Supplemental Material for "Transfer Learning with General Estimating Equations"

Notations Throughout the supplementary material, we use c and C with different subscripts to denote generic finite positive constants and may be different in different uses. The empirical measure is denoted as $\mathbb{E}_n(\cdot)$. We use $\mathbb{1}(\mathcal{A})$ as the indicator function of an event $\mathcal{A}$. For any vector $\boldsymbol{v} = (v_1, \cdots, v_d)^{\mathrm{T}}$, let $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$ and $\|\boldsymbol{v}\|_p$ denote its L^p norm. For a function $f : \mathcal{X} \to \mathbb{R}$, its supreme is denoted by $\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$, and its L_p-norm under a distribution F that generates a random variable X is denoted by $\|f\|_{L_p(F)} = (\mathbb{E}_F|f(X)|^p)^{1/p}$ for any $p \geqslant 1$. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leqslant Cb_n$. Let $\mathrm{Pdim}(F)$ be the the Pseudo dimension (Pollard, 1990) of the function class $\mathcal{F}$. The ε-covering number of the function class F with respect to the metric d is denoted as $\mathcal{N}_d(\varepsilon, \mathcal{F})$.

A Proofs for Section 3

A.1 Proof of Theorem 3.1

In the sequel, we use $\mathbb{E}_0$ and $\mathbb{E}_\tau$ to denote the expectation under the true distribution F and the regular parametric submodel F_τ, respectively. The density function for F_τ is

$$f_\tau(\boldsymbol{w}) = p^\delta(1-p)^{1-\delta} f_\tau(\boldsymbol{y}|\boldsymbol{x})^{1-\delta} q_\tau(\boldsymbol{x})^\delta p_\tau(\boldsymbol{x})^{1-\delta},$$

and the score function is given by

$$S_\tau(\boldsymbol{w}) = (1-\delta)S_\tau(\boldsymbol{y}|\boldsymbol{x}) + \delta S_\tau^1(\boldsymbol{x}) + (1-\delta)S_\tau^0(\boldsymbol{x}),$$

where $S_\tau(\boldsymbol{y}|\boldsymbol{x}) = \partial \log f_\tau(\boldsymbol{y}|\boldsymbol{x})/\partial\tau$, $S_\tau^0(\boldsymbol{x}) = \partial \log p_\tau(\boldsymbol{x})/\partial\tau$ and $S_\tau^1(\boldsymbol{x}) = \partial \log q_\tau(\boldsymbol{x})/\partial\tau$, satisfying

$$\mathbb{E}_\tau\{S_\tau(\mathbf{Y}|\boldsymbol{X})|\boldsymbol{X}\} = \mathbf{0},\ \mathbb{E}_\tau\{\delta S_\tau^1(\boldsymbol{X})\} = \mathbf{0} \text{ and } \mathbb{E}_\tau\{(1-\delta)S_\tau^0(\boldsymbol{X})\} = \mathbf{0}. \tag{A.1}$$

(i) Since $\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\} = 0$, differentiating with respect to τ gives

$$\frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\}\Big|_{\tau=0} = \frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r_0)\}\Big|_{\tau=0} + \frac{\partial}{\partial\tau}\mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r(F_\tau))\}\Big|_{\tau=0}. \tag{A.2}$$

Under Condition 2 and the mean-squared differentiability of the submodel F_τ, for any $\boldsymbol{\theta} \in \Theta_0$, the differentiation and integration operators are exchangeable (see, e.g., Ibragimov and Has' Minskii, 1981) and it holds that

$$\frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r_0)\}\Big|_{\tau=0} = \mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W}, \boldsymbol{\theta}, r_0)S_0(\boldsymbol{W})\}. \tag{A.3}$$

We now calculate the right-hand side of (A.2).

$$\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\} = \mathbb{E}_\tau\left\{\frac{1-\delta}{1-p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})r(F_\tau)\right\} = \mathbb{E}_\tau\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\}.$$

Differentiating with respect to τ gives

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}\Big|_{\tau=0} &= \frac{\partial}{\partial\tau}\mathbb{E}_\tau\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})\right\}\Big|_{\tau=0}\\ &= \mathbb{E}_0\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})S_0(\boldsymbol{W})\right\}\\ &= \mathbb{E}_0\left\{\frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})S_0(\boldsymbol{X}) + \frac{\delta}{p}\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta})S_0(\mathbf{Y}|\boldsymbol{X})\right\}\\ &= \mathbb{E}_0\left\{\frac{\delta}{p}\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})S_0(\boldsymbol{W})\right\} + \mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\mathbf{Y}|\boldsymbol{X})\right\},\end{aligned} \tag{A.4}$$

where the first term of (A.4) is from (A.1) and iterated expectation. We proceed to find a function $\mathbf{h}(\boldsymbol{W})$ such that the second term is equivalent to $\mathbb{E}_0\{\mathbf{h}(\boldsymbol{W})S_0(\boldsymbol{W})\}$. Note that

$$\begin{aligned}\mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\mathbf{Y}|\boldsymbol{X})\right\} &= \mathbb{E}_0\left[\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)\{S_0(\boldsymbol{W})-(1-\delta)S_0(\boldsymbol{X})\}\right]\\ &= \mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\boldsymbol{W})\right\} - \mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\boldsymbol{X})\right\},\end{aligned}$$

and the second term is equivalent to

$$\mathbb{E}_0\left\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r_0)S_0(\boldsymbol{X})\right\} = \mathbb{E}\left\{\frac{1-\delta}{1-p}r_0(\boldsymbol{X})\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})S_0(\boldsymbol{X})\right\} = \mathbb{E}\left\{\frac{1-\delta}{1-p}r_0(\boldsymbol{X})\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta})S_0(\boldsymbol{W})\right\}, \tag{A.5}$$

where the first equality is by the iterated expectation, and the second equality is because of (A.1). Combining (A.2)-(A.5) gives

$$\frac{\partial}{\partial\tau}\mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}\Big|_{\tau=0} = \mathbb{E}_0\left\{\boldsymbol{\varphi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}_0)S_0(\boldsymbol{W})\right\},$$

where $\boldsymbol{\eta}_0(\boldsymbol{x}) = (r_0(\boldsymbol{x}),\mathbf{m}_0(\boldsymbol{x}))$ and

$$\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{\delta}{p}\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}) - \frac{1-\delta}{1-p}r(\boldsymbol{x})\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}),$$

It is straightforward to see that $\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}_0)\} = \mathbf{0}$ for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_0$. In addition, because the set of score functions is dense in $L_2(F)$, the influence function $\boldsymbol{\varphi}$ is uniquely determined.

(ii) Let $\boldsymbol{\Psi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}) = \tilde{\mathbf{g}}(\boldsymbol{w},\boldsymbol{\theta},r) + \boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta})$. Since $\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}_0)\} = \mathbf{0}$, replacing F by F_τ gives $\mathbb{E}_\tau\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\} = \mathbf{0}$. Differentiating this identity with respect to $\tau = 0$ gives

$$\mathbf{0} = \frac{\partial}{\partial\tau}\mathbb{E}_\tau\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial\tau}\mathbb{E}_\tau\left\{\boldsymbol{\varphi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}_0)\right\} + \frac{\partial}{\partial\tau}\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\} && \text{(A.6)}\\
&= \mathbb{E}_0\left\{\boldsymbol{\varphi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}_0)S_0(\boldsymbol{W})\right\} + \frac{\partial}{\partial\tau}\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\\
&= \frac{\partial}{\partial\tau}\mathbb{E}_0\{\tilde{\mathbf{g}}(\boldsymbol{W},\boldsymbol{\theta},r(F_\tau))\}\Big|_{\tau=0} + \frac{\partial}{\partial\tau}\mathbb{E}_0\{\boldsymbol{\varphi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0} && \text{(A.7)}\\
&= \frac{\partial}{\partial\tau}\mathbb{E}_0\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\eta}(F_\tau))\}\Big|_{\tau=0},
\end{aligned}
$$

where (A.6) is from differentiation by parts and (A.7) is from the result in (i).

(iii) First, $\boldsymbol{\Psi}$ can be rewritten as

$$\boldsymbol{\Psi}(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{\delta}{p}\mathbf{g}(\boldsymbol{z},\boldsymbol{\theta}) + \left\{\frac{1-\delta}{1-p}r(\boldsymbol{x}) - \frac{\delta}{p}\right\}\{\mathbf{g}(\boldsymbol{z},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{x},\boldsymbol{\theta})\}.$$

Because $\mathbb{E}_F\{\delta\mathbf{g}(\boldsymbol{Z},\boldsymbol{\theta}_0)\} = \mathbf{0}$, we have

$$
\begin{aligned}
\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\} &= \mathbb{E}_F\left[\left\{\frac{1-\delta}{1-p}r(\boldsymbol{X}) - \frac{\delta}{p}\right\}\{\mathbf{g}(\boldsymbol{z},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{X},\boldsymbol{\theta})\}\right]\\
&= \mathbb{E}_F\left[\left\{\frac{1-\delta}{1-p}r(\boldsymbol{X}) - \frac{\delta}{p}\right\}\{\mathbf{m}_0(\boldsymbol{X},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{X},\boldsymbol{\theta})\}\right],
\end{aligned}
$$

implying that $\mathbb{E}_F\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\} = \mathbf{0}$ if either $r(\boldsymbol{x}) \overset{a.e.}{=} r_0(\boldsymbol{x})$ or $\mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}_0) \overset{a.e.}{=} \mathbf{m}_0(\boldsymbol{x},\boldsymbol{\theta}_0)$.

Let $\boldsymbol{\Delta}(\boldsymbol{x},\boldsymbol{\theta}) = \mathbf{m}_0(\boldsymbol{x},\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{x},\boldsymbol{\theta}) = (\Delta_1,\ldots,\Delta_r)^{\mathrm{T}}$. Since $\mathbb{E}_F[\{(1-p)^{-1}(1-\delta)r_0(\boldsymbol{X}) - p^{-1}\delta\}\boldsymbol{\Delta}(\boldsymbol{X},\boldsymbol{\theta})] = \mathbf{0}$, we have

$$
\begin{aligned}
|\mathbb{E}_F\{\Psi_j(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta})\}| &= \left|\mathbb{E}_F\left[\left\{\frac{1-\delta}{1-p}r_0(\boldsymbol{X}) - \frac{1-\delta}{1-p}r(\boldsymbol{X})\right\}\Delta_j(\boldsymbol{X},\boldsymbol{\theta}_0)\right]\right|\\
&= |\mathbb{E}_P[\{r_0(\boldsymbol{X}) - r(\boldsymbol{X})\}\Delta_j(\boldsymbol{X},\boldsymbol{\theta}_0)]|\\
&\leqslant \mathbb{E}_P\{|r_0(\boldsymbol{X}) - r(\boldsymbol{X})||\Delta_j(\boldsymbol{X},\boldsymbol{\theta}_0)|\}\\
&\leqslant \|r - r_0\|_{L_2(P_X)}\|m_j(\cdot,\boldsymbol{\theta}_0) - m_{0j}(\cdot,\boldsymbol{\theta}_0)\|_{L_2(P_X)}, && \text{(A.8)}
\end{aligned}
$$

which completes the proof. □

B Proofs for Section 4

B.1 Proof of Lemma 4.1

According to Fenchel dual representation (Rockafellar, 1997), each convex ϕ can be expressed by:

$$\phi(u) = \sup_{v\in\mathbb{R}}\{uv - \phi_*(v)\}.$$

By the definition of $D_\phi(Q\|P)$, we have

$$D_\phi(Q\|P) = \int\phi\left(\frac{q_0(x)}{p_0(x)}\right)p_0(\boldsymbol{x})d\boldsymbol{x}$$

$$
\begin{aligned}
&= \int \sup_{v(\boldsymbol{x})} \left(v(\boldsymbol{x}) \frac{q_0(x)}{p_0(x)} - \phi_*(v(\boldsymbol{x})) \right) p_0(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sup_{v} \int \{v(\boldsymbol{x}) q_0(\boldsymbol{x}) - \phi_*(v(\boldsymbol{x})) p_0(\boldsymbol{x})\} d\boldsymbol{x} \\
&\geqslant \sup_{v} \mathbb{E}_Q\{v(\boldsymbol{X})\} - \mathbb{E}_P\{\phi_*(v(\boldsymbol{X}))\},
\end{aligned}
$$

where the supremum in the last two equality is taken over all measurable functions from $\mathcal{X} \to \mathrm{dom}(\phi_*)$. Since for each fixed $\boldsymbol{x}$ in the third equality, $v(\boldsymbol{x})q_0(\boldsymbol{x}) - \phi_*(v(\boldsymbol{x}))p_0(\boldsymbol{x})$ is maximized at $v_*(\boldsymbol{x}) = \phi_*^{-1}(q_0(\boldsymbol{x})/p_0(\boldsymbol{x})) = \phi_*^{-1}(r_0(\boldsymbol{x}))$. By the convex duality theorem, we have $v_*(\boldsymbol{x}) = \phi'(r_0(\boldsymbol{x}))$. Therefore,

$$
\phi'(r_0) = \arg\max_{v} [\mathbb{E}_Q\{v(\boldsymbol{X})\} - \mathbb{E}_P\{\phi_*(v(\boldsymbol{X}))\}],
$$

which implies that

$$
r_0 = \arg\min_{r} [\mathbb{E}_P\{\ell_{1,\phi}(r) - \mathbb{E}_Q\{\ell_{2,\phi}(r)\}\}],
$$

where the arg min is taken over all nonnegative functions with the domain $\mathcal{X}$. □

B.2 Proof of Theorem 4.1

Our proof proceeds in several steps. In Step 1, we present an error decomposition for $\|\widehat{r} - r_0\|^2_{L_2(P)}$. In Steps 2 - 4, we investigate the deviations between the sample and population excess risks via empirical process theories. Finally, we bound the empirical estimation error by the L_2 error in Step 5. Throughout the proof, we assume $M_1 \geqslant B_1$ without loss of generality. For any $r \in \mathcal{F}_N$, we define its empirical error as $\|r - r_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n (r(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i))^2$.

Step 1: Error decomposition. Denote $\ell_1(r, \boldsymbol{x}) = \phi_*\{\phi'(r(\boldsymbol{x}))\}, \ell_2(r, \boldsymbol{x}) = -\phi'(r(\boldsymbol{x}))$. Let $\mathcal{L}_1(r) = \mathbb{E}_P\{\ell_1(r, \boldsymbol{X})\}$, $\mathcal{L}_2(r) = \mathbb{E}_Q\{\ell_2(r, \boldsymbol{X})\}$, and $\widehat{\mathcal{L}}_1(r) = n^{-1}\sum_{i=1}^n \ell_1(r, \boldsymbol{X}_i)$, $\widehat{\mathcal{L}}_2(r) = m^{-1}\sum_{i=n+1}^{n+m} \ell_2(r, \boldsymbol{X}_i)$. The population and the sample criterion function are:

$$
\mathcal{L}(r) := \mathcal{L}_1(r) + \mathcal{L}_1(r) \quad \text{and} \quad \widehat{\mathcal{L}}(r) := \widehat{\mathcal{L}}_1(r) + \widehat{\mathcal{L}}_2(r).
$$

For any $r_1, r_2 : \mathcal{X} \to [0, \infty)$, let

$$
d_\phi(r_1, r_2) := \mathcal{L}_\phi(r_1) - \mathcal{L}_\phi(r_2) \quad \text{and} \quad \widehat{d}_\phi(r_1, r_2) = \widehat{\mathcal{L}}_\phi(r_1) - \widehat{\mathcal{L}}_\phi(r_2).
$$

Given a function class $\mathcal{F}_N$, we define the best approximation for r_0 realized by $\mathcal{F}_N$ and the corresponding approximation error as:

$$
r_N := \arg\min_{r \in \mathcal{F}_N} \|r - r_0\|_\infty \quad \text{and} \quad \varepsilon_N := \|r_N - r_0\|_\infty.
$$

Note that r_N and ε_N are both deterministic and depend only on the architecture of $\mathcal{F}_N$ and the target function r_0.

By the compactness of $\boldsymbol{X}$ and Condition 4, it can be shown that there exists a positive constant L, such that for every $r, r' \in \mathcal{F}_N$,

$$|\ell_i(r, \boldsymbol{x}) - \ell_i(r', \boldsymbol{x})| \leqslant L|r(\boldsymbol{x}) - r'(\boldsymbol{x})|, \quad (i = 1, 2),$$

for all $\boldsymbol{x} \in \mathcal{X}$, and there exists positive constants c_1 and c_2 such that

$$c_1 \|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant \mathcal{L}_i(r) - \mathcal{L}_i(r_0) \leqslant c_2 \|\widehat{r} - r_0\|_{L_2(P)}^2, \quad (i = 1, 2), .$$

Therefore, we have the following error decomposition:

$$c_1 \|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant d_\phi(\widehat{r}, r_0) = d_\phi(\widehat{r}, r_N) + d_\phi(r_N, r_0) \leqslant d_\phi(\widehat{r}, r_N) + c_2 \varepsilon_N^2. \tag{B.1}$$

We next bound $d_\phi(\widehat{r}, r_N)$ by analyzing the process $\sup_{r \in \mathcal{F}_N} |d_\phi(r, r_N) - \widehat{d}_\phi(r, r_N)|$, mainly based on techniques of the local Rademacher complexity analysis of empirical risk minimization (Bartlett et al., 2005 and Koltchinskii, 2011). First, we introduce some quantities that are necessary in this approach. Let $\{\varepsilon_i\}_{i=1}^{n+m}$ be i.i.d symmetric, $\{-1, 1\}$-valued random variables that are independent of $\{\boldsymbol{X}_i\}_{i=1}^{n+m}$. For any function class $\mathcal{F}$, we define

$$\mathcal{R}_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(\boldsymbol{X}_i), \quad \mathcal{R}_m(\mathcal{F}) := \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i f(\boldsymbol{X}_i).$$

The Rademacher complexities are defined as $\bar{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}\{\mathcal{R}_n(\mathcal{F})\}$ and $\bar{\mathcal{R}}_m(\mathcal{F}) = \mathbb{E}\{\mathcal{R}_m(\mathcal{F})\}$, where the expectations are taken over both the $\boldsymbol{X}_i$s and the ε_is. The empirical Rademacher complexities, which are conditioned on the data, are denoted by $\widehat{R}_n(\mathcal{F}) = \mathbb{E}_\varepsilon\{R_n(\mathcal{F})\}$ and $\widehat{R}_m(\mathcal{F}) = \mathbb{E}_\varepsilon\{R_m(\mathcal{F})\}$. For the candidate function class $\mathcal{F}_N$, let the shifted (centered) function class be

$$\mathcal{F}_N^* := \{r - r_N : r \in \mathcal{F}_N\}.$$

The population version of the localized Rademacher complexities are defined as:

$$\bar{\mathcal{R}}_n(\delta, \mathcal{F}_N^*) := \bar{\mathcal{R}}_n\left\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_{L_2(P)} \leqslant \delta\right\} \quad \text{and} \quad \bar{\mathcal{R}}_m(\delta, \mathcal{F}_N^*) := \bar{\mathcal{R}}_m\left\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_{L_2(Q)} \leqslant \delta\right\},$$

where $\delta > 0$ is a localization scale. Similarly, the empirical localized Rademacher complexities are defined as:

$$\widehat{\mathcal{R}}_n(\delta, \mathcal{F}_N^*) := \widehat{\mathcal{R}}_n\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_n \leqslant \delta\} \quad \text{and} \quad \widehat{\mathcal{R}}_m(\delta, \mathcal{F}_N^*) := \widehat{\mathcal{R}}_m\{f : f \in \mathcal{F}_N^* \text{ and } \|f\|_m \leqslant \delta\}.$$

A crucial parameter in the localized Rademacher complexity approach is the critical radius, which is defined as δ_n and δ_m that satisfy the following inequalities:

$$\delta_n^2 \geqslant \bar{\mathcal{R}}_n(\delta_n, \mathcal{F}_N^*), \quad \delta_m^2 \geqslant \bar{\mathcal{R}}_m(\delta_m, \mathcal{F}_N^*). \tag{B.2}$$

For $j = 1$ and 2, denote the supreme deviations between $\widehat{\mathcal{L}}_j(r) - \widehat{\mathcal{L}}_j(r_0)$ and $\mathcal{L}_j(r) - \mathcal{L}_j(r_0)$ restricted in the localized ball centered at r_0 with the radius s as

$$\lambda_N^j(s) = \sup_{\|r - r_N\|_{L_2(P)} \leqslant s} \left| \left(\widehat{\mathcal{L}}_j(r) - \widehat{\mathcal{L}}_j(r_N)\right) - (\mathcal{L}_j(r) - \mathcal{L}_j(r_N)), \right| \tag{B.3}$$

and denote the supreme deviations between $d_\phi(r, r_N)$ and $\widehat{d}_\phi(r, r_N)$ restricted in $d_\phi(r, r_N)$ as

$$\lambda_N(s) = \sup_{\|r - r_N\|_{L_2(P)} \leqslant s} \left| \widehat{d}_\phi(r, r_N) - d_\phi(r, r_N) \right|, \tag{B.4}$$

where $s > 0$ is a radius to be varied.

Step 2. Tail bound of $\lambda_N(s)$. We first estimate an upper bound of the expectation of $\lambda_N(s)$ for the s in the range $[\delta_n \vee \delta_m, \infty)$. Let

$$\mathcal{G}_N^j(s) = \left\{ g : g = \ell_j(r) - \ell_j(r_0) \ \text{ for } r \in \mathcal{F}_N \text{ and } d_\phi(r, r_0) \leqslant s^2 \right\}$$

for $j = 1$ and 2. Then by standard symmetrization arguments, we have

$$\mathbb{E}\left\{\lambda_N^1(s)\right\} \leqslant 2\bar{\mathcal{R}}_n\left\{\mathcal{G}_N^1(s)\right\} \quad \text{and} \quad \mathbb{E}\left\{\lambda_N^2(s)\right\} \leqslant 2\bar{\mathcal{R}}_m\left\{\mathcal{G}_N^2(s)\right\}. \tag{B.5}$$

Since both $\phi_* \circ \phi'$ and ϕ' are L-Lipschitz continuous, by the Ledoux-Talagrand contraction inequality due to Ledoux and Talagrand (1991), it holds that $\bar{\mathcal{R}}_n\left\{\mathcal{G}_N^1(s)\right\} \leqslant 2L\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*)$ and $\bar{\mathcal{R}}_m\left\{\mathcal{G}_N^2(s)\right\} \leqslant 2L\bar{\mathcal{R}}_m(s, \mathcal{F}_N^*)$. Therefore,

$$\mathbb{E}\left\{\lambda_N^1(s)\right\} \leqslant 4L\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*) \quad \text{and} \quad \mathbb{E}\left\{\lambda_N^2(s)\right\} \leqslant 4L\bar{\mathcal{R}}_m(s, \mathcal{F}_N^*).$$

Since $\mathcal{F}_N^*$ is star-shaped around r_N (if $r \in \mathcal{F}_N^*$, then for any $\alpha \in (0,1)$, $\alpha r \in \mathcal{F}_N^*$), the function $\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*)/s$ is non-increasing with resepct to s according to Lemma 13.6 of Wainwright (2019). As $s > \delta_n$ and $\delta_n^2 > \bar{\mathcal{R}}_n\{\delta_n, \mathcal{F}_N^*\}$, it holds that $\bar{\mathcal{R}}_n(s, \mathcal{F}_N^*) \leqslant s\delta_n$. Similarly, we also have $\bar{\mathcal{R}}_m(s, \mathcal{F}_N^*) \leqslant s\delta_m$ for $s \geqslant \delta_m$, which delivers the upper bounds

$$\mathbb{E}\left\{\lambda_N^1(s)\right\} \leqslant 4Ls\delta_n \quad \text{and} \quad \mathbb{E}\left\{\lambda_N^2(s)\right\} \leqslant 4Ls\delta_m \quad (\forall s \geqslant \delta_n \vee \delta_m). \tag{B.6}$$

We next bound the deviation between $\lambda_N^j(s)$ and $\mathbb{E}\left\{\lambda_N^j(s)\right\}$ for $j = 1$ and 2. Note that for any $r \in \mathcal{F}_N$, we have $\|\ell_j(r) - \ell_j(r_N)\|_\infty \leqslant L\|r - r_N\|_\infty \leqslant 2M_1L$, by the Lipschitz condition of $\phi_* \circ \phi'$ and ϕ' and the boundness of $r \in \mathcal{F}_N$. In addition, the variance of $\ell_j(r) - \ell_j(r_N)$ can be upper bounded by

$$\begin{aligned}
\operatorname{Var}(\ell_j(r) - \ell_j(r_N)) \leqslant & \mathbb{E}\left\{(\ell_j(r) - \ell_j(r_N))^2\right\} \\
\leqslant & L^2\left(\|r - r_N\|_{L_2(P)}^2 + \|r - r_N\|_{L_2(Q)}^2\right) \\
\leqslant & 2(M_1L)^2\|r - r_N\|_{L_2(P)}^2 \leqslant 2(M_1Ls)^2,
\end{aligned} \tag{B.7}$$

where the second inequality is implied by the Lipschitz condition, the third inequality is due to $\|f\|_{L_2(Q)}^2 = \|f \cdot r_0\|_{L_2(P)}^2 \leqslant B^2\|f\|_{L_2(P)}^2$ for any $f : \mathcal{X} \to \mathbb{R}$, and the last inequality is because of the localization condition $\|r - r_N\|_{L_2(P)} \lesssim d_\phi(r, r_N) \leqslant s$. Consequently, for any $u > 0$ it holds that

$$\mathbb{P}\left\{\lambda_N^j(s) \geqslant \mathbb{E}\{\lambda_N^j(s)\} + u\right\} \leqslant 2\exp\left(\frac{-(n \wedge m)u^2}{8e\operatorname{Var}(\ell_\phi(r) - \ell_\phi(r_0)) + 8M_1Lu}\right)$$

$$\leqslant 2\exp\left(-\frac{C_t(n \wedge m)u^2}{(M_1Ls)^2 + M_1Lu}\right),$$

for some universal constant $C_t > 0$, by applying Talagrand's concerntration equality (Talagrand, 1994) and (B.7). Therefore, we have

$$\left(\mathbb{P}\left\{\lambda_N^1(s) \geqslant 4Ls\delta_n + u\right\} \vee \mathbb{P}\left\{\lambda_N^2(s) \geqslant 4Ls\delta_m + u\right\}\right) \leqslant 2\exp\left(-\frac{C_t(n \wedge m)u^2}{(M_1Ls)^2 + M_1Lu}\right),$$

for any $s \geqslant (\delta_n \vee \delta_m)$ and $u > 0$. Since

$$\lambda_N(s) \leqslant \lambda_N^1(s) + \lambda_N^2(s)$$

for any $s \geqslant 0$, we have

$$\mathbb{P}\left\{\lambda_N(s) \geqslant 4Ls(\delta_n + \delta_m) + u\right\} \leqslant 4\exp\left(-\frac{C_t(n \wedge m)u^2}{(2M_1Ls)^2 + 2M_1Lu}\right), \tag{B.8}$$

for any $s \geqslant (\delta_n \vee \delta_m)$ and $u > 0$. Denoting $\delta_N := \delta_n + \delta_m$ and setting $s = \delta_N, u = M_1L\delta_N^2$, then we have

$$\mathbb{P}\left\{\lambda_N(\delta_N) \geqslant C_1\delta_N^2\right\} \leqslant 4\exp\left(-C_2(n \wedge m)\delta_N^2\right), \tag{B.9}$$

where $C_1 = (4 + M_1)L$ and $C_2 = C_t/6$. In addition, setting $u = M_1Ls\delta_N$ yields

$$\mathbb{P}\left\{\lambda_N(s) \geqslant C_1s\delta_N\right\} \leqslant 2\exp\left(-\frac{C_tns^2\delta_N^2}{s^2 + s\delta_N}\right) \leqslant 4\exp\left(-C_2(n \wedge m)\delta_N^2\right), \tag{B.10}$$

for any $s \geqslant \delta_N$.

Let

$$\mathcal{A}_1 = \left\{\exists\, r \in \mathcal{F}_N : \|r - r_N\|_{L_2(P)} \leqslant \delta_N \quad \text{and} \quad \left|\widehat{d}_\phi(r, r_N) - d_\phi(r, r_N)\right| \geqslant C_1\delta_N^2\right\}. \tag{B.11}$$

Combining (B.4) with (B.9) yields that

$$\mathbb{P}(\mathcal{A}_1) \leqslant 4\exp\left(-C_2(n \wedge m)\delta_N^2\right). \tag{B.12}$$

The above tail bound (B.10) controls the largest deviation $\left|\widehat{d}_\phi(r, r_N) - d_\phi(r, r_N)\right|$ for r within the local ball $\|r - r_N\|_{L_2(P)} \leqslant \delta_N$. It remains to estimate an tail bound of the deviation $\left|\widehat{d}_\phi(r, r_N) - d_\phi(r, r_N)\right|$ outside this local region. We define the following event

$$\mathcal{A}_2 = \left\{\exists\, r \in \mathcal{F}_N : \|r - r_N\|_{L_2(P)} > \delta_N \quad \text{and} \quad \left|\widehat{d}_\phi(r, r_N) - d_\phi(r, r_N)\right| \geqslant 2C_1\delta_N\|r - r_N\|_{L_2(P)}\right\}$$

However, bounding $\mathbb{P}(\mathcal{A}_2)$ is more delicate, since the function r that satisfies the requirement in $\mathcal{A}_2$ is random. In the following step, we will use a "peeling" argument to address the problem.

Step 3: Bound the event $\mathcal{A}_2$ with the peeling argument. For $m \in \mathbb{N}_+$, we define the events

$$\mathcal{S}_m := \left\{ r \in \mathcal{F}_N : 2^{m-1}\delta_N < \|r - r_N\|_{L_2(P)} \leqslant 2^m \delta_N \right\}.$$

By the boundness of $r \in \mathcal{F}_N$, we have $\|r - r_N\|_{L_2(P)} \leqslant 2M_1$. Hence, any $r \in \mathcal{F}_N \cap \left\{ \|r - r_N\|_{L_2(P)} > \delta_N \right\}$ must locate in some $\mathcal{S}_m$ for $m \in [\![K]\!]$, where $K \leqslant 2\log(M_1/\delta_N) + 1$. Since $\mathcal{A}_2$ is a subset of $\cup_{m=1}^K S_m$, by the union bound we have $\mathbb{P}(\mathcal{A}_2) \leqslant \sum_{m=1}^M \mathbb{P}\left(\mathcal{A}_2 \cap S_m\right)$.

Note that if $r_m \in \mathcal{A}_2 \cap S_m$, then we can take $s_m = 2^m \delta_N$, and r_m satisfies

$$\|r_m - r_N\|_{L_2(P)} \leqslant s_m \quad \text{and} \quad \left| \widehat{d}_\phi(r_m, r_0) - d_\phi(r_m, r_0) \right| \geqslant 2C_1 \delta_N \|r - r_N\|_{L_2(P)} > C_1 \delta_N s_m,$$

where the last inequality is due to $2\|r - r_N\|_{L_2(P)} > 2^{m+1}\delta_N > s_m = 2^m \delta_N$. As a result, $\mathcal{A}_2 \cap \mathcal{S}_m \subset \left\{ \lambda_N(s_m) \geqslant C_1 s_m \delta_N \right\}$. Then according to (B.10), we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{A}_2) &\leqslant \sum_{m=1}^K \mathbb{P}(\mathcal{A} \cap \mathcal{S}_m) \leqslant 2 \sum_{m=1}^K \exp\left(-C_2(n \wedge m)\delta_N^2\right) \\ &\leqslant 4\exp(-C_2(n \wedge m)\delta_N^2 + \log K) \leqslant 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right), \end{aligned} \tag{B.13}$$

where the last inequality holds provided that

$$\frac{C_2(n \wedge m)\delta_N^2}{2} \geqslant \log\left(2\log(M_1/\delta_N) + 1\right). \tag{B.14}$$

The complement of $\mathcal{A}_2$ is composed by $\mathcal{A}_2^c = \mathcal{B}_1 \cup \mathcal{B}_2$, where

$$\mathcal{B}_1 = \left\{ r \in \mathcal{F}_N : \|r - r_N\|_{L_2(P)} \leqslant \delta_N \right\} \text{ and } \mathcal{B}_2 = \left\{ r \in \mathcal{F}_N : \left| \widehat{d}_\phi(r, r_N) - d_\phi(r, r_N) \right| < 2C_1 \delta_N \|r - r_N\|_{L_2(P)} \right\}.$$

Therefore, (B.13) implies that

$$\mathbb{P}(\mathcal{B}_1 \cup \mathcal{B}_2) \geqslant 1 - 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right).$$

If $\widehat{r} \in \mathcal{B}_1$, then we have $d(\widehat{r}, r_N) \leqslant c_2 \delta_N^2$ since $d(\widehat{r}, r_N) \leqslant c_2 \|\widehat{r} - r_N\|_{L_2(P)}^2$. Moreover, if $\widehat{r} \in \mathcal{B}_2$, since $c_1 \|\widehat{r} - r_N\|_{L_2(P)}^2 \leqslant d_\phi(\widehat{r}, r_N)$, and $\widehat{d}(\widehat{r}, r_N) \leqslant 0$ by the definition of $\widehat{r}$, we have $d_\phi(\widehat{r}, r_N) < 4c_1^{-2} C_1^2 \delta_N^2$. This together with (B.12) leads to

$$\mathbb{P}\left\{ d_\phi(\widehat{r}, r_N) < (c_2 \vee 4c_1^{-2} C_1^2)\delta_N^2 \right\} \geqslant 1 - 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right). \tag{B.15}$$

Let $C_3 = c_2 \vee 4c_1^{-2} C_1^2$ and $C_4 = C_2/2$, combining (B.1) and (B.15), we obtain

$$\mathbb{P}\left\{ c_1 \|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant C_3 \delta_N^2 + c_2 \varepsilon_N^2 \right\} \geqslant 1 - 4\exp\left(-\frac{C_2(n \wedge m)\delta_N^2}{2}\right). \tag{B.16}$$

Therefore, the estimation error $\|\widehat{r}-r_0\|_{L_2(P)}$ relies on the critical radius δ_N and the approximation error ε_N. In the next step, we provide an upper bound of the critical radius δ_N.

Step 4: Estimation of the critical radius δ_N. In this step, we first estimate the empirical critical radiuses $\widehat{\delta}_n$ and $\widehat{\delta}_m$ satisfying

$$\widehat{\delta}_n^2 \geqslant k\widehat{\mathcal{R}}_n(\widehat{\delta}_n, \mathcal{F}_N^*), \quad \widehat{\delta}_m^2 \geqslant k\widehat{\mathcal{R}}_m(\widehat{\delta}_m, \mathcal{F}_N^*), \tag{B.17}$$

where k is a fixed positive constant, $\widehat{\mathcal{R}}_n(\delta_n, \mathcal{F}_N^*)$ and $\widehat{\mathcal{R}}_n(\delta_m, \mathcal{F}_N^*)$ are localized empirical Rademacher complexities, respectively, then use Proposition 14.25 of Wainwright (2019) to obtain that

$$\mathbb{P}(C_4\delta_n \leqslant \widehat{\delta}_n \leqslant C_5\delta_n) \geqslant 1 - C_6\exp(-C_7 n\delta_n^2) \tag{B.18}$$

for some generic constants $C_4, \cdots, C_7 > 0$.

By the Dudley's chaining, we have

$$\widehat{\mathcal{R}}_n(s, \mathcal{F}_N^*) \leqslant \inf_{0<\alpha<s}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^s \sqrt{\log\left(\mathcal{N}_2(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)d\varepsilon\right)}\right\}, \tag{B.19}$$

where $\boldsymbol{X}_1^n = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$. Since for any $\|f\|_n \leqslant \max_{1\leqslant i\leqslant n}|f(\boldsymbol{X}_i)|$, we have $\mathcal{N}_2(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n) \leqslant \mathcal{N}_\infty(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)$. Since $\|f\|_\infty \leqslant 2M$ for $f \in \mathcal{F}_N^*$, according to Theorem 12.2 of Anthony and Bartlett (1999), we have

$$\log\left(\mathcal{N}_\infty(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)\right) \leqslant \mathrm{Pdim}(\mathcal{F}_N^*)\left(\frac{4eMn}{\varepsilon \mathrm{Pdim}(\mathcal{F}_N^*)}\right).$$

When $n > \mathrm{Pdim}(\mathcal{F}_N^*)$, let $\alpha = s\sqrt{\mathrm{Pdim}(\mathcal{F}_N^*)/n}$ in (B.19), we have

$$\inf_{0<\alpha<s}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^s \sqrt{\log\left(\mathcal{N}_2(\varepsilon, \mathcal{F}_N^*, \boldsymbol{X}_1^n)d\varepsilon\right)}\right\} \leqslant 16s\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N^*)}{n}\left(\log\frac{4eM}{s} + \frac{3}{2}\log n\right)}.$$

Therefore, if $s \geqslant 1/n$ and $n \geqslant (4eM)^2$, the localized empirical Rademacher complexity can be upper bounded by

$$\widehat{\mathcal{R}}_n(s, \mathcal{F}_N^*) \leqslant 32s\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N^*)}{n}\log(n)}.$$

With such result, we find that the $\widehat{\delta}_n$ satisfying $\widehat{\delta}_n^2 \geqslant \widehat{\mathcal{F}}_n(\widehat{\delta}_n, \mathcal{F}_N^*)$ can be taken as

$$\widehat{\delta}_n = 32k\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N^*)}{n}\log(n)} + u = 32k\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_N)}{n}\log(n)} + u, \tag{B.20}$$

for any $u \geqslant 0$. The empirical critical value $\hat{\delta}_m$ can be taken similarly. Using (B.16), (B.18), and (B.20), we obtain that for any $u \geqslant 0$,

$$\mathbb{P}\left\{\|\widehat{r}-r_0\|_{L_2(P)}^2 \leqslant C_8\left(\xi_N + \epsilon_N^2 + u\right)\right\} \geqslant 1 - C_9\exp(-N\xi_N - Nu), \tag{B.21}$$

for some universal constants C_8 and $C_9 > 0$, where ξ_N represents the stochastic error in the estimation and is defined as

$$\xi_N = \operatorname{Pdim}(\mathcal{F}_N)\left(\frac{\log(n)}{n} + \frac{\log(m)}{m}\right).$$

Since $N\xi_N \asymp \operatorname{Pdim}(\mathcal{F}_N)\log(N)$, we have $\exp(-N\xi_N) < C_9^{-1}$ for large enough N. Therefore, (B.21) implies that for large enough N and any $t \geqslant 0$, it holds that

$$\mathbb{P}\left\{\|\widehat{r} - r_0\|_{L_2(P)}^2 \leqslant C_8\left(\xi_N + \epsilon_N^2 + \frac{t}{N}\right)\right\} \geqslant 1 - \exp(-t).$$

Step 5: Bound the empirical error by the L_2 error.

In this step, we show that with high probability, the empirical error $\|\widehat{r} - r\|_n$ is at most twice the L_2 error if r is in a given neighboring ball around r_0.

Let $g(r) = (r - r_0)^2$ for every $r \in \mathcal{F}_N$. Then since $g(r) = (r + r_0)(r - r_0)$, we have $|g(r)| \leqslant 3M_1|r - r_0| \geqslant 9M_1^2$, implying that $g(r)$ has a Lipschitz constant of $3M_1$, and $g(r)$ is a bounded function. Furthermore, if r is restricted to a radius with $\|r - r_0\|_{L_2(P)} \leqslant \xi$ for some fixed constant $\xi > 0$, then

$$\operatorname{Var}\{g(r)\} \leqslant \mathbb{E}\{g^2(r)\} \leqslant \mathbb{E}\{(r - r_0)^4\} \leqslant 9M_1^2\xi^2.$$

By applying Theorem 2.1 of Bartlett et al. (2005), which is based on Talagrand's concentration, for every r with $\|r - r_0\|_{L_2(P)} \leqslant \xi$, it holds that

$$\begin{aligned}\|r - r_0\|_n^2 - \|r - r_0\|_{L_2(P)} &\leqslant 3\widehat{\mathcal{R}}_n(g(r) : r \in \mathcal{F}_N, \|r - r_0\|_{L_2(P)} \leqslant \xi) + 3M_1\xi\sqrt{\frac{2t}{n}} + \frac{12M_1^2t}{n} \\ &\leqslant 18M_1\widehat{\mathcal{R}}_n(\xi, \mathcal{F}_N^*) + 3M_1\xi\sqrt{\frac{2t}{n}} + \frac{12M_1^2t}{n}, \qquad \text{(B.22)}\end{aligned}$$

with probability at least $1 - e^{-t}$ where the second inequality is due to $(r - r_0) \in \mathcal{F}_N^*$, the Lipschitz continuity of $g(r)$, and iterated expectations.

Now, suppose that the radius ξ satisfies

$$\xi^2 \geqslant 36M_1\widehat{\mathcal{R}}_n(\xi, \mathcal{F}_N^*), \quad \text{and} \quad \xi^2 \geqslant \frac{72M_1^2t}{n}, \qquad \text{(B.23)}$$

then (B.22) implies that with probability at least $1 - e^{-t}$,

$$\|r - r_0\|_n^2 \leqslant \xi^2/2 + \xi^2/2 + \xi^2/6 < 2\xi^2 \text{ for all } r \text{ satisfies (B.23) and } \|r - r_0\|_{L_2(P)} \leqslant \xi.$$

As shown in the calculation of the previous step, for large enough n,

$$\xi = C_8(\xi_N + \epsilon_N^2 + \frac{t}{N})$$

satisfies the requirement in (B.23) for any given $t > 0$. This together with $\mathbb{P}(\|r - r_0\|_{L_2(P)} \leqslant \xi) > 1 - e^{-t}$ implies that $\mathbb{P}(\|r - r_0\|_n^2 \leqslant \xi) > 1 - 2e^{-t}$, which completes the proof of Theorem 4.1. □

B.3 Proof of Theorem 4.2

We will apply Yang-Barron's version of Fano's method (Yang and Barron, 1999) to derive the lower bound for the density ratio estimation.

Part 1. Let us first consider a sub-class of $\mathcal{M}^d(\beta_1, B_1)$ defined by

$$\mathcal{M}_1 = \left\{(\mathbb{P}_0, \mathbb{Q}) : \mathbb{P}_0 \text{ is the uniform distribution }, \ d\mathbb{Q}/d\mathbb{P} \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1), \ \inf_{x\in\mathcal{X}} d\mathbb{Q}(x) > c_0 > 0\right\}.$$

Then for any two distinct elements $(\mathbb{P}_0, \mathbb{Q}_1)$ and $(\mathbb{P}_0, \mathbb{Q}_2)$ in $\mathcal{M}_1$, their KL-divergence $D\left((\mathbb{P}_0, \mathbb{Q}_1)\|(\mathbb{P}_0, \mathbb{Q}_2)\right)$ can be bounded by

$$\begin{aligned} D\left((\mathbb{P}_0, \mathbb{Q}_1)\|(\mathbb{P}_0, \mathbb{Q}_2)\right) =& D(\mathbb{Q}_1\|\mathbb{Q}_2) = \int_{x\in\mathcal{X}} \log\left(\frac{d\mathbb{Q}_1(x)}{d\mathbb{Q}_2(x)}\right) d\mathbb{Q}_1(x) \\ \leqslant& \int_{x\in\mathcal{X}} \left(\frac{d\mathbb{Q}_1(x)}{d\mathbb{Q}_2(x)} - 1\right) d\mathbb{Q}_1(x) = \int_{x\in\mathcal{X}} \left(\frac{d\mathbb{Q}_1(x)}{d\mathbb{Q}_2(x)}\right)^2 d\mathbb{Q}_2(x) - 1 \\ =& \int_{x\in\mathcal{X}} \left(\frac{(d\mathbb{Q}_1(x) - d\mathbb{Q}_2(x))}{d\mathbb{Q}_2(x)}\right)^2 d\mathbb{Q}_2 \\ \leqslant& c_0^{-1} \int_{x\in\mathcal{X}} \left(d\mathbb{Q}_1(x) - d\mathbb{Q}_2(x)\right)^2 dx. \end{aligned} \tag{B.24}$$

The above bound together with $D\left((\mathbb{P}_0^{\otimes n}, \mathbb{Q}_1^{\otimes m})\|(\mathbb{P}_0^{\otimes n}, \mathbb{Q}_2^{\otimes m})\right) = mD(\mathbb{Q}_1\|\mathbb{Q}_2)$ implies that for any $\varepsilon > 0$, the ε-covering number of $\mathcal{M}_1$ in the square-root KL divergence has an upper bound:

$$\mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_1) \leqslant \mathcal{N}_{L_2(\mu)}\left(\sqrt{\frac{c_0}{m}}\varepsilon, \mathcal{Q}_1\right),$$

where $\mathcal{Q}_1$ is the function class of $\mathbb{Q}$ that is the second element of $(\mathbb{P}, \mathbb{Q}) \in \mathcal{M}_1$. By definition, we know that $\mathcal{Q}_1$ is a sub-class of $\mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$, whose covering number is known from classical theory (see e.g., Giné and Nickl, 2021). Therefore we obtain

$$\log \mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_1) \leqslant \log \mathcal{N}_{L_2(\mu)}\left(\sqrt{\frac{c_0}{m}}\varepsilon, \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)\right) \asymp \left(\frac{B\sqrt{m}}{\varepsilon}\right)^{\frac{d}{\beta}}. \tag{B.25}$$

Applying Yang-Barron's version of Fano's method, we choose $(\varepsilon_n, \delta_n)$ that satisfies

$$\varepsilon_m^2 \geqslant \mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_1) \quad \text{and} \quad \log M(2\delta_m; d, \Theta) \geqslant 4\varepsilon_m^2 + \log 2. \tag{B.26}$$

Since the estimand is the density ratio function that belongs to $\mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$, we have

$$\log M(2\delta_m; d, \Theta) \asymp (\frac{1}{\delta_m})^{\frac{d}{\beta}}. \tag{B.27}$$

With (B.25) and (B.27), $(\varepsilon_n, \delta_n)$ that ensures (B.26) can be specified as $\varepsilon_m^2 \asymp m^{\frac{d}{2\beta+d}}$ and $\delta_m^2 \asymp m^{-\frac{2\beta_1}{2\beta_1+d}}$.

According to Yang and Barron (1999), a minimax lower bound for the sub-class $\mathcal{M}_1$ is given by

$$\inf_{\widehat{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}_1} \mathbb{E}\|\widehat{r} - d\mathbb{Q}/d\mathbb{P}\|^2 \geqslant \frac{\delta_m^2}{2} \asymp m^{-\frac{2\beta_1}{2\beta_1+d}}. \tag{B.28}$$

Part 2. Let us first consider another sub-class of $\mathcal{M}^d(\beta_1, B_1)$ defined by

$$\mathcal{M}_2 = \left\{(\mathbb{P}, \mathbb{Q}_0) : \mathbb{Q}_0 \text{ is the uniform distribution },\ d\mathbb{Q}_0/d\mathbb{P} \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1),\ 0 < c_1 < d\mathbb{P}(x) < c_2 < \infty\right\}.$$

For any two distinct elements $(\mathbb{P}_1, \mathbb{Q}_0)$ and $(\mathbb{P}_2, \mathbb{Q}_0)$ in $\mathcal{M}_2$, with the same argument as in (B.24), we can obtain

$$D\left((\mathbb{P}_0, \mathbb{Q}_1)\|(\mathbb{P}_0, \mathbb{Q}_2)\right) \leqslant c_1^{-1}\int_{x\in\mathcal{X}} \left(d\mathbb{P}_1(x) - d\mathbb{P}_2(x)\right)^2 dx.$$

Since $d\mathbb{Q}_0(x) = 1$, we write $dP_i(x) = r_i^{-1}(x)$ with $r_i(x) \in \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$ for $i = 1, 2$. Then the above quantity can be upper bounded by

$$\begin{aligned} c_1^{-1}\int_{x\in\mathcal{X}} \left(d\mathbb{P}_1(x) - d\mathbb{P}_2(x)\right)^2 dx =& c_1^{-1}\int_{x\in\mathcal{X}} \left(\frac{1}{r_1(x)} - \frac{1}{r_2(x)}\right)^2 dx \\ \leqslant& c_2^4 c_1^{-1}\int_{x\in\mathcal{X}} (r_1(x) - r_2(x))^2 dx. \end{aligned} \tag{B.29}$$

Therefore, the square-root covering number of $\mathcal{M}_2$ in KL-divergence can be upper bounded by the covering number of $\mathcal{H}^{\beta_1}(\mathcal{X}, B_1)$ in the $L_2(\mu)$-norm, leading to

$$\log \mathcal{N}_{\mathrm{KL}}(\varepsilon, \mathcal{M}_2) \leqslant \log \mathcal{N}_{L_2(\mu)}\left(\sqrt{\frac{c_1}{c_2^4 n}}\varepsilon, \mathcal{H}^{\beta_1}(\mathcal{X}, B_1)\right) \asymp \left(\frac{B\sqrt{n}}{\varepsilon}\right)^{\frac{d}{\beta}}. \tag{B.30}$$

for any $\varepsilon > 0$. The rest procedure is similar to Part I and we omit here for simplicity. The conclusion is for the sub-class $\mathcal{M}_2$, a minimiax lower bound is given by

$$\inf_{\widehat{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}_2} \mathbb{E}\|\widehat{r} - d\mathbb{Q}/d\mathbb{P}\|^2 \geqslant \frac{\delta_n^2}{2} \asymp n^{-\frac{2\beta_1}{2\beta_1+d}}. \tag{B.31}$$

Since $\mathcal{M}_1$ and $\mathcal{M}_2$ are both sub-class of $\mathcal{M}^d(\beta_1, B_1)$, their minimax lower bounds are also lower bounds of $\mathcal{M}^d(\beta_1, B_1)$. Combining the results in Part I and II, we obtain:

$$\inf_{\widehat{r}} \sup_{(\mathbb{P},\mathbb{Q})\in\mathcal{M}^d(\beta_1,B_1)} \mathbb{E}\|\widehat{r} - d\mathbb{Q}/d\mathbb{P}\|^2 \gtrsim n^{-\frac{2\beta_1}{2\beta_1+d}} + m^{-\frac{2\beta_1}{2\beta_1+d}} \asymp N^{-\frac{2\beta_1}{2\beta_1+d}}, \tag{B.32}$$

which completes the proof of Theorem 4.2. □

B.4 Proof of Theorem 4.3

For any given distribution $\tilde{P}_Y$ supported on $\mathbb{R}$ with a known density $\tilde{p}_0(y)$, we let $\tilde{P} = \tilde{P}_Y \times P_{\boldsymbol{X}}$ be the distribution of $(\tilde{Y}, \boldsymbol{X})$ for $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$ and $Y \sim \tilde{P}_Y$, which is independent of $\boldsymbol{X}$, and let

$$\tilde{r}_0(y, \boldsymbol{x}) = \frac{p_0(y, \boldsymbol{x})}{p_0(\boldsymbol{x})\tilde{p}_0(y)},$$

be the true density ratio function between P and $\tilde{P}$. Then, under Conditions 5 and 6, applying Theorem 4.2 leads to

$$\mathbb{E}_N\{(\hat{\hat{r}} - \tilde{r}_0)^2\} = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}}\log(N)\right),$$

for the estimator $\hat{\hat{r}}$. Since $\hat{p}_{Y|\boldsymbol{X}} = \hat{\hat{r}}\tilde{p}_Y$ and $p_{Y|\boldsymbol{X}}(y, \boldsymbol{x}) = \tilde{r}_0\tilde{p}_Y$, where $\tilde{p}_Y$ is a bounded function, we have

$$\mathbb{E}_N\{(\hat{p}_{Y|\boldsymbol{X}} - p_{Y|\boldsymbol{X}})^2\} = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}}\log(N)\right). \tag{B.33}$$

For any $\boldsymbol{\theta}$, let $\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \int \mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)dy$ be the conditional mean function with the estimated conditional density function $\hat{p}_{Y|\boldsymbol{X}}$, then

$$\mathbb{E}_N\{\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta}) - \mathbf{m}_0(\boldsymbol{X}, \boldsymbol{\theta})\}^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{X}}\left[\int g(y, \boldsymbol{X}_i, \boldsymbol{\theta})\{\hat{p}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)\}dy\right]^2.$$

Since there exists a constant $c > 0$ such that $p_0(y|\boldsymbol{X}) > c$, we have

$$\begin{aligned}
&\left\{\int |\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})||\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|dy\right\}^2 \\
&\leqslant c^{-1}\left\{\int |\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})||\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|\sqrt{p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)}dy\right\}^2 \\
&\leqslant c^{-1}\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}_i}(y|\boldsymbol{X}_i)|^2dy\int p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)dy \\
&\leqslant c^{-1}\log^2(N)\int |\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}_i}(y|\boldsymbol{X})|^2dy + \\
&\quad + c^{-1}\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 I(\|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 > \log(N))|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|^2dy \\
&=: I_{1i} + I_{2i}, \quad \text{say.}
\end{aligned} \tag{B.34}$$

Note that as $\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)$ and $p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)$ are uniformaly bounded by a constant $M > 0$, we have

$$|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X})|^2 \leqslant 4M^2 + Mp_0(y|\boldsymbol{X}_i) \leqslant (4M^2m^{-1} + M)p_0(y|\boldsymbol{X}_i). \tag{B.35}$$

Hence, I_{2i} can be bounded by

$$I_{2i} = c^{-1}\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 I(\|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 > \log(N))|\hat{p}_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i) - p_{Y|\boldsymbol{X}}(y|\boldsymbol{X}_i)|^2dy$$

$$
\begin{aligned}
&\lesssim \int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 I(\|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 > \log(N)) p_0(y|\boldsymbol{X}_i) dy \\
&\lesssim \left\{ \left(\int \|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^4 p_0(y|\boldsymbol{X}_i) dy \right) \left(\int I(\|\mathbf{g}(y, \boldsymbol{X}_i, \boldsymbol{\theta})\|^2 > \log(N)) p_0(y|\boldsymbol{X}_i) dy \right) \right\}^{1/2} \\
&\lesssim N^{-1},
\end{aligned}
$$

which implies $\mathbb{E}_N(I_{2i}) \lesssim N^{-1}$. For the I_{1i} term, it can be seen that

$$
\mathbb{E}_N(I_{1i}) \lesssim \log^2(N) \mathbb{E}_N\{(\hat{p}_{Y|\boldsymbol{X}} - p_{Y|\boldsymbol{X}})^2\} = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}} \log^3(N) \right).
$$

Hence,

$$
\mathbb{E}_N\{\widehat{\mathbf{m}}(\boldsymbol{X}, \boldsymbol{\theta}) - \mathbf{m}_0(\boldsymbol{X}, \boldsymbol{\theta})\}^2 = \mathbb{E}_N(I_{1i}) + \mathbb{E}_N(I_{2i}) = O_p\left(N^{-\frac{2\beta_2}{2\beta_2+d+1}} \log^3(N) \right),
$$

which together with $\|\widehat{\mathbf{m}}_\kappa(\boldsymbol{X}_i, \boldsymbol{\theta}) - \widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\| = O_p(1/\sqrt{\kappa})$ complete the proof of Theorem 4.3. □

C Proofs for Section 5

C.1 Proof for the consistency of $\widehat{\boldsymbol{\theta}}$

Given the estimated $\widehat{\boldsymbol{\eta}}$, for any $\boldsymbol{\theta}$, we let $\boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \boldsymbol{\eta})$, $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = N^{-1}\sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$, and $\widehat{\Omega}(\boldsymbol{\theta}, \boldsymbol{\eta}) = N^{-1}\sum_{i=1}^N \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})\boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})^{\mathrm{T}}$. With the EL estimator $\widehat{\boldsymbol{\theta}}$, we write $\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\eta}}) = \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})$, $\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})$, and $\widehat{\Omega}(\boldsymbol{\eta}) = \widehat{\Omega}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\eta})$.

Lemma C.1. *Under Conditions 1 and 2, if the estimation errors satisfy*

$$
\mathcal{E}_N(\widehat{r}) + \mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(1) \quad \text{and} \quad \mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(N^{-\frac{1}{2}}), \tag{C.1}
$$

then we have

$$
\frac{1}{\sqrt{N}} \sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \boldsymbol{\eta}_0) + o_p(1). \tag{C.2}
$$

Proof. Note that for each $i = 1, \cdots, N$,

$$
\boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) - \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \boldsymbol{\eta}_0) = R_{1,i}(\widehat{\boldsymbol{\eta}}) + R_{2,i}(\widehat{\boldsymbol{\eta}}) + R_{3,i}(\widehat{\boldsymbol{\eta}}),
$$

where

$$
\begin{aligned}
R_{1,i}(\widehat{\boldsymbol{\eta}}) &= \left\{ \frac{\delta_i}{p} - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) \right\} \{\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i, \boldsymbol{\theta})\}, \\
R_{2,i}(\widehat{\boldsymbol{\eta}}) &= \frac{1-\delta_i}{1-p} \{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\{\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i, \boldsymbol{\theta})\},
\end{aligned}
$$

$$R_{3,i}(\widehat{\boldsymbol{\eta}}) = \frac{1-\delta_i}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\{\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta}) - \mathbf{m}_0(\boldsymbol{X}_i, \boldsymbol{\theta})\}.$$

Let $R_j(\widehat{\boldsymbol{\eta}}) = N^{-\frac{1}{2}}\sum_{i=1}^N R_{j,i}(\widehat{\boldsymbol{\eta}})$ for $j = 1, 2, 3$. Then (C.2) can be shown if $R_j(\widehat{\boldsymbol{\eta}}) = o_p(1)$ for $j = 1, 2, 3$. For the first term,

$$\begin{aligned}\mathbb{E}\{R_1^2(\widehat{\boldsymbol{\eta}})|\{\boldsymbol{X}_i\}_{i=1}^N\} &= \mathbb{E}_N\left[\left\{\frac{\delta_i}{p} - \frac{1-\delta_i}{1-p}r_0(\boldsymbol{X}_i)\right\}^2\{\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i, \boldsymbol{\theta})\}^2\right] \\ &\lesssim \mathbb{E}_N\left[\{\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta})\} - \mathbf{m}(\boldsymbol{X}_i, \boldsymbol{\theta})\}^2\right] = \mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}}) = o_p(1), \qquad \text{(C.3)}\end{aligned}$$

where the first equality is due to

$$\mathbb{E}_N\{R_{1,i}(\widehat{\boldsymbol{\eta}})R_{1,i'}(\widehat{\boldsymbol{\eta}})|\{\boldsymbol{X}_i\}_{i=1}^N\} = 0,$$

for each $i \neq i'$, by the independence of $(\boldsymbol{X}_i, \delta_i)$ and $(\boldsymbol{X}_i', \delta_i')$, and $\mathbb{E}_N\left\{\frac{\delta_i}{p} - \frac{1-\delta_i}{1-p}r_0(\boldsymbol{X}_i)|\boldsymbol{X}_i\right\} = 0$ for each $1 \leqslant i \leqslant N$. Therefore, $R_1(\widehat{\boldsymbol{\eta}}) = o_p(1)$. For the second term, we have

$$\begin{aligned}R_2(\widehat{\boldsymbol{\eta}}) &= \sqrt{N}\mathbb{E}_N\left\{\frac{1-\delta_i}{1-p}\,|\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)|\,|\widehat{\mathbf{m}}(\boldsymbol{X}_i, \boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{X}_i, \boldsymbol{\theta})|\right\} \\ &\lesssim \sqrt{N}\mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}_{\boldsymbol{\theta}_0}) = o_p(1),\end{aligned}$$

by the Cauchy-Schwarz inequality and (C.1). Finally, for the third term,

$$\begin{aligned}\mathbb{E}\{R_3^2(\widehat{\boldsymbol{\eta}})|\{\delta_i, \boldsymbol{X}_i\}_{i=1}^N\} &= \mathbb{E}_N\left[\frac{1-\delta_i}{(1-p)^2}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\{\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta}_0) - \mathbf{m}_0(\boldsymbol{X}_i, \boldsymbol{\theta})\}^2|\{\delta_i, \boldsymbol{X}_i\}_{i=1}^N\right] \\ &\lesssim \mathbb{E}_n\left[\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\mathrm{Var}(\mathbf{g}(\boldsymbol{Z}_i, \boldsymbol{\theta})|\boldsymbol{X}_i)\}_{i=1}^N\right] \\ &\lesssim \mathcal{E}_n(\widehat{r}) = o_p(1).\end{aligned}$$

Therefore, we have $R_3(\widehat{\boldsymbol{\eta}}) = o_p(1)$. Since

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) - \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{\Psi}(\boldsymbol{W}_i, \boldsymbol{\theta}, \boldsymbol{\eta}) = R_1(\widehat{\boldsymbol{\eta}}) + R_2(\widehat{\boldsymbol{\eta}}) + R_3(\widehat{\boldsymbol{\eta}}),$$

the proof of Lemma C.1 is finished. □

Lemma C.2. *Under Conditions 1 and 2, if the estimation errors satisfy*

$$\mathcal{E}_N(\widehat{r}) + \mathcal{E}_N(\widehat{\mathbf{m}}) = o_p(1) \quad \textit{and} \quad \mathcal{E}_N(\widehat{r})\mathcal{E}_N(\widehat{\mathbf{m}}) = o_p(N^{-\frac{1}{2}}), \qquad \text{(C.4)}$$

then $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$.

Proof. The EL estimator $\widehat{\boldsymbol{\theta}}$ can be written as the solution to the saddle point problem (Newey and Smith, 2004):

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta}\sup_{\lambda\in\widehat{\Lambda}_N(\boldsymbol{\theta})}\frac{1}{N}\sum_{i=1}^N \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})), \qquad \text{(C.5)}$$

where $\rho(v) = \log(1+v)$ and $\widehat{\Lambda}_N(\boldsymbol{\theta}) = \{\lambda : \lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}) \in (-1, \infty)\}$. For any $\xi \in (1/\alpha, 1/2)$ where α is defined in Condition 2 (ii), let $\tilde{\lambda} = N^{-\xi}\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})/\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\|$. By Lemma A1 of Newey and Smith (2004), $\max_{i \leqslant N} |\tilde{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_i(\widehat{\boldsymbol{\eta}})| = o_p(1)$, and $\tilde{\lambda} \in \Lambda_N(\widehat{\boldsymbol{\theta}})$ with probability approaching 1. Thus, for any $\dot{\lambda} \in (\tilde{\lambda}, 0)$. Let ρ_k be the k-th derivative function of ρ. Then since $\rho_2(0) = -1$, with probability approaching 1 we have $\rho_2(\dot{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_i(\widehat{\boldsymbol{\eta}})) \geqslant -C(i = 1, \cdots, N)$ for some positive constant C_1. In addition, by the Cauchy-Schwarz inequality, Condition 2 (iii), and the uniform weak law of large numbers it can easily be derived that $N^{-1}\sum_{i=1}^{N} \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}))^{\otimes 2} \leqslant C_2\mathbf{I}_r$ for some positive constant C_2 with probability approaching 1, meaning that the largest eigenvalue of $N^{-1}\sum_{i=1}^{N} \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$ is bounded from above with probability approaching 1. Taking the Taylor expansion for $\rho(\tilde{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))$ at 0 gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \rho(\tilde{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) &= \tilde{\lambda}\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) + \frac{1}{2}\tilde{\lambda}^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N} \rho_2(\dot{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))^{\otimes 2}\right\}\tilde{\lambda} \\ &\geqslant N^{-\xi}\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - \frac{C_1C_2}{2}\|\tilde{\lambda}\|^2 \geqslant N^{-\xi}\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - C_3N^{-2\xi}, \qquad \text{(C.6)}\end{aligned}$$

with probability approaching 1, where $C_3 = C_1C_2/2$.

By the similar arguments as Lemma A2 of Newey and Smith (2004), it can be shown that if for any $\bar{\boldsymbol{\theta}} \in \Theta$ such that $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$ and $\widehat{\boldsymbol{\Psi}}(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = O_p(N^{-\frac{1}{2}})$, then

$$\bar{\lambda} = \underset{\lambda \in \widehat{\Lambda}_N(\bar{\boldsymbol{\theta}})}{\arg\max}\, N^{-1}\frac{1}{N}\sum_{i=1}^{N} \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}))$$

exists with probability approaching 1, also it holds that

$$\sup_{\lambda \in \widehat{\Lambda}_N(\boldsymbol{\theta}_0)} \frac{1}{N}\sum_{i=1}^{N} \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) = O_p(N^{-1}), \quad \text{and} \quad \bar{\lambda} = O_p(N^{-\frac{1}{2}}). \qquad \text{(C.7)}$$

Setting $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Then, according to Lemma C.1,

$$\widehat{\boldsymbol{\Psi}}(\bar{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{\Psi}}(\bar{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) + o_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}}),$$

which shows that (C.7) holds with $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Using the definition of the saddle point $(\widehat{\boldsymbol{\theta}}, \bar{\lambda})$, the inequality (C.6), and the claim (C.7) with , we have

$$\begin{aligned}N^{-\xi}\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - C_3N^{-2\xi} &\leqslant \frac{1}{N}\sum_{i=1}^{N} \rho(\tilde{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) \\ &\leqslant \frac{1}{N}\sum_{i=1}^{N} \rho(\widehat{\lambda}^{\mathrm{T}}\boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})) \qquad \text{(C.8)} \\ &\leqslant \sup_{\lambda \in \widehat{\Lambda}_N(\boldsymbol{\theta}_0)} \frac{1}{N}\sum_{i=1}^{N} \rho(\lambda^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\eta}})) = O_p(N^{-1}),\end{aligned}$$

implying that $\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| = O_p(N^{-1+\xi}) + O_p(N^{-\xi}) = O_p(N^{-\xi})$, since $\xi < 1/2$. Now, suppose ϵ_N is an arbitrary sequence that converges to 0 and let $\tilde{\lambda} = \epsilon_N \widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})$, which implies $\tilde{\lambda} = o_p(N^{-\xi})$. Then, similar to (C.8), we have

$$\tilde{\lambda}^{\mathrm{T}} \|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| - C_3 \|\tilde{\lambda}\|^2 = O_p(N^{-1}),$$

which implies $\epsilon_N(1 - C_3\epsilon_N)\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\|^2 = O_p(N^{-1})$. Since $1 - C_3\epsilon_N = O(1)$, we have $\epsilon_N \|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\|^2 = O_p(N^{-1})$ for any sequence $\epsilon_N = o(1)$. Then it follows that $\|\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\| = O_p(N^{-\frac{1}{2}})$. Similar to Lemma C.1, it implies that $\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) = \widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) + o_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}})$.

According to the uniform weak law of large numbers,

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}, \boldsymbol{\eta}_0) - \boldsymbol{\Psi}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)\| = o_p(1),$$

which together with $\widehat{\boldsymbol{\Psi}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) = o_p(1)$ implies $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) = o_p(1)$. Since $\boldsymbol{\Psi}(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\Psi}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)$ is continuous with respect to $\boldsymbol{\theta}$, $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) = o_p(1)$ implies $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$, which establishes the consistency of $\widehat{\boldsymbol{\theta}}$. □

C.2 Proof of Theorem 5.1

The saddle point $(\widehat{\boldsymbol{\theta}}, \widehat{\lambda})$ to (C.5) satisfies $Q_{1,N}(\widehat{\boldsymbol{\theta}}, \widehat{\lambda}) = 0$ and $Q_{2,N}(\widehat{\boldsymbol{\theta}}, \widehat{\lambda}) = 0$, where

$$Q_{1,N}(\widehat{\boldsymbol{\theta}}, \widehat{\lambda}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + \widehat{\lambda}^{\mathrm{T}} \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})} \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}), \quad \text{and}$$
$$Q_{2,N}(\widehat{\boldsymbol{\theta}}, \widehat{\lambda}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + \widehat{\lambda}^{\mathrm{T}} \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})} \left(\frac{\partial \boldsymbol{\Psi}_i(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\theta}} \right)^{\mathrm{T}} \widehat{\lambda}.$$

By Taylor expansion of $Q_{1,N}(\widehat{\boldsymbol{\theta}}, \widehat{\lambda}) = 0$ and $Q_{2,N}(\widehat{\boldsymbol{\theta}}, \widehat{\lambda}) = 0$ around $(\boldsymbol{\theta}_0, 0)$, we have

$$0 = Q_{1,n}(\boldsymbol{\theta}_0, 0) + \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0, 0)}{\partial \boldsymbol{\theta}} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0, 0)}{\partial \lambda} \widehat{\lambda} + o_p(\delta_N), \quad \text{and}$$
$$0 = Q_{2,n}(\boldsymbol{\theta}_0, 0) + \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0, 0)}{\partial \boldsymbol{\theta}} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0, 0)}{\partial \lambda} \widehat{\lambda} + o_p(\delta_N),$$

where $\delta_N = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \|\widehat{\lambda}\|$, leading to

$$\begin{pmatrix} \hat{\lambda} \\ \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} = \mathbf{S}_N^{-1} \begin{pmatrix} -Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(\delta_N) \\ -Q_{2,N}(\boldsymbol{\theta}_0, 0) o_p(\delta_N) \end{pmatrix} = \mathbf{S}_N^{-1} \begin{pmatrix} -Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(\delta_N) \\ o_p(\delta_N) \end{pmatrix}, \tag{C.9}$$

where

$$\mathbf{S}_N = \begin{pmatrix} \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0, 0)}{\partial \lambda} & \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0, 0)}{\partial \boldsymbol{\theta}} \\ \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0, 0)}{\partial \lambda} & \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0, 0)}{\partial \boldsymbol{\theta}} \end{pmatrix},$$

and the partial derivatives are

$$\frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial \boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\theta}}, \qquad \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial \lambda} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Psi}_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\eta}})^{\otimes 2},$$
$$\frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial \boldsymbol{\theta}} = 0, \qquad \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial \lambda} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}.$$

Using the dominated convergence theorem, we can show that $\|\widehat{\mathbf{m}} - \mathbf{m}_0\| = o_p(1)$ implies $\|\partial\widehat{\mathbf{m}}/\partial\boldsymbol{\theta} - \partial\mathbf{m}_0/\partial\boldsymbol{\theta}\| = o_p(1)$. With the continuous mapping theorem and the law of large numbers, we have

$$\frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial \boldsymbol{\theta}} = \boldsymbol{\Gamma} + o_p(1), \qquad \frac{\partial Q_{1,N}(\boldsymbol{\theta}_0,0)}{\partial \lambda} = -\boldsymbol{\Omega} + o_p(1),$$
$$\frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial \boldsymbol{\theta}} = 0, \qquad \frac{\partial Q_{2,N}(\boldsymbol{\theta}_0,0)}{\partial \lambda} = \boldsymbol{\Gamma}^{\mathrm{T}} + o_p(1), \tag{C.10}$$

where

$$\boldsymbol{\Gamma} = \mathbb{E}\left\{\frac{\partial \boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}}\right\} \quad \text{and} \quad \boldsymbol{\Omega} = \mathbb{E}\left\{\boldsymbol{\Psi}(\boldsymbol{W},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)^{\otimes 2}\right\}.$$

From Lemma C.1, we have

$$Q_{1,N}(\boldsymbol{\theta}_0,0) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Psi}(\boldsymbol{W}_i,\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + o_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}}), \tag{C.11}$$

where the last equality is due to the CLT. Combining (C.9), (C.10), and (C.11), and using the continuous mapping theorem, we have

$$\begin{pmatrix} \hat{\lambda} \\ \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} = \left(\begin{pmatrix} -\boldsymbol{\Omega} & \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^{\mathrm{T}} & 0 \end{pmatrix}^{-1} + o_p(1)\right)\begin{pmatrix} Q_{1,N}(\boldsymbol{\theta}_0,0) + o_p(\delta_N) \\ o_p(\delta_N) \end{pmatrix}, \tag{C.12}$$

assuming that the block matrix on the right-hand side is invertible. Since $\delta_N = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \|\widehat{\lambda}\|$, we know that $\delta_N = O_P(N^{-\frac{1}{2}})$, which further implies that

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left\{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}\right\}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\sqrt{N}Q_{1,N}(\boldsymbol{\theta}_0,0) + o_p(1) \xrightarrow{d} \mathcal{N}(0,\left\{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}\right\}^{-1}),$$

which completes the proof of Theorem 5.1. □

C.3 Proof of Theorem 5.2

Since for every $\boldsymbol{\theta} \in \Theta$, the optimal empirical weight p_i is given by

$$p_i = \frac{1}{N}\frac{1}{1 + \lambda(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{\theta},\widehat{\boldsymbol{\eta}})},$$

where $\lambda(\boldsymbol{\theta})$ satisfies $Q_{1,N}(\boldsymbol{\theta}, \lambda(\boldsymbol{\theta})) = 0$, the log EL statistics with a given $\boldsymbol{\theta}$ can be written as

$$\ell_N(\boldsymbol{\theta}) = \log\{1 + \lambda(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{\Psi}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})\}.$$

With $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, solving $Q_{1,N}(\boldsymbol{\theta}_0, \lambda) = 0$ gives

$$\lambda(\boldsymbol{\theta}_0) = \Omega^{-1} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(N^{-\frac{1}{2}}).$$

Taking the expansion of $\ell_N(\boldsymbol{\theta}_0)$ leads to

$$\ell_N(\boldsymbol{\theta}_0) = -\frac{N}{2} Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)\Omega^{-1} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1). \tag{C.13}$$

Using the characteristic of $\widehat{\lambda}$ given in (C.12), and expanding $\ell_N(\widehat{\boldsymbol{\theta}})$ gives

$$\ell_N(\boldsymbol{\theta}_0) = -\frac{N}{2} Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)\mathbf{A} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1), \tag{C.14}$$

where

$$\mathbf{A} = -\boldsymbol{\Omega}^{-1}\{\mathbf{I} + \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\}.$$

Therefore, $R_N(\boldsymbol{\theta}_0)$ is equivalent to

$$\begin{aligned} R_N(\boldsymbol{\theta}_0) &= N Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)(\mathbf{A} - \boldsymbol{\Omega}^{-1}) Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1) \\ &= N Q_{1,N}^{\mathrm{T}}(\boldsymbol{\theta}_0, 0)\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1} Q_{1,N}(\boldsymbol{\theta}_0, 0) + o_p(1). \end{aligned}$$

Note that $(-\boldsymbol{\Omega})^{-\frac{1}{2}}\sqrt{N} Q_{1,N}(\boldsymbol{\theta}_0, 0)$ weakly converges to a standard normal distribution, and

$$(-\boldsymbol{\Omega})^{-\frac{1}{2}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}(-\boldsymbol{\Omega})^{-\frac{1}{2}}$$

is symmetric and idempotent with the trace equal to r. Hence, $R_N(\boldsymbol{\theta}_0) \xrightarrow{d} \chi_r^2$, which completes the proof of Theorem 5.2. □

C.4 Proof of Theorem 5.3

We present the proof for the density ratio estimation, since the conditional density estimation can be proved similarly. Throughout this proof, we take the compact covariate domain $\mathcal{X} = [0, 1]^d$ without loss of generality. The main idea for the proof, which is similar to that of Theorem 6.1 of Jiao et al. (2023), is to project the data to a low-dimensional space, where the DNN can be used to approximate the low-dimensional function.

Let $d_\delta = O(d_{\mathcal{M}} \log(d/\delta)/\delta^2)$ be an integer such that $d_{\mathcal{M}} \leqslant d_\delta < d$ for any $\delta \in (0, 1)$. According to Theorem 3.1 of Baraniuk and Wakin (2009), there exists a matrix $\mathbf{A} \in \mathbb{R}^{d_\delta \times d}$, which maps a manifold in $\mathbb{R}^d$ into a low-dimensional space $\mathbb{R}^{d_\delta}$ and approximately preserves the distance. To be more specific, such the matrix A satisfies $\mathbf{A}\mathbf{A}^{\mathrm{T}} = (d/d_\delta) I_{d_\delta}$, and

$$(1 - \delta)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leqslant \|\mathbf{A}\boldsymbol{x}_1 - \mathbf{A}\boldsymbol{x}_2\|_2 \leqslant (1 + \delta)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$$

for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}_\rho$. Using $\mathbf{A}$ as a projection operator, we have

$$\mathbf{A}(\mathcal{M}_\rho) \subset \mathbf{A}([0,1]^d) \subset \left[-\sqrt{\frac{d}{d_\delta}}, \sqrt{\frac{d}{d_\delta}}\right]^{d_\delta}.$$

We now show that for every $\mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho)$, there exists a unique $\boldsymbol{x} \in \mathcal{M}_\rho$ such that $\mathbf{A}\boldsymbol{x} = \mathbf{a}$. Suppose that $\boldsymbol{x}' \in \mathcal{M}_\rho$ is another point with $\mathbf{A}\boldsymbol{x}' = \mathbf{a}$. Then $(1-\delta)\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leqslant \|\mathbf{A}\boldsymbol{x} - \mathbf{A}\boldsymbol{x}'\|_2 \leqslant (1+\delta)\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ implies that $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 = 0$. Therefore, for any $\mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho)$, we can define $\boldsymbol{x}(\mathbf{a}) = \mathcal{S}_{\mathbf{A}}(\{\boldsymbol{x} \in \mathcal{M}_\rho, \mathbf{A}\boldsymbol{x} = \mathbf{a}\})$, where $\mathcal{S}_{\mathbf{A}}(\cdot)$ is a set function that maps a set to a unique element of this set. It can be shown that $\mathcal{S}_{\mathbf{A}} : \mathbf{A}(\mathcal{M}_\rho) \to \mathcal{M}_\rho$ is a differentiable function, because for every $\mathbf{a}_1, \mathbf{a}_2 \in \mathbf{A}(\mathcal{M}_\rho)$,

$$\frac{1}{1+\delta}\|\mathbf{a}_1 - \mathbf{a}_2\| \leqslant \|\boldsymbol{x}(\mathbf{a}_1) - \boldsymbol{x}(\mathbf{a}_2)\| \leqslant \frac{1}{1-\delta}\|\mathbf{a}_1 - \mathbf{a}_2\|,$$

and the norm of the derivative of $\mathcal{S}_{\mathbf{A}}$ is in the range $[(1+\delta)^{-1}, (1-\delta)^{-1}]$.

Given a function $f_0 : [0,1]^d \to \mathbb{R}$, with the operator $\boldsymbol{x}(\cdot)$, we can define its low-dimensional representation $\tilde{f}_0 : \mathbf{A}(\mathcal{M}_\rho) \to \mathbb{R}$ by

$$\tilde{f}_0(\mathbf{a}) = f_0(\boldsymbol{x}(\mathbf{a})), \quad \text{for every } \mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho) \subset \mathbb{R}^{d_\delta}.$$

Since $r_0 \in \mathcal{H}^{\beta_1}([0,1]^d B_1)$, we have $\tilde{f}_0 \in \mathcal{H}^{\beta}(\mathbf{A}(\mathcal{M}_\rho), B_1/(1-\delta)^{\beta_1})$. Since $\mathcal{M}_\rho$ is a compact space and $\mathbf{A}$ is a linear operator, by Whitney extension theorem (Fefferman, 2006), there exists $\tilde{F}_0 \in \mathcal{H}^{\beta_1}(E_\delta, B_1/(1-\delta)^{\beta_1})$ with $E_\delta = [-\sqrt{d/d_\delta}, \sqrt{d/d_\delta}]^{d_\delta}$, such that $\tilde{F}_0(\mathbf{a}) = \tilde{f}_0(\mathbf{a})$ for every $\mathbf{a} \in \mathbf{A}(\mathcal{M}_\rho)$. According to Theorem 3.3 of Jiao et al. (2023), for any $N, M \in \mathbb{N}_+$, there exists a function $\tilde{f} : E_\delta : \mathbb{R}$ belongs to the DNN function class with the ReLU activation function, whose width $W = 38(s+1)^2 d_\delta^{s+1} J\lceil\log_2(8J)\rceil$ and depth $D = 21(s+1)^2 M\lceil\log_2(8M)\rceil$, where $s = \lfloor\beta_1\rfloor$ such that

$$\sup_{\mathbf{a} \in E_\delta \backslash \Omega(E_\delta)} |\tilde{f}(\mathbf{a}) - \tilde{F}_0(\mathbf{a})| \leqslant 36\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta}, \tag{C.15}$$

where $\Omega(E_\delta)$ is a subset of E_δ whose Lebesgue measure is arbitrarily small, as well as $\Omega := \{\boldsymbol{x} \in \mathcal{M}_\rho : \mathbf{A}\boldsymbol{x} \in \Omega(E_\delta)\}$ does.

Let $\tilde{f}_* = \tilde{f} \circ \mathbf{A}$, meaning that $\tilde{f}_*(\boldsymbol{x}) = \tilde{f}(\mathbf{A}\boldsymbol{x})$ for every $\boldsymbol{x} \in [0,1]^d$. Then, $\tilde{f}_*$ is also a DNN whose width and depth are the same as $\tilde{f}$. For every $\boldsymbol{x} \in \mathcal{M}_\rho \backslash \Omega$ and $\mathbf{a} = \mathbf{A}\boldsymbol{x}$, by the definition of $\mathcal{M}_\rho$, there exists a $\tilde{\boldsymbol{x}} \in \mathcal{M}_\rho$ such that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leqslant \rho$. Then,

$$\begin{aligned}
|\tilde{f}_*(\boldsymbol{x}) - r_0(\boldsymbol{x})| &\leqslant |\tilde{f}(\mathbf{A}\boldsymbol{x}) - \tilde{F}_0(\mathbf{A}\boldsymbol{x})| + |\tilde{F}_0(\mathbf{A}\boldsymbol{x}) - \tilde{F}_0(\mathbf{A}\tilde{\boldsymbol{x}})| + |\tilde{F}_0(\mathbf{A}\tilde{\boldsymbol{x}}) - r_0(\boldsymbol{x})| \\
&\leqslant 36\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta} + \frac{B_1}{1-\delta}\|\mathbf{A}\boldsymbol{x} - \mathbf{A}\tilde{\boldsymbol{x}}\| + |r_0(\tilde{\boldsymbol{x}}) - r_0(\boldsymbol{x})| \\
&\leqslant 36\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta} + \frac{\rho B_1}{1-\delta}\sqrt{d/d_\delta} + \rho B_1
\end{aligned}$$

$$\leqslant (36 + C_\rho)\frac{B_1}{(1-\delta)^{\beta_1}}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta},$$

where the second inequality is by (C.15), the smoothness of $\tilde{F}_0$, and the definition of $\tilde{F}_0$. The third inequality is because $\|\mathbf{A}\| = \sqrt{d/d_\delta}$ and the smoothness of r_0. The positive constant C_ρ is taken such that $\rho \leqslant C_\rho(1-\delta)^{1-\beta}(s+1)^2\sqrt{d}d_\delta^{3s/2}(JM)^{-2\beta_1/d_\delta}(\sqrt{d/d_\delta}+1-\delta)^{-1}$. Since $P_{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure, we have

$$\|\tilde{f}_* - r_0\|^2_{L_2(P)} \leqslant (36 + C_\rho)^2\frac{B_1^2}{(1-\delta)^{2\beta_1}}(s+1)^4 d d_\delta^{3s}(JM)^{-4\beta_1/d_\delta}. \qquad \text{(C.16)}$$

As shown in the proof of Theorem 4.1,

$$\mathbb{E}\{\|\widehat{r} - r_0\|_n^2\} \leqslant C\left(\frac{\mathrm{Pdim}(\mathcal{F}_N)\log(N)}{N} + \epsilon_N^2\right),$$

for some positive constant C, where $\epsilon_N^2 = \inf_{f\in\mathcal{F}_N}\|\tilde{f}_* - r_0\|^2_{L_2(P)}$. According to Bartlett et al. (2019), for the DNN class $\mathcal{F}_N$ with width W and depth D, its pseodu-dimension is bounded by

$$\mathrm{Pdim}(\mathcal{F}_N) \leqslant C_1 W^2 D^2 \log(W^2 D),$$

where C_1 is a positive constant. The approximation error $\epsilon_N^2 \leqslant \|\tilde{f}_* - r_0\|^2_{L_2(P)}$ is bounded by the right-hand side of (C.16). Therefore,

$$\mathbb{E}\{\|\widehat{r} - r_0\|_n^2\} \leqslant C_2\left(\frac{W^2D^2\log(W^2D)\log(N)}{N} + \frac{B_1^2}{(1-\delta)^{2\beta_1}}(s+1)^4 d d_\delta^{3s}(JM)^{-4\beta_1/d_\delta}\right).$$

Choosing $J = 1$ and $M = N^{D_\delta}$ with $D_\delta = d_\delta/(2(d_\delta + 2\beta_1))$ leads to

$$\mathbb{E}\{\|\widehat{r} - r_0\|_n^2\} \leqslant C_3 d d_\delta^{3\lfloor\beta_1\rfloor} N^{-\frac{2\beta_1}{2\beta_1+d_\delta}},$$

where the positive constant C_3 does not depend on N or d, which completes the proof. □

C.5 Proof of Theorem 5.4

With our Lemma C.1 and Theorem 5.3, the proof is obtained by assigning $\alpha(k) = 0$ and $M = 1$ in Theorem 2 of Chang et al. (2015), and hence is omitted here.

D Proofs for Section 5

D.1 Proof of Theorem 6.1

Lemma D.1. *Under Conditions 1–3, 4 (iii), 9, and 10,*

$$\sqrt{N}\mathbb{E}\left\{\frac{1-\delta}{1-p}\widehat{r}(\boldsymbol{X})\mathbf{m}(\boldsymbol{X})\right\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{\delta_i}{p}\mathbf{m}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p}r_0(\boldsymbol{X}_i)\mathbf{m}(\boldsymbol{X}_i)\right\} + o_p(1), \qquad \text{(D.1)}$$

where the expectation is taken with respect to $\boldsymbol{X}$*.*

Recall that the criterion function for the estimation of r is defined as

$$\widehat{L}_N(r) = \frac{1}{N}\sum_{i=1}^{N} \ell(\delta_i, \boldsymbol{X}_i; r),$$

where

$$\ell(\delta, \boldsymbol{X}; r) = \frac{1-\delta}{1-p}\ell_1(\boldsymbol{X}; r) - \frac{\delta}{p}\ell_2(\boldsymbol{X}; r).$$

The directional derivative of $\ell(\delta, \boldsymbol{X}; r)$ with respect to r in the direction $u \in L_2(P)$ is given by

$$\begin{aligned} \frac{d}{du}\ell(\delta, \boldsymbol{X}; r)[u] &:= \lim_{t \to 0} \frac{\ell(\delta, \boldsymbol{X}; r+tu) - \ell(\delta, \boldsymbol{X}; r)}{t} \\ &= \left\{ \frac{1-\delta}{1-p}\frac{\partial}{\partial r}\ell_1(\boldsymbol{X}; r) - \frac{\delta}{p}\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r) \right\} u(\boldsymbol{X}) \\ &=: \ell^{(1)}(\delta, \boldsymbol{X}; r) u(\boldsymbol{X}), \quad \text{say.} \end{aligned} \tag{D.2}$$

According to Condition 9. (ii), we have

$$\ell^{(1)}(\delta, \boldsymbol{X}; r) = \frac{1-\delta}{1-p}\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r) r(\boldsymbol{X}) - \frac{\delta}{p}\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r).$$

The first-order approximation error for $\ell(\delta, \boldsymbol{X}; r_0)$ is denoted as

$$e(\delta, \boldsymbol{X}, r - r_0) = \ell(\delta, \boldsymbol{X}; r) - \ell(\delta, \boldsymbol{X}; r_0) - \frac{d}{du}\ell(\delta, \boldsymbol{X}; r_0)[r - r_0].$$

With the above notations, for any $r \in \mathcal{F}_N$, it holds that

$$\begin{aligned} \widehat{L}_N(r) =& \widehat{L}_N(r_0) + \{\widehat{L}_N(r) - \widehat{L}_N(r_0)\} \\ =& \widehat{L}_N(r_0) + \frac{1}{N}\sum_{i=1}^{N}\{\ell(\delta_i, \boldsymbol{X}_i; r) - \ell(\delta_i, \boldsymbol{X}_i; r_0)\} \\ =& \widehat{L}_N(r_0) + \frac{1}{N}\sum_{i=1}^{N}\left\{ \frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[r - r_0] + e(\delta_i, \boldsymbol{X}_i; r - r_0) \right\} \\ =& \widehat{L}_N(r_0) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[r - r_0] \right) + \frac{1}{N}\sum_{i=1}^{N} e(\delta_i, \boldsymbol{X}_i; r - r_0), \end{aligned} \tag{D.3}$$

where the last equality is because

$$\mathbb{E}\left\{ \frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[r - r_0] \right\} = 0. \tag{D.4}$$

We will employ the Cramer-Wald device to establish (D.1). For any $\boldsymbol{v} \in \mathbb{R}^p$ with $\|\boldsymbol{v}\| = 1$, we define $\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{x}) = \mathbf{m}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{v} \cdot (\partial \ell_2(\boldsymbol{x}, r)/\partial r)^{-1}$. For any $r \in \mathcal{F}_N$, let

$$\bar{r}(r, \epsilon_N) = (1 - \epsilon_N) r + \epsilon_N (r_0 + \tilde{m}_{\boldsymbol{v},\ell_2})$$

be a local alternative value around r and

$$\Pi_{\mathcal{F}_n}\bar{r}(r,\epsilon_N) = (1-\epsilon_N)r + \epsilon_N(r_* + \tilde{m}_*),$$

where $r_* = \arg\min_{r\in\mathcal{F}_N}\|r-r_0\|_{L_2(F)}$ and $\tilde{m}_* = \arg\min_{m\in\mathcal{F}_N}\|m-\tilde{m}_{\boldsymbol{v},\ell_2}\|_{L_2(F)}$. In the light of Condition 10, we have $\Pi_{\mathcal{F}_n}\bar{r}(r,\epsilon_N)\in\mathcal{F}_N$ and

$$\sup_{r\in\mathcal{F}_N}\|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(r,\epsilon_N) - \bar{r}_{\boldsymbol{v},\ell_2}(r,\epsilon_N)\|_{L_2(F)} = o(\epsilon_N\cdot N^{-\frac{1}{4}}). \tag{D.5}$$

By substituting r with $\widehat{r}$ and $\Pi_{\mathcal{F}_n}\bar{r}(\widehat{r},\epsilon_N)$, respectively, we obtain

$$\widehat{L}_N(\widehat{r}) = \widehat{L}_N(r_0) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\widehat{r}-r_0]\right) + \frac{1}{N}\sum_{i=1}^{N}e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0) \tag{D.6}$$

and

$$\begin{aligned}\widehat{L}_N(\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) =& \widehat{L}_N(r_0) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0]\right)\\ &+\frac{1}{N}\sum_{i=1}^{N}e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0).\end{aligned} \tag{D.7}$$

Subtracting (D.6) from (D.7) gives

$$\begin{aligned}\widehat{L}_N(\widehat{r}) =& \widehat{L}_N(\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) + \frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\widehat{r}-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)]\right)\\ &+\frac{1}{N}\sum_{i=1}^{N}\left\{e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0) - e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0)\right\}.\end{aligned} \tag{D.8}$$

We will prove later in Subsection D.2 that

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\left\{e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0) - e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0)\right\}\\ &= \epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right).\end{aligned} \tag{D.9}$$

By the definition of $\widehat{r}$, we have

$$\widehat{L}_N(\widehat{r}) - \widehat{L}_N(\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) \leqslant O(\epsilon_N^2),$$

which together with (D.8) and (D.9) yield

$$\begin{aligned}&\frac{1}{\sqrt{N}}\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i,\boldsymbol{X}_i;r_0)[\widehat{r}-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)]\right)\\ &-\epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right) \leqslant O(\epsilon_N^2).\end{aligned} \tag{D.10}$$

For the term $\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)]\right)$, we have

$$
\begin{aligned}
&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)]\right)\\
=&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)]\right) + \mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)]\right)\\
=&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)]\right) + o_p(\epsilon_N),
\end{aligned}
$$

where the last equality is due to (D.5) and the Chebyshev inequality. By the definition of $\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)$, we have

$$
\begin{aligned}
&\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)]\right)\\
=&\epsilon_N\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - r_0]\right) - \epsilon_N\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\tilde{m}_{\boldsymbol{v},\ell}]\right).
\end{aligned}
\tag{D.11}
$$

We now show that $\mathbb{G}_N\left(\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - r_0]\right) = o_p(1)$. By (D.2),

$$
\frac{d}{dr}\ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - r_0] = \ell^{(1)}(\delta_i, \boldsymbol{X}_i; r_0)\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}.
$$

Let

$$
\tilde{\mathcal{F}}_N = \left\{\ell^{(1)}(\delta, \boldsymbol{x}; r_0)\{r(\boldsymbol{x}) - r_0(\boldsymbol{x})\} : \ r \in \mathcal{F}_N, \|r - r_0\|_{L_2(F)} \leqslant \delta_N\right\},
$$

then it is evident that

$$
\log N_{[\ ]}(\epsilon, \tilde{\mathcal{F}}_N, L_2(F)) \lesssim \log N_{[\ ]}(\epsilon, \mathcal{F}_N, L_2(F))
$$

for any $\epsilon > 0$. Therefore, the bracketing number of $\tilde{F}_N$ satisfies

$$
\begin{aligned}
J_{[\ ]}(\delta_N, \tilde{\mathcal{F}}_N, L_2(F)) &= \int_0^{\delta_N} \sqrt{1 + \log N_{[\ ]}(\epsilon, \tilde{\mathcal{F}}_N, L_2(F))}d\epsilon\\
&\lesssim \int_0^{\delta_N} \sqrt{1 + \log N_{[\ ]}(\epsilon, \mathcal{F}_N, L_2(F))}d\epsilon\\
&= J_{[\ ]}(\delta_N, \mathcal{F}_N, L_2(F)) = o(1)
\end{aligned}
$$

by Condition 10 (iii). Also, for every $f \in \tilde{F}_N$, it holds that $\|f\|_\infty = O(1)$ and $\|f\|_{L_2(F)} = O(\delta_N)$. By applying Lemma 3.4.2 of van der Vaart and Wellner (1996), we have

$$
\mathbb{E}\|\mathbb{G}_N\|_{\tilde{F}_N} \lesssim J_{[\ ]}(\delta_N, \tilde{\mathcal{F}}_N, L_2(F))\left(1 + \frac{J_{[\ ]}(\delta_N, \tilde{\mathcal{F}}_N, L_2(F))}{\delta_N^2\sqrt{N}}O(1)\right) = o(1),
$$

which, by the Markov inequality, implies that

$$
\sup_{r\in\mathcal{F}_N} \mathbb{G}_N\left(\ell^{(1)}(\delta, \boldsymbol{x}; r_0)\{r(\boldsymbol{x}) - r_0(\boldsymbol{x})\}\right) = o_p(1),
\tag{D.12}
$$

meaning that

$$\epsilon_N \mathbb{G}_N \left(\frac{d}{dr} \ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - r_0] \right) = o_p(\epsilon_N).$$

In addition, plugging $\tilde{m}_{\boldsymbol{v},\ell}(\boldsymbol{X}_i) = m_{\boldsymbol{v}}(\boldsymbol{X}_i) \cdot \{\frac{\partial}{\partial r} \ell_2(\boldsymbol{X}_i, r_0)\}^{-1}$ into the directional derivative specified in (D.2) gives

$$-\mathbb{G}_N \left(\frac{d}{dr} \ell(\delta_i, \boldsymbol{X}_i; r_0)[\tilde{m}_{\boldsymbol{v},\ell}] \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{\delta_i}{p} m_{\boldsymbol{v}}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right\}.$$

Combining the above results gives

$$\begin{aligned} &\mathbb{G}_N \left(\frac{d}{dr} \ell(\delta_i, \boldsymbol{X}_i; r_0)[\widehat{r} - \Pi_{\mathcal{F}_N} \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)] \right) \\ =& \frac{\epsilon_N}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{\delta_i}{p} m_{\boldsymbol{v}}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right\} + o_p(\epsilon_N). \end{aligned}$$

Therefore, multiplying the both sides of (D.10) by $\sqrt{N}/\epsilon_N$ leads to

$$\begin{aligned} &\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{\delta_i}{p} m_{\boldsymbol{v}}(\boldsymbol{X}_i) - \frac{1-\delta_i}{1-p} r_0(\boldsymbol{X}_i) m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right\} + o_p(\epsilon_N) \\ &- \sqrt{N}(1-\epsilon_N) \mathbb{E} \left(\frac{1-\delta}{1-p} \{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X})\} m_{\boldsymbol{v}}(\boldsymbol{X}_i) \right) = o_p(1) + O_p\left(\frac{\epsilon_N}{\sqrt{N}} \right) = o_p(1), \end{aligned}$$

which completes the proof of Lemma D.1.

D.2 Proof of (D.9)

First, for any candidate r we can decompose $e(\delta, \boldsymbol{X}, r - r_0)$ as

$$\begin{aligned} &e(\delta, \boldsymbol{X}, r - r_0) \\ =& \ell(\delta, \boldsymbol{X}; r) - \ell(\delta, \boldsymbol{X}; r_0) - \frac{d}{du} \ell(\delta, \boldsymbol{X}; r_0)[r - r_0] \\ =& \frac{1}{2} \left\{ \frac{1-\delta}{1-p} \frac{\partial^2}{\partial r^2} \ell_1(\boldsymbol{X}; r_0) - \frac{\delta}{p} \frac{\partial^2}{\partial r^2} \ell_2(\boldsymbol{X}; r_0) \right\} \{r(\boldsymbol{X}) - r_0(\boldsymbol{X})\}^2 + R(\delta, \boldsymbol{X}, r), \end{aligned} \tag{D.13}$$

where the remainder term $R(\delta, \boldsymbol{X}, r)$ is

$$R(\delta, \boldsymbol{X}, r) = \frac{1}{2} \int_{r_0(\boldsymbol{X})}^{r(\boldsymbol{X})} \left\{ \frac{1-\delta}{1-p} \frac{\partial^3}{\partial r^3} \ell_1(\boldsymbol{X}; t) - \frac{\delta}{p} \frac{\partial^3}{\partial r^3} \ell_2(\boldsymbol{X}; t) \right\} \{r(\boldsymbol{X}) - t\}^2 dt,$$

and the last equality of (D.13) is due to the following Taylor's theorem

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2}(b-a)^2 + \int_a^b \frac{f'''(t)}{2}(b-t)^2 dt.$$

Let

$$\ell^{(2)}(\delta,\boldsymbol{X}) := \frac{1-\delta}{1-p}\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X};r_0) - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0).$$

Then, according to Condition 9.(i), we have

$$\begin{aligned}
&\frac{\partial}{\partial r}\ell_1(\boldsymbol{X};r_0) = r_0(\boldsymbol{X})\frac{\partial}{\partial r}\ell_2(\boldsymbol{X};r_0),\\
&\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X};r_0) = r_0(\boldsymbol{X})\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0) + \frac{\partial}{\partial r}\ell_2(\boldsymbol{X};r_0),
\end{aligned}$$

which implies that

$$\ell^{(2)}(\delta,\boldsymbol{X}) = \frac{1-\delta}{1-p}\left\{r_0(\boldsymbol{X})\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0) + \frac{\partial}{\partial r}\ell_2(\boldsymbol{X};r_0)\right\} - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0). \tag{D.14}$$

The last term in (D.8) can be written as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N\{e(\delta_i,\boldsymbol{X}_i;\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-r_0) - e(\delta_i,\boldsymbol{X}_i;\widehat{r}-r_0)\}\\
=&\frac{1}{2N}\sum_{i=1}^N\left\{\frac{1-\delta}{1-p}\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X};r_0) - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0)\right\}\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
&-\frac{1}{2N}\sum_{i=1}^N\left\{\frac{1-\delta}{1-p}\frac{\partial^2}{\partial r^2}\ell_1(\boldsymbol{X};r_0) - \frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X};r_0)\right\}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
&+\frac{1}{N}\sum_{i=1}^N\{R(\delta_i,\boldsymbol{X}_i,\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)) - R(\delta_i,\boldsymbol{X}_i,\widehat{r})\}\\
=:&E_{1,N}+E_{2,N}+E_{3,N},\quad \text{say.}
\end{aligned}$$

For the term $\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2$, we have

$$\begin{aligned}
&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
=&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) + \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2\\
=&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) + (1-\epsilon_N)(\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)) + \epsilon_N\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}^2\\
=&\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}^2 + (1-\epsilon_N)^2\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2 + \epsilon_N^2\tilde{m}_{\boldsymbol{v},\ell_2}^2(\boldsymbol{X}_i)\\
&+2(1-\epsilon_N)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\\
&+2\epsilon_N\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
&+2(1-\epsilon_N)\epsilon_N\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i).
\end{aligned}\tag{D.15}$$

Using (D.15), we can decompose $E_{1,N}+E_{2,N}$ as

$$\begin{aligned}
&E_{1,N}+E_{2,N}\\
=&\frac{1}{2N}\sum_{i=1}^N\ell^{(2)}(\delta_i,\boldsymbol{X}_i)[\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2 - \{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}^2]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{2N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}^2\\
&+\frac{(1-\epsilon_N)^2-1}{2N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}^2+\frac{\epsilon_N^2}{2N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\tilde{m}^2_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
&+\frac{1-\epsilon_N}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\\
&+\frac{\epsilon_N}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
&+\frac{\epsilon_N(1-\epsilon_N)}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
=&\frac{1}{2}\mathbb{E}[\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}^2]\{1+o_p(1)\}\\
&+\frac{\epsilon_N^2-2\epsilon_N}{2}\mathbb{E}[\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}^2]\{1+o_p(1)\}+\frac{\epsilon_N^2}{2}\mathbb{E}\{\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\tilde{m}^2_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}\{1+o_p(1)\}\\
&+(1-\epsilon_N)\mathbb{E}[\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}]\{1+o_p(1)\}\\
&+\epsilon_N\mathbb{E}\{\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}\\
&+\frac{\epsilon_N(1-\epsilon_N)}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
\leqslant&O_p(\epsilon_N^2\delta_N^2)+O_p(\epsilon_N\delta_N^2)+O_p(\epsilon_N^2)+O_p(\epsilon_N\delta_N^2)+O_p(\epsilon_N^2\delta_N)\\
&+\frac{\epsilon_N(1-\epsilon_N)}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i),
\end{aligned}
\tag{D.16}
$$

where the expectations are taken with respect to $(\delta_i,\boldsymbol{X}_i)$, and the last equality is by the uniform boundness of $\ell^{(2)}(\delta,\boldsymbol{X})$, the approximation error in (D.5), and the bounded moment of $\|\tilde{m}_{\boldsymbol{v},\ell_2}\|^2$. For the last term in (D.16), we note that

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\\
=&\frac{1}{\sqrt{N}}\mathbb{G}_N\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right)\\
&+\mathbb{E}\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right),
\end{aligned}
\tag{D.17}
$$

where the expectation is taken with respect to $(\delta_i,\boldsymbol{X}_i)$. By the stochastic equicontinuity which can be derived with the similar arguments as for (D.12), we can obtain

$$
\mathbb{G}_N\left(\ell^{(2)}(\delta_i,\boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right)=o_p(1).
\tag{D.18}
$$

In the light of (D.14) and $\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)=m_{\boldsymbol{v}}(\boldsymbol{X}_i)\cdot\{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i,r_0)\}^{-1}$, the expectation term can be written as

$$
\begin{aligned}
&\mathbb{E}\left(\ell^{(2)}(\delta_i, \boldsymbol{X}_i)\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\right)\\
=&\mathbb{E}\left(\frac{1-\delta}{1-p}\left\{r_0(\boldsymbol{X})\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0) + \frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; r_0)\right\}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}m_{\boldsymbol{v}}(\boldsymbol{X}_i) \cdot \{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i, r_0)\}^{-1}\right)\\
&-\mathbb{E}\left\{\frac{\delta}{p}\frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; r_0)\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)\}m_{\boldsymbol{v}}(\boldsymbol{X}_i) \cdot \{\frac{\partial}{\partial r}\ell_2(\boldsymbol{X}_i, r_0)\}^{-1}\right\}\\
=&\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right), \qquad \text{(D.19)}
\end{aligned}
$$

where the last equality is due to $\mathbb{E}\{(1-\delta)r_0(\boldsymbol{X})f(\boldsymbol{X})\} = \mathbb{E}\{\delta f(\boldsymbol{X})\}$ for any $f(\boldsymbol{X})$. Combining (D.16), (D.17), (D.18), and (D.19), and taking the convergence rate $\delta_N = o_p(N^{-\frac{1}{4}})$, we obtain

$$
E_{1,N} + E_{2,N} = \epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X})\}m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right). \qquad \text{(D.20)}
$$

For the term $E_{3,N}$, we let

$$
\ell^{(3)}(\delta, \boldsymbol{X}; t) = \frac{1-\delta}{1-p}\frac{\partial^3}{\partial r^3}\ell_1(\boldsymbol{X}; t) - \frac{\delta}{p}\frac{\partial^3}{\partial r^3}\ell_2(\boldsymbol{X}; t).
$$

Due to $\frac{\partial}{\partial r}\ell_1(\boldsymbol{X}, t) = t \cdot \frac{\partial}{\partial r}\ell_2(\boldsymbol{X}, t)$ imposed in Condition 9, we have

$$
\ell^{(3)}(\delta, \boldsymbol{X}; t) = \frac{1-\delta}{1-p}\left\{t \cdot \frac{\partial^3}{\partial r^3}\ell_2(\boldsymbol{X}; t) + \frac{\partial^2}{\partial r^2}\ell_2(\boldsymbol{X}; t) + \frac{\partial}{\partial r}\ell_2(\boldsymbol{X}; t)\right\} - \frac{\delta}{p}\frac{\partial^3}{\partial r^3}\ell_2(\boldsymbol{X}; t), \qquad \text{(D.21)}
$$

which is uniformly bounded by some positive constant c_ℓ according to Condition 9.(ii).

then $E_{3,N}$ can be decomposed as

$$
\begin{aligned}
E_{3,N} =&\frac{1}{N}\sum_{i=1}^{N}\{R(\delta_i, \boldsymbol{X}_i, \Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N)) - R(\delta_i, \boldsymbol{X}_i, \widehat{r})\}\\
=&\frac{1}{2N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)}\ell^{(3)}(\delta_i, \boldsymbol{X}_i; t)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - t\}^2 dt\\
&-\frac{1}{2N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)}\ell^{(3)}(\delta_i, \boldsymbol{X}_i; t)\{\widehat{r}(\boldsymbol{X}_i) - t\}^2 dt\\
=&\frac{1}{2N}\sum_{i=1}^{N}\int_{\widehat{r}(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)}\ell^{(3)}(\delta_i, \boldsymbol{X}_i; t)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - t\}^2 dt\\
&-\frac{1}{2N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)}\ell^{(3)}(\delta_i, \boldsymbol{X}_i; t)[\{\widehat{r}(\boldsymbol{X}_i) - t\}^2 - \{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - t\}^2]dt\\
=:&D_{1,N} + D_{2,N}, \quad \text{say.}
\end{aligned}
$$

For the term $D_{1,N}$, we have

$$\begin{aligned}
2|D_{1,N}| &= \frac{1}{N}\left|\sum_{i=1}^{N}\int_{\widehat{r}(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)} \ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-t\}^2dt\right| \\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}\left|\int_{\widehat{r}(\boldsymbol{X}_i)}^{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)} \{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-t\}^2dt\right| \\
&= \frac{c_\ell}{N}\sum_{i=1}^{N}(1-s_i)\left|\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\}^3\right| \quad \text{(for some } s_i\in(0,1)) \\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)|^3 \\
&\leqslant \frac{2c_\ell}{N}\sum_{i=1}^{N}\{|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)|^3+|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)|^3\},
\end{aligned}$$

where the first inequality is from the uniform boundness of $\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)$, the second equality is by applying the mean value theorem, and the last inequality is from the inequality $(a+b)^3\leqslant 2(a^3+b^3)$ for any positive a and b. From (D.5) it can be easily derived that $\max_{1\leqslant i\leqslant N}|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)|=o_p(1)$. For the term $|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)|$, we have

$$\frac{1}{N}\sum_{i=1}^{N}|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)| = \epsilon_N\frac{1}{N}\sum_{i=1}^{N}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)-\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}=O_p(\epsilon_N), \quad \text{(D.22)}$$

$$\frac{1}{N}\sum_{i=1}^{N}|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)|^2 = \epsilon_N^2\frac{1}{N}\sum_{i=1}^{N}\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)-\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)\}^2=O_p(\epsilon_N^2), \quad \text{(D.23)}$$

Using Lemma 2 of Owen (1990), it holds that $\max_{1\leqslant i\leqslant N}|\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)|=o_p(\sqrt{N})$, which together with the uniform boundness of $\widehat{r}$ and r_0 and $\epsilon_N=o_p(N^{-\frac{1}{2}})$ imply that

$$\max_{1\leqslant i\leqslant N}|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)| = \epsilon_N\max_{1\leqslant i\leqslant N}|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)-\tilde{m}_{\boldsymbol{v},\ell_2}(\boldsymbol{X}_i)|=o_p(1). \quad \text{(D.24)}$$

Therefore, $|D_{1,N}|$ can be bounded by

$$|D_{1,N}|\leqslant o_p(\epsilon_N^2\delta_N^2)+o_p(\epsilon_N^2)=o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right), \quad \text{(D.25)}$$

where the equality is due to $\epsilon_N=o(N^{-\frac{1}{2}})$ and $\delta_N=o(N^{-\frac{1}{4}})$.

For the term $D_{2,N}$, we have

$$\begin{aligned}
2|D_{2,N}| &= \frac{1}{N}\left|\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)} \ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)[\{\widehat{r}(\boldsymbol{X}_i)-t\}^2-\{\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-t\}^2]dt\right| \\
&= \frac{1}{N}\left|\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)} \ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)[\{\widehat{r}(\boldsymbol{X}_i)-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\}\{\widehat{r}(\boldsymbol{X}_i)+\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-2t\}]dt\right|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}\int_{r_0(\boldsymbol{X}_i)}^{\widehat{r}(\boldsymbol{X}_i)} \left|\widehat{r}(\boldsymbol{X}_i)-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\left|\widehat{r}(\boldsymbol{X}_i)+\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-2t\right| dt\\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}\{\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\widehat{r}(\boldsymbol{X}_i)-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\\
&\qquad\qquad \cdot\left|\widehat{r}(\boldsymbol{X}_i)+\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-2\{s_i\widehat{r}(\boldsymbol{X}_i)+(1-s_i)r_0(\boldsymbol{X}_i)\}\right|\}\quad\text{(for some } s_i\in(0,1))\\
&= \frac{c_\ell}{N}\sum_{i=1}^{N}\{\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\widehat{r}(\boldsymbol{X}_i)-\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\\
&\qquad\qquad \cdot\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)+2(1-s_i)\{\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\}\right|\}\\
&\leqslant \frac{c_\ell}{N}\sum_{i=1}^{N}\{\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left(\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|+\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|\right)\\
&\qquad\qquad \cdot\left(\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|+\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|+2\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\right)\}\\
&= \frac{c_\ell}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|^2\\
&\quad+\frac{c_\ell}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|\\
&\quad+\frac{2c_\ell}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|^2\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\\
&\quad+\frac{c_\ell}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\\
&\quad+\frac{c_\ell}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|^2\\
&\quad+\frac{2c_\ell}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|^2\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|, \qquad\qquad \text{(D.26)}
\end{aligned}
$$

where the first inequality is from the uniform boundness of $\ell^{(3)}(\delta_i,\boldsymbol{X}_i;t)$ and the second inequality is by applying the mean value theorem. By the uniform boundness of $\widehat{r}$ and r_0, the approximation error in (D.5), (D.23), $\|\widehat{r}-r_0\|_{L_2(P)}=O_p(\delta_N)$, and the Cauchy-Schwarz inequality, we can obtain

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|^2=O_p(\epsilon_N^2\delta_N^2),\\
&\frac{1}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|\left|\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\widehat{r}(\boldsymbol{X}_i)\right|=O_p(\epsilon_N^2\delta_N),\\
&\frac{1}{N}\sum_{i=1}^{N}\left|\widehat{r}(\boldsymbol{X}_i)-r_0(\boldsymbol{X}_i)\right|^2\left|\Pi_{\mathcal{F}_N}\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)-\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r},\epsilon_N)\right|=O_p(\epsilon_N\delta_N^2).
\end{aligned}
$$

By the uniform boundness of $\widehat{r}$ and r_0, $\|\widehat{r} - r_0\|_{L_2(P)} = O_p(\delta_N)$, and (D.24), we have

$$\frac{1}{N}\sum_{i=1}^{N} |\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)|\, |\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - \widehat{r}(\boldsymbol{X}_i)|^2 = O_p(\epsilon_N^2)$$
$$\frac{1}{N}\sum_{i=1}^{N} |\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X}_i)|^2\, |\bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - \widehat{r}(\boldsymbol{X}_i)| = O_p(\epsilon_N \delta_N^2),$$

where the second result is obtained from the Cauchy-Schwarz inequality. Collecting the above results and plugging them into (D.26), we can bound $|D_{2,N}|$ by

$$\begin{aligned} |D_{2,N}| &\leqslant O_p(\epsilon_N^2 \delta_N^2) + O_p(\epsilon_N \delta_N^2) + O_p(\epsilon_N \delta_N^2) + O_p(\epsilon_N^2) \\ &= o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right), \end{aligned} \tag{D.27}$$

where the equality is due to $\epsilon_N = o(N^{-\frac{1}{2}})$ and $\delta_N = o_p(N^{-\frac{1}{4}})$.

To sum up, we have shown that

$$E_{3,N} = D_{1,N} + D_{2,N} = o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right),$$

which together with the result for $E_{1,N} + E_{2,N}$ in (D.20) yield

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} \{e(\delta_i, \boldsymbol{X}_i; \Pi_{\mathcal{F}_N} \bar{r}_{\boldsymbol{v},\ell_2}(\widehat{r}, \epsilon_N) - r_0) - e(\delta_i, \boldsymbol{X}_i; \widehat{r} - r_0)\} \\ =&\epsilon_N(1-\epsilon_N)\mathbb{E}\left(\frac{1-\delta}{1-p}\{\widehat{r}(\boldsymbol{X}_i) - r_0(\boldsymbol{X})\} m_{\boldsymbol{v}}(\boldsymbol{X}_i)\right) + o_p\left(\frac{\epsilon_N}{\sqrt{N}}\right), \end{aligned}$$

which is the desired result. □

E Additional simulation results

In this part, we report additional results of the numerical simulations, including the inference for the mean of Y of the target population with the dimension of the covariate $d = 5$ in Table 1, and the inference for the mean and median Y of the target population with $d = 10$ in Table 2 and 3, respectively.

Table 1. Empirical estimation and inference results for $\theta = \mathbb{E}_Q(Y)$ of the target population with $d = 5$ based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	-0.0168	0.1322	0.0175	0.9048	0.4087
	MI	0.0203	0.1471	0.0217	0.8736	0.3716
	DRW-MI-E	-0.0135	0.1304	0.0171	0.9265	0.3824
	DRW-MI-T	-0.0125	0.1271	0.0163	0.9374	0.3791
	LDML	-0.0117	0.1426	0.0204	0.8592	0.3617
	CB	0.0370	0.1683	0.0297	0.7332	0.4204
$n = 2000$	DRW	-0.0149	0.1006	0.0103	0.9102	0.2817
	MI	-0.0182	0.1120	0.0129	0.8914	0.2546
	DRW-MI-E	-0.0118	0.0937	0.0089	0.9350	0.2972
	DRW-MI-T	-0.0121	0.0922	0.0086	0.9550	0.2935
	LDML	0.0130	0.1105	0.0124	0.9008	0.2780
	CB	0.0302	0.1319	0.0183	0.7298	0.3064
$n = 5000$	DRW	0.0105	0.0772	0.0061	0.9163	0.1708
	MI	-0.0127	0.0869	0.0078	0.9081	0.1665
	DRW-MI-E	0.0084	0.0673	0.0046	0.9437	0.1812
	DRW-MI-T	-0.0081	0.0660	0.0043	0.9481	0.1845
	LDML	-0.0119	0.0882	0.0078	0.9083	0.1713
	CB	0.0267	0.0941	0.0096	0.7510	0.1964

Table 2. Empirical estimation and inference results for $\theta = \mathbb{E}_Q(Y)$ of the target population with $d = 20$ based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	0.0815	0.3048	0.0995	0.7296	1.1592
	MI	-0.0902	0.3407	0.1242	0.7381	1.2157
	DRW-MI-E	0.0521	0.2485	0.0645	0.8168	0.9052
	DRW-MI-T	0.0347	0.2019	0.0419	0.8477	0.8895
	LDML	0.0609	0.3601	0.1334	0.7201	1.3162
	CB	-0.1308	0.2724	0.0864	0.5942	0.8125
$n = 2000$	DRW	0.0701	0.2382	0.0616	0.7640	0.8619
	MI	-0.0736	0.2619	0.0631	0.7774	0.9015
	DRW-MI-E	-0.0452	0.1829	0.0355	0.8851	0.7824
	DRW-MI-T	-0.0301	0.1681	0.0291	0.9174	0.7637
	LDML	0.0492	0.2128	0.0477	0.7831	0.8459
	CB	-0.0945	0.2209	0.0559	0.5781	0.7037
$n = 5000$	DRW	0.0539	0.1839	0.0367	0.8152	0.6729
	MI	0.0569	0.2007	0.0435	0.8347	0.7138
	DRW-MI-E	-0.0335	0.1362	0.0196	0.9214	0.6042
	DRW-MI-T	0.0304	0.1120	0.0135	0.9436	0.5814
	LDML	0.0369	0.1783	0.0331	0.8152	0.7221
	CB	-0.0901	0.1821	0.0395	0.6515	0.5981

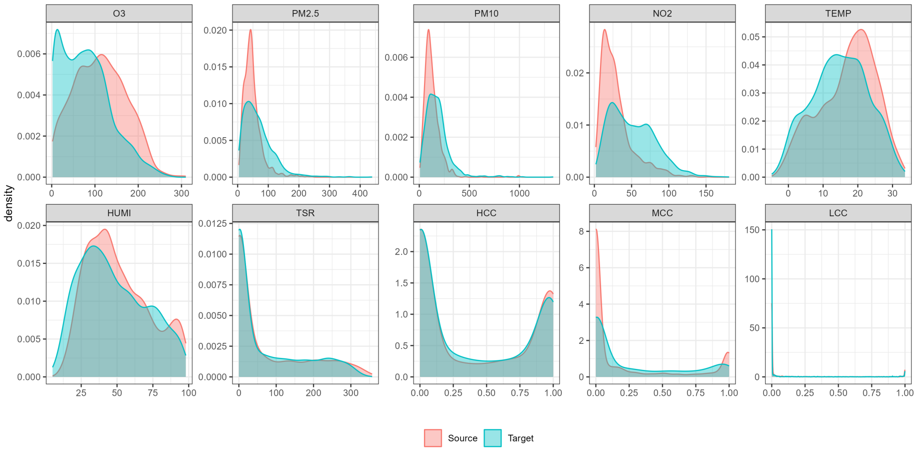
Table 3. Empirical estimation and inference results for $\theta = Q_Y^{-1}(1/2)$ of the target population with $d = 20$ based on 300 simulation replications. The five methods considered are the density ratio weighting (DRW), the multiple imputations (MI), the proposed method with both the density ratio weighting and the multiple imputations using the estimated nuisance functions (DRW-MI-E), the DRW-MI using the true nuisance functions (DRW-MI-T), the localized double machine learning (LDML), and the covariance balancing (CB). The nominal coverage probability of the confidence interval is 0.95.

	Methods	Bias	Std.dev	MSE	Coverage	Length of CI
$n = 1000$	DRW	-0.0943	0.3420	0.1258	0.7169	1.2011
	MI	-0.0962	0.3541	0.1346	0.7215	1.2142
	DRW-MI-E	0.0731	0.2685	0.0774	0.8280	1.0204
	DRW-MI-T	0.0527	0.2301	0.0557	0.8505	0.9969
	LDML	-0.0693	0.3318	0.1148	0.7119	1.2650
	CB	-0.1436	0.2817	0.1001	0.5523	0.8856
$n = 2000$	DRW	0.0815	0.2740	0.0817	0.7593	0.8619
	MI	-0.0856	0.2802	0.0858	0.7324	0.8242
	DRW-MI-E	-0.0528	0.2129	0.0481	0.8613	0.7907
	DRW-MI-T	0.0493	0.1891	0.0381	0.9038	0.7741
	LDML	0.0566	0.2547	0.0681	0.7918	0.8109
	CB	-0.1231	0.2037	0.0566	0.5390	0.7074
$n = 5000$	DRW	0.0652	0.1971	0.0431	0.8098	0.6872
	MI	-0.0690	0.2085	0.0482	0.8209	0.7524
	DRW-MI-E	-0.0341	0.1381	0.0203	0.9209	0.6507
	DRW-MI-T	-0.0318	0.1152	0.0143	0.9367	0.5901
	LDML	0.0392	0.1801	0.0339	0.8247	0.7349
	CB	-0.1056	0.1618	0.0373	0.5607	0.5890

F Additional case study results

Figure 1 in the SM illustrates the distinctions between the distributions of some key variables of the target and the source samples, which reveals that directly using the source samples to make inferences about the O_3 of the target population would introduce biases.

Figure 1. Density plots for the O_3 and covariate variables of the source and the target samples.



References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*, volume 9. Cambridge University Press, 1999.

Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005.

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

Jinyuan Chang, Song Xi Chen, and Xiaohong Chen. High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185 (1):283–304, 2015.

Charles Fefferman. Whitney's extension problem for. *Annals of mathematics*, pages 313–359, 2006.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical Estimation: Asymptotic Theory*. SpringerVerlag, New York, 1981.

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

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Art Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18 (1):90 – 120, 1990.

David Pollard. *Empirical Processes: Theory and Applications*, volume 2. Institute of Mathematical Statistics, 1990.

R Tyrrell Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1997.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

Aad W van der Vaart and Jon A Wellner. Weak convergence and empirical processes. *Springer Series in Statistics*, 126:505, 1996.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.