Supplementary material

S1 Connection to importance sampling

We briefly comment on an interesting parallel between the proposed EBWs for distributional balance and the notion of importance weights for importance sampling. The close link between propensity scores and importance sampling has been well-understood in the literature and this connection has been exploited to help make improvements in weighting methods for causal inference by utilizing variance reduction and stabilization approaches from the importance sampling literature. For an example of this literature and the deep connections between inverse probability weighting and importance sampling, see Datta and Polson [1]. Importance sampling (see, e.g., [2]) is a general technique for estimating integral quantities from a desired distribution G, using samples of another distribution H. The idea is to reweigh each sample from H by its importance weight dG/dH: the Radon-Nikodym derivative (or likelihood ratio) of G with respect to H. Clearly, such weights perfectly balance the sample from H to the desired distribution G. For distributional balance, H is the covariate distribution for the treated or control case (which we have access to), and G is the covariate distribution for the full population (which we wish to infer).

This link to importance sampling reveals two insights on the proposed distributional balance approach via EBWs. First, in order for importance weights (which are the Radon-Nikodym derivatives) to exist here, the population distribution must be absolutely continuous with respect to the treated and control distributions, which requires $f_{\mathbf{X}|A=0}(\mathbf{x}) > 0$ whenever $f(\mathbf{x}) > 0$ and $f_{\mathbf{X}|A=1}(\mathbf{x}) > 0$ whenever $f(\mathbf{x}) > 0$ for almost all $\mathbf{x} \in \mathcal{X}$, where f are (conditional) densities of the covariates. But this condition is satisfied by the positivity (or probabilistic assignment) assumption in Section 2, which requires the propensity score $\pi(\mathbf{x})$ to satisfy $0 < \pi(\mathbf{x}) < 1$. Hence, similar conditions are needed for distributional balance in both importance sampling and causal analysis. Second, it is known that under mild conditions, integral estimates under importance sampling are root-n consistent if the underlying samples from H are i.i.d. sampled [3]. The proof of Theorem 3.4 makes use of such results on importance weighting to establish the root-n consistency of the EBW estimator.

A key difference between importance weights and EBWs is that the former depends on both the population covariate density $f(\mathbf{x})$ and the propensity score $\pi(\mathbf{x})$, both of which are unknown in practice. The proposed method offers a *nonparametric* way for estimating distribution-balancing weights. It optimizes weights in the new weighted energy distance in Section 2, thereby balancing to the desired target distribution F (of which F_n is assumed to be a representative sample). When a distribution other than F is of interest (see Li et al. [4] for examples of other common target distributions), this importance sampling perspective of EBWs allows for a straight-forward modification of the criterion in Section 3 to balance to the target distribution.

S2 Technical proofs

S2.1 Proposition 2.1

Proof. For simplicity, we focus on the case where $p = 1$, but the arguments carry through for all dimensions. We further focus on the treated group (i.e. $a = 1$) without loss of generality. We begin by noting that we can express $|\varphi_n(t) - \varphi_{n,1,\mathbf{w}}(t)|^2$ in terms of $\varphi_n(t)\overline{\varphi_n(t)}$, $\varphi_n(t)\overline{\varphi_{n,1,\mathbf{w}}(t)}$, $\varphi_{n,1,\mathbf{w}}(t)\overline{\varphi_n(t)}$, and $\varphi_{n,1,\mathbf{w}}(t)\overline{\varphi_{n,1}(t)}$, where $\overline{\varphi_n(t)}$ and $\overline{\varphi_{n,1,\mathbf{w}}(t)}$ are the complex conjugates of $\varphi_n(t)$ and $\varphi_{n,1,\mathbf{w}}(t)$, respectively. For the first, we have

$$\varphi_n(t)\overline{\varphi_n(t)} = \frac{1}{n^2}\sum_{i,j}\exp\{it(X_i - X_j)\}$$
$$= \frac{1}{n^2}\sum_{i,j}\cos\{t(X_i - X_j)\} + V_1,$$

where V_1 is a term that vanishes when the integral in (8) of the main text is evaluated. Similarly, we have

$$\varphi_{n,1,\mathbf{w}}(t)\overline{\varphi_{n,1,\mathbf{w}}(t)} = \frac{1}{n_1^2}\sum_{i,j} w_i w_j A_i A_j \cos\{t(X_i - X_j)\} + V_2 \text{ and}$$
$$\varphi_{n,1,\mathbf{w}}(t)\overline{\varphi_n(t)} + \varphi_n(t)\overline{\varphi_{n,1,\mathbf{w}}(t)} = \frac{1}{n_1 n}\sum_{i,j} w_i A_i \cos\{t(X_i - X_j)\} + \frac{1}{n_1 n}\sum_{i,j} w_j A_j \cos\{t(X_i - X_j)\} + V_3.$$

Then combining terms, adding and subtracting 1 twice, by the constraints that the weights sum to n_a for $a \in \{0, 1\}$, and by Lemma 1 of [5], we have the desired result. □

S2.2 Theorem 2.2

Proof. Let $\{\widetilde{\mathbf{X}}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} F_{n,a,\mathbf{w}_n}$ and let $\widetilde{F}_{n,a,\mathbf{w}_n}$ and $\widetilde{\varphi}_{n,a,\mathbf{w}_n}$ be the empirical cdf and characteristic function of $\{\widetilde{\mathbf{X}}_i\}_{i=1}^{n_a}$. By the Glivenko-Cantelli theorem for non-identically distributed random variables (Theorem 1 of Wellner [6]), we have that $\lim_{n\to\infty}\sup_{\mathbf{x}\in\mathcal{X}}|\widetilde{F}_{n,a,\mathbf{w}_n}(\mathbf{x}) - \widetilde{F}_a(\mathbf{x})| = 0$. Similar to the proof of Theorem 2 in Székely et al. [7] (with modification, since now we need a SLLN for V-statistics of triangular arrays like Csörgö and Nasari [8], Patterson [9]), we will show that

$$\lim_{n\to\infty}\mathcal{E}(\widetilde{F}_{n,a,\mathbf{w}_n}, F_n) = \mathcal{E}(\widetilde{F}_a, F) \tag{S1}$$

almost surely. Similar to Székely et al. [7] define $D(\delta) = \{\mathbf{t} \in \mathbb{R}^p : \delta \le |\mathbf{t}|_p \le 1/\delta\}$ and $\mathcal{E}_\delta(\widetilde{F}_{n,a,\mathbf{w}_n} F_n) = \int_{D(\delta)}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathbf{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})d\mathbf{t}$. By the strong law of large numbers for V-statistics of triangular arrays [8,9], we have that the following holds almost surely

$$\lim_{n\to\infty}\mathcal{E}_\delta(\widetilde{F}_{n,a,\mathbf{w}_n}, F_n) = \mathcal{E}_\delta(\widetilde{F}_a, F_n) = \int_{D(\delta)}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_a(\mathbf{t})|^2\omega(\mathbf{t})d\mathbf{t}.$$

We note that $\lim_{\delta\to 0}\mathcal{E}_\delta(\widetilde{F}_a, F_n) = \mathcal{E}(\widetilde{F}_a, F_n)$, thus to verify (S1), we must show that

$$\limsup_{\delta\to 0}\ \limsup_{n\to\infty}|\mathcal{E}_\delta(\widetilde{F}_{n,a,\mathbf{w}_n}, F_n) - \mathcal{E}(\widetilde{F}_{n,a,\mathbf{w}_n}, F_n)| = 0. \tag{S2}$$

For each $\delta > 0$ we have

$$|\mathcal{E}_\delta(\widetilde{F}_{n,a,\mathbf{w}_n}, F_n) - \mathcal{E}(\widetilde{F}_{n,a,\mathbf{w}_n}, F_n)| \le \int_{|\mathbf{t}|_p<\delta}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathbf{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})d\mathbf{t} + \int_{|\mathbf{t}|_p>1/\delta}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathbf{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})d\mathbf{t}$$

Note that

$$|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathbf{w}_n}(\mathbf{t})|^2 = \left|\frac{1}{n}\sum_{i=1}^n \exp\{i\langle\mathbf{t}, \mathbf{X}_i\rangle\} - \frac{1}{n}\sum_{i=1}^n \exp\{i\langle\mathbf{t}, \widetilde{\mathbf{X}}_i\rangle\}\right|^2$$
$$= \left|\frac{1}{n}\sum_{i=1}^n(1 - \exp\{i\langle\mathbf{t}, \widetilde{\mathbf{X}}_i\rangle\}) - \frac{1}{n}\sum_{i=1}^n(1 - \exp\{i\langle\mathbf{t}, \mathbf{X}_i\rangle\})\right|^2$$
$$\le \frac{1}{n}\sum_{i=1}^n|1 - \exp\{i\langle\mathbf{t}, \widetilde{\mathbf{X}}_i\rangle\}|^2 + \frac{1}{n}\sum_{i=1}^n|1 - \exp\{i\langle\mathbf{t}, \mathbf{X}_i\rangle\}|^2.$$

Thus,

$$\int_{|\mathbf{t}|_p<\delta} |\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t} \le \frac{1}{n}\sum_{i=1}^{n}\int_{|\mathbf{t}|_p<\delta} |1-\exp\{i\langle\mathbf{t},\widetilde{\mathbf{X}}_i\rangle\}|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t} + \frac{1}{n}\sum_{i=1}^{n}\int_{|\mathbf{t}|_p<\delta} |1-\exp\{i\langle\mathbf{t},\mathbf{X}_i\rangle\}|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t}.$$

Similar to the arguments in the proof of Theorem 2 of Székely et al. [7], we have that $\int_{|\mathbf{t}|_p<\delta}|1-\exp\{i\langle\mathbf{t},\widetilde{\mathbf{X}}_i\rangle\}|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t} = |\widetilde{\mathbf{X}}_i|G(\widetilde{\mathbf{X}}_i\delta)$, where $G(y) = \int_{|\mathbf{t}|_p<y}\frac{1-\cos(t_1)}{|\mathbf{t}|^{1+p}}\mathrm{d}\mathbf{t}$ where t_1 is the first element of $\mathbf{t}$. Note that $\lim_{y\to 0}G(y) = 0$ and $G(y)$ is bounded. Thus, by the strong law of large numbers, $\limsup_{n\to\infty}\int_{|\mathbf{t}|_p<\delta}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t} \le \mathbb{E}\{|\widetilde{\mathbf{X}}|G(|\widetilde{\mathbf{X}}|\delta)\} + \mathbb{E}\{|\mathbf{X}|G(|\mathbf{X}|\delta)\}$. Thus, by the Lebesgue bounded convergence theorem for integrals and expectations, we have

$$\limsup_{\delta\to 0}\limsup_{n\to\infty}\int_{|\mathbf{t}|_p<\delta}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t} = 0.$$

By similar arguments, we have

$$\limsup_{\delta\to 0}\limsup_{n\to\infty}\int_{|\mathbf{t}|_p>1/\delta}|\varphi_n(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t} = 0.$$

Thus, we have shown (S1).

Then to complete the proof it remains to show that

$$\limsup_{n\to\infty}|\mathcal{E}(\widetilde{F}_{n,a,\mathrm{w}_n}F_n) - \mathcal{E}(F_{n,a,\mathrm{w}_n}F_n)| = 0. \tag{S3}$$

We denote $\mathrm{d}\omega = \omega(\mathbf{t})\mathrm{d}\mathbf{t}$. We begin by decomposing the above as

$$\begin{aligned}
&|\mathcal{E}(\widetilde{F}_{n,a,\mathrm{w}_n},F_n) - \mathcal{E}(F_{n,a,\mathrm{w}_n},F_n)| \\
&= \left|\int_{\mathbb{R}^p}\{2\varphi_n(\mathbf{t})[\varphi_{n,a,\mathrm{w}_n}(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})] + \widetilde{\varphi}^2_{n,a,\mathrm{w}_n}(\mathbf{t}) - \varphi^2_{n,a,\mathrm{w}_n}(\mathbf{t})\}\mathrm{d}\omega\right| \\
&\le 2\int_{\mathbb{R}^p}|\varphi_n(\mathbf{t})|\{|\widetilde{\varphi}_a(\mathbf{t}) - \varphi_{n,a,\mathrm{w}_n}(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\}\mathrm{d}\omega \\
&\quad + \int_{\mathbb{R}^p}|\widetilde{\varphi}_a(\mathbf{t}) + \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\cdot|\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\mathrm{d}\omega \\
&\quad + \int_{\mathbb{R}^p}|\widetilde{\varphi}_a(\mathbf{t}) + \varphi_{n,a,\mathrm{w}_n}(\mathbf{t})|\cdot|\widetilde{\varphi}_a(\mathbf{t}) - \varphi_{n,a,\mathrm{w}_n}(\mathbf{t})|\mathrm{d}\omega
\end{aligned} \tag{S4}$$

$$\begin{aligned}
&\le \int_{\mathbb{R}^p}\{2|\varphi(\mathbf{t})| + 2|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})| + 2|\widetilde{\varphi}_a(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\}\cdot|\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\mathrm{d}\omega \\
&+\int_{\mathbb{R}^p}\{2|\varphi(\mathbf{t})| + 2|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})| + 2|\widetilde{\varphi}_a(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t}) - \varphi_{n,a,\mathrm{w}_n}(\mathbf{t})|\}\cdot|\widetilde{\varphi}_a(\mathbf{t}) - \varphi_{n,a,\mathrm{w}_n}(\mathbf{t})|\mathrm{d}\omega.
\end{aligned} \tag{S5}$$

Note that $\varphi(\mathbf{t})$ is integrable due to the continuity of $\mathbf{X}$ and that $|\varphi_{n,a,\mathrm{w}_n}(\mathbf{t})| \to |\varphi(\mathbf{t})|$ and $|\widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})| \to |\varphi(\mathbf{t})|$. This ensures that the limsup of the integral converges to 0, which we will need below.

Due to the almost sure convergence of $\varphi_{n,a,\mathrm{w}_n}$ and $\widetilde{\varphi}_{n,a,\mathrm{w}_n}$ to $\widetilde{\varphi}_a$, the terms inside the integrals (S4) and (S5) both converge almost surely to 0. We first investigate (S4) and note that

$$\begin{aligned}
0 &\le 2\{2|\varphi(\mathbf{t})| + 2|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})| + 2|\widetilde{\varphi}_a(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\}\cdot\{|\widetilde{\varphi}_a(\mathbf{t})| + |\widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\} \\
&\quad - \{2|\varphi(\mathbf{t})| + 2|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})| + 2|\widetilde{\varphi}_a(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|\}\cdot|\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|.
\end{aligned} \tag{S6}$$

Note that the first term in the right hand side of (S6) converges to $8\{|\varphi(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t})|\}|\widetilde{\varphi}_a(\mathbf{t})|$ almost surely. Define $g_n(\mathbf{t}) \equiv 2|\varphi(\mathbf{t})| + 2|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})| + 2|\widetilde{\varphi}_a(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathrm{w}_n}(\mathbf{t})|$ and its almost sure limit $g(\mathbf{t}) \equiv 2\{|\varphi(\mathbf{t})| + |\widetilde{\varphi}_a(\mathbf{t})|\}$. Then an application of Fatou's lemma to the right hand side of (S6) yields

$$4\int_{\mathbb{R}^p} g(\mathbf{t})|\widetilde{\varphi}_a(\mathbf{t})|\mathrm{d}\omega \leq \liminf_{n\to\infty}\left\{2\int_{\mathbb{R}^p} g_n(\mathbf{t})|\widetilde{\varphi}_{n,a,\mathbf{w}_n}(\mathbf{t})|\mathrm{d}\omega + 2\int_{\mathbb{R}^p} g_n(\mathbf{t})|\widetilde{\varphi}_a(\mathbf{t})|\mathrm{d}\omega - \int_{\mathbb{R}^p} g_n(\mathbf{t})|\widetilde{\varphi}_a(\mathbf{t}) - \widetilde{\varphi}_{n,a,\mathbf{w}_n}(\mathbf{t})|\mathrm{d}\omega\right\}.$$

Thus we have $\limsup_{n\to\infty}(4) = 0$. A similar argument holds for (S5), and thus we have shown (S3), which concludes the proof. □

S2.3 Theorem 3.1

Proof. Similar to Amaral et al. [10], we consider weights defined by the Radon-Nikodym derivative $h_a = f_{\mathbf{X}}/f_{\mathbf{X}|A=a}$ for $a \in \{0,1\}$, where $f_{\mathbf{X}}$ is the density of $\mathbf{X}$ for the full population and $f_{\mathbf{X}|A=a}$ is the density of $\mathbf{X}$ for the treated (or control) population. We then let $\mathbf{h}_a = \{h_a(\mathbf{X}_1), \ldots, h_a(\mathbf{X}_n)\}$ be the Radon-Nikodym derivatives corresponding to the sample. We then define $\hat{h}_a(\mathbf{X}_i) = h_a(\mathbf{X}_i)/(\frac{1}{n_a}\sum_{i=1}^n I(A_i = a)h_a(\mathbf{X}_i))$ and $\hat{\boldsymbol{h}}_a = (\hat{h}_a(\mathbf{X}_1), \ldots, \hat{h}_a(\mathbf{X}_n))$. By the SLLN, $F_{n,a,\hat{\boldsymbol{h}}_a}(\mathbf{x}) = \frac{1}{n_a}\sum_{i=1}^n \hat{h}_a(\mathbf{X}_i)I(A_i = a)I(\mathbf{X}_i \leq \mathbf{x})$ converges almost everywhere to $F(\mathbf{x})$ for every continuity point $\mathbf{x}$ [10,11] for $a \in \{0,1\}$. Thus, as in the proof of Theorem 2 in Mak and Joseph [12] by the Portmanteau and dominated convergence theorems, we have

$$\lim_{n\to\infty} \mathbb{E}[|\varphi(\mathbf{t}) - \varphi_{n,a,\hat{\boldsymbol{h}}_a}(\mathbf{t})|^2] = 0 \text{ for all } \mathbf{t} \text{ for } a \in \{0,1\}, \text{ and} \tag{S7}$$

$$\lim_{n\to\infty} \mathbb{E}[|\varphi(\mathbf{t}) - \varphi_{n,a,\hat{\boldsymbol{h}}_a}(\mathbf{t})|^2] = 0 \quad \text{for all } \mathbf{t} \text{ for } a \in \{0,1\} \tag{S8}$$

where $\varphi_{n,a,\hat{\boldsymbol{h}}_a}(\mathbf{t}) = \frac{1}{n_a}\sum_{i=1}^n \hat{h}_a(\mathbf{X}_i)I(A_i = a)\exp\{i,\langle\mathbf{t},\mathbf{X}_i\rangle\}$ is a Radon-Nikodym derivative weighted ECHF for treatment arm a. Denote the expected weighted energy between the treated group and the sample population as

$$\mathbb{E}[\mathcal{E}(F_{n,a,\hat{\boldsymbol{h}}_a}, F_n)] = \mathbb{E}\left[\int_{\mathbb{R}^p} |\varphi_n(\mathbf{t}) - \varphi_{n,a,\hat{\boldsymbol{h}}_a}(\mathbf{t})|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t}\right] \text{ for } a \in \{0,1\}.$$

Note that although F is the weighted average of two conditional distribution functions, i.e. $F(\mathbf{x}) = F_1(\mathbf{x})P_1 + F_0(\mathbf{x})P_0$, due to the Theorem 2.1 and Corollary 3.1 of Van Zuijlen [13], all standard convergence properties of F_n resulting from a mixture distribution such as F this still hold. Specifically, a Glivenko-Cantelli theorem for empirical CDFs based on a mixture distribution as this holds. Thus, by the same arguments as in Mak and Joseph [12], $\lim_{n\to\infty}\mathbb{E}[\mathcal{E}(F_{n,a,\hat{\boldsymbol{h}}_a}, F_n)] = 0$ for $a \in \{0,1\}$. Define $\varphi_{n,a,\mathbf{w}_n^e}(\mathbf{t}) = \frac{1}{n_1}\sum_{i=1}^n w_i^e I(A_i = a)\exp\{i,\langle\mathbf{t},\mathbf{x}_i\rangle\}$ to be the energy-weighted ECHF for treatment arm a. By the definition of $\mathbf{w}_n^e$,

$$
\begin{aligned}
&\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_{n,0,\mathbf{w}_n^e}(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} + \int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_{n,1,\mathbf{w}_n^e}(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \\
&\le \left(\left[\int_{\mathbb{R}^p} |\varphi_n(\mathbf{t}) - \varphi_{n,0,\mathbf{w}_n^e}(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} + \left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} \right)^2 \\
&\quad + \left(\left[\int_{\mathbb{R}^p} |\varphi_n(\mathbf{t}) - \varphi_{n,1,\mathbf{w}_n^e}(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} + \left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} \right)^2 \\
&= \left([\mathcal{E}(F_{n,0,\mathbf{w}_n^e}, F_n)]^{1/2} + \left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} \right)^2 \\
&\quad + \left([\mathcal{E}(F_{n,1,\mathbf{w}_n^e}, F_n)]^{1/2} + \left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} \right)^2 \\
&\le \left([\mathbb{E}[\mathcal{E}(F_{n,0,\hat{\boldsymbol{h}}_0}, F_n)]]^{1/2} + \left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} \right)^2 \\
&\quad + \left([\mathbb{E}[\mathcal{E}(F_{n,1,\hat{\boldsymbol{h}}_1}, F_n)]]^{1/2} + \left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t}) d\mathbf{t} \right]^{1/2} \right)^2,
\end{aligned}
$$

where the first inequality holds by the Minkowski inequality. Thus, $\lim_{n\to\infty}\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F) = \lim_{n\to\infty}\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F_n) = 0$ for $a \in \{0,1\}$ since $\lim_{n\to\infty}\int_{\mathbb{R}^p}|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2\omega(\mathbf{t})d\mathbf{t} = 0$ a.s. If we choose any subsequence $\{n_k\}_{k=1}^{\infty}$ of $\mathbb{N}_+$, we have the same property that $\lim_{k\to\infty}\mathcal{E}(F_{n_k,0,\mathbf{w}_{n_k}^e}, F_n) = 0$ for $a \in \{0,1\}$. By the Riesz-Fischer Theorem, a sequence of functions f_n which converge to f in L_2 has a subsequence f_{n_k} which converges almost everywhere to f, implying the existence of a subsubsequence $\{n_k'\}_{k=1}^{\infty} \subseteq \{n_k\}_{k=1}^{\infty}$ such that $\varphi_{n_k',a,\mathbf{w}_{n_k'}^e}(\mathbf{t})$ converges to $\varphi(\mathbf{t})$ almost everywhere as $k \to \infty$ for $a \in \{0,1\}$. Since (n_k) was chosen arbitrarily, $\lim_{n\to\infty}\varphi_{n,a,\mathbf{w}_n^e}(\mathbf{t}) = \varphi(\mathbf{t})$ almost everywhere. Thus the main convergence result of Theorem 3.1 holds. That $\lim_{n\to\infty}\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F) = \lim_{n\to\infty}\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F_n) = 0$ holds almost surely is a consequence of (11) of the main text and Theorem 2.2. □

S2.4 Corollary 3.2

Proof. From (3) of the main text, the bias of $\hat{\tau}_{\mathbf{w}_n^e}$ can be written as:

$$
\begin{aligned}
|\mathbb{E}[\hat{\tau}_{\mathbf{w}_n^e}] - \tau| &= \left| \int_{\mathbf{x}\in\mathcal{X}} \mu_1(\mathbf{x}) d[F - F_{n,1,\mathbf{w}_n^e}](\mathbf{x}) - \int_{\mathbf{x}\in\mathcal{X}} \mu_0(\mathbf{x}) d[F - F_{n,0,\mathbf{w}_n^e}](\mathbf{x}) \right| \\
&\le \left| \int_{\mathbf{x}\in\mathcal{X}} \mu_1(\mathbf{x}) d[F - F_{n,1,\mathbf{w}_n^e}](\mathbf{x}) \right| + \left| \int_{\mathbf{x}\in\mathcal{X}} \mu_0(\mathbf{x}) d[F - F_{n,0,\mathbf{w}_n^e}](\mathbf{x}) \right|.
\end{aligned} \tag{S9}
$$

By Theorem 3.1, we know that $F_{n,1,\mathbf{w}_n^e}(\mathbf{x})$, the *weighted* treatment covariate distribution, converges to F, the population covariate distribution. By the Portmanteau Theorem (Theorem 2.1, [14]), it follows that:

$$
\int_{\mathbf{x}\in\mathcal{X}} \mu_1(\mathbf{x}) dF_{n,1,\mathbf{w}_n^e}(\mathbf{x}) \overset{n\to\infty}{\to} \int_{\mathbf{x}\in\mathcal{X}} \mu_1(\mathbf{x}) dF(\mathbf{x}).
$$

An analogous argument yields a similar result for the control group:

$$\int_{\mathbf{x}\in\mathcal{X}} \mu_0(\mathbf{x})\mathrm{d}F_{n,0,\mathbf{w}_n^e}(\mathbf{x}) \overset{n\to\infty}{\to} \int_{\mathbf{x}\in\mathcal{X}} \mu_0(\mathbf{x})\mathrm{d}F(\mathbf{x}).$$

Hence, from (S9), we have $\lim_{n\to\infty}|\mathbb{E}[\hat{\tau}_{\mathbf{w}_n^e}] - \tau| = 0$, which proves the claim. □

S2.5 Lemma 3.3

This follows directly from Theorem 4 of [12].

S2.6 Theorem 3.4

The proof of Theorem 3.4 requires a few lemmas.

The first lemma shows that, under i.i.d. sampling of the covariates $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim F$, the expected energy distance between F_n (its empirical distribution) and F (the population distribution) converges at a rate of $O(1/n)$:

Lemma S2.1. *Suppose* $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} F$. *Then* $\mathbb{E}[\mathcal{E}(F, F_n)] = O(1/n)$.

Lemma S2.1 By Proposition 1 of [5], we have:

$$\mathcal{E}(F, F_n) = \int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t})\mathrm{d}\mathbf{t}.$$

Taking an expectation on both sides, it follows that:

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(F, F_n)] &= \mathbb{E}\left[\int_{\mathbb{R}^p} |\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2 \omega(\mathbf{t})\mathrm{d}\mathbf{t}\right] \\
&= \int_{\mathbb{R}^p} \mathbb{E}[|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2]\omega(\mathbf{t})\mathrm{d}\mathbf{t} \ \text{(Tonelli's theorem, since the integrand is non-negative)} \\
&= \int_{\mathbb{R}^p} \frac{\mathbb{V}[\mathrm{Re}(\phi_1(\mathbf{t}))] + \mathbb{V}[\mathrm{Im}(\phi_1(\mathbf{t}))]}{n} \omega(\mathbf{t})\mathrm{d}\mathbf{t} \ (\mathbb{E}[|\varphi(\mathbf{t}) - \varphi_n(\mathbf{t})|^2] \ \text{is a variance term, since } \mathbb{E}\varphi_n(\mathbf{t}) = \varphi(\mathbf{t})) \\
&= O\left(\frac{1}{n}\right),
\end{aligned}$$

where constant terms depend on F and p.

The second lemma shows that, under the additional causal assumptions of positivity and strong ignorability as well as mild distributional assumptions on F_0 and F_1, the same convergence rate of $O(1/n)$ holds for the energy distance between $F_{n,a,\mathbf{w}_n^e}$ (the *energy-weighted* distribution for the treated or control) and F_n (the empirical covariate distribution):

Lemma S2.2. *Assume that the causal assumptions of positivity and strong ignorability hold. Let* $\mathbf{w}_n^e$ *be the solution to the energy balancing objective* (10) *of the main text. Under assumption* (*A*4), *we have* $\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F_n) = O(1/n)$ *almost surely.*

Lemma S2.2 First consider the non-normalized Radon-Nikodym derivative weights $\mathbf{w}_n^{nnrn}$. We will first show that $\mathcal{E}(F_{n,a,\mathbf{w}_n^{nnrn}}, F_n)$ is simply a degenerate two-sample V-statistic to show its convergence rate. The weights $\mathbf{w}_n^{nnrn}$ are functions of $\mathbf{x}$ in the sense that $w_i^{nnrn} = w^{nnrn}(\mathbf{X}_i) = 1/\pi(A_i, \mathbf{X}_i)$, where

$\pi(a, \mathbf{x}) = \mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x})$. Then $\mathcal{E}(F_{n,a,\mathbf{w}_n^{nnrn}}, F_n)$ is a two-sample V-statistic with kernel $h(\mathbf{x}_i, \mathbf{x}_j; \mathbf{x}_\ell, \mathbf{x}_m) = w^{nnrn}(\mathbf{x}_i)\|\mathbf{x}_i - \mathbf{x}_\ell\|_2 + w^{nnrn}(\mathbf{x}_j)\|\mathbf{x}_j - \mathbf{x}_m\|_2 - w^{nnrn}(\mathbf{x}_i)w^{nnrn}(\mathbf{x}_j)\|\mathbf{x}_i - \mathbf{x}_j\|_2 - \|\mathbf{x}_\ell - \mathbf{x}_m\|_2$. Denote $\{\widetilde{\mathbf{X}}_1, \dots, \widetilde{\mathbf{X}}_{n_a}\} = \{\mathbf{X}_i : A_i = a\}$. Then $\mathcal{E}(F_{n,a,\mathbf{w}_n^{nnrn}}, F_n)$ can be written as the following V-statistic

$$\mathcal{E}(F_{n,a,\mathbf{w}_n^{nnrn}}, F_n) = \frac{1}{n^2 n_a^2} \sum_{i=1}^{n_a} \sum_{j=1}^{n_a} \sum_{\ell=1}^{n} \sum_{m=1}^{n} h(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{X}}_j; \mathbf{X}_\ell, \mathbf{X}_m).$$

From positivity and strong ignorability it can be shown that $\mathcal{E}(F_{n,a,\mathbf{w}_n^{nnrn}}, F_n)$ is first-order degenerate in the sense that $\mathbb{E}h(\widetilde{\mathbf{x}}, \widetilde{\mathbf{X}}_j; \mathbf{X}_\ell, \mathbf{x}) = 0$ for any $\widetilde{\mathbf{x}}$ and $\mathbf{x}$. Thus, if $\mathbb{E}h^2 < \infty$, then $\mathcal{E}(F_{n,a,\mathbf{w}_n^{nnrn}}, F_n) = O(n^{-1})$ by extensions of asymptotic results for one-sample V-statistics [15,16] to multi-sample V-statistics as in Rizzo [17]. Note that this also implies that $\mathcal{E}(F_{n,a,\mathbf{w}_n^{rn}}, F_n) = O(n^{-1})$, since $\mathbf{w}_n^{nnrn}$ and $\mathbf{w}_n^{rn}$ differ only by a normalizing constant such that $\frac{1}{n}\sum_{i=1}^n w_i^{rn} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n w_i^{nnrn}\right] = 1$. By the definition of $\mathbf{w}_n^e$, we have $\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F_n) \le \mathcal{E}(F_{n,a,\mathbf{w}_n^{rn}}, F_n)$ for each n, which proves the desired result $\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F_n) = O(n^{-1})$.

The next lemma shows that, under a mild regularity condition on the energy balancing weights, the sum of the squared weights is upper bounded by $O(n)$:

Lemma S2.3. *Let $\mathbf{w}_n^e = (w_{1,n}^e, \dots, w_{n,n}^e)$ be the solution to the energy balancing objective* (10) *of the main text. Under assumptions* (A4) *and* (A5), *we have almost surely that*:

$$\sum_{i:A_i=0} \frac{{w_{i,n}^e}^2}{n_0} \le B \quad \textit{and} \quad \sum_{i:A_i=1} \frac{{w_{i,n}^e}^2}{n_1} \le B$$

for all $n > n^*$ for some $n^* > 1$ and some constant $B > 0$ that does not depend on n.

Lemma S2.3. Note that $\mathcal{E}(F_{n,1,\widetilde{\mathbf{w}}_n^e}, F_n) = O(n^{-1})$ by Lemma S2.2. We consider for simplicity the univariate case $p = 1$ and only focus on the treated group, i.e. those with $A = 1$; however, the same results apply directly for $A = 0$. For clarity of presentation, we denote $w_i \equiv w_{i,n}^e$. By the weighted energy distance duality, we have

$$\begin{aligned}
\mathcal{E}(F_{n,1,\mathbf{w}_n^e}, F_n) &= \int_{\mathbb{R}} |\phi_n(t) - \phi_{n,1,\widetilde{\mathbf{w}}_n^e}(t)|^2 \omega(t)\mathrm{d}t \\
&= \frac{1}{n^2} \int_{\mathbb{R}} \left| \sum_{i=1}^n \left(1 - w_i A_i \frac{n}{n_1}\right) \exp(itX_i) \right|^2 \omega(t)\mathrm{d}t \qquad \text{(S10)} \\
&= \frac{1}{n^2} \int_{\mathbb{R}} \sum_{i=1}^n \sum_{j=1}^n \left\{ \left(1 - w_i A_i \frac{n}{n_1} - w_j A_j \frac{n}{n_1} + w_i w_j A_i A_j \frac{n^2}{n_1^2}\right) \exp(it(X_i + X_j)) \right\} \omega(t)\mathrm{d}t.
\end{aligned}$$

Suppose that the number of weights w_i that are "near" the maximum $Cn^{1/3}$ (i.e. are of the same order with respect to n) is of order $O(n^{1/3})$. Denote the index set of these observations as $\mathcal{I}_n \equiv \{i : w_i = O(n^{1/3})\}$, and note that this supposition implies $|\mathcal{I}_n| = O(n^{1/3})$. Further suppose the "worst case" scenario that $A_i = 1$ for all $i \in \mathcal{I}_n$, $\mathrm{Re}(\exp(it(X_i + X_j))) > 0$ and $\mathrm{Im}(\exp(it(X_i + X_j))) > 0$ for all i, j, and that $\mathrm{Re}(\exp(itX_i)) < 0$, $\mathrm{Im}(\exp(itX_i)) < 0$, $\mathrm{Re}(\exp(itX_j)) < 0$, and $\mathrm{Im}(\exp(itX_j)) < 0$, so that every term in the double sum inside the integral in (S10) is positive. Then the double sum in (S10) is larger than

$$\begin{aligned}
&\sum_{i\in\mathcal{I}_n} \sum_{j\in\mathcal{I}_n} \left\{ \left(1 - w_i A_i \frac{n}{n_1} - w_j A_j \frac{n}{n_1} + w_i w_j A_i A_j \frac{n^2}{n_1^2}\right) \exp(it(X_i + X_j)) \right\} \\
&= \sum_{i\in\mathcal{I}_n} \sum_{j\in\mathcal{I}_n} \{(1 + O(n^{1/3}) + O(n^{1/3}) + O(n^{2/3}))\exp(it(X_i + X_j))\} \\
&= \sum_{i\in\mathcal{I}_n} (O(n^{1/3}) + O(n^{2/3}) + O(n^{2/3}) + O(n)) \\
&= O(n^{4/3}),
\end{aligned}$$

which implies that $\mathcal{E}(F_{n,1,\mathbf{w}^e}, F_n) = O(n^{-2/3})$, which is a contradiction to Lemma S2.2. Thus, we cannot have $|\mathcal{I}_n|$ as large as $O(n^{1/3})$. Using a similar argument, one can then show that the maximum size $\mathcal{I}_n$ can take to avoid such a contradiction is $|\mathcal{I}_n| = O(n^{1/6})$.

Assume, therefore, the worst case scenario that $|\mathcal{I}_n| = O(n^{1/6})$. To study the behavior of $\sum_{i:A_i=1} w_{i,n}^2/n_1$, we consider the set $\mathcal{J}_n = \{i : i \notin \mathcal{I}_n, w_i = O(r(n))$ where $\lim_{n\to\infty} r(n) = \infty$ and $\lim_{n\to\infty} r(n)/n^{1/3} = 0\}$. Thus, if we define $\mathcal{K}_n = \{i : w_i = O(1)\}$, then $\{i : A_i = 1\} = \mathcal{I}_n \cup \mathcal{J}_n \cup \mathcal{K}_n$. We now seek to find how large $|\mathcal{J}_n|$ can be to avoid a contradiction like the above. Consider the cross terms of $\mathcal{J}_n$ and $\mathcal{I}_n$ in (S10), which are

$$\begin{aligned}
&\sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_n}\left\{\left(1 - w_iA_i\frac{n}{n_1} - w_jA_j\frac{n}{n_1} + w_iw_jA_iA_j\frac{n^2}{n_1^2}\right)\exp(it(X_i + X_j))\right\}\\
&= \sum_{i\in\mathcal{I}_n}\sum_{j\in\mathcal{J}_n}\{(1 + O(n^{1/3}) + O(r(n)) + O(r(n)n^{1/3}))\exp(it(X_i + X_j))\}\\
&= \sum_{j\in\mathcal{J}_n}(O(n^{1/6}) + O(n^{1/2}) + O(r(n)n^{1/6}) + O(r(n)n^{1/2}))\\
&= O(|\mathcal{J}_n|r(n)n^{1/2}).
\end{aligned}$$

Thus, to avoid a contradiction to Lemma S2.2, we need $O(|\mathcal{J}_n|r(n)n^{1/2}) = O(n)$, i.e., $|\mathcal{J}_n|r(n) = O(n^{1/2})$. With this, the sum $\sum_{i:A_i=1} w_{i,n}^2$ then becomes:

$$\begin{aligned}
\sum_{i:A_i=1} w_{i,n}^2 &= \sum_{i\in\mathcal{K}_n} w_{i,n}^2 + \sum_{i\in\mathcal{J}_n} w_{i,n}^2 + \sum_{i\in\mathcal{I}_n} w_{i,n}^2\\
&= \sum_{i\in\mathcal{K}_n} O(1) + \sum_{i\in\mathcal{J}_n} O(r^2(n)) + \sum_{i\in\mathcal{I}_n} O(n^{2/3})\\
&= O(n) + O(n^{5/6}) + O(|\mathcal{J}_n|r^2(n)) = O(n),
\end{aligned}$$

where the last equality holds since $|\mathcal{J}_n|r(n)$ is at most of order $O(n^{1/2})$, $\lim_{n\to\infty} r(n)/n^{1/2} = 0$, and $|\mathcal{K}_n| = O(n)$ because $n = |\mathcal{K}_n| + |\mathcal{J}_n| + |\mathcal{I}_n|$.

From this (and the symmetry of the argument for $A = 0$), it follows that

$$\sum_{i:A_i=0}\frac{w_{i,n}^2}{n_0} \le B \quad \text{and} \quad \sum_{i:A_i=1}\frac{w_{i,n}^2}{n_1} \le B$$

for all $n > n^*$ for some $n^* > 1$, which proves the lemma.

Proof of Theorem 3.4. With these lemmas in hand, we can now tackle the main theorem. Let us condition on both $\mathbf{X}$ and A. From (3)–(5) of the main text, we can rewrite the mean squared error of $\hat{\tau}_{\mathbf{w}_n^e}$ as:

$$\begin{aligned}
\mathbb{E}_{Y|\mathbf{X},A}[(\hat{\tau}_{\mathbf{w}_n^e} - \tau)^2] &= \mathbb{V}_{Y|\mathbf{X},A}\left[\frac{1}{n_0}\sum_{i:A_i=0} w_i^e\varepsilon_i\right] + \mathbb{V}_{Y|\mathbf{X},A}\left[\frac{1}{n_1}\sum_{i:A_i=1} w_i^e\varepsilon_i\right]\\
&\quad + \left(\int\mu_1(\mathbf{x})\mathrm{d}[F - F_{n,1,\mathbf{w}_n^e}](\mathbf{x}) - \int\mu_0(\mathbf{x})\mathrm{d}[F - F_{n,0,\mathbf{w}_n^e}](\mathbf{x})\right)^2\\
&= \mathbb{V}_{Y|\mathbf{X},A}\left[\frac{1}{n_0}\sum_{i:A_i=0} w_i^e\varepsilon_i\right] + \mathbb{V}_{Y|\mathbf{X},A}\left[\frac{1}{n_1}\sum_{i:A_i=1} w_i^e\varepsilon_i\right]\\
&\quad + \left(\int\mu_1(\mathbf{x})\mathrm{d}[F_n - F_{n,1,\mathbf{w}_n^e}](\mathbf{x}) - \int\mu_1(\mathbf{x})\mathrm{d}[F_n\right.\\
&\quad \left. - F](\mathbf{x}) - \int\mu_0(\mathbf{x})\mathrm{d}[F_n - F_{n,0,\mathbf{w}_n^e}](\mathbf{x}) + \int\mu_0(\mathbf{x})\mathrm{d}[F_n - F](\mathbf{x})\right)^2\\
&\le \underbrace{\sum_{a=0}^{1}\frac{1}{n_a^2}\sum_{i:A_i=a}(w_i^e)^2\sigma_a^2(\mathbf{X}_i)}_{①} + 4\underbrace{\sum_{a=0}^{1}\left(\int\mu_a(\mathbf{x})\mathrm{d}[F_n - F_{n,a,\mathbf{w}_n^e}](\mathbf{x})\right)^2}_{②}\\
&\quad + 4\underbrace{\sum_{a=0}^{1}\left(\int\mu_a(\mathbf{x})\mathrm{d}[F - F_n](\mathbf{x})\right)^2}_{③},
\end{aligned}$$

where the last step follows from the identity $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$.

Consider first the terms in ①. Since $\sigma_a^2(\mathbf{x})$ is assumed to be bounded over $\mathcal{X}$, define $\bar{\sigma}^2 \equiv \max_{a\in\{0,1\}}\{\sup_{\mathbf{x}\in\mathcal{X}}\sigma_a^2(\mathbf{x})\}$. We have:

$$\mathbb{E}_{\mathbf{X},A}\left[\sum_{a=0}^{1}\frac{1}{n_a^2}\sum_{i:A_i=a}(w_i^e)^2\sigma_a^2(\mathbf{X}_i)\right] \leq \bar{\sigma}^2\mathbb{E}_{\mathbf{X},A}\left[\sum_{a=0}^{1}\frac{1}{n_a^2}\sum_{i:A_i=a}(w_i^e)^2\right] \text{(Lemma (2.3))}$$
$$\leq B\bar{\sigma}^2\mathbb{E}_A[Z_0+Z_1],$$

where $Z_a = 1/n_a$ if $n_a > 0$ and 0 otherwise. Note that, for $a \in \{0,1\}$, $n_a \sim \text{Bin}(n, P_a)$, where $P_a = \mathbb{P}(A=a)$. It follows that:

$$\mathbb{E}_A[Z_a] \leq \mathbb{E}_A\left[\frac{2}{n_a+1}\right] = \frac{2(1-(1-P_a)^{n+1})}{P_a(n+1)} \leq \frac{2}{P_a(n+1)} = O\left(\frac{1}{n}\right).$$

From this, we get that $\mathbb{E}_{\mathbf{X},A}[①]$ is also $O(1/n)$.

Consider next the terms in ②. For each $a \in \{0,1\}$, we have:

$$\mathbb{E}_{\mathbf{X},A}\left[\left(\int\mu_a(\mathbf{x})\mathrm{d}[F_n - F_{n,a,\mathbf{w}_n^e}](\mathbf{x})\right)^2\right] \leq \mathbb{E}_{\mathbf{X},A}\left[\sup_{\zeta\in\mathcal{H}:\|\zeta\|_{\mathcal{H}}\leq\|\mu_a\|_{\mathcal{H}}}\left(\int\zeta(\mathbf{x})\mathrm{d}[F_n - F_{n,a,\mathbf{w}_n^e}](\mathbf{x})\right)^2\right] \text{(Lemma 3.3)}$$
$$\leq C\mathbb{E}_{\mathbf{X},A}[\mathcal{E}(F_{n,a,\mathbf{w}_n^e}, F_n)]$$
$$= O\left(\frac{1}{n}\right). \text{(Lemma 2.2)}$$

Finally, consider the terms in ③. Since $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} F$, for each $a \in \{0,1\}$, we have:

$$\mathbb{E}_{\mathbf{X},A}\left[\left(\int\mu_a(\mathbf{x})\mathrm{d}[F - F_n](\mathbf{x})\right)^2\right] = \frac{\text{Var}[\mu_a(\mathbf{X})]}{n} = O\left(\frac{1}{n}\right).$$

Using the above bounds on ①, ②and ③, the desired claim is proven:

$$\mathbb{E}_{Y,\mathbf{X},A}[(\hat{\tau}_{\mathbf{w}_n^e} - \tau)^2] = \mathbb{E}_{\mathbf{X},A}\mathbb{E}_{Y|\mathbf{X},A}[(\hat{\tau}_{\mathbf{w}_n^e} - \tau)^2] = O\left(\frac{1}{n}\right). \square$$

S3 Theoretical results for penalized EBWs

In this section we consider a penalized version of the EBWs which is obtained as

$$\mathbf{w}_n^{ep} \in \operatorname{argmin}_{\mathbf{w}=(w_1,\ldots,w_n)}\left\{\mathcal{E}(F_{n,1,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,0,\mathbf{w}}, F_n) + \frac{\lambda}{n^2}\sum_{i=1}^{n}w_i^2\right\}$$
$$\text{s.t. } \sum_{i=1}^{n}w_iA_i = n_1, \sum_{i=1}^{n}w_i(1-A_i) = n_0, w_i \geq 0 \text{ for } i = 1, \ldots, n, \tag{S11}$$

where $\lambda > 0$ is a fixed constant. Similarly to the unpenalized EBWs, we have

Theorem S3.1. *Assume that* $\mathbb{E}(\|\mathbf{X}\|_2 \mid A=a) < \infty, \mathbb{E}\|\mathbf{X}\|_2 < \infty, \mathbb{E}[\pi_a^{-2}(X)] < \infty$ *for* $a \in \{0,1\}$, *and that the assumptions presented in Section* 2.1 *hold. Let* $\mathbf{w}_n^{ep}$ *be as defined in* (S11). *Then, for* $a \in \{0,1\}$,

$$\lim_{n\to\infty} F_{n,a,\mathbf{w}_n^{ep}}(\mathbf{x}) \equiv \lim_{n\to\infty}\frac{1}{n_a}\sum_{i=1}^{n}w_i^{ep}I(\mathbf{X}_i \leq \mathbf{x}, A_i = a) = F(\mathbf{x}) \tag{S12}$$

almost surely for every continuity point $\mathbf{x} \in \mathcal{X}$. *Furthermore,*

$$\lim_{n\to\infty} \mathcal{E}(F_{n,a,\mathbf{w}_n^{ep}}, F_n) = 0$$

holds almost surely.

Theorem S3.1. The proof follows the same arc as the proof of Theorem 3.1, however the key inequality in the proof of Theorem 3.1 is $\mathcal{E}(F_{n,1,\mathbf{w}_n^e}, F_n) \leq \mathbb{E}[\mathcal{E}(F_{n,a,\hat{\boldsymbol{h}}_a}, F_n)]$, whereas for the penalized weights this inequality is not guaranteed to hold. Instead, we have $\mathcal{E}(F_{n,1,\mathbf{w}_n^{ep}}, F_n) + \frac{\lambda}{n^2}\sum_{i=1}^n {w^{ep}_i}^2 \leq \mathbb{E}[\mathcal{E}(F_{n,a,\hat{\boldsymbol{h}}_a}, F_n)] + \frac{\lambda}{n^2}\sum_{i=1}^n \hat{h}_a(\mathbf{X}_i)^2$. Since $\mathbb{E}[\pi_a(X)^{-2}] < \infty$ for $a \in \{0,1\}$, by the SLLN, $\frac{\lambda}{n}\sum_{i=1}^n \hat{h}_a(\mathbf{X}_i)^2$ converges a.s. to a constant and thus $\frac{\lambda}{n^2}\sum_{i=1}^n \hat{h}_a(\mathbf{X}_i)^2$ converges to 0 a.s. The rest of the proof follows in the same manner as the proof of Theorem 3.1.

We also have the following result regarding the root-n consistency of the resulting weighted average estimate of the ATE. With this, we now state the result on root-n consistency:

Theorem S3.2. *Assume the same conditions in Theorem* S3.1. *Let $\mathcal{H}$ be the native space induced by the radial kernel $\Phi(\cdot) = -\|\cdot\|_2$ on $\mathcal{X}$. Suppose the following mild conditions hold*:

(A1) $\mu_0(\cdot) \in \mathcal{H}$ *and* $\mu_1(\cdot) \in \mathcal{H}$,

(A2) $\mathrm{Var}[\mu_0(\mathbf{X})] < \infty$ *and* $\mathrm{Var}[\mu_1(\mathbf{X})] < \infty$,

(A3) $\sigma_0^2(\mathbf{x})$ and $\sigma_1^2(\mathbf{x})$ *are bounded over* $\mathbf{x} \in \mathcal{X}$,

(A4) $\mathbb{E}[h_0^2(\mathbf{X}, \mathbf{X}', \mathbf{X}'', \mathbf{X}''')] < \infty$ *and* $\mathbb{E}[h_1^2(\mathbf{X}, \mathbf{X}', \mathbf{X}'', \mathbf{X}''')] < \infty$, *where* $\mathbf{X}, \mathbf{X}', \mathbf{X}'', \mathbf{X}''' \overset{i.i.d.}{\sim} F$ *and, with* $\pi_0(\mathbf{x}) \coloneqq 1 - \pi(\mathbf{x})$ *and* $\pi_1(\mathbf{x}) \coloneqq \pi(\mathbf{x})$, *the kernel h_a is defined for $a = 0, 1$ as*:

$$h_a(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{a}) = \frac{1}{\pi_a(\mathbf{x})}\|\mathbf{x} - \mathbf{z}\|_2 + \frac{1}{\pi_a(\mathbf{y})}\|\mathbf{y} - \mathbf{a}\|_2 - \frac{1}{\pi_a(\mathbf{x})\pi_a(\mathbf{y})}\|\mathbf{x} - \mathbf{y}\|_2 - \|\mathbf{a} - \mathbf{z}\|, \tag{S13}$$

Then the proposed EBW estimator $\hat{\tau}_{\mathbf{w}_n^{ep}}$ is root-n consistent, i.e.:

$$\sqrt{\mathbb{E}_{\mathbf{X},A,Y}[(\hat{\tau}_{\mathbf{w}_n^{ep}} - \tau)^2]} = O(n^{-1/2}). \tag{S14}$$

Theorem S3.2. The proof follows the same structure as the proof of Theorem 3.4, however the key difference is in how the term ① from the proof of Theorem 3.4 is handled. We note that since $\mathcal{E}(F_{n,1,\mathbf{w}_n^{ep}}, F_n) + \frac{\lambda}{n^2}\sum_{i=1}^n {w^{ep}_i}^2 \leq \mathbb{E}[\mathcal{E}(F_{n,a,\hat{\boldsymbol{h}}_a}, F_n)] + \frac{\lambda}{n^2}\sum_{i=1}^n \hat{h}_a(\mathbf{X}_i)^2$and since both terms on the right of the inequality converge to 0 a.s. and are both $O\left(\frac{1}{n}\right)$, then since both terms on the left are strictly positive, they are necessarily $O\left(\frac{1}{n}\right)$. As in the proof of Theorem 3.4, we have

$$\mathbb{E}_{\mathbf{X},A}\left[\sum_{a=0}^{1} \frac{1}{n_a^2} \sum_{i:A_i=a} (w_i^e)^2 \sigma_a^2(\mathbf{X}_i)\right] \leq \bar{\sigma}^2 \mathbb{E}_{\mathbf{X},A}\left[\sum_{a=0}^{1} \frac{1}{n_a^2} \sum_{i:A_i=a} (w_i^e)^2\right].$$

Since the term on the right of the inequality is $O\left(\frac{1}{n}\right)$, the remainder of the proof follows as in the proof of Theorem 3.4.

S4 Additional simulation results

S4.1 Additional details for simulationst

In this section we provide specific details of all of the propensity score models and outcome models used in the simulations in Section 4.2 of the main text. The propensity score models are described in Table S1. The average proportion of those treated in propensity models I, II, III, IV, and V are 0.35, 0.31, 0.50, 0.51, and 0.51, respectively. The conditional mean functions of the outcome given the covariates and treatment for outcome models (A–E) are provided in Table S2.

S4.2 Detailed simulation results

Table S3 contains a summary of the results averaged across propensity models (I-VI) and dimension settings ($p \in \{10, 25\}$). Each entry in the table is the average rank of each method in terms of RMSE and bias for each combination of outcome model and dimension; i.e. the method with the smallest RMSE for a particular setting receives a “1” and the method with the largest RMSE receives a “7.”

S4.3 Details for weighted energy distance toy examples

In this section we outline the details for the toy examples in Section 2.3 of the main text. In the first example, we generate a 1-dimensional covariate of sample size 250, which impacts treatment assignment for a binary via a logistic model under three scenarios: (1) $\text{logit}(\pi(X)) = -1 + X$, (2) $\text{logit}(\pi(X)) = -1 + X + 2X^2/3$, and (3) $\text{logit}(\pi(X)) = -1 + X + 2X^2/3 - X^3/3$. In each scenario, the response is generated as $Y = X + X^3 - 1/(0.1 + 0.1X^2) + \varepsilon$, where $\varepsilon \sim N(0, \sqrt{2})$. For each scenario, we construct inverse probability weights based off of 3 logistic regression models, which consider only a linear term in X (denoted as “IPW (1)”), a linear plus quadratic term (denoted as “IPW (2)”), and up to the cubic term (denoted as “IPW (3)”), respectively. For each set of weights $\mathbf{w}$, we compute the sum of the energy distances between each treatment group and the combined sample, i.e $\mathcal{E}(F_{n,0,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,1,\mathbf{w}}, F_n)$ and compute the bias of (2) for for τ using each set of weights (Tables S4 and S5).

In a second toy example, we consider a two dimensional example where the true assignment mechanism depends on first and second moments of the covariates. In particular, we generate treatment assignments from $\text{logit}(\pi(X)) = -1 + X_1 + 0.5X_1^2 - X_2 - 0.5X_2^2$. The response is generated as $Y = X_1 - 1/(0.1 + 0.1X_1^2) - X_2 + 1/(0.1 + 0.1X_2^2) + \varepsilon$. We consider a collection of methods to estimate weights, including logistic regression, the method of Imai and Ratkovic [18], and the method of Chan et al. [19], each with (i) just first order moments included for balancing or estimation and additionally (ii) all first and second order moments included. The

Table S1: Propensity models used in the simulation studies. The average proportion of those treated in propensity models I, II, III, IV, and V are 0.35, 0.31, 0.50, 0.51, and 0.51, respectively

Model	$\eta = \text{logit}\{\mathbb{P}(A = 1\vert\mathbf{X})\}=$
I	$2X_1X_2I(\vert X_1\vert > 1, \vert X_2\vert > 1) + 2X_2X_3I(\vert X_2\vert < 1, \vert X_3\vert < 1)$
	$+2X_3X_4I(\vert X_3\vert > 1, \vert X_4\vert > 1) + 2X_4X_1I(\vert X_1\vert < 1, \vert X_4\vert < 1)$
	$+I(\vert X_1\vert > 0.5, \vert X_2\vert > 0.5, \vert X_3\vert > 0.5, \vert X_4\vert > 0.5)$
	$+I(\vert X_1\vert < 0.25, \vert X_2\vert > 0.25, \vert X_3\vert < 0.25, \vert X_4\vert > 0.25)$
II	$-2 + \log\vert X_1-X_2\vert - \log\vert X_2-X_3\vert + \vert(X_3-X_4)X_1X_2\vert^{1/2}$
III	$-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4 - X_5 + 0.5X_6 - 0.25X_7 - 0.1X_8$
IV	$c\sum_{i=1}^{3}\sum_{j=i}^{4}(-1)^{2j-i}X_iX_j$, where c is chosen such that $\text{SD}(\eta) = 5$
V	$-2 + 2X_1X_2 + (X_1-X_2)^2 - 2X_3X_4 - (X_3 + X_5)^2$
VI	$\vert X_1-2X_2\vert\cdot\vert X_2-2X_3\vert - \vert X_3-2X_4\vert\cdot\vert X_4-2X_5\vert + X_6 - 0.5X_7 - 0.25X_8$

Table S2: The coefficients in Model D above are $\boldsymbol{\beta} = (0.8, 0.25, 0.6, -0.4, -0.8, -0.5, 0.7)$

Model	$\boldsymbol{\mu} = \mathbb{E}[Y \vert \mathbf{X}, A]=$
A	$210 + 27.4\vert X_1\vert + 13.7\vert X_2\vert + 13.7\vert X_3\vert + 13.7\vert X_4\vert$
B	$X_1X_2^3X_3^2X_4 + X_4\vert X_1\vert^{1/2}$
C	$2\sum_{j=1}^{4}(1-X_jI(X_j > 0)A)\cdot(X_j-2X_{j+1})$
D	$\sum_{j=1}^{7}X_j\beta_j + \beta_2X_2^2 + \beta_4X_4^2 + \beta_7X_7^2 + 0.5\beta_1X_1X_3 + 0.7\beta_2X_2X_4 + 0.7\beta + 2X_2X_4 + 0.5\beta_3X_3X_5$
	$+0.7\beta_4X_4X_6 + 0.5\beta_5X_5X_7 + 0.5\beta_1X_1X_6 + 0.7\beta_2X_2X_3 + 0.5\beta_3X_3X_4 + 0.5\beta_4X_4X_5 + 0.5\beta_5X_5X_6$
E	$210 + (1.5A-0.5)(27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4)$

Table S3: Displayed are the ranks among all methods tested of each method in terms of RMSE and bias averaged over all response models (A–E) for $n = 250$ and over the dimension settings $p = 10$ and $p = 25$

Y Model:	**A**		**B**		**C**		**D**		**E**	
	Mean rank		**Mean rank**		**Mean rank**		**Mean rank**		**Mean rank**	
Method	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
Unweighted	5.1	4.8	6.2	4.8	4.8	4.7	6.0	4.8	3.8	**2.2**
EBW	2.3	2.5	3.4	4.0	**2.0**	**2.2**	3.2	3.5	**1.5**	4.2
iEBW	**1.1**	**1.3**	**1.8**	**2.8**	2.4	2.9	**1.8**	**2.4**	3.2	4.5
KCB	2.6	2.9	2.9	3.7	2.7	2.7	2.8	3.0	3.2	3.2
IPW	6.4	6.1	5.0	4.2	6.3	5.7	5.2	5.0	6.9	4.3
CBPS	5.3	5.0	4.8	4.0	5.2	5.4	5.3	4.6	4.8	3.7
Cal	5.2	5.3	3.8	4.7	4.6	4.5	3.8	4.7	4.7	5.8

weights of all methods are then used for weighted estimates of τ. We then compare the weighted energy distances and absolute biases of (2) based on these weights in Figure 1(b) of the main text.

S4.4 Details for value function optimization toy example

In this section we detail the setup for the example involving estimation of individualized treatment rules (ITRs) via value function optimization. To demonstrate the effectiveness of using energy balancing weights in optimal ITR estimation, we provide an illustrative example under two data-generating scenarios. For both scenarios we generate outcomes as $Y = g(\mathbf{X}) + \widetilde{A}\Delta(\mathbf{X})/2 + \varepsilon$, where $g(\mathbf{X})$ are the main effects of $\mathbf{X}$, $\widetilde{A} = 2A - 1$, and $\Delta(\mathbf{X}) = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$ is the treatment-covariate interaction, $\varepsilon \sim N(0, 1)$, and $\mathbb{R}^{10} \ni \mathbf{X} \overset{i.i.d.}{\sim} \text{Unif(-1,1)}$. Both scenarios are motivated by the simulation studies of Zhao et al. [20] but generate A from a logistic regression model with terms depending on up to third order polynomials in a subset of the predictors and $g(\mathbf{X})$ contains non-linear terms in the predictors. Scenario 1 uses $g(\mathbf{X}) = 8 - \sum_{j=1}^{3}(-1)^j\{X_j + 10X_j^3 - 1/(0.1 + 0.1X_j^2)\}$, $\Delta(\mathbf{X}) = X_2 - 0.25X_1^2 - X_4 + 0.25X_3^2$, and $\text{logit}(\pi(\mathbf{X})) = -1 - \sum_{j=1}^{3}(-1)^j\{(7/4)X_j + (7/6)X_j^2 + (7/12)X_j^3\}$. Scenario 2 uses $g(\mathbf{X}) = 8 + 0.5(X_1 + 10X_1^3 - 1/(0.1 + 0.1X_1^2))$, $\Delta(\mathbf{X}) = -1 - X_1^3 + \exp(X_3^2 + X_5) + 0.6X_6 - (X_7 + X_8)^2$, and $\text{logit}(\pi(\mathbf{X})) = -1 + (7/4)X_1 + (7/6)X_1^2 + (7/12)X_1^3$. We utilize the OWL method to obtain estimates $\hat{d}$, which

Table S4: Displayed are results for n = 250 and p = 10 averaged over 1,000 independent simulated datasets. The average proportion of those treated in propensity models I, II, III, IV, V, and VI are 0.35, 0.31, 0.50, 0.51, 0.51, and 0.50, respectively

	Propensity model:	**I**		**II**		**III**		**IV**		**V**		**VI**	
	Method	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
Y Model: A	Unweighted	21.375	−21.110	23.362	22.893	3.851	0.032	7.772	6.752	35.122	34.983	12.423	11.786
	EBW	8.450	−8.173	6.908	6.714	2.133	−0.094	1.521	**0.200**	13.542	13.387	6.673	6.403
	iEBW	**5.239**	**5.010**	5.372	5.204	**1.790**	**0.005**	**1.338**	−0.496	**9.607**	**9.475**	**5.091**	**4.841**
	KCB	7.305	−6.380	**4.015**	**3.413**	3.025	0.789	1.463	−0.446	15.210	14.643	7.070	6.549
	IPW	21.462	−21.188	23.371	22.827	8.083	0.930	7.864	6.766	35.114	34.971	12.731	12.028
	CBPS	21.442	−21.174	23.194	22.688	4.324	0.612	7.822	6.759	35.083	34.942	12.655	11.994
	Cal	21.459	−21.184	23.129	22.597	3.567	0.744	7.854	6.764	35.033	34.892	12.791	12.112
Y Model: B	Unweighted	13.678	−0.377	30.446	0.861	29.859	−24.281	17.670	0.870	17.529	0.474	18.164	−0.855
	EBW	8.892	−0.263	9.596	0.520	28.707	−22.923	11.779	0.996	8.679	0.237	12.226	−0.717
	iEBW	**4.428**	−0.198	**5.740**	**0.419**	23.214	−18.442	**8.824**	0.790	4.077	0.109	**9.191**	−0.376
	KCB	8.965	−0.267	9.334	0.446	17.595	−14.202	9.744	**0.567**	8.945	0.278	9.809	**0.281**
	IPW	8.979	−0.238	12.352	0.817	**15.395**	**11.546**	17.420	0.972	8.962	0.265	72.360	−4.526
	CBPS	8.929	−0.241	10.603	0.934	23.757	−13.231	13.973	0.830	9.074	0.281	14.939	−0.602
	Cal	4.643	**0.191**	6.762	1.027	44.189	−31.440	18.345	1.138	**1.702**	**0.100**	14.375	0.580
Y Model: C	Unweighted	5.406	5.366	5.584	−5.203	1.239	0.179	2.493	−2.117	6.253	−6.099	1.381	−0.466
	EBW	3.666	3.606	1.243	−0.802	**1.385**	1.013	0.906	−0.195	**0.993**	**0.527**	**0.964**	**0.003**
	iEBW	3.784	3.727	**1.030**	**0.497**	1.991	1.787	**0.888**	0.254	1.280	1.107	1.103	0.588
	KCB	**3.313**	**3.145**	1.159	−0.548	1.732	1.329	0.894	**0.078**	1.224	−0.786	1.039	0.071
	IPW	5.406	5.376	5.649	−5.214	2.415	1.247	2.524	−2.148	6.239	−6.082	1.670	−1.042
	CBPS	5.410	5.381	5.530	−5.130	2.145	1.692	2.491	−2.123	6.215	−6.063	1.532	−0.850
	Cal	5.412	5.383	5.509	−5.087	1.387	**0.731**	2.486	−2.106	6.193	−6.040	1.638	−0.983
Y Model: D	Unweighted	1.193	−0.456	1.827	1.077	8.525	−8.456	1.226	−0.358	1.386	0.708	**1.401**	**0.724**
	EBW	0.543	**0.039**	0.575	0.192	3.627	−3.411	0.577	−0.336	0.563	−0.186	1.773	1.696
	iEBW	**0.450**	0.066	0.478	0.134	2.908	−2.693	0.487	−0.264	**0.491**	**0.180**	1.519	1.445

(Continued)

Table S4: *Continued*

	Propensity model:	I		II		III		IV		V		VI	
	Method	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
	KCB	0.530	−0.127	**0.467**	**0.128**	2.667	−2.223	**0.430**	**0.184**	0.615	0.209	1.611	1.511
	IPW	0.896	−0.505	1.457	1.096	3.144	−2.059	0.799	−0.359	0.986	0.680	2.325	2.201
	CBPS	0.978	−0.505	1.477	1.061	3.183	−2.829	0.893	−0.346	1.103	0.687	1.958	1.765
	Cal	0.880	−0.524	1.344	1.065	**2.096**	**1.378**	0.785	−0.354	0.985	0.687	2.303	2.195
Y Model: E	Unweighted	**2.463**	**0.004**	3.209	**0.039**	4.105	−3.469	**2.013**	−0.029	2.466	**0.011**	2.458	**0.103**
	EBW	2.764	0.811	**2.469**	−0.694	**2.517**	0.176	2.130	0.118	**2.300**	−0.079	**2.247**	0.462
	iEBW	2.868	0.975	2.504	−0.606	2.776	0.594	2.187	0.188	2.376	0.130	2.375	0.858
	KCB	2.846	0.757	2.493	−0.480	2.812	**0.013**	2.215	0.161	2.351	0.244	2.382	0.389
	IPW	3.112	0.395	3.454	−1.957	5.277	0.157	2.928	0.035	3.962	−1.895	4.606	0.267
	CBPS	2.817	0.358	3.002	−1.450	3.317	−0.517	2.209	**0.010**	3.393	−1.813	2.473	0.265
	Cal	3.786	1.279	2.811	−1.389	3.482	1.680	2.201	0.265	2.951	−0.620	2.336	0.384

Table S5: Displayed are results for n = 250 and p = 25 averaged over 1,000 independent simulated datasets

	Propensity model:	I		II		III		IV		V		VI	
	Method	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
Y Model: A	Unweighted	21.203	−20.951	23.159	22.709	3.845	**0.036**	7.455	6.419	34.913	34.764	12.252	11.627
	EBW	15.116	−14.852	14.214	13.906	2.789	−0.388	3.026	1.630	24.808	24.663	11.005	10.670
	iEBW	**11.547**	**11.300**	**12.374**	**12.093**	**2.466**	−0.201	**2.352**	**0.688**	**21.365**	**21.223**	**9.941**	**9.642**
	KCB	15.877	−15.534	14.928	14.271	3.036	−0.354	3.834	2.348	27.159	26.898	10.764	10.257
	IPW	21.305	−21.014	23.009	22.430	8.154	0.602	7.602	6.382	34.859	34.694	12.644	11.932
	CBPS	21.327	−21.056	22.883	22.385	3.969	0.301	7.538	6.409	34.830	34.675	12.483	11.829
	Cal	21.324	−21.031	22.776	22.183	3.504	0.607	7.596	6.379	34.829	34.669	12.675	11.967
Y Model: B	Unweighted	15.619	−0.082	32.211	0.375	30.937	−24.117	20.220	0.163	20.323	−0.013	18.792	−0.765
	EBW	9.217	**0.011**	9.470	**0.355**	29.218	−22.972	14.260	0.359	8.191	−0.199	14.823	−0.428
	iEBW	3.483	0.026	**5.725**	0.383	22.282	−17.571	11.303	0.389	2.315	−0.137	12.119	**0.175**
	KCB	10.093	−0.132	11.084	0.775	17.597	−14.152	**10.572**	0.353	10.082	0.187	**11.119**	−0.293
	IPW	9.927	−0.097	18.046	0.509	**17.506**	**11.581**	29.761	0.485	10.180	**0.010**	128.009	−8.725
	CBPS	10.142	−0.096	12.701	0.798	21.586	−14.104	16.355	**0.025**	12.731	0.085	19.754	−0.683
	Cal	**3.075**	−0.062	8.836	0.912	39.914	−27.307	23.360	0.428	**1.333**	0.054	19.856	0.428
Y Model: C	Unweighted	5.404	5.362	5.544	−5.171	**1.240**	**0.246**	2.458	−2.060	6.174	−6.012	1.367	−0.431
	EBW	**4.884**	**4.843**	2.687	−2.364	1.845	1.605	1.153	−0.602	2.456	−2.268	1.119	−0.366
	iEBW	4.914	4.876	**2.353**	**2.016**	2.730	2.601	**0.927**	**0.075**	**1.160**	**0.832**	**1.016**	0.147
	KCB	5.024	4.975	3.215	−2.649	2.249	2.045	1.349	−0.674	3.440	−3.143	1.144	**0.096**
	IPW	5.423	5.390	5.547	−5.059	2.680	1.433	2.457	−2.039	6.107	−5.936	1.663	−0.988
	CBPS	5.414	5.383	5.431	−5.035	2.311	1.942	2.447	−2.044	6.105	−5.945	1.508	−0.780
	Cal	5.438	5.405	5.427	−4.937	1.567	1.031	2.420	−1.997	6.085	−5.920	1.621	−0.922
Y Model: D	Unweighted	1.273	−0.515	1.761	1.057	8.495	−8.428	1.245	**0.307**	1.362	0.707	**1.410**	**0.782**
	EBW	0.692	−0.196	0.908	0.534	4.348	−4.185	0.818	−0.480	0.639	0.080	2.164	2.066
	iEBW	**0.602**	**0.108**	**0.817**	**0.465**	3.438	−3.263	**0.760**	−0.452	**0.588**	**0.002**	2.096	2.010
	KCB	0.881	−0.363	1.039	0.582	4.993	−4.799	0.863	−0.369	0.894	0.390	1.691	1.481

(Continued)

Table S5: *Continued*

	Propensity model:	I		II		III		IV		V		VI	
	Method	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
	IPW	0.981	−0.524	1.478	1.002	3.497	−2.240	0.879	−0.323	1.013	0.694	2.326	2.174
	CBPS	1.039	−0.540	1.442	1.027	3.866	−3.645	0.922	−0.318	1.075	0.682	1.945	1.740
	Cal	0.916	−0.543	1.338	0.996	**2.089**	**1.424**	0.832	−0.317	0.999	0.697	2.303	2.177
Y Model: E	Unweighted	**2.464**	**0.063**	3.106	**0.158**	4.077	−3.469	**1.968**	−0.061	2.343	**0.064**	2.407	**0.074**
	EBW	2.738	0.981	**2.662**	−0.687	**2.416**	−0.357	2.022	0.126	**2.311**	−0.185	**2.164**	0.435
	iEBW	2.751	0.958	2.715	−0.511	2.607	−0.246	2.074	0.167	2.362	−0.184	2.182	0.605
	KCB	2.605	0.262	2.814	−0.653	3.141	−1.728	2.038	**0.038**	2.344	−0.568	2.208	0.131
	IPW	3.648	0.436	4.151	−1.970	5.690	**0.102**	3.332	−0.070	4.362	−1.978	4.691	0.330
	CBPS	2.626	0.339	3.041	−1.159	3.260	−1.062	2.237	−0.044	3.121	−1.555	2.329	0.236
	Cal	3.377	1.038	2.949	−1.388	3.416	1.571	2.150	0.227	2.852	−0.833	2.310	0.408

Table S6: Displayed are the median, mean, standard deviation, and maximum RMSEs for each method across the 100 simulation settings using the RHC data

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
	Constant treatment effect					
Median RMSE	4.6293	1.6949	2.1071	1.3496	0.6848	0.5298
Mean RMSE	5.8204	2.0779	2.5354	1.7885	0.8336	0.6894
SD RMSE	4.2938	1.5037	1.8716	1.4943	0.6261	0.5247
Max RMSE	22.3084	6.7853	7.7406	5.2275	2.6797	2.3131
	Heterogeneous treatment effect					
Median RMSE	7.4944	3.7809	4.3439	2.4788	1.2268	1.0147
Mean RMSE	9.4230	3.9293	4.8919	3.2720	1.4285	1.2008
SD RMSE	6.9536	2.8326	3.6563	2.8277	1.0519	0.8788
Max RMSE	36.1208	12.5586	15.1127	9.6681	4.5404	3.7934

uses inverse weighting by the propensity score and adds $\lambda_n||d||^2$ to the objective. For OWL, the propensity score is misspecified to only include linear terms in the covariates. We also estimate d^* by minimizing (21) plus $\lambda_n||d||^2$. We denote this as OWL (EBW) for weights given by (10) and OWL (iEBW) for weights given by (16). We simulate 1,000 independent datasets and compute the average value function $\widehat{\mathbb{E}}[Y(\hat{d})]$ evaluated on a large independent dataset in addition to the missclassification rate in estimating $I(d^*(X) > 0)$ on the independent dataset.

S4.5 Details for RHC simulation and an additional simulation

We now define the outcome model used in the simulation using the RHC data from Section 6.3 the main text. The outcome model is based on outcome model D from Table S1. Outcome model D depends on 7 covariates,

Table S7: Estimates of the ATE and standard errors for the mechanical power data. Standard errors were computed for all methods using the nonparametric bootstrap with 1,000 replications. Also displayed are various measures of discrepancy between the distributions of covariates for the mechanical power high and mechanical power low groups. We also display the mean and max RIMSE statistic for marginal univariate and bivariate CDF differences, as in Figure 4. In addition, we display summary statistics of SMDs for marginal means and SMDs for all polynomials up to order 5 and pairwise interactions (denoted SMD(2))

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
$\hat{\tau}_w$	0.0405	0.0768	0.0997	0.0868	0.0729	0.0683
SE($\hat{\tau}_w$)	0.0149	0.0243	0.1724	0.0276	0.0198	0.0191
Energy dist (10)	42.7399	3.8453	52.7812	3.0272	1.4990	1.6333
Energy dist (16)	112.0910	6.3409	102.8196	4.1975	3.0852	2.8396
Mean RIMSE, 1d	0.0880	0.0136	0.0314	0.0147	0.0125	**0.0098**
Max RIMSE, 1d	0.3985	0.0873	0.0979	0.0602	0.0924	**0.0692**
Mean RIMSE, 2d	0.0892	0.0151	0.0404	0.0156	0.0131	**0.0105**
Max RIMSE, 2d	0.2617	0.0445	0.1443	0.0572	0.0407	**0.0297**
Mean |SMD|	0.2192	0.0012	0.1024	**0.0001**	0.0105	0.0068
Max |SMD|	1.1430	0.0163	2.4637	**0.0054**	0.0803	0.0524
Mean |SMD(2)|	0.1755	0.013	0.1016	0.0121	0.0146	**0.0104**
Max |SMD(2)|	1.1872	0.246	4.8639	0.2456	0.1456	**0.0888**

however the outcome model we use in this section uses an application of this model to multiple sets of 7 covariates from the RHC dataset. Define the mean function from outcome model D of Table S1 to be $f_D(\mathbf{x}^{1:7})$, where $1:7$ indicates that the first through seventh covariates are used in the mean model. We now define the outcome model of our simulation to be

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i \ \text{ for } \ i = 1, \ldots, 5{,}735,$$

where $f(\mathbf{x}_i) = \sum_{k=0}^{8} f_D(\mathbf{x}_i^{(7k+1):(7(k+1))})$ and ε are i.i.d $N(0, 5)$ random variables. Thus, 63 of the 65 covariates have an impact on the response. The design matrix and the treatment assignment vector are fixed throughout all simulations. Since the ordering of the covariates results in a different outcome model, since the 65 covariates are from the RHC dataset, we create new outcome models by uniformly at random permuting the columns of the design matrix. For each column permutation, we replicate the simulation 1,000 times and record the RMSE of each method for that permutation. Since the above outcome model used in the main text has a constant treatment effect of zero, we also include an outcome model with a treatment effect that varies with the covariates $\mathbf{X}$. The heterogeneous treatment effect model is

$$Y_i = f(\mathbf{x}_i) + A_i(f(\mathbf{x}_i) - \overline{f(\mathbf{x}_i)}) + \varepsilon_i \ \text{ for } \ i = 1, \ldots, 5{,}735,$$

where $f(\mathbf{x}_i)$ is defined as above and $\overline{f(\mathbf{x}_i)} = \sum_{i=1}^{n} f(\mathbf{x}_i)/n$ and ε are i.i.d $N(0, 5)$ random variables. The interaction between treatment and covariates is centered so that the sample average treatment effect is always 0, but varies significantly with $\mathbf{x}$. The median, average, standard devation, and maximum RMSEs over the 100 permutations of covariates for both the constant treatment effect setting and the heterogeneous treatment effect setting are displayed in Table S6. Both EBW and iEBW perform quite well, with iEBW with the lowest RMSEs on average, by median, with the lowest variability from permutation to permutation, and with the smallest worst-case RMSE.

S5 Remaining data analyses using the MIMIC-III critical care database

In this section we present the remaining two studies based on the MIMIC-III Critical Care Database.

Table S8: Displayed are the median, mean, standard deviation, and maximum RMSEs for each method across the 100 simulation settings using the mechanical power data

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
	Constant treatment effect					
Median RMSE	10.7097	5.4671	22.0340	3.8651	3.0773	2.3889
Mean RMSE	13.0120	7.1981	52.1948	5.2540	3.6800	2.5932
SD RMSE	9.2960	5.6951	61.4692	4.6933	2.7329	1.8679
Max RMSE	38.6099	22.2681	247.9993	21.9211	10.1050	7.6787
	Heterogeneous treatment effect					
Median RMSE	18.6460	7.5122	23.3952	6.6768	6.1644	5.2506
Mean RMSE	22.6549	9.2626	53.3998	9.6951	6.9088	5.8760
SD RMSE	16.1860	7.2928	61.1497	9.1299	5.0759	4.2979
Max RMSE	67.2217	29.4037	247.4224	43.2159	19.7568	16.7650

Table S9: Estimates of the ATE and standard errors for the echocardiography data. Standard errors were computed for all methods using the nonparametric bootstrap with 1,000 replications. Also displayed are various measures of discrepancy between the distributions of covariates for the echocardiography and control groups. We also display the mean and max RIMSE statistic for marginal univariate and bivariate CDF differences, as in Figure 4. In addition, we display summary statistics of SMDs for marginal means and SMDs for all polynomials up to order 5 and pairwise interactions (denoted SMD(2))

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
$\hat{\tau}_w$	0.0064	0.0317	0.0445	0.0284	0.0305	0.0309
SE($\hat{\tau}_w$)	0.0113	0.0206	0.0137	0.0113	0.0091	0.0088
Energy dist (10)	5.7673	0.4411	0.5970	0.3039	0.1994	0.2038
Energy dist (16)	17.2943	0.9708	1.261	0.8013	0.5137	0.5048
Mean RIMSE, 1d	0.0282	0.0095	0.0094	0.0095	0.0074	**0.0072**
Max RIMSE, 1d	0.0944	0.0214	0.0207	0.0211	0.0185	**0.0183**
Mean RIMSE, 2d	0.0424	0.0070	0.0078	0.0069	0.0051	**0.0048**
Max RIMSE, 2d	0.2683	0.0284	0.0249	0.0273	0.0173	**0.0140**
Mean |SMD|	0.0776	0.0005	0.0096	**0.0000**	0.0029	0.0022
Max |SMD|	0.2773	0.0153	0.0307	**0.0000**	0.0203	0.0133
Mean |SMD(2)|	0.1043	0.0074	0.0137	0.0071	0.0057	**0.0047**
Max |SMD(2)|	0.5062	0.0784	0.1675	0.0704	0.0431	**0.0385**

S5.1 Mechanical power of ventilation data

We use the MIMIC-III database to study the impact of a large degree of mechanical power of ventilation on outcomes. Our study and the construction of the cohort from the MIMIC-III database is based the original study of Neto et al. [21] and is based on the code provided by the authors located at https://github.com/alistairewj/mechanical-power. Neto et al. [21] treat mechanical power as a continuous treatment, however, we treat it as binary (whether mechanical power of ventilation of greater than 25 Joules per minute) for the purpose of demonstrating the use of our proposed EBWs. The study contains 5,014 patients, 1,298 of whom received a mechanical power of ventilation of greater than 25 Joules per minute, the amount of energy generated by the mechanical ventilator. The outcome is an indicator of in-hospital mortality. In all, the dimension of the design matrix of confounders is 86.

All methods explored in the main text were applied to adjust for the 86 confounders. Estimated treatment effects and balance statistics are displayed in Table S7. The KCB approach yielded constant weights of 1 regardless of the tuning parameter. From the univariate standardized mean differences (SMDs), Cal and CBPS balance marginal means the most effectively, however iEBW balances means of interactions and polynomials the best, with the smallest worst case mean imbalance and the best average imbalance. iEBW balances marginal distributions the most effectively on average and in the worse case scenario, with Cal a close second, followed by EBW and CBPS. iEBW balances bivariate distributions the best on average and in the worse case, followed by Cal and EBW. Among non-EBW approaches, Cal yields the smallest weighted energy distances, which is in alignment with its ability to balance marginal univariate and bivariate distributions for this data. The point estimates from each approach, including the unweighted analysis, suggest that mechanical power larger than 25 Joules/min harms patients in terms of in-hospital mortality, however iEBW and EBW suggest less harm than do other approaches. All approaches yield 95% confidence intervals that do not contain 0, except IPW, which has an extraordinarily large standard error compared with other approaches. iEBW and EBW yield the shortest length confidence intervals, suggesting a significant increase in in-hospital mortality from mechanical power greater than 25 Joules/min despite their attenuated estimate of the impact on mortality. These findings align qualitatively with the analysis conducted by Neto et al. [21].

Table S10: Displayed are the median, mean, standard deviation, and maximum RMSEs for each method across the 100 simulation settings using the echocardiography data

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
	Constant treatment effect					
Median RMSE	4.1580	1.3328	1.6951	1.3463	1.2802	0.9243
Mean RMSE	4.4938	1.7167	1.9567	1.7288	1.6174	1.2176
SD RMSE	3.4166	1.3789	1.5090	1.3744	1.3003	0.9296
Max RMSE	14.6757	5.7763	7.5684	5.7088	5.2813	3.8240
	Heterogeneous treatment effect					
Median RMSE	6.1837	1.7658	2.1917	2.1114	1.8577	1.5091
Mean RMSE	6.6822	2.4121	2.6146	2.6639	2.2572	1.6587
SD RMSE	5.0821	1.8060	1.9278	2.1047	1.6669	1.1206
Max RMSE	21.8241	7.8449	9.4596	8.6978	6.9643	4.8831

As mentioned, we also use the MPV data to conduct simulation studies, wherein we fix the confounders and treatment assignment and simulate outcomes. The median, average, standard deviation, and maximum RMSEs over the 100 permutations of covariates for both the constant treatment effect setting and the heterogeneous treatment effect setting are displayed in Table S8. We note that the rankings of each method in terms of their RMSEs across the simulation settings align with their weighted energy distances in Table S7, with iEBW performing best in terms of median, mean, and worst-case RMSE across all settings for both the constant treatment effect setup and the heterogeneous treatment effect setup, followed by EBW.

S5.2 Transthoracic echocardiography data

We use the MIMIC-III database to analyse a study of the effect of transthoracic echocardiography on 28 day mortality in sepsis patients originally conducted by Feng et al. [22]. Our construction of the study cohort from the MIMIC-III database is based on the code provided by Feng et al. [22] located at https://github.com/nus-mornin-lab/echo-mimiciii. The study contains information on 6361 patients, 3262 of whom received transthoracic echocardiography. The outcome is an indicator of mortality within 28 days of admission to the ICU. In all, the dimension of the design matrix of confounders is 77.

All methods explored in the main text were applied to adjust for the 77 confounders. Estimated treatment effects and balance statistics are displayed in Table S9. The KCB approach yielded constant weights of 1 regardless of the tuning parameter. From the univariate standardized mean differences (SMDs), Cal, CBPS, and iEBW balance marginal means the most effectively, however iEBW and EBW balance means of interactions and polynomials the best, with the smallest worst case mean imbalance and the best average imbalance. EBW and iEBW balance marginal distributions the most effectively on average and in the worse case scenario, with Cal a close second, followed by EBW and CBPS. EBW and iEBW balance bivariate distributions the best on average and in the worse case, followed by Cal and CBPS. The point estimates for all methods of the effect of echocardiography all indicate a potential reduction of 28 day mortality, with EBW, iEBW, Cal, and CBPS all suggesting a similar effect and IPW suggesting a stronger effect. 95% confidence intervals for all methods do not contain zero, except for CBPS which has a larger standard error. EBW and iEBW result in the smallest standard error and thus shortest length confidence interval. These findings align with the original analysis conducted in Feng et al. [22].

We also use the echocardiography data to conduct simulation studies, wherein we fix the confounders and treatment assignment and simulate outcomes. The median, average, standard deviation, and maximum RMSEs over the 100 permutations of covariates for both the constant treatment effect setting and the heterogeneous treatment effect setting are displayed in Table S10. We note that the rankings of each method in terms of their RMSEs across the simulation settings closely align with their weighted energy distances in Table S9, with iEBW performing best in terms of median, mean, and worst-case RMSE across all settings for both the constant treatment effect setup and the heterogeneous treatment effect setup, followed by EBW. Here, CBPS performs slightly better than Cal, unlike with the RHC, IAC, and MPV datasets.

S5.3 Simulation comparing with matching methods for estimation of the ATT

In this section we explore a comparison of energy balancing weights and various distance-based matching methods in estimation of the average treatment effect on the treated (ATT). We compared against nearest neighbor matching ("NN Matching") using the Mahalanobis distance as the matching criterion and also used the generalized full matching ("Full Matching") of Sävje et al. [23] also using the Mahalanobis distance as the matching distance; generalized full matching is a generalization of full matching [24] that is also computationally feasible for all datasets investigated.

We use the same data-generating setup as the heterogeneous treatment effect settings in the simulations for each of the four real data-based simulations. The only difference is that the estimand is defined as the ATT. As before, 100 outcome models are simulated from each of the RHC, echocardiography, IAC, and mechanical

Table S11: Displayed are the median, mean, standard deviation, and maximum RMSEs for each method across the 100 simulation settings focusing on estimation of the ATT

	RHC data			
	Unweighted	**NN matching**	**Full matching**	**EBW**
Median RMSE	4.6293	4.6484	2.6560	0.6230
Mean RMSE	5.8204	4.8012	3.1338	0.8206
SD RMSE	4.2938	2.7918	2.3384	0.6154
Max RMSE	22.3084	12.5596	11.1715	2.4829
	Echocardiography data			
Median RMSE	4.1580	3.4675	3.3596	1.3428
Mean RMSE	4.4938	4.4525	4.0710	2.0292
SD RMSE	3.4166	3.5902	3.2859	1.8390
Max RMSE	14.6757	17.2410	13.1938	7.6528
	IAC data			
Median RMSE	8.0151	8.0418	6.7149	5.6295
Mean RMSE	9.1296	9.4448	7.6033	6.2174
SD RMSE	6.7091	7.0132	5.3071	4.3376
Max RMSE	32.3715	33.1432	24.5695	21.3706
	Mechanical power data			
Median RMSE	10.7097	13.2411	10.5449	3.5021
Mean RMSE	13.0120	14.8983	10.9030	4.1407
SD RMSE	9.2960	9.3011	7.9816	2.7425
Max RMSE	38.6099	40.8559	31.0700	13.6858

power datasets. For each of the 100 outcome models, 1,000 datasets are generated and the root mean squared error is reported across the 1,000 replications for each method. The methods are evaluated in the same manner as for the previous simulation studies: the RMSEs are summarized across the 100 outcome models in terms of the median, mean, max, and standard deviation of the RMSE. The results are summarized in Table S11. For all four datasets, the energy balancing weights that target the ATT result in the lowest average and median RMSE in addition to the smallest worst-case RMSE across the 100 outcome models.

References

[1] Datta J, Polson N.. Inverse probability weighting: the missing link between survey sampling and evidence estimation. 2022. arXiv: http://arXiv.org/abs/arXiv:220414121.
[2] Geweke J. Bayesian inference in econometric models using Monte Carlo integration. Econometrica. 1989;57(6):1317–39.
[3] Robert C, Casella G. Monte Carlo statistical methods. Springer Science & Business Media; 2013.
[4] Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Amer Stat Assoc. 2018;113(521):390–400.
[5] Székely GJ, Rizzo ML. Energy statistics: A class of statistics based on distances. J Stat Plan Inference. 2013;143(8):1249–72.
[6] Wellner JA. A Glivenko-Cantelli theorem for empirical measures of independent but non-identically distributed random variables. Stochastic Process Appl. 1981;11(3):309–12.
[7] Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35(6):2769–94.
[8] Csörgö M, Nasari MM. Asymptotics of Randomly Weighted u-and v-statistics: Application to Bootstrap. J Multivariate Anal. 2013;121:176–92.
[9] Patterson RF. Strong convergence'for U-statistics in arrays of row-wise exchangeable random variables. Stochastic Anal Appl. 1989;7(1):89–102.
[10] Amaral S, Allaire D, Willcox K. Optimal L_2-norm empirical importance weights for the change of probability measure. Stat Comput. 2017;27(3):625–43.
[11] Tokdar ST, Kass RE. Importance sampling: a review. Wiley Interdiscipl Rev Comput Stat. 2010;2(1):54–60.
[12] Mak S, Joseph VR. Support points. Ann Stat. 2018;46(6A):2562–92.
[13] Van Zuijlen MC. Properties of the empirical distribution function for independent nonidentically distributed random variables. Ann Probabil. 1978;6(2):250–66.
[14] Billingsley P. Convergence of probability measures. John Wiley & Sons; 1993.
[15] Serfling RJ. Approximation theorems of mathematical statistics. John Wiley & Sons; 1980.
[16] Korolyuk VS, Borovskich YV. Theory of U-statistics, mathematics and its applications. vol. 273. Dordrecht: Kluwer Academic Publishers Group; 1994.
[17] Rizzo ML. A test of homogeneity for two multivariate populations. Proceedings of the American Statistical Association, Physical and Engineering Sciences Section. 2002.
[18] Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc Ser B (Stat Methodol). 2014;76(1):243–63.
[19] Chan KCG, Yam SCP, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. J R Stat Soc Ser B (Stat Methodol). 2016;78(3):673–700.
[20] Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Amer Stat Assoc. 2012;107(499):1106–18.
[21] Neto AS, Deliberato RO, Johnson AE, Bos LD, Amorim P, Pereira SM, et al. Mechanical power of ventilation is associated with mortality in critically ill patients: an analysis of patients in two observational cohorts. Intensive Care Med. 2018;44(11):1914–22.
[22] Feng M, McSparron JI, Kien DT, Stone DJ, Roberts DH, Schwartzstein RM, et al. Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database. Intensive Care Med. 2018;44(6):884–92.
[23] Sävje F, Higgins MJ, Sekhon JS. Generalized full matching. Political Anal. 2021;29(4):423–47.
[24] Rosenbaum PR. A characterization of optimal designs for observational studies. J R Stat Soc Ser B (Methodol). 1991;53(3):597–610.