Research Article

Jared D. Huling* and Simon Mak

Energy balancing of covariate distributions

https://doi.org/10.1515/jci-2022-0029
received April 26, 2022; accepted October 03, 2023

Abstract: Bias in causal comparisons has a correspondence with distributional imbalance of covariates between treatment groups. Weighting strategies such as inverse propensity score weighting attempt to mitigate bias by either modeling the treatment assignment mechanism or balancing specified covariate moments. This article introduces a new weighting method, called energy balancing, which instead aims to balance weighted covariate distributions. By directly targeting distributional imbalance, the proposed weighting strategy can be flexibly utilized in a wide variety of causal analyses without the need for careful model or moment specification. Our energy balancing weights (EBW) approach has several advantages over existing weighting techniques. First, it offers a model-free and robust approach for obtaining covariate balance that does not require tuning parameters, obviating the need for modeling decisions of secondary nature to the scientific question at hand. Second, since this approach is based on a genuine measure of distributional balance, it provides a means for assessing the balance induced by a given set of weights for a given dataset. We demonstrate the effectiveness of this EBW approach in a suite of simulation experiments, and in studies on the safety of right heart catheterization and on three additional studies using electronic health record data.

Keywords: energy distance, causal inference, covariate balance, observational data, weighting

MSC 2020: 62D20, 62G05

1 Introduction

Studying the causal effect of a treatment or intervention is a central goal in many scientific disciplines. In randomized controlled trials, estimation of causal effects is possible since randomization ensures that treated units and control units are comparable [1]. However, for many pressing questions, it is impossible or impractical to randomize treatment assignments. Researchers are thus left with existing sources of observational data to answer these questions. In these observational studies, it is of prime interest to make unconfounded comparisons between treatment groups, a common example being estimation of the average treatment effect (ATE). Yet in observational settings, natural selection processes into the treatment, resulting in imbalance in the covariate distributions between treatment groups. This may then introduce substantial bias in naive comparisons of the outcome of interest between groups.

In observational studies based on complex electronic health records (EHRs) or administrative health data, the differences between treated and untreated units are typically substantial and difficult to characterize. One motivating application for our work is the study by Connors et al. [2], which explored the impact of right heart catheterization (RHC), a diagnostic procedure designed to guide therapy, on mortality among intensive care unit (ICU) patients. In this study, patients who received RHC are different from those who did not receive RHC in highly complex ways. As an example among many such differences, both younger patients and older patients are less likely to receive RHC; thus, a simple correction for the average age does not characterize

*** Corresponding author: Jared D. Huling**, Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, MN, United States, e-mail: huling@umn.edu
Simon Mak: Department of Statistical Science, Duke University, Durham, North Carolina, United States, e-mail: sm769@duke.edu

the differences between those who received RHC and those who did not. Similar complex differences between treated and untreated patients can also be observed in three other motivating studies arising from clinical care and EHRs using the MIMIC-III critical care database [3]; we will demonstrate how the proposed method tackles these four challenging applications later.

There is a vast literature on adjustment methods for correcting for imbalances between treated and untreated units to reduce estimation bias in treatment effects. Weighting methods are a class of adjustment methods that control for confounding by re-weighting the treatment and control groups to look similar. Inverse-probability weighting (IPW) methods [4–9], which have origins in survey sampling, are by far the most commonly used weighting approaches. IPW methods model the treatment assignment mechanism (or propensity score, see Rosenbaum and Rubin [10]), and inverse weight each sample by the probability of receiving its assigned treatment. Inverse weighting by the true underlying propensity score controls for confounding, as it re-weights the covariate distributions of the treatment groups to that of the overall population.

In practice, IPW methods require positing and fitting a model for the propensity score. While model fitting is no unfamiliar task for a statistician, it has been noted in the literature that even mild model misspecifications for the propensity score can result in substantial bias in estimating the treatment effect [11]. Hearkening to George Box's maxim, "all models are wrong, but some are useful," it is often quite difficult in practice to obtain a useful propensity score model, especially in the presence of many covariates. Recent work has focused on mitigating this issue, by either (i) including conditions on the estimation of the propensity model which encourage moment balance of the covariates [12] or by (ii) altogether avoiding direct modeling of the propensity score and instead estimating weights that explicitly balance moments of the covariates either exactly [13,14] or approximately [15]. A large class of such estimators is explored Chan et al. [14] and further expanded on by Zhao [16] and has a long history [17]. Recent works [18,19] have attempted to mitigate this issue by focusing on nonparametric approaches to moment balancing, yet they require a careful choice of kernel function and tuning of multiple hyperparameters.

To make progress in addressing these issues, we show in this article that estimation bias for the ATE has a link to imbalance in the covariate *distributions*. This shows that balancing the full covariate distributions (and not just lower order moments) of the treatment groups to the full population provides a robust way for mitigating confounding. Although this is understood in the literature [16,20], in this article, we establish and make use of this link in a general manner, by (i) introducing a metric that evaluates how well a set of weights mitigates imbalance for a given dataset and (ii) developing a new method for estimating weights by minimizing this metric, yielding good distributional balance for a given dataset. The distributional balancing property of these weights allows for robust empirical performance; they tend to perform well in practice without relying on modeling assumptions for the propensity score or which moments are imbalanced. Our approach is designed to yield good distributional balance between treatment groups without the need of carefully tuning hyperparameters. Hirshberg et al. [21] study the theoretical properties of use of integral probability metrics for construction of weights; however, our work focuses on a specific choice that is suitable to wide use in practice. Despite the lack of need for tuning, our proposal works well empirically in a wide variety of settings.

We introduce weights that are explicitly constructed to balance the covariate distributions of the treatment groups to a target distribution (usually the full population). We do so by leveraging the energy distance presented in Székely and Rizzo [22]. The energy distance is a measure based on powers of the Euclidean distance and was originally introduced as a means to replace standard nonparametric goodness-of-fit tests in high dimensions. The energy distance has an exact duality with a norm on the characteristic functions, enabling its use to compare two (or more) distributions or the distributions of two samples. We show that a *weighted* energy distance still retains this duality, making it a rigorously justified and reliable metric to compare between multiple sets of weights for a given dataset. From this, we propose the so-called *energy balancing weights* (EBWs), which are defined as the weights which minimize the weighted energy distance between treatment groups and the full sample, subject to constraints that mitigate variability in the weights. We prove that EBWs asymptotically ensure full distributional balance and result in root-n consistent estimation of the ATE. Our emphasis is on the robust performance of our approach in practice, and our asymptotic

analysis serves primarily to justify the use of the proposed weights. We analyze four challenging observational studies and use each of these datasets to conduct realistic and highly challenging simulations, demonstrating the effectiveness of EBWs in practice with minimal user input required. Two of the studies we present in this article and the remaining two are presented in the Supplementary Material.

Although we focus primarily on a simple estimand, namely, the average treatment effect (ATE) the proposed weights can be used for a wide variety of causal estimands since the distributional re-balancing property of EBWs enables flexible control of confounding. We show that they can be used for the estimation of a wide variety of causal quantities, such as the ATE and individualized treatment rules (ITRs) [23,24]. With minor modifications, they can also be used for estimation of the average treatment effect on the treated (ATT) and for estimating treatment effects for multi-category treatments. Despite the fact that EBWs are not specifically designed to match low-order moments, in practice, they often result in better marginal mean balance of covariates than propensity score methods even in high dimensions (50–100 covariates), as seen in Section 6. EBWs are also quite stable in practice, rarely resulting in large weights, an issue that plagues standard propensity score methods [11], making it less critical to impose constraints that induce weight variability. EBWs are constructed without using outcome information; however, as for other weighting approaches, variance can be reduced via augmented estimators that make use of outcome regression models, such as the augmented estimators in the studies by Wong and Chan [18], Zhao [16], and Athey et al. [25]. However, we do not explore such techniques here.

The remainder of this article is organized as follows. Section 2 motivates the need for distributional balance and introduces the weighted energy distance. Section 3 presents the proposed EBWs and discusses their computation and asymptotic properties. Section 5 compares the performance of EBWs with other weighting methods in simulation studies. Section 6 discusses an application of EBWs in a study of RHC and further uses these data to explore highly challenging simulations that use the natural data-generating processes of the study. Section 7 further demonstrates the utility of EBWs by analyzing data from a complex application involving EHR data. Section 8 concludes with a discussion and future work.

2 Distributional balance and weighted energy distance

2.1 Setup

Consider a sample $\{(Y_i, A_i, \mathbf{X}_i)\}_{i=1}^n$ of size n from a population, where Y_i is the outcome of the ith unit, $A_i \in \{0, 1\}$ is a binary indicator of receiving a treatment, and $\mathbf{X}_i \in \mathcal{X} = \mathbb{R}^p$ is a p-dimensional vector of covariates. Further denote $n_1 = \sum_{i=1}^n A_i$ and $n_0 = n - n_1$. In this article, we are interested in estimating the average causal effect of the treatment on the outcome. A formal definition of a causal effect often involves the use of so-called potential outcomes [26–29]. The potential outcome $Y(a)$ is the outcome that would have been observed under level a of the treatment. As each individual only receives one level of the treatment at a given time, only one potential outcome, either $Y_i(0)$ or $Y_i(1)$, for each individual is observable. We assume the standard stable unit treatment value assumption (SUTVA), which posits that the potential outcomes for each unit are unaffected by the potential outcomes of other units and that only one version of the treatment exists. Under SUTVA, the observed outcome is consistent with the potential outcomes in that $Y_i = Y_i(A_i)$. We further assume the assignment mechanism is *strongly unconfounded* in the sense that $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid \mathbf{X}$, which requires that there are no unmeasured confounders. The $\perp\!\!\!\perp$ notation of Dawid [30] denotes (conditional) independence. We further assume *positivity* (or probabilistic assignment) in that the propensity score $\pi(\mathbf{x}) \equiv \mathbb{P}(A = 1 \mid \mathbf{X} = \mathbf{x})$ [10] satisfies $0 < \pi(\mathbf{x}) < 1$, so that everyone has a *chance* of receiving the treatment. Positivity, together with no unmeasured confounders, constitute *strong ignorability* [10].

Let us denote $\mu_a(\mathbf{X}_i) \equiv \mathbb{E}(Y(a) \mid \mathbf{X}_i)$ as the conditional mean function and $\sigma_a^2(\mathbf{X}_i) \equiv \mathbb{V}(Y_i(a) \mid \mathbf{X}_i)$ as the conditional variance function of the response for $a \in \{0, 1\}$. We consider scenarios where data have been collected from an observational study, and thus, the treatment groups are not comparable due to imbalances

in the distributions of their baseline covariates. These differences can be characterized through $\pi(\mathbf{x})$ or through $F_a(\mathbf{x}) \equiv \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid A = a)$, the cumulative distribution function (CDF) of covariates $\mathbf{X}$ conditional on treatment level a.

Using the notation of potential outcomes, the (population) ATE is defined as $\tau \equiv \mathbb{E}(Y(1) - Y(0))$. This can be rewritten as follows:

$$\tau = \int_{\mathbf{x} \in \mathcal{X}} [\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})] \mathrm{d}F(\mathbf{x}), \tag{1}$$

where $F(\mathbf{x}) = Pr(\mathbf{X} \leq \mathbf{x})$ is the CDF of covariates $\mathbf{X}$ marginalized over the treatment groups. In other words, $F(\mathbf{x}) = F_1(\mathbf{x})P_1 + F_0(\mathbf{x})P_0$, where $P_a \equiv \mathbb{P}(A = a) \in (0, 1)$ is the probability of being assigned treatment level $a \in \{0, 1\}$.

2.2 Weighted average estimates and distributional balance

We restrict our focus to weighted averages as estimates of the ATE. Given a vector of weights $\mathbf{w} = (w_1, \ldots, w_n)$, we study estimators of the form:

$$\hat{\tau}_{\mathbf{w}} = \frac{1}{n_1} \sum_{i=1}^{n} w_i Y_i A_i - \frac{1}{n_0} \sum_{i=1}^{n} w_i Y_i (1 - A_i). \tag{2}$$

The most commonly used example of (2), inverse propensity score weighting, uses $w_i = (A_i \pi(\mathbf{X}_i) n_1/n + (1 - A_i)(1 - \pi(\mathbf{X}_i)) n_0/n)^{-1}$. As presented in Imai and Ratkovic [12] and Li et al. [20], the weights in (2) are often normalized by treatment group, i.e., $\sum_{i=1}^{n} w_i I(A_i = a) = n_a$ for $a \in \{0, 1\}$, to improve precision [11,16] at the cost of a small bias. When these weights are constructed as mentioned earlier and normalized by treatment group, the resulting estimator is called the Hájek estimator [31] and may yield reduced mean squared error over its nonnormalized version [32].

Given any weight vector $\mathbf{w}$, we can express the error of $\hat{\tau}_{\mathbf{w}}$ as follows:

$$\hat{\tau}_{\mathbf{w}} - \tau = \int_{\mathcal{X}} \mu_1(\mathbf{x}) \mathrm{d}[F_{n,1,\mathbf{w}} - F_n](\mathbf{x}) - \int_{\mathcal{X}} \mu_0(\mathbf{x}) \mathrm{d}[F_{n,0,\mathbf{w}} - F_n](\mathbf{x}) \tag{3}$$

$$- \int_{\mathcal{X}} [\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})] \mathrm{d}[F - F_n](\mathbf{x}) \tag{4}$$

$$+ \frac{1}{n_1} \sum_{i=1}^{n} w_i \varepsilon_i A_i - \frac{1}{n_0} \sum_{i=1}^{n} w_i \varepsilon_i (1 - A_i), \tag{5}$$

where $\varepsilon_i \equiv Y_i(A_i) - \mu_{A_i}(\mathbf{X}_i)$, $F_n(\mathbf{x}) = \sum_{i=1}^{n} I(\mathbf{X}_i \leq \mathbf{x})/n$ is the empirical CDF (ECDF) of the combined sample $\{\mathbf{X}_i\}_{i=1}^{n}$, and $F_{n,a,\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{n} w_i I(\mathbf{X}_i \leq \mathbf{x}, A_i = a)/n_a$ is the *weighted* ECDF for treatment level $a \in \{0, 1\}$. In observational studies, F_n is not impacted by the weights, so the error term (4) is irreducible. However, this term goes to 0 as long as the sample is representative of the desired population. Since (5) always has mean 0, the bias of $\hat{\tau}_{\mathbf{w}}$ in essence depends on the properties of (3): the difference of integrals with respect to $F_{n,a,\mathbf{w}} - F_n$. Thus, the systematic source of error from the weighted estimator $\hat{\tau}_{\mathbf{w}}$ can be completely controlled by reducing the imbalance between the weighted ECDFs $F_{n,a,\mathbf{w}}$ and the ECDF F_n. Unlike the decomposition in Zhao [16], the decomposition of $\hat{\tau}_{\mathbf{w}} - \tau$ above holds even if the treatment effect is not constant over $\mathbf{x}$. Further, none of the methodology or results of this article require such a constant treatment effect to justify the validity of our methods. We note that the systematic source of bias in equation (3) can be zero without $F_{n,a,\mathbf{w}}$ being balanced to F_n, depending on $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$. However, knowledge of such an event *a priori* is impossible without knowledge of the mean potential outcome functions; as such, a measure of distributional imbalance is critical to characterize the degree of bias for a given dataset. The terms in (5) drives the variability of $\hat{\tau}_{\mathbf{w}}$ and can be straightforwardly mitigated with a measure of dispersion of the weights [33]; we discuss this more at the end of Section 3; however, this is not the focus of this article.

The notion that the covariate distributions should be balanced to obtain a good estimate of τ is not new (see Imai and Ratkovic [12] and Li et al. [20], among many others). In fact, it is well understood that the weights resulting from correctly specified propensity score models asymptotically balance covariate distributions. However, slight model misspecifications of the propensity score may result in poor performance of (2) [11]. Further, two units with the same propensity score do not necessarily have the same covariate values, so in finite samples, propensity score weights may not be optimal. Other approaches that aim to estimate weights by balancing prespecified *moments* of the covariates [12,14] tend to be more robust than directly modeling $\pi(\mathbf{x})$. Yet, the term (3) makes it explicit that bias is directly related to distributional imbalance and that, holding μ_a fixed, weights that yield less *distributional* imbalance will in general result in less bias. A natural conclusion is that $\mathbf{w}$ should be estimated to directly balance each $F_{n,a,\mathbf{w}}$ to F_n. In the following, we introduce a new distance metric between $F_{n,a,\mathbf{w}}$ and F_n, which enables one to characterize how well a set of weights re-balances the weighted distribution $F_{n,a,\mathbf{w}}$ to F_n.

2.3 Weighted energy distance

We introduce next a new measure of the distributional balance induced by a set of weights. This measure is based on the *energy distance*, which is a metric on distributions [34]. Due to the link between estimation bias and distributional imbalance, our measure can be used to evaluate the degree of bias one expects from a given set of weights and a given dataset. We will later leverage this measure to construct distributional balance weights that minimize this metric for a given dataset.

The energy distance (as surveyed in Székely and Rizzo [34]) is defined as follows. Let G and H be two finite-mean distribution functions on $\mathcal{X}$, and let $\mathbf{Z}, \mathbf{Z}' \overset{i.i.d}{\sim} G$ and $\mathbf{V}, \mathbf{V}' \overset{i.i.d}{\sim} H$. The energy distance between distributions G and H is defined as follows:

$$\mathcal{E}(G, H) \equiv 2\mathbb{E}||\mathbf{Z} - \mathbf{V}||_2 - \mathbb{E}||\mathbf{Z} - \mathbf{Z}'||_2 - \mathbb{E}||\mathbf{V} - \mathbf{V}'||_2, \tag{6}$$

where $||\cdot||_2$ is the Euclidean norm. When both G and H are ECDFs, i.e., G_n is the ECDF of $\{\mathbf{Z}_i\}_{i=1}^n \subseteq \mathcal{X}$ and H_m is the ECDF of $\{\mathbf{V}_i\}_{i=1}^m \subseteq \mathcal{X}$, the energy distance $\mathcal{E}(G_n, H_m)$ can be expressed as follows:

$$\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}||\mathbf{Z}_i - \mathbf{V}_j||_2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{Z}_i - \mathbf{Z}_j||_2 - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}||\mathbf{V}_i - \mathbf{V}_j||_2. \tag{7}$$

The energy distance has been used within a wide variety of statistical methods, e.g., for testing equivalence of distributions, for testing statistical independence [35], and for generating samples from a target distribution [36]. The energy distance is simple to compute as it only involves sums of Euclidean distances, it is nonnegative, and it can be shown to be equivalent to a distance between the characteristic functions of G and H, indicating that a value of 0 occurs if and only if G and H are the same distribution and a positive value indicates how similar they are. The name "energy" comes from the fact that the population energy distance (or limit of the empirical energy distance) has an intricate connection with gravitational potential energy (which depends on the distance between two bodies of mass). Two key attractive properties of the energy distance are that it tends to perform well in measuring the distance between two distributions in moderately high-dimensional settings, and it does not require the choice of a kernel or any tuning parameters.

We propose a *weighted* modification of this energy distance, which measures the distance between a weighted distribution, i.e., the weighted covariate ECDFs $F_{n,0,\mathbf{w}}$ and $F_{n,1,\mathbf{w}}$ for the control and treated, and a target distribution, i.e., the combined covariate ECDF F_n. The *weighted energy distance* between $F_{n,a,\mathbf{w}}$ and F_n is defined as follows:

$$\begin{aligned}\mathcal{E}(F_{n,a,\mathbf{w}}, F_n) \equiv{}& \frac{2}{n_a n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i I(A_i = a)||\mathbf{X}_i - \mathbf{X}_j||_2 \\ &- \frac{1}{n_a^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j I(A_i = A_j = a)||\mathbf{X}_i - \mathbf{X}_j||_2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{X}_i - \mathbf{X}_j||_2.\end{aligned}$$

In other words, $\mathcal{E}(F_{n,a,\mathbf{w}}, F_n)$ is the energy distance between the ECDF of a sample $\{\mathbf{X}_i\}_{i=1}^n$ and a weighted ECDF of a subsample $\{\mathbf{X}_i\}_{i:A_i=a}$.

We show this new weighted energy distance is indeed a distance between $F_{n,a,\mathbf{w}}$ and F_n. Let $\langle \mathbf{t}, \mathbf{s} \rangle$ be the inner product of vectors $\mathbf{t}$ and $\mathbf{s}$ and define the empirical characteristic function (ECHF) of $\{\mathbf{X}_i\}_{i=1}^n$ and the weighted ECHFs of $\{\mathbf{X}_i\}_{i:A_i=a}$ as follows:

$$\varphi_n(\mathbf{t}) \equiv \frac{1}{n}\sum_{i=1}^{n} \exp\{i\langle \mathbf{t}, \mathbf{X}_i\rangle\} \quad \text{and} \quad \varphi_{n,a,\mathbf{w}}(\mathbf{t}) \equiv \frac{1}{n_a}\sum_{i=1}^{n} w_i I(A_i = a) \exp\{i\langle \mathbf{t}, \mathbf{X}_i\rangle\},$$

respectively. The following proposition establishes the *distance property* of the weighted energy distance.

Proposition 2.1. *Let* $\mathbf{w}$ *be a vector of weights such that* $\sum_{i=1}^n w_i I(A_i = a) = n_a$ *for* $a \in \{0, 1\}$ *and* $w_i > 0$ *for* $i = 1, \ldots, n$. *Then*

$$\mathcal{E}(F_{n,a,\mathbf{w}}, F_n) = \int_{\mathbb{R}^p} |\varphi_n(\mathbf{t}) - \varphi_{n,a,\mathbf{w}}(\mathbf{t})|^2 \omega(\mathbf{t}) \mathrm{d}\mathbf{t} \quad \textit{for } a \in \{0, 1\}, \tag{8}$$

where $\omega(\mathbf{t}) = 1/(C_p ||\mathbf{t}||_2^{1+p})$, $C_p = \pi^{(1+p)/2}/\Gamma((1+p)/2)$ *is a constant, and* $\Gamma(\cdot)$ *is the complete gamma function. Thus,* $\mathcal{E}(F_{n,a,\mathbf{w}}, F_n) \geq 0$ *with equality to zero iff* $\varphi_{n,a,\mathbf{w}}(\mathbf{t}) = \varphi_n(\mathbf{t})$ *for all* $\mathbf{t}$.

Thus, the weighted energy distance is a distance between the weighted distribution of interest and the target distribution, and is thus a *bona fide* measure of distributional balance of covariates induced by a set of weights. This proportion extends the duality results of Proposition 1 from Székely and Rizzo [34] and Theorem 1 from Székely et al. [35] for the weighted energy distance at hand. A subtle point is that Proposition 2.1, combined with the decomposition presented in Section 2.2, makes clear that $\mathcal{E}(F_{n,a,\mathbf{w}}, F_n)$ more closely aligns with evaluating how well a set of weights estimates the *sample* average treatment effect [32] than the population ATE τ that is our main focus. Yet, as long as F_n is representative of F, the weighted energy distance still aligns well with the population ATE.

We now show that the weighted energy distance converges to the energy distance when the weights yield a well-defined limiting distribution.

Theorem 2.2. *Assume* $\mathbb{E}(||\mathbf{X}||_2 \mid A = a) < \infty$ *and* $\mathbb{E}||\mathbf{X}||_2 < \infty$. *Further assume, for a sequence of weights* $\{\mathbf{w}_n\}_{n=1}^\infty$ *with* $\sum_{i=1}^n w_i I(A_i = a) = n_a$ *for* $a \in \{0, 1\}$ *and* $w_i > 0$ *for* $i = 1, \ldots, n$ *for each* n (*so that* $F_{n,a,\mathbf{w}_n}$ *are well-defined as CDFs*), *that* $\lim_{n\to\infty} \varphi_{n,a,\mathbf{w}_n}(\mathbf{t}) \to \widetilde{\varphi}_a(\mathbf{t})$ *almost surely for every* $\mathbf{t} \in \mathbb{R}^p$, *where* $\widetilde{\varphi}_a$ *is some integrable characteristic function with associated CDF* $\widetilde{F}_a(\mathbf{x})$. *Then almost surely we have*

$$\lim_{n\to\infty} \mathcal{E}(F_{n,a,\mathbf{w}_n}, F_n) = \mathcal{E}(\widetilde{F}_a, F). \tag{9}$$

Theorem 2.2 shows that the weighted energy distance converges to the limiting energy distance. If the limiting distribution implied by a set of weights is $\widetilde{F}_0 = \widetilde{F}_1 = F$, then $\mathcal{E}(\widetilde{F}_0, F) + \mathcal{E}(\widetilde{F}_1, F) = 0$. Proposition 2.1 and Theorem 2.2 together imply that weights with smaller values of the sum $\mathcal{E}(F_{n,1,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,0,\mathbf{w}}, F_n)$ yield better distributional balance of covariates. Due to the link between imbalance and bias (see the following section), this also implies that better balance will yield estimates with smaller values for the term (3).

2.4 Bias and distributional imbalance

We now demonstrate this connection between bias and distributional imbalance (as measured by the weighted energy distance) using two illustrative examples. A more formal presentation is provided later in Section 3.2 when proving asymptotic properties of EBWs.

In the first illustrative example, we generate a one-dimensional covariate of sample size 250, which impacts treatment assignment via a logistic model under each of three scenarios: (1) $\mathrm{logit}(\pi(X)) = -1 + X$,

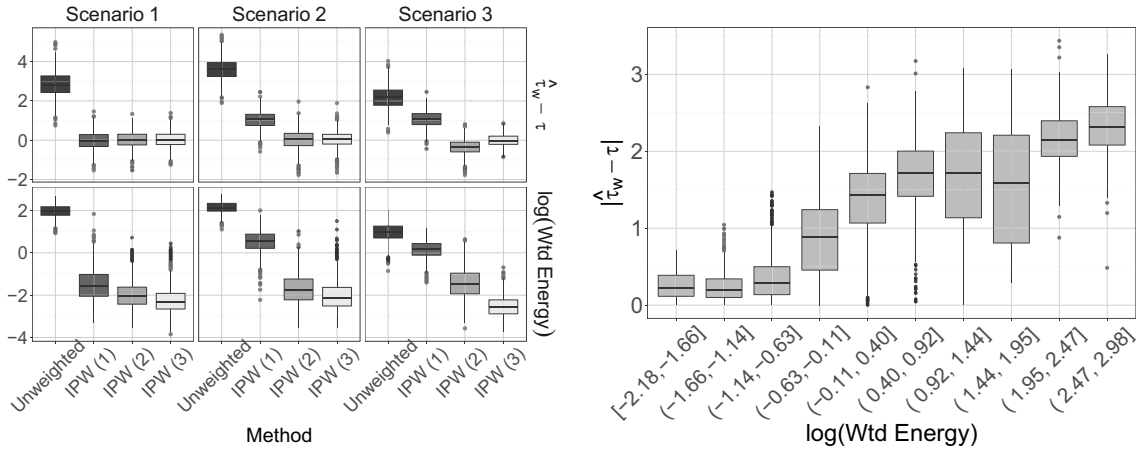

Figure 1: (a, left) Energy distances and biases for IPW estimates based on weights from the three fitted logistic regression models; (b, right) Boxplots of the biases for IPW estimates versus weighted energy distance based on weights estimated by several methods, each with different combinations of moments included for balancing or estimation.

(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 IPWs based on three logistic regression models, which consider only a linear term in X (denoted as "IPW (1)"), a linear plus quadratic term ("IPW (2)"), and up to the cubic term ("IPW (3)"), respectively. Thus, for Scenarios 2 and 3, at least one of the fitted models is misspecified. For each set of weights $\mathbf{w}$, we compute $\mathcal{E}(F_{n,0,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,1,\mathbf{w}}, F_n)$, i.e., the sum of the energy distances between each treatment group and the combined sample, and compute the error $\hat{\tau}_{\mathbf{w}} - \tau$ of (2) for τ.

Figure 1(a) displays the energy distances and biases over 1,000 replications of the experiment. We see that the energy distances are the largest in all scenarios for the unweighted estimator ((2) with all weights equal to 1). For scenarios where the weights are estimated using a misspecified model (IPW (1) in Scenario 2 and IPW (1) and (2) in Scenario 3), the energy distances are much larger than for weights based on correctly (or over-specified) models. Correspondingly, the bias is pronounced for the misspecified models. Thus, the weighted energy distance can be a useful tool to compare between different models, as weights with smaller weighted energy distances tend to yield estimates with smaller error.

In the second example, we consider a two-dimensional example where the true assignment mechanism depends on first and second moments of the covariates. We consider several methods for estimate weights: logistic regression, the method of Imai and Ratkovic [12], and the method of Chan et al. [14], each with different moments included for balancing or estimation. We then compare their weighted energy distances and absolute errors of (2) over 1,000 replications. Figure 1(b) displays the distances and errors for each dataset and method. We see that, in general, weights with lower energy distance have a much smaller magnitude of bias in estimating the ATE.

2.5 Other measures of distributional imbalance

There are, of course, many ways of measuring distance between distributions in the literature, including the Kolmogorov-Smirnov statistic, f-divergences (e.g., the Kullback-Leibler divergence and the Hellinger distance), the Wasserstein distance, and the maximum mean discrepancy (MMD). The energy distance has several advantages for characterizing distributional imbalance. First, while in general there is no uniformly most powerful nonparametric test for the difference between two distributions, the energy distance is often sensitive to differences in distributions, unlike the Kolmogorov-Smirnov statistic. Second, unlike the energy distance, f-divergences such as the Kullback–Leibler divergence and Hellinger distance do not metrize weak

convergence; this property ensures that the distance remains stable under perturbations of the support of the distributions being measured [37]. Third, the energy distance is easy and efficient to compute, unlike the Wasserstein distance. It also works reasonably well in moderately high dimensions [36], whereas the Wasserstein distance suffers from the curse of dimensionality, in the sense that empirical Wassterstein distances converge to the true Wasserstein distance at rate $O(n^{-1/p})$ [37,38]. Finally, while there is a direct link between MMD and the energy distance [39], the energy distance does not require careful tuning of hyperparameters and tends to work well across a wide variety of scenarios.

3 Energy balancing weights

3.1 Definition

We will now use the proposed weighted energy distance to estimate weights which (i) match the distribution of covariates of the treated group to the distribution of covariates of the full population, and (ii) match the distribution of covariates of the control group to the distribution of covariates of the full population.

To achieve this, we define the EBWs to be

$$\begin{aligned} \mathbf{w}_n^e \in & \underset{\mathbf{w}=(w_1,\ldots,w_n)}{\operatorname{argmin}} \{\mathcal{E}(F_{n,1,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,0,\mathbf{w}}, F_n)\} \\ & \text{s.t. } \sum_{i=1}^n w_i A_i = n_1, \sum_{i=1}^n w_i(1-A_i) = n_0, w_i \geq 0 \text{ for } i = 1, \ldots, n. \end{aligned} \tag{10}$$

Thus, the EBWs $\mathbf{w}_n^e$ minimize the statistical energy between the treatment and control groups to the full population. Due to the duality result of Proposition 2.1, minimizing statistical energy is directly equivalent to balancing covariate distributions. The constraints $\sum_{i=1}^n w_i I(A_i = a) = n_a$ serve several purposes: they (i) preserve the sample size of the weighted pseudo-population to that of the study population, (ii) stabilize the estimated weights, and (iii) ensure that $F_{n,0,\mathbf{w}}$ and $F_{n,1,\mathbf{w}}$ are valid distribution functions. Also note that, due to the bias decomposition in (3) and the duality result in Proposition (2.1), the weights $\mathbf{w}_n^e$ are explicitly designed to minimize the key component of the finite-sample bias of $\hat{\tau}_{\mathbf{w}_n^e}$, in a manner agnostic to the functional forms of $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$. If the functional forms of $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ are known, it is certainly possible to construct a different set of weights to better reduce bias, by emphasizing balance in regions of $\mathcal{X}$ where either $\mu_0(\mathbf{X})$ or $\mu_1(\mathbf{X})$ are pronounced. However, this information is rarely (if ever) known, and hence this is difficult to achieve in practice except by good fortune.

To illustrate the effectiveness of EBWs for distributional balance, we consider data generated under Scenario 3 of the toy example in Figure 1(a). Figure 2 shows the difference between the weighted ECDF using EBWs of the covariate in the treatment group and the ECDF of the combined (i.e., treated and untreated) sample, for varying sample sizes n. This is compared with the weighted ECDF using weights from the true data-generating propensity score, the estimated propensity score under the correct model, and the unweighted ECDF of the treatment group. As sample size increases, the difference vanishes for EBWs, but much more slowly for both the true and estimated propensity score weights. This demonstrates the improved distributional balance provided by the proposed EBWs, which should then translate to a greater reduction of bias.

EBWs can be naturally extended to handle a wide variety of scenarios, such as for treatments with more than two levels, estimation of the ATT, and for the estimation of optimal ITRs. In the Supplementary Material, we show how these extensions are manifested and empirically demonstrate the benefit of using EBWs for ITR estimation.

A few key distinguishing features of our proposal from the works of Wong and Chan [18] and Kallus [19] are (1) its direct focus on distributional balance rather than moment balance. Despite the connection between balancing moments of an infinite dimensional class of functions and distributional balance, this feature helps alleviate modeling decisions about what moments to balance of what space of moments to balance and has

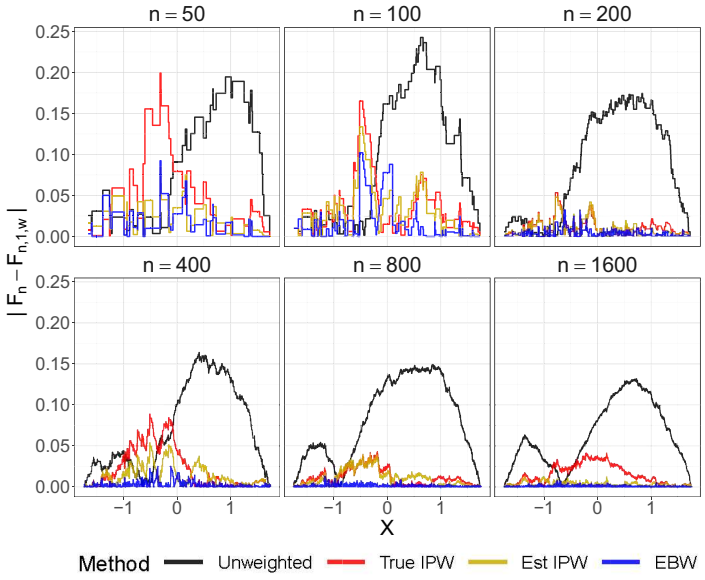

Figure 2: Displayed are the absolute differences between the ECDF of the combined sample and the weighted ECDF of the covariate in the treatment group based on energy weights. Also displayed the same for the unweighted ECDF of the treated group and the weighted ECDF based on the true and estimated propensity scores.

advantages in terms of interpretability and (2) by focusing on a measure that does not require tuning parameters to characterize distributional differences, our approach can be applied broadly by practitioners of varying degrees of statistical sophistication.

3.2 Asymptotic properties

Next, we show two desirable properties of the proposed EBWs. We first show that the weighted ECDFs based on EBWs indeed converge to the population CDF F of $\mathbf{X}$.

Theorem 3.1. *Assume that* $\mathbb{E}(\|\mathbf{X}\|_2 \mid A = a) < \infty$ *and* $\mathbb{E}\|\mathbf{X}\|_2 < \infty$ *and that the assumptions presented in Section* 2.1 *hold. Let* $\mathbf{w}_n^e$ *be as defined in* (10)*. Then, for* $a \in \{0, 1\}$,

$$\lim_{n\to\infty} F_{n,a,\mathbf{w}_n^e}(\mathbf{x}) \equiv \lim_{n\to\infty} \frac{1}{n_a} \sum_{i=1}^{n} w_i^e I(\mathbf{X}_i \le \mathbf{x}, A_i = a) = F(\mathbf{x}) \tag{11}$$

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

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

holds almost surely.
Thus, EBWs result in the almost sure convergence of the weighted ECDFs of the treated (and untreated) group to the underlying covariate distribution F.

The consistency of $\hat{\tau}_{\mathbf{w}_n^e}$ follows immediately from Theorem 3.1:

Corollary 3.2. *Suppose the conditions of Theorem* 3.1 *hold, and that the treatment and control mean response functions* $\mu_0(\mathbf{x})$ *and* $\mu_1(\mathbf{x})$ *are bounded and continuous on* $\mathcal{X}$*. Then* $\hat{\tau}_{\mathbf{w}_n^e}$ *is a consistent estimator of* τ.

Next, we show that the ATE estimator $\hat{\tau}_{\mathbf{w}_n^e}$ is root-n consistent. To do so, we make use of the following lemma, which provides a connection between bias and distributional balance:

Lemma 3.3. *Let $\mathcal{H}$ be the native space induced by the radial kernel $\Phi(\cdot) = -\|\cdot\|_2$ on $\mathcal{X}$, and suppose $\mu_a(\cdot) \in \mathcal{H}$ for $a \in \{0, 1\}$. Then, for any weights $\mathbf{w}$ satisfying $\sum_{i=1}^n w_i A_i = n_1$, $\sum_{i=1}^n w_i(1 - A_i) = n_0$, $w_i \geq 0$, we have for $a \in \{0, 1\}$:*

$$\left[\int \mu_a(\mathbf{x}) \mathrm{d}[F_{n,a,\mathbf{w}} - F_n](\mathbf{x})\right]^2 \leq C_a \mathcal{E}(F_{n,a,\mathbf{w}}, F_n),$$

where $C_a \geq 0$ is a constant depending on only μ_a.

Lemma 3.3 provides a connection between the systematic bias in (3) and the weighted energy distance $\mathcal{E}(F_{n,a,\mathbf{w}}, F_n)$. Under the mild condition that the conditional mean function $\mu_a(\mathbf{x})$ is found in $\mathcal{H}$ (this is discussed further at the end of the section), this lemma shows that the weighted energy distance is a key component in an upper bound on the systematic bias in (3). We note that Lemma 3.3 applies to any set of auto-normalized weights, justifying the use of the weighted energy distance to compare between different sets of weights. The proposed EBWs $\mathbf{w}_n^e$, which minimize $\mathcal{E}(F_{n,a,\mathbf{w}}, F_n)$, may therefore yield lower estimation bias via this upper bound, which is in line with the empirical observations in Section 2.4. With this, we now state the result on root-n consistency:

Theorem 3.4. *Assume the same conditions in Theorem* 3.1. *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*:

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

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

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

(4) $\mathbb{E}[h_0^2(\mathbf{X}, \mathbf{X}', \mathbf{X}'', \mathbf{X}''')] < \infty$ *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 follows*:

$$h_a(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{a}) = \frac{1}{\pi_a(\mathbf{x})}\|\mathbf{x} - \mathbf{z}\|_2 + \frac{1}{\pi_a(\mathbf{y})}\|\mathbf{y} - \mathbf{a}\|_2 - \frac{1}{\pi_a(\mathbf{x})\pi_a(\mathbf{y})}\|\mathbf{x} - \mathbf{y}\|_2 - \|\mathbf{a} - \mathbf{z}\|, \tag{12}$$

(5) *The EBWs $\mathbf{w}_n^e = (w_{i,n}^e)_{i=1}^n$ in* (10) *satisfy $w_{i,n}^e \leq Cn^{1/3}$ for some constant $C > 0$ independent of n.*

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

$$\sqrt{\mathbb{E}_{\mathbf{X},A,Y}[(\hat{\tau}_{\mathbf{w}_n^e} - \tau)^2]} = O(n^{-1/2}). \tag{13}$$

We give a brief discussion of Assumptions (A1)–(A5). Assumption (A1) concerns the regularity of the conditional mean functions μ_0 and μ_1. It can be shown that the Sobolev space $W_{\lceil (p+1)/2 \rceil, 2}(\mathcal{X})$ – the space of functions with square-integrable $r < \lceil (p + 1)/2 \rceil$-th differentials – is contained within the native space $\mathcal{H}(\mathcal{X})$ (Theorem 10.42 of [40]), so (A1) can be viewed as a smoothness assumption on the conditional ATE $\mu_1 - \mu_0$ (a similar assumption is made in [18]). Assumptions (A2) and (A3) require the conditional mean functions to have finite variance and the conditional variance function to be bounded, respectively. Assumption (A4) require the kernels h_0 and h_1 to have finite second moments; these kernels can be seen as a "modified" Euclidean kernel weighted by the true propensity scores. Assumption (A5) assumes all EBWs to be bounded above by $Cn^{1/3}$ for some positive constant $C > 0$; this assumption has been used in several weighting-based covariate balancing methods [18,25]. In practice, (A5) can always be checked after optimizing for the EBWs in (10). One can change the optimization procedure (10) to explicitly enforce (A5), but we have never encountered "exploding weights," which violate (A5) in practice; EBWs in our simulations and data analysis typically satisfy (A5) with $C = 1$. As we discuss in Section 3.4, if one is willing to add a penalty on the dispersion of the weights, the explicit assumption on the maximum of the weights is unnecessary and root-n consistency still holds, as shown in the Supplement.

While Theorem 3.4 proves the desired root-n consistency of the proposed EBW estimator $\hat{\tau}_{\mathbf{w}_n^e}$, it unfortunately does not shed light on its asymptotic variance, which is useful for constructing confidence intervals on

τ. We recommend the use of bootstrapped confidence intervals for the EBW estimator. We explore a connection to importance sampling in the Supplementary Material.

3.3 Optimization

The optimization problem (10) for computing EBWs (and the later optimization problem (16) for obtaining three-way EBWs) can be viewed as quadratic programs with linear (in)equality constraints. There has been much work on efficient algorithms for solving such programs [41], including interior point methods [42], augmented Lagrangian techniques [43], and extensions of the simplex algorithm [44]. A recent development is the operator splitting approach in Stellato et al. [45], which provides a reliable alternative for nonpositive definite quadratic programs.

In our implementation, we made use of well-maintained interior-point cone programming solvers in the `R` package `cccp` [46] for optimizing the EBW formulations (10) and (16). Such solvers follow a two-stage procedure: finding an initial feasible solution for $\mathbf{w}$, then iteratively refining this solution by traversing the interior of the feasible region. Detailed algorithmic steps can be found in Wright et al. [47]. Interior-point algorithms enjoy empirical and theoretical advantages for solving large-scale quadratic programs (see further discussion in Gondzio [48] on its scalability and complexity), and we have observed excellent performance of such algorithms for EBW optimization. In practice, we recommend that the covariates should be normalized so that each covariate has mean zero and unit variance, before being used as inputs for the optimization problem. An efficient implementation of these methods for EBW optimization is provided in our `R` package `ebw`, which will be released on comprehensive R archive network in the near future.

3.4 Controlling weight variability

In our experience, the weights resulting from the optimization criterion (10) rarely, if ever, result in large weights; however, in the literature there has been an emphasis on methods that afford explicit control on the variability of weights [33]. To allow for such within our framework, one can simply add a penalty $\lambda n^{-2}\sum_{i=1}^{n} w_i^2$ to the criterion (10) for some fixed $\lambda > 0$. This penalty can be thought of as a penalty on the inverse of the effective sample size of the weights [33]. In our experience, since the energy distance criterion is not nearly as variable as using a kernelized distance using a universal kernel such as the Gaussian kernel, careful tuning of λ is not at all critical and we advocate for a simple choice such as $\lambda = 1$ or not using any penalty at all. On the other hand, if one were to replace the energy distance with a kernelized distance using a universal kernel, the choice of λ becomes critical. In the Supplement, we show analogs of Theorems 3.1 and 3.4 hold when using a penalized energy distance criterion, although the conditions regarding the magnitude of the weights in the latter theorem can be relaxed.

3.5 Three-way EBWs

The EBWs in (10) are designed to balance the distributions of covariates between each treatment group to that of the combined sample $\{\mathbf{X}\}_{i=1}^{n}$. As such, the treatment group should be asymptotically balanced to the control group. Yet for finite samples, EBWs do not necessarily guarantee good distributional balance between the treatment and control arms, which can be important in practice. Consider the following re-expression of the two terms in (3):

$$\int_{\mathbf{x}\in\mathcal{X}} [\mu_1(\mathbf{x}) + \mu_0(\mathbf{x})]\mathrm{d}[F_{n,1,\mathbf{w}} - F_{n,0,\mathbf{w}}](\mathbf{x}) \tag{14}$$

$$- \int_{\mathbf{x}\in\mathcal{X}} \mu_1(\mathbf{x})\mathrm{d}[F_n - F_{n,0,\mathbf{w}}](\mathbf{x}) + \int_{\mathbf{x}\in\mathcal{X}} \mu_0(\mathbf{x})\mathrm{d}[F_n - F_{n,1,\mathbf{w}}](\mathbf{x}). \tag{15}$$

Terms (14) and (15) shed light on how the choice of $\mathbf{w}$ impacts estimation error for the ATE. In particular, this error term depends not only on (i) how close the weighted ECDFs $F_{n,0,\mathbf{w}}$ and $F_{n,1,\mathbf{w}}$ are to that of the combined sample F_n but also (ii) how close the weighted ECDF for the control group $F_{n,0,\mathbf{w}}$ is to the weighted ECDF of the treated group $F_{n,1,\mathbf{w}}$. The EBWs in Section 3 take care of the imbalance in (i), but not the imbalance in (ii) between treatment and control. With finite samples, imbalance in (ii) can result in (14) dominating (15), depending on the properties of $\mu_0(\mathbf{x})$, $\mu_1(\mathbf{x})$, and $\mu_0(\mathbf{x}) + \mu_1(\mathbf{x})$. Indeed, the importance of this *three-way balance* is recognized in the literature, and is emphasized in Chan et al. [14] as an important component in constructing globally efficient estimators based on moment balancing. In our case, the target is the three-way *distributional* balance, i.e., balance in (i) and (ii).

We propose an extension of EBWs, the *improved EBWs* (iEBWs), to help improve covariate balance between the treatment and control groups. These improved weights are defined as follows:

$$\begin{aligned} \mathbf{w}_n^{ei} \in \operatorname*{argmin}_{\mathbf{w}} &\{\mathcal{E}(F_{n,1,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,0,\mathbf{w}}, F_n) + \mathcal{E}(F_{n,0,\mathbf{w}}, F_{n,1,\mathbf{w}})\} \\ &\text{s.t. } \sum_{i=1}^n w_i A_i = n_1, \sum_{i=1}^n w_i(1-A_i) = n_0, w_i \geq 0 \text{ for } i = 1, \ldots, n, \end{aligned} \tag{16}$$

where

$$\begin{aligned} \mathcal{E}(F_{n,1,\mathbf{w}}, F_{n,0,\mathbf{w}}) \equiv{}& \frac{2}{n_1 n_0} \sum_{i=1}^n \sum_{j=1}^n w_i w_j A_i (1-A_j) \|\mathbf{X}_i - \mathbf{X}_j\|_2 \\ &- \frac{1}{n_1^2} \sum_{i=1}^n \sum_{j=1}^n w_i w_j A_i A_j \|\mathbf{X}_i - \mathbf{X}_j\|_2 - \frac{1}{n_0^2} \sum_{i=1}^n \sum_{j=1}^n w_i w_j (1-A_i)(1-A_j) \|\mathbf{X}_i - \mathbf{X}_j\|_2 \end{aligned}$$

is the energy distance between the weighted ECDFs for treated and control. Thus, the iEBWs $\mathbf{w}_n^{ei}$ aim to minimize imbalance not only between treatment arms and the full population but also between the treatment arm and the control arm. Note that $\mathcal{E}(F_{n,1,\mathbf{w}}, F_{n,0,\mathbf{w}})$ still retains the properties of a weighted energy distance, in the sense that $\mathcal{E}(F_{n,0,\mathbf{w}}, F_{n,1,\mathbf{w}}) = \int_{\mathbb{R}^p} |\varphi_{n,1,\mathbf{w}}(\mathbf{t}) - \varphi_{n,0,\mathbf{w}}(\mathbf{t})|^2 \omega(\mathbf{t}) \mathrm{d}\mathbf{t}$, as Proposition 2.1 can be trivially extended to the case where both arguments of the energy distance are weighted.

4 Extensions and applications of EBWs

4.1 Estimation of the ATT

A common target of estimation is the (population) ATT, $\tau^{(1)} \equiv \mathbb{E}(Y(1) - Y(0) \mid A = 1)$, which is the mean difference in potential outcomes among those who are actually treated. Due to the unconfoundedness assumption, we can write

$$\tau^{(1)} = \int_{\mathbf{x}\in\mathcal{X}} [\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})] \mathrm{d}F_1(\mathbf{x}), \tag{17}$$

which suggests that a plug-in estimator can be obtained by replacing $F_1(\mathbf{x})$ with a suitable energy-weighted ECDF. To do so, we define the new weights $\mathbf{w}_n^{e1} = (w_1^{e1}, \ldots, w_n^{e1})$, where $w_i^{e1} = 1$ for $\{i : A_i = 1\}$ and

$$\begin{aligned} \{w_i^{e1}\}_{i:A_i=0} \in \operatorname*{argmin}_{\mathbf{w}} &\mathcal{E}(F_{n,0,\mathbf{w}}, F_{n,1}) \\ &\text{s.t. } \sum_{i=1}^n w_i(1-A_i) = n_0, w_i \geq 0 \text{ for } i = 1, \ldots, n, \end{aligned} \tag{18}$$

where $F_{n,1}(\mathbf{x}) = \frac{1}{n_1}\sum_{i=1}^{n} A_i I(\mathbf{X}_i \le \mathbf{x})$. Thus, $\mathbf{w}_n^{e1}$ balances the covariate distribution for the control group to the covariate distribution of the treated group. Similar to Theorem 3.1, we have $\lim_{n\to\infty} F_{n,0,\mathbf{w}_n^{e1}}(\mathbf{x}) = F_1(\mathbf{x})$ a.s. for every continuity point $\mathbf{x} \in \mathcal{X}$. With the new weights $\mathbf{w}_n^{e1}$, the natural plug-in estimate for the ATT $\tau^{(1)}$ is $\hat{\tau}_{\mathbf{w}_n^{e1}}$ (i.e., the weighted estimator in (2) with $\mathbf{w} = \mathbf{w}_n^{e1}$).

4.2 Multi-category treatments

EBWs can also be constructed for multi-category treatments. When the treatment A_i takes multiple values, i.e. $A_i \in \mathcal{A} = \{a_1, \ldots, a_K\}$. Denote $n_a = \sum_{i=1}^{n} I(A_i = a)$, for $a \in \mathcal{A}$. Each unit has K potential outcomes $(Y_i(a_1), \ldots, Y_i(a_K))$, one for each level. The unconfoundedness assumption in this setting becomes $\{Y_i(a_1), \ldots, Y_i(a_K)\} \perp\!\!\!\perp A \mid \mathbf{X}$. Following Lopez and Gutman [49], the positivity assumption required is now $0 < \pi(a, \mathbf{x}) = \mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x}) < 1$ for all $a \in \mathcal{A}$ and all possible $\mathbf{x} \in \mathcal{X}$. The standard IPW estimator for $\mathbb{E}[Y(a)]$ involves inverse weighting each sample i by $\pi(A_i, \mathbf{X}_i)$. Instead, we propose to estimate $\mathbb{E}[Y(a)]$ or any causal contrast $\tau^{(a-a')} \equiv \mathbb{E}[Y(a) - Y(a')]$ for $a, a' \in \mathcal{A}$ with EBWs.

We define the EBWs for the multiple treatment case as follows:

$$\mathbf{w}_n^{em} \in \underset{\mathbf{w}=(w_1,\ldots,w_n)}{\operatorname{argmin}} \sum_{a\in\mathcal{A}} \mathcal{E}(F_{n,a,\mathbf{w}}, F_n)$$
$$\text{s.t.} \sum_{i=1}^{n} w_i I(A_i = a) = n_a \text{ for all } a \in \mathcal{A} \text{ and } w_i \ge 0 \text{ for } i = 1, \ldots, n.$$

Improved EBWs, which encourage covariate balance between all pairs of treatment options, can be defined similarly as (16), where an additional weighted energy distance between each pair of treatment options is added to the objective. Given any two treatment options $a, a' \in \mathcal{A}$, we can then estimate the causal contrast $\tau^{(a-a')}$ with

$$\hat{\tau}_{\mathbf{w}_n^{em}}^{(a-a')} = \frac{1}{n_a}\sum_{i=1}^{n} w_i Y_i I(A_i = a) - \frac{1}{n_{a'}}\sum_{i=1}^{n} w_i Y_i I(A_i = a').$$

4.3 Estimation of ITRs

As many treatments exhibit heterogeneous effects for different patients, there is great interest in tailoring treatment decisions to patients. A main line of work in this area is the development of statistical methods aimed at finding an optimal ITR, which maps patient characteristics to treatment decisions. Thus, the immediate goal is to estimate a mapping $d : \mathcal{X} \mapsto \{0, 1\}$ which optimizes the expected potential outcomes under the distribution induced by d. Following Qian and Murphy [23], and Zhao et al. [24] and assuming that larger values of the outcome are preferred, the optimal ITR is defined as

$$d^* \in \underset{d}{\operatorname{argmax}}\, \mathbb{E}[Y(d)] = \underset{d}{\operatorname{argmax}}\, \mathbb{E}\left[\frac{YI(A = d(\mathbf{X}))}{\pi(A, \mathbf{X})}\right] = \underset{d}{\operatorname{argmin}}\, \mathbb{E}\left[\frac{YI(A \ne d(\mathbf{X}))}{\pi(A, \mathbf{X})}\right], \tag{19}$$

where $\pi(a, \mathbf{X}) = \mathbb{P}(A = a \mid \mathbf{X})$ and the second equality holds due to the causal assumptions outlined in Section 2.1. The optimal ITR d^* has the property that $d^*(\mathbf{x}) = a \;\Rightarrow\; \mu_a(\mathbf{x}) > \mu_{1-a}(\mathbf{x})$. The last term in (19) appears as a weighted classification task due to the weighted 0–1 loss. With observed data, the objective becomes to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{\pi(A_i, \mathbf{X}_i)} I(A_i \ne d(\mathbf{X}_i)). \tag{20}$$

Due to the nonconvexity of (20), in practice $I(A_i \ne d(\mathbf{X}_i))$ is replaced with a surrogate, such as the hinge function $(1 - (2A_i - 1)d(\mathbf{X}_i))^+$ (see, e.g., the outcome weighted learning (OWL) method of Zhao et al. [24]).

Table 1: Displayed are the value functions and misclassification rates for the optimal ITR estimation example averaged over 1,000 independent simulated datasets

	Scenario 1			Scenario 2		
Method	**Value (SD)**		**Misclass**	**Value (SD)**		**Misclass**
OWL (EBW)	3.168	(0.253)	0.283	2.921	(0.166)	0.228
OWL (iEBW)	**3.198**	(**0.204**)	**0.277**	**2.931**	(**0.165**)	**0.223**
OWL (PS)	2.671	(0.327)	0.344	2.706	(0.196)	0.295

In Scenario 1, the optimal value is 4.74 with 56% of units with optimal assignments of $A = 1$; in Scenario 2, the optimal value is 3.19 with 50% of units with optimal assignments of $A = 1$. The bold values indicate the best performance across all methods for a given setting.

Yet, this objective still requires estimation of the propensity score $\pi(\mathbf{X})$ and involves plugging in this estimate to (20), which subjects it to the same issues of propensity score weighted estimates of τ. To see how EBWs can be used in place of $\pi(A, \mathbf{X})$, we can express (19) as $\text{argmin}_d \int_{\mathbf{x}\in\mathcal{X}} \sum_{a\in\{0,1\}} \mu_a(\mathbf{x}) I(d(\mathbf{x}) \neq a) \mathrm{d}F(\mathbf{x})$. Thus, $d^*(\mathbf{x})$ is a functional of $F(\mathbf{x})$. This suggests a plug-in estimator by replacing $F(\mathbf{x})$ with energy-weighted ECDFs (either $F_{n,a,\mathbf{w}_n^e}(\mathbf{x})$ or $F_{n,a,\mathbf{w}_n^{ei}}(\mathbf{x})$). Thus, we propose to estimate the optimal ITR as follows:

$$\hat{d}^* \in \underset{d}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} Y_i w_i^e \phi(A_i, d(\mathbf{X}_i)), \tag{21}$$

where w_i^e could be replaced with improved weights w_i^{ei} and $\phi(a, d)$ is a convex surrogate for $I(a \neq d)$, such as the hinge function or logistic loss. As in Zhao et al. [24], to prevent overfitting, a penalty $\lambda_n ||d||^2$ can be added to (21) to control complexity of the estimated ITR.

To demonstrate the effectiveness of using EBWs 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. [24], 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 nonlinear terms in the predictors. Full details of the experiment are given in the Supplementary Material. We utilize the OWL method to obtain estimates $\hat{d}$, which uses inverse weighting via the propensity score and adds a penalty $\lambda_n ||d||^2$ 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 $d^*(\mathbf{X})$ on the independent dataset. The results are given in Table 1, which indicates that EBWs can yield better performing ITRs.

5 Simulation studies

To evaluate the finite sample performance and operating characteristics of our proposed estimators, we conducted a large-scale simulation study across a wide variety of data-generating scenarios. Since existing techniques, such as empirical calibration balancing and the covariate balancing propensity score, work exceedingly well when the correct moments are specified to be balanced, we primarily consider simulation settings where the relationships between covariates and both the treatment status and outcome regression model are nonlinear. We consider a wide range of scenarios for the propensity score and outcome regression models, several of which are examples taken from the existing works. Outcome models B and E are taken from [18], outcome model D is taken from Wong and Chan [50], and outcome model A is a slight modification of an

outcome model from Kang and Schafer [11]. Outcome model C is designed to be linear in $\mathbf{X}$ for the untreated group, but nonlinear and with interaction terms for the treated group. Propensity model III is from Kang and Schafer [11] but with the main effects having twice the dimension as by the Kang and Schafer [11] example. The remainder of the propensity models have interactions, smooth nonlinear terms, or nonsmooth nonlinear terms. We generate a p-dimensional covariate vector $\mathbf{Z} = (Z_i, \ldots, Z_p)$ from a p-dimensional mean zero multivariate Gaussian distribution with $\mathrm{Cov}(Z_i, Z_j) = -0.75^{|i-j|}$, which results in positive and negative correlations between predictors. In covariate setup 1, we let the observed covariates be $\mathbf{X} = \mathbf{Z}$; however, similar to the widely-tested setup by Kang and Schafer [11], in covariate setup 2, we define $\mathbf{X} = (X_1, \ldots, X_p)$, where $X_1 = \exp(Z_1/2)$, $X_2 = Z_2/(1 + \exp(Z_1)) + 10$, $X_3 = (Z_1 Z_3/25 + 0.6)^3$, $X_4 = 20 + (Z_2 + Z_4)^2$, $X_5 = \exp(Z_5/2)$, $X_6 = Z_6/(1 + \exp(Z_5)) + 10$, $X_7 = (Z_1 Z_7/25 + 0.6)^3$, and $X_8 = 5 + (Z_6 + Z_8)^2$. To align with the setup of Kang and Schafer [11], we utilize covariate setup 2 for any setting with propensity model III. To align with the setup of Wong and Chan [18], we utilize covariate setup 2 for any setting with outcome model B or E. All other settings use covariate setup 1.

We compare our proposed EBW and the iEBWs (16), both with no penalty on the squares of the weights, with several widely used alternatives. First, we compare with inverse propensity score weights (denoted "IPW"). In addition, we compare with the covariate balancing propensity score weights (denoted "CBPS"), the empirical calibration balancing weights (denoted "Cal") with exponential tilting weights. For all methods that require specification of a model for the treatment assignment or moments to balance, only first-order terms in $\mathbf{X}$ were used, as in Wong and Chan [18]. This reflects a setting where the analyst misspecifies that only first moments should be balanced. We also utilize the kernel-based functional covariate balancing approach of Wong and Chan [18], denoted as "KCB" for kernel covariate balancing with the second-order Sobolev kernel. As a baseline for comparison, we investigate a naïve unweighted estimator that simply compares the means between treated groups and denote this as "Unweighted." For all estimators, we use weights normalized by treatment group. Thus, the IPW approach is the Hájek estimator as opposed to the more unstable, nonnormalized Horvitz-Thompson estimator. For each setting, we generate 1,000 independent datasets and evaluate each method based on the square root of the mean-square error (RMSE) and bias in estimating the ATE. Simulations for IPW, Cal, and CBPS are conducted using the `WeightItR` package [51] and that of KCB using the `ATE.ncb` package.

For the sake of brevity of presentation, we present a summary of the results across all outcome models. More detailed results are presented in the Supplementary Material. Table 2 contains a summary of the results averaged across outcome models (A–E) 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." The Supplementary Material contains a similar summary, but averaged

Table 2: Displayed are the ranks among all methods tested of each method in terms of RMSE and bias averaged over all response models (I–VI) for $n = 250$ and over the dimension settings $p = 10$ and $p = 25$. The bold values indicate the best performance across all methods for a given setting

Propensity model:	**I**		**II**		**III**		**IV**		**V**		**VI**	
	Mean rank		**Mean rank**		**Mean rank**		**Mean rank**		**Mean rank**		**Mean rank**	
Method	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**
Unweighted	4.7	3.7	6.5	4.9	5.0	4.5	4.8	4.1	6.2	5.3	3.9	3.1
EBW	2.5	2.8	2.4	2.7	3.2	4.2	2.5	3.6	**1.9**	2.6	2.5	3.8
iEBW	**2.2**	**2.5**	**1.5**	**1.7**	**3.0**	3.9	**1.7**	3.5	**1.9**	**2.0**	**2.0**	3.1
KCB	3.0	3.4	2.5	2.5	3.2	4.3	2.7	**2.5**	3.0	3.6	2.5	**2.3**
IPW	5.5	5.0	6.4	5.9	5.0	**2.9**	6.3	5.4	5.8	5.1	6.8	6.0
CBPS	5.1	4.8	5.1	5.3	4.7	3.9	5.1	3.7	5.6	5.3	4.9	4.2
Cal	5.0	5.8	3.6	5.0	3.9	4.3	4.9	5.2	3.6	4.1	5.4	5.5

over propensity models and dimensions. In general, iEBW tends to yield the best rank in terms of performance across the settings, with EBW and KCB yielding similarly low ranks.

6 RHC data

6.1 Description of data

A study by Connors et al. [2] was conducted to investigate the effectiveness of RHC, a diagnostic procedure for critically ill patients in ICUs. Since RHC is more relevant for certain forms of intensive care than others, there is substantial imbalance in patient characteristics in those treated with RHC and those who did not receive RHC. The original analysis was based on propensity score matching, and the data have been subsequently re-analyzed in many other works [7,20,52,53]. The study consists of data on 5,735 individuals, 2,184 of whom received RHC, and the remaining 3,551 did not receive RHC. The outcome is an indicator of survival at 30 days after admission. A panel of experts convened to discuss which variables contribute to a decision to use RHC, resulting in a large set of covariates to study (72 in total, 21 of which are continuous, 25 binary, and 26 dummy variables originating from 6 categorical covariates). The dataset is publicly available at: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/rhc.html. There are substantial empirical differences in the distributions of many of these covariates between treatment groups (RHC vs. no RHC). In Section 6.2, we study the effect of RHC on 30-day survival. However, since there is no ground truth available, in Section 6.3, we use the RHC data to conduct a realistic simulation that demonstrates the effectiveness of EBWs.

6.2 Analysis of RHC data

We used 65 of the available covariates, as in the analysis of the same dataset in Rosenbaum [53], leaving out date-related covariates. Using the 65 covariates, we applied the weighting methods used in Section 5 (except the method of Wong and Chan [18] as the code returned constant weights of 1 regardless of the tuning parameters used) to estimate weights to balance the treated groups. To first investigate how well each method balances the marginal means of each covariate, we evaluate the absolute standardized mean differences for each covariate and p-values for the difference in weighted means between treated and control groups, which are displayed in Figure 3. Empirical calibration balancing is designed to balance all specified moments exactly, and since we balance the first moments, these have exact balance. Both EBWs still result in tight moment balance despite the fact that this was not an explicit goal. In Figure 4, we investigate how well each method balances the distributions of all 1- and two-dimensional projections of the covariates. To do this, we compare the weighted ECDFs for each projection by evaluating the square root of the integrated mean-square error (RIMSE) of the weighted ECDFs between the two treatment groups. Both EBWs yield the smallest RIMSEs on average. However, it is important to note that lower-order projections of the distribution are not the explicit focus of EBWs and thus, the EBWs also likely balance other aspects of the distributions of covariates between the treatment groups. We display further measures of discrepancy of the covariates between the treatment groups in Table 3. In addition to displaying the average RIMSE values and absolute SMDs, in this table, we display the weighted energy values for each of the weights, including the unweighted energy distances. By definition, EBWs have the minimum weighted energy distances; however, it can be seen that the empirical calibration balancing weights have a relatively small weighted energy distance, indicating its use is likely sensible for the RHC data. The estimates of the ATE of RHC and its standard errors are displayed in Table 3. Standard errors are computed using the nonparametric bootstrap with 1,000 replications. The EBW-based estimates have the smallest standard errors, and, perhaps more interestingly, yield an estimate of the effect of RHC that is slightly less deleterious than estimates in the literature, despite it still being a highly significant effect.

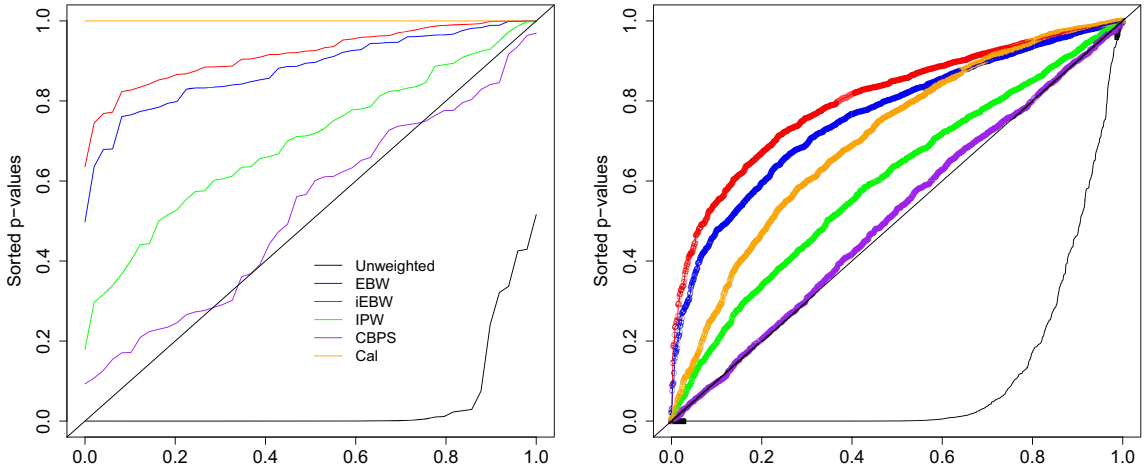

Figure 3: Displayed are p-values for tests of the difference in weighted means between treatment groups marginally for each covariate (left) and for each pairwise interaction of covariates (right). For continuous covariates, weighted t-tests are used and weighted Chi-squared tests are used for discrete covariates. For an RCT, the sorted p-values are expected to roughly fall along the diagonal – lines above the diagonal indicate an improvement in moment balance over random assignment.

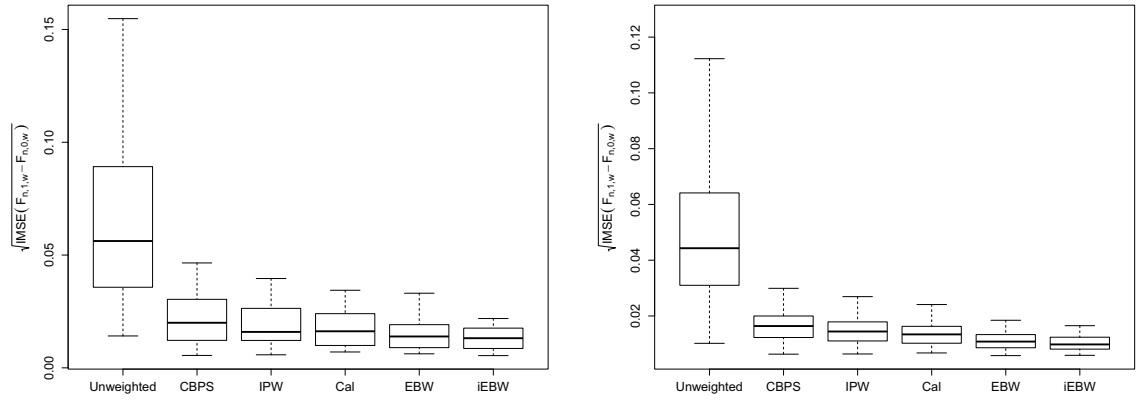

Figure 4: This figure illustrates the differences in the *marginal* univariate weighted ECDFs between the treated and control populations. In particular, let $x_j \in \mathcal{X}_j$ denote the j component of the covariate vector $\mathbf{x}$ and let $F_{n,a,j}(x)$ denote its empirical CDF on treatment arm a. Similarly denote the weighted versions of this quantity. Here, we are displaying how well each method balances the marginal empirical CDFs for the treated and control arms. We do so by evaluating an estimate of $\left\{\int_{x_j \in \mathcal{X}_j} [F_{n,1,j,\boldsymbol{w}} - F_{n,0,j,\boldsymbol{w}}](x_j) \mathrm{d}x_j\right\}^{1/2}$ obtained by integration over a grid of values for all $j = 1, \ldots, 65$. The results across all covariates are displayed in the left two plots above. The rightmost plots similarly display the RIMSEs for all possible 65 choose 2 *bivariate* CDFs.

Table 3: Estimates of the ATE and standard errors for the RHC 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 RHC and non-RHC groups. In addition to weighted energy distances, we display mean and max "|SMD|," which are the average and maximum, respectively, absolute standardized marginal mean difference of covariates

	Unwtd	CBPS	IPW	Cal	EBW	iEBW
$\hat{\tau}_{\mathbf{w}}$	0.0736	0.0576	0.0528	0.0547	0.0499	0.0470
SE($\hat{\tau}_{\mathbf{w}}$)	0.0132	0.0142	0.0155	0.0145	0.0120	0.0117
Energy dist (10)	8.1996	0.8572	0.7804	0.4250	0.3236	0.3377
Energy dist (16)	23.7172	1.8560	1.5756	1.0045	0.7536	0.7270
Mean |SMD|	0.1648	0.0274	0.0182	0.0000	0.0063	0.0043
Max |SMD|	0.5826	0.1139	0.0778	0.0000	0.0239	0.0168
SD |SMD|	0.1276	0.0206	0.0147	0.0000	0.0050	0.0036

6.3 RHC data simulation

In this section, we fix the covariates and treatment assignments from the RHC data, and simulate responses with confounding. This produces a realistic and highly challenging simulation scenario with high dimension. We use the key functional form from outcome model D from Section 5. Since this outcome model involves only 7 covariates, we use the functional form from this outcome model, apply it over 8 separate groups of 7 covariates in the RHC data, and take the sum of all groups of 7 covariates as the mean of the outcome. We then simulate 1,000 independent datasets using this procedure and each time estimate the ATE and record each method's RMSE in estimating the true ATE. Since under this model the ordering of the covariates changes the nature of the confounding, we randomly permute the columns 100 times and repeat the simulation 1,000 times for each permutation and each time record the RMSE over the 1,000 replications. Further details are provided in the Supplementary Material. The RMSEs for each of the 100 column permutations are displayed in Figure 5. Both EBW and iEBW consistently yield among the smallest RMSE across the permutation settings. Since the outcome model in this simulation involves a constant treatment effect, and thus is not impacted by potential issues of covariate overlap, and in the Supplementary Material, we additionally consider an outcome model with a treatment effect that depends on $\mathbf{X}$; the pattern of the results is consistent with the findings using a constant treatment effect, albeit the advantage of EBWs is slightly more pronounced with the heterogeneous treatment effect scenario.

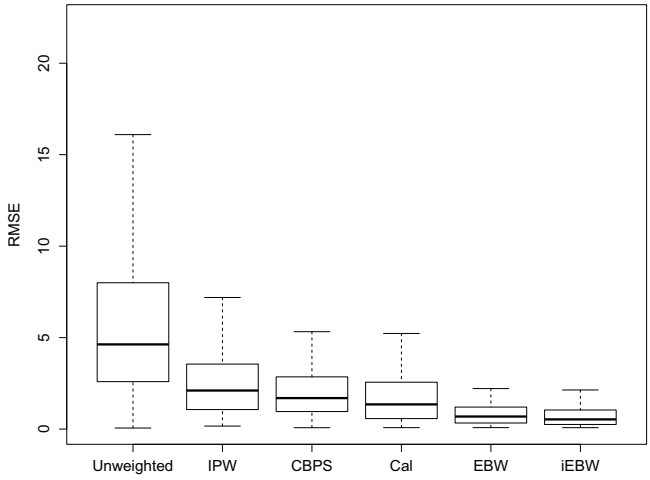

Figure 5: RMSEs for each method across 100 outcome models using the RHC data.

7 Analysis of the MIMIC-III critical care database

We analyze the effectiveness of three treatments based on separate subpopulations of the MIMIC-III v1.4 critical care database [3]: a study of the effect of indwelling arterial catheters (IACs) on mortality in patients with respiratory failure [54], a study of the effect of transthoracic echocardiography (echo) on mortality in sepsis patients [55], and a study of mechanical power of ventilation (MPV) on mortality in critically ill patients [56]. Each study is based on the existing studies utilizing MIMIC-III. The degree of confounding in each study varies, with the IAC and MPV studies exhibiting a great degree of confounding and the echo study with minimal confounding. For all studies, missing values were imputed using missForest [57]. We present the IAC study in this section and the remaining two studies in the Supplementary Material. For each study, we present treatment effect estimates and balance statistics using each method used in the main text. For all studies, we also use the covariates and treatment assignments of the observed data to conduct simulation studies, using the same approach used for the simulation based on the RHC data. In essence, with these simulation studies we preserve the treatment assignment mechanism of the observed data and simulate

outcomes that involve a high likelihood of confounding under this real-world treatment assignment mechanism. The simulation studies investigate scenarios with a constant treatment effect over $\mathbf{X}$ and with a treatment effect that varies with $\mathbf{X}$ but results in a (population and sample) ATE of 0. The outcome models used are the same as described in the Supplementary Material. Each dataset has a differing number of confounders; however, the outcome model across all datasets involves 63 covariates impacting the response for any given scenario. As in the setup for the RHC data simulation, the columns are permuted 100 times, resulting in 100 separate outcome models with different covariates impacting the response in different ways.

7.1 IAC data

In this section, we replicate a study originally conducted Hsu et al. [54] based on the MIMIC-III critical care database to study the effect of indwelling arterial catherization on 28 day mortality. The data are based on the queries provided in https://github.com/MIT-LCP/mimic-codeand contains information on 2,522 mechanically-ventilated patients, 1,298 of whom received the treatment, IAC. The outcome is an indicator of 28 day mortality from time of admission. Pre-treatment covariates likely to be confounders include demographics, lab values, calculated risk scores, missingness indicators, and more, totaling to a design matrix with $p = 81$. We leveraged the same approach used in the RHC study to control for the confounders via various weighting approaches to estimate the ATE of IAC on 28-day mortality. The KCB approach yielded constant weights of 1 regardless of the tuning parameter. Table 4 presents information on the estimated effect of IAC on mortality based on each weighting method in addition to various balance statistics for the 81 covariates, including marginal mean balance, univariate and bivariate distributional balance, and weighted energy statistics. EBW and iEBW yield the best univariate and bivariate distributional balance, while Cal and CBPS result in the best marginal mean balance. Note that Cal does not achieve exact marginal mean balance due to numerical issues. All approaches except inverse weighting by propensity score (IPW) yield 95% confidence intervals that contain zero, albeit with point estimates that suggest a potential benefit of IAC. IPW results in a significant estimated benefit of IAC on 28 day mortality, with a point estimate of the benefit that is substantially larger than for other methods.

In addition, we conduct a simulation study based on the IAC data with precisely the same response-generating mechanism as for the simulation on the RHC data in Section 6.3. Although the dimension of the IAC data is higher, the number of dimensions that impact the response for each data-generating setting is 63, the

Table 4: Estimates of the ATE and standard errors for the IAC 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 IAC 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)). The bold values indicate the best performance across all methods for a given setting

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
$\hat{\tau}_{\mathbf{w}}$	0.0012	−0.0343	−0.0744	−0.0150	−0.0150	−0.0130
$\mathrm{SE}(\hat{\tau}_{\mathbf{w}})$	0.015	0.0217	0.0446	0.0165	0.0118	0.0114
Energy dist (10)	4.5625	0.7829	31.336	0.5612	0.3869	0.4028
Energy dist (16)	13.6796	1.6922	62.6704	1.3854	0.9614	0.9304
Mean RIMSE, 1d	0.0362	0.0130	0.0670	0.0126	0.0113	**0.0111**
Max RIMSE, 1d	0.0863	0.0277	0.1646	0.0256	**0.0235**	0.0245
Mean RIMSE, 2d	0.0416	0.0075	0.0619	0.0071	0.0067	**0.0062**
Max RIMSE, 2d	0.3295	0.0283	0.2028	0.0284	0.0248	**0.0189**
Mean |SMD|	0.0732	0.0023	0.0990	**0.0001**	0.0060	0.0045
Max $\lvert \mathrm{SMD} \rvert$	0.2996	0.0203	1.2906	**0.0028**	0.0289	0.0212
Mean |SMD(2)|	0.0788	0.0081	0.0932	0.0075	0.0093	**0.0073**
Max $\lvert \mathrm{SMD}(2) \rvert$	0.6801	0.1782	1.2906	0.1412	0.1502	**0.0998**

Table 5: Displayed are the median, mean, standard deviation, and maximum RMSEs for each method across the 100 simulation settings using the IAC data. The bold values indicate the best performance across all methods for a given setting

	Unweighted	CBPS	IPW	Cal	EBW	iEBW
	Constant treatment effect					
Median RMSE	8.0151	3.2899	7.9435	3.4895	3.0276	**1.8319**
Mean RMSE	9.1296	3.8542	9.5545	3.7609	3.5113	**2.2087**
SD RMSE	6.7091	2.5588	7.5263	2.5293	2.5028	**1.5785**
Max RMSE	32.3715	11.0479	39.1095	12.0521	12.4231	**7.2066**
	Heterogeneous treatment effect					
Median RMSE	11.9037	3.8587	15.1933	3.7128	3.0732	**1.7355**
Mean RMSE	13.5594	4.4330	18.4590	4.1689	3.9079	**1.8688**
SD RMSE	9.9665	3.1225	14.6490	2.8963	2.8718	**1.3415**
Max RMSE	48.0820	12.3672	75.3899	12.8563	13.8571	**5.6715**

same as in Section 6.3. The results in terms of RMSE in estimating the ATE across the 100 data-generating scenarios for both the constant and heterogeneous treatment effect settings, each averaged over 1,000 replications, are displayed in Table 5. The iEBW approach results in the smallest mean, median, and worst-case RMSE, followed by EBW, which is closely followed by Cal and CBPS. IPW in this case results in nearly worse performance than no weighting. We also conducted the same simulation study but with a treatment effect that varies in $\mathbf{X}$.

8 Discussion

We have introduced a new metric, the weighted energy distance, which measures the distributional balance induced by a set of weights and thus can be used to determine which set of weights is likely to result in low bias when estimating a causal quantity. Building on the weighted energy distance, we have introduced the EBWs that minimize this distance to achieve distributional balance. The energy balancing weights are robust and reliable across many functional forms of confounding and further rarely result in large weights. Due to the distributional balancing of the energy balancing weights, they can be utilized to estimate a wide variety of causal estimands which can be represented as a statistical functional of the population distribution function of the covariates. While we focused entirely on the weighted energy distance, the connection between the energy distance and distances between embeddings of probability measures into reproducing kernel Hilbert spaces [39] opens up the possibility of more effective distributional balancing weights if more is known about the functional form of confounding. In particular, if the analyst believes lower order projections of the distribution should be balanced with priority over higher order aspects of the distribution, the use of a kernel which emphasizes these projections, such as the sparsity-inducing kernel in Mak and Joseph [58], could be used.

Acknowledgements: The authors would like to thank the anonymous referees for their helpful and constructive feedback. Dr. Mak was funded by NSF CSSI Frameworks 2004571, NSF DMS 2316012, and NSF DMS 2316012.

Conflict of interest: Authors state no conflict of interest.

References

[1] Dasgupta T, Pillai NS, Rubin DB. Causal inference from 2K factorial designs by using potential outcomes. J R Stat Soc Ser B (Stat Methodol). 2015;77(4):727–53.

[2] Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, et al. The effectiveness of right heart catheterization in the initial care of critically Ill patients. J Amer Med Assoc. 1996;276(11):889–97.
[3] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035.
[4] Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. J Amer Stat Assoc. 1995;90(429):122–9.
[5] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66:315–31.
[6] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–60.
[7] Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services Outcomes Res Meth. 2001;2(3–4):259–78.
[8] Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003;71(4):1161–89.
[9] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. Rev Econom Stat. 2004;86(1):4–29.
[10] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
[11] Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22(4):523–39.
[12] Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc Ser B (Stat Methodol). 2014;76(1):243–63.
[13] Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis. 2012;20(1):25–46.
[14] 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.
[15] Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. J Amer Stat Assoc. 2015;110(511):910–22.
[16] Zhao Q. Covariate balancing propensity score by tailored loss functions. Ann Stat. 2019 Apr;47(2):965–93.
[17] Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann Math Stat. 1940;11(4):427–44.
[18] Wong RK, Chan KCG. Kernel-based covariate functional balancing for observational studies. Biometrika. 2017;105(1):199–213.
[19] Kallus N. Generalized optimal matching methods for causal inference. J Machine Learn Res. 2020;21(1):2300–53.
[20] Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Amer Stat Assoc. 2018;113(521):390–400.
[21] Hirshberg DA, Maleki A, Zubizarreta JR. Minimax linear estimation of the retargeted mean. 2019. arXiv: http://arXiv.org/abs/arXiv:190110296.
[22] Székely GJ, Rizzo ML. Testing for equal distributions in high dimension. InterStat. 2004;5(1–6):1249–72.
[23] Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat. 2011;39(2):1180.
[24] 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.
[25] Athey S, Imbens GW, Wager S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. J R Stat Soc Ser B (Stat Methodol). 2018;80(4):597–623.
[26] Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. translated in Statistical Science. vol. 1923; 1990. p. 465–472.
[27] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
[28] Rubin DB. Bayesian inference for causal effects: The role of randomization. Ann Stat. 1978;6(1):34–58.
[29] Hernan MA, Robins JM. Causal inference. Boca Raton, FL: CRC; 2019.
[30] Dawid AP. Some misleading arguments involving conditional independence. J R Stat Soc Ser B (Methodological). 1979;41(2):249–52.
[31] Hájek J. Comment on a paper by D. Basu. Foundations of Statistical Inference; 1971. p. 236.
[32] Ding P, Li F. Causal Inference: A Missing Data Perspective. Stat Sci. 2018 May;33(2):214–37.
[33] Chattopadhyay A, Hase CH, Zubizarreta JR. Balancing vs modeling approaches to weighting in practice. Stat Med. 2020;39(24):3227–54.
[34] Székely GJ, Rizzo ML. Energy statistics: A class of statistics based on distances. J Stat Plan Inference. 2013;143(8):1249–72.
[35] Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35(6):2769–94.
[36] Mak S, Joseph VR. Support points. Ann Stat. 2018;46(6A):2562–92.
[37] Genevay A. Entropy-regularized optimal transport for machine learning. Université Paris Dauphine and Ecole Normale Supérieure; 2019.
[38] Weed J, Bach F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli. 2019;25(4A):2620–48.
[39] Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann Stat. 2013 Oct;41(5):2263–91.

[40] Wendland H. Scattered data approximation. vol. 17. United Kingdom: Cambridge University Press; 2004.
[41] Nocedal J, Wright SJ. Numerical optimization. New York, NY: Springer; 1999.
[42] Andersen M, Dahl J, Liu Z, Vandenberghe L, Sra S, Nowozin S, et al. Interior-point methods for large-scale cone programming. Optim Machine Learn. 2011;5583.
[43] Delbos F, Gilbert JC. Global linear convergence of an augmented Lagrangian algorithm for solving convex quadratic optimization problems. J Convex Anal. 2003;12:45–69.
[44] Murty KG, Yu FT. Linear complementarity, linear and nonlinear programming. vol. 3. Berlin, Germany: Citeseer; 1988.
[45] Stellato B, Banjac G, Goulart P, Bemporad A, Boyd S. OSQP: An operator splitting solver for quadratic programs. Math Program Comput. 2020;12:637–72.
[46] Pfaff B. cccp: Cone Constrained Convex Problems; 2015. R package version 0.2-4.
[47] Wright SJ. Primal-dual interior-point methods. United States of America: SIAM; 1997.
[48] Gondzio J. Interior point methods 25 years later. Europ J Operat Res. 2012;218(3):587–601.
[49] Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: a review and new ideas. Stat Sci. 2017;32(3):432–54.
[50] Cannas M, Arpino B. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. Biometric J. 2019;61:1049–72.
[51] Greifer N. WeightIt: Weighting for Covariate Balance in Observational Studies; 2020. R package version 0.9.0.
[52] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96(1):187–99.
[53] Rosenbaum PR. Optimal matching of an optimally chosen subset in observational studies. J Comput Graph Stat. 2012;21(1):57–71.
[54] Hsu DJ, Feng M, Kothari R, Zhou H, Chen KP, Celi LA. The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis. Chest. 2015;148(6):1470–6.
[55] 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.
[56] 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.
[57] Stekhoven DJ, Bühlmann P. MissForest non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.
[58] Mak S, Joseph VR. Projected support points: a new method for high-dimensional data reduction. 2017, arXiv: http://arXiv.org/abs/arXiv:170806897.