

APPENDIX

A DETAILED EXPERIMENTAL RESULTS

A.1 SIMULATION STUDY

We conduct a simulation study to illustrate the proposed methods. By extending our motivating example in Section 2.1, we consider the following data-generating process:

$$\begin{aligned}
 F &\sim \text{Bernoulli}(0.5), \\
 \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} &\sim N \left(\begin{bmatrix} 2(2F-1) \\ 2(2F-1) \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{bmatrix} \right), \\
 X_4 &\sim \text{Uniform}[-3, 3], \quad X_5 \sim \chi^2(1), \quad X_6 \sim \text{Bernoulli}(0.5), \\
 A &= \mathbb{1}\{F + FX_1 + 2X_2 - 2X_3 \\
 &\quad - X_4 - 0.5X_5 + X_6 + \delta > 0\}, \\
 Y^a &= (X_1 + X_2 + X_5)^2 + \Lambda \left(\frac{X_1 - 2(2a-1)}{2} \right) + \varepsilon',
 \end{aligned}$$

for $\delta \sim N(0, 5)$, $\varepsilon' \sim N(0, 1)$. To assess robustness, we additionally define two transformed covariate sets: $X^* = (X_1X_3, X_2^2, X_4, X_5, X_6)$, $X^{**} = (X_1X_3, X_2^2, X_4, 0.5 \log(X_5), \text{Sigmoid}(X_6))$. The above setup can be viewed as an extension of the simulation design by Hainmueller (2012).

Baselines. We select five baseline methods that are widely recognized in the literature on heterogeneous treatment effects. Three of these methods focus on accurately estimating the CATE using distinct approaches: inverse probability weighting (IPW) (Zhou & Zhu, 2021), outcome regression with random forests (Wager & Athey, 2018), and doubly robust estimation (Kennedy, 2020). The remaining two methods identify subgroups by constructing terminal nodes of a decision tree through recursive partitioning. We include the method proposed by Loh et al. (2015), as it has been reported to be the best-performing approach in the comparative study conducted by Loh et al. (2019).

In the baseline methods, subgroups are identified by applying a threshold to the CATE estimates, classifying individuals based on whether the estimated treatment effect is greater or less than zero. When applying tree-based methods, we effectively constrain all models to a stump (a decision tree with only two terminal nodes), ensuring a binary partitioning of the sample.

Hyperparameters. In our methods, we simply set $R = 2$. We set $m = 100$, $\lambda = 1$ and initialize the tolerance levels at $\delta, \delta' = 0.1$. We then progressively decrease δ and δ' until the optimization problem becomes infeasible. When employing kernel basis functions, following Hazlett (2020) and Kim et al. (2024c), we balance only the first B eigenvectors scaled by the inverse of their corresponding eigenvalues, rather than the full n columns of the kernel matrix. With our setting of $n = 2000$, we use $B = 50, 100$. However, since no significant differences are observed between the two cases, we report results only for $B = 50$.

Optimizer and Computing resources. We use a MOSEK optimizer in R (Rmosek package v11.0.20). All experiments were conducted on a Mac Studio equipped with an Apple M4 Max chip, featuring 32 CPU cores and 64 GPU cores, and 512GB memory.

Performance and subgroup fairness. Here, we compare covariate distributions within subgroup and subgroup fairness across different methods. For illustrative purposes, and given the comparable performance across similar methods, we report results only for the first baseline (B1) and two of the proposed methods: SB1 (power series) and SB2 (kernel 1). S_1 and S_2 denote the subgroups corresponding to positive and negative treatment effects, respectively. Figures 5–13 correspond to the setup with X^* , while Figures 14–22 pertain to the setup with X^{**} . We observe that the proposed methods utilizing the kernel basis achieve the desired levels of covariate balance and fairness with greater stability.

In Table 2, we also evaluate performance of our methods against CATE-based baselines describe above in terms of the Root Mean Squared Error (RMSE), along with the average unfairness ($\mathbb{P}_n\{\widehat{\text{uf}} \cdot S_r\}$)

Methods\RMSE	X	X^*	X^{**}	UF
B1 (Zhou & Zhu, 2021)	0.06 (± 0.03)	0.16 (± 0.06)	0.29 (± 0.05)	0.58 (± 0.11)
B2 (Loh et al., 2015)	0.09 (± 0.03)	0.19 (± 0.06)	0.33 (± 0.09)	0.52 (± 0.10)
B3 (Wager & Athey, 2018)	0.05 (± 0.02)	0.17 (± 0.06)	0.38 (± 0.06)	0.61 (± 0.12)
B4 (Kennedy, 2020)	0.04 (± 0.03)	0.19 (± 0.05)	0.22 (± 0.08)	0.62 (± 0.15)
B5 (Athey & Imbens, 2016)	0.07 (± 0.05)	0.15 (± 0.01)	0.38 (± 0.09)	0.69 (± 0.10)
SB1 (power series)	0.06 (± 0.01)	0.07 (± 0.02)	0.16 (± 0.03)	0.02 (± 0.01)
SB2 (kernel 1)	0.06 (± 0.01)	0.08 (± 0.02)	0.10 (± 0.02)	0.01 (± 0.01)
SB3 (kernel 2)	0.07 (± 0.01)	0.09 (± 0.02)	0.08 (± 0.02)	0.01 (± 0.01)

Table 2: RMSE of subgroup estimation across different methods. One-standard-deviation (1σ) errors are reported beneath each value. Best results in each column are highlighted in bold.

across subgroups/settings. We generate a sample of size $n=2000$ per iteration, and repeat the simulation 100 times per setting. In the proposed subgroup balancing (SB) methods, we utilize power series expansion up to the fourth moment (SB1), kernel basis functions (SB2), and kernel basis with an additional balancing condition equation 11 (SB3). The results demonstrate that the proposed methods not only achieve the desired level of fairness within subgroups but also provide more accurate subgroup effect estimates compared to the baselines.

A.2 CASE STUDY

The data used in this study are derived from the ACTG 175 randomized trial (Hammer et al., 1996) and are publicly available in the `speff2trial` R package. The treatment variable A indicates whether patients received combination therapy ($A = 1$) or zidovudine alone ($A = 0$). The outcome Y is the CD4 count measured 96 weeks post-baseline. The baseline covariates X include age, weight, Karnofsky score, race, gender, hemophilia, homosexual activity, drug use, symptomatic HIV status, and prior zidovudine and antiretroviral use. See Table 3 for description of each covariate information. The analysis is conducted on a sample of $n = 1342$ patients with complete outcome and covariate data. We use the kernel basis with the additional balancing condition (SB2) for our analysis, with the same setup as in our simulation study.

Basic demographics	
Male sex — No. (%)	Age — yr
Race or ethnic group — No. (%)	
White (non-Hispanic), Black (non-Hispanic), Hispanic, Other	
Risk factors — No. (%)	
Homosexuality, Injection-drug use, Hemophilia	
Karnofsky score = 100 — No. (%)	Symptomatic HIV infection — No. (%)
CD4 cell count — cells/mm ³	
Median length of prior ART — (IQR)	

Table 3: Baseline characteristics of study participants.

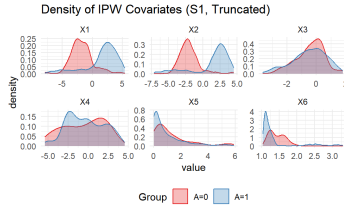


Figure 5: Covariate distributions for S_1 (X^* , B1).

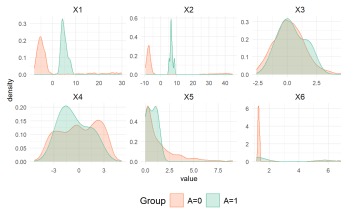


Figure 6: Covariate distributions for S_2 (X^* , B1).



Figure 7: Sensitive variable distributions (X^* , B1).

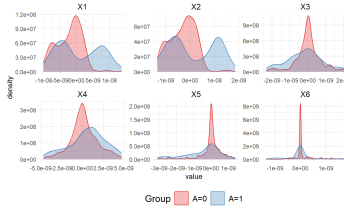


Figure 8: Covariate distributions for S_1 (X^* , SB1).

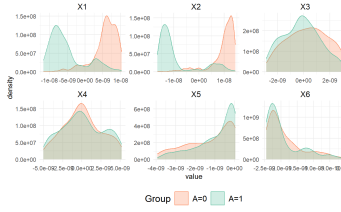


Figure 9: Covariate distributions for S_2 (X^* , SB1).

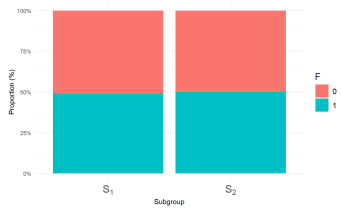


Figure 10: Sensitive variable distributions (X^* , SB1).

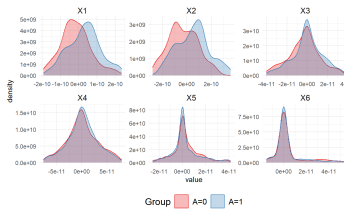


Figure 11: Covariate distributions for S_1 (X^* , SB2).

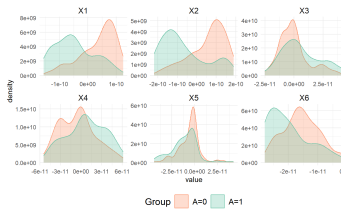


Figure 12: Covariate distributions for S_2 (X^* , SB2).

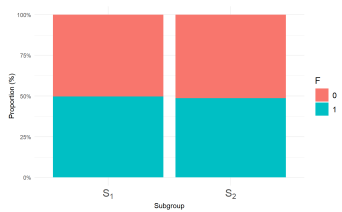


Figure 13: Sensitive variable distributions (X^* , SB2).

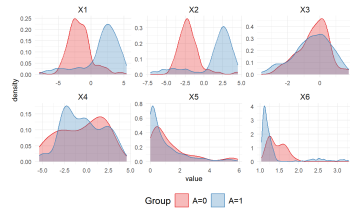


Figure 14: Covariate distributions for S_1 (X^{**} , B1).

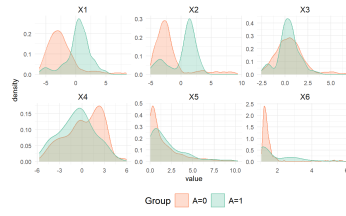


Figure 15: Covariate distributions for S_2 (X^{**} , B1).

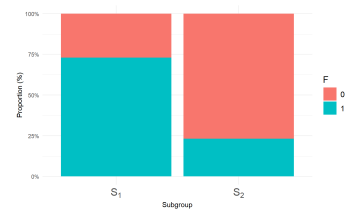


Figure 16: Sensitive variable distributions (X^{**} , B1).

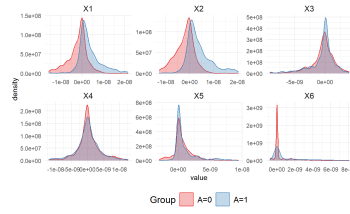


Figure 17: Covariate distributions for S_1 (X^{**} , SB1).

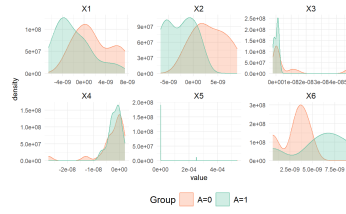


Figure 18: Covariate distributions for S_2 (X^{**} , SB1).

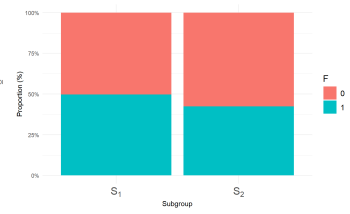


Figure 19: Sensitive variable distributions (X^{**} , SB1).

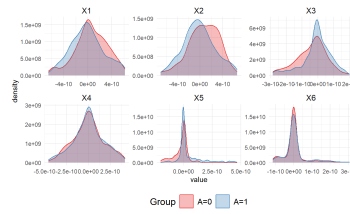


Figure 20: Covariate distributions for S_1 (X^{**} , SB2).

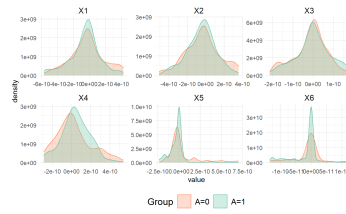


Figure 21: Covariate distributions for S_2 (X^{**} , SB2).

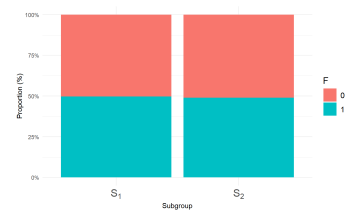


Figure 22: Sensitive variable distributions (X^{**} , SB2).

B PROOFS

B.1 PROOF OF THEOREM 4.1

Proof. For notational convenience, we let d denote the L_q distance here. By the consistency assumption (C1), we have $Y = AY^1 + (1 - A)Y^0$. Then under the additive model equation 8, for any $r \in \{1, \dots, R\}$, we obtain the following based on the similar logic used in Yang et al. (2021, Proposition 2).

$$\begin{aligned}
d(\tau_r, \hat{\tau}_r) &= \left\| \sum_{i=1}^n A_i S_{ir} w_i Y_i - \sum_{i=1}^n (1 - A_i) S_{ir} w_i Y_i - \tau_r \right\|_q \tag{14} \\
&= \left\| \sum_{j=1}^B \beta_{rj} \left\{ \sum_{i=1}^n A_i S_{ir} w_i \phi_j(X_i) - \sum_{i=1}^n (1 - A_i) S_{ir} w_i \phi_j(X_i) \right\} + \beta_r \left\{ \sum_{i=1}^n A_i S_{ir} w_i - \sum_{i=1}^n (1 - A_i) S_{ir} w_i \right\} \right. \\
&\quad \left. + \tau_r \left(\sum_{i=1}^n A_i S_{ir} w_i - 1 \right) + \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} - \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q \\
&= \left\| \sum_{j=1}^B \beta_{rj} \left\{ \sum_{i=1}^n A_i S_{ir} w_i \phi_j(X_i) - \sum_{i=1}^n (1 - A_i) S_{ir} w_i \phi_j(X_i) \right\} + \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} - \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q \\
&\leq \delta \sum_{j=1}^B \|\beta_{rj}\|_q + \left\| \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} \right\|_q + \left\| \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q, \tag{15}
\end{aligned}$$

where the second equality follows by the normalization equation 7b, the third by the balancing condition equation 7c and the triangle inequality.

For the second term in equation 15, it follows that

$$\begin{aligned}
\mathbb{E} \left\| \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} \right\|_q &\leq \mathbb{E} \left\| \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} \right\|_1 \\
&\leq \mathbb{E} \left\| \sum_{i=1}^n w_i \epsilon_{1i} \right\|_1 \\
&= \sum_{t=1}^p \mathbb{E} \left| \sum_{i=1}^n w_i \epsilon_{1it} \right| \\
&\leq \sqrt{\sum_{i=1}^n w_i^2 \sum_{t=1}^p \mathbb{E} \left\{ \sqrt{\sum_{i=1}^n \epsilon_{1it}^2} \right\}} \\
&\leq \sqrt{\sum_{i=1}^n w_i^2 \sum_{t=1}^p \sqrt{\sum_{i=1}^n \mathbb{E}(\epsilon_{1it}^2)}} \\
&= \sqrt{\sum_{i=1}^n w_i^2 \sum_{t=1}^p \sqrt{n \text{var}(\epsilon_{1it})}},
\end{aligned}$$

where the third inequality follows by Cauchy–Schwarz inequality, the fourth by Jensen’s inequality, and the last equality by the i.i.d. assumption. The same logic applies to $\mathbb{E} \left\| \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q$ and we get

$$\mathbb{E} \left\| \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q \leq \sqrt{\sum_{i=1}^n w_i^2 \sum_{t=1}^p \sqrt{n \text{var}(\epsilon_{0it})}},$$

and thus

$$\mathbb{E} \{d(\tau_r, \hat{\tau}_r)\} \leq \delta \sum_{j=1}^B \|\beta_{rj}\|_q + \sqrt{n \sum_{i=1}^n w_i^2 \sum_{t=1}^p \left\{ \sqrt{\text{var}(\epsilon_{1it})} + \sqrt{\text{var}(\epsilon_{0it})} \right\}}. \quad (16)$$

The condition that each $\text{var}(\epsilon_{ait}) \leq \sigma_\infty^2$ yields the result. Notice that setting $q = 1$ in the above result gives an upper bound for the L_1 error.

Next, provided that for any pair of (a, t) , $a \in \{0, 1\}$, $t \in \{1, \dots, p\}$, the variables ϵ_{ait} , $i = 1, \dots, n$, are independent, mean-zero sub-Gaussian with parameter σ_a , by the Hoeffding bound, we have

$$\mathbb{P} \left\{ \left| \sum_{i=1}^n w_i \epsilon_{ait} \right| \geq \kappa \right\} \leq 2 \exp \left(-\frac{\kappa^2}{2\sigma_a^2 \sum_{i=1}^n w_i^2} \right),$$

for all $\kappa \geq 0$. Hence, with probability at least $1 - \xi$, it holds that

$$\left| \sum_{i=1}^n w_i \epsilon_{ait} \right| \leq \sigma_a \sqrt{2 \log \left(\frac{2}{\xi} \right) \sum_{i=1}^n w_i^2}.$$

Then it is immediately follows that with probability at least $1 - \xi$,

$$\left\| \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} \right\|_q + \left\| \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q \leq p(\sigma_0 + \sigma_1) \sqrt{2 \log \left(\frac{2}{\xi} \right) \sum_{i=1}^n w_i^2}. \quad (17)$$

Hence, by equation 15, we have that

$$d(\tau_r, \hat{\tau}_r) \leq \delta \sum_{j=1}^B \|\beta_{rj}\|_q + p(\sigma_0 + \sigma_1) \sqrt{2 \log \left(\frac{2}{\xi} \right) \sum_{i=1}^n w_i^2}. \quad (18)$$

□

Bounds for $|\hat{D}_{\text{sep}} - D_{\text{sep}}|$ follow directly from the preceding result. We first introduce the auxiliary lemma, followed by the proof of the main result.

Lemma B.1. *For real-valued vectors v_1, v_2, v_3, v_4 , and the associated distance function d , we have*

$$|d(v_1, v_2) - d(v_3, v_4)| \leq d(v_1, v_3) + d(v_2, v_4).$$

Proof. Since D is distance measure, by triangle inequality it follows

$$\begin{aligned} d(v_1, v_2) &\leq d(v_1, v_3) + d(v_3, v_4) + d(v_4, v_2), \\ d(v_3, v_4) &\leq d(v_3, v_1) + d(v_1, v_2) + d(v_2, v_4), \end{aligned}$$

and consequently we obtain

$$|d(v_1, v_2) - d(v_3, v_4)| \leq d(v_1, v_3) + d(v_2, v_4).$$

□

Now we are in a position to prove our main theorem.

Proof. Note that

$$\left| \hat{D}_{\text{disp}} - D_{\text{disp}} \right| = \left| \sum_{r \neq r'} \{d(\tau_r, \tau_{r'}) - d(\hat{\tau}_r, \hat{\tau}_{r'})\} \right|$$

$$\begin{aligned}
&\leq \sum_{r \neq r'} |\{d(\tau_r, \tau_{r'}) - d(\hat{\tau}_r, \hat{\tau}_{r'})\}| \\
&\leq \sum_{r \neq r'} \{d(\tau_r, \hat{\tau}_r) + d(\tau_{r'}, \hat{\tau}_{r'})\} \\
&\leq R \sum_{r=1}^R d(\tau_r, \hat{\tau}_r), \tag{19}
\end{aligned}$$

where the second inequality follows by Lemma B.1.

Hence, from equation 16, we obtain that

$$\mathbb{E} \left| \hat{D}_{\text{disp}} - D_{\text{disp}} \right| \leq \delta R \sum_{r=1}^R \sum_{j=1}^B \|\beta_{rj}\|_q + R^2 \sqrt{n \sum_{i=1}^n w_i^2 \sum_{t=1}^p \left\{ \sqrt{\mathbb{E}(\epsilon_{1t})^2} + \sqrt{\mathbb{E}(\epsilon_{0t})^2} \right\}}.$$

Further, by plugging equation 18 back into equation 19 and equation 15, we conclude that with probability at least $1 - \xi$ the following bound holds:

$$\left| \hat{D}_{\text{disp}} - D_{\text{disp}} \right| \leq \delta R \sum_{r=1}^R \sum_{j=1}^B \|\beta_{rj}\|_q + R^2 p (\sigma_0 + \sigma_1) \sqrt{2 \log \left(\frac{2}{\xi} \right) \sum_{i=1}^n w_i^2}.$$

□

B.2 PROOF OF THEOREM 4.2

We first show that under the given conditions the sample set $\{X_1, \dots, X_n\}$ becomes an ε -cover of \mathcal{X} with high probability approaching 1: i.e.,

$$\mathbb{P}(\{X_1, \dots, X_n\} \text{ is an } \varepsilon\text{-cover of } \mathcal{X}) \geq 1 - \delta,$$

or equivalently, an event

$$\mathcal{E}_{n,\varepsilon,\delta} = \left\{ \omega \in \Omega : \{X_1(\omega), \dots, X_n(\omega)\} \text{ forms an } \varepsilon\text{-cover of } \mathcal{X} \right\}$$

occurs with probability at least $1 - \delta$. Here, an ε -cover means:

$$\forall x \in \mathcal{X}, \min_{1 \leq i \leq n} d(x, X_i(\omega)) \leq \varepsilon, \quad \forall \varepsilon > 0.$$

Lemma B.2 (Coverage in Probability). *Suppose that \mathcal{X} is a compact metric space, and that we draw X_1, \dots, X_n i.i.d. from \mathbb{P} where \mathbb{P} has full support on \mathcal{X} . Then, the set $\{X_1, \dots, X_n\}$ forms an ε -cover of \mathcal{X} with probability approaching 1.*

Proof. Because \mathcal{X} is compact, for each $\varepsilon > 0$ there is a finite set $\{z_1, \dots, z_M\} \subset \mathcal{X}$ such that

$$\mathcal{X} \subset \bigcup_{j=1}^M \mathbb{B}(z_j, \varepsilon),$$

where $\mathbb{B}(x, r)$ denotes the open ball of radius r around x . Since \mathbb{P} has full support, each ball $\mathbb{B}(z_j, \varepsilon)$ has $\mathbb{P}(\mathbb{B}(z_j, \varepsilon)) > 0$. For each j , the probability that no X_i lands in $B(z_j, \varepsilon)$ is

$$\left[1 - \mathbb{P}(\mathbb{B}(z_j, \varepsilon)) \right]^n.$$

Hence, the probability that each $\mathbb{B}(z_j, \varepsilon)$ contains at least one X_i is at least

$$1 - \sum_{j=1}^M \left[1 - \mathbb{P}(\mathbb{B}(z_j, \varepsilon)) \right]^n,$$

which converges to 1 as n grows, as each term $[1 - \mathbb{P}(\mathbb{B}(z_j, \varepsilon))]^n \rightarrow 0$ as $n \rightarrow \infty$.

If at least one X_i lands in each $B(z_j, \varepsilon)$, then the entire \mathcal{X} is covered by the sample's ε -balls:

$$\mathcal{X} \subset \bigcup_{j=1}^M B(z_j, \varepsilon) \subset \bigcup_{i=1}^n B(X_i, 2\varepsilon).$$

Hence with probability approaching 1, the set $\{X_1, \dots, X_n\}$ forms a 2ε -cover of \mathcal{X} , i.e.,

$$\mathbb{P}(\{X_1, \dots, X_n\} \text{ is a } 2\varepsilon\text{-cover of } \mathcal{X}) \rightarrow 1.$$

Since ε is chosen arbitrarily, this yields the desired result. \square

The above property shown in Lemma B.2 is often called *coverage in probability*. Now, we shall show that for any function $f \in \mathcal{H}_K$, there always exists a partial sum $\sum_{i=1}^n \alpha_i K(x, X_i)$ that can approximate f arbitrarily well in probability.

Lemma B.3. Consider $f \in \mathcal{H}_K$, and $X_1, \dots, X_n \stackrel{i.i.d}{\sim} \mathbb{P}$ where \mathbb{P} has full support on a compact metric space \mathcal{X} . Then, for each $x \in \mathcal{X}$, there exists a partial sum

$$S_n(x) = \sum_{i=1}^n \alpha_i K(x, X_i)$$

such that $|f(x) - S_n(x)| = o_{\mathbb{P}}(1)$, for some coefficients $\{\alpha_i\}$ depending on both n and $\{X_1, \dots, X_n\}$.

Proof. As $f \in \mathcal{H}_K$, by construction, we have the H_K -norm convergence:

$$\left\| f - \sum_{j=1}^m \beta_j K(\cdot, z_j) \right\|_{H_K} \rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

Pick an arbitrary small $\epsilon > 0$. Then there exist finite partial sums of the form $\sum_{j=1}^{m'} \beta_j K(\cdot, z_j)$, for each $z_j \in \mathcal{X}$ and some $M < \infty$, such that

$$\left\| f - \sum_{j=1}^M \beta_j K(\cdot, z_j) \right\|_{H_K} \leq \epsilon.$$

If a sample $\{X_1(\omega), \dots, X_n(\omega)\}$ is an ε -cover, each chosen center z_j is within distance ε of some $X_{i_j}(\omega)$. Thus, we may construct S_n from $\sum_{j=1}^M \beta_j K(\cdot, z_j)$ by sliding each $z_j, j = 1, \dots, M$, to the nearest sample point X_{i_j} . Also by Lipschitz continuity of K , when $d(z_j, X_{i_j}) \leq \varepsilon$, we have

$$\|K(\cdot, z_j) - K(\cdot, X_{i_j})\|_{H_K} \leq L_K \varepsilon.$$

Now, letting

$$S_n(\omega; \cdot) = \sum_{i=1}^n \alpha_i(\omega) K(\cdot, X_i(\omega)),$$

where each α_i is obtained via linear mapping from $\{\beta_j\}_{j=1}^M$, on the event $\mathcal{E}_n(\omega) = \{\omega : \{X_1(\omega), \dots, X_n(\omega)\} \text{ is a } \varepsilon\text{-cover}\}$, we have that

$$\begin{aligned} \left\| \sum_{j=1}^M \beta_j K(\cdot, z_j) - S_n(\omega; \cdot) \right\|_{H_K} &= \left\| \sum_{j=1}^M \beta_j (K(\cdot, z_j) - K(\cdot, X_{i_j})) \right\|_{H_K} \\ &\leq \sqrt{\sum_{j=1}^M \beta_j^2} \left\| K(\cdot, z_j) - K(\cdot, X_{i_j}) \right\|_{H_K} \end{aligned}$$

$$\leq L_K \varepsilon \sqrt{\sum_{j=1}^M \beta_j^2},$$

where the second and the third inequalities follow by Cauchy–Schwarz and Lipschitz continuity of K . This allows us to pick partial sums with centers $X_{i_j}(\omega)$ that are close (in $\|\cdot\|_{H_K}$) to the “ideal” partial sums with centers z_j . Consequently, we have that

$$\begin{aligned} \|f - S_n\|_{H_K} &\leq \left\| f - \sum_{j=1}^M \beta_j K(\cdot, z_j) \right\|_{H_K} + \left\| \sum_{j=1}^M \beta_j K(\cdot, z_j) - S_n \right\|_{H_K} \\ &\leq \epsilon + L_K \varepsilon \sqrt{\sum_{j=1}^M \beta_j^2}. \end{aligned}$$

Let $\epsilon' = \epsilon + L_K \varepsilon \sqrt{\sum_{j=1}^M \beta_j^2}$. By Lemma B.2, it follows that $\mathbb{P}(\mathcal{E}_n) \rightarrow 1$. Thus, we have

$$\mathbb{P}\left(\exists S_n(\omega; \cdot) \text{ s.t. } \|f - S_n(\omega; \cdot)\|_{H_K} < \epsilon'\right) \geq P(\mathcal{E}_n) \rightarrow 1. \quad (20)$$

Since ϵ and ε are arbitrary, this shows that one may find a partial sum $S_n(\omega; \cdot)$ in the span of $\{K(\cdot, X_i)\}$ that approximates f arbitrarily well in $\|\cdot\|_{H_K}$, with high probability.

Now, for each $x \in \mathcal{X}$, by the reproducing property

$$f(x) - S_n(\omega; x) = \langle f - S_n(\omega; \cdot), K(\cdot, x) \rangle_{H_K}.$$

Hence by Cauchy–Schwarz

$$|f(x) - S_n(\omega; x)| \leq \|f - S_n(\omega; \cdot)\|_{H_K} \cdot \|K(\cdot, x)\|_{H_K}.$$

By the boundedness assumption: $K(x, x) \leq C$ for all $x \in \mathcal{X}$, it follows that $\|K(\cdot, x)\|_{H_K}^2 = K(x, x) \leq C \implies \|K(\cdot, x)\|_{H_K} \leq \sqrt{C}$. Hence,

$$|f(x) - S_n(\omega; x)| \leq \sqrt{C} \|f - S_n(\omega; \cdot)\|_{H_K}.$$

From equation 20, we have

$$\mathbb{P}\left(\|f - S_n(\omega; \cdot)\|_{H_K} \geq \epsilon'\right) \rightarrow 0 \text{ as } n \rightarrow \infty,$$

for any $\epsilon' > 0$. Take $\epsilon' = \frac{\epsilon''}{\sqrt{C}}$ for arbitrarily small $\epsilon'' > 0$. Then

$$\mathbb{P}\left(|f(x) - S_n(\omega; x)| > \epsilon''\right) \leq \mathbb{P}\left(\|f - S_n(\omega; \cdot)\|_{H_K} \geq \epsilon'\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Thus, for each fixed $x \in \mathcal{X}$, $|f(x) - S_n(\omega; x)| \rightarrow 0$ in probability: i.e.,

$$|f(x) - S_n(x)| = o_{\mathbb{P}}(1).$$

□

We are now ready to prove Theorem 4.2.

Proof. Based on the same logic used to obtain equation 15, it is immediate to see that for the the L_q distance d , it follows

$$d(\tau_r, \hat{\tau}_r) \leq \left\| \sum_{i=1}^n A_i S_{ir} w_i m_r(X_i) - \sum_{i=1}^n (1 - A_i) S_{ir} w_i m_r(X_i) \right\|_q + \left\| \sum_{i=1}^n A_i S_{ir} w_i \epsilon_{1i} \right\|_q + \left\| \sum_{i=1}^n (1 - A_i) S_{ir} w_i \epsilon_{0i} \right\|_q,$$

where we let $m(x, S_r = 1) \equiv m_r(x)$ for notational simplicity. By Lemma B.3, there exists $S_n = \sum_{j=1}^n \alpha_j K(x, X_j)$ such that $|m_r(x) - S_n(x)| = o_{\mathbb{P}}(1)$, $\forall x \in \mathcal{X}$. Thus, for the first term on the RHS, we have

$$\begin{aligned} & \left\| \sum_{i=1}^n A_i S_{ir} w_i m_r(X_i) - \sum_{i=1}^n (1 - A_i) S_{ir} w_i m_r(X_i) \right\|_q \\ & \leq \left\| \sum_{i=1}^n A_i S_{ir} w_i \left\{ m_r(X_i) - \sum_{j=1}^n K_{ij} \right\} - \sum_{i=1}^n (1 - A_i) S_{ir} w_i \left\{ m_r(X_i) - \sum_{j=1}^n \alpha_j K_{ij} \right\} \right\|_q \\ & \quad + \left\| \sum_{j=1}^n \alpha_j \left\{ \sum_{i=1}^n A_i S_{ir} w_i K_{ij} - \sum_{i=1}^n (1 - A_i) S_{ir} w_i K_{ij} \right\} \right\|_q \\ & = O \left(\delta \sqrt{\sum_{j=1}^n \alpha_j^2} \right) + o_{\mathbb{P}}(1), \end{aligned}$$

where the last equality follows by the balancing condition equation 7c and the Cauchy–Schwarz inequality.

By the same Hoeffding’s bound argument used in the proof of Theorem 4.1, it follows that $\forall a$,

$$\left\| \sum_{i=1}^n \mathbb{1}(A_i = a) w_i \epsilon_{ai} \right\|_q = O_{\mathbb{P}} \left(\sqrt{\sum_{i=1}^n w_i^2} \right).$$

Putting all the pieces together, finally we obtain that

$$d(\tau_r, \hat{\tau}_r) = O \left(\delta \sum_{j=1}^n \alpha_j \right) + O_{\mathbb{P}} \left(\sqrt{\sum_{i=1}^n w_i^2} \right) + o_{\mathbb{P}}(1),$$

which also yields the same bounds for $|\hat{D}_{\text{sep}} - D_{\text{sep}}|$, based on the same logic used in the proof of Theorem 4.1. \square

B.3 PROOF OF THEOREM 4.3

Proof. It suffices to focus on the estimand:

$$\tau_{a,r} = \mathbb{E} \{ \mu_{a,r}(X) \mid S_r = 1 \} = \frac{\mathbb{E} \{ \mu_{a,r}(X) S_r \}}{\mathbb{E}(S_r)},$$

and the case that $d = 1$. (The result for the general L_q distance follows directly from the structure of the previous proof.)

Under the outcome model equation 10, we consider the following decomposition: for $\forall a, r$,

$$\begin{aligned} \hat{\tau}_{a,r} - \tau_{a,r} &= \underbrace{\sum_{i=1}^n \mathbb{1}(A_i = a) S_{ir} w_i \mu_{a,r}(X_i) - \frac{\frac{1}{n} \sum_{i=1}^n S_{ir} \mu_{a,r}(X_i)}{\mathbb{P}_n(S_r)}}_{(i)} \\ & \quad + \underbrace{\sum_{i=1}^n \mathbb{1}(A_i = a) S_{ir} w_i \epsilon_i}_{(ii)} \\ & \quad + \underbrace{\frac{\mathbb{P}_n \{ \mu_{a,r}(X) S_r \}}{\mathbb{P}_n(S_r)} - \frac{\mathbb{E} \{ \mu_{a,r}(X) S_r \}}{\mathbb{E}(S_r)}}_{(iii)}. \end{aligned}$$

We will analyze the above three terms sequentially.

(i) Since $\mu_{a,r} \in \mathcal{H}_k$, by Lemma B.3, under the given conditions we may write

$$\mu_{a,r}(x) = \sum_{i=1}^n \alpha_i K(X_i, x) + o_{\mathbb{P}}(1),$$

for some $\{\alpha_i\}$ depending on n and $\{X_i\}$. Hence, we get

$$\begin{aligned} & \sum_{i=1}^n \mathbb{1}(A_i = a) S_{ir} w_i \mu_{a,r}(X_i) - \frac{\frac{1}{n} \sum_{i=1}^n S_{ir} \mu_{a,r}(X_i)}{\mathbb{P}_n(S_r)} \\ &= \sum_{j=1}^n \alpha_j \left\{ \sum_{i=1}^n \mathbb{1}(A_i = a) S_{ir} w_i K_{ij} - \frac{\frac{1}{n} \sum_{i=1}^n S_{ir} K_{ij}}{\mathbb{P}_n(S_r)} \right\} \\ &+ \sum_{i=1}^n \mathbb{1}(A_i = a) S_{ir} w_i \left\{ \mu_{a,r}(X_i) - \sum_{j=1}^n \alpha_j K_{ij} \right\} - \frac{\frac{1}{n} \sum_{i=1}^n S_{ir} \left\{ \mu_{a,r}(X_i) - \sum_{j=1}^n \alpha_j K_{ij} \right\}}{\mathbb{P}_n(S_r)} \\ &= O \left(\delta \sqrt{\sum_{j=1}^n \alpha_j^2} \right) + o_{\mathbb{P}}(1), \end{aligned}$$

where the last equality follows by the balancing condition equation 11, the Cauchy–Schwarz inequality, and the fact that $\frac{1}{\mathbb{P}_n(S_r)} \leq \frac{1}{m_0} < \infty$.

(ii). By the same Hoeffding’s inequality argument as in the proof of Theorem 4.1, the term in (ii) is $O_{\mathbb{P}} \left(\sqrt{\sum_{i=1}^n w_i^2} \right)$.

(iii). It is immediate to see that by the weak law of large numbers and the continuous mapping theorem, the terms in (iii) converge to zero in probability.

Putting all the pieces together yields the desired result that

$$d(\tau_r, \hat{\tau}_r) = O \left(\delta \sqrt{\sum_{j=1}^n \alpha_j^2} \right) + O_{\mathbb{P}} \left(\sqrt{\sum_{i=1}^n w_i^2} \right) + o_{\mathbb{P}}(1),$$

□

B.4 PROOF OF PROPOSITION 4.4

Before proving Proposition 4.4, we provide some technical results first.

Definition B.1. Let $\lfloor \beta \rfloor$ denote the greatest integer strictly less than $\beta \in \mathbb{R}$. Given a vector $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{R}^d$ and a multivariate function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, define $D^{(\alpha)} f = \frac{\partial^{\alpha_1 + \dots + \alpha_d}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} f$, the order- α partial derivatives of f . Then we define the Hölder class $\Sigma(\beta, L)$ on \mathbb{R}^d defined as the set of $l = \lfloor \beta \rfloor$ times differentiable functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that for some positive numbers β, L ,

$$|D^{(l)} f(v) - D^{(l)} f(w)| \leq L \|v - w\|_1^{\beta-l}, \quad \forall v, w \in \mathbb{R}^d.$$

We also define the associated Hölder class of densities by

$$\mathcal{P}_{\Sigma}(\beta, L) := \left\{ p \mid p \geq 0, \int p(v) dv = 1, \text{ and } p \in \Sigma(\beta, L) \right\}.$$

Definition B.2. For an integer $l \geq 1$, we say that k is a kernel of order l if

$$\int k(u) du = 1, \quad \int u^j k(u) du = 0, \quad j = 1, \dots, l.$$

See Tsybakov (2010, Section 1.2) for examples of the kernel of order l .

Lemma B.4. Let $p_{r,h} = \mathbb{E}[\hat{p}_{r,h}]$, and suppose $p_r(s) \leq p_{\max}$ for some $p_{\max} < \infty$, $\forall x \in \mathcal{X}$. Then under the same assumptions as Proposition 4.4,

$$\mathbb{E} \left[\{ \hat{p}_{r,h}(x) - p_{r,h}(x) \}^2 \right] \leq \frac{C_{K,p_{\max}}}{n\mathbb{P}(S_r = 1)h^d},$$

where $C_{K,p_{\max}}$ is a constant depending only on $\|K\|_2$ and p_{\max} .

Proof. Note that $(\hat{p}_{r,h}(x) - p_{r,h}(x))^2$ can be expanded as

$$\begin{aligned} (\hat{p}_{r,h}(x) - p_{r,h}(x))^2 &= \left(\frac{\sum_{i=1}^n T_{h,x}(X_i) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \mathbb{1}(n_r > 0) - p_{r,h}(x) \right)^2 \\ &= \left(\frac{\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right)^2 \mathbb{1}(n_r > 0) + p_{r,h}(x)^2 \mathbb{1}(n_r = 0). \end{aligned} \quad (21)$$

For the first term of equation 21, we have that

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right)^2 \mathbb{1}(n_r > 0) \mid S \right] \\ &= \frac{\mathbb{E} \left[\left(\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1) \right)^2 \mid S \right]}{(\sum_{i=1}^n \mathbb{1}(S_{ir} = 1))^2} \mathbb{1}(n_r > 0) \\ &= \frac{\sum_{i=1}^n \mathbb{E} \left[(T_{h,x}(X_i) - p_{r,h}(x))^2 \mid S \right] \mathbb{1}(S_{ir} = 1)}{(\sum_{i=1}^n \mathbb{1}(S_{ir} = 1))^2} \mathbb{1}(n_r > 0) \\ &\quad + \frac{\sum_{i \neq j} \mathbb{E} \left[(T_{h,x}(X_i) - p_{r,h}(x))(T_{h,x}(X_j) - p_{r,h}(x)) \mid S \right] \mathbb{1}(S_{ir} = 1) \mathbb{1}(S_{jr} = 1)}{(\sum_{i=1}^n \mathbb{1}(S_{ir} = 1))^2} \mathbb{1}(n_r > 0). \end{aligned} \quad (22)$$

Then by Tsybakov (2010, Proposition 1.1),

$$\mathbb{E} \left[(T_{h,x}(X_i) - p_{r,h}(x))^2 \mid S \right] \leq \text{var}(T_{h,x}(X_i)) \leq \frac{p_{\max} \|K\|_2^2}{h^d}. \quad (23)$$

Also, for $i \neq j$, it follows that

$$\begin{aligned} \mathbb{E} \left[(T_{h,x}(X_i) - p_{r,h}(x))(T_{h,x}(X_j) - p_{r,h}(x)) \mid S \right] &= \mathbb{E} \left[(T_{h,x}(X_i) - p_{r,h}(x))(T_{h,x}(X_j) - p_{r,h}(x)) \right] \\ &= \mathbb{E} [T_{h,x}(X_i) - p_{r,h}(x)] \mathbb{E} [T_{h,x}(X_j) - p_{r,h}(x)] \\ &= 0. \end{aligned} \quad (24)$$

Applying equation 23 and equation 24 to equation 22 yields

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right)^2 \mathbb{1}(n_r > 0) \mid S \right] \\ &\leq \frac{\sum_{i=1}^n \frac{p_{\max} \|K\|_2^2}{h^d} \mathbb{1}(S_{ir} = 1)}{(\sum_{i=1}^n \mathbb{1}(S_{ir} = 1))^2} \mathbb{1}(n_r > 0) \\ &= \frac{p_{\max} \|K\|_2^2}{h^d} \frac{\mathbb{1}(n_r > 0)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)}. \end{aligned} \quad (25)$$

Next, applying Györfi et al. (2002, Lemma 4.1) to equation 25 gives the bound for the first term of equation 21 as

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right)^2 \mathbb{1}(n_r > 0) \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\left(\frac{\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right)^2 \mathbb{1}(n_r > 0) \mid S \right] \right] \end{aligned}$$

$$\begin{aligned}
&\leq \frac{p_{\max} \|K\|_2^2}{h^d} \mathbb{E} \left[\frac{\mathbb{1}(n_r > 0)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right] \\
&\leq \frac{2p_{\max} \|K\|_2^2}{(n+1)h^d \mathbb{P}(S_r = 1)}. \tag{26}
\end{aligned}$$

Also, by applying Proposition 1.1. from Tsybakov (2010), the second term of equation 21 can be calculated as

$$\mathbb{E} [p_{r,h}(x)^2 \mathbb{1}(n_r = 0)] = p_{r,h}(x)^2 \mathbb{P}(n_r = 0) \leq \frac{p_{\max} \|K\|_2^2}{h^d} (1 - \mathbb{P}(S_r = 1))^n. \tag{27}$$

Now, applying equation 26 and equation 27 to equation 21 gives

$$\begin{aligned}
&\mathbb{E} [(\hat{p}_{r,h}(x) - p_{r,h}(x))^2] \\
&= \mathbb{E} \left[\left(\frac{\sum_{i=1}^n (T_{h,x}(X_i) - p_{r,h}(x)) \mathbb{1}(S_{ir} = 1)}{\sum_{i=1}^n \mathbb{1}(S_{ir} = 1)} \right)^2 \mathbb{1}(n_r > 0) \right] + \mathbb{E} [p_{r,h}(x)^2 \mathbb{1}(n_r = 0)] \\
&\leq \frac{p_{\max} \|K\|_2^2}{h^d} \left(\frac{2}{(n+1)\mathbb{P}(S_r = 1)} + (1 - \mathbb{P}(S_r = 1))^n \right) \\
&\leq \frac{p_{\max} \|K\|_2^2}{h^d} \left(\frac{2}{n\mathbb{P}(S_r = 1)} + \exp(-n\mathbb{P}(S_r = 1)) \right) \\
&\leq \frac{3p_{\max} \|K\|_2^2}{n\mathbb{P}(S_r = 1)h^d} \\
&= \frac{C_{K,p_{\max}}}{n\mathbb{P}(S_r = 1)h^d}.
\end{aligned}$$

□

Lemma B.5. *Under the same assumptions of Proposition 4.4,*

$$|p_{r,h}(x) - p_r(x)|^2 \lesssim h^{2\beta}.$$

Proof. The proof mimics the proof of Tsybakov (2010, Proposition 1.2). Since $p_r \in \mathcal{P}_\Sigma(\beta, L)$ and K is a kernel of order l , using the multi-index notation, by Taylor's theorem one would obtain that $\forall y \in \mathcal{Y}$,

$$\begin{aligned}
|p_{r,h}(x) - p_r(x)| &\leq \int |K(u)| \sum_{\substack{|\alpha|=l \\ \alpha \in \mathbb{Z}_0^{+d}}} \frac{l}{\alpha!} \|uh\|_1^\alpha \int_0^1 (1-\tau)^{(l-1)} |D^\alpha p_r(y + \tau uh) - D^\alpha p_r(y)| d\tau du \\
&\leq \int |K(u)| \sum_{\substack{|\alpha|=l \\ \alpha \in \mathbb{Z}_0^{+d}}} \frac{l}{\alpha!} \|uh\|_1^l \left\{ \int_0^1 (1-\tau)^{(l-1)} L \|\tau uh\|_1^{\beta-l} d\tau \right\} du \\
&\leq h^\beta \left(\sum_{\substack{|\alpha|=l \\ \alpha \in \mathbb{Z}_0^{+d}}} \frac{Ll}{\alpha!} \int |K(u)| \|u\|_1^\beta du \int_0^1 (1-\tau)^{(l-1)} \tau^{\beta-l} d\tau \right)
\end{aligned}$$

Hence, provided that $\int |K(u)| \|u\|_1^\beta du < \infty$, we have $|p_{r,h}(x) - p_r(x)|^2 \lesssim h^{2\beta}$, where we use the shorthand $a_n \lesssim b_n$ to denote $a_n \leq Cb_n$ for some universal constant $C > 0$.

□

Remark 1. *The condition $\int |K(u)| \|u\|_1^\beta du < \infty$ does, in fact, imply the bounded-density condition $p_r(x) \leq p_{\max} < \infty$, $\forall x \in \mathcal{X}$; it could be shown that one may set $p_{\max} = O\left(\int |K(u)| \|u\|_1^\beta du + \|K\|_\infty\right)$.*

Now we turn to the proof of Proposition 4.4.

Proof. To analyze the proposed estimator in equation 13, we first compute the L_2 error bounds for the kernel density estimator $\hat{p}_{r,h}$. Note that under the given conditions there exists a constant $p_{\max} < \infty$ such that

$$\sup_{x \in \mathcal{X}} \sup_{p_r \in \mathcal{P}_\Sigma(\beta, L)} p_r(x) \leq p_{\max}.$$

(Remark 1). Thus, by Lemmas B.4 and B.5, it immediately follows that $\forall x \in \mathcal{X}$ and $\forall r \in \{1, \dots, R\}$,

$$\begin{aligned} \mathbb{E} \left[\{\hat{p}_{r,h}(x) - p_r(x)\}^2 \right] &\leq \mathbb{E} \left[\{\hat{p}_{r,h}(x) - p_{r,h}(x)\}^2 \right] + \{p_{r,h}(x) - p_r(x)\}^2 \\ &\lesssim \frac{1}{nh^d} + h^{2\beta}. \end{aligned}$$

Hence, under the condition that $n^{-1}h^{-d} + h^{2\beta} = o(1)$, we obtain $\hat{p}_{r,h}(x) \xrightarrow{p} p_r(x)$. Also by the given condition, $\mathbb{P}_n\{S_r\} \xrightarrow{p} \mathbb{P}(S_r = 1)$. Then by the continuous mapping theorem, we get the desired consistency of $\hat{\mathbb{P}}(S_r = 1 \mid X = x)$ and \hat{r}^* with respect to the distribution \mathbb{P} where the pmf p_S is defined with the parameter $\mathbf{p}^* = \arg \max_{\mathbf{p}} \sum_{r \neq r'} d(\tau_r, \tau_{r'}; p_S)$.

□