SUPPLEMENT TO "CONDITIONAL DENSITY ESTIMATION WITH DEEP NEURAL NETWORKS"

Chenxuan He[1], Yuan Gao[2], Liping Zhu[1] and Jian Huang[3]

[1]*Renmin University of China,* [2]*Nankai University*

and [3]*The Hong Kong Polytechnic University*

Supplementary Material

The Supplementary Material contains additional simulation results and provides proofs for each result stated in the paper.

S1 Additional Simulation Results

In Table S1, we compare the unnormalized estimator $\hat{f}_{n,\text{aug}}$ in (5.1) with the post-normalized estimator $\tilde{f}_{n,\text{aug}}$ in (5.2) introduced in Section 5. The results indicate that the unnormalized and post-normalized estimators achieve similar excess risk across the settings considered.

S2 Proof of Lemma 1

Recall the kernel $K(u)$ is symmetric and of the order β. We define $s = \lceil \beta \rceil - 1$ and $r = \beta - s$. Then we have $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$ and $\int_{\mathbb{R}} u^l K(u)\mathrm{d}u = 0$ for $l = 1, \ldots, s$. The expectation of

Table S1: Normalized and unnormalized NN results for global fixed $\mathbf{x}$. Standard errors are shown in parentheses. Values are multiplied by 1000 for Setting 1.1 and Setting 2.1 and by 100 for Setting 1.2 and Setting 2.2, except for a and n. All results are based on 200 replications.

		Setting 1.1		Setting 1.2	
a	n	NN	NN$_{un}$	NN	NN$_{un}$
	500	$0.53_{(0.65)}$	$0.67_{(0.50)}$	$0.76_{(0.74)}$	$1.84_{(0.83)}$
-0.3	1000	$0.40_{(0.57)}$	$0.48_{(0.36)}$	$0.73_{(0.68)}$	$1.67_{(0.87)}$
	2000	$0.35_{(0.36)}$	$0.38_{(0.29)}$	$0.71_{(0.61)}$	$1.49_{(0.93)}$
	500	$0.47_{(0.56)}$	$0.73_{(0.57)}$	$0.79_{(0.73)}$	$1.79_{(0.86)}$
0	1000	$0.38_{(0.49)}$	$0.54_{(0.43)}$	$0.79_{(0.69)}$	$1.64_{(0.89)}$
	2000	$0.30_{(0.34)}$	$0.40_{(0.37)}$	$0.72_{(0.61)}$	$1.45_{(0.92)}$
	500	$0.45_{(0.58)}$	$0.67_{(0.52)}$	$0.78_{(0.71)}$	$1.83_{(0.82)}$
0.3	1000	$0.36_{(0.47)}$	$0.48_{(0.37)}$	$0.78_{(0.68)}$	$1.70_{(0.89)}$
	2000	$0.32_{(0.42)}$	$0.36_{(0.30)}$	$0.67_{(0.60)}$	$1.48_{(0.92)}$
		Setting 2.1		Setting 2.2	
a	n	NN	NN$_{un}$	NN	NN$_{un}$
	500	$0.83_{(0.72)}$	$0.70_{(0.63)}$	$0.12_{(0.11)}$	$0.19_{(0.16)}$
-0.3	1000	$0.48_{(0.40)}$	$0.36_{(0.28)}$	$0.09_{(0.08)}$	$0.21_{(0.17)}$
	2000	$0.33_{(0.26)}$	$0.22_{(0.13)}$	$0.06_{(0.08)}$	$0.42_{(0.38)}$
	500	$0.56_{(0.64)}$	$0.49_{(0.52)}$	$0.13_{(0.09)}$	$0.19_{(0.16)}$
0	1000	$0.27_{(0.25)}$	$0.22_{(0.19)}$	$0.10_{(0.07)}$	$0.21_{(0.16)}$
	2000	$0.20_{(0.17)}$	$0.14_{(0.10)}$	$0.07_{(0.07)}$	$0.40_{(0.33)}$
	500	$0.37_{(0.36)}$	$0.36_{(0.33)}$	$0.13_{(0.08)}$	$0.19_{(0.14)}$
0.3	1000	$0.16_{(0.13)}$	$0.15_{(0.11)}$	$0.11_{(0.07)}$	$0.21_{(0.15)}$
	2000	$0.12_{(0.10)}$	$0.10_{(0.06)}$	$0.08_{(0.06)}$	$0.36_{(0.27)}$

$K_b(Y-y)$ given $\mathbf{X}=\mathbf{x}$ satisfies the following expressions

$$
\begin{aligned}
&\mathbb{E}_Y\{K_b(Y-y)\mid\mathbf{X}=\mathbf{x}\}\\
=&\int f_*(y'\mid\mathbf{x})K_b(y-y')\mathrm{d}y'\\
\overset{(i)}{=}&\int f_*(y+bz\mid\mathbf{x})K(z)\mathrm{d}z\\
\overset{(ii)}{=}&\int\Big\{\sum_{a=0}^{s-1}(bz)^a f_*^{(a)}(y\mid\mathbf{x})/a!+(bz)^s f_*^{(s)}(y+bz\lambda\mid\mathbf{x})/s!\Big\}K(z)\mathrm{d}z\\
\overset{(iii)}{=}&f_*(y\mid\mathbf{x})+b^s\int z^s\big\{f_*^{(s)}(y+bz\lambda\mid\mathbf{x})-f_*^{(s)}(y\mid\mathbf{x})\big\}K(z)\mathrm{d}z/s!\\
\overset{(iv)}{\le}&f_*(y\mid\mathbf{x})+b^s\int z^s\mathcal{B}|bz\lambda|^r K(z)\mathrm{d}z/s!\\
\overset{(v)}{\le}&f_*(y\mid\mathbf{x})+b^\beta\int\mathcal{B}|z|^\beta\lambda^r K(z)\mathrm{d}z/s!\\
\overset{(vi)}{=}&f_*(y\mid\mathbf{x})+C_K b^\beta,
\end{aligned}
$$

where (i) is derived from the transformation $y' = y + bz$ and the symmetric property of $K(u)$, (ii) is given by a Taylor expansion and $\lambda \in (0,1)$, (iii) is due to the property of the β-order kernel, (iv) is due to the property of the Hölder class, (v) is by taking out b^r, and (vi) is derived from the property of the β-order kernel.

S3 Proof of Lemma 2

For the sample $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, we formulate the outcome as $\{K_b(Y_i - y)\}_{i=1}^n$ for a given y. The regression problem can be further formulated as $K_b(Y_i - y) = f_*(y \mid \mathbf{X}_i) + \varepsilon_i$. In addition, since $\{Y_i\}_{i=1}^n$ are independent, the outcomes $\{K_b(Y_i - y)\}_{i=1}^n$ are also independent for a given y. Throughout the proofs, we use the notation $\|\hat{f}_n - f_*\|_2$ to denote the L_2-norm of the difference between two functions, as defined in Definition 1 of the main text. For clarity, for any fixed y, we consider the functions $\mathbf{x} \mapsto \hat{f}_n(y \mid \mathbf{x})$ and $\mathbf{x} \mapsto f_*(y \mid \mathbf{x})$, and define

$$\|\hat{f}_n - f_*\|_2^2 = \int |\hat{f}_n(y \mid \mathbf{x}) - f_*(y \mid \mathbf{x})|^2 \mathrm{d}\nu(\mathbf{x}) = \mathbb{E}_{\mathbf{X}}\big\{\hat{f}_n(y \mid \mathbf{X}) - f_*(y \mid \mathbf{X})\big\}^2,$$

where the expectation is taken over the distribution of $\mathbf{X}$ and ν denotes its probability measure.

We perform the error decomposition as below

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2 \\
\overset{(i)}{=}&\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2 + 2\mathbb{E}_n(\hat{f}_n - f_*)^2\right\} \\
\overset{(ii)}{=}&\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2 + 2\mathbb{E}_n\{\hat{f}_n - K_b(Y_i - y) + K_b(Y_i - y) - f_*\}^2\right] \\
\overset{(iii)}{=}&\mathbb{E}_{\mathcal{D}_n}\Big(\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2 + 2\mathbb{E}_n\{L(\hat{f}_n, y)\} + 2\mathbb{E}_n\{L(f_*, y)\} \\
&-4\mathbb{E}_n[\{\hat{f}_n - K_b(Y_i - y)\}\{f_* - K_b(Y_i - y)\}]\Big) \\
\overset{(iv)}{\leq}&\mathbb{E}_{\mathcal{D}_n}\Big(\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2 + 2\mathbb{E}_n\{L(f_n, y)\} + 2\mathbb{E}_n\{L(f_*, y)\} \\
&-4\mathbb{E}_n[\{\hat{f}_n - f_n + f_n - K_b(Y_i - y)\}\{f_* - K_b(Y_i - y)\}]\Big) \\
\overset{(v)}{=}&\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2 + 2\mathbb{E}_n(f_n - f_*)^2 + 4\mathbb{E}_n\{\varepsilon_i(\hat{f}_n - f_n)\}\right] \\
\overset{(vi)}{=}&\underbrace{\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2\right\}}_{=:T_1} + 4\underbrace{\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\left\{\varepsilon_i(\hat{f}_n - f_n)\right\}\right]}_{=:T_2} + 2\underbrace{\mathbb{E}_{\mathbf{X}}(f_n - f_*)^2}_{=:T_3} \\
=&T_1 + 4T_2 + 2T_3,
\end{aligned}
$$

where (i) and (ii) are derived from adding and subtracting same terms, (iii) is derived from splitting the squared term, (iv) follows from $\mathbb{E}_n\{L(\hat{f}_n, y)\} \leq \mathbb{E}_n\{L(f_n, y)\}$ since $\hat{f}_n$ is the empirical risk minimizer, (v) is due to $\varepsilon_i = K_b(Y_i - y) - f_*$ and $\mathbb{E}_n(f_n - f_*)^2 = \mathbb{E}_n\{L(f_n, y)\} + \mathbb{E}_n\{L(f_*, y)\} - 2\mathbb{E}_n[\{f_n - K_b(Y_i - y)\}\{f_* - K_b(Y_i - y)\}]$, and (vi) follows from that f_n and f_* are not stochastic.

S4 Proof of Lemma 3

Our proof of Lemma 3 is inspired by the proof of Lemma 4 from Schmidt-Hieber (2020). Given $\|\hat{f}_n\|_\infty \leq \mathcal{B}$, $\|f_*\|_\infty \leq \mathcal{B}$, and $\mathcal{B} > 1$, our goal is to bound the error $\mathbb{E}_{\mathcal{D}_n}\{\mathbb{E}_{\mathbf{X}}(\hat{f}_n -$

$f_*)^2 - 2\mathbb{E}_n(\hat{f}_n - f_*)^2\}$. Given an empirical $(1/n)$-covering number $\mathcal{N}_\infty(1/n, \mathcal{F}_n, \mathcal{D}_n)$ shown in Definition 2, there exists a collection of functions $\{f_j \in \mathcal{F}_n, j = 1, \ldots, \mathcal{N}_\infty(1/n, \mathcal{F}_n, \mathcal{D}_n)\}$ such that for all $f \in \mathcal{F}_n$, there exists a j^* such that $\|f - f_{j^*}\|_{L_\infty(\nu_n)} \leq 1/n$. For simplicity, let $\mathcal{N}_\infty$ denote $\sup_{\mathcal{D}_n} \mathcal{N}_\infty(1/n, \mathcal{F}_n, \mathcal{D}_n)$ throughout this proof.

Let $\mathbb{E}'_n$ be the empirical mean of a dataset $\mathcal{D}'_n$ independent of $\mathcal{D}_n$, and define $r_j = (\log \mathcal{N}_\infty/n)^{1/2} \vee \|f_j - f_*\|_2$ and $g_n = \max_j |n\{\mathbb{E}'_n(f_j - f_*)^2 - \mathbb{E}_n(f_j - f_*)^2\}/(\mathcal{B}r_j)|$. To bound the term T_1, we first consider the following expressions

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\{\mathbb{E}_{\mathbf{X}}(\hat{f}_n - f_*)^2 - \mathbb{E}_n(\hat{f}_n - f_*)^2\} \\
\overset{(i)}{=}&\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n}\{\mathbb{E}'_n(\hat{f}_n - f_*)^2 - \mathbb{E}_n(\hat{f}_n - f_*)^2\} \\
\overset{(ii)}{\leq}&\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n}\left|\mathbb{E}'_n(f_{j^*} - f_*)^2 - \mathbb{E}_n(f_{j^*} - f_*)^2\right| + \frac{9\mathcal{B}}{n} \\
\overset{(iii)}{\leq}& \mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n} \max_j \left|\mathbb{E}'_n(f_j - f_*)^2 - \mathbb{E}_n(f_j - f_*)^2\right|/(\mathcal{B}r_j) \times \mathcal{B}r_j + \frac{9\mathcal{B}}{n}, \\
\overset{(iv)}{\leq}&\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n} \frac{g_n}{n} \times \mathcal{B}\left\{(\log \mathcal{N}_\infty/n)^{1/2} + \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2} + \frac{1}{n}\right\} + \frac{9\mathcal{B}}{n} \\
\overset{(v)}{\leq}&\frac{1}{n}\mathcal{B}\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n} g_n \left\{(\log \mathcal{N}_\infty/n)^{1/2} + \frac{1}{n}\right\} + \frac{1}{n}\mathcal{B}\left(\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n} g_n^2\right)^{1/2} \times \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2} + \frac{9\mathcal{B}}{n},
\end{aligned} \tag{S4.1}$$

where (i) is due to the independence between $\hat{f}_n$ and $\mathcal{D}'_n$, (ii) is derived from the definition of covering numbers, $\|f_{j^*}\|_\infty \leq \mathcal{B}$, and $\mathcal{B} > 1$, (iii) is by taking the maximum, (iv) is given by the definition of r_j and the property of covering numbers, and (v) is by the Cauchy-Schwarz inequality. It remains to obtain a bound for g_n which involves the following the Bernstein inequality.

Lemma 1. *(Bernstein inequality, Vershynin, 2018, Theorem 2.8.4) Let* $U_1, \ldots, U_n$ *be inde-*

pendent mean-zero random variables such that $|U_i| \leq K$ for all i. Then, for every $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} U_i\right| \geq t\right) \leq 2\exp\left\{-t^2/(2\sigma^2+2Kt/3)\right\}.$$

Here $\sigma^2 = \sum_{i=1}^{n} EU_i^2$ is the variance of the sum.

Let us define $g_{i,j} = [\{f_j(y \mid \mathbf{X}_i') - f_*(y \mid \mathbf{X}_i')\}^2 - \{f_j(y \mid \mathbf{X}_i) - f_*(y \mid \mathbf{X}_i)\}^2]/(\mathcal{B}r_j)$ and rewrite $g_n = \max_j |\sum_i g_{i,j}|$. The variance of $g_{i,j}$ satisfies

$$\begin{aligned}\operatorname{var}(g_{i,j}) &\overset{(i)}{=} 2\operatorname{var}\left[\{f_j(y \mid \mathbf{X}_i) - f_*(y \mid \mathbf{X}_i)\}^2\right]/(\mathcal{B}^2 r_j^2)\\ &\overset{(ii)}{\leq} 2\mathbb{E}\left[\{f_j(y \mid \mathbf{X}_i) - f_*(y \mid \mathbf{X}_i)\}^4\right]/(\mathcal{B}^2 r_j^2)\\ &\overset{(iii)}{\leq} 8\mathcal{B}^2\mathbb{E}\{f_j(y \mid \mathbf{X}_i) - f_*(y \mid \mathbf{X}_i)\}^2/(\mathcal{B}^2 r_j^2)\\ &\overset{(iv)}{\leq} 8,\end{aligned}$$

where (i) follows from the independence between $\mathbf{X}_i$ and $\mathbf{X}_i'$, (ii) follows from $\operatorname{var}(\eta) \leq \mathbb{E}(\eta^2)$ for a random variable η, (iii) is derived from $|\{f_j(y \mid \mathbf{X}_i) - f_*(y \mid \mathbf{X}_i)\}| \leq 2\mathcal{B}$, and (iv) follows from the definition of r_j. The tail probability of g_n satisfies

$$\begin{aligned}\mathbb{P}\left(g_n \geq t\right) =&\mathbb{P}\left(\max_j \left|\sum_{i=1}^{n} g_{i,j}\right| \geq t\right)\\ \overset{(i)}{\leq}&\mathcal{N}_\infty \max_j \mathbb{P}\left(\left|\sum_{i=1}^{n} g_{i,j}\right| \geq t\right)\\ \overset{(ii)}{\leq}&2\mathcal{N}_\infty \max_j \exp\left[-t^2/\{16n + 16\mathcal{B}t/(3r_j)\}\right]\\ \overset{(iii)}{\leq}&2\mathcal{N}_\infty \exp\left[-t^2\Big/\left\{16n + 16\mathcal{B}tn^{1/2}\Big/\left(3\sqrt{\log\mathcal{N}_\infty}\right)\right\}\right],\end{aligned}$$

where (i) follows from the union bound, (ii) follows from Lemma 1 and $g_{i,j} \leq 8\mathcal{B}/r_j$, and (iii) is due to $r_j \geq (\log\mathcal{N}_\infty/n)^{1/2}$. Therefore, the expectation of g_n satisfies the following bounds

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n} g_n = & \int_{t>0} \mathbb{P}(g_n \geq t)\mathrm{d}t \\
& \overset{(i)}{\leq} a_n + \int_{a_n}^{\infty} 2\mathcal{N}_\infty \exp\left[-t^2 \Big/ \left\{16n + 16\mathcal{B}tn^{1/2} \Big/ \left(3\sqrt{\log \mathcal{N}_\infty}\right)\right\}\right] \mathrm{d}t \\
& \overset{(ii)}{\leq} a_n + \int_{a_n}^{\infty} 2\mathcal{N}_\infty \exp\left\{-3t\sqrt{\log \mathcal{N}_\infty} \Big/ \left(32\mathcal{B}n^{1/2}\right)\right\} \mathrm{d}t \\
& \overset{(iii)}{\leq} a_n + \frac{64\mathcal{B}n^{1/2}}{3\sqrt{\log \mathcal{N}_\infty}} \mathcal{N}_\infty \exp\left\{-3a_n\sqrt{\log \mathcal{N}_\infty} \Big/ \left(32\mathcal{B}n^{1/2}\right)\right\} \\
& \overset{(iv)}{\leq} 32\mathcal{B}(n\log \mathcal{N}_\infty)^{1/2}/3 + 64\mathcal{B}(n/\log \mathcal{N}_\infty)^{1/2}/3,
\end{aligned}$$

where (i) follows from the truncation at $t = a_n$ and the bound of the tail probability, (ii) is satisfied when $t \geq 3(n\log \mathcal{N}_\infty)^{1/2}/\mathcal{B}$, (iii) follows from taking the integration, and (iv) is derived by choosing $a_n = 32\mathcal{B}(n\log \mathcal{N}_\infty)^{1/2}/3 \geq 3(n\log \mathcal{N}_\infty)^{1/2}/\mathcal{B}$ as $\mathcal{B} \geq 1$.

The expectation of g_n^2 can be similarly bounded as follows

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n,\mathcal{D}'_n} g_n^2 = & \int_{t>0} \mathbb{P}(g_n^2 \geq t)\mathrm{d}t \\
& \overset{(i)}{\leq} a_n^2 + \int_{a_n^2}^{\infty} 2\mathcal{N}_\infty \exp\left[-t \Big/ \left\{16n + 16\mathcal{B}(tn)^{1/2} / \left(3\sqrt{\log \mathcal{N}_\infty}\right)\right\}\right] \mathrm{d}t \\
& \overset{(ii)}{\leq} a_n^2 + \int_{a_n^2}^{\infty} 2\mathcal{N}_\infty \exp\left\{-3\sqrt{t\log \mathcal{N}_\infty} \Big/ \left(32\mathcal{B}\sqrt{n}\right)\right\} \mathrm{d}t \\
& \overset{(iii)}{\leq} a_n^2 + 4\mathcal{N}_\infty \exp\left\{-\frac{3\sqrt{\log \mathcal{N}_\infty}a_n}{32\mathcal{B}\sqrt{n}}\right\} \left[(32\mathcal{B}a_n\sqrt{n}/\left(3\sqrt{\log \mathcal{N}_\infty}\right) + \{32\mathcal{B}/3\}^2 \times n/\log \mathcal{N}_\infty\right] \\
& \overset{(iv)}{\leq} a_n^2 + 4\left[(32\mathcal{B}a_n\sqrt{n}/\left(3\sqrt{\log \mathcal{N}_\infty}\right) + \{32\mathcal{B}/3\}^2 \times n/\log \mathcal{N}_\infty\right] \\
& \overset{(v)}{\leq} (32\mathcal{B}/3)^2 \times (n\log \mathcal{N}_\infty) + 4(32\mathcal{B}/3)^2 n + 4(32\mathcal{B}/3)^2 \times n/\log \mathcal{N}_\infty \\
& \overset{(vi)}{\leq} 2^{10}\mathcal{B}^2(n\log \mathcal{N}_\infty)/9 + 2^{12}\mathcal{B}^2 n/9 + 2^{12}\mathcal{B}^2 n/(9\log \mathcal{N}_\infty),
\end{aligned}$$

where (i) follows from the truncation at $t = a_n^2$ and the bound of the tail probability, (ii) is satisfied when $\sqrt{t} \ge 3\sqrt{n\log\mathcal{N}_\infty}/\mathcal{B}$, (iii) follows from taking the integration, and (iv), (v), and (vi) are derived by choosing $a_n = 32\mathcal{B}\sqrt{n\log\mathcal{N}_\infty}/3 \ge 3\sqrt{n\log\mathcal{N}_\infty}/\mathcal{B}$ since $\mathcal{B}\ge 1$.

Combining the bounds of $\mathbb{E}_{\mathcal{D}_n,\mathcal{D}_n'} g_n$ and $\mathbb{E}_{\mathcal{D}_n,\mathcal{D}_n'} g_n^2$ with the inequality (S4.1), we obtain that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathbf{X}}(\hat f_n - f_*)^2 - \mathbb{E}_n(\hat f_n - f_*)^2\right\}\\
\le&\frac{1}{n}\mathcal{B}\{32\mathcal{B}(n\log\mathcal{N}_\infty)^{1/2}/3 + 64\mathcal{B}(n/\log\mathcal{N}_\infty)^{1/2}/3\}\left\{(\log\mathcal{N}_\infty/n)^{1/2}+\frac{1}{n}\right\}\\
&+\frac{1}{n}\mathcal{B}\left\{2^{10}\mathcal{B}^2(n\log\mathcal{N}_\infty)/9 + 2^{12}\mathcal{B}^2 n/9 + 2^{12}\mathcal{B}^2 n/(9\log\mathcal{N}_\infty)\right\}^{1/2}\times\left(\mathbb{E}_{\mathcal{D}_n}\|\hat f_n - f_*\|_2^2\right)^{1/2} + \frac{9\mathcal{B}}{n}.
\end{aligned}$$

Since $|a-b|\le 2a^{1/2}c+d$ implies $a-2b\le 2d+4c^2$ for positive numbers a,b,c, and d, we get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathbf{X}}(\hat f_n - f_*)^2 - 2\mathbb{E}_n(\hat f_n - f_*)^2\right\}\\
\le&\frac{2\mathcal{B}^2}{n^2}\left\{2^{10}\mathcal{B}^2(n\log\mathcal{N}_\infty)/9 + 2^{12}\mathcal{B}^2 n/9 + 2^{12}\mathcal{B}^2 n/(9\log\mathcal{N}_\infty)\right\}\\
&+\frac{2}{n}\mathcal{B}\{32\mathcal{B}(n\log\mathcal{N}_\infty)^{1/2}/3 + 64\mathcal{B}(n/\log\mathcal{N}_\infty)^{1/2}/3\}\left\{(\log\mathcal{N}_\infty/n)^{1/2}+\frac{1}{n}\right\}+\frac{18\mathcal{B}}{n}\\
=&\frac{\log\mathcal{N}_\infty}{n}\left(2^{11}\mathcal{B}^4/9+2^6\mathcal{B}^2/3\right)+\frac{1}{n}\left(2^{13}\mathcal{B}^4/9+18\mathcal{B}+2^7\mathcal{B}^2/3\right)+\frac{1}{n\log\mathcal{N}_\infty}\left(2^{13}\mathcal{B}^4/9\right)\\
&+\frac{(\log\mathcal{N}_\infty)^{1/2}}{n^{3/2}}2^6\mathcal{B}^2/3+\frac{1}{n^{3/2}(\log\mathcal{N}_\infty)^{1/2}}2^7\mathcal{B}^2/3\\
\lesssim&\frac{\log\mathcal{N}_\infty}{n},
\end{aligned}$$

where the last inequality holds as $\mathcal{N}_\infty\ge 3$. This completes the proof of Lemma 3.

S5 Proof of Lemma 4

We prove Lemma 4 by five steps. In the first step, we decompose the term T_2 into three terms. In the second, third, and fourth steps, we bound these three terms respectively. Then we obtain a bound of the term T_2 by combining these three bounds in the last step.

Step 1: Error decomposition. Since ε_i is not centered as we discussed before, we decompose T_2 as given below

$$
\begin{aligned}
T_2 =& \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\{\varepsilon_i(\hat{f}_n - f_n)\}\right] \\
=& \underbrace{\mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_n\left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)\right]\right)}_{=:T_4} + \underbrace{\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\left\{\mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)(\hat{f}_n - f_n)\right\}\right]}_{=:T_5}.
\end{aligned}
$$

For the term T_4, if we directly apply the functional Bernstein inequality (Wainwright, 2019, Theorem 3.27) to it, the tail probability bound will degenerate to one as the bandwidth b goes to zero. As an alternative, we truncate ε_i at a constant $C_\varepsilon > \mathcal{B}$. We further decompose the term T_4 as below

$$
\begin{aligned}
T_4 =& \underbrace{\mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_n\left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i| \geq C_\varepsilon)}\right]\right)}_{=:T_6} \\
&+ \underbrace{\mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_n\left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i| < C_\varepsilon)}\right]\right)}_{=:T_7}.
\end{aligned}
$$

To bound the term T_2, it suffices to bound the terms T_5, T_6, and T_7, respectively.

Step 2: Bounding the term T_5. The term T_5 can be bounded as the following

$$
\begin{aligned}
T_5 =& \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\{\mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)(\hat{f}_n - f_n)\}\right] \\
\overset{(i)}{\leq}& \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\left|\mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)(\hat{f}_n - f_*)\right|\right] + \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\left|\mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)(f_* - f_n)\right|\right] \\
\overset{(ii)}{\leq}& 3\left\{\mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\right\}^2 + \mathbb{E}_{\mathbf{X}}(f_n - f_*)^2/4 + \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_n(\hat{f}_n - f_*)^2\right\}/8 \\
\overset{(iii)}{=}& 3C_K^2 b^{2\beta} + \mathbb{E}_{\mathbf{X}}(f_n - f_*)^2/4 + \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_n(\hat{f}_n - f_*)^2\right\}/8, \qquad \text{(S5.1)}
\end{aligned}
$$

where (i) is based on the triangular inequality, (ii) is due to $ab \leq a^2 + b^2/4$ and $ab \leq 2a^2 + b^2/8$, and (iii) is derived from Lemma 1. The remaining term $\mathbb{E}_{\mathbf{X}}(f_n - f_*)^2/4$ is the approximation error which can be incorporated into the term T_3 in Lemma 2, and the term $\mathbb{E}_{\mathcal{D}_n}\big\{\mathbb{E}_n(\hat{f}_n - f_*)^2\big\}/8$ is the empirical excess risk that can be incorporated into the term T_1 in Lemma 2.

Step 3: Bounding the term T_6. We first consider bounding the tail probability as follows

$$\begin{aligned}
\mathbb{E}\left\{1_{(|\varepsilon_i| \geq C_\varepsilon)}\right\} =& \mathbb{P}\left(|\varepsilon_i| \geq C_\varepsilon\right) \\
\overset{(i)}{=}& \mathbb{P}\{K_b(Y_i - y) \geq C_\varepsilon + f_*(y \mid \mathbf{X}_i)\} \\
\overset{(ii)}{\leq}& \mathbb{P}\left[K\{(Y_i - y)/b\} \geq bC_\varepsilon\right] \\
\overset{(iii)}{\leq}& \mathbb{P}[C_e \exp\{-|Y_i - y|/b\} \geq bC_\varepsilon] \\
=& \ \mathbb{P}\left[|Y_i - y| \leq b \log\{C_e/(bC_\varepsilon)\}\right] \\
\overset{(iv)}{\leq}& 2\mathcal{B}b \log\{C_e/(bC_\varepsilon)\},
\end{aligned}$$

where (i) is derived by noting $0 \leq f_* \leq \mathcal{B}$ and $C_\varepsilon > \mathcal{B}$, (ii) is due to the definition of the kernel function, (iii) follows from Assumption 1-(iii) with the constant $C_e > 0$, and (iv) is due to the density function $f_*(y \mid \mathbf{X}_i) \in [0, \mathcal{B}]$. Thus, the term T_6 is bounded by

$$\begin{aligned}
T_6 =& \mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_n\left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i| \geq C_\varepsilon)}\right]\right) \\
\overset{(i)}{\leq}& \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\big\{\big|\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\big| \cdot \big|\hat{f}_n - f_n\big|1_{(|\varepsilon_i| \geq C_\varepsilon)}\big\}\right] \\
\overset{(ii)}{\leq}& 2\mathcal{B}\mathbb{E}_{\mathcal{D}_n}\left\{\big|\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\big|1_{(|\varepsilon_i| \geq C_\varepsilon)}\right\}, \\
\overset{(iii)}{\leq}& C_t \log(1/b) b, \qquad \text{(S5.2)}
\end{aligned}$$

where (i) follows from taking the absolute value and (ii) is due to $|\hat{f}_n| \leq \mathcal{B}$ and $|f_n| \leq \mathcal{B}$.

Since the expectation of ε_i and the tail probability $\mathbb{P}(|\varepsilon_i| \geq C_\varepsilon)$ are bounded, we get (iii) with the constant C_t.

Step 4: Bounding the term T_7. Let us define $v_j = (\log \mathcal{N}_\infty / n)^{1/2} \vee \|f_j - f_n\|_2$ and $h_{i,j} = \{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(f_j - f_n)1_{(|\varepsilon_i|<C_\varepsilon)}/v_j$, and let $h_n = \max_j |\sum_i h_{i,j}|$. By the definition of covering numbers, there exists a $j^* \in \{1, \ldots, \mathcal{N}_\infty\}$ such that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i|<C_\varepsilon)}\Big] \\
\leq &\frac{2C_\varepsilon}{n} + \frac{1}{n}\sum_{i=1}^{n}\big|\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(f_{j^*} - f_n)1_{(|\varepsilon_i|<C_\varepsilon)}\big|.
\end{aligned}$$

The term T_7 can be bounded by

$$\begin{aligned}
T_7 =&\mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_n\Big[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i|<C_\varepsilon)}\Big]\right) \\
\leq &\frac{2C_\varepsilon}{n} + \mathbb{E}_{\mathcal{D}_n}\left(\sum\nolimits_i |h_{i,j^*}| v_{j^*}/n\right) \\
\overset{(i)}{\leq} &\frac{2C_\varepsilon}{n} + \mathbb{E}_{\mathcal{D}_n}\frac{h_n v_{j^*}}{n} \\
\overset{(ii)}{\leq} &\frac{2C_\varepsilon}{n} + \mathbb{E}_{\mathcal{D}_n}\left[\frac{h_n}{n}\left\{\|\hat{f}_n - f_*\|_2 + (\log \mathcal{N}_\infty/n)^{1/2} + \frac{1}{n}\right\}\right] \\
\overset{(iii)}{\leq} &\frac{2C_\varepsilon}{n} + \mathbb{E}_{\mathcal{D}_n}\left[\frac{h_n}{n}\left\{(\log \mathcal{N}_\infty/n)^{1/2} + \frac{1}{n}\right\}\right] + \frac{1}{n}\left\{\mathbb{E}_{\mathcal{D}_n}(h_n^2)\right\}^{1/2} \times \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2},
\end{aligned}$$

where (i) follows from the definition of h_n, (ii) is due to the definition of v_{j^*}, and (iii) follows from the Cauchy-Schwarz inequality.

According to the definition of $h_{i,j}$, it holds that $|h_{i,j}| \leq 4\mathcal{B}C_\varepsilon/v_j$. Now we consider the variance of $h_{i,j}$ as follows

$$\begin{aligned}\operatorname{var}(h_{i,j}) =&\operatorname{var}\left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(f_j - f_n)1_{(|\varepsilon_i|<C_\varepsilon)}/v_j\right]\\ \le& 4C_\varepsilon^2\mathbb{E}\{(f_j - f_n)^2\}/\|f_j - f_n\|_2^2 = 4C_\varepsilon^2.\end{aligned}$$

Since ε_i is truncated and h_n is bounded by the definition of v_j, it is ready to apply the Bernstein inequality in Lemma 1, and the tail probability satisfies

$$\begin{aligned}\mathbb{P}(h_n > t) =&\mathbb{P}\left(\max_j\left|\sum_{i=1}^n \left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(f_j - f_n)1_{(|\varepsilon_i|<C_\varepsilon)}\right]/v_j\right| > t\right)\\ \overset{(i)}{\le}& 2\mathcal{N}_\infty \exp\left[-t^2/\left\{8C_\varepsilon^2 n + 8C_\varepsilon\mathcal{B}t/(3v_j)\right\}\right]\\ \overset{(ii)}{\le}& 2\mathcal{N}_\infty \exp\left[-t^2\Big/\left\{8C_\varepsilon^2 n + 8C_\varepsilon\mathcal{B}t\sqrt{n}\Big/\left(3\sqrt{\log\mathcal{N}_\infty}\right)\right\}\right],\end{aligned}$$

where (i) follows from the union bound argument and (ii) is due to the definition that $v_j \ge (\log\mathcal{N}_\infty/n)^{1/2}$. Consequently, the expectation of h_n can be bounded as follows

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}(h_n) \overset{(i)}{\le}& \int_{t>0}\mathbb{P}\left(\max_j\left|\sum_{i=1}^n\left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i|<C)}\right]/v_{j^*}\right| > t\right)\mathrm{d}t\\ \overset{(ii)}{\le}& a_n + \int_{a_n}^\infty 2\mathcal{N}_\infty\exp\left[-t^2\Big/\left\{8C_\varepsilon^2 n + 8C_\varepsilon\mathcal{B}t\sqrt{n}\Big/\left(3\sqrt{\log\mathcal{N}_\infty}\right)\right\}\right]\mathrm{d}t\\ \overset{(iii)}{\le}& a_n + \int_{a_n}^\infty 2\mathcal{N}_\infty\exp\{-3t(\log\mathcal{N}_\infty/n)^{1/2}/(16C_\varepsilon\mathcal{B})\}\mathrm{d}t\\ \overset{(iv)}{\le}& a_n + \frac{32C_\varepsilon\mathcal{B}}{3(\log\mathcal{N}_\infty/n)^{1/2}}\mathcal{N}_\infty\exp\left\{-3a_n\left(\log\mathcal{N}_\infty/n\right)^{1/2}/(16C_\varepsilon\mathcal{B})\right\}\\ \overset{(v)}{\le}& 16C_\varepsilon\mathcal{B}(n\log\mathcal{N}_\infty)^{1/2}/3 + 32C_\varepsilon\mathcal{B}(n/\log\mathcal{N}_\infty)^{1/2}/3,\end{aligned}$$

where (i) is due to the property of the expectation, (ii) follows from the truncation at $t = a_n$ and the tail probability of Bernstein inequality, (iii) is satisfied when $t \ge 3C_\varepsilon(n\log\mathcal{N}_\infty)^{1/2}/\mathcal{B}$,

(iv) follows from an integration, and (v) is derived from taking $a_n = 16C_\varepsilon\mathcal{B}(n\log\mathcal{N}_\infty)^{1/2}/3 \geq 3C_\varepsilon(n\log\mathcal{N}_\infty)^{1/2}/\mathcal{B}$.

Similarly, the expectation of h_n^2 can be bounded as follows

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}(h_n^2) &\overset{(i)}{\leq} \int_{t>0} \mathbb{P}\left(\max_j \left|\sum_{i=1}^n \left[\{\varepsilon_i - \mathbb{E}_{\mathcal{D}_n}(\varepsilon_i \mid \mathbf{X}_i)\}(\hat{f}_n - f_n)1_{(|\varepsilon_i|<C)}\right]/v_{j^*}\right| > t^{1/2}\right) \mathrm{d}t \\
&\overset{(ii)}{\leq} a_n^2 + \int_{a_n^2}^{\infty} 2\mathcal{N}_\infty \exp\left[-t\Big/\left\{8C_\varepsilon^2 n + 8C_\varepsilon\mathcal{B}(tn)^{1/2}\Big/\left(3\sqrt{\log\mathcal{N}_\infty}\right)\right\}\right] \mathrm{d}t \\
&\overset{(iii)}{\leq} a_n^2 + \int_{a_n^2}^{\infty} 2\mathcal{N}_\infty \exp\{-3t^{1/2}(\log\mathcal{N}_\infty/n)^{1/2}/(16C_\varepsilon\mathcal{B})\}\mathrm{d}t \\
&\overset{(iv)}{\leq} a_n^2 + 4\mathcal{N}_\infty \exp\left\{\frac{-3a_n(\log\mathcal{N}_\infty/n)^{1/2}}{16C_\varepsilon\mathcal{B}}\right\}\left(\frac{16C_\varepsilon\mathcal{B}a_n}{3(\log\mathcal{N}_\infty/n)^{1/2}} + \frac{2^8C_\varepsilon^2\mathcal{B}^2n}{9\log\mathcal{N}_\infty}\right) \\
&\overset{(v)}{\leq} \frac{2^8C_\varepsilon^2\mathcal{B}^2n\log\mathcal{N}_\infty}{9} + \frac{2^{10}C_\varepsilon^2\mathcal{B}^2n}{9} + \frac{2^{10}C_\varepsilon^2\mathcal{B}^2n}{9\log\mathcal{N}_\infty},
\end{aligned}
$$

where (i) is due to the property of the expectation, (ii) follows from the Bernstein inequality, (iii) is satisfied by choosing $\sqrt{t} \geq 3C_\varepsilon(n\log\mathcal{N}_\infty)^{1/2}/\mathcal{B}$, (iv) follows from an integration, and (v) is derived from taking $a_n = 16C_\varepsilon\mathcal{B}(n\log\mathcal{N}_\infty)^{1/2}/3 \geq 3C_\varepsilon(n\log\mathcal{N}_\infty)^{1/2}/\mathcal{B}$.

Combining the results, we can get

$$
\begin{aligned}
T_7 \leq& \frac{2C_\varepsilon}{n} + \mathbb{E}_{\mathcal{D}_n}\left[\frac{h_n}{n}\left\{(\log\mathcal{N}_\infty/n)^{1/2} + \frac{1}{n}\right\}\right] + \frac{1}{n}\left\{\mathbb{E}_{\mathcal{D}_n}\left(h_n^2\right)\right\}^{1/2} \times \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2} \\
\leq& \frac{2C_\varepsilon}{n} + \frac{16C_\varepsilon\mathcal{B}\log\mathcal{N}_\infty}{3n} + \frac{32C_\varepsilon\mathcal{B}}{3n} + \frac{8C_\varepsilon\mathcal{B}(\log\mathcal{N}_\infty)^{1/2}}{3n^{3/2}} + \frac{16C_\varepsilon\mathcal{B}}{3n^{3/2}\sqrt{\log\mathcal{N}_\infty}} \\
&+ \frac{1}{n}\left(\frac{2^8C_\varepsilon^2\mathcal{B}^2n\log\mathcal{N}_\infty}{9} + \frac{2^{10}C_\varepsilon^2\mathcal{B}^2n}{9} + \frac{2^{10}C_\varepsilon^2\mathcal{B}^2n}{9\log\mathcal{N}_\infty}\right)^{1/2} \times \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2}. \quad \text{(S5.3)}
\end{aligned}
$$

Step 5: Combining error bounds. Combining the bounds of T_5, T_6, and T_7 given by

(S5.1), (S5.2), and (S5.3), we can bound the term T_2 as follows

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_n\{\varepsilon_i(\hat{f}_n - f_n)\}\right] \\
\leq & 3C_K^2 b^{2\beta} + \mathbb{E}_{\mathbf{X}}(f_n - f_*)^2/4 + \mathbb{E}_{\mathcal{D}_n}\big\{\mathbb{E}_n(\hat{f}_n - f_*)^2\big\}/8 \\
& + C_t \log(1/b) b + \frac{2C_\varepsilon}{n} + \frac{16 C_\varepsilon \mathcal{B} \log \mathcal{N}_\infty}{3n} + \frac{32 C_\varepsilon \mathcal{B}}{3n} + \frac{8 C_\varepsilon \mathcal{B} (\log \mathcal{N}_\infty)^{1/2}}{3n^{3/2}} + \frac{16 C_\varepsilon \mathcal{B}}{3n^{3/2}\sqrt{\log \mathcal{N}_\infty}} \\
& + \frac{1}{n}\left(\frac{2^8 C_\varepsilon^2 \mathcal{B}^2 n \log \mathcal{N}_\infty}{9} + \frac{2^{10} C_\varepsilon^2 \mathcal{B}^2 n}{9} + \frac{2^{10} C_\varepsilon^2 \mathcal{B}^2 n}{9 \log \mathcal{N}_\infty}\right)^{1/2} \times \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2} \\
\lesssim & b^{2\beta} + \log(1/b) b + \frac{1}{4}\|f_n - f_*\|^2 + \frac{1}{8}\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_n(\hat{f}_n - f_*)^2\right\} \\
& + \frac{\log \mathcal{N}_\infty}{n} + \left(\frac{\log \mathcal{N}_\infty}{n}\right)^{1/2}\left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2}.
\end{aligned}
$$

S6 Proof of Theorem 1

Combining all the terms together, we can get the excess risk

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2 \\
\leq & T_1 + 4T_2 + 2T_3 \\
\leq & \frac{\log \mathcal{N}_\infty}{n}\left(2^{11}\mathcal{B}^4/9 + 2^6 \mathcal{B}^2/3\right) + \frac{1}{n}\left(2^{12}\mathcal{B}^4/9 + 18\mathcal{B} + 2^7\mathcal{B}^2/3\right) + \frac{1}{n \log \mathcal{N}_\infty}\left(2^{12}\mathcal{B}^4/9\right) \\
& + \frac{(\log \mathcal{N}_\infty)^{1/2}}{n^{3/2}} 2^6 \mathcal{B}^2/3 + \frac{1}{n^{3/2}(\log \mathcal{N}_\infty)^{1/2}} 2^7 \mathcal{B}^2/3 \\
& + 12 C_K^2 b^{2\beta} + \mathbb{E}_{\mathbf{X}}(f_n - f_*)^2 + \mathbb{E}_{\mathcal{D}_n}\big\{\mathbb{E}_n(\hat{f}_n - f_*)^2\big\}/2 \\
& + 4C_t \log(1/b) b + \frac{8C_\varepsilon}{n} + \frac{2^6 C_\varepsilon \mathcal{B} \log \mathcal{N}_\infty}{3n} + \frac{2^7 C_\varepsilon \mathcal{B}}{3n} + \frac{2^5 C_\varepsilon \mathcal{B}(\log \mathcal{N}_\infty)^{1/2}}{3n^{3/2}} + \frac{2^6 C_\varepsilon \mathcal{B}}{3n^{3/2}\sqrt{\log \mathcal{N}_\infty}} \\
& + \frac{4}{n}\left(\frac{2^8 C_\varepsilon^2 \mathcal{B}^2 n \log \mathcal{N}_\infty}{9} + \frac{2^{10} C_\varepsilon^2 \mathcal{B}^2 n}{9} + \frac{2^{10} C_\varepsilon^2 \mathcal{B}^2 n}{9 \log \mathcal{N}_\infty}\right)^{1/2} \times \left(\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\right)^{1/2} + 2T_3.
\end{aligned}
$$

Since $|a - b| \leq 2a^{1/2}c + d$ implies $a - 2b \leq 2d + 4c^2$ for positive numbers a, b, c, d, we match each term in the above inequality as follows

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\\
\leq&\frac{\log\mathcal{N}_\infty}{n}\left(2^{12}\mathcal{B}^4/9 + 2^7\mathcal{B}^2/3\right) + \frac{1}{n}\left(2^{13}\mathcal{B}^4/9 + 36\mathcal{B} + 2^8\mathcal{B}^2/3\right) + \frac{1}{n\log\mathcal{N}_\infty}\left(2^{13}\mathcal{B}^4/9\right)\\
&+\frac{(\log\mathcal{N}_\infty)^{1/2}}{n^{3/2}}2^7\mathcal{B}^2/3 + \frac{1}{n^{3/2}(\log\mathcal{N}_\infty)^{1/2}}2^8\mathcal{B}^2/3 + 24C_K^2 b^{2\beta} + \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_n(\hat{f}_n - f_*)^2\right\}\\
&+2^3C_t\log(1/b)b + \frac{2^4C_\varepsilon}{n} + \frac{2^7C_\varepsilon\mathcal{B}\log\mathcal{N}_\infty}{3n} + \frac{2^8C_\varepsilon\mathcal{B}}{3n} + \frac{2^6C_\varepsilon\mathcal{B}(\log\mathcal{N}_\infty)^{1/2}}{3n^{3/2}} + \frac{2^7C_\varepsilon\mathcal{B}}{3n^{3/2}\sqrt{\log\mathcal{N}_\infty}}\\
&+\frac{2^6}{n^2}\left(\frac{2^8C_\varepsilon^2\mathcal{B}^2n\log\mathcal{N}_\infty}{9} + \frac{2^{10}C_\varepsilon^2\mathcal{B}^2n}{9} + \frac{2^{10}C_\varepsilon^2\mathcal{B}^2n}{9\log\mathcal{N}_\infty}\right) + 6T_3.
\end{aligned}$$

Incorporating the term $\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_n(\hat{f} - f_*)^2\right\}$ at right by Lemma 3, we can get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2\\
\leq&\frac{\log\mathcal{N}_\infty}{n}\left(2^{12}\mathcal{B}^4/3 + 2^7\mathcal{B}^2\right) + \frac{1}{n}\left(2^{13}\mathcal{B}^4/3 + 108\mathcal{B} + 2^8\mathcal{B}^2\right) + \frac{1}{n\log\mathcal{N}_\infty}\left(2^{13}\mathcal{B}^4/3\right)\\
&+\frac{(\log\mathcal{N}_\infty)^{1/2}}{n^{3/2}}2^7\mathcal{B}^2 + \frac{1}{n^{3/2}(\log\mathcal{N}_\infty)^{1/2}}2^8\mathcal{B}^2 + 48C_K^2 b^{2\beta}\\
&+2^4C_t\log(1/b)b + \frac{2^5C_\varepsilon}{n} + \frac{2^8C_\varepsilon\mathcal{B}\log\mathcal{N}_\infty}{3n} + \frac{2^9C_\varepsilon\mathcal{B}}{3n} + \frac{2^7C_\varepsilon\mathcal{B}(\log\mathcal{N}_\infty)^{1/2}}{3n^{3/2}} + \frac{2^8C_\varepsilon\mathcal{B}}{3n^{3/2}\sqrt{\log\mathcal{N}_\infty}}\\
&+\frac{1}{n}\left(\frac{2^{15}C_\varepsilon^2\mathcal{B}^2\log\mathcal{N}_\infty}{9} + \frac{2^{17}C_\varepsilon^2\mathcal{B}^2}{9} + \frac{2^{17}C_\varepsilon^2\mathcal{B}^2}{9\log\mathcal{N}_\infty}\right) + 12T_3\\
\lesssim&\frac{\log\mathcal{N}_\infty}{n} + b^{2\beta} + b\log(1/b) + T_3.
\end{aligned}$$

Then we complete the proof of Theorem 1.

S7 The VC-dimension and the pseudodimension for neural networks

We first provide the formal definition of the VC-dimension and pseudodimension.

Definition 1. (Shattering,VC-dimension, Györfi et al., 2002, Definitions 9.5, 9.6) Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$, and let $n \in \mathbb{N}$. For $\mathcal{D}_n^* = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, where $\mathbf{z}_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$, define $s(\mathcal{A}, \mathcal{D}_n^*) = |\{A \cap \mathcal{D}_n^* : A \in \mathcal{A}\}|$. The n-th shatter coefficient of $\mathcal{A}$ is defined as

$$S(\mathcal{A}, n) = \max_{\mathcal{D}_n^*} s(\mathcal{A}, \mathcal{D}_n^*). \tag{S7.1}$$

Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$ with $\mathcal{A} \neq \emptyset$. The Vapnik–Chervonenkis dimension (VC-dimension) of $\mathcal{A}$ is defined by

$$\text{VCdim}(\mathcal{A}) = \sup\{n \in \mathbb{N} : S(\mathcal{A}, n) = 2^n\}. \tag{S7.2}$$

Definition 2. (Pseudodimension, Bartlett et al., 2019, Definition 2) Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The pseudodimension of $\mathcal{F}$, denoted as $\text{Pdim}(\mathcal{F})$, is the largest integer m for which there exist $\{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ in $\mathcal{X}^m$ and $\{a_1, \ldots, a_m\}$ in $\mathbb{R}^m$ such that for any $(b_1, ..., b_m) \in \{0, 1\}^m$, there exists $f \in \mathcal{F}$ such that

$$\forall i \in \{1, \ldots, m\}, f(\mathbf{z}_i) > a_i \Longleftrightarrow b_i = 1.$$

The VC-dimension and pseudodimension are both defined to characterize the complexity of a functional class. In our context, we specify the functional class $\mathcal{F}_n$ as the deep ReLU neural network functional class and aim to bound the covering numbers $\sup_{\mathcal{D}_n} \mathcal{N}_\infty(1/n, \mathcal{F}_n, \mathcal{D}_n)$. Following Bartlett et al. (2019), for a functional class $\mathcal{F}$, we define $\text{VCdim}(\mathcal{F}) = \text{VCdim}\{\text{sgn}(\mathcal{F})\}$, where

$$\text{sgn}(\mathcal{F}) = \{\text{sgn}(f) : f \in \mathcal{F}\}, \tag{S7.3}$$

and $\text{sgn}(x)$ is an indicator function of whether x is positive or not. It is known that

$\mathrm{VCdim}(\mathcal{F}) \leq \mathrm{Pdim}(\mathcal{F})$ (Anthony and Bartlett, 1999, Theorem 14.1). Moreover, if $\mathcal{F}$ is a neural network functional class, we can identify another functional class $\mathcal{F}'$ with one more parameter and one more layer, such that $\mathrm{Pdim}(\mathcal{F}) \leq \mathrm{VCdim}(\mathcal{F}')$. Thus, we can conclude that $\mathrm{Pdim}(\mathcal{F}) \asymp \mathrm{VCdim}(\mathcal{F})$. We aim to obtain a bound for the covering number $\sup_{\mathcal{D}_n} \mathcal{N}_\infty(1/n, \mathcal{F}_n, \mathcal{D}_n)$. By Theorem 12.2 in Anthony and Bartlett (1999) and the VC-dimension bounds, the VC-dimension satisfies

$$\mathcal{S}\mathcal{T}\log(\mathcal{S}/\mathcal{T}) \lesssim \mathrm{VCdim}(\mathcal{F}_n) \lesssim \mathcal{S}\mathcal{T}\log(\mathcal{S}).$$

According to Bartlett et al. (2019), for $n \geq \mathrm{Pdim}(\mathcal{F})$,

$$\begin{aligned} \log\left\{\sup_{\mathcal{D}_n}\mathcal{N}_\infty(1/n, \mathcal{F}_n, \mathcal{D}_n)\right\}/n \leq & \mathrm{Pdim}(\mathcal{F})\log\{en^2\mathcal{B}_0/\mathrm{Pdim}(\mathcal{F})\}/n \\ \lesssim & \mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n)/n. \end{aligned} \tag{S7.4}$$

This inequality implies Corollary 1.

S8 Proof of Corollary 2

Given the error bound in Corollary 1 and approximation error in Lemma 5, to get the consistency, we let $\mathcal{S} \to \infty$, and $\mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n)/n \to 0$ as $n \to \infty$.

To get the optimal convergence rate, we need to balance the stochastic error term and the approximation error term in

$$\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2 \lesssim \frac{1}{n}\mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n) + n^{-2\beta/(2\beta+d)}\log n + (NM)^{-4\beta/d}.$$

From Lemma 5, the network width and depth satisfy $\mathcal{W} = O(N \log N)$ and $\mathcal{T} = O(M \log M)$. Moreover, for a feedforward network, the size satisfies $\mathcal{S} = O(\mathcal{W}^2 \mathcal{T})$. Therefore,

$$\begin{aligned}\frac{1}{n}\mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n) &\lesssim \frac{1}{n}\mathcal{W}^2\mathcal{T}^2\log(\mathcal{S})\log(n)\\ &\lesssim \frac{1}{n}N^2M^2\log^4(n),\end{aligned}$$

where we used $\log N \lesssim \log n$ and $\log M \lesssim \log n$ for polynomial choices of N and M. Thus, to attain the minimax rate (up to logarithmic factors), it is sufficient to choose (N, M) so that

$$\frac{(NM)^2}{n} \asymp n^{-2\beta/(2\beta+d)} \quad \text{and} \quad (NM)^{-4\beta/d} \asymp n^{-2\beta/(2\beta+d)},$$

which is achieved by taking $NM \asymp n^{d/(4\beta+2d)}$. We provide two illustrative specifications corresponding to different width/depth scalings.

Fixed depth, increasing width. Set M fixed and $N \asymp n^{d/(4\beta+2d)}$. Then Lemma 5 implies $\mathcal{W} = O(n^{d/(4\beta+2d)} \log n)$ and $\mathcal{T} = O(1)$, and thus $\mathcal{S} = O(\mathcal{W}^2\mathcal{T}) = O(n^{d/(2\beta+d)} \log^2 n)$. The first term of (4.3) becomes

$$\mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n)/n \lesssim n^{d/(2\beta+d)-1}\log^4(n) = n^{-2\beta/(2\beta+d)}\log^4(n).$$

Fixed width, increasing depth. Alternatively, set N fixed and $M \asymp n^{d/(4\beta+2d)}$ (so that still $NM \asymp n^{d/(4\beta+2d)}$). Then $\mathcal{W} = O(1)$ while $\mathcal{T} = O(M \log M) = O(n^{d/(4\beta+2d)} \log n)$, and hence $\mathcal{S} = O(\mathcal{W}^2\mathcal{T}) = O(n^{d/(4\beta+2d)}\log n)$. Therefore,

$$\begin{aligned}\mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n)/n &\lesssim \{n^{d/(4\beta+2d)}\log n\}\{n^{d/(4\beta+2d)}\log n\}\log n\log n/n\\ &\lesssim n^{d/(2\beta+d)-1}\log^4(n) = n^{-2\beta/(2\beta+d)}\log^4(n).\end{aligned}$$

In both cases, the remaining two terms in (4.3) can be controlled by $n^{-2\beta/(2\beta+d)}$ (up to logarithmic factors), which yields (4.4). Motivated by the common viewpoint that depth can be more important than width for ReLU networks (e.g., Vardi et al. (2022)), we present the increasing-depth / bounded-width regime in Corollary 2 as the main-text illustration.

S9 Proof of Corollary 3

According to Jiao et al. (2023), under Assumption 3, by applying Theorem 6.2 in Jiao et al. (2023) and Corollary 1, we obtain that, given N, M, and the bandwidth $b = n^{-1/(2\beta+d_{\mathcal{M}})} \wedge n^{-2\beta/(2\beta+d_{\mathcal{M}})}$, the excess risk can be bounded by

$$\mathbb{E}_{\mathcal{D}_n}\|\hat{f}_n - f_*\|_2^2 \lesssim \mathcal{S}\mathcal{T}\log(\mathcal{S})\log(n)/n + n^{-2\beta/(2\beta+d_{\mathcal{M}})}\log n + \lceil\beta\rceil^4 d d_{\mathcal{M}}^{3\lceil\beta\rceil-2}(NM)^{-4\beta/d_{\mathcal{M}}}.$$

By balancing the stochastic error and approximation error, we further establish a bound of order $n^{-2\beta/(2\beta+d_{\mathcal{M}})}\log^4 n$.

S10 Convergence Rate of the NW Estimator for Conditional Density Estimation

In this section, we provide a self-contained derivation showing that the Nadaraya–Watson (NW) type kernel estimator for conditional density estimation can achieve the minimax optimal rate when appropriate higher-order kernels are employed.

Let $K(\cdot)$ and $L(\cdot)$ be kernel functions and b_y, b_x be bandwidths for y and $\mathbf{x}$, respectively. The NW estimator can be written as the ratio of a joint density estimator and a marginal density estimator:

$$\hat{f}_{NW}(y \mid \mathbf{x}) := \frac{\hat{f}_{XY}(\mathbf{x}, y)}{\hat{f}_X(\mathbf{x})} = \frac{\sum_{i=1}^n K_{b_y}(Y_i - y)\, L_{b_x}(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n L_{b_x}(\mathbf{X}_i - \mathbf{x})},$$

where $\hat{f}_{XY}(\mathbf{x}, y) = n^{-1}\sum_{i=1}^n K_{b_y}(Y_i - y)\, L_{b_x}(\mathbf{X}_i - \mathbf{x})$ and $\hat{f}_X(\mathbf{x}) = n^{-1}\sum_{i=1}^n L_{b_x}(\mathbf{X}_i - \mathbf{x})$, with $K_{b_y}(u) = K(u/b_y)/b_y$ and $L_{b_x}(\mathbf{v}) = L(\mathbf{v}/b_x)/b_x^d$.

Assume that $f_*(y \mid \mathbf{x})$ belongs to the β-Hölder class on $[0,1]^{d+1}$, the kernels K and L are of order β (i.e., $\int u^j K(u)\, \mathrm{d}u = 0$ for $j = 1, \dots, \lfloor \beta \rfloor - 1$ and similarly for L), and $f_X(\mathbf{x})$ is bounded away from zero on $[0,1]^d$.

Using a standard change-of-variables and Taylor expansion argument (similar to the proof of Lemma 1, but applied to both $\mathbf{x}$ and y), the kernel estimators satisfy:

$$\mathbb{E}\{\hat{f}_{XY}(\mathbf{x}, y)\} = f_{XY}(\mathbf{x}, y) + O(b_x^\beta + b_y^\beta), \qquad \mathbb{E}\{\hat{f}_X(\mathbf{x})\} = f_X(\mathbf{x}) + O(b_x^\beta).$$

A Taylor expansion of the ratio $\hat{f}_{XY}(\mathbf{x}, y)/\hat{f}_X(\mathbf{x})$ around its population counterpart $f_{XY}(\mathbf{x}, y)/f_X(\mathbf{x}) = f_*(y \mid \mathbf{x})$ yields a pointwise bias of $O(b_x^\beta + b_y^\beta)$ and thus a squared bias of $O(b_x^{2\beta} + b_y^{2\beta})$.

Moreover, since $\int K_{b_y}^2 = O(1/b_y)$ and $\int L_{b_x}^2 = O(1/b_x^d)$, we have $\mathrm{var}\{\hat{f}_{XY}(\mathbf{x}, y)\} =$

$O\{(nb_x^d b_y)^{-1}\}$. The ratio structure, combined with the assumption that $f_X(\mathbf{x}) \gtrsim 1$, implies $\mathrm{var}\{\hat{f}_{NW}(y \mid \mathbf{x})\} = O\{(nb_x^d b_y)^{-1}\}$.

Consequently, the integrated mean squared error (IMSE) satisfies:

$$\mathbb{E} \iint \left\{\hat{f}_{NW}(y \mid \mathbf{x}) - f_*(y \mid \mathbf{x})\right\}^2 \mathrm{d}y \, \mathrm{d}\mathbf{x} \lesssim b_x^{2\beta} + b_y^{2\beta} + \frac{1}{nb_x^d b_y}.$$

Balancing the bias and variance terms by setting $b_x \asymp b_y \asymp n^{-1/(2\beta+d+1)}$ yields the minimax optimal rate:

$$\text{IMSE} \lesssim n^{-2\beta/(2\beta+d+1)}.$$

This rate matches the minimax lower bound for conditional density estimation when the risk is integrated over both $\mathbf{x}$ and y, where the effective nonparametric dimension is $d+1$. Note that achieving this rate requires kernels of order β; the original analysis by Hyndman et al. (1996) focused on the case $\beta = 2$ with second-order kernels.

Bibliography

Anthony, M. and P. L. Bartlett (1999). *Neural Network Learning: Theoretical Foundations.* Cambridge: Cambridge University Press.

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research 20*(63), 1–17.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. New York, NY: Springer.

Hyndman, R. J., D. M. Bashtannyk, and G. K. Grunwald (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics 5*(4), 315–336.

Jiao, Y., G. Shen, Y. Lin, and J. Huang (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics 51*(2), 691–716.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics 48*(4), 1875–1897.

Vardi, G., G. Yehudai, and O. Shamir (2022). Width is less important than depth in ReLU neural networks. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 1249–1281. PMLR.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Chenxuan He

Institute of Statistics and Big Data, Renmin University of China, Beijing, China

E-mail: hechenxuan@ruc.edu.cn

Yuan Gao

LPMC & KLMDASR, School of Statistics and Data Science, Nankai University, Tianjin, China

E-mail: yuangao@nankai.edu.cn

Liping Zhu

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing, China

E-mail: zhu.liping@ruc.edu.cn

Jian Huang

Department of Data Science and AI and Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

E-mail: j.huang@polyu.edu.hk

Chenxuan He and Yuan Gao are co-first authors.