

8 Web Appendix

8.1 On the Identifying Assumptions for the Target Average Treatment Effect

We now discuss the assumptions for identification of τ , as given in Assumption 1 in Section 2. We can weaken Assumption 1(b) to only require $E_{\mathbb{P}_1}\{Y(z)|\mathbf{X}\} = E_{\mathbb{P}_0}\{Y(z)|\mathbf{X}\}$, $z \in \{0, 1\}$, i.e., conditional mean exchangeability of the potential outcomes across treatment groups. Also, we can weaken Assumption 1(d) to require $E_{\mathbb{T}}\{Y(1) - Y(0)|\mathbf{X}\} = E_{\mathbb{P}}\{Y(1) - Y(0)|\mathbf{X}\}$, i.e., conditional mean exchangeability of the treatment effect across the study and target populations. In this paper, we consider somewhat stronger assumptions as they allow us to identify general features of the distributions of the potential outcomes in the target population, e.g., $E_{\mathbb{T}}[g\{Y(z)\}]$, $z \in \{0, 1\}$, and $\tilde{g}[E_{\mathbb{T}}\{Y(1)\}, E_{\mathbb{T}}\{Y(0)\}]$ for measurable functions $g(\cdot)$ and $\tilde{g}(\cdot)$.

By Assumption 1(d), controlling for the observed covariates is sufficient to remove differences in potential outcome distributions between the study and the target populations. Importantly, because Assumptions 1(c) and 1(d) are based completely on the distributions of $\{Y(1), Y(0), \mathbf{X}\}$ in the study and the target populations, they can pertain to a variety of generalization and transportation settings. These include those where the study and target populations are disjoint (e.g., transportation) and those where the study population is nested within the target (e.g., generalization).

8.2 Connection to inverse propensity weighting

8.2.1 The dual optimization problem

In this section, we formally discuss the equivalence between one-step balancing weights and inverse propensity modeling weights. Recall that the primal optimization problem for the one-step balancing weights is given by

$$\operatorname{argmin}_{\mathbf{w}} \left\{ \sum_{i:Z_i=z} \psi(w_i) : \left| \sum_{i:Z_i=z} w_i B_k(\mathbf{X}_i) - \bar{B}_k^* \right| \leq \delta_k, \ k = 1, 2, \dots, K; \sum_{i:Z_i=z} w_i = 1 \right\} \quad (8)$$

Throughout this section, we consider generalization in the nested study setting, because traditional inverse propensity weighting methods have only been developed for settings with a formal study selection process and selection indicator variable. Here, the study population \mathcal{P} is nested within the target population \mathcal{T} , with D indicating selection from \mathcal{T} into \mathcal{P} . We assume that the study sample of size n is nested within a random sample of size n^* from \mathcal{T} . For a unit in group $Z = z$ with covariate vector \mathbf{x} , let $\hat{w}(\mathbf{x})$ and $w^{\text{IP}}(\mathbf{x})$ be its one-step balancing weight and the true inverse propensity weight, respectively, where the one-step balancing weights are obtained via (8). In Theorem 8.1, we show that the one-step balancing weights estimate the inverse propensity weights under a specific functional form of the inverse propensity weights and a loss function. Theorem 8.1 generalizes the results of Wang and Zubizarreta (2020) and Chattopadhyay et al. (2020), where similar connections between minimal and inverse propensity weights have been established in missing data and causal inference settings, respectively.

Theorem 8.1.

- (a) The dual problem of (8) is equivalent to the empirical loss minimization problem with L_1 regularization:

$$\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad (n^*)^{-1} \sum_{i=1}^{n^*} \left[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} + \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \right] + |\boldsymbol{\lambda}|^\top \boldsymbol{\delta}, \quad (9)$$

where $\boldsymbol{\lambda}$ is a $K \times 1$ vector of dual variables corresponding to the K balancing constraints, $|\boldsymbol{\lambda}|$ is the vector of component-wise absolute values of $\boldsymbol{\lambda}$, and $\rho(t) = t/n^* - t(h')^{-1}(t) - h((h')^{-1}(t))$, with $h(t) = \psi(1/n^* - t)$.

- (b) If $\hat{\mathbf{w}}$ and $\boldsymbol{\lambda}_z^\dagger$ are solutions to the primal and dual forms of (8), respectively, then for $i : Z_i = z$,

$$\hat{w}_i = \rho'\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}_z^\dagger\}. \quad (10)$$

- (c) If $\tilde{\boldsymbol{\lambda}}_z \in \underset{\boldsymbol{\lambda}}{\text{argmin}} E_{\mathbb{T}}[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} + \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} | \mathbf{X}_i = \mathbf{x}]$, then $\tilde{\boldsymbol{\lambda}}_z$ satisfies

$$\rho'\{\mathbf{B}(\mathbf{x})^\top \tilde{\boldsymbol{\lambda}}_z\} = \{n^* \pi(\mathbf{x}) \mathbb{P}(Z_i = z | \mathbf{X}_i = \mathbf{x})\}^{-1} = (n^*)^{-1} w^{\text{IP}}(\mathbf{x}). \quad (11)$$

Parts (a) and (b) of Theorem 8.1 provide an alternative approach to obtain the one-step balancing weights via the dual form of the optimization problem in (8), and part (c) links the form of the dual objective to inverse propensity weighting. Formally, suppose the true inverse propensity weights in group $Z = z$ have the functional form $w^{\text{IP}}(\mathbf{x}) = n^* \rho' \{ \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda} \}$ for some parameter $\boldsymbol{\lambda}$, and we estimate $\boldsymbol{\lambda}$ by minimizing the regularized empirical loss function in (9). Equation 10 implies that the resulting estimated inverse propensity weights are proportional to the one-step balancing weights in the sense that $\hat{w}_i^{\text{IP}} = n^* \hat{w}_i$. Importantly, estimation of the inverse propensity weights here does not involve separate specifications of the treatment assignment and the study selection models. Theorem 8.1 thus shows that our proposed method is equivalent to a one-step estimation method of the inverse propensity weights via a loss function that directly addresses covariate balance relative to the target and simultaneously ensures that the weights are less dispersed.

8.2.2 Proof of Theorem 8.1

We first consider the optimization problem without the normalization constraint and with $\delta_k > 0$ for $k \in \{1, 2, \dots, K\}$. The k th balancing constraint in the optimization problem in (9) can be written as,

$$\begin{aligned}
& \left| \sum_{i:Z_i=z} w_i B_k(\mathbf{X}_i) - \bar{B}_k^* \right| \leq \delta_k \\
\implies & \left| \sum_{i=1}^{n^*} \mathbb{1}(Z_i = z) D_i w_i B_k(\mathbf{X}_i) - (n^*)^{-1} \sum_{i=1}^{n^*} B_k(\mathbf{X}_i) \right| \leq \delta_k \\
\implies & \left| \sum_{i=1}^{n^*} \{ (n^*)^{-1} - \mathbb{1}(Z_i = z) D_i w_i \} B_k(\mathbf{X}_i) \right| \leq \delta_k \\
\implies & \left| \sum_{i=1}^{n^*} \xi_i B_k(\mathbf{X}_i) \right| \leq \delta_k
\end{aligned} \tag{12}$$

where $\xi_i = 1/n^* - \mathbb{1}(Z_i = z) D_i w_i$. Thus, for the units in group $Z_i = z$ of the sample, $w_i = (1/n^* - \xi_i)$.

For the objective function $\sum_{i:Z_i=z} \psi(w_i)$, we have

$$\sum_{i:Z_i=z} \psi(w_i) = \sum_{i=1}^{n^*} \mathbb{1}(Z_i = z) D_i h(\xi_i), \tag{13}$$

where $h(x) = \psi(1/n^* - x)$. We let $\underline{\mathbf{A}}$ be a $K \times n$ matrix whose (i, j) th element is $B_i(\mathbf{X}_j)$; $\underline{\mathbf{Q}} = (\underline{\mathbf{A}}^\top, -\underline{\mathbf{A}}^\top)^\top$; and $\mathbf{d} = (\boldsymbol{\delta}, \boldsymbol{\delta})^\top$. We can write the primal problem as

$$\begin{aligned} & \underset{\boldsymbol{\xi}}{\text{minimize}} && \sum_{i=1}^{n^*} \mathbb{1}(Z_i = z) D_i h(\xi_i) \\ & \text{subject to} && \underline{\mathbf{Q}} \boldsymbol{\xi} \leq \mathbf{d} \end{aligned} \tag{14}$$

This gives us a convex optimization problem in $\boldsymbol{\xi}$ with linear constraints. The Lagrange dual function of the primal problem in Equation 14 is given by $\inf_{\boldsymbol{\xi}} \{ \sum_{i=1}^{n^*} \mathbb{1}(Z_i = z) D_i h(\xi_i) + \boldsymbol{\lambda}^\top \underline{\mathbf{Q}} \boldsymbol{\xi} \} - \boldsymbol{\lambda}^\top \mathbf{d}$ (see, e.g., [Boyd and Vandenberghe 2004](#) Chapter 5). Let \mathbf{Q}_i be the i th column of $\underline{\mathbf{Q}}$. The dual objective function can be written as

$$\begin{aligned} & - \sup_{\boldsymbol{\xi}} \{ - \sum_{i=1}^{n^*} \mathbb{1}(Z_i = z) D_i h(\xi_i) - \boldsymbol{\lambda}^\top \underline{\mathbf{Q}} \boldsymbol{\xi} \} - \boldsymbol{\lambda}^\top \mathbf{d} \\ & = - \sum_{i=1}^{n^*} \sup_{\xi_i} \{ - (\mathbf{Q}_i^\top \boldsymbol{\lambda}) \xi_i - \mathbb{1}(Z_i = z) D_i h(\xi_i) \} - \boldsymbol{\lambda}^\top \mathbf{d} \\ & = \sum_{i=1}^{n^*} \{ -h_i^*(-\mathbf{Q}_i^\top \boldsymbol{\lambda}) \} - \boldsymbol{\lambda}^\top \mathbf{d}, \end{aligned}$$

where $h_i^*(\cdot)$ is the convex conjugate of $\mathbb{1}(Z_i = z) D_i h(\cdot)$. Thus, the dual problem is given by

$$\begin{aligned} & \underset{\boldsymbol{\lambda}}{\text{maximize}} && \sum_{i=1}^{n^*} \{ -h_i^*(-\mathbf{Q}_i^\top \boldsymbol{\lambda}) \} - \boldsymbol{\lambda}^\top \mathbf{d} \\ & \text{subject to} && \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{15}$$

Since the last K components of \mathbf{Q}_i are reflected versions of the first K components, by symmetry we can write the dual problem as,

$$\begin{aligned} & \underset{\boldsymbol{\lambda}}{\text{maximize}} && \sum_{i=1}^{n^*} \{ -h_i^*(\mathbf{Q}_i^\top \boldsymbol{\lambda}) \} - \boldsymbol{\lambda}^\top \mathbf{d} \\ & \text{subject to} && \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{16}$$

Now,

$$\begin{aligned}
h_i^*(t) &= \sup_{\xi_i} \{t\xi_i - \mathbb{1}(Z_i = z)D_i h(\xi_i)\} \\
&= \sup_{w_i} [\{1/n^* - \mathbb{1}(Z_i = z)D_i w_i\}t - \mathbb{1}(Z_i = z)D_i h\{1/n^* - \mathbb{1}(Z_i = z)D_i w_i\}] \\
&= \sup_{w_i} [\{1/n^* - \mathbb{1}(Z_i = z)D_i w_i\}t - \mathbb{1}(Z_i = z)D_i h(1/n^* - w_i)] \\
&= \{1/n^* - \mathbb{1}(Z_i = z)D_i \hat{w}_i(t)\}t - \mathbb{1}(Z_i = z)D_i h(1/n^* - \hat{w}_i(t)), \tag{17}
\end{aligned}$$

where $\hat{w}_i(t)$ satisfies

$$\left. \partial/\partial w_i [\{1/n^* - \mathbb{1}(Z_i = z)D_i w_i\}t - \mathbb{1}(Z_i = z)D_i h(1/n^* - w_i)] \right|_{w_i = \hat{w}_i(t)} = 0. \tag{18}$$

Solving for w_i , we get $\hat{w}_i(t) = 1/n^* - (h')^{-1}(t)$. Therefore, the dual problem boils down to,

$$\begin{aligned}
&\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad \sum_{i=1}^{n^*} \left\{ -\mathbb{1}(Z_i = z)D_i \rho(\mathbf{Q}_i^\top \boldsymbol{\lambda}) + 1/n^* \mathbf{Q}_i^\top \boldsymbol{\lambda} \right\} + \boldsymbol{\lambda}^\top \mathbf{d} \\
&\text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}, \tag{19}
\end{aligned}$$

where $\rho(t) = t/n^* - t(h')^{-1}(t) + h\{(h')^{-1}(t)\}$. Note that $\rho'(t) = 1/n^* - (h')^{-1}(t)$. Therefore,

$$\hat{w}_i(t) = \rho'(t). \tag{20}$$

Following the proof structure of Theorem 1 in [Wang and Zubizarreta \(2020\)](#), we write $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_+^\top, \boldsymbol{\lambda}_-^\top)^\top$, where $\boldsymbol{\lambda}_+$ and $\boldsymbol{\lambda}_-$ are $K \times 1$ vectors. Denoting \mathbf{A}_i as the i th column of \mathbf{A} , we write the dual objective as,

$$g(\boldsymbol{\lambda}) = \sum_{i=1}^{n^*} \left\{ -\mathbb{1}(Z_i = z)D_i \rho\{\mathbf{A}_i^\top (\boldsymbol{\lambda}_+ - \boldsymbol{\lambda}_-)\} + (n^*)^{-1} \mathbf{A}_i^\top (\boldsymbol{\lambda}_+ - \boldsymbol{\lambda}_-) \right\} + (\boldsymbol{\lambda}_+^\top + \boldsymbol{\lambda}_-^\top) \boldsymbol{\delta}. \tag{21}$$

Denote $\boldsymbol{\lambda}^\dagger = (\boldsymbol{\lambda}_+^{\dagger\top}, \boldsymbol{\lambda}_-^{\dagger\top})^\top$ as the dual solution. Let, if possible, the j th component of $\boldsymbol{\lambda}_+^\dagger$ and $\boldsymbol{\lambda}_-^\dagger$ be both strictly positive, for some $j \in \{1, 2, \dots, K\}$. Define,

$$\boldsymbol{\lambda}^{\dagger\dagger} = (\boldsymbol{\lambda}_{+\top}^\dagger - (0, 0, \dots, \underbrace{\min(\lambda_{+,j}^\dagger, \lambda_{-,j}^\dagger)}_{j\text{th place}}, 0, \dots, 0)^\top, \boldsymbol{\lambda}_{-\top}^\dagger - (0, 0, \dots, \underbrace{\min(\lambda_{+,j}^\dagger, \lambda_{-,j}^\dagger)}_{j\text{th place}}, 0, \dots, 0)^\top) \tag{22}$$

Notice that $g(\boldsymbol{\lambda}^{\dagger\dagger}) = g(\boldsymbol{\lambda}^{\dagger}) - 2\delta_j \min(\lambda_{+,j}^{\dagger}, \lambda_{-,j}^{\dagger}) < g(\boldsymbol{\lambda}^{\dagger})$ since $\delta_j > 0$. This leads to a contradiction since $\boldsymbol{\lambda}^{\dagger}$ minimizes $g(\boldsymbol{\lambda})$. This implies that at least one of $\lambda_{+,j}^{\dagger}, \lambda_{-,j}^{\dagger}$ equals zero. From Equation 21 we see that the dual problem has the following unconstrained form.

$$\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad (n^*)^{-1} \sum_{i=1}^{n^*} \left[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\mathbf{B}(\mathbf{X}_i)^{\top} \boldsymbol{\lambda}\} + \{\mathbf{B}(\mathbf{X}_i)^{\top} \boldsymbol{\lambda}\} \right] + |\boldsymbol{\lambda}|^{\top} \boldsymbol{\delta}, \quad (23)$$

where $|\boldsymbol{\lambda}|$ is the vector of co-ordinate wise absolute values of $\boldsymbol{\lambda}$.

Now, consider the primal problem

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i:Z_i=z} \psi(w_i), \\ & \text{subject to} \quad \left| \sum_{i:Z_i=z} w_i B_k(\mathbf{X}_i) - \bar{B}_k^* \right| \leq \delta_k, \quad k = 1, 2, \dots, K_0. \\ & \quad \quad \quad \left| \sum_{i:Z_i=z} w_i B_k(\mathbf{X}_i) - \bar{B}_k^* \right| = 0, \quad k = K_0 + 1, \dots, K. \end{aligned} \quad (24)$$

Let $\tilde{\mathbf{B}}(\mathbf{x}) = (B_1(\mathbf{x}), \dots, B_{K_0}(\mathbf{x}))^{\top}$, $\tilde{\tilde{\mathbf{B}}}(\mathbf{x}) = (B_{K_0+1}(\mathbf{x}), \dots, B_K(\mathbf{x}))^{\top}$, and $\tilde{\boldsymbol{\delta}} = (\delta_1, \dots, \delta_{K_0})^{\top}$. We note that for $B_k(\mathbf{x}) = 1$, the corresponding equality constraint boils down to the normalization constraint. Using similar steps as before, we see that the dual of 24 is

$$\underset{\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}}{\text{minimize}} \quad (n^*)^{-1} \sum_{i=1}^{n^*} \left[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\tilde{\mathbf{B}}(\mathbf{X}_i)^{\top} \tilde{\boldsymbol{\lambda}} + \tilde{\tilde{\mathbf{B}}}(\mathbf{X}_i)^{\top} \tilde{\boldsymbol{\nu}}\} + \{\tilde{\mathbf{B}}(\mathbf{X}_i)^{\top} \tilde{\boldsymbol{\lambda}} + \tilde{\tilde{\mathbf{B}}}(\mathbf{X}_i)^{\top} \tilde{\boldsymbol{\nu}}\} \right] + |\tilde{\boldsymbol{\lambda}}|^{\top} \tilde{\boldsymbol{\delta}}. \quad (25)$$

Let $\boldsymbol{\lambda} = (\tilde{\boldsymbol{\lambda}}^{\top}, \tilde{\boldsymbol{\nu}}^{\top})^{\top}$. Since $\mathbf{B}(\mathbf{x}) = (\tilde{\mathbf{B}}(\mathbf{x})^{\top}, \tilde{\tilde{\mathbf{B}}}(\mathbf{x})^{\top})^{\top}$ and $\boldsymbol{\delta} = (\tilde{\boldsymbol{\delta}}^{\top}, \mathbf{0}^{\top})^{\top}$, the dual problem in 25 can be written as:

$$\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad (n^*)^{-1} \sum_{i=1}^{n^*} \left[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\mathbf{B}(\mathbf{X}_i)^{\top} \boldsymbol{\lambda}\} + \{\mathbf{B}(\mathbf{X}_i)^{\top} \boldsymbol{\lambda}\} \right] + |\boldsymbol{\lambda}|^{\top} \boldsymbol{\delta}, \quad (26)$$

which has the same form as in 23. This proves part (a) of Theorem 4.1. Moreover, Equation 20 implies that the optimal solutions of the primal problem satisfies

$$\hat{w}_i = \rho'\{\mathbf{B}(\mathbf{X}_i)^{\top} \boldsymbol{\lambda}^{\dagger}\}, \quad (27)$$

proving part (b). Finally, for part (c), we consider the conditional expected loss in (23).

$$\begin{aligned}
& E_{\mathbb{T}} \left[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} + \{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} \middle| \mathbf{X}_i \right] \\
&= E_{\mathbb{T}} \left[-n^* \mathbb{1}(Z_i = z) D_i \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} \middle| \mathbf{X}_i \right] + \{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} \\
&= -n^* \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} P_{\mathbb{T}}(Z_i = z | \mathbf{X}_i) \pi(\mathbf{X}_i) + \{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} =: \phi(\boldsymbol{\lambda})
\end{aligned} \tag{28}$$

We now minimize this expected loss $\phi(\boldsymbol{\lambda})$ wrt $\boldsymbol{\lambda}$. The minimizer $\boldsymbol{\lambda}^*$ satisfies the following:

$$\rho'\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}^*\} = \{n^* \pi(\mathbf{X}_i) P_{\mathbb{T}}(Z_i = z | \mathbf{X}_i)\}^{-1} \tag{29}$$

This implies the solutions \hat{w}_i implicitly estimate the inverse propensity weights. This completes the proof.

8.3 Connection to inverse odds weighting

In this section, we consider the generalization problem in a nested design setting, where the target population is the population of study non-participants. As before, we denote \mathcal{P} and \mathcal{T} as the study and target population, respectively. Let \mathcal{Q} be the population represented by the study participants and non-participants, with associated probability measure \mathbb{Q} (and density q). Also, let n^{**} be the total number of study participants and non-participants. As before, D is the indicator of being a study participant. It follows that $\mathbb{P} = \mathbb{Q}|D = 1$ and $\mathbb{T} = \mathbb{Q}|D = 0$. In particular, we can write

$$\tau = E_{\mathbb{T}}\{Y(1) - Y(0)\} = E_{\mathbb{Q}}\{Y(1) - Y(0) | D = 0\} \tag{30}$$

We now derive a connection between the one-step balancing weights and the inverse odds weights in this setting. The k th balancing constraint in the optimization problem in (9) can be written

as,

$$\begin{aligned}
& \left| \sum_{i:Z_i=z} w_i B_k(\mathbf{X}_i) - \bar{B}_k^* \right| \leq \delta_k \\
\Rightarrow & \left| \sum_{i=1}^{n^{**}} \mathbb{1}(Z_i = z) D_i w_i B_k(\mathbf{X}_i) - (n^*)^{-1} \sum_{i=1}^{n^{**}} (1 - D_i) B_k(\mathbf{X}_i) \right| \leq \delta_k \\
\Rightarrow & \left| \sum_{i=1}^{n^{**}} \{ (n^*)^{-1} (1 - D_i) - \mathbb{1}(Z_i = z) D_i w_i \} B_k(\mathbf{X}_i) \right| \leq \delta_k \\
\Rightarrow & \left| \sum_{i=1}^{n^*} \xi_i B_k(\mathbf{X}_i) \right| \leq \delta_k
\end{aligned} \tag{31}$$

where $\xi_i = (n^*)^{-1} (1 - D_i) - \mathbb{1}(Z_i = z) D_i w_i$. So, for the units in the group $Z_i = z$ of the sample, $w_i = \{(1 - D_i)/n^* - \xi_i\} = -\xi_i$. For the objective function $\sum_{i:Z_i=z} \psi(w_i)$ (where ψ is a convex function of the weights), we have

$$\sum_{i:Z_i=z} \psi(w_i) = \sum_{i=1}^{n^{**}} \mathbb{1}(Z_i = z) D_i h(\xi_i), \tag{32}$$

where $h(x) = \psi(-x)$. Now, let $\underline{\mathbf{A}}$ be a $K \times n$ matrix whose (i, j) th element is $B_i(\mathbf{X}_j)$; $\underline{\mathbf{Q}} = (\mathbf{A}^\top, -\mathbf{A}^\top)^\top$; and $\mathbf{d} = (\boldsymbol{\delta}, \boldsymbol{\delta})^\top$. We can write the primal problem as

$$\begin{aligned}
& \underset{\mathbf{w}}{\text{minimize}} && \sum_{i=1}^{n^{**}} \mathbb{1}(Z_i = z) D_i h(\xi_i) \\
& \text{subject to} && \underline{\mathbf{Q}} \boldsymbol{\xi} \leq \mathbf{d}
\end{aligned} \tag{33}$$

gives us a convex optimization problem in $\boldsymbol{\xi}$ with linear constraints. Let \mathbf{Q}_i be the i th column of $\underline{\mathbf{Q}}$. The corresponding dual problem is given by,

$$\begin{aligned}
& \underset{\boldsymbol{\lambda}}{\text{maximize}} && \sum_{i=1}^{n^{**}} \{-h_i^*(\mathbf{Q}_i^\top \boldsymbol{\lambda})\} - \boldsymbol{\lambda}^\top \mathbf{d} \\
& \text{subject to} && \boldsymbol{\lambda} \geq \mathbf{0}
\end{aligned} \tag{34}$$

where,

$$\begin{aligned}
h_i^*(t) &= \sup_{\xi_i} \{t\xi_i - \mathbb{1}(Z_i = z)D_i h(\xi_i)\} \\
&= \sup_{w_i} [\{(1 - D_i)/n^* - \mathbb{1}(Z_i = z)D_i w_i\}t - \mathbb{1}(Z_i = z)D_i h((1 - D_i)/n^* - \mathbb{1}(Z_i = z)D_i w_i)] \\
&= \sup_{w_i} [\{(1 - D_i)/n^* - \mathbb{1}(Z_i = z)D_i w_i\}t - \mathbb{1}(Z_i = z)D_i h(-w_i)] \\
&= \{(1 - D_i)/n^* - \mathbb{1}(Z_i = z)D_i \hat{w}_i(t)\}t - \mathbb{1}(Z_i = z)D_i h(-\hat{w}_i(t))
\end{aligned} \tag{35}$$

where $\hat{w}_i(t)$ satisfies

$$\partial/\partial w_i \left[\{(1 - D_i)/n^* - \mathbb{1}(Z_i = z)D_i w_i\}t - \mathbb{1}(Z_i = z)D_i h(-w_i) \right] \Big|_{w_i = \hat{w}_i(t)} = 0.$$

Solving for w_i , we get $\hat{w}_i(t) = -(h')^{-1}(t)$.

Therefore, the dual problem boils down to:

$$\begin{aligned}
&\underset{\boldsymbol{\lambda}}{\text{minimize}} && \sum_{i=1}^{n^{**}} \left\{ -\mathbb{1}(Z_i = z)D_i \rho(\mathbf{Q}_i^\top \boldsymbol{\lambda}) + (n^*)^{-1}(1 - D_i)\mathbf{Q}_i^\top \boldsymbol{\lambda} \right\} + \boldsymbol{\lambda}^\top \mathbf{d} \\
&\text{subject to} && \boldsymbol{\lambda} \geq \mathbf{0}
\end{aligned} \tag{36}$$

where $\rho(t) = -t(h')^{-1}(t) + h((h')^{-1}(t))$. Note that $\rho'(t) = -(h')^{-1}(t)$. Therefore,

$$\hat{w}_i(t) = \rho'(t). \tag{37}$$

Using similar steps as in the proof of Theorem 4.1 we obtain the following form of the dual problem.

$$\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad (n^{**})^{-1} \sum_{i=1}^{n^{**}} \left[-n^{**} \mathbb{1}(Z_i = z)D_i \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} + (n^*)^{-1}n^{**}(1 - D_i)\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} \right] + |\boldsymbol{\lambda}|^\top \boldsymbol{\delta} \tag{38}$$

Also, if \hat{w}_i s are the optimal solutions of the primal problem and $\boldsymbol{\lambda}^\dagger$ is the optimal solution of the

dual problem, then we have

$$\hat{w}_i = \rho' \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}^\dagger \}. \quad (39)$$

Let us consider the conditional expected loss in (38).

$$\begin{aligned} & E_{\mathbb{Q}} \left[-n^{**} \mathbb{1}(Z_i = z) D_i \rho \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \} + (n^*)^{-1} n^{**} (1 - D_i) \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \} \middle| \mathbf{X}_i \right] \\ &= E_{\mathbb{Q}} \left[-n^{**} \mathbb{1}(Z_i = z) D_i \rho \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \} \middle| \mathbf{X}_i \right] + (n^*)^{-1} n^{**} (1 - \pi(\mathbf{X}_i)) \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \} \\ &= -n^{**} \rho \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \} P_{\mathbb{Q}}(Z_i = z | \mathbf{X}_i) \pi(\mathbf{X}_i) + (n^*)^{-1} n^{**} \{ 1 - \pi(\mathbf{X}_i) \} \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda} \} \\ &=: \phi(\boldsymbol{\lambda}) \end{aligned} \quad (40)$$

We now minimize this expected loss $\phi(\boldsymbol{\lambda})$ wrt $\boldsymbol{\lambda}$. The minimizer $\boldsymbol{\lambda}^{**}$ satisfies the following–

$$\rho' \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}^{**} \} = \{ 1 - \pi(\mathbf{X}_i) \} / \{ n^* \pi(\mathbf{X}_i) P_{\mathbb{Q}}(Z_i = z | \mathbf{X}_i) \}. \quad (41)$$

This implies, the solutions \hat{w}_i implicitly estimate the inverse odds weights.

8.4 Connection to linear regression

In this section, we discuss the connection between one-step weights and linear regression. For simplicity, we consider the generalization problem in a nested design setting, where the study sample is nested within a random sample of size n^* from the target population. Extensions of the results to other settings hold analogously. Also, without loss of generality, we focus on estimating $E_{\mathbb{T}}\{Y(1)\}$ (the results hold similarly for $E_{\mathbb{T}}\{Y(0)\}$).

Recall that, by Assumption 1, $m_1(\mathbf{x}) = E_{\mathbb{T}}\{Y(1) | \mathbf{X} = \mathbf{x}\} = E_{\mathbb{P}}\{Y(1) | \mathbf{X} = \mathbf{x}\}$. The linear regression imputation approach for generalization (our outcome modeling approach; see Dahabreh et al. (2019)) first fits a linear outcome model $Y_i^{\text{obs}} = \beta_0 + \beta_1^\top \mathbf{X}_i + \epsilon_i$ in the treatment group and estimates $m_1(\mathbf{x})$ as $\hat{m}_1(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1^\top \mathbf{x}$, where the coefficients are estimated by ordinary least squares (OLS). The regression model can also incorporate other transformations of \mathbf{X} . The

regression imputation estimator of $E_{\mathbb{T}}\{Y(1)\}$ is given by,

$$\hat{E}_{\mathbb{T}}\{Y(1)\} = \frac{1}{n^*} \sum_{i=1}^{n^*} \{\hat{\beta}_0 + \hat{\beta}_1^\top \mathbf{X}_i\} = \hat{\beta}_0 + \hat{\beta}_1^\top \bar{\mathbf{X}}^*, \quad (42)$$

where $\bar{\mathbf{X}}^*$ is the target profile, i.e., the mean of the covariates in the full sample.

We note that procedurally, this approach is equivalent to the multi-regression imputation (MRI) approach in [Chattopadhyay and Zubizarreta \(2023\)](#). Hence, by Proposition 2 of [Chattopadhyay and Zubizarreta \(2023\)](#), it follows that, $\hat{\beta}_0 + \hat{\beta}_1^\top \bar{\mathbf{X}}^* = \sum_{i:Z_i=1} w_i Y_i^{\text{obs}}$, where the weights w_i sum to one in the treatment group.

Moreover, by Proposition 3 of the same paper, the weights w_i are the weights of minimum variance that add up to one and *exactly* balances the mean of the covariates in the treatment group, relative to the target profile $\bar{\mathbf{X}}^*$. In other words, these weights are equivalent to the one-step weights in [\(7\)](#), where — (i) the L_2 norm of the weights are minimized, (ii) The basis functions are identity functions, (iii) $\delta_k = 0$ for all k , and (iv) weights are allowed to be negative.

8.5 Convergence of one-step weights

The equivalence in Theorem [8.1](#) allows us to establish several asymptotic properties of the one-step balancing weights and their resulting estimators. In this section, we formally show that under Assumption [2](#) and [4](#), the one-step weights converge uniformly to the inverse propensity weights.

For convenience, we restate the regularity conditions in Assumption 4 below.

Assumption 4. For $z \in \{0, 1\}$,

- (a) There exist constants c_0, c_1, c_2 with $0 < c_0 < 1/2$ and $c_1 < c_2 < 0$, such that $c_1 \leq n^* \rho''(v) \leq c_2$ for all v in a neighborhood of $\mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}_{1z}^*$. Also, $c_0 \leq 1/(n^* \rho'(v)) \leq 1 - c_0$ for all $v = \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}, \mathbf{x} \in \mathcal{X}, \boldsymbol{\lambda}$.
- (b) $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{B}(\mathbf{x})\|_2 \leq CK^{1/2}$ and $\|E_{\mathbb{T}}\{\mathbf{B}(\mathbf{X})\mathbf{B}(\mathbf{X})^\top\}\|_F \leq C$, for some constant $C > 0$, where $\|\cdot\|_F$ denotes the Frobenius norm.

- (c) $K = O\{(n^*)^\alpha\}$ for some $0 < \alpha < 2/3$.
- (d) For some constant $C > 0$, $\lambda_{\min}\left[E_{\mathbb{T}}\{D\mathbb{1}(Z = z)\mathbf{B}(\mathbf{X})\mathbf{B}(\mathbf{X})^\top\}\right] > C$, where $\lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalue of \mathbf{A} .
- (e) $\|\boldsymbol{\delta}\|_2 = O_P\left[K^{1/4}\{(\log K)/n^*\}^{1/2} + K^{-r_z+1/2}\right]$.

Assumption 4(a) bounds the slope and curvature of the function $\rho(\cdot)$, allowing us to translate the convergence of the dual solution to the convergence of the weights. This condition is satisfied for typical choices of convex objective functions $\psi(\cdot)$, e.g., those corresponding to entropy balancing (Hainmueller 2012) and stable balancing weights (Zubizarreta 2015). Assumption 4(b) restricts the rate of growth of the norm of the basis functions, and 4(c) specifies the rate of growth of the number of basis functions K relative to n^* . Assumption 4(d) is a technical condition required to ensure non-singularity of the covariance matrices of the basis functions within each treatment group. Finally, 4(e) controls the degree of approximate balancing in terms of K and n^* .

Theorem 8.2 (Uniform convergence). Under Assumptions 1, 2, and 4, the one-step balancing weights in group $Z = z \in \{0, 1\}$ satisfy

$$\sup_{\mathbf{x} \in \mathcal{X}} |n^* \hat{w}(\mathbf{x}) - w^{\text{IP}}(\mathbf{x})| = O_P\left[K^{3/4}\{(\log K)/n^*\}^{1/2} + K^{1-r_z}\right] = o_P(1). \quad (43)$$

Below we provide a proof of Theorem 8.2. All probabilities and expectations in this proof are computed with respect to the probability measure \mathbb{T} . We focus on the primal problem in group z , for $z \in \{0, 1\}$. For simplicity, let us denote $\tilde{Z} = \mathbb{1}(Z_i = z)D_i$. The dual optimization problem can be written as,

$$\begin{aligned} & \underset{\boldsymbol{\lambda}}{\text{minimize}} \quad G(\boldsymbol{\lambda}), \text{ where} \\ & G(\boldsymbol{\lambda}) = (n^*)^{-1} \sum_{i=1}^{n^*} \left[-n^* \tilde{Z}_i \rho\{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} + \{\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}\} \right] + |\boldsymbol{\lambda}|^\top \boldsymbol{\delta}. \end{aligned} \quad (44)$$

Let $\boldsymbol{\lambda}^\dagger$ be a solution to the dual problem. We first consider the L_∞ distance between the scaled one-step balancing weights and inverse probability weights. In the following, we use C, C', C'' as

generic positive constants whose values may change from one step to the next.

$$\begin{aligned}
& \sup_{\mathbf{x} \in \mathcal{X}} |n^* \hat{w}(\mathbf{x}) - w^{\text{IP}}(\mathbf{x})| \\
&= \sup_{\mathbf{x} \in \mathcal{X}} |n^* \rho' \{ \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}^\dagger \} - n^* \rho' \{ g_z^*(\mathbf{x}) \}| \\
&\leq \sup_{\mathbf{x} \in \mathcal{X}} |n^* \rho' \{ \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}^\dagger \} - n^* \rho' \{ \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}_{1z}^* \}| + \sup_{\mathbf{x} \in \mathcal{X}} |n^* \rho' \{ \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}_{1z}^* \} - n^* \rho' \{ g_z^*(\mathbf{x}) \}| \\
&\leq C \sup_{\mathbf{x} \in \mathcal{X}} | \mathbf{B}(\mathbf{x})^\top (\boldsymbol{\lambda}^\dagger - \boldsymbol{\lambda}_{1z}^*) | + C \sup_{\mathbf{x} \in \mathcal{X}} | \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}_{1z}^* - g_z^*(\mathbf{x}) | \\
&\leq CK^{1/2} \| \boldsymbol{\lambda}^\dagger - \boldsymbol{\lambda}_{1z}^* \|_2 + O(K^{-rz}).
\end{aligned} \tag{45}$$

The first equality is due to Assumption 2. The first inequality is due to the triangle inequality. The second inequality follows from applying using the mean value theorem and Assumption 4(a). The final inequality is due to the Cauchy-Schwarz inequality and Assumptions 2 and 4(b).

Let us now consider the following Lemma.

Lemma 8.3. There exists a dual solution $\boldsymbol{\lambda}^\dagger$ such that $\| \boldsymbol{\lambda}^\dagger - \boldsymbol{\lambda}_{1z}^* \|_2 = O_p[K^{1/4} \{ (\log K)/n^* \}^{1/2} + K^{-rz+1/2}]$.

Given Lemma 8.3, Equation 45 completes the proof of Theorem 4.2. So, it suffices to prove Lemma 8.3. For this, we require the following lemmas.

Lemma 8.4 (Bernstein's inequality for random matrices). Let $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_{n^*}$ be $d_1 \times d_2$ independent random matrices with $E(\mathbf{W}_j) = \mathbf{0}$ and $\| \mathbf{W}_j \|_2 \leq R_{n^*}$ a.s., where $\| \cdot \|_2$ denotes the spectral norm. Let $\sigma_{n^*}^2 := \max \left\{ \| \sum_{j=1}^{n^*} E(\mathbf{W}_j \mathbf{W}_j^\top) \|_2, \| \sum_{j=1}^{n^*} E(\mathbf{W}_j^\top \mathbf{W}_j) \|_2 \right\}$. Then, for all $t \geq 0$,

$$\mathbb{P} \left(\left\| \sum_{j=1}^{n^*} \mathbf{W}_j \right\|_2 \geq t \right) \leq (d_1 + d_2) \exp \left[(t^2/2) / \{ \sigma_{n^*}^2 + (R_{n^*} t)/3 \} \right]. \tag{46}$$

Proof of Lemma 8.4. See Tropp et al. (2015).

Lemma 8.5. $\left\| (n^*)^{-1} \sum_{j=1}^{n^*} \{ 1 - \tilde{Z}_j w^{\text{IP}}(\mathbf{X}_j) \} \mathbf{B}(\mathbf{X}_j) \right\|_2 = O_P[K^{1/4} \{ (\log K)/n^* \}^{1/2}]$.

Proof of Lemma 8.5. We will use Lemma 8.4 to prove this. Let us denote

$$\underline{\mathbf{W}}_j = (n^*)^{-1} \{1 - \tilde{Z}_j w^{\text{IP}}(\mathbf{X}_j)\} \mathbf{B}(\mathbf{X}_j), \quad \text{for } j \in \{1, 2, \dots, n^*\}. \quad (47)$$

First, by unconfoundedness, $E(\underline{\mathbf{W}}_j) = \mathbf{0}$. Second, we have

$$\begin{aligned} \|\underline{\mathbf{W}}_j\|_2 &= (n^*)^{-1} |1 - \tilde{Z}_j n^* \rho' \{g_z^*(\mathbf{X}_j)\}| \times \|\mathbf{B}(\mathbf{X}_j)\|_2 \\ &\leq [1 + |n^* \rho' \{g_z^*(\mathbf{X}_j)\}|] (n^*)^{-1} \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{B}(\mathbf{x})\|_2 \\ &\leq CK^{1/2} (n^*)^{-1} [C' + O(K^{-r_z})C] \leq (C' K^{1/2})/n^*. \end{aligned} \quad (48)$$

Here the second inequality is obtained by applying Assumption 4(b) on the second term in the product and using the mean value theorem on $n^* \rho' \{g_z^*(\mathbf{X}_j)\}$ about $\mathbf{B}(\mathbf{X}_j)^\top \boldsymbol{\lambda}_{1z}^*$, followed by assumptions 4(a) and 2. Next, we consider

$$\begin{aligned} \left\| \sum_{j=1}^{n^*} E(\underline{\mathbf{W}}_j^\top \underline{\mathbf{W}}_j) \right\|_2 &= \sum_{j=1}^{n^*} E \left[(n^*)^{-2} \{1 - \tilde{Z}_j w^{\text{IP}}(\mathbf{X}_j)\}^2 \mathbf{B}(\mathbf{X}_j)^\top \mathbf{B}(\mathbf{X}_j) \right] \\ &\leq C(n^*)^{-2} \sum_{j=1}^{n^*} E \left\{ \mathbf{B}(\mathbf{X}_j)^\top \mathbf{B}(\mathbf{X}_j) \right\} \\ &= C(n^*)^{-2} \text{trace} \left[E \left\{ \mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top \right\} \right]. \end{aligned} \quad (49)$$

Here first inequality follows from applying the mean value theorem along with assumptions 4(a) and 2, similar to the steps in Equation 48. Now, let $\lambda_1, \dots, \lambda_K$ be the eigenvalues of a non-negative definite matrix \mathbf{A} . By the Cauchy-Schwarz inequality, $\text{trace}(\mathbf{A}) \leq K^{1/2} (\lambda_1^2 + \dots + \lambda_K^2)^{1/2} = K^{1/2} \|\mathbf{A}\|_F$. Thus, from Equation 49, we get

$$\left\| \sum_{j=1}^{n^*} E(\underline{\mathbf{W}}_j^\top \underline{\mathbf{W}}_j) \right\|_2 \leq (CK^{1/2})/n^* \left\| E \left[\mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top \right] \right\|_F \leq (C' K^{1/2})/n^*, \quad (50)$$

where the last inequality holds due to Assumption 4(b). Next, we consider

$$\begin{aligned}
\left\| \sum_{j=1}^{n^*} E(\mathbf{W}_j \mathbf{W}_j^\top) \right\|_2 &\leq \sum_{j=1}^{n^*} \left\| E \left[(n^*)^{-2} \{1 - \tilde{Z}_j w^{\text{IP}}(\mathbf{X}_j)\}^2 \mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top \right] \right\|_2 \\
&\leq C(n^*)^{-2} \sum_{j=1}^{n^*} \|E\{\mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}\|_2 \\
&\leq C(n^*)^{-2} \sum_{j=1}^{n^*} \|E\{\mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}\|_F \leq C''/n^*
\end{aligned} \tag{51}$$

Here the first inequality is due to the triangle inequality. The second inequality follows from upper bounding $\{1 - \tilde{Z}_j w^{\text{IP}}(\mathbf{X}_j)\}^2$ as before and using monotonicity of spectral norms. The third inequality holds since spectral norm is dominated by the Frobenius norm. Finally, the fourth inequality holds due to Assumption 4(b)f. Therefore, from equations 50 and 51 we get,

$$\sigma_{n^*}^2 := \max \left\{ \left\| \sum_{j=1}^{n^*} E\{\mathbf{W}_j \mathbf{W}_j^\top\} \right\|_2, \left\| \sum_{j=1}^{n^*} E\{\mathbf{W}_j^\top \mathbf{W}_j\} \right\|_2 \right\} \leq (CK^{1/2})/n^*. \tag{52}$$

Using Lemma 8.4, we get,

$$\mathbb{P} \left(\left\| \sum_{j=1}^{n^*} \mathbf{W}_j \right\|_2 \geq t \right) \leq (K+1) \exp \left[(t^2/2) / \left\{ CK^{1/2}(n^*)^{-1} + C'K^{1/2}t(3n^*)^{-1} \right\} \right] \tag{53}$$

Finally, we observe that due to Assumption 4(c), the right hand side of Equation 53 goes to zero if $t = \bar{C}K^{1/4}(\log K)^{1/2}(n^*)^{-1/2}$, for some constant $\bar{C} > 0$. This implies, $\|\sum_{j=1}^{n^*} \mathbf{W}_j\|_2 = O_P \left\{ K^{1/4}(\log K)^{1/2}(n^*)^{-1/2} \right\}$. This completes the proof.

Lemma 8.6. With probability tending to one, $\lambda_{\min} \left(\sum_{j: \tilde{Z}_j=1} (n^*)^{-1} \mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top \right) \geq \tilde{C}$ for some constant $\tilde{C} > 0$.

Proof of Lemma 8.6. Let $\mathbf{D} := \sum_{j: \tilde{Z}_j=1} (n^*)^{-1} \mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top = (n^*)^{-1} \sum_{j=1}^{n^*} \{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)\} \{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)^\top\}$ and $\mathbf{D}^* := E\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}$. We will first use Lemma 8.4 to show that $\|\mathbf{D} - \mathbf{D}^*\|_2 = o_P(1)$. To this end, denote $\mathbf{W}_j = (n^*)^{-1} [\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)\} \{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)^\top\} - E\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}]$. By construc-

tion, $E(\underline{\mathbf{W}}_j) = \mathbf{0}$. Moreover,

$$\begin{aligned} \|\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)\}\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)^\top\}\|_2 &\leq \|\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)\}\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j)^\top\}\|_F \\ &\leq C' \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{B}(\mathbf{x})\|_2^2 \leq C'' K, \end{aligned} \quad (54)$$

where the last inequality holds due to Assumption 4(b). Also,

$$\begin{aligned} \|E\{\tilde{Z}_j \mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}\|_2 &\leq \|E\{\mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}\|_2 \\ &\leq \|E\{\mathbf{B}(\mathbf{X}_j) \mathbf{B}(\mathbf{X}_j)^\top\}\|_F \leq C', \end{aligned} \quad (55)$$

where the first inequality is due to the monotonicity of the spectral norm and the last inequality holds due to Assumption 4(b). This implies,

$$\|\underline{\mathbf{W}}_j\|_2 \leq \{C(K+1)\}/n^*, \quad (56)$$

for some constant $C > 0$. Next, we compute $\sigma_{n^*}^2$. After some algebra, it follows that,

$$\begin{aligned} &\left\| \sum_{j=1}^{n^*} E(\underline{\mathbf{W}}_j \underline{\mathbf{W}}_j^\top) \right\|_2 \\ &\leq \sum_{j=1}^{n^*} \|E(\underline{\mathbf{W}}_j \underline{\mathbf{W}}_j^\top)\|_2 \\ &\leq (n^*)^{-1} \left(CK \|E\{\mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top\}\|_2 + \|E\{\tilde{Z}_1 \mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top\} E\{\tilde{Z}_1 \mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top\}\|_2 \right) \\ &\leq (n^*)^{-1} \left[CK \|E\{\mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top\}\|_2 + \|E\{\mathbf{B}(\mathbf{X}_1) \mathbf{B}(\mathbf{X}_1)^\top\}\|_2^2 \right] \\ &\leq \{C'(K+1)\}/n^*, \end{aligned} \quad (57)$$

for some large $C' > 0$. Here the first inequality is due to the triangle inequality; the second inequality is due to Assumption 4(b) and monotonicity of the spectral norm; the third inequality is due to the submultiplicativity of the spectral norm; and the final inequality is due to Assumption

4(b). Now, since \mathbf{W}_j is symmetric we have

$$\left\| \sum_{j=1}^{n^*} E(\mathbf{W}_j^\top \mathbf{W}_j) \right\|_2 \leq \{C'(K+1)\}/n^*, \quad (58)$$

implying $\sigma_{n^*}^2 \leq \{C'(K+1)\}/n^*$. Therefore, by Lemma 8.4, we get

$$\begin{aligned} \mathbb{P} \left(\left\| \sum_{j=1}^{n^*} \mathbf{W}_j \right\|_2 \geq t \right) &\leq 2K \exp \left[(t^2/2) / \left\{ C'(K+1)(n^*)^{-1} + C(K+1)t(2n^*)^{-1} \right\} \right] \\ &\leq 2K \exp \left[n^* t^2 / \{C''K(1+t)\} \right], \end{aligned} \quad (59)$$

for a large constant $C'' > 0$. We note that the right hand side goes to zero for $t = \bar{C}\{(K \log K)/n^*\}^{1/2}$ for a constant $\bar{C} > 0$. Therefore, we have

$$\|\mathbf{D} - \mathbf{D}^*\|_2 = O_p \left[\{(K \log K)/n^*\}^{1/2} \right] = o_p(1), \quad (60)$$

where the last equality holds due to Assumption 4(c). Now, Weyl's inequality implies

$$\lambda_{\min}(\mathbf{D}) \geq \lambda_{\min}(\mathbf{D}^*) - \|\mathbf{D} - \mathbf{D}^*\|_2 \geq C - \|\mathbf{D} - \mathbf{D}^*\|_2, \quad (61)$$

where the last inequality holds due to Assumption 4(d). Since $\|\mathbf{D} - \mathbf{D}^*\|_2 = o_p(1)$, we have for n^* large enough, $\lambda_{\min}(\mathbf{D}) \geq C/2 > 0$. This completes the proof.

Proof of Lemma 8.3. We follow the proof structure of Fan et al. (2016) and Wang and Zubizarreta (2020). All the subsequent probabilities and expectations are taken with respect to \mathbb{T} . First, let $r = C^* \left\{ (K^{1/4} \log K)/n^* + K^{-rz+1/2} \right\}$ for a sufficiently large constant $C^* > 0$. Let $\mathbf{\Delta} = \mathbf{\lambda} - \mathbf{\lambda}_{1z}^*$. Also, let $\mathcal{C} = \{\mathbf{\Delta} \in \mathbb{R}^K : \|\mathbf{\Delta}\|_2 \leq r\}$. To show that there exists a $\mathbf{\lambda}^\dagger$ such that $\|\mathbf{\lambda}^\dagger - \mathbf{\lambda}_{1z}^*\|_2 = O_p \left\{ (K^{1/4} \log K)/n^* + K^{-rz+1/2} \right\}$, it is enough to show that there exists a $\mathbf{\Delta}^\dagger \in \mathbb{R}^K$ such that $\mathbb{P}(\mathbf{\Delta}^\dagger \in \mathcal{C}) \xrightarrow{n^* \rightarrow \infty} 1$.

Now, the dual objective can be written as,

$$G(\mathbf{\lambda}_{1z}^* + \mathbf{\Delta}) = (n^*)^{-1} \sum_{i=1}^{n^*} \left[-n^* \tilde{Z}_i \rho \{ \mathbf{B}(\mathbf{X}_i)^\top (\mathbf{\lambda}_{1z}^* + \mathbf{\Delta}) \} + \{ \mathbf{B}(\mathbf{X}_i)^\top (\mathbf{\lambda}_{1z}^* + \mathbf{\Delta}) \} \right] + |\mathbf{\lambda}_{1z}^* + \mathbf{\Delta}|^\top \boldsymbol{\delta}. \quad (62)$$

Since $f(\cdot)$ is convex, $\rho(\cdot)$ is concave. It follows that $G(\boldsymbol{\lambda}_{1z}^* + \boldsymbol{\Delta})$ is convex in $\boldsymbol{\Delta}$. Moreover, $G(\boldsymbol{\lambda}_{1z}^* + \boldsymbol{\Delta})$ is also continuous in $\boldsymbol{\Delta}$. Therefore, to show $P(\boldsymbol{\Delta}^\dagger \in \mathcal{C}) \xrightarrow{n^* \rightarrow \infty} 1$, it is enough show that

$$P\left(\inf_{\boldsymbol{\Delta} \in \partial\mathcal{C}} G(\boldsymbol{\lambda}_{1z}^* + \boldsymbol{\Delta}) - G(\boldsymbol{\lambda}_{1z}^*) > 0\right) \xrightarrow{n^* \rightarrow \infty} 1, \quad (63)$$

where $\partial\mathcal{C}$ is the boundary set of \mathcal{C} given by $\partial\mathcal{C} = \{\boldsymbol{\Delta} \in \mathbb{R}^K : \|\boldsymbol{\Delta}\|_2 = r\}$.

Now, fix $\boldsymbol{\Delta} \in \partial\mathcal{C}$. Using multivariate Taylor's theorem, we seen that for some intermediate $\tilde{\boldsymbol{\lambda}}$,

$$\begin{aligned} & G(\boldsymbol{\lambda}_{1z}^* + \boldsymbol{\Delta}) - G(\boldsymbol{\lambda}_{1z}^*) \\ &= \boldsymbol{\Delta}^\top \left[(n^*)^{-1} \sum_{i=1}^{n^*} \{-n^* \tilde{Z}_i \rho'(\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}_{1z}^*) \mathbf{B}(\mathbf{X}_i) + \mathbf{B}(\mathbf{X}_i)\} \right] \\ &+ \boldsymbol{\Delta}^\top \left[(n^*)^{-1} \sum_{i=1}^{n^*} \{-n^* \tilde{Z}_i \rho''(\mathbf{B}(\mathbf{X}_i)^\top \tilde{\boldsymbol{\lambda}}) \mathbf{B}(\mathbf{X}_i) \mathbf{B}(\mathbf{X}_i)^\top\} \right] \boldsymbol{\Delta} / 2 + (|\boldsymbol{\lambda}_{1z}^* + \boldsymbol{\Delta}| - |\boldsymbol{\lambda}_{1z}^*|)^\top \boldsymbol{\delta} \\ &\geq -\|\boldsymbol{\Delta}\|_2 \left\| (n^*)^{-1} \sum_{i=1}^{n^*} \{-n^* \tilde{Z}_i \rho'(\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}_{1z}^*) \mathbf{B}(\mathbf{X}_i) + \mathbf{B}(\mathbf{X}_i)\} \right\|_2 + (\boldsymbol{\Delta}^\top \underline{\mathbf{M}} \boldsymbol{\Delta}) / 2 - |\boldsymbol{\Delta}|^\top \boldsymbol{\delta}, \quad (64) \end{aligned}$$

where $\underline{\mathbf{M}} = (n^*)^{-1} \sum_{i=1}^{n^*} \{-n^* \tilde{Z}_i \rho''(\mathbf{B}(\mathbf{X}_i)^\top \tilde{\boldsymbol{\lambda}}) \mathbf{B}(\mathbf{X}_i) \mathbf{B}(\mathbf{X}_i)^\top\}$. Here the last inequality is due to the Cauchy-Schwarz inequality (for the first term) and the triangle inequality (for the third term).

By Cauchy Schwarz, we get,

$$\begin{aligned} & G(\boldsymbol{\lambda}_{1z}^* + \boldsymbol{\Delta}) - G(\boldsymbol{\lambda}_{1z}^*) \\ &\geq (\boldsymbol{\Delta}^\top \underline{\mathbf{M}} \boldsymbol{\Delta}) / 2 - r \left(\left\| (n^*)^{-1} \sum_{i=1}^{n^*} \{-n^* \tilde{Z}_i \rho'(\mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}_{1z}^*) \mathbf{B}(\mathbf{X}_i) + \mathbf{B}(\mathbf{X}_i)\} \right\|_2 + \|\boldsymbol{\delta}\|_2 \right), \quad (65) \end{aligned}$$

since $\|\Delta\|_2 = r$. Now,

$$\begin{aligned}
& \left\| (n^*)^{-1} \sum_{i=1}^{n^*} \{-n^* \tilde{Z}_i \rho' \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}_{1z}^* \} \mathbf{B}(\mathbf{X}_i) + \mathbf{B}(\mathbf{X}_i) \} \right\|_2 \\
& \leq \left\| (n^*)^{-1} \sum_{i=1}^{n^*} \{-\tilde{Z}_i n^* \rho' \{ g_z^*(\mathbf{X}_i) \} \mathbf{B}(\mathbf{X}_i) + \mathbf{B}(\mathbf{X}_i) \} \right\|_2 + \\
& \left\| (n^*)^{-1} \sum_{i=1}^{n^*} -\tilde{Z}_i [n^* \rho' \{ g_z^*(\mathbf{X}_i) \} - n^* \rho' \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}_{1z}^* \}] \mathbf{B}(\mathbf{X}_i) \right\|_2 \\
& \leq \left\| (n^*)^{-1} \sum_{i=1}^{n^*} \{1 - \tilde{Z}_i w^{\text{IP}}(\mathbf{X}_i)\} \mathbf{B}(\mathbf{X}_i) \right\|_2 + \sup_{\mathbf{x} \in \mathcal{X}} |g_z^*(\mathbf{x}) - \mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}_{1z}^*| \left\| (n^*)^{-1} \sum_{i=1}^{n^*} -\tilde{Z}_i \mathbf{B}(\mathbf{X}_i) \right\|_2 \\
& \leq \left\| (n^*)^{-1} \sum_{i=1}^{n^*} \{1 - \tilde{Z}_i w^{\text{IP}}(\mathbf{X}_i)\} \mathbf{B}(\mathbf{X}_i) \right\|_2 + O(K^{-r_z})(n^*)^{-1} \|\mathbf{B}(\mathbf{X}_i)\|_2 \\
& \leq O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} \right\} + CK^{-r_z+1/2} = O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\}. \quad (66)
\end{aligned}$$

Here, the first inequality is due to the triangle inequality, and the second inequality is due to the mean value theorem. The third inequality is due to Assumption 2 and the triangle inequality. Finally, the fourth inequality is due to Lemma 8.5 and Assumption 4(b). Equation 66 combined with Assumption 4(e) implies that with probability tending to one,

$$\begin{aligned}
& G(\boldsymbol{\lambda}_{1z}^* + \Delta) - G(\boldsymbol{\lambda}_{1z}^*) \\
& \geq (\Delta^\top \underline{\mathbf{M}} \Delta)/2 - r O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\} \\
& = (2n^*)^{-1} \sum_{i=1}^{n^*} [-n^* \tilde{Z}_i \rho''(\mathbf{B}(\mathbf{X}_i)^\top \tilde{\boldsymbol{\lambda}}) \{\Delta^\top \mathbf{B}(\mathbf{X}_i)\}^2] - r O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\} \\
& \geq C(n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i \{\Delta^\top \mathbf{B}(\mathbf{X}_i)\}^2 - r O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\} \\
& = C \Delta^\top \left\{ \sum_{i: \tilde{Z}_i=1} (n^*)^{-1} \mathbf{B}(\mathbf{X}_i) \mathbf{B}(\mathbf{X}_i)^\top \right\} \Delta - r O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\} \\
& \geq Cr^2 \lambda_{\min} \left(\sum_{i: \tilde{Z}_i=1} (n^*)^{-1} \mathbf{B}(\mathbf{X}_i) \mathbf{B}(\mathbf{X}_i)^\top \right) - r O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\} \\
& \geq C'' r^2 - r O_P \left\{ (K^{1/4} \log K)/(n^*)^{1/2} + K^{-r_z+1/2} \right\} \\
& = C'' (C^*)^2 \left(K^{1/4} \{(\log K)/n^*\}^{1/2} + K^{-r_z+1/2} \right)^2 \\
& \quad - C^* O_P \left(\left\{ K^{1/4} \{(\log K)/n^*\}^{1/2} + K^{-r_z+1/2} \right\}^2 \right) > 0. \quad (67)
\end{aligned}$$

Here the second inequality holds due to Assumption 4(a). The third inequality holds since for a square matrix \mathbf{A} , $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq \lambda_{\min}(\mathbf{A}) \|\mathbf{x}\|_2^2$. The fourth inequality is due to Lemma 8.6. Finally, the fifth inequality holds for a choice of C^* large enough. This completes the proof of Lemma 8.3 and Theorem 4.2.

8.6 Proofs of propositions and theorems

8.6.1 Proof of Theorem 1

We first show that when Assumption 3 holds, $\sum_{i:Z_i=z} \hat{w}_i Y_i(z) \xrightarrow[n \rightarrow \infty]{P} E_{\mathbb{T}}\{Y(z)\}$, for $z \in \{0, 1\}$. For simplicity, we denote $\tilde{Z}_i = D_i \mathbb{1}(Z_i = z)$. Also, denote $\bar{\mathbf{B}}^* = (n^*)^{-1} \sum_{i=1}^{n^*} \mathbf{B}(\mathbf{X}_i)$. Writing $Y_i(z) = m_z(\mathbf{X}_i) + \epsilon_{iz}$, where $E_{\mathbb{T}}(\epsilon_{iz} | \mathbf{X}_i) = 0$, we get the following decomposition.

$$\begin{aligned}
& \left| \sum_{i:Z_i=z} \hat{w}_i Y_i(z) - E_{\mathbb{T}}\{Y(z)\} \right| \\
& \leq \left| \sum_{i=1}^{n^*} \tilde{Z}_i \hat{w}_i \{m_z(\mathbf{X}_i) - \boldsymbol{\lambda}_{2z}^{*\top} \mathbf{B}(\mathbf{X}_i)\} \right| + \left| \boldsymbol{\lambda}_{2z}^{*\top} \left\{ \sum_{i:Z_i=z} \hat{w}_i \mathbf{B}(\mathbf{X}_i) - \bar{\mathbf{B}}^* \right\} \right| \\
& + \left| \boldsymbol{\lambda}_{2z}^{*\top} \bar{\mathbf{B}} - (n^*)^{-1} \sum_{i=1}^{n^*} m_z(\mathbf{X}_i) \right| + \left| (n^*)^{-1} \sum_{i=1}^{n^*} m_z(\mathbf{X}_i) - E_{\mathbb{T}}\{m_z(\mathbf{X}_i)\} \right| + \left| \sum_{i=1}^{n^*} \tilde{Z}_i \hat{w}_i \epsilon_{iz} \right| \\
& \leq \sup_{\mathbf{x} \in \mathcal{X}} |m_z(\mathbf{x}) - \boldsymbol{\lambda}_{2z}^{*\top} \mathbf{B}(\mathbf{x})| \sum_{i=1}^{n^*} \tilde{Z}_i |\hat{w}_i| + |\boldsymbol{\lambda}_{2z}^*|^\top \boldsymbol{\delta} + \sup_{\mathbf{x} \in \mathcal{X}} |m_z(\mathbf{x}) - \boldsymbol{\lambda}_{2z}^{*\top} \mathbf{B}(\mathbf{X}_i)| \\
& + o_P(1) + \left| \sum_{i=1}^{n^*} \tilde{Z}_i \hat{w}_i \epsilon_{iz} \right| \\
& \leq O(K^{-s_z}) |(n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i n^* \rho' \{\boldsymbol{\lambda}^{\dagger\top} \mathbf{B}(\mathbf{X}_i)\}| + \|\boldsymbol{\lambda}_{2z}^*\|_2 \|\boldsymbol{\delta}\|_2 + O(K^{-s_z}) + o_P(1) + \left| \sum_{i=1}^{n^*} \tilde{Z}_i \hat{w}_i \epsilon_{iz} \right| \\
& = o_P(1) + \left| (n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i n^* \rho' \{\boldsymbol{\lambda}^{\dagger\top} \mathbf{B}(\mathbf{X}_i)\} \epsilon_{iz} \right|. \tag{68}
\end{aligned}$$

Here the first inequality is due to the triangle inequality. In the second inequality, we bound the imbalances $|\sum_{i:Z_i=z} \hat{w}_i \mathbf{B}(\mathbf{X}_i) - \bar{\mathbf{B}}^*|$ by $\boldsymbol{\delta}$ (component-wise). The third inequality holds due to the Cauchy-Schwarz inequality (for the second term) and Assumption 3 (for the first and third terms). The final equality is due to assumptions 3 and 4(a). Now,

$$(1 - c_0)^{-1} (n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i \epsilon_{iz} \leq (n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i n^* \rho' \{\boldsymbol{\lambda}^{\dagger\top} \mathbf{B}(\mathbf{X}_i)\} \epsilon_{iz} \leq c_0^{-1} (n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i \epsilon_{iz}. \tag{69}$$

Both the upper and lower bounds converge to a constant times $E_{\mathbb{T}}(\tilde{Z}_i \epsilon_{iz}) = E_{\mathbb{T}}\{\pi(\mathbf{X}_i)P_{\mathbb{P}}(Z_i = 1|\mathbf{X}_i)E_{\mathbb{T}}(\epsilon_{iz}|\mathbf{X}_i)\} = 0$, by Assumption 1. Therefore, $\left|(n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i n^* \rho' \{\boldsymbol{\lambda}^{\dagger \top} \mathbf{B}(\mathbf{X}_i)\} \epsilon_{iz}\right| = o_P(1)$.

We now show that when Assumption 2 holds, $\sum_{i:Z_i=z} \hat{w}_i Y_i(z) \xrightarrow[n \rightarrow \infty]{P} E_{\mathbb{T}}\{Y(z)\}$.

$$\begin{aligned}
& \left| \sum_{i:Z_i=z} w_i Y_i^{\text{obs}} - E_{\mathbb{T}}\{Y(z)\} \right| \\
& \leq \left| \sum_{i:Z_i=z} w_i Y_i(z) - (n^*)^{-1} \sum_{i:Z_i=z} w_i^{\text{IP}} Y_i(z) \right| + \left| (n^*)^{-1} \sum_{i:Z_i=z} w_i^{\text{IP}} Y_i(z) - E_{\mathbb{T}}\{Y(z)\} \right| \\
& \leq \sup_{\mathbf{x} \in \mathcal{X}} \left| n^* w(\mathbf{x}) - w^{\text{IP}}(\mathbf{x}) \right| \left\{ (n^*)^{-1} \sum_{i:Z_i=z} |Y_i(z)| \right\} + o_P(1) \\
& = o_P(1),
\end{aligned} \tag{70}$$

where the last step holds due to Theorem 4.2. This completes the proof.

8.6.2 Proof of Theorem 2

For convenience, we restate Assumption 5 below.

Assumption 5. For $z \in \{0, 1\}$,

- (a) $E_{\mathbb{T}}\{Y^2(z)\} < \infty$.
- (b) Let $g_z^*(\cdot) \in \mathcal{G}_z$. \mathcal{G}_z satisfies $\log N_{[]} \{\epsilon, \mathcal{G}_z, L_2(P)\} \leq C_1 (1/\epsilon)^{1/k_1}$ for some constants $C_1 > 0$ and $k_1 > 1/2$, where $N_{[]} \{\epsilon, \mathcal{G}_z, L_2(P)\}$ is the covering number of \mathcal{G}_z by epsilon brackets.
- (c) Let $m_z(\cdot) \in \mathcal{M}_z$. \mathcal{M}_z satisfies $\log N_{[]} \{\epsilon, \mathcal{M}_z, L_2(P)\} \leq C_2 (1/\epsilon)^{1/k_2}$ for some constants $C_2 > 0$ and $k_2 > 1/2$, where $N_{[]} \{\epsilon, \mathcal{M}_z, L_2(P)\}$ is the covering number of \mathcal{M}_z by epsilon brackets.
- (d) $(n^*)^{\{2(r_z + s_z - 0.5)\}^{-1}} = o(K)$, where r_z, s_z are the constants in assumptions 2 and 3, respectively.

The conditions in Assumption 5 are similar to Assumption 2 in Wang and Zubizarreta (2020) and Assumption 4.1 in Fan et al. (2016). In particular, Assumption 5(a) ensures existence of the second moment of $Y(z)$ with respect to the target distribution. Assumptions 5(b) and (c) control the complexity of the function classes \mathcal{G}_z and \mathcal{M}_z . Finally, Assumption 5(d) puts further restriction on the growth rate of the number of basis functions K as a function of n^* .

All probabilities and expectations in this proof are computed with respect to the probability measure \mathbb{T} . We first decompose the Hajek estimator $T - \tau = \{\sum_{i:Z_i=1} \hat{w}_i Y_i^{\text{obs}} - \sum_{i:Z_i=0} \hat{w}_i Y_i^{\text{obs}}\} - E_{\mathbb{T}}\{Y(1) - Y(0)\}$ as follows.

$$T - \tau = S + R_0 + R_1 + R_2 + \tilde{R}_0 + \tilde{R}_1 + \tilde{R}_2, \quad (71)$$

where

$$\begin{aligned} S &= (n^*)^{-1} \sum_{i=1}^{n^*} \{m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i)\} + (n^*)^{-1} \sum_{i=1}^{n^*} (D_i Z_i) / \{e(\mathbf{X}_i) \pi(\mathbf{X}_i)\} \{Y_i^{\text{obs}} - m_1(\mathbf{X}_i)\} \\ &\quad - (n^*)^{-1} \sum_{i=1}^{n^*} (D_i (1 - Z_i)) / \{[1 - e(\mathbf{X}_i)] \pi(\mathbf{X}_i)\} \{Y_i^{\text{obs}} - m_0(\mathbf{X}_i)\} - \tau, \\ R_0 &= \sum_{i=1}^{n^*} D_i Z_i [\hat{w}_i - \{n^* e(\mathbf{X}_i) \pi(\mathbf{X}_i)\}^{-1}] \{Y_i(1) - m_1(\mathbf{X}_i)\}, \\ R_1 &= \sum_{i=1}^{n^*} (D_i Z_i \hat{w}_i - (n^*)^{-1}) \{m_1(\mathbf{X}_i) - \boldsymbol{\lambda}_{21}^{*\top} \mathbf{B}(\mathbf{X}_i)\}, \\ R_2 &= \sum_{i=1}^{n^*} (D_i Z_i \hat{w}_i - (n^*)^{-1}) \{\boldsymbol{\lambda}_{21}^{*\top} \mathbf{B}(\mathbf{X}_i)\}, \\ \tilde{R}_0 &= \sum_{i=1}^{n^*} D_i (1 - Z_i) [\hat{w}_i - \{n^* \{1 - e(\mathbf{X}_i)\} \pi(\mathbf{X}_i)\}^{-1}] \{Y_i(0) - m_0(\mathbf{X}_i)\}, \\ \tilde{R}_1 &= \sum_{i=1}^{n^*} \{D_i (1 - Z_i) \hat{w}_i - (n^*)^{-1}\} \{m_0(\mathbf{X}_i) - \boldsymbol{\lambda}_{20}^{*\top} \mathbf{B}(\mathbf{X}_i)\}, \\ \tilde{R}_2 &= \sum_{i=1}^{n^*} \{D_i (1 - Z_i) \hat{w}_i - (n^*)^{-1}\} \{\boldsymbol{\lambda}_{20}^{*\top} \mathbf{B}(\mathbf{X}_i)\}. \end{aligned}$$

By central limit theorem, it follows that,

$$\sqrt{n^*} S \xrightarrow[n^* \rightarrow \infty]{d} \mathcal{N}(0, V), \quad (72)$$

where

$$V = \text{Var} \left(m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) + [D_i Z_i \{Y_i(1) - m_1(\mathbf{X}_i)\}] / \{e(\mathbf{X}_i) \pi(\mathbf{X}_i)\} \right. \quad (73)$$

$$\left. - [D_i (1 - Z_i) \{Y_i(0) - m_0(\mathbf{X}_i)\}] / \{[1 - e(\mathbf{X}_i)] \pi(\mathbf{X}_i)\} \right). \quad (74)$$

V is same as the semiparametric efficiency bound for the target average treatment effect for nested

designs (See [Dahabreh et al. 2019](#), [Li et al. 2021](#)).

For notational convenience, we denote $\tilde{Z}_i = Z_i D_i$. We now consider

$$\begin{aligned}\sqrt{n^*} R_0 &= \sqrt{n^*} \left[\sum_{i=1}^{n^*} \tilde{Z}_i \{ \hat{w}_i - (n^*)^{-1} w_i^{\text{IP}} \} \{ Y_i(1) - m_1(\mathbf{X}_i) \} \right] \\ &= \sqrt{n^*} \left((n^*)^{-1} \sum_{i=1}^{n^*} \tilde{Z}_i \{ Y_i(1) - m_1(\mathbf{X}_i) \} [n^* \rho' \{ \mathbf{B}(\mathbf{X}_i)^\top \boldsymbol{\lambda}^\dagger \} - w_i^{\text{IP}}] \right).\end{aligned}\quad (75)$$

For a function $g_1(\cdot)$, let us define the function

$$f_0(\tilde{Z}, Y(1), \mathbf{X}) := \tilde{Z} \{ Y(1) - m_1(\mathbf{X}) \} \left[n^* \rho' \{ g_1(\mathbf{X}) \} - w^{\text{IP}}(\mathbf{X}) \right] \quad (76)$$

and the corresponding empirical process \mathbb{G}_n given by,

$$\mathbb{G}_n(f_0) = (n^*)^{1/2} \left\{ (n^*)^{-1} \sum_{i=1}^{n^*} f_0(\tilde{Z}_i, Y_i(1), \mathbf{X}_i) - E \{ f_0(\tilde{Z}, Y(1), \mathbf{X}) \} \right\}. \quad (77)$$

First, $E \{ f_0(\tilde{Z}, Y(1), \mathbf{X}) \} = 0$ by Assumption 1. Now, consider a class of functions \mathcal{F} defined as

$$\mathcal{F} = \{ f_0 : \sup_{\mathbf{x} \in \mathcal{X}} |g_1(\mathbf{x}) - g_1^*(\mathbf{x})| \leq \delta_0 \}, \quad (78)$$

where $\delta_0 = C[K^{3/4}\{(\log K)/n^*\}^{1/2} + K^{1-r_z}]$. From the proof of Theorem 4.2, it follows that $\sup_{\mathbf{x} \in \mathcal{X}} |\mathbf{B}(\mathbf{x})^\top \boldsymbol{\lambda}^\dagger - g_1^*(\mathbf{x})| \leq \delta_0$. Hence,

$$\sqrt{n^*} |R_0| \leq \sup_{f_0 \in \mathcal{F}} |\mathbb{G}_n(f_0)| \quad (79)$$

By the Markov inequality, $P(\sup_{f_0 \in \mathcal{F}} |\mathbb{G}_n(f_0)| \geq C) \leq C^{-1} E \{ \sup_{f_0 \in \mathcal{F}} |\mathbb{G}_n(f_0)| \}$, for $C > 0$. Thus, to show $\sqrt{n^*} |R_0| \xrightarrow[n^* \rightarrow \infty]{P} 0$, it is enough to show that $E \{ \sup_{f_0 \in \mathcal{F}} |\mathbb{G}_n(f_0)| \} \xrightarrow[n^* \rightarrow \infty]{P} 0$. Now, assumptions 2 and [4\(a\)](#) imply that for $f_0 \in \mathcal{F}$, $|f_0(Z, Y(1), \mathbf{X})| \leq C' |Y(1) - m_1(\mathbf{X})| \delta_0$ for a constant $C' > 0$. So the function $F_0(Z, Y(1), \mathbf{X}) := C' |Y(1) - m_1(\mathbf{X})| \delta_0$ is an envelope of \mathcal{F} , with $\|F_0\|_{P,2} := [E \{ F_0^2(Z, Y(1), \mathbf{X}) \}]^{1/2} \leq C \delta_0$ for some $C > 0$ by Assumption [5\(a\)](#). By the maximal inequality

(see [Van der Vaart](#) [2000](#) Chapter 19) we have

$$E\left\{\sup_{f_0 \in \mathcal{F}} |\mathbb{G}_n(f_0)|\right\} \lesssim J_{[]}(\|F_0\|_{P,2}, \mathcal{F}, L_2(P)), \quad (80)$$

where $J_{[]}(\|F_0\|_{P,2}, \mathcal{F}, L_2(P)) := \int_0^{\|F_0\|_{P,2}} [\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))]^{1/2} d\epsilon$ is the bracketing integral and \lesssim indicates less than up to a constant. We now use similar steps as in [Fan et al.](#) [\(2016\)](#) and [Wang and Zubizarreta](#) [\(2020\)](#) to bound the log of the bracketing number. Define $\mathcal{F}_0 = \{f_0 : \sup_{\mathbf{x} \in \mathcal{X}} |g_1(\mathbf{x}) - g_1^*(\mathbf{x})| \leq C\}$ for some constant $C > 0$. It follows that, $\log N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \lesssim \log N_{[]}(\epsilon, \delta_0 \mathcal{F}_0, L_2(P)) = \log N_{[]}(\epsilon/\delta_0, \mathcal{F}_0, L_2(P)) \lesssim \log N_{[]}(\epsilon/\delta_0, \mathcal{G}_1, L_2(P)) \lesssim (\delta_0/\epsilon)^{1/k_1}$, where the final inequality holds due to Assumption [5\(c\)](#). This implies,

$$J_{[]}(\|F_0\|_{P,2}, \mathcal{F}, L_2(P)) \lesssim \int_0^{C\delta_0} (\delta_0/\epsilon)^{1/(2k_1)} d\epsilon \lesssim \delta_0/\{1 - 1/(2k_1)\}, \quad (81)$$

where in the last step we used $2k_1 > 1$. The right hand side converges to zero as n^* goes to ∞ .

Thus, $\sqrt{n^*} R_0 \xrightarrow[n^* \rightarrow \infty]{P} 0$. Following similar steps, we can show $\sqrt{n^*} \tilde{R}_0 \xrightarrow[n^* \rightarrow \infty]{P} 0$.

We will now show that $\sqrt{n^*} R_1 \xrightarrow[n^* \rightarrow \infty]{P} 0$ where $R_1 = \sum_{i=1}^{n^*} (\tilde{Z}_i \hat{w}_i - (n^*)^{-1}) \{m_1(\mathbf{X}_i) - \boldsymbol{\lambda}_{21}^{*\top} \mathbf{B}(\mathbf{X}_i)\}$.

Define the function

$$f_1(\tilde{Z}, \mathbf{X}) := [n^* \tilde{Z} \rho' \{g_1(\mathbf{X})\} - 1] \{m_1(\mathbf{X}) - \boldsymbol{\lambda}_{21}^* \mathbf{B}(\mathbf{X})\}, \quad (82)$$

and the corresponding empirical process \mathbb{G}_n given by,

$$\mathbb{G}_n(f_1) = (n^*)^{1/2} \left\{ (n^*)^{-1} \sum_{i=1}^{n^*} f_1(\tilde{Z}_i, \mathbf{X}_i) - E\{f_1(\tilde{Z}, \mathbf{X})\} \right\} \quad (83)$$

Denote $\Delta(\mathbf{x}) = m_1(\mathbf{x}) - \boldsymbol{\lambda}_{21}^{*\top} \mathbf{B}(\mathbf{x})$. Now, consider a class of functions \mathcal{F}_1 defined as,

$$\mathcal{F}_1 = \{f_1 : \sup_{\mathbf{x} \in \mathcal{X}} |g_1(\mathbf{x}) - g_1^*(\mathbf{x})| \leq \delta_1, \sup_{\mathbf{x} \in \mathcal{X}} |\Delta(\mathbf{x})| \leq \delta_2\}, \quad (84)$$

where $\delta_1 = C[K^{3/4}\{(\log K)/n^*\}^{1/2} + K^{1-r_z}]$, $\delta_2 = CK^{-s_1}$ for some constant $C > 0$. As before,

using the proof of Theorem 4.2, we get

$$(n^*)^{1/2}|R_1| \leq \sup_{f_1 \in \mathcal{F}_1} |\mathbb{G}_n(f_1) + \sqrt{n^*}E\{f_1(\tilde{Z}, \mathbf{X})\}| \leq \sup_{f_1 \in \mathcal{F}_1} |\mathbb{G}_n(f_1)| + (n^*)^{1/2} \sup_{f_1 \in \mathcal{F}_1} |E\{f_1(\tilde{Z}, \mathbf{X})\}| \quad (85)$$

We show that each term on the right hand side is $o_P(1)$. For the first term, by the Markov inequality it suffices to show that $E\{\sup_{f_1 \in \mathcal{F}_1} |\mathbb{G}_n(f_1)|\} \xrightarrow{n^* \rightarrow \infty} 0$. Now, assumptions 2, 4(a), and 3 imply that for $f_1 \in \mathcal{F}_1$, $|f_1(Z, \mathbf{X})| \leq C'\delta_2$ for a constant $C' > 0$. So the function $F_1(Z, \mathbf{X}) := C'\delta_2$ is an envelope of \mathcal{F}_1 , with $\|F_1\|_{P,2} \leq C'\delta_2$. By the maximal inequality,

$$E\{\sup_{f_1 \in \mathcal{F}_1} |\mathbb{G}_n(f_1)|\} \lesssim J_{[]}(\|F_1\|_{P,2}, \mathcal{F}_1, L_2(P)), \quad (86)$$

where

$$J_{[]}(\|F_1\|_{P,2}, \mathcal{F}_1, L_2(P)) \lesssim \int_0^{C'\delta_2} \{\log N_{[]}(\epsilon, \mathcal{F}_1, L_2(P))\}^{1/2} d\epsilon. \quad (87)$$

Define $\mathcal{F}_0 := \{f_1 : \sup_{\mathbf{x} \in \mathcal{X}} |g_1(\mathbf{x}) - g_1^*(\mathbf{x})| \leq C, \sup_{\mathbf{x} \in \mathcal{X}} |\Delta(\mathbf{x})| \leq 1\}$ for some constant $C > 0$, $\mathcal{H}_{10} := \{\gamma \in \mathcal{G}_1 + g_1^* : \sup_{\mathbf{x} \in \mathcal{X}} |\gamma(\mathbf{x})| \leq C\}$, $\mathcal{H}_{20} := \{\Delta \in \mathcal{M}_1 - \boldsymbol{\lambda}_{21}^{*\top} \mathbf{B}(\mathbf{x}) : \sup_{\mathbf{x} \in \mathcal{X}} |\Delta(\mathbf{x})| \leq 1\}$. Using similar steps as in Fan et al. (2016) and Wang and Zubizarreta (2020) to bound the log of this bracketing number, we get $\log N_{[]}(\epsilon, \mathcal{F}_1, L_2(P)) \lesssim \log N_{[]}(\epsilon/\delta_2, \mathcal{F}_0, L_2(P)) \lesssim \log N_{[]}(\epsilon/\delta_2, \mathcal{H}_{10}, L_2(P)) + \log N_{[]}(\epsilon/\delta_2, \mathcal{H}_{20}, L_2(P)) \lesssim \log N_{[]}(\epsilon/\delta_2, \mathcal{G}_1, L_2(P)) + \log N_{[]}(\epsilon/\delta_2, \mathcal{M}_1, L_2(P)) \leq (\delta_2/\epsilon)^{1/k_1} + (\delta_2/\epsilon)^{1/k_2}$, where the final inequality holds due to assumptions 5(b) and 5(c). Thus,

$$J_{[]}(\|F_1\|_{P,2}, \mathcal{F}_1, L_2(P)) \lesssim \delta_2/\{1 - 1/(2k_1)\} + \delta_2/\{1 - 1/(2k_2)\}, \quad (88)$$

since $2k_1 > 1$ and $2k_2 > 1$. The right hand side converges to zero as n^* goes to infinity. Thus,

$$\sup_{f_1 \in \mathcal{F}_1} |\mathbb{G}_n(f_1)| = o_P(1).$$

Now, let $\mathcal{H}_1 = \{g \in \mathcal{G}_1 : \sup_{\mathbf{x} \in \mathcal{X}} |g_1(\mathbf{x}) - g_1^*(\mathbf{x})| \leq \delta_1\}$ and $\mathcal{H}_2 = \{\Delta \in \mathcal{M}_1 - \boldsymbol{\lambda}_{21}^{*\top} \mathbf{B} : \sup_{\mathbf{x} \in \mathcal{X}} |\Delta(\mathbf{x})| \leq$

$\delta_2\}$.

$$\begin{aligned}
(n^*)^{1/2} \sup_{f_1 \in \mathcal{F}_1} |E\{f_1(\tilde{Z}, \mathbf{X})\}| &= (n^*)^{1/2} \sup_{g_1 \in \mathcal{H}_1, \Delta \in \mathcal{H}_2} |E(\{n^* \tilde{Z} \rho' \{g_1(\mathbf{X})\} - 1\} \Delta(\mathbf{X}))| \\
&= (n^*)^{1/2} \sup_{g_1 \in \mathcal{H}_1, \Delta \in \mathcal{H}_2} \left| E \left[\left([n^* \rho' \{g_1(\mathbf{X})\}] / [n^* \rho' \{g_1^*(\mathbf{X})\}] - 1 \right) \Delta(\mathbf{X}) \right] \right| \\
&\lesssim (n^*)^{1/2} \sup_{g_1 \in \mathcal{H}_1, \Delta \in \mathcal{H}_2} \left\{ \sup_{\mathbf{x} \in \mathcal{X}} |g_1(\mathbf{x}) - g_1^*(\mathbf{x})| \right\} \left\{ \sup_{\mathbf{x} \in \mathcal{X}} |\Delta(\mathbf{x})| \right\} \\
&\lesssim (n^*)^{1/2} \delta_1 \delta_2 \\
&\lesssim K^{-s_1+3/4} \log K + (n^*)^{1/2} K^{-r_1-s_1+1/2}.
\end{aligned} \tag{89}$$

Here the first inequality holds by applying the mean value theorem and assumptions 4(a), 2, and 4(a). The second inequality holds by definition of \mathcal{H}_1 and \mathcal{H}_2 . Finally, by assumptions 3 and 5(d), the right hand side of Equation 89 goes to zero as n^* goes to infinity.

Thus, $(n^*)^{1/2} R_1 \xrightarrow[n^* \rightarrow \infty]{P} 0$. Following similar steps, we can show $(n^*)^{1/2} \tilde{R}_1 \xrightarrow[n^* \rightarrow \infty]{P} 0$. Finally, we consider $R_2 = \sum_{i=1}^{n^*} (\tilde{Z}_i \hat{w}_i - (n^*)^{-1}) \{\boldsymbol{\lambda}_{21}^* \mathbf{B}(\mathbf{X}_i)\}$. We observe that,

$$\begin{aligned}
(n^*)^{1/2} |R_2| &\leq \|\boldsymbol{\lambda}_{21}^*\|_2 \left\| \sum_{i: \tilde{Z}_i=1} \hat{w}_i \mathbf{B}(\mathbf{X}_i) - (n^*)^{-1} \sum_{i=1}^{n^*} \mathbf{B}(\mathbf{X}_i) \right\|_2 \\
&\leq \|\boldsymbol{\lambda}_{21}^*\|_2 \|\boldsymbol{\delta}\|_2 = o(1)
\end{aligned} \tag{90}$$

The first inequality is due to Cauchy-Schwarz; the second inequality holds by construction of the weights, and the third equality holds by Assumption 3. This implies, $(n^*)^{1/2} R_2 \xrightarrow[n^* \rightarrow \infty]{P} 0$. Following similar steps, we can show $(n^*)^{1/2} \tilde{R}_2 \xrightarrow[n^* \rightarrow \infty]{P} 0$. This completes the proof of the theorem.

8.7 Details for the Simulation Study

In this section, we include additional results on the performance of the one-step and two-step estimators. We also include their performance when the estimand is $\mathbb{E}[Y(1) - Y(0)|S = 0]$ (transportability) rather than $\mathbb{E}[Y(1) - Y(0)|S = 1]$ (generalizability).

For the settings in the main text, Table 4 shows the bias of the Hájek estimators of the target average treatment effect under each weighting method, based on 800 simulations. The one-step estimators

tend to perform better than the corresponding two-step estimators across the three outcome models, in both the randomized and observational study settings. In the randomized study, the one-step estimators reduce the absolute bias by 85% in the misspecified cases and by 79% in the correctly specified cases relative to the corresponding two-step estimators, on average. In the observational study, the biases are typically higher than their experimental counterparts, particularly for the misspecified cases. Here, the one-step estimators reduce the absolute bias by 10% in the misspecified cases and by 92% in the correctly specified cases relative to the corresponding two-step estimators, on average. In the correctly specified cases, biases are reduced substantially under the one-step method because by construction, the one-step method balances the right functions of covariates relative to the target.

Table 4: Bias of the Hájek estimator of the target average treatment effect using different weighting methods in both the randomized and observational study settings.

Weighting Method	Randomized Study Setting			Observational Study Setting		
	Outcome	Outcome	Outcome	Outcome	Outcome	Outcome
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Two-Step Method 1	0.67	3.30	5.02	-9.26	-6.45	-4.77
One-Step Method 1	-0.02	0.53	2.04	-8.62	-8.00	-6.46
Two-Step Method 2	0.79	3.59	3.15	-32.12	-27.84	-32.71
One-Step Method 2	-0.03	0.50	-0.43	-17.16	-16.21	-23.97
Two-Step Method 3	0.18	0.20	0.14	0.63	0.91	1.05
One-Step Method 3	0.01	-0.03	-0.06	-0.10	-0.04	0.03

For the transportability design, we slightly modify the setup in Section 5. There are $n_{\text{study}} = 500$ units in the study sample and $n_{\text{target}} = 5000$ units in the target sample. Let $S = 1$ indicate a unit is in the study sample and $S = 0$ indicate a unit is in the target sample. For units with $S = 1$, the four independent latent covariates are $U_1, U_2, U_3, U_4 \sim \mathcal{N}(0, 1)$. For units with $S = 0$, they are $U_1 \sim \mathcal{N}(1.2, 1)$, $U_2 \sim \mathcal{N}(-0.4, 1)$, $U_3 \sim \mathcal{N}(0.3, 1)$, and $U_4 \sim \mathcal{N}(0.1, 1)$. This creates similar correlations between the latent covariates and S as in the setting detailed in Section 5. The remaining setup is the same as that in Section 5.

Table 5 shows the root-mean-squared errors and Table 6 shows the mean bias of the Hájek estimators of the target average treatment effect. Figure 4 shows how the effective sample sizes and maximum normalized weights vary across simulations. Overall, the pattern of results is similar as in Section 5, albeit with a more dramatic improvement in performance by the one-step weights, likely due to the more limited covariate overlap. A few exceptions occur for the observational study, where the two-step weights produce less biased estimates than the one-step weights for outcome model 1 (i.e., no treatment effect heterogeneity), though the one-step weights still show improved root-mean-squared error.

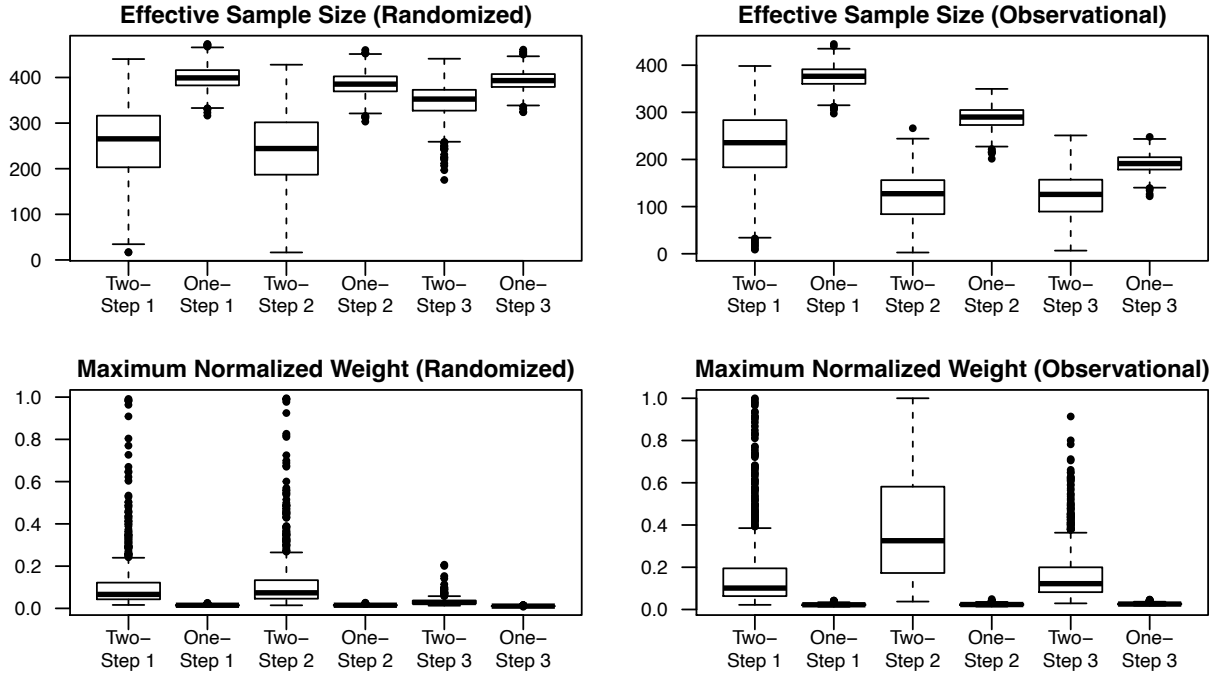
Table 5: Root-mean-squared error of the Hájek estimator of the target average treatment effect using different weighting methods in both the randomized and observational study settings (transportability setting).

Weighting Method	Randomized Study Setting			Observational Study Setting		
	Outcome	Outcome	Outcome	Outcome	Outcome	Outcome
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Two-Step 1	30.35	45.61	48.11	29.56	45.78	47.33
One-Step 1	2.32	2.52	3.24	5.64	5.15	5.59
Two-Step 2	31.88	46.71	49.74	31.90	46.02	48.44
One-Step 2	1.56	1.81	2.42	14.05	12.98	20.97
Two-Step 3	8.81	19.15	18.57	12.81	22.27	23.20
One-Step 3	0.52	0.81	1.22	0.72	0.95	1.33

Table 6: Bias of the Hájek estimator of the target average treatment effect using different weighting methods in both the randomized and observational study settings (transportability setting).

Weighting Method	Randomized Study Setting			Observational Study Setting		
	Outcome	Outcome	Outcome	Outcome	Outcome	Outcome
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Two-Step 1	-0.29	27.29	55.09	-0.37	28.08	55.13
One-Step 1	-0.06	-0.07	-0.07	-5.12	-4.47	3.21
Two-Step 2	-0.38	26.87	54.85	-3.39	24.10	55.07
One-Step 2	-0.02	0.00	-0.04	-13.95	-12.83	7.69
Two-Step 3	-0.27	16.04	32.80	0.03	15.92	31.17
One-Step 3	-0.02	-0.02	-0.09	0.21	0.29	-0.06

Figure 4: Effective sample sizes and maximum normalized weights across weighting methods in both the randomized and observational study settings (transportability setting).



8.8 Simulation study with heteroscedastic outcomes

In this section, we compare the one-step and two-step estimators in settings where the potential outcomes are heteroscedastic. We use the same setting as in Section 5 with four independent unobserved covariates distributed as $U_1, U_2, U_3, U_4 \sim \mathcal{N}(0, 1)$, and four observed covariates generated as $X_1 = \exp(U_1/2)$, $X_2 = U_2/\{1 + \exp(U_1)\} + 10$, $X_3 = (U_1 U_3/25 + 0.6)^3$, and $X_4 = (U_2 + U_4 + 20)^2$. D is the binary indicator for selection into the study, and Z is the binary treatment indicator. The true model for the probability of selection into the study is $\text{pr}(D = 1|\mathbf{U}) = \text{expit}(-U_1 + 0.5U_2 - 0.25U_3 - 0.1U_4)$ so that, marginally, $\text{pr}(D = 1) = 0.5$. The total cohort size is 1000. For the randomized study setting, $\text{pr}(Z = 1|\mathbf{U}) = 0.5$, and for the observational setting, $\text{pr}(Z = 1|\mathbf{U}) = \text{expit}(U_1 + 2U_2 - 2U_3 - U_4)$.

We consider three different models for $Y(0)$ and $Y(1)$. Under Model- j ($j \in \{1, 2, 3\}$), $Y(0) = 210 + 27.4U_1 + 13.7U_2 + 13.7U_3 + 13.7U_4 + h_j(\mathbf{U})\epsilon_0$, where $\epsilon_0 \sim \mathcal{N}(0, 5^2)$. This model allows for heteroscedasticity, since the conditional variance varies as a function of the covariates, i.e., $\text{Var}_{\mathbb{T}}(Y(0)|\mathbf{U} = \mathbf{u}) = 25h_j^2(\mathbf{u})$. We set $h_1(\mathbf{u}) = 2u_1$, $h_2(\mathbf{u}) = 2(u_1 + u_2)$, and $h_3(\mathbf{u}) = 2(u_1 + u_2 + u_3 + u_4)$. Similarly, there are three models for $Y(1)$: Model 1 is given by $Y(1) = 210 + 27.4U_1 + 13.7U_2 + 13.7U_3 + 13.7U_4 + h_1(\mathbf{U})\epsilon_1$; Model 2 by $Y(1) = 210 + 41.1U_1 + 13.7U_2 + 13.7U_3 + 13.7U_4 + h_2(\mathbf{U})\epsilon_1$; and Model 3 by $Y(1) = 210 + 41.1U_1 + 27.4U_2 + 27.4U_3 + 13.7U_4 + h_3(\mathbf{U})\epsilon_1$; where $\epsilon_1 \sim \mathcal{N}(0, 5^2)$ and $h_j(\cdot)$ s are the same as those for the models of $Y(0)$.

We compare three versions of one-step weighting to three versions of two-step weighting as specified in Section 5. Table 7 shows the root-mean-squared errors of the Hájek estimators, based on 800 simulations. We observe that the one-step weights outperform the two-step weights across the three outcome models in both the randomized and observational study settings.

Table 7: Root-mean-squared error of the Hájek estimator of the target average treatment effect using different weighting methods in both the randomized and observational study settings under heteroscedastic outcomes.

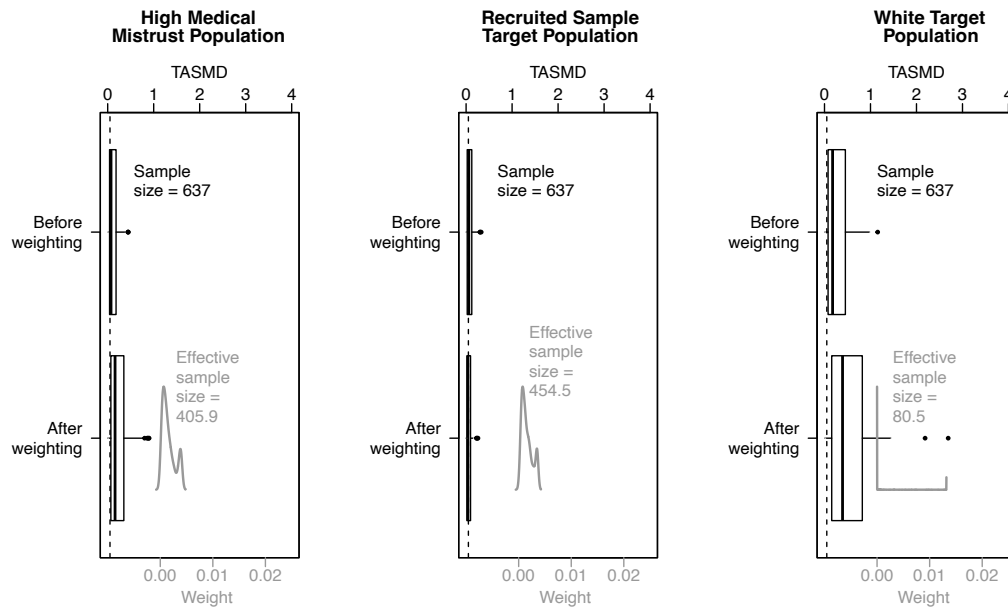
Weighting Method	Randomized Study Setting			Observational Study Setting		
	Outcome Model 1	Outcome Model 2	Outcome Model 3	Outcome Model 1	Outcome Model 2	Outcome Model 3
Two-Step 1	19.23	23.21	26.33	23.83	24.96	25.72
One-Step 1	3.03	3.57	4.75	9.23	8.69	7.93
Two-Step 2	20.83	25.25	28.09	44.45	42.65	48.17
One-Step 2	2.42	3.02	3.84	17.34	16.41	24.41
Two-Step 3	5.07	5.73	6.58	9.68	12.64	13.92
One-Step 3	0.89	1.52	2.03	1.22	1.73	2.99

8.9 Additional Case Study Results

In this section, we present results from the case study in Section 5, albeit with the weights implemented via the two-step method. To calculate the two-step weights, we fit logistic regression models for treatment and study selection, and we trim each set of weights at their 90th percentiles.

Figure 5 summarizes the performance of the two-step weighting method for achieving balance relative to the various target covariate profiles. The figure also summarizes the dispersion of the weights via density plots and effective sample sizes. Compared to the weights in Figure 2, each set of two-step weights is less evenly dispersed and has higher variance and lower effective sample size, reflecting the one-step method’s explicit optimization for these criteria. Covariate balance is also worse than in the one-step approach, again because the one-step weights explicitly target covariate balance. This pattern becomes more stark as the profile becomes more difficult to target (i.e., as there is less overlap between the target and study populations).

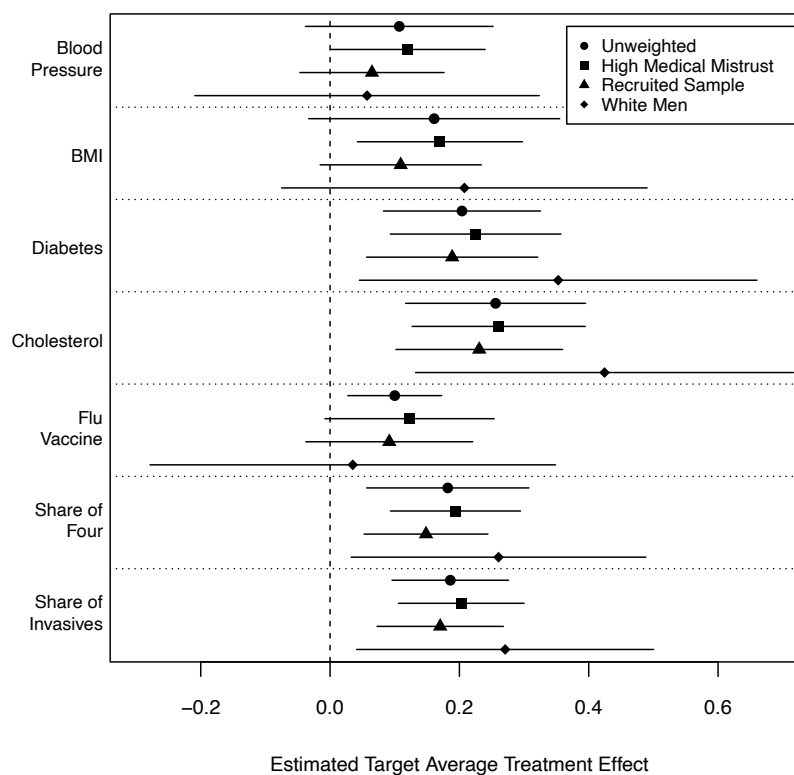
Figure 5: Distributions of target absolute standardized mean differences and effective sample sizes for three target populations (two-step).



TASMD = target absolute standardized mean difference. The black vertical dashed line in each plot marks a TASMD of 0.05, signifying the heuristic that a $\text{TASMD} < 0.05$ indicates good balance. SD = standard deviation.

Figure 6 presents the Hájek estimates of the target average treatment effect for each outcome along with bootstrapped confidence intervals. The results are similar as those due to the one-step weights in Figure 3, however, due to the higher variability of the two-step weights, the confidence intervals are much wider.

Figure 6: Estimates of the target average treatment effect for various outcome variables and target populations.



8.10 R Code

In this section, we provide instructions on how to implement the one-step weights using `sbw` package for R. First, however, we recommend installing `gurobi`, an optimizer which increases the performance of `sbw`. Instructions on installation can be found on <https://www.gurobi.com>.

Next, install the `sbw` package in R via the code `install.packages("sbw")`. Now, in this section, we provide example code that would be used to weight the data set from the applied study corresponding to the randomized participants. We weight these data toward a covariate profile constructed from the covariate means of the entire recruited sample.

First, read in the trial data (`oakland_analysis_final.dta`) and the recruited sample data (`oakland_analysis_selection.dta`) and list the covariates to balance.

```

> library(sbw)
> oakland.final <- read_dta("oakland_analysis_final.dta")
> oakland.selection <- read_dta("oakland_analysis_selection.dta")
> T1.vars <- c("good_sa_health",
>              "any_health_prob",
>              "ER_2years",
>              "nights_hosp_2years",
>              "hosp_visits_2years",
>              "med_mistrust",
>              "has_PCP",
>              "uninsured",
>              "age",
>              "married",
>              "unemployed",
>              "HSless",
>              "low_income",
>              "benefits")

```

For the sake of this analyses, we have imputed missing values with the mean for continuous covariates and created an additional missing category for categorical covariates. For continuous imputed covariates, we also add a dummy variable indicating missingness. These additional variables are added to the list to balance. For the sake of space, we omit including the code that performs these imputations. We assume that the final list of covariates is included in the list `T1.vars.imp`. We assume the data sets `oakland.selection` and `oakland.final` have been updated accordingly.

Next, we define the balance tolerances. We define each covariate's tolerance as 0.1 times the covariate standard deviation in the recruited sample.

```

> sd_targets <- apply(as.matrix(oakland.selection[T1.vars.imp]), 2, sd)
> tols <- 0.1 * sd_targets

```

Next we subset the trial data by treatment group (i.e., values of `black_dr`) — as we want to weight each treatment group toward the target profile. Then we define the various inputs to the `sbw` function. For this first implementation, we manually set the tolerances, restrict the weights to be positive (`wei_pos = TRUE`), and restrict the weights to sum to one (`wei_sum = TRUE`).

```

> t_ind <- "black_dr"
> dat.1 <- subset(dat, dat[t_ind][,1] == 1)
> dat.0 <- subset(dat, dat[t_ind][,1] == 0)
> bal_cov <- T1.vars.imp

```

```

> bal_alg <- FALSE
> bal_tol <- tols
> bal_std <- "manual"
> bal <- list(bal_cov = bal_cov, bal_alg = bal_alg, bal_tol = bal_tol, bal_std = bal_std)
> wei <- list(wei_sum = TRUE, wei_pos = TRUE)
> par_tar <- colMeans(oakland.selection[T1.vars.imp])

```

Next, we balance each treatment group and combine the weighted data into a single data set with both treatment groups.

```

> sbw.results.1 <- sbw(dat = dat.1, bal = bal, par = list(par_est = "aux", par_tar
= par_tar), sol = list(sol_nam = "gurobi"),wei = wei)
> sbw.results.0 <- sbw(dat = dat.0, bal = bal, par = list(par_est
= "aux", par_tar = par_tar), sol = list(sol_nam = "gurobi"),wei = wei)
> weighted.df.1 <- sbw.results.1$dat_weights
> weighted.df.0 <- sbw.results.0$dat_weights
> weighted.df <- rbind(weighted.df.1, weighted.df.0)

```

Using the weighted data, we can compute the TATE directly via weighted means. For the sake of demonstration, we evaluate the TATE for the outcome that measures whether the participant elected to receive a flu shot after their doctor's visit.

```

> mean.1 <- weighted.mean(weighted.df.1["post_flu"][,1], weighted.df.1["sbw_weights"][,1])
> mean.0 <- weighted.mean(weighted.df.0["post_flu"][,1], weighted.df.0["sbw_weights"][,1])

```

We omit the code to compute standard errors and confidence intervals via bootstrapping. In addition to supplying the tolerances manually, one could also implement the algorithm that selects the tolerances from a grid of options in a data-adaptive way. Code to implement this method appears below.

```

> t_ind <- "black_dr"
> dat.1 <- subset(dat, dat[t_ind][,1] == 1)
> dat.0 <- subset(dat, dat[t_ind][,1] == 0)
> bal_cov <- T1.vars.imp
> bal_alg <- TRUE
> bal_std <- "manual"
> bal_gri <- c(0.0001, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1)
> bal <- list(bal_cov = bal_cov, bal_alg = bal_alg, bal_tol = NULL, bal_std = bal_std,
bal_gri = bal_gri)
> wei <- list(wei_sum = TRUE, wei_pos = TRUE)
> par_tar <- colMeans(oakland.selection[T1.vars.imp])

```

```

> sbw.results.1 <- sbw(dat = dat.1, bal = bal, par = list(par_est = "aux", par_tar
= par_tar), sol = list(sol_nam = "gurobi"),wei = wei)
> sbw.results.0 <- sbw.results.1 <- sbw(dat = dat.0, bal = bal, par = list(par_est
= "aux", par_tar = par_tar), sol = list(sol_nam = "gurobi"),wei = wei)
> weighted.df.1 <- sbw.results.1$dat_weights
> weighted.df.0 <- sbw.results.0$dat_weights
> weighted.df <- rbind(weighted.df.1, weighted.df.0)

```