y_n)(1 - \phi_n)1_{A_r} \right].$$

Define

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

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

$$D_n = E_{X_n} \left\{ P_{n,f^o} \left[ \frac{\int_{\mathcal{F}} 1\{f : \|f - f^o\|_n > \sqrt{2}er\} p_{n,f} d\Pi(f)}{\int_{\mathcal{F}} p_{n,f} d\Pi(f)} (1 - \phi_n)1_{A_r} \right] \right\}$$

$$= E_{X_n} \left\{ P_{n,f^o} \left[ \frac{\int_{f \in \mathcal{F} : \|f - f^o\|_n > \sqrt{2}er} p_{n,f}/p_{n,f^o} d\Pi(f) \exp(n\tilde{c}_n r^2/\sigma^2 + \Xi_n(\tilde{r}))(1 - \phi_n)1_{A_r} \right] \right\}$$

$$\leq E_{X_n} \left\{ \int_{f \in \mathcal{F} : \|f - f^o\|_n > \sqrt{2}er} P_{n,f}[(1 - \phi_n)1_{A_r}] \exp(n\tilde{c}_n r^2/\sigma^2 + \Xi_n(\tilde{r})) d\Pi(f) \right\}$$

$$\leq \exp \left( \frac{n\tilde{c}_n^2 r^2}{\sigma^2} + \Xi_n(\tilde{r}) - \frac{n\tilde{c}_n^2 r^2}{4\sigma^2} \right).$$

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

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

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

$$\leq -\sum_{\ell=1}^{L} \log(\Pi(W^{(\ell)} : \|W^{(\ell)} - W^{*(\ell)}\|_F \leq \tilde{\delta}_{2,n}/\hat{G}))$$

$$-\sum_{\ell=1}^{L} \log(\Pi(b^{(\ell)} : \|b^{(\ell)} - b^{*(\ell)}\|_2 \leq \tilde{\delta}_{2,n}/\hat{G}))$$

$$\leq \sum_{\ell=1}^{L} m_\ell m_{\ell+1} \log(\hat{R}\hat{G}/(\tilde{\delta}_{2,n}/2)) + \sum_{\ell=1}^{L} m_\ell \log(\hat{R}\hat{G}/(\tilde{\delta}_{2,n}/2)).$$

(S-8)

Step 5. Finally, we combine the results obtained above.
\[ \mathbb{E} \left[ \Pi(\|f - f^o\|_n \geq \sqrt{2}\epsilon_n r | Y_n) \right] \leq 9 \exp \left[ -\frac{1}{4\sigma^2} n\epsilon_n^2 r^2 + \sum_{\ell=1}^{L} m_\ell \epsilon_n + \sum_{\ell=1}^{L} m_\ell \log \left( 1 + 4\sqrt{2}\tilde{G}\tilde{R} \right) + \sum_{\ell=1}^{L} m_\ell \log \left( 1 + 4\sqrt{2}\tilde{G}\tilde{R}_b \right) \right] + \exp(-n\delta_{1,n}^2 r^2/(8\sigma^2)) + \exp(-n\delta_{1,n}^2 (\tilde{r}^2 - 1)^2/(11\tilde{R}_\infty^2)) + \exp \left( \frac{n}{\sigma^2} \epsilon_n r^2 + \Xi_n(\tilde{r}) - \frac{n\delta_{1,n}^2 r^2}{4\sigma^2} \right). \]  
(S-9)

Now, let \( 1 \leq \tilde{r} \leq r \). Then, since \( \epsilon_n \geq \delta_{2,n} \) and \( r \geq 1 \), we have that

\[ \max \left\{ \log \left( \frac{2\tilde{G}\tilde{R}}{\delta_{2,n}} \right), \log \left( 1 + 4\sqrt{2}\tilde{G}R \right) \right\} \leq \log \left( 1 + \frac{4\sqrt{2}\tilde{G}R}{\delta_{2,n}} \right), \]

for all \( R' > 0 \). Now, we set \( \delta_{2,n} \) to satisfy

\[ \frac{n\delta_{2,n}^2}{\sigma^2} \geq \sum_{\ell=1}^{L} m_\ell m_{\ell+1} \log \left( 1 + \frac{4\sqrt{2}\tilde{G}\tilde{R}}{\delta_{2,n}} \right) + \sum_{\ell=1}^{L} m_\ell \log \left( 1 + \frac{4\sqrt{2}\tilde{G}\tilde{R}_b}{\delta_{2,n}} \right) \geq \Xi_n(\tilde{r}), \]

which can be satisfied by

\[ \delta_{2,n}^2 = \frac{2\sigma^2}{n} \sum_{\ell=1}^{L} m_\ell m_{\ell+1} \log \left( 1 + \frac{4\sqrt{2}\tilde{G} \max\{\tilde{R}, \tilde{R}_b\} \sqrt{n}}{\sigma^2 \sum_{\ell=1}^{L} m_\ell m_{\ell+1}} \right). \]

Then, by noticing \( n\delta_{2,n}^2 \leq n\epsilon_n^2 \) and Eq. (S-8), the RHS of Eq. (S-9) is upper bounded by

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

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

\[ \exp(-n\epsilon_n^2 r^2/(8\sigma^2)) + \exp(-n\epsilon_n^2 r^2/(8\sigma^2)) + 10 \exp(-n\epsilon_n^2 r^2/(8\sigma^2)) \]

\[ \leq \exp(-n\epsilon_n^2 r^2/(8\sigma^2)) + 11 \exp(-n\epsilon_n^2 r^2/(8\sigma^2)). \]

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

\[ P_{D,n} \left( \int \frac{p_{n,f}(Y_n)}{p_{n,f^o}(Y_n)} \Pi(df) \geq \exp(-n\epsilon_n^2 r^2/(8\sigma^2)) \Pi(f : \|f - f^o\|_\infty \leq \delta_{2,n} \tilde{r}) \right) \]

\[ \geq 1 - \exp(-n\epsilon_n^2 r^2/(8\sigma^2)) - \exp(-n\epsilon_n^2 \min\{\tilde{r}^2 - 1, \tilde{r}^2 - 1\}/(11\tilde{R}_\infty^2)). \]

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

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

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 \( \|f - f^o\|_n \leq \delta_{1,n} \tilde{r} + \|f - f^o\|_\infty \) with high probability. Note that

\[ \|f - f^o\|_n \leq \|f - f^o\|_\infty + \|f^o - f^o\|_\infty \leq \|f - f^o\|_\infty + \|f^o - f^o\|_n. \]

Hence, we just need to show \( \|f^o - f^o\|_n^2 \leq \tilde{r}^2_1 + \tilde{r}^2 \tilde{r}^2 \leq 1 \) \((1 + \tilde{r}^2)\delta_{1,n}^2 \) with high probability for appropriately chosen \( \tilde{r}^2 \). This can be shown by Bernstein’s inequality:

\[ P \left( \|f^o - f^o\|_2^2 \leq \delta_{1,n}^2 \right) \leq \exp \left( -\frac{n\epsilon_n^2 \delta_{1,n}^2}{2(v + \tilde{r}^2 \|f^o - f^o\|_\infty \delta_{1,n}^2/3) \delta_{1,n}^2} \right). \]
where \( v = \mathbb{E}_X[(f^*(X) - f^0(X))^2 - \|f^* - f^0\|_{L_2(P_X)}^2] \). Now \( v \leq \mathbb{E}_X[(f^*(X) - f^0(X))^4] \leq \|f^* - f^0\|_\infty^2 \|f^* - f^0\|_{L_2(P_X)}^2 \). This yields that

\[
P \left( \|f^* - f^0\|_{L_2(P_X)}^2 + \tilde{r}^2 \tilde{\delta}^2_{1,n} \leq \|f^* - f^0\|_n^2 \right) \leq \exp \left( -3n \min \{ \tilde{r}^2, \tilde{r}' \} \frac{\tilde{\delta}^2_{1,n}}{8 \|f^* - f^0\|_\infty^2} \right). \tag{S-11}
\]

Since \( \|f^* - f^0\|_\infty \leq 2 \hat{R}_\infty \), the RHS is further bounded by \( \exp \left( -\frac{3n \min \{ \tilde{r}^2, \tilde{r}' \} \tilde{\delta}^2_{1,n}}{32 \hat{R}_\infty^2} \right) \).

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

\[
\|f - f^0\|_n \leq \|f - f^*\|_\infty + \sqrt{\|f^* - f^0\|_{L_2(P_X)}^2 + \tilde{r}^2 \tilde{\delta}^2_{1,n}} \leq \|f - f^*\|_\infty + \sqrt{1 + \tilde{r}^2} \tilde{\delta}_{1,n}
\]

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. \( \square \)

### 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:

\[
E_{D_n} \left[ \Pi(f : \|f - f^0\|_{L_2(P_X)} \geq \epsilon n r | D_n) \right], \tag{S-12}
\]

for sufficiently large \( r \geq 1 \).

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

\[
I = E_{D_n} \left[ 1_{A_1} \right],
II = E_{D_n} \left[ 1_{A_1} \Pi(f : \sqrt{2}\|f - f^0\|_n > \epsilon n r, \|f\|_\infty \leq \hat{R}_\infty | D_n) \right],
III = E_{D_n} \left[ 1_{A_1} \Pi(f : \|f - f^0\|_{L_2(P_X)} > \epsilon n r \geq \sqrt{2}\|f - f^0\|_n, \|f\|_\infty \leq \hat{R}_\infty | D_n) \right],
IV = E_{D_n} \left[ 1_{A_1} \Pi(f : \|f\|_\infty > \hat{R}_\infty | D_n) \right].
\]

The term I and II are already evaluated in Section B.1, that is, \( I + II \) 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 \( \|f - f^*\|_n \) and the expected norm \( \|f - f^0\|_{L_2(P_X)} \), which can be done by Bernstein’s inequality. Following the same argument to derive Eq. (S-11), it holds that

\[
P \left( \|f - f^0\|_{L_2(P_X)} \geq \sqrt{2}\|f - f^0\|_n \right) \leq \exp \left( -\frac{n\|f - f^0\|_{L_2(P_X)}^2}{11 \hat{R}_\infty^2} \right).
\]

Therefore, we arrive at the following bound of III:

\[
III \leq E_{X_n} \left[ \mathbb{P}_{n,f} \left[ \int_{f \in \mathcal{F} : \|f - f^0\|_{L_2(P_X)} > \epsilon n r \geq \sqrt{2}\|f - f^0\|_n} \frac{p_{n,f} / p_{n,f^0}}{d\Pi(f) \exp(n \epsilon^2 \tilde{r}^2 / \sigma^2 + \Xi_n(\tilde{r})) 1_{A_1}} \right] \right]
\leq \exp(n \epsilon^2 \tilde{r}^2 / \sigma^2 + \Xi_n(\tilde{r})) \int_{f \in \mathcal{F} : \|f - f^0\|_{L_2(P_X)} > \epsilon n r} P(\|f - f^0\|_{L_2(P_X)} \geq \sqrt{2}\|f - f^0\|_n) d\Pi(f)
\leq \exp \left( \frac{n \epsilon^2 \tilde{r}^2}{\sigma^2} + \Xi_n(\tilde{r}) - \frac{n \epsilon^2 \tilde{r}^2}{11 \hat{R}_\infty^2} \right)
\leq \exp \left( \frac{2n \epsilon^2 \tilde{r}^2}{\sigma^2} - \frac{n \epsilon^2 \tilde{r}^2}{11 \hat{R}_\infty^2} \right).
\]
Finally, since all \( f \in \mathcal{F} \) satisfies \( \|f\|_{\infty} \leq \tilde{R}_\infty \), \( IV = 0 \).

Combining the results we arrive at
\[
E_{D_n}\left[ \Pi(f : \|f - f^0\|_{L_2(p_X)} \geq \epsilon_n r | D_n) \right] \leq \exp \left[ -\frac{n \delta^2_n \min\{(\hat{r}^2 - 1)^2, \hat{r}^2 - 1\}}{11 \tilde{R}^2_\infty} \right] + 12 \exp \left( -n \epsilon_n^2 \hat{r}^2/(8\sigma^2) \right),
\]
for all \( \hat{r} \geq 1 \) and \( r \geq \max\{12, 33\tilde{R}^2_\infty/\sigma^2\}\hat{r}^2 \). This concludes the proof of Theorem 2.

\section{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 \((\xi_n)_{n=1}^\infty\) be i.i.d. Gaussian sequence with mean 0 and variance \( \sigma^2 \), and \((x_i)_{i=1}^n \subset \mathcal{X}\) be a given set of input variables. Then, for a set \( \mathcal{F} \) of functions from \( \mathcal{X} \) to \( \mathbb{R} \) which is separable with respect to \( L_\infty \)-norm and \( \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \frac{1}{n} \bar{\xi}_i f(x_i) \right\} < \infty \) almost surely, it holds that for every \( r > 0 \),
\[
P \left( \sup_{f \in \mathcal{F}} \left\| \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\| \geq \mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \left\| \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\| + r \right\} \right) \leq \exp(-nr^2/(2\sigma^2)||\tilde{F}||^2)
\]
where \( ||\tilde{F}||^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(x_i)^2 \). Here the probability is taken with respect to \((\xi_n)_{n=1}^\infty\).

Remind that every \( f \in \mathcal{F} \) satisfies \( ||f||_{\infty} \leq ||f||_{\infty} \leq \tilde{R}_\infty \). Hence \( ||f||_{\infty} \leq \tilde{R}_\infty \). For an observation \((x_i)_{i=1}^n\), let \( \mathcal{G}_\delta = \{f - f^* | ||f - f^*||_{\infty} \leq \delta, f \in \mathcal{F}\} \). It is obvious that \( \mathcal{G}_\delta \) is separable with respect to \( L_\infty \)-norm. Then, by the Gaussian concentration inequality, we have that
\[
P \left( \sup_{f \in \mathcal{G}_\delta} \left\| \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\| \geq \mathbb{E} \left\{ \sup_{f \in \mathcal{G}_\delta} \left\| \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\| + r \right\} \right) \leq \exp(-nr^2/(2\sigma^2))
\]
for every \( r > 0 \). By applying this inequality for \( \delta_j = 2^{j-1} \sigma/\sqrt{n} \) for \( j = 1, \ldots, \lfloor \log_2(\tilde{R}_\infty \sqrt{n}/\sigma) \rfloor \) and using the uniform bound, we can show that, for every \( r > 0 \), with probability \( \exp(-nr^2/2\sigma^2)\), it holds that
\[
\left\{ \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) - f^*(x_i) \right\} \geq \mathbb{E} \left\{ \sup_{f \in \mathcal{G}_\delta} \left\{ \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\} \right\} + 2\delta r
\]
uniformly for all \( f \in \mathcal{G}_\delta \) where \( \delta \) is any positive real satisfying \( \delta \geq \sigma/\sqrt{n} \).

**Lemma 5.** There exists a universal constant \( C \) such that for any \( \delta \) it holds that
\[
\mathbb{E} \left\{ \sup_{f \in \mathcal{G}_\delta} \left\| \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\| \right\} \leq C \delta \sqrt{\frac{\sum_{\ell=1}^L m_{\ell} m_{\ell+1}}{n} \log \left( \frac{4G \max\{\tilde{R}, \tilde{R}_b\}}{\delta} \right)}.
\]

**Proof.** Since \( f \mapsto \frac{1}{n} \sum_{i=1}^n \xi_i f(x_i) \) 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
\[
\mathbb{E} \left\{ \sup_{f \in \mathcal{G}_\delta} \left\{ \sum_{i=1}^n \frac{1}{n} \xi_i f(x_i) \right\} \right\} \leq 4\sqrt{2} \frac{\sigma}{\sqrt{n}} \int_0^{2\delta} \sqrt{\log(2N(\epsilon, \mathcal{G}_\delta, \| \cdot \|_n))} d\epsilon.
\]
Since \( \log N(\epsilon, \mathcal{G}_\delta, \| \cdot \|_n) \leq \log N(\epsilon, \mathcal{F}, \| \cdot \|_\infty) \leq 2 \sum_{\ell=1}^L m_{\ell} m_{\ell+1} \log \left( 1 + \frac{4G \max\{\tilde{R}, \tilde{R}_b\}}{\epsilon} \right) \), the right hand side is bounded by
\[
\int_0^{2\delta} \sqrt{\log(2N(\epsilon, \mathcal{F}, \| \cdot \|_n))} d\epsilon \leq \int_0^{2\delta} \sqrt{\log(2) + 2 \sum_{\ell=1}^L m_{\ell} m_{\ell+1} \log \left( 1 + \frac{4G \max\{\tilde{R}, \tilde{R}_b\}}{\epsilon} \right)} d\epsilon
\]
\[
\leq C \delta \sqrt{\frac{\sum_{\ell=1}^L m_{\ell} m_{\ell+1}}{n} \log \left( 1 + \frac{4G \max\{\tilde{R}, \tilde{R}_b\}}{\delta} \right)}.
\]
Therefore, by substituting $\delta \leftarrow \left( \| f - f^* \|_n \vee \sigma \sqrt{\frac{\sum_{\ell=1}^{L} m_{\ell+1}}{n}} \right)$ and $r \leftarrow \sigma r / \sqrt{n}$, the following inequality holds:

$$- \frac{1}{n} \sum_{i=1}^{n} \xi_i (f(x_i) - f^*(x_i))$$

$$\leq C\sigma \left( \| f - f^* \|_n \vee \sigma \sqrt{\frac{\sum_{\ell=1}^{L} m_{\ell+1}}{n}} \right) \sqrt{\frac{\sum_{\ell=1}^{L} m_{\ell+1}}{n} \log_+ \left( 1 + \frac{4\sqrt{n} \max \{ \hat{R}, \tilde{R}_b \} \sqrt{m} }{\sigma \sqrt{\sum_{\ell=1}^{L} m_{\ell+1}} } \right)}$$

$$+ 2 \left( \| f - f^* \|_n \vee \sigma \sqrt{\frac{\sum_{\ell=1}^{L} m_{\ell+1}}{n}} \right) \sigma \frac{r}{\sqrt{n}}$$

$$\leq \frac{1}{4} \left( \| f - f^* \|_n \vee \sigma \sqrt{\frac{\sum_{\ell=1}^{L} m_{\ell+1}}{n}} \right)^2$$

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

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

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

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

$$\hat{f} := \argmin_{f \in F} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

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

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \leq \frac{1}{n} \sum_{i=1}^{n} (y_i - f^*(x_i))^2$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} y_i (f^*(x_i) - \hat{f}(x_i)) + \| \hat{f} \|_n^2 - \| f^* \|_n^2 \leq 0$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} (\xi_i + f^o(x_i))(f^*(x_i) - \hat{f}(x_i)) + \| \hat{f} \|_n^2 - \| f^* \|_n^2 \leq 0$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} \xi_i (f^*(x_i) - \hat{f}(x_i)) + \frac{2}{n} \sum_{i=1}^{n} f^o(x_i)(f^*(x_i) - \hat{f}(x_i)) + \| \hat{f} \|_n^2 - \| f^* \|_n^2 \leq 0$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} \xi_i (f^*(x_i) - \hat{f}(x_i)) + \| \hat{f} - f^o \|_n^2 \leq \| f^* - f^o \|_n^2.$$

Therefore, we have

$$- \frac{1}{4} \left( \| \hat{f} - f^* \|_n \vee \sqrt{\frac{\sigma^2 \sum_{\ell=1}^{L} m_{\ell+1}}{n}} \right)^2 - \Psi_{r,n} + \| \hat{f} - f^o \|_n^2 \leq \| f^* - f^o \|_n^2. \quad (S-13)$$
Let us assume $\|\hat{f} - f^*\|^2_n \geq \frac{\sigma^2 \sum_{l=1}^{L} m_l m_{l+1}}{n}$. Then, by Eq. (S-13), we have

$$-\frac{1}{4} \|\hat{f} - f^*\|^2_n - \Psi_{r,n} \leq \|\hat{f} - f^\circ\|^2_n \leq \|f^* - f^\circ\|^2_n$$

$$\Rightarrow -\frac{1}{4} \|\hat{f} - f^*\|^2_n - \Psi_{r,n} + \frac{1}{2} \|\hat{f} - f^*\|^2_n - \|f^* - f^\circ\|^2_n \leq \|f^* - f^\circ\|^2_n$$

$$\Rightarrow \frac{1}{4} \|\hat{f} - f^*\|^2_n \leq 2\|f^* - f^\circ\|^2_n + \Psi_{r,n}.$$ (S-14)

Otherwise, we trivially have $\|\hat{f} - f^*\|^2_n \leq \frac{\sigma^2 \sum_{l=1}^{L} m_l m_{l+1}}{n}$.

Combining the inequalities, it holds that

$$\|\hat{f} - f^*\|^2_n \leq 8\|f^* - f^\circ\|^2_n + 4\Psi_{r,n} + \frac{\sigma^2 \sum_{l=1}^{L} m_l m_{l+1}}{n}.$$ (S-15)

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

**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 $X$. Then, for a set $F$ of functions from $X$ to $\mathbb{R}$ which is separable with respect to $L_\infty$-norm and $\|f\|_\infty \leq \bar{R}$ for all $f \in F$, it holds that for every $r > 0$,

$$P \left( \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 - E[f]^2 \right) \geq C \left\{ E \left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 - E[f]^2 \right] \right\} \geq \frac{\sqrt{\bar{R}}}{n} \right) \leq \exp(-r)$$

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

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

$$\sup_{f \in G^\delta} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 - E[f]^2 \right\} \geq C_1 \left\{ E \left[ \sup_{f \in G^\delta} \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 - E[f]^2 \right] \right\} \geq \frac{\delta \bar{R}^2}{n} \right) \leq \exp(-r)$$

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 $\delta > 0$,

$$E \left[ \sup_{f \in G^\delta} \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 - E[f]^2 \right] \leq C \left[ \delta \bar{R}^2 + \sqrt{\sum_{l=1}^{L} m_l m_{l+1}} \log_+ \left( 1 + \frac{4\bar{G} \max\{\bar{R}, \bar{R}_b\}}{\delta} \right) \right].$$

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

$$E \left[ \sup_{f \in G^\delta} \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 - E[f]^2 \right] \leq 2E \left[ \sup_{f \in G^\delta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)^2 \right] \leq 2E \left[ \sum_{i=1}^{n} \epsilon_i f(x_i)^2 \right] \leq 2E \left[ \sum_{i=1}^{n} \epsilon_i f(x_i)^2 \right]$$
Let $\Phi$ (see, for example, Lemma 2.3.1 in van der Vaart and Wellner (1996)). Since $\|f\|_\infty \leq 2\hat{R}_\infty$ for all $f \in \mathcal{G}_j$, the contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12) gives an upper bound of the RHS as

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

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

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

This concludes the proof. \(\square\)

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

$$
\left| \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2 - \mathbb{E}[(f - f^*)^2] \right| \leq C_1 \left[ (C_2 \delta \hat{R}_\infty \sqrt{\Phi_n}) \vee (\hat{R}_\infty \Phi_n) + \delta \sqrt{\frac{\hat{R}_\infty^2 r}{n}} + \frac{r \hat{R}_\infty^2}{n} \right],
$$

where $\delta$ is any positive real such that $\delta^2 \geq \mathbb{E}[(f - f^*)^2]$ 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 $f = \hat{f}$ to Eq. (S-15) gives that

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

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

$$\| f^* - f^o \|^2_n \leq (1 + \bar{\nu}') \| f^* - f^o \|^2_{L_2(P_X)} \leq (1 + \bar{\nu}') \delta_1^2 \psi_n,$$

with probability $1 - \exp \left( - \frac{3n \sigma^2 + \bar{\nu}'^2 \bar{\nu}^2}{32\hat{R}_\infty^2} \right)$ for every $\bar{\nu}' > 0$.

Combining all inequalities, we obtain that

$$
\| \hat{f} - f^* \|^2_{L_2(P_X)} \leq 2C_2 \hat{R}_\infty^2 \left( \Phi_n + \frac{r}{n} \right) + 16(1 + \bar{\nu}') \delta_1^2 \psi_n + 4\Psi_{r,n} + \frac{2(\sigma^2 + \hat{R}_\infty^2)}{n} \sum_{\ell=1}^{L} m_{\ell} m_{\ell+1}.
$$

This gives a bound for the distance between $\hat{f}$ and $f^*$. However, what we want is a bound on the distance from the true function $f^o$ to $\hat{f}$. This can be accomplished by noticing that $\| \hat{f} - f^o \|^2_{L_2(P_X)} \leq 2\| \hat{f} - f^* \|^2_{L_2(P_X)} + \| f^o -}
\[ f^* \|_{L^2(P_X)}^2 \leq 2 \| \hat{f} - f^* \|_{L^2(P_X)}^2 + 2 \hat{\delta}^2_{1,n}, \]

and conclude that

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

More concisely, letting

\[ \alpha(U) := U^2 \sum_{\ell=1}^L m_\ell m_{\ell+1} \frac{\log \left( 1 + \frac{4\sqrt{\bar{G}}_{\text{max}}(\bar{R}, \bar{R}_b)}{U \sqrt{\sum_{\ell=1}^L m_\ell m_{\ell+1}}} \right)}{n}, \]

the right side is further upper bounded as

\[ \| \hat{f} - f^o \|_{L^2(P_X)}^2 \leq C_3 \left\{ \alpha(\hat{R}_\infty) + \alpha(\sigma) + \frac{(\hat{R}_\infty^2 + \sigma^2)}{n} \left[ \log \left( \frac{\sqrt{n}}{\min\{\sigma/\hat{R}_\infty, 1\}} \right) + r \right] + (1 + \tilde{r}') \hat{\delta}^2_{1,n} \right\} \]

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

References


