Non-parametric Estimation via Neural Network

Suh, Ko, Huo

August 14, 2020

1 Preliminary notations

In this document, we consider a non-parametric regression model, $Y_i = f_0(X_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We are interested in finding the bound of prediction risk

$$R(\widehat{f}_n, f_0) := E_{f_0}\left[\left(\widehat{f}_n(X) - f_0(X)\right)^2\right],$$

with $X \overset{D}{=} X_1$ being independent of sample (X_i, Y_i). Let $\widehat{f}_n$ denote any local or global minimizer of empirical risk $\frac{1}{n}\sum_{i=1}^n \left(Y_i - f(X_i)\right)^2$ over the neural network class $\mathcal{F}(L, p, s, F)$. Note also that the subscript f_0 in E_{f_0} indicates that the expectation is taken with respect to a sample generated from f_0. The sequence $\Delta_n(\widehat{f}_n, f_0)$ measures the difference between the expected empirical risk of $\widehat{f}_n$ and the global minimum over all networks in the class $\mathcal{F}(L, p, s, F)$, which is defined as follows:

$$\Delta_n(\widehat{f}_n, f_0) := E_{f_0}\left[\frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{f}_n(X_i)\right)^2 - \inf_{f\in\mathcal{F}(L,p,s,F)} \frac{1}{n}\sum_{i=1}^n \left(Y_i - f(X_i)\right)^2\right].$$

Note that $\Delta_n(\widehat{f}_n, f_0) \geq 0$ and $\Delta_n(\widehat{f}_n, f_0) = 0$ if $\widehat{f}_n$ is an empirical minimizer. The goal of this document is to obtain both lower and upper bounds of the prediction risk in terms of $\Delta_n(\widehat{f}_n, f_0)$.

2 Risk Bound w.r.t. Δ_n?

In this Section, we provide the Theorem on upper and lower bound of prediction risk $R(\widehat{f}_n, f_0)$.

Theorem 2. (Hieber, 20.) Consider the d-variate non-parametric regression model with unknown regression function f_0 satisfying $\|f_0\|_\infty \leq F$ for some $F \geq 1$. Let $\widehat{f}_n$ be any estimator taking values in the class $\mathcal{F}(L, p, s, F)$ and let $\Delta_n(\widehat{f}_n, f_0)$ be the quantity defined above. For any $\varepsilon \in (0, 1]$, there exists a constant C_ε, only depending on ε, such that with

$$\tau_{\varepsilon,n} := C_\varepsilon F^2 \frac{(s+1)\log(n(s+1)^L)p_0 p_{L+1}}{n},$$

$$\begin{aligned}(1-\varepsilon)^2\Delta_n(\widehat{f}_n, f_0) - \tau_{\varepsilon,n} &\leq R(\widehat{f}, f_0) \\ &\leq (1+\varepsilon)^2\left(\inf_{f\in\mathcal{F}(L,p,s,F)} E_{f_0}[\|f - f_0\|_n^2] + \Delta_n(\widehat{f}_n, f_0)\right) + \tau_{\varepsilon,n}.\end{aligned}$$

Proof of the Theorem 2. is involved with following two Lemmas : Lemma 4 and Lemma 5.

Lemma 4. (Hieber, 20.) Consider the d-variate non-parametric regression model with unknown regression function f_0 satisfying $\|f_0\|_\infty \leq F$ for some $F \geq 1$. Let $\widehat{f_n}$ be any estimator taking values in the class $\mathcal{F}(L, p, s, F)$ and let $\Delta_n(\widehat{f}_n, f_0)$ be the quantity defined above. Assume $\{f_0\} \cup \mathcal{F} \subset \{f : [0,1]^d \to [-F, F]\}$ for some $F \geq 1$. If $\mathcal{N}_n := \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty) \geq 3$, then,

$$(1-\varepsilon)^2\Delta_n - F^2\frac{18\log\mathcal{N}_n + 76}{n\varepsilon} - 38\delta \leq R(\widehat{f}, f_0) \leq (1+\varepsilon)^2\left[\inf_{f\in\mathcal{F}} E_{f_0}[\|f - f_0\|_n^2] + F^2\frac{18\log\mathcal{N}_n + 72}{n\varepsilon} + 32\delta F + \Delta_n\right]$$

for all $\delta, \varepsilon \in (0, 1]$.

Proof sketch of Lemma 4. Proof of Lemma 4 can be divided into several substeps. Here, we present key ideas for each substep, and all detailed proof procedures are deferred in next Section.

1. First key idea for the proof of Lemma 4 is to construct an empirical counterpart of prediction risk, where we denote it as $\widehat{R}_n(\widehat{f}_n, f_0) := E_X[\frac{1}{n}\sum_{j=1}^n \big(\widehat{f}_n(X_i) - f_0(X_i)\big)^2]$, and relate it with the risk $R_n(\widehat{f}_n, f_0)$ via certain inequality. Note that since $\widehat{f}_n$ is a stochastic term, $R(\widehat{f}_n, f_0) \neq \widehat{R}_n(\widehat{f}_n, f_0)$. Generate i.i.d. random variables $X_1^{'}, \ldots, X_n^{'}$ with the same distribution as X and independent of $(X_i)_{i=1,\ldots,n}$. Then, we can obtain a following inequality :

$$\begin{aligned}\left|R(\widehat{f}_n, f_0) - \widehat{R}_n(\widehat{f}_n, f_0)\right| &= \left|E\left[\frac{1}{n}\sum_{i=1}^n\left\{\big(\widehat{f}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(\widehat{f}(X_i) - f_0(X_i)\big)^2\right\}\right]\right| \\ &\leq \frac{F}{n}R(\widehat{f}, f_0)^{1/2}\big(36n\log\mathcal{N}_n + 2^8 n\big) + \frac{F}{n}\big(6\log\mathcal{N}_n + 11\big) + 26\delta F.\end{aligned}$$

   Bernstein's inequality and concept of δ-covering number are used for derivation.

2. Second key idea is to obtain both the lower and upper bound of risk $R(\widehat{f}_n, f_0)$ in terms of $\widehat{R}_n(\widehat{f}_n, f_0)$ by using a following inequality : Let a, b, c, d be positive real numbers, such that $|a - b| \leq 2\sqrt{a}c + d$. It can be easily shown for any $\varepsilon \in (0, 1]$,

$$(1-\varepsilon)b - d - \frac{c^2}{\varepsilon} \leq a \leq (1+\varepsilon)(b+d) + \frac{(1+\varepsilon)^2}{\varepsilon}c^2.$$

   In our case, we set $a = R(\widehat{f}, f_0)$, $b = \widehat{R}_n(\widehat{f}, f_0)$, $c = F(9n\log\mathcal{N}_n + 64n)^{1/2}/n$, and $d = F(6\log\mathcal{N}_n + 11)/n + 26\delta F$.

3. Lastly, a remaining task for obtaining the bounds in Lemma 4. is to get the upper and lower bound of empirical version of prediction risk, $\widehat{R}_n(\widehat{f}_n, f_0)$, with respect to Δ_n. Here, we use the definition of Δ_n. For any fixed $f \in \mathcal{F}$,

$$E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n\big(\widehat{f}(X_i) - f_0(X_i)\big)^2\right] \leq E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n\Big(f_0(X_i) - f(X_i)\Big)^2\right] + \Delta_n + \frac{2}{n}E_{f_0}\left[\sum_{i=1}^n\varepsilon_i\widehat{f}(X_i)\right].$$

Observe that δ-entropy of class $\mathcal{F}$ (i.e., $\log\mathcal{N}_n$) appears both at the upper and lower bound of the risk $R(\widehat{f}, f_0)$ in Lemma 4. In order to derive the bounds in Theorem 2, an upper bound on the covering number, $\mathcal{N}_n$, should be obtained.

Lemma 5. (Hieber, 20.) If $V := \prod_{\ell=0}^{L+1}(p_\ell + 1)$, then for any $\delta > 0$,

$$\log \mathcal{N}\left(\delta, \mathcal{F}(L, p, s, \infty), \|\cdot\|_\infty\right) \leq (s+1)\log\left(2\delta^{-1}(L+1)V^2\right).$$

Proof sketch of Lemma 5. The key idea for the proof of Lemma 5 is to discretize the parameter space $(\mathbf{v}_k, \mathbf{W}_k)_k$ of neural network f with certain grid size ε. Let $(\mathbf{v}_k^*, \mathbf{W}_k^*)_k$ be the network parameter on the constructed grid and f^* be the network value evaluated at the grid point, $(\mathbf{v}_k^*, \mathbf{W}_k^*)_k$. Fix $\varepsilon > 0$. Let $f, f^* \in \mathcal{F}$ be two networks, such that all parameters are ε away from each other. Then, the proof of Lemma 5 can be divided into following four steps :

1. Calculate the upper-bound of $|f(X) - f^*(X)|$ in terms of ε, V and L. Actually, we can get the bound $|f(X) - f^*(X)| \leq \varepsilon V(L+1)$. By setting $\varepsilon = \frac{\delta}{2V(L+1)}$, for any $\delta > 0$, we can get $|f(X) - f^*(X)| \leq \frac{\delta}{2} < \delta$.

2. Note that the total number of parameters in the network can be bounded by V as follows:

$$\sum_{j=0}^{L} p_\ell p_{\ell+1} + \sum_{j=1}^{L} p_\ell = \sum_{j=0}^{L}(p_\ell + 1)p_{\ell+1} - p_{L+1} \leq \prod_{\ell=0}^{L+1}(p_\ell + 1) := V.$$

3. Since the absolute value of parameters in the network are bounded by 1, it is obvious that the total number of grid points in parameter space $(\mathbf{v}_k, \mathbf{W}_k)_k$ is bounded by $\frac{2V^2(L+1)}{\delta}$.

4. Since $\mathcal{F}(L, p, s, F)$ is a class of neural network functions with at most s parameters, we can obtain a bound for the covering number of class $\mathcal{F}$ as follows,

$$\mathcal{N}\left(\delta, \mathcal{F}(L, p, s, \infty), \|\cdot\|_\infty\right) \leq \sum_{s^* \leq s}\left(\frac{2V^2(L+1)}{\delta}\right)^{s^*} \leq \left(\frac{2V^2(L+1)}{\delta}\right)^{s+1}.$$

All we need to do is to derive the bound of $|f(X) - f(X^*)| \leq \varepsilon V(L+1)$ for which we defer the proof in Section 4.

3 Proof of Lemma 4.

3.1 Step 0.

First, we divide the cases into two:1) when $\log \mathcal{N}_n \geq n$ and 2) when $\log \mathcal{N}_n \leq n$. In this step, when $\log \mathcal{N}_n \geq n$ holds, we will show that the upper and lower bounds in Lemma 4. hold trivially. For the upper bound, note that the risk $R(\widehat{f}, f_0)$ is bounded by $4F^2$. The assumptions $\log \mathcal{N}_n \geq n$ and $\varepsilon \in (0, 1]$ make the claimed upper bound in Lemma 4. already greater than $18F^2$. For the lower bound, first observe that $R(\widehat{f}, f_0) \geq 0$, then it is enough to show that when $\log \mathcal{N}_n \geq n$, the claimed lower bound in Lemma 4 is less than or equal to 0. It is easy to see that the following is true, for any fixed empirical risk minimizer, $\tilde{f}$, we have :

$$\begin{aligned}
\widehat{R}_n(\widehat{f}_n, f_0) - \widehat{R}_n(\tilde{f}_n, f_0) &= E_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}(X_i) - f_0(X_i)\right)^2 - \left(\tilde{f}(X_i) - f_0(X_i)\right)^2\right] \\
&= E_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}(X_i) - Y_i + \varepsilon_i\right)^2 - \left(\tilde{f}(X_i) - Y_i + \varepsilon_i\right)^2\right] \\
&= \Delta_n + E_{f_0}\left[\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i \widehat{f}(X_i)\right] - E_{f_0}\left[\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i \tilde{f}(X_i)\right].
\end{aligned}$$

From this equation, we can see that Δ_n is bounded by $8F^2$, since $F \geq 1$. The lower bound of Lemma 4 is bounded by $(1-\varepsilon)^2\Delta_n - 18F^2 \leq -10F^2 \leq 0$. So the lower bound holds trivially.

3.2 Step 1.

From this step, we assume $\log \mathcal{N}_n \le n$. Recall that the definition of δ-covering number $\mathcal{N}\big(\delta, \mathcal{F}(L,p,s,F), \|\cdot\|_\infty\big)$ is the minimal number of $\|\cdot\|_\infty$-balls with radius δ that covers $\mathcal{F}(L,p,s,F)$. By construction there exists a (random) j^* such that $\|\widehat{f} - f_{j^*}\|_\infty \le \delta$ for any $\widehat{f}$ achieved by any optimization methods. Generate i.i.d. random variables $X_1^{'}, \ldots, X_n^{'}$ with the same distribution as X and independent of $(X_i)_{i=1,\ldots,n}$, then we can get a following inequality under the assumptions: $\|f_j\|_\infty, \|f_0\|_\infty, \delta \le F$.

$$
\begin{aligned}
\left|R(\widehat{f}, f_0) - R_n(\widehat{f}, f_0)\right| &= \left|E_{f_0}\left[\frac{1}{n}\sum_{i=1}^n \big(\widehat{f}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(\widehat{f}(X_i) - f_0(X_i)\big)^2\right]\right| \\
&\le E_{f_0}\left|\frac{1}{n}\sum_{i=1}^n \underbrace{\big(f_{j^*}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(f_{j^*}(X_i) - f_0(X_i)\big)^2}_{:= g_{j^*}(X_i, X_i')}\right| + 9\delta F \\
&= E_{f_0}\left[\left|\frac{1}{n}\sum_{i=1}^n \frac{g_{j^*}(X_i, X_i^{'})}{r_{j^*}F}\big(r_{j^*}F\big)\right|\right] + 9\delta F \\
&\le \frac{F}{n} E_{f_0}\left[\left|\max_j \sum_{i=1}^n \left(\frac{g_j(X_i, X_i^{'})}{r_j F}\right)\right| r_{j^*}\right] + 9\delta F
\end{aligned}
$$

For second inequality, we use a following relation :

$$
\begin{aligned}
&\big(\widehat{f}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(\widehat{f}(X_i) - f_0(X_i)\big)^2 \\
&= \big(\widehat{f}(X_i^{'}) - f_{j^*}(X_i^{'}) + f_{j^*}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(\widehat{f}(X_i) - f_{j^*}(X_i) + f_{j^*}(X_i) - f_0(X_i)\big)^2 \\
&= \big(f_{j^*}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(f_{j^*}(X_i) - f_0(X_i)\big)^2 + \underbrace{\big(\widehat{f}(X_i^{'}) - f_{j^*}(X_i^{'})\big)^2}_{\le \delta^2 \le \delta F} - \underbrace{\big(\widehat{f}(X_i) - f_{j^*}(X_i)\big)^2}_{\ge 0} \\
&\quad - 2\underbrace{\big(\widehat{f}(X_i^{'}) - f_{j^*}(X_i^{'})\big)}_{\le \delta}\underbrace{\big(f_{j^*}(X_i^{'}) - f_0(X_i^{'})\big)}_{\ge -2F} - 2\underbrace{\big(\widehat{f}(X_i) - f_{j^*}(X_i)\big)}_{\le \delta}\underbrace{\big(f_{j^*}(X_i) - f_0(X_i)\big)}_{\ge -2F} \\
&\le \big(f_{j^*}(X_i^{'}) - f_0(X_i^{'})\big)^2 - \big(f_{j^*}(X_i) - f_0(X_i)\big)^2 + 9\delta F.
\end{aligned}
$$

Now we set $r_{j^*} := \sqrt{n^{-1}\log \mathcal{N}_n} \vee E^{1/2}\big[(f_{j^*}(X) - f_0(X))^2 | (X_i, Y_i)_i\big]$. Note that we have a following bound for r_{j^*} by using the notion of δ-covering number once again:

$$
r_{j^*} \le \sqrt{\frac{\log \mathcal{N}_n}{n}} + E^{1/2}\left[(\widehat{f}(X) - f_0(X))^2 \big| (X_i, Y_i)_i\right] + \delta
$$

Denoting $T := \big|\max_j \sum_{i=1}^n g_j(X_i, X_i^{'})/r_j F\big|$ and $U := E^{1/2}\big[(\widehat{f}(X) - f_0(X))^2 | (X_i, Y_i)_i\big]$, and using Cauchy-Schwartz inequality for random variables U and T (i.e., $E[UT] \le E^{1/2}[U^2]E^{1/2}[T^2]$), we have:

$$
\begin{aligned}
\left|R(\widehat{f}, f_0) - R_n(\widehat{f}, f_0)\right| &\le \frac{F}{n} E_{f_0}\left[TU + \left(\delta + \sqrt{\frac{\log \mathcal{N}_n}{n}}\right)T\right] + 9\delta F \\
&\le \frac{F}{n} R(\widehat{f}, f_0)^{1/2} E^{1/2}[T^2] + \frac{F}{n}\left(\delta + \sqrt{\frac{\log \mathcal{N}_n}{n}}\right)E[T] + 9\delta F. \qquad (1)
\end{aligned}
$$

Now, the first and second moment of T need to be controlled. (i.e.,$E[T]$ and $E[T^2]$). First observe that $Eg_j(X_i, X_i^{'})/r_j = 0$, $\mathrm{Var}(g_j(X_i, X_i^{'})/r_j) \le 8F^2$ and $|g_j(X_i, X_i^{'})/r_j| \le 4F^2/r_j$. We can use Bernstein's inequality which states that for independent and centered random variables $U_1, \ldots, U_n$, satisfying $|U_i| \le M$,

$P(|\sum_{i=1}^n U_i| \geq t) \leq 2\exp(-t^2/[2Mt/3 + 2\sum_{i=1}^n Var(U_i)])$.

$$\begin{aligned} P(T \geq t) = P\Big(\Big|\max_j \sum_{i=1}^n \Big(\frac{g_j(X_i, X_i^{'})}{r_j F}\Big)\Big| \geq t\Big) &\leq P\Big(\max_j \Big|\sum_{i=1}^n \Big(\frac{g_j(X_i, X_i^{'})}{r_j}\Big)\Big| \geq Ft\Big) \\ &\leq \mathcal{N}_n \max_j P\Big(\Big|\sum_{i=1}^n \Big(\frac{g_j(X_i, X_i^{'})}{r_j}\Big)\Big| \geq Ft\Big) \\ &\leq 2\mathcal{N}_n \max_j \exp\Big(-\frac{(Ft)^2}{2\cdot 4F^2 t/(3r_j) + 2\cdot 8F^2 n}\Big) \\ &= 2\mathcal{N}_n \max_j \exp\Big(-\frac{t^2}{8t/(3r_j) + 16n}\Big) \wedge 1 \end{aligned}$$

The first term in the denominator dominates for large enough t: Specifically, for $t \geq 6\sqrt{n\log\mathcal{N}_n}$, we have

$$\exp\Big(-\frac{t^2}{8t/(3r_j) + 16n}\Big) \leq \exp\Big(-\frac{t}{16/(3r_j)}\Big) \leq \exp\Big(-\frac{3t\sqrt{\log\mathcal{N}_n}}{16\sqrt{n}}\Big),$$

where we use the inequality $r_j \geq \sqrt{n^{-1}\log\mathcal{N}_n}$ in the last inequality. We therefore find following inequality for the two moments of T:

$$\begin{aligned} E[T] = \int_0^\infty P(T \geq t)dt &\leq 6\sqrt{n\log\mathcal{N}_n} + \int_{6\sqrt{n\log\mathcal{N}_n}}^\infty 2\mathcal{N}_n \exp\Big(-\frac{3t\sqrt{\log\mathcal{N}_n}}{16\sqrt{n}}\Big)dt \\ &\leq 6\sqrt{n\log\mathcal{N}_n} + \frac{32}{3}\sqrt{\frac{n}{\log\mathcal{N}_n}}. \end{aligned}$$

Here, in second inequality, for $3 \leq \mathcal{N}_n$, we use :

$$2\mathcal{N}_n \int_{6\sqrt{n\log\mathcal{N}_n}}^\infty \exp\Big(-\frac{3t\sqrt{\log\mathcal{N}_n}}{16\sqrt{n}}\Big)dt = \Big(\mathcal{N}_n^{-\frac{7}{2}}\Big)\frac{32}{3}\sqrt{\frac{n}{\log\mathcal{N}_n}} \leq \frac{32}{3}\sqrt{\frac{n}{\log\mathcal{N}_n}}.$$

Similarly $E[T^2]$ can be controlled as follows:

$$\begin{aligned} E[T^2] = \int_0^\infty P(T^2 \geq u)du &= \int_0^\infty P(T \geq \sqrt{u})du \\ &\leq 36n\log\mathcal{N}_n + \int_{36n\log\mathcal{N}_n}^\infty 2\mathcal{N}_n \exp\Big(-\frac{3\sqrt{u}\sqrt{\log\mathcal{N}_n}}{16\sqrt{n}}\Big)du \\ &\leq 36n\log\mathcal{N}_n + 2^8 n. \end{aligned}$$

In the third inequality, we use the identity : $\int_{b^2}^\infty e^{-\sqrt{u}a}du = 2(ab+1)e^{-ab}/b^2$. By setting $a = \frac{3\sqrt{\log\mathcal{N}_n}}{16\sqrt{n}}$ and $b = 6\sqrt{n\log\mathcal{N}_n}$ and using $3 \leq \mathcal{N}_n$ once again, we have :

$$\int_{36n\log\mathcal{N}_n}^\infty 2\mathcal{N}_n \exp\Big(-\frac{3\sqrt{u}\sqrt{\log\mathcal{N}_n}}{16\sqrt{n}}\Big)du = \Big(\frac{1}{2} + \frac{4}{9\log\mathcal{N}_n}\Big)\mathcal{N}_n^{-\frac{1}{8}}\cdot 2^8 n \leq 2^8 n.$$

Lastly, plugging the bound of $E[T]$ and $E[T^2]$ in (1) and using the relation $1 \leq \log\mathcal{N}_n \leq n$, we get :

$$\Big|R(\widehat{f}, f_0) - R_n(\widehat{f}, f_0)\Big| \leq \frac{F}{n}R(\widehat{f}, f_0)^{1/2}\big(36n\log\mathcal{N}_n + 2^8 n\big)^{1/2} + \frac{F}{n}\big(6\log\mathcal{N}_n + 11\big) + 26\delta F. \tag{2}$$

3.3 Step 2.

Let a, b, c, d be positive real numbers, such that $|a-b| \leq 2\sqrt{a}c + d$. It can be easily shown for any $\varepsilon \in (0, 1]$,

$$(1-\varepsilon)b - d - \frac{c^2}{\varepsilon} \leq a \leq (1+\varepsilon)(b+d) + \frac{(1+\varepsilon)^2}{\varepsilon}c^2. \tag{3}$$

Upper bound of (3) can be obtained as follows:

$$\begin{aligned} a - b \leq 2\sqrt{a}c + d &= 2\sqrt{\frac{\varepsilon}{1+\varepsilon}a}\sqrt{\frac{1+\varepsilon}{\varepsilon}}c + d \\ &\leq \frac{\varepsilon}{1+\varepsilon}a + \frac{1+\varepsilon}{\varepsilon}c^2 + d. \end{aligned}$$

Lower bound of (3) can be obtained as follows:

$$\begin{aligned} b - a \leq 2\sqrt{a}c + d &= 2\sqrt{\frac{\varepsilon}{1-\varepsilon}a}\sqrt{\frac{1-\varepsilon}{\varepsilon}}c + d \\ &\leq \frac{\varepsilon}{1-\varepsilon}a + \frac{1-\varepsilon}{\varepsilon}c^2 + d. \end{aligned}$$

Rearranging the terms and using the fact $\varepsilon \in (0, 1]$ yield the claim. Now, set $a = R(\widehat{f}, f_0)$, $b = \widehat{R}_n(\widehat{f}, f_0)$, $c = F(9n \log \mathcal{N}_n + 64n)^{1/2}/n$, and $d = F(6 \log \mathcal{N}_n + 11)/n + 27\delta F$. Finally, we relate the risk $R(\widehat{f}, f_0) = E[(\widehat{f}(X) - f_0(X))^2]$ to its empirical counterpart $\widehat{R}_n(\widehat{f}, f_0)$ via the inequalities,

$$\begin{aligned} (1-\varepsilon)\widehat{R}_n(\widehat{f}, f_0) - \frac{F^2}{n\varepsilon}(15 \log \mathcal{N}_n + 75) - 26\delta F &\leq R(\widehat{f}, f_0) \\ &\leq (1+\varepsilon)\left(\widehat{R}_n(\widehat{f}, f_0) + (1+\varepsilon)\frac{F^2}{n\varepsilon}(12 \log \mathcal{N}_n + 70) + 26\delta F\right). \end{aligned}$$

Note that we have used $2 \leq (1+\varepsilon)/\varepsilon$ for obtaining the upper-bound.

3.4 Step 3.

Given an estimator $\tilde{f}$ taking values in $\mathcal{F}$, let j' be such that $\|\tilde{f} - f_{j'}\|_\infty \leq \delta$. First, using the independence assumption of ε_i and X_i, it can be easily seen that $E[\varepsilon_i f_0(X_i)] = 0$ for deterministic f_0. Secondly, it is also easy to see that $|E[\sum_{i=1}^n \varepsilon_i(\tilde{f}(X_i) - f_{j'}(X_i))]| \leq n\delta$ by using the fact $E[|\varepsilon_i|] = \sqrt{\frac{2}{\pi}}$. So the bound of can be obtained as follows :

$$\begin{aligned} \left| E\left[\frac{2}{n}\sum_{i=1}^n \varepsilon_i \tilde{f}(X_i)\right]\right| &= \left| E\left[\frac{2}{n}\sum_{i=1}^n \varepsilon_i\Big(\tilde{f}(X_i) - f_0(X_i)\Big)\right]\right| \\ &\leq \left| E\left[\frac{2}{n}\sum_{i=1}^n \varepsilon_i\Big(f_{j'}(X_i) - f_0(X_i)\Big)\right]\right| + 2\delta \\ &= \frac{2}{\sqrt{n}}\left| E\left[\frac{\sum_{i=1}^n \varepsilon_i(f_{j'}(X_i) - f_0(X_i))}{\sqrt{n}\|f_{j'} - f_0\|_n}\|f_j - f_0\|_n\right]\right| + 2\delta \\ &\leq \frac{2}{\sqrt{n}}E\left[\left|\frac{\sum_{i=1}^n \varepsilon_i(f_{j'}(X_i) - f_0(X_i))}{\sqrt{n}\|f_{j'} - f_0\|_n}\right|\|f_{j'} - f_0\|_n\right] + 2\delta \\ &\leq \frac{2}{\sqrt{n}}E\left[\left|\underbrace{\frac{\sum_{i=1}^n \varepsilon_i(f_{j'}(X_i) - f_0(X_i))}{\sqrt{n}\|f_{j'} - f_0\|_n}}_{:=\xi_{j'}}\right|\big(\|\tilde{f} - f_0\|_n + \delta\big)\right] + 2\delta \end{aligned}$$

Here, note that random variable $\xi_{j'}$ follows standard normal (i.e., $\xi_{j'} \sim \mathcal{N}(0,1)$). Next, we further need to control the term $E\left[|\xi_{j'}|\big(\|\tilde{f}-f_0\|_n+\delta\big)\right]$ as follows:

$$\begin{aligned}
E\left[|\xi_{j'}|\big(\|\tilde{f}-f_0\|_n+\delta\big)\right] &= E\left[|\xi_{j'}|\|\tilde{f}-f_0\|_n\right]+\delta E[|\xi_{j'}|] \\
&\le E\left[|\xi_{j'}|\|\tilde{f}-f_0\|_n\right]+\delta \\
&\le \widehat{R}_n^{1/2}(\tilde{f},f_0)E^{1/2}[\xi_{j'}^2]+\delta \\
&\le \widehat{R}_n^{1/2}(\tilde{f},f_0)\sqrt{3\log\mathcal{N}_n+1}+\delta \\
&\le \left(\widehat{R}_n^{1/2}(\tilde{f},f_0)+\delta\right)\sqrt{3\log\mathcal{N}_n+1}
\end{aligned}$$

In first inequality, the fact $E[|\xi_{j'}|]=\sqrt{\frac{2}{\pi}}\le 1$ is used. For second inequality, Cauchy-Schwarz inequality is employed. For third inequality, the fact $E[\xi_{j'}^2]\le 3\log\mathcal{N}_n+1$ is used. In the last inequality, we use $1\le\log\mathcal{N}_n$. Here, we provide a simple proof for the fact :

$$E[\xi_{j'}^2]\le 3\log\mathcal{N}_n+1, \quad \xi_{j'}\sim\mathcal{N}(0,1).$$

Proof. Let $Z=\max_{j=1,\ldots,\mathcal{N}_n}\xi_{j'}^2$. Since $Z\le\sum_{j=1}^{\mathcal{N}_n}\xi_j^2$, we have $E[Z]\le\mathcal{N}_n$. For $\mathcal{N}_n\in\{1,2,3\}$, it can be checked that $\mathcal{N}_n\le 3\log(\mathcal{N}_n)+1$ and the result holds in this case. Now, we consider the case where $\mathcal{N}_n\ge 4$. Mill's ratio gives $P(|\xi_1|\ge\sqrt{t})=2P(\xi_1\ge\sqrt{t})\le\frac{2}{\sqrt{2\pi t}}e^{-t/2}$. For any $T\ge 0$ and by using union bound,

$$\begin{aligned}
E[Z] &= \int_0^\infty P(Z\ge t)dt \\
&\le T+\int_T^\infty P(Z\ge t)dt \le T+\mathcal{N}_n\int_T^\infty P(\xi_1^2\ge t)dt \\
&\le T+\mathcal{N}_n\int_T^\infty P(\xi_1^2\ge t)dt \le T+\mathcal{N}_n\int_T^\infty\frac{2}{\sqrt{2\pi t}}e^{-t/2}dt \\
&\le T+\frac{2\mathcal{N}_n}{\sqrt{2\pi T}}\int_T^\infty e^{-t/2}dt = T+\frac{4\mathcal{N}_n}{\sqrt{2\pi T}}e^{-T/2}.
\end{aligned}$$

Set $T=2\log(\mathcal{N}_n)$, we find

$$E[Z]\le 2\log(\mathcal{N}_n)+\frac{2}{\sqrt{\pi\log(\mathcal{N}_n)}}.$$

This yields the claim for $\mathcal{N}_n\ge 4$. □

Further, because of $\log\mathcal{N}_n\le n$, we have $2n^{-1/2}\delta\sqrt{3\log\mathcal{N}_n+1}\le 4\delta$. By combining the facts, we finally get a following inequality:

$$\left|E\left[\frac{2}{n}\sum_{i=1}^n\varepsilon_i\tilde{f}(X_i)\right]\right|\le 2\sqrt{\frac{\widehat{R}_n(\tilde{f},f_0)\big(3\log\mathcal{N}_n+1\big)}{n}}+6\delta.$$

3.5 Step 4.

By the definition of Δ_n, for any fixed $f \in \mathcal{F}$, we have $E[\frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{f}(X_i))^2] \le E[\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2] + \Delta_n$. We want to get a bound for $\widehat{R}_n(\widehat{f}, f_0)$ for any fixed $f \in \mathcal{F}$,

$$\begin{aligned}
\widehat{R}_n(\widehat{f}, f_0) &= E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n \big(\widehat{f}(X_i) - f_0(X_i)\big)^2\right] \\
&= E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n \big(\widehat{f}(X_i) + \varepsilon_i - Y_i\big)^2\right] \\
&= E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n \Big((Y_i - \widehat{f}(X_i))^2 + 2\varepsilon_i(Y_i - \widehat{f}(X_i)) + \varepsilon_i^2\Big)\right] \\
&\le E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n \big(Y_i - f(X_i)\big)^2\right] + \Delta_n - \frac{2}{n}E_{f_0}\left[\sum_{i=1}^n \varepsilon_i(Y_i - \widehat{f}(X_i))\right] + E_{f_0}[\varepsilon_1^2] \\
&\le E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n \Big(f_0(X_i) - f(X_i) + \varepsilon_i\Big)^2\right] + \Delta_n + \frac{2}{n}E_{f_0}\left[\sum_{i=1}^n \varepsilon_i\Big(\widehat{f}(X_i) - f_0(X_i)\Big)\right] - E_{f_0}[\varepsilon_1^2] \\
&\le E_{f_0}\left[\frac{1}{n}\sum_{j=1}^n \Big(f_0(X_i) - f(X_i)\Big)^2\right] + \Delta_n + \frac{2}{n}E_{f_0}\left[\sum_{i=1}^n \varepsilon_i\widehat{f}(X_i)\right].
\end{aligned}$$

Note that we have used the relation $E[\varepsilon_i f(X_i)] = 0$ at second to the last inequality. Finally, combining the upper bound we get on $\frac{2}{n}E_{f_0}\left[\sum_{i=1}^n \varepsilon_i\widehat{f}(X_i)\right]$ at Step 3., we obtain the upper-bound for $\widehat{R}_n(\widehat{f}, f_0)$ as follows:

$$\widehat{R}_n(\widehat{f}, f_0) \le E_{f_0}[\|f - f_0\|_n^2] + 2\sqrt{\frac{\widehat{R}_n(\tilde{f}, f_0)\big(3\log\mathcal{N}_n + 1\big)}{n}} + 6\delta + \Delta_n.$$

Lastly using the upper bound of (3) and by setting $a = \widehat{R}_n(\widehat{f}, f_0)$, $b = 0$, $c = \sqrt{\frac{3\log\mathcal{N}_n+1}{n}}$, and $d = E_{f_0}[\|f - f_0\|_n^2] + 6\delta + \Delta$, we get the following upper bound of $\widehat{R}_n(\widehat{f}, f_0)$ for any $\varepsilon \in (0, 1]$:

$$\widehat{R}_n(\widehat{f}, f_0) \le (1+\varepsilon)\bigg(\inf_{f\in\mathcal{F}} E_{f_0}[\|f - f_0\|_n^2] + F^2\frac{6\log\mathcal{N}_n + 2}{n\varepsilon} + 6\delta + \Delta_n\bigg).$$

As for the lower bound of $\widehat{R}_n(\widehat{f}, f_0)$, first it is trivial to see that

$$\begin{aligned}
\widehat{R}_n(\widehat{f}, f_0) - \widehat{R}_n(\tilde{f}, f_0) &= \Delta_n + E_{f_0}\left[\frac{2}{n}\sum_{i=1}^n \varepsilon_i\widehat{f}(X_i)\right] - E_{f_0}\left[\frac{2}{n}\sum_{i=1}^n \varepsilon_i\tilde{f}(X_i)\right] \\
&\ge \Delta_n - 2\sqrt{\frac{\widehat{R}_n(\widehat{f}, f_0)\big(3\log\mathcal{N}_n + 1\big)}{n}} - 2\sqrt{\frac{\widehat{R}_n(\tilde{f}, f_0)\big(3\log\mathcal{N}_n + 1\big)}{n}} - 12\delta \\
&\ge \Delta_n - \frac{\varepsilon}{1-\varepsilon}\widehat{R}_n(\widehat{f}, f_0) - \frac{3\log\mathcal{N}_n + 1}{\varepsilon n} - \widehat{R}_n(\tilde{f}, f_0) - 12\delta.
\end{aligned}$$

Here in the third inequality, we have used following inequalities, for any $\varepsilon \in (0, 1]$,

$$\begin{aligned}
2\sqrt{\frac{\widehat{R}_n(\widehat{f}, f_0)\big(3\log\mathcal{N}_n + 1\big)}{n}} &\le \frac{\varepsilon}{1-\varepsilon}\widehat{R}_n(\widehat{f}, f_0) + \frac{1-\varepsilon}{\varepsilon}\bigg(\frac{3\log\mathcal{N}_n + 1}{n}\bigg), \\
2\sqrt{\frac{\widehat{R}_n(\tilde{f}, f_0)\big(3\log\mathcal{N}_n + 1\big)}{n}} &\le \widehat{R}_n(\tilde{f}, f_0) + \frac{3\log\mathcal{N}_n + 1}{n}.
\end{aligned}$$

Rearranging the term, for any $\varepsilon \in (0, 1]$,

$$\widehat{R}_n(\widehat{f}, f_0) \geq (1 - \varepsilon)\bigg(\Delta_n - \frac{3 \log \mathcal{N}_n + 1}{n\varepsilon} - 12\delta\bigg).$$