Supplementary Material for "Deep regression learning with optimal loss function"

Xuancheng Wang*, Ling Zhou* and Huazhen Lin†‡

New Cornerstone Science Laboratory,
Center of Statistical Research and School of Statistics,
Southwestern University of Finance and Economics, Chengdu, China

This supplement contains four sections. Section S.1 presents the notations and conditions needed in the proofs of Theorems 1-2. Section S.2 presents Proposition S.1, the proofs of Theorems 1-2 and Proposition 1. Section S.3 presents related lemmas used in the proof of Theorems and Corollaries. Section S.4 presents some results of simulation studies.

Contents

*Co-first authors.

†Corresponding author. Email address: *linhz@swufe.edu.cn*.

‡The research was supported by National Key R&D Program of China (No.2022YFA1003702), National Natural Science Foundation of China (Nos. 11931014 and 12271441), and New Cornerstone Science Foundation.

S.1 Notations and conditions

S.1.1 Notations

Feedforward neural network. Let $\mathcal{G}$ be a function class consisting of ReLU neural networks, that is, $\mathcal{G} := \mathcal{G}_{\mathcal{D},\mathcal{U},\mathcal{W},\mathcal{S},\mathcal{B}}$, where the input data is the predictor X, forming the first layer, and the output is the last layer of the network; Such a network $\mathcal{G}$ has $\mathcal{D}$ hidden layers and a total of $(\mathcal{D}+2)$ layers. Denote the width of layer j by d_j, $j = 0, \cdots, \mathcal{D}, \mathcal{D}+1$ with $d_0 = d$ representing the dimension of the input X, and $d_{\mathcal{D}+1} = 1$ representing the dimension of the response Y. The width $\mathcal{W}$ is defined as the maximum width among the hidden layers, i.e., $\mathcal{W} = \max(d_1, ..., d_{\mathcal{D}})$. The size $\mathcal{S}$ is defined as the total number of parameters in the network $\mathcal{G}$, given by $\mathcal{S} = \sum_{i=0}^{\mathcal{D}} d_{i+1} \times (d_i + 1)$; The number of neurons $\mathcal{U}$ is defined as the total number of computational units in the hidden layers, given by $\mathcal{U} = \sum_{i=1}^{\mathcal{D}} d_i$. Further, we assume every function $g \in \mathcal{G}$ satisfies $|g|_\infty \leq \mathcal{B}$ with $\mathcal{B}$ being a positive constant.

Covering number. Given a δ-uniform covering of $\mathcal{G}$, we denote the centers of the balls by $g_q, q = 1, \cdots, \mathcal{N}_{2n}$, where $\mathcal{N}_{2n} = \mathcal{N}_{2n}(\delta, \|\cdot\|_\infty, \mathcal{G}|_{\boldsymbol{x}})$ is the uniform covering number with radius δ under the norm $\|\cdot\|_\infty$. By the definition of covering, there exists a q^* such that $\|\hat{g} - g_{q^*}\|_\infty \leq \delta$ on $\boldsymbol{x} \in (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{X}'_1, \cdots, \boldsymbol{X}'_n)$. Let $A \preceq B$ represent $A \leq cB$ for a postive constant c.

The definitions of $\hat{g}$ and $\hat{g}_{oracle}$. For any independent and identically distributed (i.i.d.) samples $D_n = \{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ with the sample size n. With $g \in \mathcal{G}$, given $f(\cdot)$ known, we define the oracle estimator as

$$\hat{g}_{oracle} = \arg\min_{g \in \mathcal{G}} \mathcal{R}_n(g) := \arg\min_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(-\log f(Y_i - g(\boldsymbol{X}_i))\right) \right\}. \tag{S.1}$$

The proposed estimator is defined by

$$\begin{aligned} \tilde{g} &= \hat{g} - \frac{1}{n}\sum_{i=1}^{n} \hat{g}(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n} Y_i, \\ \hat{g} &\in \arg\min_{g \in \mathcal{G}} \hat{\mathcal{R}}_n(g) := \arg\min_{g \in \mathcal{G}} n^{-1} \sum_{i=1}^{n} \left(-\log \hat{f}_g(Y_i - g(\boldsymbol{X}_i))\right) \\ &:= \arg\min_{g \in \mathcal{G}} n^{-1} \sum_{i=1}^{n} \left(-\log \frac{1}{n} \sum_{j=1}^{n} \mathcal{K}_h(Y_j - g(\boldsymbol{X}_j), Y_i - g(\boldsymbol{X}_i))\right), \end{aligned} \tag{S.2}$$

where $\mathcal{K}_h(y_1, y_2) = K(\frac{y_1 - y_2}{h})/h$, h is a bandwidth and $K(\cdot)$ is a kernel function.

The following notations are needed in the proofs of Theorems 1-2. Define $Y_i = g(\boldsymbol{X}_i) + \epsilon_i$ with $\mathbb{E}(\epsilon_i) = 0$, and

$$S(g, \boldsymbol{Z}_i) = -\log f(Y_i - g(\boldsymbol{X}_i)) + \log f(Y_i - g^*(\boldsymbol{X}_i)),$$

for any g and the sample D_n where g^* is defined as

$$g^* := \arg\min_{g} \mathcal{R}(g) = \arg\min_{g} \mathbb{E}\left(-\log f\left(Y_i - g(\boldsymbol{X}_i)\right)\right).$$

Let D_n' be another sample independent of D_n, then the excess risk of $\hat{g}$ takes the following form:

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g^*) = \mathbb{E}_{D_n'}\left(n^{-1}\sum_{i=1}^{n} S(\hat{g}, \boldsymbol{Z}_i')\right),$$

and its expected excess risk is

$$\mathbb{E}\left(\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)\right) = \mathbb{E}_{D_n}\left[\mathbb{E}_{D_n'}\left(n^{-1}\sum_{i=1}^{n} S(\hat{g}, \boldsymbol{Z}_i')\right)\right].$$

In the following, we write

$$\begin{aligned}
L_r(g - g_0) &= \int (g(\boldsymbol{X}) - g_0(\boldsymbol{X}))^r f_{\boldsymbol{x}}(\boldsymbol{X}) d\boldsymbol{X}, \\
L(g, \boldsymbol{Z}_i) &= \mathbb{E}_{D_n'}\{S(g, \boldsymbol{Z}_i')\} - 2S(g, \boldsymbol{Z}_i), \quad \text{for } g \in \mathcal{G}, \\
\tilde{f}_{g_1,h}(g(\boldsymbol{x}) - y) &:= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h\left(g_1(\boldsymbol{X}_i) - Y_i, g(\boldsymbol{x}) - y\right)\right].
\end{aligned}$$

S.1.2 Conditions

Denote $f^{(r)}(\cdot)$ to be the rth derivative of f, and $f_{\boldsymbol{x}}(\cdot)$ to be the density function of covariates $\boldsymbol{X}$, who is supported on a bounded set, and for simplicity, we assume this bounded set to be $[0,1]^d$. In the rest of the paper, the symbol c denotes a positive constant which may vary across different contexts. The following conditions are required for establishing the rate of the excess risk:

(C1) Kernel: Let $U_r = \int K(t)t^r dt$ and $v_r = \int K^2(t)t^r dt$. Assume the kernel function $K(\cdot)$ has a bounded second derivative, $U_0 = 1$ and $U_1 = 0$.

(C2) Bandwidth: $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

(C3) Density function f: (C3a) For any $\zeta > 0$, there exists a $\beta_\zeta > 0$ such that $\mathbb{E}(|\log f(\epsilon) I(f(\epsilon) < \beta_\zeta)|) < \zeta$. For $\zeta = O(n^{-1}\log n)$, there exists a $\beta_\zeta = O(n^{-r})$ for some constant $r \geq 1$ satisfying the above inequality. (C3b) Assume the density function $f(\cdot)$ has a continuous first-order derivative and satisfies $\mathbb{E}(|f^{(1)}/f|) < \mathcal{B}$. (C3c) Assume the density function $f(\cdot)$ has a continuous third-order derivative and satisfies $\mathbb{E}|f^{(r_1)}/f|^{r_2} < \mathcal{B}$ for $r_1 = 1, 2, 3$, $r_2 = 1, 2$.

(C4) Function class for g and g^*: For any function $g \in \mathcal{G}$ and the true function g^*, we assume $\|g\|_\infty < \mathcal{B}$ and $\|g^*\|_\infty < \mathcal{B}$.

Condition (C1) is a mild condition for the kernel function, which is easy to be satisfied when the kernel function is a symmetric density function. Condition (C2) is the most commonly used assumption for the bandwidth. Condition (C3a) requires a tail condition on the log transformation of the density. Conditions (C3b) and (C3c) require bounded moment for the density and its derivatives to avoid tail-related problems. Clearly, when the errors have bounded supports, conditions (C3b) and (C3c) are automatically satisfied. For errors with unbounded supports, Gaussian errors and several sub-Gaussian errors also meet these conditions. An example is a variable whose characteristic function takes the form: $\varphi(t) \propto \exp(-\gamma t^2)\psi_b(t)$, where γ is a positive constant and $\psi_b(t)$ is a polynomial function of t with order $b > 0$. In particular, a variable with the characteristic function $\varphi(t) = \exp(-t^2/2)(1 - \alpha t^2 + \beta t^4)$ for $\alpha \geq \sqrt{2\beta}$ is classified as a strictly sub-Gaussian variable, as defined in Proposition 5.1 in Bobkov et al. (2024) . Condition (C4) is a bounded condition for the function class $\mathcal{G}$ and the true function g^*.

S.2 Propositions and Proofs

S.2.1 Proposition S.1

Proposition S.1. *For any two functions g and g_1 satisfy the following model: $Y = g(\boldsymbol{X}) + \epsilon = g_1(\boldsymbol{X}) + \epsilon_1$, where $\boldsymbol{X}$ and ϵ are independent, and $\boldsymbol{X}$ and ϵ_1 are independent. Denote F_ϵ as the distribution function of ϵ. Then it follows that for any $\boldsymbol{x} \in [0,1]^d$ and $m \in R$,*

$$g_1(\boldsymbol{x}) - g(\boldsymbol{x}) \equiv c, \quad \text{and} \quad F_{\epsilon_1}(m) = F_\epsilon(m - c).$$

for some constant c.

Proof of Proposition S.1. Denote $\delta(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g(\boldsymbol{x})$. Then, we have

$$Y = g_1(\boldsymbol{X}) + \epsilon_1 = g(\boldsymbol{X}) + \left(g_1(\boldsymbol{X}) - g(\boldsymbol{X}) + \epsilon\right).$$

Denote $F_\epsilon(m \mid \boldsymbol{X} = \boldsymbol{x}) := P(\epsilon \le m \mid \boldsymbol{X} = \boldsymbol{x})$ and $F_\epsilon(m) := P(\epsilon \le m)$. By the independent assumption on ϵ and $\boldsymbol{X}$, it follows that

$$\begin{aligned}
&F_{\epsilon_1}(m \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\left[I\left(\epsilon \le m - g_1(\boldsymbol{X}) + g(\boldsymbol{X})\right) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= F_\epsilon(m - \delta(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}) = F_\epsilon(m - \delta(\boldsymbol{x})),
\end{aligned}$$

and

$$F_{\epsilon_1}(m) = \int F_\epsilon(m - \delta(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} = \int F_\epsilon(m - \delta(\boldsymbol{x})) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}.$$

By the independent assumption on ϵ_1 and $\boldsymbol{X}$, it holds that

$$F_{\epsilon_1}(m \mid \boldsymbol{X} = \boldsymbol{x}) = F_{\epsilon_1}(m),$$

which leads to

$$F_\epsilon(m - \delta(\boldsymbol{x})) = \int F_\epsilon(m - \delta(\boldsymbol{x})) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}.$$

Then, it follows that

$$\delta(\boldsymbol{x}) \equiv c.$$

That is, $g(\boldsymbol{x})$ is identifiable up to a constant, and the density of ϵ_1 is the same as ϵ with a constant mean shift.

S.2.2 Proof of Theorem 1

Let $g^*_{\mathcal{G}}$ be the estimator in the function class that $g^*_{\mathcal{G}} = \arg\min_{g\in\mathcal{G}} \mathcal{R}(g)$. By the definition of the empirical risk minimizer, we have

$$\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i - \hat{g}_{oracle}(\boldsymbol{X}_i))\right] \geq \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i - g^*_{\mathcal{G}}(\boldsymbol{X}_i))\right].$$

Then, it follows that

$$\begin{aligned}
&\mathbb{E}\left(\mathcal{R}(\hat{g}_{oracle}) - \mathcal{R}(g^*)\right)\\
&= \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{D'_n}(-\log f(Y'_i - \hat{g}_{oracle}(\boldsymbol{X}'_i))) - \mathbb{E}_{D'_n}(-\log f(Y'_i - g^*(\boldsymbol{X}'_i)))\right]\right\}\\
&= \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{D'_n}\left(n^{-1}\sum_{i=1}^{n} S(\hat{g}_{oracle}, \boldsymbol{Z}'_i)\right) - 2S(\hat{g}_{oracle}, \boldsymbol{Z}_i)\right\}\right]\\
&\quad -2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i - \hat{g}_{oracle}(\boldsymbol{X}_i)) - \log f(Y_i - g^*_{\mathcal{G}}(\boldsymbol{X}_i))\right\}\right]\\
&\quad +2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n} S(g^*_{\mathcal{G}}, \boldsymbol{Z}_i)\right]\\
&\leq I_1 + 2\left(\mathcal{R}(g^*_{\mathcal{G}}) - \mathcal{R}(g^*)\right).
\end{aligned}$$

Next, we will give an upper bound of I_1 and handle it with truncation and classical chaining technique of empirical processes.

Upper bound for I_1.

Write $\hat{g}_o := \hat{g}_{oracle}$. Let $L(g, \boldsymbol{Z}_i) = \mathbb{E}_{D'_n}\{S(g, \boldsymbol{Z}'_i)\} - 2S(g, \boldsymbol{Z}_i)$. According to the definition of $S(g, \boldsymbol{Z}_i)$, we have

$$\left|L(\hat{g}_o, \boldsymbol{Z}_i) - L(g_{q^*}, \boldsymbol{Z}_i)\right| = \left|\mathbb{E}_{D'_n}\{S(\hat{g}_o, \boldsymbol{Z}'_i)\} - 2S(\hat{g}_o, \boldsymbol{Z}_i) - \mathbb{E}_{D'_n}\{S(g_{q^*}, \boldsymbol{Z}'_i)\} + 2S(g_{q^*}, \boldsymbol{Z}_i)\right|,$$

$$\left|S(\hat{g}_o, \boldsymbol{Z}_i) - S(g_{q^*}, \boldsymbol{Z}_i)\right| = \left|\log f(\hat{g}_o(\boldsymbol{X}_i) - Y_i) - \log f(g_{q^*}(\boldsymbol{X}_i) - Y_i)\right|.$$

Then given the definition of g_{q^*}, it follows from

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\{S(\hat{g}_o, \boldsymbol{Z}_i)\} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\{S(g_{q^*}, \boldsymbol{Z}_i)\}\right| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}|S(\hat{g}_o, \boldsymbol{Z}_i) - S(g_{q^*}, \boldsymbol{Z}_i)| \\
=\ & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left|\log f(\hat{g}_o(\boldsymbol{X}_i) - Y_i) - \log f(g_{q^*}(\boldsymbol{X}_i) - Y_i)\right| \\
=\ & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left|\frac{\dot{f}(\epsilon)(\hat{g}_o(\boldsymbol{X}_i) - g_{q^*}(\boldsymbol{X}_i))}{f(\epsilon)}\right| \\
\leq\ & \mathcal{B}\delta,
\end{aligned}$$

where the last inequality follows from Condition (C3b) and

$$\begin{aligned}
& \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|L(\hat{g}_o, \boldsymbol{Z}_i) - L(g_{q^*}, \boldsymbol{Z}_i)\right|\right\} \\
& = \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{D'_n}\{S(\hat{g}_o, \boldsymbol{Z}'_i)\} - 2S(\hat{g}_o, \boldsymbol{Z}_i) - \mathbb{E}_{D'_n}\{S(g_{q^*}, \boldsymbol{Z}'_i)\} + 2S(g_{q^*}, \boldsymbol{Z}_i)\right|\right\} \\
& \leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{D'_n}\{S(\hat{g}_o, \boldsymbol{Z}'_i)\} - \mathbb{E}_{D'_n}\{S(g_{q^*}, \boldsymbol{Z}'_i)\}\right| + 2\left|S(\hat{g}_o, \boldsymbol{Z}_i) - S(g_{q^*}, \boldsymbol{Z}_i)\right|\right\} \\
& \leq 3\mathcal{B}\delta,
\end{aligned}$$

that

$$\mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L(\hat{g}_o, \boldsymbol{Z}_i)\right\} \leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L(g_{q^*}, \boldsymbol{Z}_i)\right\} + 3\mathcal{B}\delta. \tag{S.3}$$

Let $0 < \beta_n$ be a positive number who may depend on the sample size n. Denote $T_{\beta_n} f = f$ if $f \geq \beta_n$ and $T_{\beta_n} f = \beta_n$ otherwise. Define the function $g^*_{\beta_n}$ by

$$g^*_{\beta_n}(\boldsymbol{x}) = \arg \min_{g:|g|_\infty < \mathcal{B}} \mathbb{E}\left(-\log T_{\beta_n} f(Y - g(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\right).$$

For any $g \in \mathcal{G}$, we let $S_{\beta_n}(g, \boldsymbol{Z}_i) = -\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) + \log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)$. Then, we have

$$\begin{aligned}
\mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} &= \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left\{\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) - \log f(g(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log T_{\beta_n} f(g^*(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log f(g^*(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*(\boldsymbol{X}_i) - Y_i)\right\} \\
&\leq \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left|\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) - \log f(g(\boldsymbol{X}_i) - Y_i)\right| \\
&\quad + \mathbb{E}\left|\log f(g^*(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*(\boldsymbol{X}_i) - Y_i)\right| \\
&\leq \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + 2\mathbb{E}\left|(\log \beta_n - \log f(\epsilon)) I(f(\epsilon) < \beta_n)\right| \\
&\leq \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + 4\mathbb{E}\left\{\left|\log f(\epsilon) I(f(\epsilon) < \beta_n)\right|\right\}, \\
\mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} &= \mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left\{\log f(g(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i) - \log f(g^*(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i) - \log f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)\right\} \\
&\leq \mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left|\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) - \log f(g(\boldsymbol{X}_i) - Y_i)\right| \\
&\quad + \mathbb{E}\left|\log f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)\right| \\
&\leq \mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} + 4\mathbb{E}\left\{\left|\log f(\epsilon) I(f(\epsilon) < \beta_n)\right|\right\},
\end{aligned}$$

so

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{S(g,\boldsymbol{Z}_i)\right\}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{S_{\beta_n}(g,\boldsymbol{Z}_i)\right\}\right| \leq 4\mathbb{E}\left\{\Big|\log f(\epsilon)I(f(\epsilon)<\beta_n)\Big|\right\},$$

which lead to

$$\left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(L(g_{q^*},\boldsymbol{Z}_i)-L_{\beta_n}(g_{q^*},\boldsymbol{Z}_i)\right)\right]\right| \leq 12\mathbb{E}\left\{\Big|\log f(\epsilon)I(f(\epsilon)<\beta_n)\Big|\right\}. \tag{S.4}$$

On the other hand, for any $g\in\mathcal{G}$, we have

$$\begin{aligned}
&|S_{\beta_n}(g,\boldsymbol{Z}_i)|\leq 2|\log\beta_n|,\\
&\sigma_S^2(g):=\mathrm{Var}(S_{\beta_n}(g,\boldsymbol{Z}_i))\leq\mathbb{E}\left\{S_{\beta_n}^2(g,\boldsymbol{Z}_i)\right\}\leq 2|\log\beta_n|\mathbb{E}(S_{\beta_n}(g,\boldsymbol{Z}_i)).
\end{aligned}$$

Following the Bernstein inequality, for any $t>0$, let $u=t/2+\sigma_S^2(g)/4|\log\beta_n|$, we have

$$\begin{aligned}
&P\left\{\frac{1}{n}\sum_{i=1}^{n}L_{\beta_n}(g_q,\boldsymbol{Z}_i)>t\right\}\\
&=P\left\{\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q,\boldsymbol{Z}_i')\}-\frac{2}{n}\sum_{i=1}^{n}S_{\beta_n}(g_q,\boldsymbol{Z}_i)>t\right\}\\
&=P\left\{\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q,\boldsymbol{Z}_i')\}-\frac{1}{n}\sum_{i=1}^{n}S_{\beta_n}(g_q,\boldsymbol{Z}_i)>\frac{t}{2}+\frac{1}{2}\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q,\boldsymbol{Z}_i')\}\right\}\\
&\leq P\left\{\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q,\boldsymbol{Z}_i')\}-\frac{1}{n}\sum_{i=1}^{n}S_{\beta_n}(g_q,\boldsymbol{Z}_i)>\frac{t}{2}+\frac{1}{2}\frac{\sigma_S^2(g)}{2|\log\beta_n|}\right\}\\
&\leq\exp(-\frac{nu^2}{2\sigma_S^2(g)+8u|\log\beta_n|/3})\\
&\leq\exp(-\frac{nu^2}{8u|\log\beta_n|+8u|\log\beta_n|/3})\\
&\leq\exp(-\frac{1}{8+8/3}\frac{nu}{|\log\beta_n|})\\
&\leq\exp(-\frac{1}{16+16/3}\frac{nt}{|\log\beta_n|})\\
&=\exp\left(-\frac{Cnt}{|\log\beta_n|}\right),
\end{aligned}$$

This leads to a tail probability bound of $\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)$, that is

$$P\left\{\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i) > t\right\} \leq 2\mathcal{N}_{2n}\exp\left(-\frac{Cnt}{|\log\beta_n|}\right).$$

Then for $a_n > 0$,

$$\begin{aligned}
\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)\right] &\leq a_n + \int_{a_n}^{\infty} P\left\{\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i) > t\right\} dt \\
&\leq a_n + \int_{a_n}^{\infty} 2\mathcal{N}_{2n}\exp\left(-\frac{Cnt}{|\log\beta_n|}\right) dt \\
&\leq a_n + 2\mathcal{N}_{2n}\exp\left(-a_n\frac{Cn}{|\log\beta_n|}\right)\frac{|\log\beta_n|}{Cn}.
\end{aligned}$$

Choosing $a_n = \log 2\mathcal{N}_{2n}\frac{|\log\beta_n|}{Cn}$, the above inequality leads to

$$\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)\right] \leq \frac{C|\log\beta_n|(\log 2\mathcal{N}_{2n}+1)}{n}. \tag{S.5}$$

Combining inequalities (S.3), (S.4), (S.5), we have

$$\begin{aligned}
I_1 &= \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^n L(\hat{g}_o, \boldsymbol{Z}_i)\right\} \\
&\leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^n L(g_{q^*}, \boldsymbol{Z}_i)\right\} + 3\mathcal{B}\delta \\
&\leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)\right\} + 12\mathbb{E}\left\{\left|\log f(\epsilon)I(f(\epsilon) < \beta_n)\right|\right\} + 3\mathcal{B}\delta \\
&\leq \frac{C|\log\beta_n|(\log 2\mathcal{N}_{2n}+1)}{n} + 12\mathbb{E}\left\{\left|\log f(\epsilon)I(f(\epsilon) < \beta_n)\right|\right\} + 3\mathcal{B}\delta.
\end{aligned} \tag{S.6}$$

Let $\beta_n = n^{-r}$. Using the above inequalities and Condition (C3a), we obtain that

$$\mathbb{E}\left(\mathcal{R}(\hat{g}_{oracle}) - \mathcal{R}(g^*)\right) \preceq \frac{\log n(\log 2\mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{G}|_{\boldsymbol{x}}) + 2)}{n} + \big(\mathcal{R}(g^*_{\mathcal{G}}) - \mathcal{R}(g^*)\big).$$

S.2.3 Proof of Theorem 2

Recall that $g^*_{\mathcal{G}}$ is the estimator in the function class that $g^*_{\mathcal{G}} = \arg\min_{g\in\mathcal{G}} \mathcal{R}(g)$ and

$$\hat{f}_g(z) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(Y_j - g(\boldsymbol{X}_j), z). \tag{S.7}$$

Since $\hat{g} \in \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\left(-\log \hat{f}_g(Y_i - g(\boldsymbol{X}_i))\right)$, we write $\hat{g}_c(\cdot) = \hat{g}(\cdot) - \int_{\boldsymbol{x}} \hat{g}(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} + \mathbb{E}(Y)$. It then follows from

$$\hat{f}_g(Y_i - g(\boldsymbol{X}_i)) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}_h\big(Y_i - g(\boldsymbol{X}_i), Y_j - g(\boldsymbol{X}_j)\big),$$

and $\mathcal{K}_h(y, x) = K((y-x)/h)/h$ that

$$\frac{1}{n}\sum_{i=1}^{n}\left(-\log \hat{f}_{\hat{g}}(Y_i - \hat{g}(\boldsymbol{X}_i))\right) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(-\log \hat{f}_{\hat{g}_c}(Y_i - \hat{g}_c(\boldsymbol{X}_i))\right).$$

- We first show that

$$\begin{aligned}\mathbb{E}\left(\mathcal{R}(\hat{g}_c) - \mathcal{R}(g^*)\right) \;\preceq\; & n^{-1}\log n \log \mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{G}|_{\boldsymbol{x}}) \\ & + \left(\mathcal{R}(g^*_{\mathcal{G}}) - \mathcal{R}(g^*)\right) + \left(\|g^*_{\mathcal{G}} - g^*\|^2_\infty + h^2\right).\end{aligned} \tag{S.8}$$

By the definition of the empirical risk minimizer, we have

$$\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{\hat{g}_c}(Y_i - \hat{g}_c(\boldsymbol{X}_i))\right] \geq \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{g^*_{\mathcal{G}}}(Y_i - g^*_{\mathcal{G}}(\boldsymbol{X}_i))\right]. \tag{S.9}$$

Then, it follows that

$$
\begin{aligned}
&\mathbb{E}\left(\mathcal{R}(\hat{g}_c) - \mathcal{R}(g^*)\right) \\
&= \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{D'_n}(-\log f(Y'_i - \hat{g}_c(\boldsymbol{X}'_i))) - \mathbb{E}_{D'_n}(-\log f(Y'_i - g^*(\boldsymbol{X}'_i)))\right]\right\} \\
&= \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{D'_n}\left(n^{-1}\sum_{i=1}^{n} S(\hat{g}_c, \boldsymbol{Z}'_i)\right) - 2S(\hat{g}_c, \boldsymbol{Z}_i)\right\}\right] \\
&\quad -2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i - \hat{g}_c(\boldsymbol{X}_i)) - \log\left(\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h\left(\hat{g}_c(\boldsymbol{X}_j) - Y_j, \hat{g}_c(\boldsymbol{X}_i) - Y_i\right)\right)\right\}\right] \\
&\quad -2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log \hat{f}_{\hat{g}_c}\left(Y_i - \hat{g}_c(\boldsymbol{X}_i)\right) - \log \hat{f}_{g^*_{\mathcal{G}}}\left(Y_i - g^*_{\mathcal{G}}(\boldsymbol{X}_i)\right)\right\}\right] \\
&\quad -2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h\left(g^*_{\mathcal{G}}(\boldsymbol{X}_j) - Y_j, g^*_{\mathcal{G}}(\boldsymbol{X}_i) - Y_i\right)\right) - \log f(Y_i - g^*_{\mathcal{G}}(\boldsymbol{X}_i))\right\}\right] \\
&\quad +2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n} S(g^*_{\mathcal{G}}, \boldsymbol{Z}_i)\right] \\
&\leq \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{D'_n}\left(n^{-1}\sum_{i=1}^{n} S(\hat{g}_c, \boldsymbol{Z}'_i)\right) - 2S(\hat{g}_c, \boldsymbol{Z}_i)\right\}\right] \\
&\quad -2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i - \hat{g}_c(\boldsymbol{X}_i)) - \log\left(\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h\left(\hat{g}_c(\boldsymbol{X}_j) - Y_j, \hat{g}_c(\boldsymbol{X}_i) - Y_i\right)\right)\right\}\right] \\
&\quad -2\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h\left(g^*_{\mathcal{G}}(\boldsymbol{X}_j) - Y_j, g^*_{\mathcal{G}}(\boldsymbol{X}_i) - Y_i\right)\right) - \log f(Y_i - g^*_{\mathcal{G}}(\boldsymbol{X}_i))\right\}\right] \\
&\quad +2\left(\mathcal{R}(g^*_{\mathcal{G}}) - \mathcal{R}(g^*)\right) \\
&= I_1 + I_2 + I_3 + I_4.
\end{aligned}
$$

Next, we will give an upper bound of I_r for $r = 1, 2, 3$ and handle it with truncation and classical chaining technique of empirical processes.

Upper bound for I_1. According to the definition of $L(g, \boldsymbol{Z}_i)$ and $S(g, \boldsymbol{Z}_i)$, we have

$$\left|L(\hat{g}_c, \boldsymbol{Z}_i) - L(g_{q^*}, \boldsymbol{Z}_i)\right| = \left|\mathbb{E}_{D'_n}\{S(\hat{g}_c, \boldsymbol{Z}'_i)\} - 2S(\hat{g}_c, \boldsymbol{Z}_i) - \mathbb{E}_{D'_n}\{S(g_{q^*}, \boldsymbol{Z}'_i)\} + 2S(g_{q^*}, \boldsymbol{Z}_i)\right|,$$

$$\left|S(\hat{g}_c, \boldsymbol{Z}_i) - S(g_{q^*}, \boldsymbol{Z}_i)\right| = \left|\log f(\hat{g}_c(\boldsymbol{X}_i) - Y_i) - \log f(g_{q^*}(\boldsymbol{X}_i) - Y_i)\right|.$$

Then given the definition of g_{q^*}, it follows from

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left\{S(\hat{g}_c, \boldsymbol{Z}_i)\right\} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left\{S(g_{q^*}, \boldsymbol{Z}_i)\right\}\right| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}|S(\hat{g}_c, \boldsymbol{Z}_i) - S(g_{q^*}, \boldsymbol{Z}_i)| \\
=\ & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}|\log f(\hat{g}_c(\boldsymbol{X}_i) - Y_i) - \log f(g_{q^*}(\boldsymbol{X}_i) - Y_i)| \\
=\ & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left|\frac{f^{(1)}(\epsilon)((\hat{g}_c(\boldsymbol{X}_i) - Y_i) - (g_{q^*}(\boldsymbol{X}_i) - Y_i))}{f(\epsilon)}\right| \\
\leq\ & \mathcal{B}\delta,
\end{aligned}$$

where the last inequality follows from Condition (C3c), and

$$\begin{aligned}
& \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|L(\hat{g}_c, \boldsymbol{Z}_i) - L(g_{q^*}, \boldsymbol{Z}_i)\right|\right\} \\
& = \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{D'_n}\{S(\hat{g}_c, \boldsymbol{Z}'_i)\} - 2S(\hat{g}_c, \boldsymbol{Z}_i) - \mathbb{E}_{D'_n}\{S(g_{q^*}, \boldsymbol{Z}'_i)\} + 2S(g_{q^*}, \boldsymbol{Z}_i)\right|\right\} \\
& \leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{D'_n}\{S(\hat{g}_c, \boldsymbol{Z}'_i)\} - \mathbb{E}_{D'_n}\{S(g_{q^*}, \boldsymbol{Z}'_i)\}\right| + 2\left|S(\hat{g}_c, \boldsymbol{Z}_i) - S(g_{q^*}, \boldsymbol{Z}_i)\right|\right\} \\
& \leq 3\mathcal{B}\delta,
\end{aligned}$$

that

$$\mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L(\hat{g}_c, \boldsymbol{Z}_i)\right\} \leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L(g_{q^*}, \boldsymbol{Z}_i)\right\} + 3\mathcal{B}\delta. \tag{S.10}$$

Let $0 < \beta_n$ be a positive number who may depend on the sample size n. Denote $T_{\beta_n} f = f$ if $f \geq \beta_n$ and $T_{\beta_n} f = \beta_n$ otherwise. Define the function $g^*_{\beta_n}$ by

$$g^*_{\beta_n}(\boldsymbol{x}) = \arg \min_{g:|g|_\infty < \mathcal{B}} \mathbb{E}\left(-\log T_{\beta_n} f(Y - g(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\right).$$

For any $g \in \mathcal{G}$, we let $S_{\beta_n}(g, \boldsymbol{Z}_i) = -\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) + \log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)$ Then, we have

$$\begin{aligned}
\mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} &= \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left\{\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) - \log f(g(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log T_{\beta_n} f(g^*(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log f(g^*(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*(\boldsymbol{X}_i) - Y_i)\right\} \\
&\leq \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left|\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) - \log f(g(\boldsymbol{X}_i) - Y_i)\right| \\
&\quad + \mathbb{E}\left|\log f(g^*(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*(\boldsymbol{X}_i) - Y_i)\right| \\
&\leq \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + 2\mathbb{E}\left|(\log \beta_n - \log f(\epsilon)) I(f(\epsilon) < \beta_n)\right| \\
&\leq \mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} + 4\mathbb{E}\left\{\left|\log f(\epsilon) I(f(\epsilon) < \beta_n)\right|\right\}, \\
\mathbb{E}\left\{S_{\beta_n}(g, \boldsymbol{Z}_i)\right\} &= \mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left\{\log f(g(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i) - \log f(g^*(\boldsymbol{X}_i) - Y_i)\right\} \\
&\quad + \mathbb{E}\left\{\log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i) - \log f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)\right\} \\
&\leq \mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} + \mathbb{E}\left|\log T_{\beta_n} f(g(\boldsymbol{X}_i) - Y_i) - \log f(g(\boldsymbol{X}_i) - Y_i)\right| \\
&\quad + \mathbb{E}\left|\log f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i) - \log T_{\beta_n} f(g^*_{\beta_n}(\boldsymbol{X}_i) - Y_i)\right| \\
&\leq \mathbb{E}\left\{S(g, \boldsymbol{Z}_i)\right\} + 4\mathbb{E}\left\{\left|\log f(\epsilon) I(f(\epsilon) < \beta_n)\right|\right\},
\end{aligned}$$

so

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{S(g,\boldsymbol{Z}_i)\right\}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{S_{\beta_n}(g,\boldsymbol{Z}_i)\right\}\right| \quad\leq\quad 4\mathbb{E}\left\{\left|\log f(\epsilon)I(f(\epsilon)<\beta_n)\right|\right\},$$

which lead to

$$\left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(L(g_{q^*},\boldsymbol{Z}_i)-L_{\beta_n}(g_{q^*},\boldsymbol{Z}_i)\right)\right]\right|\leq 12\mathbb{E}\left\{\left|\log f(\epsilon)I(f(\epsilon)<\beta_n)\right|\right\}. \quad \text{(S.11)}$$

On the other hand, for any $g\in\mathcal{G}$, we have

$$\begin{aligned}
&|S_{\beta_n}(g,\boldsymbol{Z}_i)|\leq 2|\log\beta_n|,\\
&\sigma_S^2(g):=\mathrm{Var}(S_{\beta_n}(g,\boldsymbol{Z}_i))\leq\mathbb{E}\left\{S_{\beta_n}^2(g,\boldsymbol{Z}_i)\right\}\leq 2|\log\beta_n|\mathbb{E}(S_{\beta_n}(g,\boldsymbol{Z}_i)).
\end{aligned}$$

Following the Bernstein inequality, for any $t>0$, let $u=t/2+\sigma_S^2(g)/4|\log\beta_n|$, we have

$$
\begin{aligned}
& P\left\{\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_q, \boldsymbol{Z}_i) > t\right\} \\
&= P\left\{\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{Z}_i')\} - \frac{2}{n}\sum_{i=1}^{n} S_{\beta_n}(g_q, \boldsymbol{Z}_i) > t\right\} \\
&= P\left\{\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{Z}_i')\} - \frac{1}{n}\sum_{i=1}^{n} S_{\beta_n}(g_q, \boldsymbol{Z}_i) > \frac{t}{2} + \frac{1}{2}\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{Z}_i')\}\right\} \\
&\leq P\left\{\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{Z}_i')\} - \frac{1}{n}\sum_{i=1}^{n} S_{\beta_n}(g_q, \boldsymbol{Z}_i) > \frac{t}{2} + \frac{1}{2}\frac{\sigma_S^2(g)}{2|\log\beta_n|}\right\} \\
&\leq \exp(-\frac{nu^2}{2\sigma_S^2(g) + 8u|\log\beta_n|/3}) \\
&\leq \exp(-\frac{nu^2}{8u|\log\beta_n| + 8u|\log\beta_n|/3}) \\
&\leq \exp(-\frac{1}{8+8/3}\frac{nu}{|\log\beta_n|}) \\
&\leq \exp(-\frac{1}{16+16/3}\frac{nt}{|\log\beta_n|}) \\
&= \exp\left(-\frac{Cnt}{|\log\beta_n|}\right),
\end{aligned}
$$

This leads to a tail probability bound of $\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)$, that is

$$
P\left\{\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i) > t\right\} \quad \leq \quad 2\mathcal{N}_{2n}\exp\left(-\frac{Cnt}{|\log\beta_n|}\right).
$$

Then for $a_n > 0$,

$$
\begin{aligned}
\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)\right] \quad &\leq \quad a_n + \int_{a_n}^{\infty} P\left\{\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i) > t\right\} dt \\
&\leq \quad a_n + \int_{a_n}^{\infty} 2\mathcal{N}_{2n}\exp\left(-\frac{Cnt}{|\log\beta_n|}\right) dt \\
&\leq \quad a_n + 2\mathcal{N}_{2n}\exp\left(-a_n\frac{Cn}{|\log\beta_n|}\right)\frac{|\log\beta_n|}{Cn}.
\end{aligned}
$$

Choosing $a_n = \log 2\mathcal{N}_{2n} \frac{|\log \beta_n|}{Cn}$, the above inequality leads to

$$\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)\right] \leq \frac{C|\log \beta_n|(\log 2\mathcal{N}_{2n}+1)}{n}. \tag{S.12}$$

Combining inequalities (S.10), (S.11), (S.12), we have

$$\begin{aligned} I_1 &= \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L(\hat{g}_c, \boldsymbol{Z}_i)\right\} \\ &\leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L(g_{q^*}, \boldsymbol{Z}_i)\right\} + 3\mathcal{B}\delta \\ &\leq \mathbb{E}_{D_n}\left\{\frac{1}{n}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{Z}_i)\right\} + 12\mathbb{E}\left\{\left|\log f(\epsilon) I(f(\epsilon) < \beta_n)\right|\right\} + 3\mathcal{B}\delta \\ &\leq \frac{C|\log \beta_n|(\log 2\mathcal{N}_{2n}+1)}{n} + 12\mathbb{E}\left\{\left|\log f(\epsilon) I(f(\epsilon) < \beta_n)\right|\right\} + 3\mathcal{B}\delta. \end{aligned}$$

Upper bound for I_2 and I_3. Recall that $\tilde{f}_{g_1,h} = \mathbb{E}\left(\mathcal{K}_h(Y_i - g_1(\boldsymbol{X}_i), x)\right)$. For I_2, we have

$$\begin{aligned} I_2 &= \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i - \hat{g}_c(\boldsymbol{X}_i)) - \log \tilde{f}_{\hat{g}_c,h}(Y_i - \hat{g}_c(\boldsymbol{X}_i))\right\}\right] \\ &\quad + \mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log \tilde{f}_{\hat{g}_c,h}(Y_i - \hat{g}_c(\boldsymbol{X}_i)) - \log\left(\frac{1}{n_1}\sum_{j=1}^{n}\mathcal{K}_h\left(\hat{g}_c(\boldsymbol{X}_j) - Y_j, \hat{g}_c(\boldsymbol{X}_i) - Y_i\right)\right)\right\}\right] \\ &= I_{2,1} + I_{2,2}. \end{aligned}$$

Denote $f^{(r)}(\epsilon)$ as the rth derivative of $f(\epsilon)$. We first show that for $\|g_1 - g^*\|_\infty = o_p(1)$ and $h \to 0$, we have

$$\begin{aligned} \tilde{f}_{g_1,h}(z) &= U_0 f(z) + f^{(1)}(z)\left[U_0 L_1(g_1 - g^*) + U_1 h\right] \\ &\quad + 0.5 f^{(2)}(z)\left[U_0 L_2(g_1 - g^*) + 2U_1 h L_1(g_1 - g^*) + U_2 h^2\right](1 + o(1)) \end{aligned} \tag{S.13}$$

Note that

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h(Y_j - g_1(\boldsymbol{X}_j), y - g(\boldsymbol{x}))\right\}\\
&= \int \frac{1}{h}K\left(\frac{Y - g_1(\boldsymbol{X}) - y + g(\boldsymbol{x})}{h}\right) f(Y - g^*(\boldsymbol{X}))f_{\boldsymbol{x}}(\boldsymbol{X})dYd\boldsymbol{X}\\
&= \int K(t)f(g_1(\boldsymbol{X}) - g^*(\boldsymbol{X}) + y - g(\boldsymbol{x}) + th)f_{\boldsymbol{x}}(\boldsymbol{X})dtd\boldsymbol{X}\\
&= \int K(t)\left\{f\left(g_1(\boldsymbol{X}) - g^*(\boldsymbol{X}) + y - g(\boldsymbol{x})\right) + f^{(1)}\left(g_1(\boldsymbol{X}) - g^*(\boldsymbol{X}) + y - g(\boldsymbol{x})\right)th\right.\\
&\quad\left.+f^{(2)}\left(g_1(\boldsymbol{X}) - g^*(\boldsymbol{X}) + y - g(\boldsymbol{x})\right)\frac{1}{2}t^2h^2 + o(h^2)\right\} f_{\boldsymbol{x}}(\boldsymbol{X})dtd\boldsymbol{X}\\
&= U_0\left(f(y - g(\boldsymbol{x})) + f^{(1)}(y - g(\boldsymbol{x}))L_1(g_1 - g^*)\right.\\
&\quad\left.+f^{(2)}(y - g(\boldsymbol{x}))\frac{1}{2}L_2(g_1 - g^*) + O(L_3(g_1 - g^*))\right)\\
&\quad+U_1h\left(f^{(1)}(y - g(\boldsymbol{x})) + f^{(2)}(y - g(\boldsymbol{x}))L_1(g_1 - g^*) + O(L_2(g_1 - g^*))\right)\\
&\quad+\frac{U_2}{2}h^2\left(f^{(2)}(y - g(\boldsymbol{x})) + O(L_1(g_1 - g^*))\right)(1 + o(1)),
\end{aligned}
$$

Then, using $U_0 = 0$ and $U_1 = 1$, we obtain that

$$
\begin{aligned}
|I_{2,1}| &= \left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i - \hat{g}_c(\boldsymbol{X}_i)) - \log \tilde{f}_{\hat{g}_c,h}(Y_i - \hat{g}_c(\boldsymbol{X}_i))\right\}\right]\right| \qquad \text{(S.14)}\\
&= \left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log f(Y_i - \hat{g}_c(\boldsymbol{X}_i)) - \log(f(Y_i - \hat{g}_c(\boldsymbol{X}_i) + f^{(1)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)L_1(\hat{g}_c - g^*))\right.\right.\right.\\
&\quad\left.\left.\left.+f^{(2)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)0.5\left\{L_2(\hat{g}_c - g^*) + U_2h^2\right\})\right\}\right]\right|\\
&= \left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log(1 + \frac{f^{(1)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}L_1(\hat{g}_c - g^*))\right.\right.\right.\\
&\quad\left.\left.\left.+\frac{f^{(2)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}0.5\left\{L_2(\hat{g}_c - g^*) + U_2h^2\right\}\right\}\right]\right|.
\end{aligned}
$$

Let

$$
\begin{aligned}
t_1 &= \mathbb{E}\left|\frac{f^{(1)}(\epsilon)}{f(\epsilon)}\right|, \quad t_4 = \mathbb{E}\left|\frac{f^{(2)}(\epsilon)}{f(\epsilon)}\right|, \\
t_2 &= \mathbb{E}\left|\frac{f^{(2)}\big(\epsilon\big)f\big(\epsilon\big) - (f^{(1)}\big(\epsilon\big))^2}{f^2\big(\epsilon\big)}\right|, \quad t_3 = \mathbb{E}\left|\frac{f^{(3)}\big(\epsilon\big)f\big(\epsilon\big) - f^{(2)}\big(\epsilon\big)f^{(1)}\big(\epsilon\big)}{f^2\big(\epsilon\big)}\right|.
\end{aligned}
$$

Note that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}L_1(\hat{g}_c - g^*)\right\} \\
& \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}L_1(g_{q^*} - g^*)\right\} \\
& \quad + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left|\frac{f^{(1)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}L_1(\hat{g}_c - g^*) - \frac{f^{(1)}\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}L_1(g_{q^*} - g^*)\right| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}L_1(g_{q^*} - g^*)\right\} \\
& \quad + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left|\frac{f^{(2)}\big(\epsilon\big)f\big(\epsilon\big) - (f^{(1)}\big(\epsilon\big))^2}{f^2\big(\epsilon\big)}\right|\|\hat{g}_c(\boldsymbol{X}_i) - g_{q^*}(\boldsymbol{X}_i)\|_\infty L_1(|\hat{g}_c - g^*|) \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}L_1(g_{q^*} - g^*)\right\} + t_2\mathcal{B}\delta,
\end{aligned}
$$

and

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}L_1(g_{q^*} - g^*)\right\} \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}L_1(\hat{g}_c - g^*)\right\} + t_2\mathcal{B}\delta, \\
& \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left[\frac{f^{(2)}\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i - \hat{g}_c(\boldsymbol{X}_i)\big)}0.5\left\{L_2(\hat{g}_c - g^*) + U_2h^2\right\}\right]\right. \\
& \qquad\qquad \left. -\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left[\frac{f^{(2)}\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i - g_{q^*}(\boldsymbol{X}_i)\big)}0.5\left\{L_2(g_{q^*} - g^*) + U_2h^2\right\}\right]\right| \\
\leq\ & (t_3\mathcal{B}^2 + t_3h^2)\delta.
\end{aligned}
$$

Then, it follows from expression (S.14) and the above inequalities and Condition

(C3c), we have

$$
\begin{aligned}
|I_{2,1}| &\leq \left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{f^{(1)}\big(Y_i-\hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i-\hat{g}_c(\boldsymbol{X}_i)\big)}L_1(\hat{g}_c-g^*)\right.\right.\\
&+ \left.\left.\frac{1}{n}\sum_{i=1}^{n}\frac{f^{(2)}\big(Y_i-\hat{g}_c(\boldsymbol{X}_i)\big)}{f\big(Y_i-\hat{g}_c(\boldsymbol{X}_i)\big)}0.5\left\{L_2(\hat{g}_c-g^*)+U_2h^2\right\}(1+o_p(1))\right]\right|\\
&\leq \left|\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i-g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i-g_{q^*}(\boldsymbol{X}_i)\big)}L_1(g_{q^*}-g^*)\right\}\right|\\
&\quad+\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}_{D_n}\left[\frac{f^{(2)}\big(Y_i-g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i-g_{q^*}(\boldsymbol{X}_i)\big)}0.5\left\{L_2(g_{q^*}-g^*)+U_2h^2\right\}\right]\right|(1+o_p(1))\\
&\quad+(t_2\mathcal{B}+t_3\mathcal{B}^2+t_3h^2)\delta\\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\big(Y_i-g^*(\boldsymbol{X}_i)\big)}{f\big(Y_i-g^*(\boldsymbol{X}_i)\big)}\right\}\mathbb{E}_{D_n}\left\{L_1(g_{q^*}-g^*)\right\}\right|+\|g_{q^*}-g^*\|_\infty^2\\
&\quad+\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left[\frac{f^{(2)}\big(Y_i-g_{q^*}(\boldsymbol{X}_i)\big)}{f\big(Y_i-g_{q^*}(\boldsymbol{X}_i)\big)}0.5\left\{L_2(g_{q^*}-g^*)+U_2h^2\right\}\right]\right|(1+o_p(1))\\
&\quad+(t_2\mathcal{B}+t_3\mathcal{B}^2+t_3h^2)\delta\\
&\preceq \|g^*_{\mathcal{G}}-g^*\|_\infty^2+h^2+\delta, \qquad\qquad (\text{S.15})
\end{aligned}
$$

where the last inequality follows from the definition of g_{q^*}, i.e., $\|g_{q^*}-\hat{g}_c\|_\infty<\delta$ and $g_{q^*}\in\mathcal{G}$, and the unbiasedness of the score function, i.e., $\mathbb{E}_{D_n}\left\{\frac{f^{(1)}\left(Y_i-g^*(\boldsymbol{X}_i)\right)}{f\left(Y_i-g^*(\boldsymbol{X}_i)\right)}\right\}=0$.

For $I_{2,2}$, let $L_{\mathcal{K}}(g_1,g,\boldsymbol{Z}_i)=\log\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\mathcal{K}_h(g_1(\boldsymbol{X}_j)-Y_j,g(\boldsymbol{X}_i)-Y_i)\right)-\log\tilde{f}_{g_1,h}(Y_i-$

$g(\boldsymbol{X}_i))$. Then, we have that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left(L_{\mathcal{K}}(\hat{g}_c,\hat{g}_c,\boldsymbol{Z}_i)\right) \\
& = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left[\log\left(\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h(\hat{g}_c(\boldsymbol{X}_j)-Y_j,\hat{g}_c(\boldsymbol{X}_i)-Y_i)\right)-\log\tilde{f}_{\hat{g}_c,h}(Y_i-\hat{g}_c(\boldsymbol{X}_i))\right] \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left[\log\left(1+\frac{\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h(\hat{g}_c(\boldsymbol{X}_j)-Y_j,\hat{g}_c(\boldsymbol{X}_i)-Y_i)-\tilde{f}_{\hat{g}_c,h}(Y_i-\hat{g}_c(\boldsymbol{X}_i)}{\tilde{f}_{\hat{g}_c,h}(Y_i-\hat{g}_c(\boldsymbol{X}_i)}\right)\right] \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\left[\frac{\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h(\hat{g}_c(\boldsymbol{X}_j)-Y_j,\hat{g}_c(\boldsymbol{X}_i)-Y_i)-\tilde{f}_{\hat{g}_c,h}(Y_i-\hat{g}_c(\boldsymbol{X}_i))}{\tilde{f}_{\hat{g}_c,h}(Y_i-\hat{g}_c(\boldsymbol{X}_i))}(1+o_p(1))\right] \\
\le \;& \Bigg(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\frac{\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h(g^*(\boldsymbol{X}_j)-Y_j,g^*(\boldsymbol{X}_i)-Y_i)-\tilde{f}_{g^*,h}(Y_i-g^*(\boldsymbol{X}_i))}{\tilde{f}_{g^*,h}(Y_i-g^*(\boldsymbol{X}_i))}\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\frac{-\frac{1}{n}\sum_{j=1}^{n}\dot{\mathcal{K}}_h(g^*(\boldsymbol{X}_j)-Y_j,g^*(\boldsymbol{X}_i)-Y_i)-\dot{\tilde{f}}_{g^*,h}(Y_i-g^*(\boldsymbol{X}_i))}{\dot{\tilde{f}}_{g^*,h}(Y_i-g^*(\boldsymbol{X}_i))}\right|\|\hat{g}_c-g^*\|_\infty\Bigg) \\
& \times(1+o_p(1)).
\end{aligned}
$$

Write $\boldsymbol{Z}_i = Y_i - g^*(\boldsymbol{X}_i)$. Let $U(\boldsymbol{Z}_j,\boldsymbol{Z}_i) = \frac{\mathcal{K}_h(\boldsymbol{Z}_j,\boldsymbol{Z}_i)-\tilde{f}_{g^*,h}(\boldsymbol{Z}_i)}{\tilde{f}_{g^*,h}(\boldsymbol{Z}_i)}\times 0.5+\frac{\mathcal{K}_h(\boldsymbol{Z}_i,\boldsymbol{Z}_j)-\tilde{f}_{g^*,h}(\boldsymbol{Z}_j)}{\tilde{f}_{g^*,h}(\boldsymbol{Z}_j)}\times 0.5$. Clearly, using the U-statistics theory (Theorems 1 and 3 of Chapter 1.3 in Lee

2019), we have

$$
\begin{aligned}
\tilde{U}(\boldsymbol{Z}_j) &= \mathbb{E}_{\boldsymbol{Z}_i}\left[U(\boldsymbol{Z}_j, \boldsymbol{Z}_i \| \boldsymbol{Z}_j)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}_i}\left[\frac{\mathcal{K}_h(\boldsymbol{Z}_i, \boldsymbol{Z}_j) - \tilde{f}_{g^*,h}(\boldsymbol{Z}_i)}{\tilde{f}_{g^*,h}(\boldsymbol{Z}_i)} \times 0.5\right] \\
&= 0.5 \times \int \frac{\mathcal{K}_h(\boldsymbol{Z}_i, \boldsymbol{Z}_j) - \tilde{f}_{g^*,h}(\boldsymbol{Z}_i)}{\tilde{f}_{g^*,h}(\boldsymbol{Z}_i)} f(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \\
&= 0.5 \times \int \frac{\mathcal{K}_h(\boldsymbol{Z}_i, \boldsymbol{Z}_j)}{\tilde{f}_{g^*,h}(\boldsymbol{Z}_i)} f(\boldsymbol{Z}_i) d\boldsymbol{Z}_i - 0.5 \\
&= 0.5 \times \int \frac{1}{h} \frac{\mathcal{K}(\frac{\boldsymbol{Z}_i - \boldsymbol{Z}_j}{h})}{\tilde{f}_{g^*,h}(\boldsymbol{Z}_i)} f(\boldsymbol{Z}_i) d\boldsymbol{Z}_i - 0.5 \\
&= 0.5 \times \int \frac{\mathcal{K}(t)}{\tilde{f}_{g^*,h}(th + \boldsymbol{Z}_j)} f(th + \boldsymbol{Z}_j) dt - 0.5 \\
&= 0.5 \times \int \mathcal{K}(t) \frac{f(th + \boldsymbol{Z}_j)}{\tilde{f}_{g^*,h}(th + \boldsymbol{Z}_j)} dt - 0.5 \\
&= 0.5 \times \int \mathcal{K}(t) \Big(\tilde{f}_{g^*,h}(th + \boldsymbol{Z}_j) - f^{(1)}\big(th + \boldsymbol{Z}_j\big) L_1(g^* - g^*) \\
&\quad - f^{(2)}\big(th + \boldsymbol{Z}_j\big) 0.5 \left\{L_2(g^* - g^*) + U_2 h^2\right\}\Big) \Big/ \Big(\tilde{f}_{g^*,h}(th + \boldsymbol{Z}_j)\Big) dt - 0.5 \\
&= 0.5 \times \int \mathcal{K}(t) \left(1 - \frac{f^{(2)}}{\tilde{f}_{g_{q^*},h}} \big(th + \boldsymbol{Z}_j\big) 0.5 U_2 h^2\right) dt - 0.5 \\
&= -\left(\frac{f^{(2)}}{\tilde{f}_{g_{q^*},h}} \big(\boldsymbol{Z}_j\big) 0.5 U_2 h^2 (1 + o(1))\right),
\end{aligned}
$$

which leads to

$$
Var(\tilde{U}(\boldsymbol{Z}_j)) \leq h^4.
$$

Then, we have

$$
\mathbb{E}_{D_n}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n} U(\boldsymbol{Z}_j, \boldsymbol{Z}_i)\right| \preceq n^{-1/2} h^2.
$$

Note that

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h^{(1)}(Y_j-g_1(\boldsymbol{X}_j),y-g(\boldsymbol{x}))\right\}\\
&=\int\frac{1}{h^2}K^{(1)}\left(\frac{Y-g_1(\boldsymbol{X})-y+g(\boldsymbol{x})}{h}\right)f(Y-g^*(\boldsymbol{X}))f_{\boldsymbol{x}}(\boldsymbol{X})dYd\boldsymbol{X}\\
&=\int\frac{1}{h}K^{(1)}(t)f(g_1(\boldsymbol{X})-g^*(\boldsymbol{X})+y-g(\boldsymbol{x})+th)f_{\boldsymbol{x}}(\boldsymbol{X})dtd\boldsymbol{X}\\
&=\int K^{(1)}(t)\left\{h^{-1}f\left(g_1(\boldsymbol{X})-g^*(\boldsymbol{X})+y-g(\boldsymbol{x})\right)+f^{(1)}\left(g_1(\boldsymbol{X})-g^*(\boldsymbol{X})+y-g(\boldsymbol{x})\right)t\right.\\
&\left.+f^{(2)}\left(g_1(\boldsymbol{X})-g^*(\boldsymbol{X})+y-g(\boldsymbol{x})\right)\frac{1}{2}t^2h+o(h)\right\}f_{\boldsymbol{x}}(\boldsymbol{X})dtd\boldsymbol{X}\\
&=h^{-1}\int K^{(1)}(t)dt\left(f(y-g(\boldsymbol{x}))+f^{(1)}(y-g(\boldsymbol{x}))L_1(g_1-g^*)\right.\\
&\left.+f^{(2)}(y-g(\boldsymbol{x}))\frac{1}{2}L_2(g_1-g^*)+O(L_3(g_1-g^*))\right)\\
&+\int K^{(1)}(t)tdt\left(f^{(1)}(y-g(\boldsymbol{x}))+f^{(2)}(y-g(\boldsymbol{x}))L_1(g_1-g^*)+O(L_2(g_1-g^*))\right)\\
&+\frac{\int K^{(1)}(t)t^2dt}{2}h\left(f^{(2)}(y-g(\boldsymbol{x}))+O(L_1(g_1-g^*))\right)(1+o(1)).
\end{aligned}
$$

Then, after similar calculations, we have that

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{D_n}\frac{-\frac{1}{n}\sum_{j=1}^{n}\dot{\mathcal{K}}_h(g^*(\boldsymbol{X}_j)-Y_j,g^*(\boldsymbol{X}_i)-Y_i)-\dot{\hat{f}}_{g^*,h}(Y_i-g^*(\boldsymbol{X}_i))}{\dot{\hat{f}}_{g^*,h}(Y_i-g^*(\boldsymbol{X}_i))}\right|\\
&\qquad\times\|\hat{g}_c-g^*\|_\infty\leq n^{-1/2}h\|\hat{g}_c-g^*\|_\infty.
\end{aligned}
$$

Thus, we can obtain that

$$
|I_{2,2}| \;\preceq\; \frac{1}{\sqrt{n}}(h^2+h\|\hat{g}_c-g^*\|_\infty). \tag{S.16}
$$

Combining inequalities (S.15) and (S.16), it follows from $h\to 0$,

$$
|I_2| \;\leq\; h^2+n^{-1/2}h^2+\delta+\|g^*_{\mathcal{G}}-g^*\|^2_\infty.
$$

Similarly, we can obtain that $|I_3| = O(|I_2|)$. Thus, equation (S.8) holds from condition (C3a).

- We then show that

$$\|\tilde{g}(\boldsymbol{x}) - \hat{g}_c(\boldsymbol{x})\|_\infty = O_p(n^{-1/2}).$$

In particular,

–

$$\begin{aligned}
\tilde{g}(\cdot) &= \hat{g}_c(\cdot) + \int \hat{g}(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} - \mathbb{E}(Y) - \frac{1}{n}\sum_{i=1}^{n} \hat{g}(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n} Y_i \\
&= \hat{g}_c(\cdot) + \left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}(Y)\right) + \left(\int \hat{g}(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} - \frac{1}{n}\sum_{i=1}^{n} \hat{g}(\boldsymbol{X}_i)\right) \\
&:= \hat{g}_c(\cdot) + I_1 + I_2.
\end{aligned}$$

– Given $\mathbb{E}Y^2 < \infty$, it follows from central limit theorem that

$$I_1 := \frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}(Y) = O_p(n^{-1/2}). \tag{S.17}$$

– Define $F_{n,\boldsymbol{x}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} I(\boldsymbol{X}_i \le \boldsymbol{x})$. Let $F_{\boldsymbol{x}}$ be the distribution function of $\boldsymbol{x}$. Based on the empirical measure, we have

$$\begin{aligned}
|I_2| &:= \Big|\frac{1}{n}\sum_{i=1}^{n} \hat{g}(\boldsymbol{X}_i) - \int \hat{g}(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}\Big| \\
&= \Big|\int \hat{g}(\boldsymbol{x}) dF_n(\boldsymbol{x}) - \int \hat{g}(\boldsymbol{x}) dF(\boldsymbol{x})\Big| \\
&= \Big|\int \hat{g}(\boldsymbol{x}) d\left(F_{n,\boldsymbol{x}}(\boldsymbol{x}) - F_{\boldsymbol{x}}(\boldsymbol{x})\right)\Big| \\
&\le \frac{\mathcal{B}}{n^{1/2}},
\end{aligned} \tag{S.18}$$

where the last inequality follows from the upper bounded condition of $\hat{g}$ and the $n^{1/2}$ convergence rate of empirical distribution. Similar conclusions can also be found in Bickel and Ritov (2003). They claimed that functionals of a nonparametric regression function is able to be estimated efficiently, i.e., at $n^{1/2}$ convergence rate (Last paragraph on Page 1050).

- Last, we show that

$$\mathbb{E}\left(\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*)\right) \simeq \mathbb{E}\left(\mathcal{R}(\hat{g}_c) - \mathcal{R}(g^*)\right) + O(n^{-1}).$$

Particularly,

$$\begin{aligned}
\mathbb{E}\left(\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*)\right) &= \mathbb{E}\left(\mathcal{R}(\hat{g}_c + I_1 + I_2) - \mathcal{R}(g^*)\right) \\
&= \mathbb{E}\left(\mathcal{R}(\hat{g}_c) - \mathcal{R}(g^*)\right) + \mathbb{E}\left(\mathcal{R}(\hat{g}_c + I_1 + I_2) - \mathcal{R}(\hat{g}_c)\right) \\
&\simeq \mathbb{E}\left(\mathcal{R}(\hat{g}_c) - \mathcal{R}(g^*)\right) + O_p(n^{-1}),
\end{aligned}$$

where the last approximation follows from

$$\mathbb{E}\left(\mathcal{R}(g_1) - \mathcal{R}(g_2)\right) \simeq \|g_1 - g_2\|_\infty^2,$$

and $\|I_1 + I_2\|_\infty^2 = O_p(n^{-1})$ by (S.17) and (S.18).

Then, Theorem 2 follows.

S.2.4 Proof of Proposition 1

According to Theorem 2, the variance of $\tilde{g}$ equals to the variance of $\hat{g}_{oracle}$ (Substituting f with $\hat{f}$ and rescaling does not introduce additional variance). Thus, it suffices to show that for any asymptotically unbiased $\breve{g}$,

$$\mathrm{Var}(\breve{g}) \geq \mathrm{Var}(\hat{g}_{oracle}) = \mathrm{Var}(\tilde{g})(1 + o(1)).$$

According to Stoica and Marzetta (2001), when J is singular, any biased estimator $\check{g}(\boldsymbol{x}) := g(\boldsymbol{x}; \check{\boldsymbol{\theta}})$ with finite variance C must satisfy (inequality (16) in Stoica and Marzetta (2001))

$$C := \operatorname{Var}(\check{g}(\boldsymbol{x})) \geq nHJ^{\dagger}H^{\top}, \tag{S.19}$$

where

$$H = \frac{1}{n}\left(\frac{\partial\left(\mathbb{E}\{\check{g}(\boldsymbol{x})\} - g(\boldsymbol{x};\boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}^{\top}} + \frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\top}} + \mathbb{E}(\check{g}(\boldsymbol{x}))o(1)\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

and $B^{\dagger}$ is the Moore-Penrose generalized inverse of a matrix B.

Let's consider the asymptotically unbiased estimator $g(x, \check{\boldsymbol{\theta}})$ obtained by minimizing a loss function ℓ under the framework of FNN. That is, minimizing the loss $\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \boldsymbol{X}_i; g(\cdot; \boldsymbol{\theta}))$. Using $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} = 0$, we can obtain that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}))}{\partial g(\boldsymbol{x};\boldsymbol{\theta})} |_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial g(\boldsymbol{x};\boldsymbol{\theta})} |_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) \frac{\partial \boldsymbol{\theta}}{\partial g(\boldsymbol{x};\boldsymbol{\theta})} |_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} \\
&= 0.
\end{aligned}$$

Then it follows that

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}))}{\partial g(\boldsymbol{x};\boldsymbol{\theta})} |_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)} + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2} \left(g(\boldsymbol{x};\check{\boldsymbol{\theta}}) - g(\boldsymbol{x};\boldsymbol{\theta}^*)\right) \\
&\quad + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 \ell(Y_i, \boldsymbol{X}_i; g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^3} \left(g(\boldsymbol{x};\check{\boldsymbol{\theta}}) - g(\boldsymbol{x};\boldsymbol{\theta}^*)\right)^2 (1 + o_p(1)),
\end{aligned} \tag{S.20}$$

which leads to

$$
\begin{aligned}
& g(\boldsymbol{x};\check{\boldsymbol{\theta}}) \\
= \;& g(\boldsymbol{x};\boldsymbol{\theta}^*) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)} \\
& -\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right\}^{-1}\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^3}\left(g(\boldsymbol{x};\check{\boldsymbol{\theta}})-g(\boldsymbol{x};\boldsymbol{\theta}^*)\right)^2(1+o_p(1)).
\end{aligned}
\tag{S.21}
$$

Note that the approximation bias of $g(\boldsymbol{x};\boldsymbol{\theta})$ to $g(\boldsymbol{x})$, expressed by $g(\boldsymbol{x};\boldsymbol{\theta}^*) - g(\boldsymbol{x})$, is determined by the function class $g(\boldsymbol{x})$ belongs to and the framework of FNN, and tends to zero if $g(\cdot;\check{\boldsymbol{\theta}})$ achieves nearly minimax optimal rate (Jiao et al. 2023). Therefore, an asymptotically unbiased FNN estimator achieving nearly minimax optimal rate actually satisfies $Eg(\boldsymbol{x};\check{\boldsymbol{\theta}}) - g(\boldsymbol{x},\boldsymbol{\theta}^*) = o(1)$. Then it follows that $\mathbb{E}\left(\frac{\partial\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right) = o(1)$ from (S.21). Furthermore, from (S.21), we have

$$
\begin{aligned}
& \mathbb{E}\left(g(\boldsymbol{x};\check{\boldsymbol{\theta}})-g(\boldsymbol{x};\boldsymbol{\theta}^*)\right)^2 \\
= \;& \left\{\mathbb{E}\left(\frac{\partial^2\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}\left\{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)^2\right\} \\
& \times\left\{\mathbb{E}\left(\frac{\partial^2\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}(1+o(1)) \\
= \;& \left[\mathbb{E}\left(\frac{\partial\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)\left\{\mathbb{E}\left(\frac{\partial^2\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}\right]^2 \\
& +\frac{1}{n}\left\{\mathbb{E}\left(\frac{\partial\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}-\mathbb{E}\left(\frac{\partial\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)\right)^2\right\} \\
& \times\left\{\mathbb{E}\left(\frac{\partial^2\ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-2}(1+o(1)).
\end{aligned}
\tag{S.22}
$$

Substituting (S.22) into (S.21), it then follows from $\mathbb{E}\left(\frac{\partial \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)=o(1)$ that

$$
\begin{aligned}
Bias \;:=\;& \mathbb{E}\{g(\boldsymbol{x};\check{\boldsymbol{\theta}})\}-g(\boldsymbol{x};\boldsymbol{\theta}^*)\\
=\;& -\mathbb{E}\left(\frac{\partial \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)\left\{\mathbb{E}\left(\frac{\partial^2 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}\\
& -\left\{\mathbb{E}\left(\frac{\partial^2 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}\frac{1}{2}\mathbb{E}\left(\frac{\partial^3 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^3}\right)\\
& \times\mathbb{E}\left(g(\boldsymbol{x};\check{\boldsymbol{\theta}})-g(\boldsymbol{x};\boldsymbol{\theta}^*)\right)^2(1+o(1))\\
=\;& -\mathbb{E}\left(\frac{\partial \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)\left\{\mathbb{E}\left(\frac{\partial^2 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}\\
& -\frac{1}{2n}\left\{\mathbb{E}\left(\frac{\partial^2 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-1}\mathbb{E}\left(\frac{\partial^3 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^3}\right)\\
& \times\left\{\mathbb{E}\left(\frac{\partial \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}-\mathbb{E}\left(\frac{\partial \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}\right)\right)^2\right\}\\
& \times\left\{\mathbb{E}\left(\frac{\partial^2 \ell(Y_i,\boldsymbol{X}_i;g(\cdot;\boldsymbol{\theta}^*))}{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)^2}\right)\right\}^{-2}(1+o(1)).
\end{aligned}
$$

Thus, if the loss function has bounded third order derivatives with respect to g, it holds that

$$\frac{\partial Bias}{\partial \boldsymbol{\theta}^*}=o(1).$$

Then it follows from (S.19) that

$$C=\mathrm{Var}(\breve{g}(\boldsymbol{x}))\geq n\left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right)J^\dagger\left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}:=V.$$

We next show that $V>0$. Utilizing the eigenvector/eigenvalue representation of J, we have

$$J=U\Lambda U^\top=\begin{bmatrix}U_1 & U_2\end{bmatrix}\begin{bmatrix}\Lambda_1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}U_1^\top\\ U_2^\top\end{bmatrix},$$

where U is orthonormal, $\Lambda_1 \in \mathbb{R}^{r\times r}$ is diagonal and positive definite. Then, it follows from the definition of Moore-Penrose inverse $J^\dagger$ that

$$J^\dagger = U_1 \Lambda_1^{-1} U_1^\top,$$

Thus, combining with $\partial g(\boldsymbol{x};\boldsymbol{\theta})/\partial \boldsymbol{\theta} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \neq 0$, we have

$$V = n\left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) U_1\Lambda_1^{-1}U_1^\top \left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) > 0. \tag{S.23}$$

On the other hand, note that, the oracle estimator

$$\hat{g}_{oracle} = \arg\min_{g\in\mathcal{G}} \left\{ \frac{1}{n}\sum_{i=1}^n (-\log f(Y_i - g(\boldsymbol{X}_i))) \right\}.$$

Then, it follows from

$$\begin{aligned}
0 \quad &\equiv \quad \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{oracle}} \\
&= \quad \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \\
&\quad + \left[\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right] (\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*) + O_p(\|\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*\|_2^2),
\end{aligned}$$

that

$$\begin{aligned}
&\left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) J^\dagger \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \\
&= -\left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) J^\dagger \left[\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right] (\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*)(1 + o_p(1)) \\
&= \left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) J^\dagger J (\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*)(1 + o_p(1)) \\
&= \left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) (\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*)(1 + o_p(1)),
\end{aligned}$$

where the last equality follows from equation (18) in Stoica and Marzetta (2001). According to Theorem 1, the oracle estimator $\|\hat{g}_{oracle} - g(\boldsymbol{x};\boldsymbol{\theta}^*)\|_\infty^2 \to 0$ under some conditions on the function class and network width and depth (Corollary 1). Thus, we can obtain that

$$\begin{aligned}
& g(\boldsymbol{x};\hat{\boldsymbol{\theta}}_{oracle}) \\
= \;& g(\boldsymbol{x};\boldsymbol{\theta}^*) + \frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*) + O_p(\|\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*\|_2^2) \\
= \;& g(\boldsymbol{x};\boldsymbol{\theta}^*) + n\left(\frac{1}{n}\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right) J^\dagger \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(Y_i - g(\boldsymbol{X}_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) |_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} + O_p(\|\hat{\boldsymbol{\theta}}_{oracle} - \boldsymbol{\theta}^*\|_2^2).
\end{aligned}$$

It then follows that the variance of $\hat{g}_{oracle}$ achieves the bound V, i.e.,

$$\mathrm{Var}(\hat{g}_{oracle}) = V(1+o(1)).$$

Thus, we can obtain that for any asymptotically unbiased FNN based estimator,

$$\mathrm{Var}(\breve{g}) \geq \mathrm{Var}(\hat{g}_{oracle}).$$

Then Proposition 1 follows.

S.3 Lemmas

Lemma S.1. *(Approximation error, (Theorem 3.3 in Jiao et al. 2023))*

Given Hölder smooth functions $g^* \in \mathcal{H}_\beta([0,1]^d, B_0)$, for any $D \in \mathbb{N}^+$ and $W \in \mathbb{N}^+$, there exists a function $g^*_{\mathcal{G}}$ implemented by a ReLU feedforward neural network with width $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1} W \lceil \log_2(8W) \rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 D \lceil \log_2(8D) \rceil$ such that

$$\left|g^* - g^*_{\mathcal{G}}\right| \leq 18 B_0 (\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor + \max\{\beta,1\}/2} (WD)^{-2\beta/d},$$

for all $x \in [0,1]^d \setminus \Omega([0,1]^d, K, \delta)$ where

$$\Omega([0,1]^d, K, \delta) = \bigcup_{i=1}^d \{x = [x_1, \cdots, x_d]^T : x_i \in \bigcup_{k=1}^{K-1} (k/K - \delta, k/K)\},$$

with $K = \lfloor WD \rfloor$ and δ is an arbitrary number in $(0, 1/3K]$.

Lemma S.2. *(Bounding the covering number, (Theorem 12.2 in Anthony et al. 1999) and (Theorems 3 and 7 in Bartlett et al. 2019))*

Let ReLU feedforward neural network $\mathcal{G}$ be a set of real functions from a domain $\mathcal{X}$ to the bounded interval $[0, \mathcal{B}]$. There exists a universal constant C such that the following holds. Given any $\mathcal{D}, \mathcal{S}$ with $\mathcal{S} > C\mathcal{D} > C^2$, there exists network class $\mathcal{G}$ with $\leq \mathcal{D}$ layers and $\leq \mathcal{S}$ parameters with VC-dimension $\geq \mathcal{S}\mathcal{D}\log(\mathcal{S}/\mathcal{D})/C$ and given $\delta > 0$

$$\mathcal{N}_{2n}(\delta, \|\cdot\|_\infty, \mathcal{G}|_{\boldsymbol{x}}) \leq \sum_{i=1}^{\mathcal{S}\mathcal{D}\log(\mathcal{S}/\mathcal{D})/C} \binom{2n}{i} \left(\frac{\mathcal{B}}{\delta}\right) = O\big(\mathcal{S}\mathcal{D}\log(\mathcal{S})\big).$$

Lemma S.3. *(Intrinsic dimensionality, Shen (2020))*

Let f be a continuous functions on $[0,1]^d$ and $\mathcal{M} \subseteq [0,1]^d$ be a compact $d_{\mathcal{M}}$-dimensional Riemannian submanifold. For any $D \in \mathbb{N}^+$, $W \in \mathbb{N}^+$, $\epsilon \in (0,1)$ and $\delta \in (0,1)$, there exists a function g^*_ϕ implemented by a ReLU feedforward neural network with width $\mathcal{W} = 3^{d_\delta} \max\{d_\delta \lfloor W^{1/d_\delta} \rfloor, W + 1\}$ and depth $\mathcal{D} = 12D + 14 + 2d_\delta$ such that

$$\left\| g^*(x) - g^*_\phi(x) \right\| \leq 2\omega_g\left(\frac{2\epsilon}{1-\delta}\sqrt{\frac{d}{d_\delta}} + 2\epsilon\right) + 19\sqrt{d}\,\omega_g\left(\frac{2\sqrt{d}}{(1-\delta)\sqrt{d_\delta}} W^{-2/d_\delta} D^{-2/d_\delta}\right)$$

for any $x \in \mathcal{M}_\epsilon$, where $\mathcal{M} := \{x \in [0,1]^d : \inf\{\|x - y\|_2 : y \in \mathcal{M}\} \leq \epsilon\}$, for $\epsilon \in (0,1)$ and d_δ is an integer such that $d_{\mathcal{M}} \leq d_\delta \leq d$.

S.4 Results in numerical studies

		LS	LAD	Huber	Cauchy	Turkey	EML
				$n = 256, d = 100$			
Normal	PE	**1.1169**	1.3431	1.2402	1.3904	1.3421	1.1181
	SD	**0.0473**	0.0591	0.0518	0.0653	0.0583	0.0479
Mixture Gaussian	PE	10.2935	9.8321	10.1538	9.5132	9.3231	**8.2394**
	SD	0.8363	0.7035	0.8111	0.6683	0.6051	**0.5582**
Student-t	PE	10.7079	9.1759	10.2351	8.8393	9.9832	**8.0394**
	SD	3.9652	2.7836	2.8532	2.7563	2.8165	**2.3942**
Heteroscedasticit	PE	18.6772	17.5086	18.4008	17.1440	16.6032	**14.5063**
	SD	1.4171	1.2793	1.2759	1.1432	1.1234	**0.9011**
				$n = 256, d = 500$			
Normal	PE	**1.1175**	1.3967	1.3531	1.4973	1.4196	1.1193
	SD	**0.0487**	0.0610	0.0523	0.0671	0.0610	0.0491
Mixture Gaussian	PE	10.5336	9.8851	10.4314	9.6366	9.9842	**8.3159**
	SD	0.8630	0.7246	0.8616	0.7502	0.8166	**0.6142**
Student-t	PE	11.3261	9.5716	10.8394	9.3742	10.0432	**8.4261**
	SD	4.1326	2.8154	2.9862	2.7738	2.9143	**2.5142**
Heteroscedasticit	PE	18.7571	17.6223	18.4314	17.3645	16.7491	**14.5154**
	SD	1.4224	1.2831	1.3238	1.2243	1.2032	**0.9132**
				$n = 1024, d = 500$			
Normal	PE	1.1112	1.2548	1.1942	1.3613	1.2332	**1.0972**
	SD	**0.0461**	0.0568	0.0503	0.0609	0.0515	0.0466
Mixture Gaussian	PE	10.1931	9.806	9.9978	9.3856	9.3031	**8.2185**
	SD	0.7147	0.6712	0.6990	0.5343	0.5343	**0.5154**
Student-t	PE	9.5032	8.1356	8.7142	8.0797	8.3921	**7.1264**
	SD	3.924	2.5163	2.7168	2.4766	2.6032	**2.3012**
Heteroscedasticit	PE	17.9693	17.1379	17.6407	16.8798	16.5712	**14.4745**
	SD	1.2401	1.1957	1.2394	1.1396	1.1073	**0.8879**

Table S1: The mean and standard deviation of prediction error of g_5 when using six methods for four error distributions with sample sizes $n = 256, 1024$, testing sample size $t = 2048$ and input dimensions $d = 100, 500$.

		LS	LAD	Huber	Cauchy	Turkey	EML
				$n = 256, d = 200$			
Normal	PE	**1.1603**	1.3703	1.1909	1.3904	1.3450	1.1632
	SD	0.0592	0.0911	0.0685	0.0879	0.0905	**0.0464**
Mixture Gaussian	PE	10.4945	10.1405	10.4869	9.7103	9.5147	**8.3187**
	SD	0.9168	0.8779	0.9067	0.8695	0.8453	**0.7168**
Student-t	PE	13.9393	11.7430	13.3536	11.7075	11.273	**10.0289**
	SD	10.1406	9.9717	10.0178	9.1084	9.0983	**8.0537**
Heteroscedasticit	PE	18.8234	17.8664	18.6526	16.8545	16.6812	**14.4935**
	SD	1.2655	1.13	1.1842	1.0594	1.0032	**0.8785**
				$n = 256, d = 600$			
Normal	PE	**1.1835**	1.5035	1.2898	1.4856	1.5304	1.1912
	SD	0.0651	0.1508	0.0732	0.0988	0.1768	**0.0537**
Mixture Gaussian	PE	11.4058	10.2008	10.534	9.7404	9.9544	**8.3311**
	SD	1.1526	0.9003	0.9527	0.8813	0.8651	**0.7556**
Student-t	PE	14.6317	12.6555	13.7298	12.3261	13.3261	**11.1408**
	SD	11.9669	10.8725	11.8775	10.852	11.8499	**9.6821**
Heteroscedasticit	PE	20.1975	18.3254	19.0039	17.2722	16.7207	**14.5214**
	SD	1.4803	1.2841	1.5119	1.1595	1.0528	**0.8937**
				$n = 1024, d = 600$			
Normal	PE	1.1593	1.3111	1.2323	1.3772	1.3347	**1.1082**
	SD	0.0531	0.0842	0.0548	0.0907	0.0895	**0.0416**
Mixture Gaussian	PE	10.3174	9.5994	10.2685	9.4418	9.3905	**8.2696**
	SD	0.8825	0.863	0.8779	0.8407	0.7982	**0.6617**
Student-t	PE	12.8441	11.1071	12.6051	11.0201	12.4215	**9.4071**
	SD	9.8297	8.8399	9.7424	8.8095	9.7735	**7.7424**
Heteroscedasticit	PE	18.1690	17.4382	18.2861	16.6768	16.5519	**14.3442**
	SD	1.1637	1.0451	1.0982	1.0028	0.8494	**0.78**

Table S2: The mean and standard deviation of prediction error of g_{10} when using six methods for four error distributions with sample sizes $n = 256, 1024$, testing sample size $t = 2048$ and input dimensions $d = 200, 600$.

References

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. (1999). *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301.

Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be" plugged-in". *The Annals of Statistics*, 31(4):1033–1053.

Bobkov, S., Chistyakov, G., and Götze, F. (2024). Strictly subgaussian probability distributions. *Electronic Journal of Probability*, 29:1–28.

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

Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.

Shen, Z. (2020). Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5).

Stoica, P. and Marzetta, T. L. (2001). Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing*, 49(1):87–90.