Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Independence Weights for Causal Inference with Continuous Treatments

Jared D. Huling, Noah Greifer & Guanhua Chen

To cite this article: Jared D. Huling, Noah Greifer & Guanhua Chen (2023): Independence Weights for Causal Inference with Continuous Treatments, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2213485

To link to this article: https://doi.org/10.1080/01621459.2023.2213485

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2023.2213485

Independence Weights for Causal Inference with Continuous Treatments

Jared D. Huling [a], Noah Greifer [b], and Guanhua Chen [c]

[a]Division of Biostatistics, University of Minnesota, Minneapolis, MN; [b]Institute for Quantitative Social Science, Harvard University, Cambridge, MA; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

Studying causal effects of continuous treatments is important for gaining a deeper understanding of many interventions, policies, or medications, yet researchers are often left with observational studies for doing so. In the observational setting, confounding is a barrier to the estimation of causal effects. Weighting approaches seek to control for confounding by reweighting samples so that confounders are comparable across different treatment values. Yet, for continuous treatments, weighting methods are highly sensitive to model misspecification. In this article we elucidate the key property that makes weights effective in estimating causal quantities involving continuous treatments. We show that to eliminate confounding, weights should make treatment and confounders independent on the weighted scale. We develop a measure that characterizes the degree to which a set of weights induces such independence. Further, we propose a new model-free method for weight estimation by optimizing our measure. We study the theoretical properties of our measure and our weights, and prove that our weights can explicitly mitigate treatment-confounder dependence. The empirical effectiveness of our approach is demonstrated in a suite of challenging numerical experiments, where we find that our weights are quite robust and work well under a broad range of settings. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received March 2022
Accepted April 2023

KEYWORDS
Balancing weights; Confounding; Distance covariance; Electronic health records; Observational data

1. Introduction

Confounding is a major barrier to studying causal effects of treatments or exposures from observational data. Considerable work has focused on the development of approaches for studying causal effects of binary or otherwise discrete-valued treatments from observational data. With continuous treatments, however, the choices of methods for confounding control are far more limited, and clear guidance that can help practitioners choose among the available methods is lacking. A common approach to reduce confounding by observed variables is using the propensity score, which was initially proposed for binary treatments Rosenbaum and Rubin (1983) and has been generalized to the setting of continuous treatments (Hirano and Imbens 2004; Imai and Van Dyk 2004; Galvao and Wang 2015; Zhu, Coffman, and Ghosh 2015; Kennedy et al. 2017). With binary treatments, the causal effect of interest is often the average treatment effect, which can be estimated as a difference in the weighted averages of the treatment group outcomes, where the weights are proportional to the inverse of the propensity score (Robins, Rotnitzky, and Zhao 1994; Robins, Hernan, and Brumback 2000); this method is known as inverse probability weighting (IPW). With continuous treatments, the interest is often in estimation of the causal dose-response functionals (Robins, Hernan, and Brumback 2000; van der Laan and Robins 2003) such as the causal average dose-response function (ADRF), which can be estimated using a weighted regression of the outcome on the treatment, where the weights are proportional to the inverse of the conditional density of the treatment given the covariates, the generalized propensity score (GPS).

In the binary treatment setting, IPW estimators can be unstable due to extreme weights and susceptible to model misspecification (Kang and Schafer 2007; Fan et al. 2021). These issues carry over to IPW estimators for continuous treatments and are substantially more challenging to address. A key reason is that IPW estimation via the GPS requires inverse weighting by a conditional density estimate, not just a conditional probability. Even if the conditional mean of the treatment given covariates is correctly specified, GPS weights can fail to perform well if the *distribution* of the conditional density is misspecified (Naimi et al. 2014). The difficulty of correctly specifying a conditional distribution is exacerbated with increased dimension of the pretreatment covariates to be controlled for, that is, the confounders. The tailored use of flexible machine learning estimation approaches such as that proposed in Zhu, Coffman, and Ghosh (2015) can in many cases yield substantial improvements, however, as shown in our simulations and data analysis, they are still susceptible to the issues of GPS weighting and can perform poorly and/or yield large weights in practice. The ReGPS approach of Colangelo and Lee (2020) directly estimates the inverse GPS without the need for correct specification of a model for the GPS.

As misspecification of a conditional density model can be difficult to diagnose and assess, several works have focused on

CONTACT Jared D. Huling *huling@umn.edu* Biostatistics, University of Minnesota Twin Cities, United States, Minnesota, Minneapolis.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

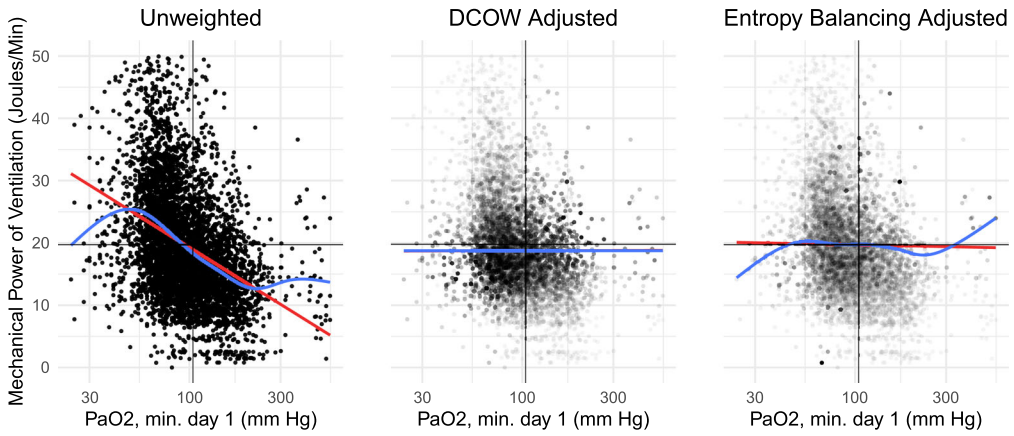

Figure 1. Shown are plots of the relationship between the minimum PaO_2 on day 1 in the ICU (on a log base 10 scale) and the treatment, including an unadjusted plot (left) and plots adjusted by DCOWs and entropy balancing weights (right two plots). In the adjusted plots, the transparency of each point is proportional to its assigned weight, with lighter points indicating less weight. The blue line is a weighted nonparametric regression of the treatment on PaO_2 (on the log base 10 scale) and the red line is a weighted linear regression.

directly estimating weights to reduce the correlation magnitude between marginal moments of (pretreatment) covariates and the treatment. Approaches along this line of work include the generalized covariate balancing propensity score (CBPS) approach of Fong, Hazlett, and Imai (2018) which is an extension of the CBPS approach for discrete treatments (Imai and Ratkovic 2014), covariate association eliminating weights (Yiu and Su 2018), and entropy balancing weights (Vegetabile et al. 2021; Tübbicke 2022). Because these approaches focus on estimation of weights directly as opposed to estimating a conditional density explicitly and inverting it, they tend to be more effective empirically than direct modeling of the GPS. While intuitively appealing, these approaches require careful choices of which moments of *both* the covariates and the treatment to "decorrelate." Yet, there is no guidance on specifying the right moments necessary to mitigate bias in estimation of the ADRF due to confounding. Missing important moments can leave substantial residual dependence between the covariates and treatment, see for example, Figure 1 from our analysis of electronic health record data. Our simulations show that the choice of moments is indeed critically important in practice and that numerical instability can arise when too many moments are used. The tension between including enough moments to reduce bias and the instability of weights as more moments are included can make these methods challenging to use in practice. The general setup of Ai et al. (2021) relies on sieve/series estimators; little finite-sample guidance is provided. Kernel Optimal Orthogonality Weighting (KOW) (Kallus and Santacatterina 2019) and a generalization of KOW proposed in Martinet (2020) are kernel-based nonparametric extensions of direct weights estimation ideas. They focus on estimating weights to decorrelate over function spaces of treatment and covariates such that there is no need to choose models or which moments to decorrelate. Yet, careful tuning is still required when flexible kernels are used, and, on the other hand, when inflexible kernels are used, there is no guarantee that the resulting weights fully mitigate confounding. Thus, the user is often left with a difficult choice of which kernel to use and unclear guidance on how to choose the kernel's tuning parameters. Further, no theoretical justification of the approach of Kallus and Santacatterina (2019) is provided. The theoretical results of Martinet (2020) are limited to the convergence of the weighted distribution functions and do not explore properties involving estimation of the ADRF. More extensive discussion of the existing literature can be found in the Section E of the supplementary material.

Our work aims to achieve several goals. First, we provide clarity on the role of weights in estimation of the ADRF. To do so, we provide a general decomposition of the error of a weighted nonparametric estimator of the ADRF and demonstrate that, under broad conditions, the ideal weights should induce complete *independence* between the treatment variable and pretreatment covariates to guarantee mitigation of confounding bias when estimating the ADRF. While already intuitively understood in the literature, our decomposition precisely quantifies the impact of this dependence on the estimation error in finite samples. We also show that dependence plays a key role in other estimands, such as the causal quantile dose-response function. Second, we develop a measure based on energy statistics (Székely, Rizzo, and Bakirov 2007; Székely and Rizzo 2013) that allows one to assess how well a set of weights is able to induce independence, where smaller values of our measure indicate the weights yield less treatment-covariates dependence and a value of zero indicates complete independence between the treatment and covariates in the weighted data. Huling and Mak (2020) developed a weighted energy distance to mitigate distributional imbalance of covariates in a discrete-treatment setting. In their setting, the energy distance is used to measure distributional imbalance of covariates between different treatment groups, whereas in the setting of this work, a modified distance covariance is used to measure statistical *dependence* between treatment and confounders. Distributional imbalance plays an important role in confounding control with discrete treatments, but this concept does not naturally generalize to continuous treatments. However, removing statistical dependence between discrete treatments and confounders implies distributional balance has been achieved. As such, our work involves a more general notion of confounding control that can in principle be applied to discrete treatments.

Finally, we propose a new approach for estimating weights, which we call the distance covariance optimal weights (DCOWs), by optimizing our measure. The proposed weights

directly aim to mitigate dependence between the treatment and confounders; our error decomposition illustrates that the DCOWs reduce finite sample dependence and thus source of error due to confounding in a weighted nonparametric estimate of the ADRF. In other words, the DCOWs aim to create a pseudo population where treatment and confounders are statistically independent. Our weight construction approach does not require modeling a conditional density, careful tuning of hyperparameters, or choosing which moments of covariates and treatment to decorrelate, making it readily accessible and easy to use for practitioners with varying degrees of statistical sophistication.

We provide some theoretical results for our proposal, showing that our weights indeed reduce dependence between treatment and covariates and fully induce independence asymptotically. Further, we show that with a small penalty on the variability of the weights, our proposal results in the same convergence rate as a nonparametric regression estimate of the ADRF in a scenario with no confounding. Although adding a penalty to reduce weight variability involves the inclusion of a tuning parameter, our proposed approach rarely results in weights with large variability even without penalization. Careful tuning of the parameter that controls weight variability is rarely necessary, as evidenced by our simulation studies, which investigate a wide variety of scenarios with strong and complex confounding and scenarios with moderately high-dimensional confounding, in all of which we fix the tuning parameter to its default value.

Our proposed weights can be used beyond simple weighted nonparametric estimators of the ADRF. Kennedy et al. (2017) and Díaz and van der Laan (2013) extended the idea of doubly robust estimation to continuous treatments, allowing for estimates that combine outcome regression models (Imbens 2004; Hill 2011) and the conditional density models. This allows for relaxed dependence on the correctness of the regression and conditional density models. However, doubly robust estimators are not immune to highly variable weights and their finite sample performance can suffer if the conditional density model is misspecified. We show that our weights can enhance doubly robust estimators. Pairing a reasonable outcome regression model with our weights in a doubly robust fashion can be effective in estimating the ADRF.

The remainder of this article is organized as follows. We investigate the role of dependence between the treatment and covariates in the estimation of the ADRF in Section 2 and we develop a criterion that assesses how much dependence is mitigated by a set of weights, propose a new weight estimation strategy, and provide some corresponding theory in Section 3. We demonstrate the effectiveness of our approach in finite samples with a suite of challenging simulation studies in Section 4 and illustrate the use of our approach in a real-world study of electronic health record data in Section 5. We conclude the article with some discussion.

2. Confounding, Weighting, and Dependence

2.1. Setup, Notation, and Assumptions

The observable quantities we consider consist of the random triplet $(\mathbf{X}, A, Y)$, where $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ is a vector of pretreatment covariates, $A \in \mathcal{A} \subseteq \mathbb{R}$ is a continuous-valued treatment variable indicating the assigned dose for a unit, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is an outcome of interest. The variate $(\mathbf{X}, A, Y)$ has a joint distribution $F_{\mathbf{X},A,Y}$ with respect to a dominating measure. We denote the marginal density of the treatment and covariates as $f_A(a)$ and $f_{\mathbf{X}}(\mathbf{x})$, respectively, the conditional density of the treatment given $\mathbf{X}$ as $f_{A|\mathbf{X}}(a|\mathbf{x})$, and their joint density as $f_{\mathbf{X},A}(a, \mathbf{x})$. Similarly, corresponding distribution functions are denoted $F_A(a) = \mathbb{P}(A \leq a)$, $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$, $F_{A|\mathbf{X}}(a|\mathbf{x}) = \mathbb{P}(A \leq a \mid \mathbf{X} = \mathbf{x})$, and $F_{\mathbf{X},A}(\mathbf{x}, a) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}, A \leq a)$. Our observed data consists of n i.i.d. samples $(\mathbf{X}_i, A_i, Y_i)_{i=1}^n$ from $(\mathbf{X}, A, Y)$. Note that we drop the subscripts on the density and cumulative distribution functions when there is no ambiguity.

We work under the potential outcomes framework, wherein the potential outcome function $Y(a)$ for $a \in \mathcal{A}$ is the outcome that would be observed if A were set to the value a. The causal quantity of interest in this article is the mean potential outcome function, also called the causal *average dose-response function* (ADRF), which is $\mu(a) \equiv \mathbb{E}[Y(a)]$, for $a \in \mathcal{A}$.

The causal ADRF $\mu(a)$ can be identified, or expressed in terms of observational data, under standard causal assumptions. These assumptions, which we employ throughout this article, are (a) *consistency*, which posits that $A = a$ implies $Y = Y(a)$, (b) *positivity*, which states that all values of the treatment are possible across the covariate space in the sense that $f(a|\mathbf{X} = \mathbf{x}) \geq \nu > 0$ for all $\mathbf{x} \in \mathcal{X}$ for some constant ν, and (c) *ignorability* of the assignment mechanism: $Y(a) \perp\!\!\!\perp A \mid \mathbf{X}$ for all $a \in \mathcal{A}$, where $\perp\!\!\!\perp$ denotes (conditional) independence. Under assumptions (a)–(c), the dose-response function is identified as $\mu(a) = \mathbb{E}_{\mathbf{X}}(\mathbb{E}[Y \mid \mathbf{X}, A = a]) = \mathbb{E}_{\mathbf{X}}[\mu(\mathbf{X}, a)]$. Estimation of the dose-response function via regression-based estimation of the mean function $\mu(\mathbf{X}, a) \equiv \mathbb{E}[Y \mid \mathbf{X}, A = a]$, however, can be highly challenging. Misspecification of the regression function can result in poor estimation of $\mu(a)$, and nonparametric estimation of the regression function is also difficult, especially when $\mathbf{X}$ is not low dimensional. Instead, this article focuses on weighting-based estimators of $\mu(a)$. A benefit of weighting estimators is that the dose-response function can be flexibly estimated by univariate (weighted) nonparametric regression, whereas with regression-based estimation, the need to incorporate covariates in a regression may make flexible estimation of the ADRF more difficult due to the additional dimensions. A conceptual benefit of weighting methods over regression-based methods is that they provide a clear separation between design and analysis phases of a study. This separation is critical when substantial or iterative model-building is required to control for confounding.

2.2. Ignorability, Independence, and the GPS

To reliably estimate the causal dose response function using observational data, sources of structural bias should be mitigated, among which the bias due to confounding is the most common. Consider the setting where the covariate vector $\mathbf{X}$ contains all confounders in studying the causal relationship between A and Y. Blocking the backdoor path $A \leftarrow \mathbf{X} \rightarrow Y$ mitigates confounding. One way of blocking the backdoor path is by removing the arrow between A and $\mathbf{X}$ (i.e., making

A independent of $\mathbf{X}$). In a randomized trial, the independence between A and $\mathbf{X}$ holds due to randomization. This motivates us to create a pseudo-population mimicking the one we would observe under such a trial by reweighting the subjects in an observational study such that A is independent of $\mathbf{X}$ in the pseudo-population. Weighting by the generalized propensity score $f_{A|\mathbf{X}}(A \mid \mathbf{X})$ (GPS) (Hirano and Imbens 2004) achieves such and extends the pioneering work of Rosenbaum and Rubin (1983) for binary treatments to continuous treatments.

For continuous treatments, stabilized GPS weights are computed as $f_A(A)/f_{A|\mathbf{X}}(A \mid \mathbf{X})$ (Robins, Hernan, and Brumback 2000), which naturally arise in estimating equations for the dose-response function via semiparametric theory (Kennedy et al. 2017). Estimating the weights requires correct specification not only of the mean of the conditional density of the treatment, but also of its higher order properties such as *shape*. When any of these is misspecified, bias can result in estimates of the ADRF (Naimi et al. 2014; Zhu, Coffman, and Ghosh 2015), indicating a particularly sensitive reliance on correct modeling of the conditional distribution of the treatment given the covariates. Furthermore, similar to standard propensity score weights, GPS modeling can yield extreme weights, leading to unstable estimation. Weight trimming/capping may alleviate the problem of large weights but can be seen as ad hoc and may change the estimand (Crump et al. 2009): the estimated ADRF will correspond to the population represented by the newly weighted sample rather than to the original target population.

The stabilized GPS weights $f_A(A)/f_{A|\mathbf{X}}(A \mid \mathbf{X})$ have several key properties that have motivated work to improve upon the GPS weights. Namely, when weighting by the GPS weights in the population sense, they (a) result in independence of $\mathbf{X}$ and A, (b) preserve the marginal distributions of $\mathbf{X}$ and A, and (c) have mean 1. These properties are listed more explicitly in Section A of the supplementary material. Instead of indirectly estimating the weights by estimating the GPS, a nascent line of work has involved methods which directly estimate weights designed to satisfy the above three properties via balancing criteria. For example, work has focused on estimation of weights that induce zero marginal correlation between treatment and covariates (Fong, Hazlett, and Imai 2018; Yiu and Su 2018; Vegetabile et al. 2021). Although they are more robust than GPS weights, these approaches rely on both the correct choice of moments of the covariates and the choice of moments of the treatment variable to decorrelate. Yet, there is little guidance in deciding which set of moments in both covariates and treatment to focus on, as these choices depend on the form of the true potential outcome-generating model.

The aforementioned three properties of GPS weights are intuitively appealing, but it is not immediately clear which of or in what manner these properties are important in mitigating bias in a weighted nonparametric estimator of the ADRF. To justify which properties are crucial in weights estimation, we derive the relationship between the properties of weighted dose-response function estimators using generic weights and the systematic source of error of the weighted estimator for the ADRF. We demonstrate that the ability of a set of weights to induce independence between $\mathbf{X}$ and A is critical for reducing the bias of a weighted estimator.

2.3. A General Error Decomposition for Weighted Nonparametric Estimators of the ADRF

In this section we aim to provide an explicit mechanistic connection between dependence and bias in weighted estimates of the ADRF. Although it is understood that using weights constructed from a well-estimated conditional density is consistent (Kennedy et al. 2017), it is unclear what role weights play more generally. It is not immediately clear what the connection is between the balancing criteria that aim to "decorrelate" moments of covariates and moments of the treatment variable and the systematic bias in estimating the ADRF. In the following, we investigate the precise source of the systematic bias of an estimator and illuminate the role of the weights in influencing the bias. We focus on weighted Nadaraya-Watson estimators of the ADRF for clarity of presentation, though the key message applies to weighted local polynomial regression and other weighted nonparametric regression.

The response can be expressed as $Y_i = \mu(\mathbf{X}_i, A_i) + \varepsilon_i$, where $\varepsilon_i \equiv Y_i(A_i) - \mu(\mathbf{X}_i, A_i)$. By construction, ε_i have mean zero but are not necessarily identically distributed. Given any set of weights $\mathbf{w} = (w_1, \ldots, w_n)$ and a kernel $K_h(A_i - a_0) = K(\frac{A_i - a_0}{h})/h$ centered at $A = a_0$ with bandwidth $h > 0$, the weighted Nadaraya-Watson (NW) estimator of the ADRF at $A = a_0$ is

$$\widehat{\mu}^{\mathbf{w}}_{NW}(a_0) = \frac{\sum_{i=1}^n Y_i w_i K_h(A_i - a_0)}{\sum_{i=1}^n K_h(A_i - a_0)}. \tag{1}$$

This class of estimators of the causal ADRF is motivated by the identification results of Colangelo and Lee (2020), who showed under certain causal conditions and assumptions regarding the kernel K_h that $\mu(a_0) = \lim_{h\to 0} \mathbb{E}\big[YK_h(A - a_0)/f_{A|\mathbf{X}}(a_0 \mid \mathbf{X})\big]$, which implies the use of the inverse of $f_{A|\mathbf{X}}(A|\mathbf{X})$ as weights, since $\mu(a_0) = \lim_{h\to 0} \mathbb{E}\left[YK_h(A - a_0)w^*(\mathbf{X}, A)\right] / \mathbb{E}\left[K_h(A - a_0)\right] = \lim_{h\to 0} \mathbb{E}\left[YK_h(A - a_0)w^*(\mathbf{X}, a_0)\right] / f_A(a_0)$, where $w^*(\mathbf{x}, a) = f_A(a)/f_{A|\mathbf{X}}(a \mid \mathbf{X} = \mathbf{x})$.

Given *any* weights $\mathbf{w}$, the error of (1) at $A = a_0$ can be decomposed as

$$\begin{aligned}
&\widehat{\mu}^{\mathbf{w}}_{NW}(a_0) - \mu(a_0) \\
&= \int_{\mathcal{X}}\int_{\mathcal{A}} \mu(\mathbf{x}, a_0)\mathrm{d}\left[F^n_{\mathbf{X},A,\mathbf{w}} - F^n_{\mathbf{X}}F^n_A\right](\mathbf{x}, a) \qquad (2)\\
&\quad + \int_{\mathcal{X}} \mu(\mathbf{x}, a_0)\mathrm{d}\left[F^n_{\mathbf{X}} - F_{\mathbf{X}}\right](\mathbf{x}) + \left(\frac{f_A(a_0)}{\widehat{f}^n_{A,h}(a_0)} - 1\right) \\
&\qquad \times \int_{\mathcal{X}} \mu(\mathbf{x}, a_0)\mathrm{d}F^n_{\mathbf{X}}(\mathbf{x}) \\
&\quad + \left(\frac{f_A(a_0)}{\widehat{f}^n_{A,h}(a_0)} - 1\right)\int_{\mathcal{X}}\int_{\mathcal{A}} \mu(\mathbf{x}, a_0)\mathrm{d}\left[F^n_{\mathbf{X},A,\mathbf{w}} - F^n_{\mathbf{X}}F^n_A\right](\mathbf{x}, a) \\
&\quad + \widehat{f}^{n\,-1}_{A,h}(a_0)\int_{\mathcal{X}}\int_{\mathcal{A}} [\mu(\mathbf{x}, a)K_h(a - a_0) \\
&\qquad - \mu(\mathbf{x}, a_0)f_A(a_0)]\,\mathrm{d}F^n_{\mathbf{X},A,\mathbf{w}}(\mathbf{x}, a) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i \widehat{f}^{n\,-1}_{A,h}(a_0)K_h(A_i - a_0), \qquad (3)
\end{aligned}$$

where $\widehat{f}^n_{A,h}(a_0) = \int_{\mathcal{A}} K_h(a - a_0)\mathrm{d}F^n_A(a)$ is a kernel density estimate of $f_A(a_0)$, $F^n_{\mathbf{X}}(\mathbf{x}) =$

$n^{-1}\sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})$ is the empirical cumulative distribution function (CDF) of $\{\mathbf{X}_i\}_{i=1}^n$, $F_A^n(a) = n^{-1}\sum_{i=1}^n I(A_i \leq a)$ is the empirical CDF of $\{A_i\}_{i=1}^n$, and $F_{\mathbf{X},A,\mathbf{w}}^n(\mathbf{x},a) = n^{-1}\sum_{i=1}^n w_i I(\mathbf{X}_i \leq \mathbf{x}, A_i \leq a)$ is the weighted empirical CDF of $\{\mathbf{X}_i, A_i\}_{i=1}^n$ using weights $\mathbf{w}$. We provide a derivation of this decomposition in Section B of the supplementary material. The second term on the right is due to sampling variability only and has mean zero and converges at rate $n^{-1/2}$ if the sample is representative of the super-population. The expectations of the third, fourth, and fifth terms on the right above go to 0 when $h \to 0$ regardless of the weights $\mathbf{w}$. The last term has mean zero regardless of both the weights and the bandwidth, though its variability is impacted by the weights. We note, however, that the third through fifth terms are not guaranteed to converge to zero without additional conditions on the variability of the weights.

On the other hand, the first term (2) on the right above is the source of systematic bias of the weighted estimator unrelated to kernel smoothing. In other words, taking limits of the bandwidth of the kernel to 0 and sample size to infinity does not make (2) vanish. This term also provides insight into why using the weights we propose later performs well in finite sample settings when used in treatment effect estimators, as targeting this term can help decrease the magnitude of the systematic component of the bias of an estimator. If a given set of weights induces finite-sample independence of $\mathbf{X}$ and A in the sense that $F_{\mathbf{X},A,\mathbf{w}}^n(\mathbf{x},a) = F_{\mathbf{X}}^n(\mathbf{x})F_A^n(a)$ for all $\mathbf{x} \in \mathcal{X}$, $a \in \mathcal{A}$, then the source of bias (2) of $\hat{\mu}^{\mathbf{w}}(a_0)$ will be zero. The mean-squared error of the estimator will, however, depend primarily on both the bias term (2) and the variance of (3). Mitigating the variance of (3) merely amounts to controlling the squares of the weights; however, providing a measure that can characterize (2) is nontrivial and none exists in the literature. The term (2) is bounded by the distance between $F_{\mathbf{X},A,\mathbf{w}}^n(\mathbf{x},a)$ and $F_{\mathbf{X}}^n(\mathbf{x})F_A^n(a)$ provided that $\mu(\mathbf{x},a_0)$ is bounded. Without modeling the response function, constructing a measure that bounds (2) is critical for assessing a set of weights.

Remark 1. The role of weights in their ability to induce independence between treatment and covariates is not unique to estimation of the ADRF and applies to a wide variety of estimands. Consider estimation of the causal dose-response quantile function $q_{Y(a_0)}(\alpha) = \inf\{y : F_{Y(a_0)}(y) \leq \alpha\}$, where $F_{Y(a_0)}(y) = \mathbb{P}(Y(a_0) \leq y) = \mathbb{E}_{\mathbf{X}}\left\{\mathbb{P}(Y \leq y|\mathbf{X}, A = a_0)\right\} = \mathbb{E}_{\mathbf{X}}\left\{F_{Y|\mathbf{X},A}(y|\mathbf{X}, A = a_0)\right\}$. By replacing Y_i with $I(Y_i \leq y)$ in (1), we can show that the estimation error of $F_{Y(a_0)}(y)$ also depends on how well weights mitigate dependence between A and $\mathbf{X}$. More details are included in Section B of the supplementary material.

In practice, the estimator $\widehat{\mu}_{NW}^{\mathbf{w}}(a_0)$ in (1) may be unstable, as the weights only appear in the numerator, so the estimated ADRF may lie outside the range of the observed values of the response. Instead, a more stable estimator is the following weighted average of the responses

$$\widehat{\mu}_{NWs}^{\mathbf{w}}(a_0) = \frac{\sum_{i=1}^n Y_i w_i K_h(A_i - a_0)}{\sum_{i=1}^n w_i K_h(A_i - a_0)}, \tag{4}$$

which can be viewed as the minimizer of a weighted least squares problem where the ith weight is $w_i K_h(A_i - a_0)$. The estimator (4) is also a valid estimator of the ADRF as long as the denominator divided by $\sum_{i=1}^n w_i$ is a consistent estimator of $f_A(a_0)$; in this case the key source of systematic bias still depends on the term (2).

3. Measuring and Controlling Weighted Dependence with Energy Statistics

3.1. A Criterion to Evaluate the Quality of a Set of Weights

Having established the relationship between the dependence of A and $\mathbf{X}$ and the error in a weighted nonparametric estimate of the ADRF, we now construct a criterion that can assess how well a given set of weights induces independence on the weighted scale, that is, we aim to characterize and bound the distance between $F_{\mathbf{X},A,\mathbf{w}}^n$ and $F_{\mathbf{X}}^n F_A^n$. We do so by building on the ideas of distance covariance (Székely, Rizzo, and Bakirov 2007), which is a measure of dependence between two random vectors of arbitrary dimensions. The population distance covariance is zero if and only if the vectors are independent. Hence, a *weighted* distance covariance will be a useful component for our measure. Let $\mathbf{w} = (w_1, \ldots, w_n)$ and define the weighted distance covariance to be

$$\mathcal{V}_{n,\mathbf{w}}^2(\mathbf{X}, A) = \frac{1}{n^2}\sum_{k,\ell=1}^n w_k w_\ell C_{k\ell} D_{k\ell}, \tag{5}$$

where

$$c_{k\ell} = \|\mathbf{X}_k - \mathbf{X}_\ell\|_2, \quad \bar{c}_{k\cdot} = \frac{1}{n}\sum_{\ell=1}^n c_{k\ell}, \quad \bar{c}_{\cdot\ell} = \frac{1}{n}\sum_{k=1}^n c_{k\ell},$$

$$\bar{c}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,\ell=1}^n c_{k\ell}, \quad C_{k\ell} = c_{k\ell} - \bar{c}_{k\cdot} - \bar{c}_{\cdot\ell} + \bar{c}_{\cdot\cdot},$$

for $k, \ell = 1, \ldots, n$. Similarly define $d_{k\ell} = |A_k - A_\ell|$, $\bar{d}_{k\cdot} = \frac{1}{n}\sum_{\ell=1}^n d_{k\ell}$, $\bar{d}_{\cdot\ell} = \frac{1}{n}\sum_{k=1}^n d_{k\ell}$, and $D_{k\ell} = d_{k\ell} - \bar{d}_{k\cdot} - \bar{d}_{\cdot\ell} + \bar{d}_{\cdot\cdot}$. The quantity (5) simplifies to the original distance covariance when weights are all 1. Since the original distance covariance is always nonzero, (5) is also always nonzero if the weights are positive. We now provide further insight and motivation of the form of $\mathcal{V}_{n,\mathbf{w}}^2(\mathbf{X}, A)$ and its interpretation in terms of weighted distributions.

Letting $i = \sqrt{-1}$, we define the (weighted) empirical characteristic functions as $\varphi_{\mathbf{X},A,\mathbf{w}}^n(\mathbf{t},s) = \frac{1}{n}\sum_{j=1}^n w_j \exp\{i\mathbf{t}^\mathsf{T}\mathbf{X}_j + isA_j\}$, $\varphi_{\mathbf{X},\mathbf{w}}^n(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^n w_j \exp\{i\mathbf{t}^\mathsf{T}\mathbf{X}_j\}$, $\varphi_{A,\mathbf{w}}^n(s) = \frac{1}{n}\sum_{j=1}^n w_j \exp\{isA_j\}$, and empirical characteristic functions $\varphi_{\mathbf{X},A}^n(\mathbf{t},s)$, $\varphi_{\mathbf{X}}^n(\mathbf{t})$ and $\varphi_A^n(s)$ are defined accordingly.

Theorem 3.1. Let $\mathbf{w} = (w_1, \ldots, w_n)$ be a vector of weights such that $\sum_{i=1}^n w_i = n$ and $w_i \geq 0$ for all $i = 1, \ldots, n$. Then $\mathcal{V}_{n,\mathbf{w}}^2(\mathbf{X}, A) \geq 0$ and

$$\mathcal{V}_{n,\mathbf{w}}^2(\mathbf{X}, A) = \int_{\mathbb{R}^{p+1}} |\varphi_{\mathbf{X},A,\mathbf{w}}^n(\mathbf{t},s) - \varphi_{\mathbf{X},\mathbf{w}}^n(\mathbf{t})\varphi_{A,\mathbf{w}}^n(s) + (\varphi_{\mathbf{X},\mathbf{w}}^n(\mathbf{t}) - \varphi_{\mathbf{X}}^n(\mathbf{t}))(\varphi_{A,\mathbf{w}}^n(s) - \varphi_A^n(s))|^2 \omega(\mathbf{t},s)\mathrm{d}\mathbf{t}\,\mathrm{d}s \tag{6}$$

where $\omega(\mathbf{t},s) = (c_p c_1 \|\mathbf{t}\|_2^{1+p}|s|^2)^{-1}$ with $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$, and $\Gamma(\cdot)$ is the complete gamma function.

Based on (6), it is clear that if $F^n_{\mathbf{X},A,\mathbf{w}} = F^n_{\mathbf{X}}F^n_A$, then $\mathcal{V}^2_{n,\mathbf{w}}(\mathbf{X}, A) = 0$. However, the converse is not necessarily true. Yet, if the weights preserve the marginal distribution of treatment and covariates, that is, $F^n_{\mathbf{X},\mathbf{w}} = F^n_{\mathbf{X}}$ and $F^n_{A,\mathbf{w}} = F^n_A$, then $\mathcal{V}^2_{n,\mathbf{w}}(\mathbf{X}, A) = 0$ implies that $F^n_{\mathbf{X},A,\mathbf{w}} = F^n_{\mathbf{X}}F^n_A$. In other words, if one can add additional terms to (5) that also measure the distance between $F^n_{\mathbf{X},\mathbf{w}}$ and $F^n_{\mathbf{X}}$ along with that between $F^n_{A,\mathbf{w}}$ and F^n_A, then (5) can be used to construct a measure that determines the distance between $F^n_{\mathbf{X},A,\mathbf{w}}$ and $F^n_{\mathbf{X}}F^n_A$, that is, a measure for the weighted dependence between $\mathbf{X}$ and A. We leverage the weighted energy distance proposed in Huling and Mak (2020) to help construct such a measure.

Applied to our setting, the weighted energy distance between $F^n_{\mathbf{X},\mathbf{w}}$ and $F^n_{\mathbf{X}}$ is

$$\begin{aligned}\mathcal{E}(F^n_{\mathbf{X},\mathbf{w}}, F^n_{\mathbf{X}}) \equiv \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n w_i\|\mathbf{X}_i - \mathbf{X}_j\|_2 \\ - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n w_iw_j\|\mathbf{X}_i - \mathbf{X}_j\|_2 \\ - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\|\mathbf{X}_i - \mathbf{X}_j\|_2.\end{aligned}$$

Due to Proposition 1 of Huling and Mak (2020), it can be shown that $\mathcal{E}(F^n_{\mathbf{X},\mathbf{w}}, F^n_{\mathbf{X}}) = \int_{\mathbb{R}^p}|\varphi^n_{\mathbf{X}}(\mathbf{t}) - \varphi^n_{\mathbf{X},\mathbf{w}}(\mathbf{t})|^2\omega(\mathbf{t})\mathrm{d}\mathbf{t}$, 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. The weighted energy distance $\mathcal{E}(F^n_{A,\mathbf{w}}, F^n_A)$ between $F^n_{A,\mathbf{w}}$ and F^n_A can be similarly defined.

Our proposed measure of the level of independence induced by a set of weights is defined as

$$\mathcal{D}(\mathbf{w}) = \mathcal{V}^2_{n,\mathbf{w}}(\mathbf{X}, A) + \mathcal{E}(F^n_{\mathbf{X},\mathbf{w}}, F^n_{\mathbf{X}}) + \mathcal{E}(F^n_{A,\mathbf{w}}, F^n_A). \quad (7)$$

The following result demonstrates that $\mathcal{D}(\mathbf{w})$ indeed achieves its stated goal.

Theorem 3.2. Let $\mathbf{w} = (w_1, \ldots, w_n)$ be such that $w_i > 0$ and $\sum_{i=1}^n w_i = n$. Then $\mathcal{D}(\mathbf{w}) \geq 0$ with equality to zero if and only if $\varphi^n_{\mathbf{X},A,\mathbf{w}}(\mathbf{t}, s) = \varphi^n_{\mathbf{X}}(\mathbf{t})\varphi^n_A(s)$, $\varphi^n_{\mathbf{X},\mathbf{w}}(\mathbf{t}) = \varphi^n_{\mathbf{X}}(\mathbf{t})$, and $\varphi^n_{A,\mathbf{w}}(s) = \varphi^n_A(s)$ for all $(\mathbf{t}, s) \in \mathbb{R}^{p+1}$. Further, $\int|\varphi^n_{\mathbf{X},A,\mathbf{w}}(\mathbf{t}, s) - \varphi^n_{\mathbf{X}}(\mathbf{t})\varphi^n_A(s)|^2\omega(\mathbf{t}, s)\mathrm{d}\mathbf{t}\mathrm{d}s \leq 3\mathcal{D}(\mathbf{w})$.

Thus, smaller values of $\mathcal{D}(\mathbf{w})$ indicate smaller potential for dependence between $\mathbf{X}$ and A after weighting and better preservation of the marginal distributions, and larger values indicate the opposite. $\mathcal{D}(\mathbf{w}) = 0$ implies the weights induce complete independence between $\mathbf{X}$ and A and that the marginal distributions of $\mathbf{X}$ and A are exactly preserved.

We also have the following result, which shows how the proposed distance acts as a bound on integration errors over a class of functions.

Lemma 3.3. Let $\mathcal{H}$ be the native space induced by the radial kernel $\Phi(\cdot,\cdot) = -\|\cdot\|_2 \times -|\cdot|$ on $\mathcal{X}\times\mathcal{A}$ equipped with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|g\|_{\mathcal{H}} = \sqrt{\langle g, g\rangle_{\mathcal{H}}}$ for any $g(\cdot,\cdot) \in \mathcal{H} = \mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{A}}$, where $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{A}}$ are defined in Theorem 4 of Mak and Joseph (2018). Then, for any weights $\mathbf{w}$ satisfying $\sum_{i=1}^n w_i = n$, $w_i \geq 0$, we have

$$\left[\int_{\mathcal{X}}\int_{\mathcal{A}} g(\mathbf{x}, a)\mathrm{d}\left[F^n_{\mathbf{X},A,\mathbf{w}} - F^n_{\mathbf{X}}F^n_A\right](\mathbf{x}, a)\right]^2 \leq C_g\mathcal{D}(\mathbf{w}), \quad (8)$$

where $C_g = 3\|g\|^2_{\mathcal{H}} \geq 0$ is a constant depending on only g.

For any $a \in \mathcal{A}$ if $\mu(\cdot, a) \in \mathcal{H}_{\mathcal{X}}$, we can see that $\mathcal{D}(\mathbf{w})$, modulo a constant, acts as a bound on the systematic bias term (2) as long as $\mu(\cdot, a)$ is sufficiently smooth. Thus, if $\mu(\mathbf{x}, a)$ is contained in $\mathcal{H}$, then we can expect weights with smaller $\mathcal{D}(\mathbf{w})$ to lead to smaller systematic bias. The space $\mathcal{H}$ is a reasonably broad class of functions as it contains the Sobolev space of functions with square-integrable functions with $r < \lceil(p+1)/2\rceil$th differentials (Mak and Joseph 2018; Huling and Mak 2020). Our goal for the next section is to define weights that are optimal in terms of our criterion. The weights that minimize $\mathcal{D}(\mathbf{w})$ will result in mitigation of the dependence of $\mathbf{X}$ and A induced by nonrandom selection into treatment.

In contrast to the measure (7) comprised of the distance covariance term (6), it is natural to wonder whether it would instead be more appropriate to simply define a distance as $\int|\varphi^n_{\mathbf{X},A,\mathbf{w}}(\mathbf{t}, s) - \varphi^n_{\mathbf{X}}(\mathbf{t})\varphi^n_A(s)|^2\omega(\mathbf{t}, s)\mathrm{d}\mathbf{t}\mathrm{d}s$ and construct a relationship between this distance and Euclidean norms computable from data. However, we have found that the resulting quantity can be empirically problematic and unreliable. Further, we have found that *optimization* of such a quantity cannot be achieved reliably by existing algorithms and is thus not suitable for the proactive construction of weights. Our proposed quantity, while more complicated, does not exhibit any of these issues in the sense that it reliably measures dependence, and, as we will demonstrate, it is straightforward to optimize with existing quadratic programming software.

3.2. A New Proposal: Distance Covariance Optimal Independence Weights

We define the *distance covariance optimal weights* (DCOWs) to be

$$\mathbf{w}^d_n \in \underset{\mathbf{w}=(w_1,\ldots,w_n)}{\operatorname{argmin}} \mathcal{D}(\mathbf{w}) \text{ such that } \sum_{i=1}^n w_i = n, \text{ and } w_i \geq 0 \text{ for } i = 1, \ldots, n. \quad (9)$$

The name reflects that the weights are constructed as the optimizers of our distance-covariance-based criterion. Due to Theorem 3.2, the DCOWs $\mathbf{w}^d_n$ are designed to minimize dependence between $\mathbf{X}$ and A on the weighted scale while keeping the weighted marginal distributions of $\mathbf{X}$ and A close to those of the unweighted data. The constraint $\sum_{i=1}^n w_i = n$ ensures that $F^n_{\mathbf{X},A,\mathbf{w}^d_n}$ is a valid distribution function. Since $\mathcal{D}(\mathbf{w})$ tracks with the dependence induced by a set of weights, the DCOWs can be thought of as *optimal independence weights* (i.e., optimal with respect to achieving independence between the treatment and covariates).

Although, as we will show later, the DCOWs result in consistent weighted dose-response estimators, they may not guarantee optimal convergence rates without additional constraints. Instead, a small change to our criterion to include penalization

of the squares of the weights provides control of the variability of the weights without sacrificing their bias-reduction property. This additional penalty is akin to focusing more directly on mean squared error instead of bias and can be interpreted as discouraging the effective sample size (ESS) after weighting from being too small. In particular, the ESS is typically approximated as $(\sum_i w_i)^2/\sum_i w_i^2$ (Kish 1965), which, due to our constraints, is $n^2/\sum_i w_i^2$ and is precisely the inverse of our proposed penalty. Further, combining a "balance" criterion with a means of weight variability mitigation is in line with the recommendations of Chattopadhyay, Hase, and Zubizarreta (2020).

We now define the *penalized distance covariance optimal weights* (PDCOWs) to be

$$\mathbf{w}_n^{pd} \in \underset{\mathbf{w}=(w_1,\ldots,w_n)}{\operatorname{argmin}} \; \mathcal{D}(\mathbf{w}) + \lambda \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 \text{ such that } \sum_{i=1}^{n} w_i = n, w_i \geq 0 \text{ for } 0 < \lambda < \infty, i = 1, \ldots, n. \quad (10)$$

Here, the tuning parameter λ is any positive constant and can be chosen by the user to achieve a desired ESS. A lemma provided in Section F of the supplementary material similar to Lemma 3.3 shows that the penalized version of our criterion acts as a bound on the term in the left-hand side of (8) plus the squares of the weights, which is more akin to a bound on the root mean squared error than bias as in Lemma 3.3. Although having a nonzero, positive value of λ is necessary for the convergence rate guarantee of Theorem 3.5, in practice we have found that minimal or even no penalization at all works well because the unpenalized weights, the DCOWs, tend to be quite stable. In all analyses described later, we use only the DCOWs with no weight penalization.

Both the DCOWs and PDCOWs can be used in a wide variety of estimators for various causal estimands involving continuous treatments, not just Nadaraya-Watson-based estimators and not just estimators of the ADRF. The weights can be used either in a simple weighted nonparametric estimator of the dose-response function or to supplement any doubly robust estimator of such. For the former, our weights provide a fully nonparametric and empirically robust estimation approach that requires only mild moment conditions on the covariates and treatment for estimation consistency. For the latter, such an estimator using our weights is guaranteed to be consistent regardless of the correctness of the outcome model, while it still enjoys efficiency gains if the outcome model is well-specified.

Remark 2. The optimization problems (9) and (10) can be formulated as quadratic programming problems with linear constraints, making them straightforward to implement with commercial and open-source solvers such as OSQP (Stellato et al. 2020). If in practice additional emphasis on correlations of particular moments of covariates and the treatment is of importance, our framework can accommodate that by adding additional linear constraints. The details are deferred to Sections C and D of the supplementary material.

In a later section, we provide more formal statements on the consistency of dose-response function estimators that use our proposed distance covariance optimal weights.

3.3. Asymptotic Properties

We first show that the distance covariance optimal weights do induce complete independence asymptotically. Throughout this section we define $w^*(\mathbf{x}, a) = f_A(a)/f_{A|\mathbf{X}}(a|\mathbf{X} = \mathbf{x})$ to be the "true" normalized density weights.

Theorem 3.4. Let $\mathbf{w}_n^d$ be the distance covariance optimal weights defined in (9). Then if $\mathbb{E}\|\mathbf{X}\|_2 < \infty$ and $\mathbb{E}|A| < \infty$ we have

$$\lim_{n\to\infty} F^n_{\mathbf{X},A,\mathbf{w}^d}(\mathbf{x}, a) = F_{\mathbf{X}}(\mathbf{x})F_A(a) \quad (11)$$

almost surely for every continuity point $(\mathbf{x}, a) \in \mathbb{R}^{p+1}$ and further that $\lim_{n\to\infty} F^n_{\mathbf{X},\mathbf{w}^d}(\mathbf{x}) = F_{\mathbf{X}}(\mathbf{x})$ for every continuity point $\mathbf{x} \in \mathbb{R}^p$ and $\lim_{n\to\infty} F^n_{A,\mathbf{w}^d}(\mathbf{x}, a) = F_A(a)$ for every continuity point $a \in \mathbb{R}$. If, additionally $\mathbb{E}w^{*2}(\mathbf{X}, A) < \infty$ holds, then the same result holds for $\mathbf{w}^{pd}$.

Theorem 3.4 is in some sense the most important property of the DCOWs, as it demonstrates the feasibility of using these weights not just in estimation of the ADRF, but also in the estimation of many causal estimands that require independence. In particular, if the source of confounding bias has the form $\int_{\mathcal{X}}\int_{\mathcal{A}} g(\mathbf{x}, a)\mathrm{d}\left[F^n_{\mathbf{X},A,\mathbf{w}} - F^n_{\mathbf{X}}F^n_A\right](\mathbf{x}, a)$ for some function $g \in \mathcal{H}$, then the use of our weights can be justified due to Lemma 3.3 and Theorem 3.4.

We now show that for the particular task of estimating the ADRF, using the DCOWs in a weighted Nadaraya-Watson estimator results in consistent estimation of the ADRF.

Theorem 3.5. Assume that the kernel $K(\cdot)$ is symmetric, second order, that is, it meets the conditions that $\int uK(u)\mathrm{d}u = 0$, $\int K(u)\mathrm{d}u = 1$, and $0 < \int u^2K(u)\mathrm{d}u < \infty$, and is bounded differentiable. Further, assume that the moment conditions required in Theorem 3.4 hold and that $\mu(\mathbf{x}, a_0)$ and is bounded and continuous on $\mathcal{X} \times \mathcal{A}$ and has second order derivatives, $f_A(a_0)$ is bounded and has second order derivatives, $1/f_A(a_0)$ is uniformly bounded. When $h \to 0$, $nh \to \infty$, then for $\mathbf{w} = \mathbf{w}^d$ and $\mathbf{w} = \mathbf{w}^{pd}$

$$\lim_{n\to\infty} \widehat{\mu}^{\mathbf{w}}_{NW}(a_0) = \lim_{n\to\infty} \widehat{\mu}^{\mathbf{w}}_{NWs}(a_0) = \mu(a_0) \quad (12)$$

in probability for all continuity points $a_0 \in \mathcal{A}$.

Thus, both the DCOWs and PDCOWs result in consistent estimation of the causal ADRF using either the stabilized or unstabilized estimator. It can also be shown (supplementary material Section F.3) that $\widehat{\mu}^{\mathbf{w},DR}_{NW}(a_0)$ is still consistent even if $\widehat{\mu}(\mathbf{x}, a_0)$ is inconsistent for $\mu(\mathbf{x}, a_0)$ as long as $\widehat{\mu}(\mathbf{x}, a_0)$ converges to any finite function uniformly almost surely.

Remark 3. Our distance metric has some relationship with maximum mean discrepancy based distances and the kernel-based independence test via the results in Sejdinovic et al. (2013), where our distance induces a particular kernel $\Phi(\cdot, \cdot)$, defined in our Lemma 3.3. However, despite this connection, our Theorem 3.5 does not require the response function $\mu(\mathbf{x}, a_0)$ to be in the native space induced by $\Phi(\cdot, \cdot)$. Thus, while our method has some connection with kernel-based distances, our weights result in consistent estimation of the ADRF without correct specification of the kernel $\Phi(\cdot, \cdot)$.

The following shows the convergence rate of the penalized distance covariance optimal weights under additional mild moment conditions on the covariates, treatment, and $w^*(\mathbf{X}, A)$. This theorem builds on a key lemma on the squares of the weights presented in Section F of the supplementary material.

Theorem 3.6. Assume the conditions required in Theorem 3.5 hold, that the moment conditions (A1) and (A2) listed in the Appendix hold, and that $\mathbb{E}w^{*2}(\mathbf{X}, A) < \infty$. Additionally assume $\mathbb{E}[\varepsilon^2 \mid \mathbf{X} = \mathbf{x}, A = a_0] < c$ for some c uniformly over $\mathbf{x} \in \mathcal{X}$. Then

$$\widehat{\mu}_{NW}^{\mathbf{w}^{pd}}(a_0) - \mu(a_0) = O_p(1/\sqrt{nh} + h^2). \tag{13}$$

This convergence rate is the standard rate for unweighted Nadaraya-Watson estimators of a univariate regression function; thus, the convergence rate of the weighted estimator based on our weights is unaffected by the nonparametric nature of the estimation of $\mathbf{w}^{pd}$.

3.4. Augmented Estimation with Independence Weights

Another class of estimators for causal ADRFs are doubly robust/augmented estimators such as in Kennedy et al. (2017) and Colangelo and Lee (2020), which combine sample weights and an outcome model $\widehat{\mu}(\mathbf{x}, a_0)$, ideally a consistent estimator of $\mu(\mathbf{x}, a_0)$. Though the term "doubly robust" is reflective of the property that if either the weights or outcome model is correctly specified, the estimator will be consistent, a more consequential property of doubly robust estimators is that the error rates of each model are multiplied.

Although in the previous section we showed that the use of the DCOWs alone results in the ideal convergence rate for the ADRF, DCOWs can be enhanced by using doubly robust/augmented estimators, or conversely, that the use of DCOWs can significantly enhance doubly robust estimators. The DCOWs assure the analyst that the estimator will converge at the right rate regardless of whether the outcome model is correctly specified, but allow for using an outcome model to provide an opportunity to fine-tune performance by reducing residual variance, resulting in an estimator that works well empirically.

Here, for simplicity of presentation, we focus on the following Nadaraya-Watson-based augmented estimator based on any estimator $\hat{\mu}(\mathbf{x}, a_0)$ of $\mu(\mathbf{x}, a_0)$ as

$$\widehat{\mu}_{NW}^{\mathbf{w},DR}(a_0) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(\mathbf{X}_i, a_0) + \frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}(\mathbf{X}_i, a_0))w_i K_h(A_i - a_0)}{\sum_{i=1}^{n} K_h(A_i - a_0)}.$$

In Section B of the supplementary material, we derive a decomposition of the error $\widehat{\mu}_{NW}^{\mathbf{w},DR}(a_0) - \mu(a_0)$ and show that the systematic bias term not related to smoothing is

$$\int_{\mathcal{X}} \{\mu(\mathbf{x}, a_0) - \widehat{\mu}(\mathbf{x}, a_0)\} \int_{\mathcal{A}} \mathrm{d}\left[F^n_{\mathbf{X},A,\mathbf{w}} - F^n_{\mathbf{X}} F^n_A\right](\mathbf{x}, a). \tag{14}$$

Lemma 3.3 implies that (14) is less than or equal to $3\|\mu(\mathbf{x}, a_0) - \widehat{\mu}(\mathbf{x}, a_0)\|_{\mathcal{H}}\mathcal{D}(\mathbf{w})$ provided that $\mu(\cdot, a_0) - \widehat{\mu}(\cdot, a_0) \in \mathcal{H}$. With the DCOWs, we provide a set of weights $\mathbf{w}^d$ that makes $\mathcal{D}(\mathbf{w})$ as small as possible, though it may not be exactly zero for a finite sample.

We now formalize the above claims and provide asymptotic results for the augmented estimator $\widehat{\mu}_{NW}^{\mathbf{w},DR}(a_0)$ using a slightly modified version of our PDCOWs; this modification is motivated by a technical condition and in practice has little or no impact on the weights. The modified penalized distance covariance optimal weights are defined as

$$\widetilde{\mathbf{w}}_n^{pd} \in \underset{\mathbf{w}=(w_1,\ldots,w_n)}{\operatorname{argmin}} \ \mathcal{D}(\mathbf{w}) + \lambda\frac{1}{n^2}\sum_{i=1}^{n} w_i^2 \ \text{ such that } \sum_{i=1}^{n} w_i = n, Bn^{1/3} \geq w_i \geq 0 \ \text{ for } \ 0 < \lambda, B < \infty, \ i = 1, \ldots, n. \tag{15}$$

Here, B is a pre-specified positive constant. We have found that in practice the maximum weight rarely, if ever, comes near $Bn^{1/3}$ with $B = 1$ even without the constraint on the max weight. Thus, this additional constraint does little to change the empirical behavior of the PDCOWs. Further, we show in Section Section F of the supplementary material and in the following that the key asymptotic results of the PDCOWs (e.g., Theorem 3.4) also hold for $\widetilde{\mathbf{w}}_n^{pd}$.

We next show that the modified weights paired with an augmented estimator based on a correctly-specified outcome model result in asymptotic normality of the resulting causal ADRF.

Theorem 3.7. Let $\widetilde{\mathbf{w}}_n^{pd}$ be the distance covariance optimal weights defined in (15). Let $K(\cdot)$ be a kernel with conditions listed in the statement of Theorem 3.5. Assume the moment conditions (A1) and (A2) listed in the Appendix hold, that $\mathbb{E}\|\mathbf{X}\|_2 < \infty$ and $\mathbb{E}|A| < \infty$, that $\mathbb{E}w^{*2}(\mathbf{X}, A) < \infty$, that $1/f_A(a_0)$ is uniformly bounded, $f_A(a_0)$ is bounded and has second order derivatives, and that $\mu(\mathbf{x}, a_0)$ is bounded and continuous on $\mathcal{X} \times \mathcal{A}$ and has second order derivatives. Further, assume that $\mathbb{E}|\varepsilon_i|^3 < \infty$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$ for all i. Assume that the outcome regression model satisfies $\mu(\cdot, a_0) - \hat{\mu}(\cdot, a_0) \in \mathcal{H}_{\mathcal{X}}$ for each n, $||\mu - \hat{\mu}||_{\mathcal{H}} = O_p(1)$, and $\int_{\mathcal{X}}(\mu(\mathbf{x}, a_0) - \hat{\mu}(\mathbf{x}, a_0))^2 \mathrm{d}F_{\mathbf{X}}(\mathbf{x}) = o_p(1)$. Then

$$\frac{\sqrt{nh}f_A(a_0)}{\sigma\sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i^2 K_h^2(A_i - a_0)}}\left(\hat{\mu}_{NW}^{\mathbf{w},DR}(a_0) - \mu(a_0) - h^2\kappa_2 B(a_0)\right) \xrightarrow{d} \mathcal{N}(0, 1) \tag{16}$$

as $h \to 0$, $nh \to \infty$, and $nh^5 = O_p(1)$ for $\mathbf{w} = \widetilde{\mathbf{w}}_n^{pd}$, where $\kappa_2 = \int u^2 K(u)\mathrm{d}u$, $B(a_0) \equiv \int_{\mathcal{X}} B(\mathbf{x}, a_0)\mathrm{d}F_{\mathbf{X}}(\mathbf{x})$, and $B(\mathbf{x}, a_0) \equiv \frac{\partial^2}{\partial a_0^2}\mu(\mathbf{x}, a_0)/2 + \frac{\partial}{\partial a_0}\mu(\mathbf{x}, a_0)\frac{\partial}{\partial a_0}f_A(a_0)/f_A(a_0)$.

The conditions required regarding $\hat{\mu}(\cdot, a_0)$ and ε_i are analogous to those required in Theorem 3 of Wong and Chan (2017). We note that normalization by the squares of the weights in (16) is necessary as our results do not rely on a proof of $\widetilde{\mathbf{w}}_n^{pd}$ or their squares to converge to anything in particular. We note,

however, that the expression in (16) can be simplified if it is possible to show that $\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{w}_i^{pd} - w^*(\mathbf{X}_i, A_i)\right)^2$ converges to 0 in probability. In particular, it would simplify to a form similar to the asymptotic distribution of the augmented ADRF estimator in Colangelo and Lee (2020). We provide an informal investigation into the convergence of the PDCOWs to the true GPS weights in Section J of the supplementary material.

4. Numerical Experiments

We evaluate our proposed methodology using two tracks of simulation experiments. The first track uses existing data to conduct simulation studies. In this approach, we fix the confounding structure of a complex dataset and simulate outcomes under a wide range of outcome models. In the second track of simulation experiments, we generate synthetic data under the data-generating setup of Vegetabile et al. (2021), which amounts to a markedly different data-generating process from the first set of simulation experiments.

4.1. Comparator Methods

We use the following methods to estimate weights. We use a naïve method which uses weights equal to identity (unweighted). We use stabilized GPS weights computed four ways: a linear regression model for estimating $\mathbb{E}(A|\mathbf{X})$ (i.e., the conditional mean of dose given covariates) and a normal conditional density ("GPS normal"); a gamma regression model for estimating $\mathbb{E}(A|\mathbf{X})$ and a gamma conditional density ("GPS gamma"); a generalized boosted model for estimating $\mathbb{E}(A|\mathbf{X})$, where the number of trees was chosen to minimize the weighted average absolute correlation between the treatment and covariates as in Zhu, Coffman, and Ghosh (2015), and a normal conditional density ("GBM"); and a Bayesian Additive Regression Trees model for estimating $\mathbb{E}(A|\mathbf{X})$ and a normal conditional density ("BART"). For methods that estimate weights by directly inducing a lack of correlation between moments of the treatment and covariates, we use the covariate balancing generalized propensity score of Fong, Hazlett, and Imai (2018) and the entropy balancing approach of Tübbicke (2022) and Vegetabile et al. (2021). Among this class of methods, we only consider these two as other approaches to estimating weights that decorrelate pre-specified moments behave largely similarly to each other (Vegetabile et al. 2021; Tübbicke 2022). We use the exactly identified version of the covariate balancing generalized propensity score ("CBPS"). We use versions of the entropy balancing approach that decorrelate either all first order moments ("Entropy (1)"), all first order moments and squared terms in continuous covariates ("Entropy (2)"), or all first order moments, pairwise interactions, and squared terms in continuous covariates ("Entropy (int)"). Any resulting weights greater than 500 when standardized to sum to n are truncated at 500. We use our proposed DCOWs ("DCOW") and the proposed DCOWS where we further induce exact decorrelation of first order moments ("DCOW (dm)") as discussed in Section D of the supplementary material, both using the dimension adjustment described also in that section. For each method, the weights are used in a weighted local linear regression used to estimate the ADRF. For the GPS (normal), GPS (gamma), DCOW, and DCOW (dm) methods, we also use the doubly robust estimator of Kennedy et al. (2017) with a an outcome model that is linear in the covariates with additive first order terms; the methods are labeled as "GPS (normal,DR)," "GPS (gamma,DR)," "DCOW (DR)," and "DCOW (dm,DR)," respectively.

4.2. Simulation using National Medical Expenditure Survey Data

The National Medical Expenditure Survey (NMES) relates medical expenditures with degree of smoking among U.S. citizens (Johnson et al. 2003). The NMES dataset contains information on 9708 individuals. The outcome variable is the total medical expenditures in dollars and the treatment $A \in [0.05, 216]$ is the amount of smoking in pack years. We limit the data to those with $A \leq 80$ (i.e., $\mathcal{A} = [0.05, 80]$), leaving 9368 units. We limit to those with such values, as the number of patients who smoke more than 80 pack years is exceptionally rare. Two of the covariates are continuous and the remaining are categorical, with an overall dimension equal to 18 after converting categorical variables to dummy variables. In our simulation, we leave the treatment level (pack years) and covariates intact and simulate outcomes for each unit from the following model: $Y = m(\mathbf{X};\boldsymbol{\theta}_1) + f(A)(1 + \delta(\mathbf{X};\boldsymbol{\theta}_2))) + \varepsilon$, where $m(\mathbf{X};\boldsymbol{\theta}_1)$ are main effects with parameters $\boldsymbol{\theta}_1$ generated as described in Section H of the supplementary material, $\delta(\mathbf{X};\boldsymbol{\theta}_2)$ are mean 0 interaction effects, ε are iid $N(0,4)$ random variables, and the treatment effect curve $f(A) = A/4 + \frac{2}{(A/100+1/2)^3} - (A - 40)^2/100$. The main effect function involves interactions and up to squared terms in the continuous covariates. For the constant treatment effect setting, $\delta(\mathbf{X};\boldsymbol{\theta}_2) = 0$, and for the heterogeneous treatment effect setting it is a nonzero but mean 0 function involving linear and interaction effects in covariates. Notably, the parameters in $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ also make some covariates have no contribution to the main effects and interactions, respectively. We generate 100 different draws of the coefficients ($\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$) in the outcome model above, allowing for the simulation study to explore a wide variety of outcome models. For each of these 100 outcome model draws, we replicate the simulation experiment 1000 times. For each replication, a random subsample of size $n < 9368$ is drawn without replacement from the 9368 units, and outcomes for these units are generated from the outcome model. The simulation process is repeated for each of the sample sizes $n = 100, 200, 400, 800, 1600, 3200$.

We also consider a simulation setting with the heterogeneous treatment effect where 50 additional variables are added to the covariate vector so that the overall dimension is 68. The 50 additional variables are generated so that they are correlated with both the response and treatment while preserving the original values of the response and treatment; details are provided in Section H of the supplementary material.

All methods are evaluated with a measure of the mean absolute bias (MAB) and integrated root mean squared error (IRMSE), both of which are used for evaluation of estimates of the ADRF in Kennedy et al. (2017). MAB and IRMSE are

Table 1. Mean absolute bias (MAB) and integrated root mean squared error (IRMSE) for the constant treatment effect setting.

	$n = 100$		$n = 200$		$n = 400$		$n = 800$		$n = 1600$		$n = 3200$	
Method	MAB	IRMSE	MAB	IRMSE	MAB	IRMSE	MAB	IRMSE	MAB	IRMSE	MAB	IRMSE
Unweighted	11.461	17.488	11.283	14.753	11.252	13.182	11.246	12.297	11.254	11.799	11.237	11.491
GPS (normal)	10.145	27.180	15.033	29.187	18.197	28.056	20.786	27.650	21.832	25.703	22.611	24.432
GPS (gamma)	9.340	23.994	9.892	19.269	9.993	15.831	10.112	13.978	10.084	12.693	10.006	11.162
GPS (normal,DR)	8.315	30.551	11.081	33.342	13.216	30.293	15.320	27.636	15.324	21.199	15.769	18.312
GPS (gamma,DR)	7.045	26.437	6.866	16.463	6.905	12.044	6.920	10.168	6.768	8.985	6.745	7.612
CBPS	9.871	25.398	12.463	23.375	14.055	22.016	14.064	18.802	14.970	17.546	15.712	16.847
GBM	8.190	23.428	7.551	17.298	7.105	12.498	7.379	10.432	7.524	9.225	7.746	8.589
BART	6.550	17.206	7.275	14.423	8.023	12.326	8.638	11.292	9.033	10.659	9.446	10.366
Entropy (1)	—	—	9.409	17.880	9.166	12.946	9.161	11.321	9.023	10.258	8.952	9.537
Entropy (2)	—	—	8.909	25.080	9.343	15.520	8.927	12.360	8.503	10.456	8.231	9.245
Entropy (2,int)	—	—	—	—	—	—	—	—	8.998	253.786	10.974	16.780
DCOW	4.684	12.204	3.866	8.383	3.495	6.245	3.252	4.947	2.992	4.075	2.750	3.416
DCOW (dm)	4.522	16.465	3.907	8.988	3.544	6.364	3.276	5.018	3.001	4.147	2.775	3.497
DCOW (DR)	**3.904**	**9.284**	**3.335**	**6.455**	**2.663**	**4.586**	**2.196**	**3.388**	**1.919**	**2.661**	**1.753**	**2.189**
DCOW (dm,DR)	3.905	11.567	3.459	6.919	2.828	4.806	2.356	3.568	2.057	2.817	1.873	2.320

defined as

$$\text{MAB} = \int_{\mathcal{A}} \left| \frac{1}{S} \sum_{s=1}^{S} \widehat{\mu}_s(a) - \mu(a) \right| \widehat{f}_A(a) \mathrm{d}a,$$

$$\text{IRMSE} = \int_{\mathcal{A}} \left[\frac{1}{S} \sum_{s=1}^{S} \{\widehat{\mu}_s(a) - \mu(a)\}^2 \right]^{1/2} \widehat{f}_A(a) \mathrm{d}a,$$

where $\widehat{f}_A(a)$ is a kernel density estimate of the marginal density of the treatment variable, s indexes the simulation replications, and $\mathcal{A}$ is a trimmed version of the support of the treatment variable that excludes pack years greater than 80. We calculate the MAB and IRMSE statistics for each of the 100 different outcome model settings and then average them over the 100 settings.

The MAB and IRMSE results for the constant treatment effect setting are displayed in Table 1 and the results for the heterogeneous treatment effect setting are displayed in Section H of the supplementary material as they track closely with the former setting. Results for any method are not shown if no numerical solution is found in more than 75% of the replications. The standard DCOWs performed the best in terms of both MAB and IRMSE across all sample sizes among all non-doubly robust estimators, only being outperformed by the doubly robust estimator that uses DCOWs as weights. DCOWs with exact first order moment decorrelation performed similarly to, but slightly worse than standard DCOWs, for both the non-doubly robust estimator and the doubly robust estimator, though the performance was much worse for $n = 100$, as the exact constraints may have been too stringent for the sample size. The estimators using standard GPS weights, both doubly robust and non-doubly robust, performed poorly in terms of both MAB and IRMSE for small to moderate sample sizes, though the doubly robust estimator with gamma regression-based weights performed well in terms of MAB and IRMSE for larger sample sizes. The machine learning approaches to GPS estimation (GBM and BART) performed relatively poorly in terms of MAB and IRMSE for small sample sizes, with BART performing better than GBM for small sample sizes, but with similar but slightly better performance for GBM for large sample sizes. Among the GPS-based methods, GBM and BART generally yielded the lowest MAB and IRMSE across the sample size settings, except for the doubly robust gamma GPS estimator. Among the moment balancing approaches, entropy balancing with decorrelation induced only for first and second order moments and not interactions performed the best in terms of MAB and IRMSE, though for small sample sizes, entropy balancing frequently failed to arrive at a solution. For the heterogeneous setting, entropy balancing with decorrelation induced only for first order moments performed the best, likely due to numerical instability of adding more moment constraints. The CBPS estimator yielded worse performance than did entropy balancing overall, though it did not face the convergence problems of entropy balancing for small sample sizes. Entropy balancing with higher order moment constraints failed to arrive at a solution more frequently and performed poorly even for large sample sizes, likely due to the strictness of the exact moment constraints.

The results for the heterogeneous effect setting with 50 additional noise variables are displayed in Section H of the supplementary material. Again, DCOWs performed the best among all methods in terms of both MAB and IRMSE for all sample sizes except $n = 100$, where the doubly robust gamma GPS estimate was slightly better in terms of IRMSE.

Additional simulation results under the same setting of Vegetabile et al. (2021) can be found in Section I of the supplementary material, wherein our weights also yielded the best performance across all sample size settings.

5. Analysis of Mechanical Power Data

We use the Medical Information Mart for Intensive Care III (MIMIC-III) database (Johnson et al. 2016) to study the impact of a large degree of mechanical power of ventilation on mortality among critically ill patients in an ICU using electronic health record (EHR) data. Our study and the construction of the cohort from the MIMIC-III database are based on the original study of Neto et al. (2018) and the code provided by the authors located at *https://github.com/alistairewj/mechanical-power*. Since there are widely-used formal guidelines that influence ventilation management among patients with respiratory distress (Papazian et al. 2019), many observable factors in the EHR data are highly related to the mechanical power of ventilation. Many of

Table 2. Summary statistics (mean, standard deviation, median, 95th percentile, and maximum) of the absolute weighted correlations of the first five powers of mechanical power of ventilation and all marginal moments of covariates, pairwise interactions of covariates, and up to 5th order polynomials of covariates.

	Unweighted	GPS (normal)	GPS (gamma)	CBPS	GBM	BART	Entropy (1)	Entropy (2)	DCOW	DCOW (dm)
(7)	21.041	2.574	4.269	4.713	12.992	6.166	1.093	4.262	0.220	0.237
$\mathcal{D}(\mathbf{w})$	21.041	2.349	4.014	4.328	12.960	5.710	1.005	4.083	0.144	0.151
$\mathcal{E}(F^n_{A,\mathbf{w}}, F^n_A)$	0.000	0.027	0.003	0.010	0.003	0.043	0.030	0.001	0.051	0.059
$\mathcal{E}(F^n_{\mathbf{X},\mathbf{w}}, F^n_{\mathbf{X}})$	0.000	0.198	0.251	0.374	0.029	0.414	0.058	0.178	0.026	0.027
ESS	4933	2048	1832	1017	3892	782	2185	478	2232	2057
mean(|corr|)	0.070	0.014	0.037	0.025	0.058	0.045	0.007	0.021	0.007	0.003
sd(|corr|)	0.082	0.015	0.036	0.025	0.061	0.043	0.008	0.022	0.006	0.003
median(|corr|)	0.041	0.010	0.026	0.018	0.036	0.035	0.005	0.016	0.005	0.002
95-pctl(|corr|)	0.248	0.044	0.105	0.063	0.200	0.118	0.020	0.055	0.018	0.009
max(|corr|)	0.788	0.200	0.854	0.470	0.568	0.939	0.123	0.507	0.084	0.045

NOTE: Summaries are over 8284 $\times$ 5 weighted correlations of covariate moments and mechanical power moments.

these factors are also closely related to patient mortality. The guidelines involve consideration of both factors individually and many interactions among these factors. In addition, as the exposure, the mechanical power of ventilation, is itself a complex summary of multiple manipulable elements of a ventilator, the guidelines are not directly related to the exposure, but have a strong indirect effect on the power of ventilation. Thus, the dependence between observable factors and the exposure level is highly complex.

Patients included in the study were at least 16 years of age and received invasive mechanical ventilation for at least 48 hr. This restriction was used in the original analysis of Neto et al. (2018) and focused on the population that requires extended use of ventilation. Patients who die within this 48-hour period are substantially sicker, and it is unclear whether manipulation of ventilation settings for such a population may induce meaningful changes to outcomes. The study contains 5014 patients, and the treatment variable of interest is the amount of energy generated by the mechanical ventilator measured by the mean of the largest and smallest mechanical power of ventilation in Joules per minute in the second 24-hour period in the ICU, as in Neto et al. (2018). The outcome is an indicator of in-hospital mortality. As in the original analysis of Neto et al. (2018), we limit our analysis to patients receiving less than or equal to 50 Joules per minute, resulting in a sample size of 4933. We include 73 pretreatment covariates (some of which are discrete), resulting in a total dimension of 97 of the vector $\mathbf{X}$ of potential confounders. All methods considered in the simulation section were then applied to construct weights aiming to control for dependence between the 97 covariates and mechanical power. Both GPS approaches, GBM, and BART resulted in several extraordinarily large weights; these weights were truncated to mitigate extreme variation. We attempted to construct entropy balancing weights that either exactly or approximately (within a correlation tolerance of 0.1) balance pairwise interactions, but no numerical solution was found.

For all methods, balance statistics, including our developed criterion (7) and weighted correlations between first-order moments and pairwise interactions of covariates and mechanical power, are summarized in Table 2. In terms of weighted marginal correlations, the version of our DCOWs that induces first-order marginal correlations between covariates and mechanical power to be zero performs the best, with the standard DCOWs and entropy balancing weights a close second. In terms of our proposed criterion (7), by definition, the DCOWs yield the smallest value. However, it is notable that only a small price is paid in terms of both (7) and effective sample size (ESS) in order to exactly decorrelate marginal covariate moments and treatment. Among methods that do not target independence, entropy balancing has the smallest value of (7), indicating that decorrelating first-order moments in this particular dataset mitigates a vast majority of the dependence between covariates and mechanical power of ventilation. We also note that exactly decorrelating second-order moments using entropy balancing results in further instability, and thus worse mitigation of dependence. However, exact decorrelation of marginal covariate moments via entropy balancing results in much larger standard errors of the resulting ADRF, as evidenced in Figure 2. Interestingly, in terms of both marginal weighted correlations and our proposed independence metric, the flexible machine learning approaches (GBM and BART) perform significantly worse than a normal model for the conditional density.

For a concrete example of how DCOWs mitigate dependence, we illustrate the unadjusted marginal dependence of the PaO_2/FiO_2 ratio and the mechanical power of ventilation, where PaO_2 is the partial pressure of oxygen in the arterial blood and FiO_2 is the fraction of inspired oxygen. PaO_2/FiO_2 ratio is a pretreatment covariate that characterizes acute hypoxemia, defines the presence and severity of acute respiratory distress syndrome (ADRS), and plays a critical role in ventilation guidelines (Papazian et al. 2019). Both PaO_2 and PaO_2/FiO_2 ratio are also strongly associated with mortality and are thus critical confounders in this study. We display weighted marginal dependence of PaO_2 and the mechanical power of ventilation with our proposed weights and with the entropy-balancing weights in Figure 1; we also show this relationship for the PaO_2/FiO_2 ratio in Section K of the supplementary material. Compared with methods that aim to exactly decorrelate specified moments, our proposed weights can handle nonlinear dependence between covariates and treatment even in datasets with moderately high dimensions. While entropy balancing accounts for a significant proportion of dependence, there remains residual nonlinear dependence.

For each method, we construct pointwise confidence bands of the ADRF using a nonparametric bootstrap similar to Wang and Wahba (1995). Weighted local linear regression estimates of the ADRF of mechanical power of ventilation on in-hospital mortality and 95% confidence bands are displayed in Figure 2.

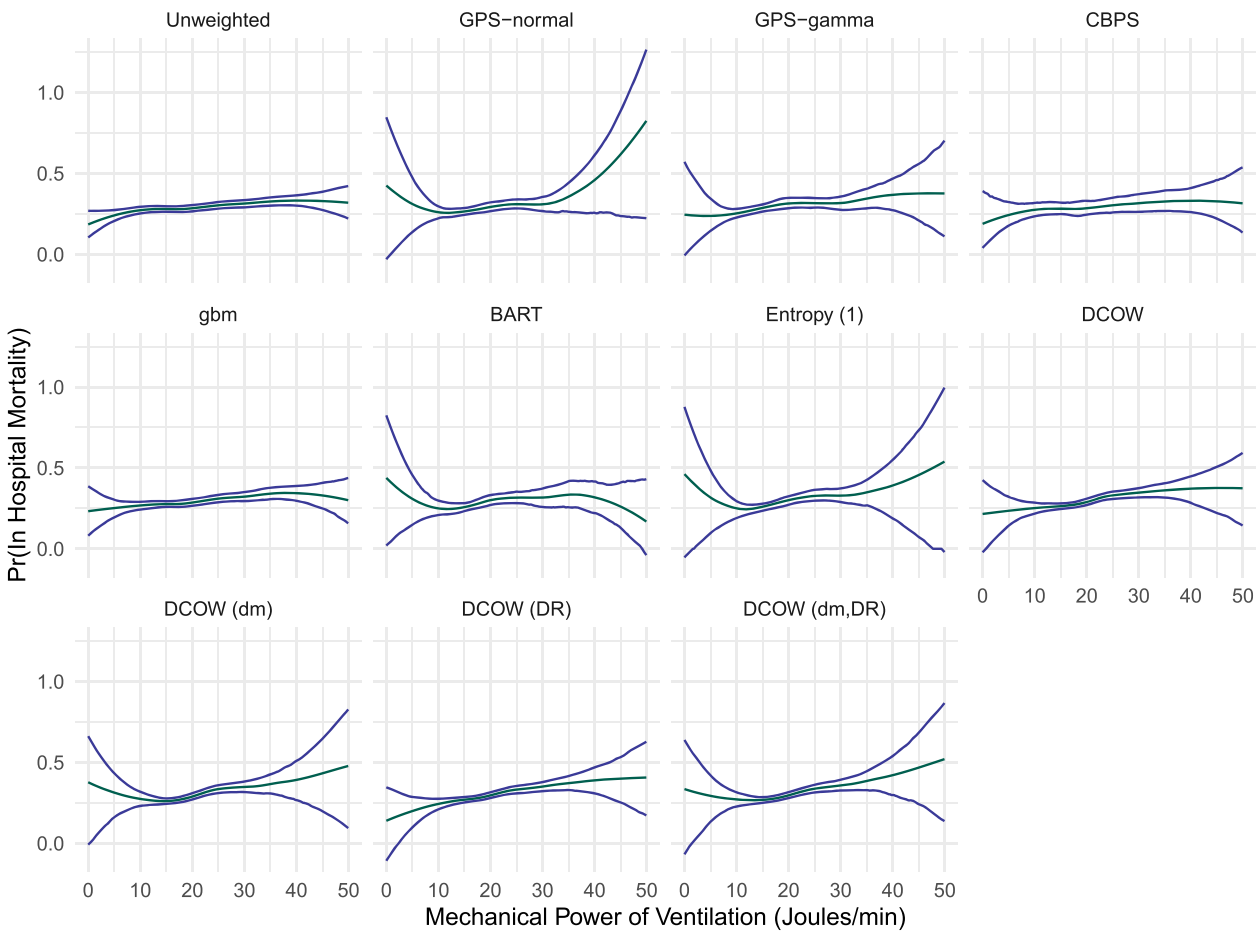

Figure 2. Shown are weighted local linear regression estimates of the ADRF of mechanical power on the probability of in-hospital mortality and pointwise 95% confidence bands estimated via a nonparametric bootstrap. 38 replications of the bootstrap resulted in no numerical solution for entropy balancing weights and are left out of the confidence interval calculation.

Alternatively, it is possible to conduct inference using the results of Theorem 3.7; however, doing so requires careful selection of the bandwidth parameter h, such as by using undersmoothing, which involves choosing a smaller bandwidth than the optimal one. Rigorous assessment of the use of Theorem 3.7 for inference is left as future work. The entropy balancing approach with higher order moments was not included due to the small effective sample size and a large number of bootstrap replications for which no solution was found. Both the entropy balancing weights and the GPS estimated with a normal model exhibit reasonable control over the dependence between covariates and treatment but result in unacceptably high variances in the estimate of the ADRF.

6. Discussion

In this article, we provide a detailed inspection of the role of weights in weighted nonparametric estimates of causal quantities involving continuous-valued treatments from observational data. This inspection shows clearly that the key source of bias depends on the degree to which the weights induce independence between the treatment and confounders. We then provide a measure that characterizes how well a set of weights mitigates the dependence between the treatment and confounders. This measure does not require any tuning parameters, making it straightforward to deploy in practice. Given some light smoothness conditions on the outcome data-generating model, this measure acts as an upper bound on the key source of systematic bias in a weighted nonparametric estimate of the ADRF. Our proposed weights, the DCOWS, which minimize our measure of dependence, provide an empirically robust means for estimating weights, as they directly target independence between the treatment and confounders. These weights are a natural complement to doubly robust estimators, as they provide an anchor for the doubly robust estimator: since they are guaranteed to be consistent on their own, the consistency of the doubly robust estimator does not critically depend on the correctness of the outcome model. Thus, the outcome model can be safely used as a tool to reduce variability in the estimate of the ADRF.

In contrast to other weighting approaches aimed at removing bias due to confounding with continuous treatments, the DCOWs enjoy a number of benefits that make them particularly attractive for applied use: they do not require any modeling of the relationship between the treatment or outcome and the covariates, they do not require the choice of specific features of the covariate distribution to balance, they do not require parameter tuning or cross-fitting, they perform well empirically (in addition to their strong asymptotic properties), and they can be readily implemented without specialized software

using existing quadratic programming solvers. We have released an open-source implementation of our method in the R statistical computing language that can be used off-the-shelf to estimate the weights, available at *https://github.com/jaredhuling/independenceWeights*. Although our method, like all methods that adjust for confounding by measured confounders, requires that a sufficient set of confounding variables has been collected by the researcher, its ease of use helps alleviate the analytical burden of estimating the causal effects of continuous treatments in the presence of confounding by measured variables.

Appendix: Regularity Conditions

For $(\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_6, A_6) \overset{\text{iid}}{\sim} F_{\mathbf{X},A}$, define the 6th order kernel

$$k((\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_6, A_6)) = w^*(\mathbf{X}_1, A_1) g_{\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) \times g_A(A_1, A_2, A_5, A_6) w^*(\mathbf{X}_2, A_2)$$

with $w^*(\mathbf{x}, a) = \frac{f_A(a)}{f_{A|\mathbf{X}}(a|\mathbf{X}=\mathbf{x})}$, $g_{\mathbf{X}}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) = \|\mathbf{X}_1 - \mathbf{X}_2\|_2 - \|\mathbf{X}_1 - \mathbf{X}_3\|_2 - \|\mathbf{X}_2 - \mathbf{X}_4\|_2 + \|\mathbf{X}_3 - \mathbf{X}_4\|_2$, and $g_A(A_1, A_2, A_3, A_4) = |A_1 - A_2| - |A_1 - A_3| - |A_2 - A_4| + |A_3 - A_4|$. Further define the 4th order kernels

$$\begin{aligned} &k_{\mathbf{X}}((\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_4, A_4)) \\ &= w^*(\mathbf{X}_1, A_1)\|\mathbf{X}_1 - \mathbf{X}_3\|_2 + w^*(\mathbf{X}_2, A_2)\|\mathbf{X}_2 - \mathbf{X}_4\|_2 \\ &\quad - w^*(\mathbf{X}_1, A_1) w^*(\mathbf{X}_2, A_2)\|\mathbf{X}_1 - \mathbf{X}_2\|_2 - \|\mathbf{X}_3 - \mathbf{X}_4\|_2 \quad \text{and} \\ &k_A((\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_4, A_4)) \\ &= w^*(\mathbf{X}_1, A_1)|A_1 - A_3| + w^*(\mathbf{X}_2, A_2)|A_2 - A_4| \\ &\quad - w^*(\mathbf{X}_1, A_1) w^*(\mathbf{X}_2, A_2)|A_1 - A_2| - |A_3 - A_4|. \end{aligned}$$

The following assumptions are required for several Lemmas and Theorems presented that rely on V-statistics. These conditions amount to finite moment conditions on squares of the Euclidean norms of $\mathbf{X}$, A and their products with the weights $w^*(\mathbf{X}, A)$.

(A1) $\mathbb{E}[k^2((\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_6, A_6))] < \infty$
(A2) $\mathbb{E}[k_{\mathbf{X}}^2((\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_4, A_4))] < \infty$ and
$\mathbb{E}[k_A^2((\mathbf{X}_1, A_1), \ldots, (\mathbf{X}_4, A_4))] < \infty$

Supplementary Materials

Supplementary Material: The supplementary material contains details for error decompositions of weighted nonparametric estimators, computational details of the proposed method, extended discussion of the existing literature, proofs of the theoretical results, and additional simulation studies. (pdf)

Acknowledgments

We would like to thank the Associate Editor and anonymous referees for their constructive feedback and suggestions Jue Hou for the helpful discussion.

Funding

Chen's effort was partially supported by NSF grant DMS-2054346 and the University of Wisconsin School of Medicine and Public Health from the Wisconsin Partnership Program.

ORCID

Jared D. Huling http://orcid.org/0000-0003-0670-4845
Guanhua Chen http://orcid.org/0000-0002-9314-2037

References

Ai, C., Linton, O., Motegi, K., and Zhang, Z. (2021), "A Unified Framework for Efficient Estimation of General Treatment Models," *Quantitative Economics*, 12, 779–816. [2]

Chattopadhyay, A., Hase, C. H., and Zubizarreta, J. R. (2020), "Balancing vs Modeling Approaches to Weighting in Practice," *Statistics in Medicine*, 39, 3227–3254. [7]

Colangelo, K., and Lee, Y.-Y. (2020), "Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments," arXiv preprint arXiv:2004.03036. [1,4,8,9]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [4]

Díaz, I., and van der Laan, M. J. (2013), "Targeted Data Adaptive Estimation of the Causal Dose–Response Curve," *Journal of Causal Inference*, 1, 171–192. [3]

Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y., and Yang, X. (2021), "Optimal Covariate Balancing Conditions in Propensity Score Estimation," *Journal of Business & Economic Statistics*, 41, 97–110. [1]

Fong, C., Hazlett, C., and Imai, K. (2018), "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements," *The Annals of Applied Statistics*, 12, 156–177. [2,4,9]

Galvao, A. F., and Wang, L. (2015), "Uniformly Semiparametric Efficient Etimation of Treatment Effects with a Continuous Treatment," *Journal of the American Statistical Association*, 110, 1528–1542. [1]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [3]

Hirano, K., and Imbens, G. W. (2004), *The Propensity Score with Continuous Treatments*, pp. 73–84, Wiley. [1,4]

Huling, J. D., and Mak, S. (2020), "Energy Balancing of Covariate Distributions," arXiv preprint arXiv:2004.13962. [2,6]

Imai, K., and Ratkovic, M. (2014), "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society*, Series B, 76, 243–263. [2]

Imai, K., and Van Dyk, D. A. (2004), "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866. [1]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29. [3]

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016), "Mimic-iii, A Freely Accessible Critical Care Database," *Scientific Data*, 3, 1–9. [10]

Johnson, E., Dominici, F., Griswold, M., and Zeger, L. S. (2003), "Disease Cases and their Medical Costs Attributable to Smoking: An Analysis of the National Medical Expenditure Survey," *Journal of Econometrics*, 112, 135–151. [9]

Kallus, N., and Santacatterina, M. (2019), "Kernel Optimal Orthogonality Weighting: A Balancing Approach to Estimating Effects of Continuous Treatments," arXiv preprint arXiv:1910.11972. [2]

Kang, J. D., and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 523–539. [1]

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017), "Nonparametric Methods for Doubly Robust Estimation of Continuous Treatment Effects," *Journal of the Royal Statistical Society*, Series B, 79, 1229–1245. [1,3,4,8,9]

Kish, L. (1965), *Survey Sampling* (Vol. 26), New York: Wiley. [7]

Mak, S., and Joseph, V. R. (2018), "Support Points," *The Annals of Statistics*, 46, 2562–2592. [6]

Martinet, G. (2020), "A Balancing Weight Framework for Estimating the Causal Effect of General Treatments," arXiv preprint arXiv:2002.11276. [2]

Naimi, A. I., Moodie, E. E., Auger, N., and Kaufman, J. S. (2014), "Constructing Inverse Probability Weights for Continuous Exposures: A Comparison of Methods," *Epidemiology*, 25, 292–299. [1,4]

Neto, A. S., Deliberato, R. O., Johnson, A. E., Bos, L. D., Amorim, P., Pereira, S. M., Cazati, D. C., Cordioli, R. L., Correa, T. D., Pollard, T. J. et al. (2018), "Mechanical Power of Ventilation is Associated with Mortality in Critically Ill Patients: An Analysis of Patients in Two Observational Cohorts," *Intensive Care Medicine*, 44, 1914–1922. [10,11]

Papazian, L., Aubron, C., Brochard, L., Chiche, J.-D., Combes, A., Dreyfuss, D., Forel, J.-M., Guérin, C., Jaber, S., Mekontso-Dessap, A. et al. (2019), "Formal Guidelines: Management of Acute Respiratory Distress Syndrome," *Annals of Intensive Care*, 9, 1–18. [10,11]

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560. [1,4]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [1]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1,4]

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013), "Equivalence of Distance-based and RKHS-based Statistics in Hypothesis Testing," *The Annals of Statistics*, 41, 2263–2291. [7]

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020), "Osqp: An Operator Splitting Solver for Quadratic Programs," *Mathematical Programming Computation*, 12, 637–672. [7]

Székely, G. J., and Rizzo, M. L. (2013), "Energy Statistics: A Class of Statistics based on Distances," *Journal of Statistical Planning and Inference*, 143, 1249–1272. [2]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [2,5]

Tübbicke, S. (2022), "Entropy Balancing for Continuous Treatments," *Journal of Econometric Methods*, 11, 71–89. [2,9]

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer. [1]

Vegetabile, B. G., Griffin, B. A., Coffman, D. L., Robbins, M. W., Cefalu, M., and McCaffrey, D. F. (2021), "Nonparametric Estimation of Population Average Dose-Response Curves using Entropy Balancing Weights for Continuous Exposures," *Health Services and Outcomes Research Methodology*, 21, 69–110. [2,4,9,10]

Wang, Y., and Wahba, G. (1995), "Bootstrap Confidence Intervals for Smoothing Splines and their Comparison to Bayesian Confidence Intervals," *Journal of Statistical Computation and Simulation*, 51, 263–279. [11]

Wong, R. K., and Chan, K. C. G. (2017), "Kernel-based Covariate Functional Balancing for Observational Studies," *Biometrika*, 105, 199–213. [8]

Yiu, S., and Su, L. (2018), "Covariate Association Eliminating Weights: A Unified Weighting Framework for Causal Effect Estimation," *Biometrika*, 105, 709–722. [2,4]

Zhu, Y., Coffman, D. L., and Ghosh, D. (2015), "A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments," *Journal of Causal Inference*, 3, 25–40. [1,4,9]