Permutation Weighting

David Arbour [*1] **Drew Dimmery** [*2 3] **Arjun Sondhi** [4]

Abstract

A commonly applied approach for estimating causal effects from observational data is to is to apply weights which render treatments independent of observed pre-treatment covariates. This property is known as balance; in its absence, estimated causal effects may be arbitrarily biased. In this work we introduce *permutation weighting*, a method for estimating balancing weights using a standard binary classifier (regardless of cardinality of treatment). A large class of probabilistic classifiers may be used in this method; the choice of loss for the classifier implies the particular definition of balance. We bound bias and variance in terms of the excess risk of the classifier, show that these disappear asymptotically, and demonstrate that our classification problem directly minimizes imbalance. Additionally, hyper-parameter tuning and model selection can be performed with standard cross-validation methods. Empirical evaluations indicate that permutation weighting provides favorable performance in comparison to existing methods.

1. Introduction

Observational causal inference methods infer causal effects in the absence of an explicit randomization mechanism. Given observed treatments, outcomes, and a sufficient set of confounding pretreatment covariates, identification of the causal effect is made possible by rendering treatment independent of the covariates (Rubin, 2011; Pearl, 2009). Inverse propensity score weighting (IPW) is a common way to accomplish this, where outcomes are weighted by the inverse probability of receiving the observed treatment given covariates (Hernán and Robins, 2010). If these probabilities correctly represent the conditional distribution, then the weighted data will have independence between treatment and covariates. This property is known as *balance*; for example, in a binary or categorical treatment setting, all treatment groups would have the same weighted distribution of covariates. Unlike in design-based causal inference, where the relationship between treatment and covariates is known by design, propensity scores often must be estimated from observed data. Under model misspecification, however, there are no guarantees of balance, and there may remain arbitrary dependencies between treatment and covariates. Nevertheless, IPW has become widely used in a variety of fields, e.g., epidemiology (Cole and Hernán, 2008), economics (Hirano et al., 2003) and computer science (Dudík et al., 2011).

In this paper, we present permutation weighting (PW), a method for estimating balancing weights for general treatment types by solving a binary classification problem. In contrast to prior work, where the target distribution is implicitly defined via the balance objective, PW explicitly represents the balanced dataset by permuting observed treatments, emulating the target randomized control trial (RCT) (Hernán and Taubman, 2008). As a result, the problem of inferring balancing weights reduces to estimating importance sampling weights between the observed and permuted data. We estimate these importance sampling weights using classifier-based density ratio estimation (Qin, 1998; Cheng et al., 2004; Bickel et al., 2007). This procedure is amenable to general treatment types—binary, multi-valued or continuous—and reduces them all to the same simple binary classification problem which can be solved with off-the-shelf methods. The choice of classifier and specification of the classification problem implies the balance condition. Existing methods (Imai and Ratkovic, 2014; Hazlett, 2016; Fong et al., 2018; Zhao, 2019) with balance constraints correspond to particular choices of loss and feature representations for this classifier. We show that minimizing error in our classification problem directly minimizes the bias and variance of the causal estimator, and imbalance. This property also implies that cross-validation can be used to tune classifier hyperparameters (section 4.1) and choose between balancing weight specifications using standard software (section 4.3).

[*] Equal contribution [1] Adobe Research, San Jose, CA, USA [2] Work carried out while at Facebook Core Data Science, Menlo Park, CA, USA [3] Forschungsverbund Data Science, University of Vienna, Vienna, AT [4] Flatiron Health, New York, NY, USA. Correspondence to: David Arbour <darbour26@gmail.com>.

Proceedings of the 38^{th} *International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

To summarize, this paper makes three contributions to the literature on balancing weights:

1. We show how to use standared probabilistic classifiers for weight estimation.
2. We tie causal estimation to classification error (defined through proper scoring rules), providing justification for using cross-validation for hyperparameter tuning and the selection of balance criteria.
3. The capability to model arbitrary treatment types within the same theoretical and practical framework.

The rest of the paper is structured as follows. Section 2 introduces necessary background and the problem setting of causal inference, balance and weighting methods. Section 3 introduces permutation weighting and in Section 4 we discuss properties of the method. Finally, we evaluate the efficacy of our method for causal inference on binary and continuous treatments in Section 5.

2. Problem Statement and Related Work

We first fix notation used throughout. We denote random variables using upper case, constant values and observations drawn from random variables in lower case, and denote a set with boldface. We will refer to estimates of quantities using hats, e.g., $\hat{w}$ is an estimate of w. Let $\mathcal{D}$ be a dataset consisting of treatments $\mathbf{A}$ defined over a domain $\mathcal{A}$, real valued outcomes $Y \in \mathbb{R}$, and a set of covariates $\mathbf{X}$ defined over a domain $\mathcal{X}$. Note that in our setup, we make no assumption on the cardinality of treatment. Finally, we denote a *potential outcome* as $Y(\mathbf{a})$, which represents an outcome that would have been observed if treatment **a** had been assigned.

We assume the following properties of the observed data throughout this work:

A1. *Weak unconfoundedness (Hirano and Imbens, 2004), i.e., $Y(\boldsymbol{a}) \perp\!\!\!\perp \boldsymbol{A} \mid \mathbf{X} \quad \forall \boldsymbol{a} \in \mathcal{A}$*

A2. *Positivity over treatment status, i.e., there exists a positive constant c such that for all $\boldsymbol{a} \in \mathcal{A}$, $\pi(\boldsymbol{a} \mid \mathbf{X} = \mathrm{x}) \geq c \quad \forall \mathrm{x} \in \mathcal{X}$*

A3. *SUTVA (Imbens and Rubin, 2015): Units' potential outcomes are independent of the realized treatment status of all other units.*

The causal estimand we focus our attention on is the *dose-response function*, $\mathbb{E}[Y(\mathbf{a})]$, i.e., the expected value of the outcome after intervening and assigning treatment to value **a**. This is a general construct that does not presuppose a specific type, e.g. binary, for treatment. Further, the identification of the dose-response function implies identification of many common treatment contrasts of interest. For example, the average treatment effect under binary treatments $\mathcal{A} = \{0, 1\}$, is given as $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

2.1. Balance

A common approach to obtain an unbiased estimate of the dose-response function is to render treatments independent from the confounding variables, $\mathbf{X}$ (Pearl, 2009; Rubin, 2011). Within the causal inference literature, this independence is often referred to as the *balance* condition. We define a general notion of imbalance as some divergence between the observed joint distribution, $p(\mathbf{A}, \mathbf{X})$, and the product distribution, $p(\mathbf{A})p(\mathbf{X})$.

In the binary treatment setting, where balance is most commonly considered, this reduces to performing a two sample test between covariates under treatment and control, i.e. $D(\phi(\mathbf{X}_C), \phi(\mathbf{X}_T)) = 0$, where D denotes some divergence, $\phi(\cdot)$ is some function, and $\mathbf{X}_C$ and $\mathbf{X}_T$ refer to instances of $\mathbf{X}$ associated with control and treatment, respectively.

When treatment is not binary, e.g., continuous or multi-valued, the balance condition must be described explicitly in terms of independence between $\mathbf{A}$ and $\mathbf{X}$, rather than indirectly via the two sample condition. While there are a number of definitions, we will focus on divergences which can be described with an L_p norm of the form

$$\|\mathbb{E}\left[\phi(\mathbf{X}) \otimes \psi(\mathbf{A})\right] - \mathbb{E}\left[\phi(\mathbf{X})\right] \otimes \mathbb{E}\left[\psi(\mathbf{A})\right]\|_p, \qquad (1)$$

where $\otimes$ is the Kronecker product, ϕ and ψ are arbitrary functions, and p is the order of the L_p norm. It may help build intuition to note that when ϕ and ψ are the identity function and $\mathbf{X}$ is univariate, this value is some norm of the covariance between $\mathbf{X}$ and $\mathbf{A}$. In a more general setting when the functions are contained in some reproducing kernel Hilbert space, equation 1 corresponds to the Hilbert-Schmidt independence criterion (Gretton et al., 2005).

2.2. Importance Weighting

A common method for estimating the dose-response function is weighting by the inverse of the conditional probability of receiving treatment given observed covariates, i.e., inverse propensity score weighting (IPW) (Rosenbaum and Rubin, 1983; Imai and Van Dyk, 2004). Weighting by the inverse of this score provides the standard Hájek (or "self-normalized") estimator, which reweights data such that there is no relationship between $\mathbf{A}$ and $\mathbf{X}$, providing identification of causal effects. This is based on the insight that:

$$\mathbb{E}[Y(\mathbf{a})] = \mathbb{E}\left[\frac{y_i \mathbb{1}(\mathbf{a}_i = \mathbf{a})}{p(\mathbf{a} \mid \mathbf{x}_i)}\right] \Big/ \mathbb{E}\left[\frac{\mathbb{1}(\mathbf{a}_i = \mathbf{a})}{p(\mathbf{a} \mid \mathbf{x}_i)}\right],$$

i.e. reweighting individual units provides unbiased estimates of the dose-response surface. In the binary or categorical treatment case, this allows direct aggregation of effects through a weighted average that is consistent for the dose-response. In the continuous case, consistency requires approaches such as Kennedy et al. (2016), which uses local regression to aggregate units together with similar observed dosage. To improve efficiency, many practitioners use the Hájek (1964) estimator which renormalizes weighted averages based on the sum of the weights rather than the number of units; this improves variance in exchange for a small bias which disappears quickly with increasing sample size. When the marginal distribution of treatment is far from uniform, both inverse propensity score weighting and the Hájek estimator can have high variance. To remedy this, Robins (1997) proposed inverse-propensity stabilized weighting (IPSW) which modifies IPW by placing the marginal density of treatment in the numerator, i.e.,

$$\mathbb{E}[Y(\mathbf{a})] = \mathbb{E}\left[\frac{y_i p(\mathbf{a})\mathbb{1}(\mathbf{a}_i = \mathbf{a})}{p(\mathbf{a} \mid \mathbf{x}_i)}\right] \Big/ \mathbb{E}\left[\frac{p(\mathbf{a})\mathbb{1}(\mathbf{a}_i = \mathbf{a})}{p(\mathbf{a} \mid \mathbf{x}_i)}\right].$$

When the conditional distribution has been correctly specified in the propensity score estimation procedure, IPW results in the balance condition (Rosenbaum and Rubin, 1983), i.e. the weighted distribution of $\mathbf{X}$ is the same for all values of $\mathbf{A}$. However, when the conditional distribution is *not* well specified, either in terms of the functional form or the assumed sufficient set of pretreatment covariates, inverse propensity score weighting may fail to produce balance on the observed covariates, and the resulting causal estimate may be badly biased (Harder et al., 2010; Kang and Schafer, 2007).

In this work we revisit the definition of IPSW as

$$\mathbb{E}\left[y_i \frac{p(\mathbf{a}_i)}{p(\mathbf{a}_i \mid \mathbf{x}_i)}\mathbb{1}(\mathbf{a}_i = \mathbf{a})\right] = \mathbb{E}\left[y_i \frac{p(\mathbf{a}_i)p(\mathbf{x}_i)}{p(\mathbf{a}_i, \mathbf{x}_i)}\mathbb{1}(\mathbf{a}_i = \mathbf{a})\right] \tag{2}$$

which makes plain that the weights given by IPSW define importance sampling weights where the target distribution is the distribution under balance. To be explicit, the goal of IPSW is to transform expectations over the observed joint distribution of $\mathbf{A}$ and $\mathbf{X}$ to expectations over $\mathbf{A}$ and $\mathbf{X}$ in which they appear as if generated from an RCT (the "target trial"). However, the importance sampling weights under IPSW are constructed indirectly by separately estimating the conditional and marginal treatment densities. The contribution of this work is a method, permutation weighting, which estimates this quantity directly via a probabilistic classification problem which we describe in the next section. Direct estimation provides more than just intuitive appeal. Unlike IPSW, direct estimation of the importance sampling ratio explicitly seeks to minimize imbalance, which we show in section 4. We also show that this approach leads to bounds on bias and variance of the dose-response estimates based on classifier error. The result is that bias is reduced under direct estimation of the density ratio, even in the case of misspecification.

2.3. Balancing Weights

Covariate balancing weights (Hainmueller, 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015) seek to remedy the problems of imbalance under misspecification by optimizing a balance condition directly. The promise of these techniques is that even when the propensity score model is misspecified, the method will still reduce confounding bias by optimizing for its respective balance condition. As the balancing weight literature grows, proposed methodologies are differentiated largely by two aspects: (1) the choice of the distance employed as a measure of balance, and (2) the optimization procedure. This presents a challenge for practitioners, since the appropriate measure of balance is application-specific and many of the proposed optimization procedures (e.g. Zubizarreta (2015); Hainmueller (2012)), have hyperparameters (e.g., δ for stable balancing weights, and the strength of the entropy penalty for entropy balancing) which must be manually specified and can significantly affect performance. In addition, the aforementioned work focuses on the binary treatment regime, but many applied problems are not simple dichotomous treatments. While Fong et al. (2018) provides a linear-balancing method for general treatments, this task still requires hard choices for practitioners about how to specify balance and how to parameterize the conditional distribution of treatment. Thus, providing a unified framework for comparing balancing weight estimators is critical for effective application. Zhao (2019) provides one step in this direction by unifying many existing balancing weights for binary treatments by considering proper scoring rules, but does not provide guidance for model selection.

3. Permutation Weighting

We now introduce permutation weighting, which allows for the direct estimation of the importance sampler defined by equation 2. Permutation weighting consists of two steps:

1. The original dataset is stacked with a dataset in which $\mathbf{A}$ has been permuted. The permuted dataset is equivalent to fixed-margin randomization of treatment, so represents a distribution where $\mathbf{A}$ and $\mathbf{X}$ are independent. That is, it obeys the balance condition *by design*. In what follows, we denote the distribution that the observed data is drawn from as P, and the product distribution resulting from permutation as Q.

2. The importance sampling weights, $\hat{w}(\mathbf{a}_i, \mathbf{x}_i)$, are constructed by estimating the density ratio between P and Q.

In order to estimate the density ratios (step 2), we employ classifier-based density ratio estimation (Qin, 1998; Cheng et al., 2004; Bickel et al., 2007), which transforms the problem of density ratio estimation into binary classification by building a training set from the concatenation of the observed and permuted datasets. $\mathbf{A}$ and $\mathbf{X}$ are used as features, and a label, $C \in \{0, 1\}$, is given to denote the membership of the instance to the observed or the permuted dataset, respectively. A probabilistic classifier learns to recover $p(C = 1 \mid \mathbf{A}, \mathbf{X})$. We denote the true conditional probability as η and the estimated conditional probabilities from the classifier as $\hat{\eta}$. To aid discussion, with some abuse of notation, we will also refer to the classifier which produced the conditional probabilities as $\hat{\eta}$. After training the classifier, assuming equally sized observed and permuted datasets, the importance weights are recovered by taking the density of the distribution of $\mathbf{A}$ and $\mathbf{X}$ in the permuted dataset (dQ) over the density of the observed joint distribution (dP) (Bickel et al., 2007):

$$w(\mathbf{a}_i, \mathbf{x}_i) = \frac{\eta(\mathbf{a}_i, \mathbf{x}_i)}{1 - \eta(\mathbf{a}_i, \mathbf{x}_i)} = \frac{p(C = 1 \mid \mathbf{a}_i, \mathbf{x}_i)}{p(C = 0 \mid \mathbf{a}_i, \mathbf{x}_i)}$$

$$= \frac{p(C = 1, \mathbf{a}_i, \mathbf{x}_i)}{p(C = 0, \mathbf{a}_i, \mathbf{x}_i)} \frac{(p(C = 1)dQ + p(C = 0)dP)}{(p(C = 1)dQ + p(C = 0)dP)} = \frac{dQ}{dP}$$

When dQ breaks dependence between $\mathbf{A}$ and $\mathbf{X}$ (e.g. permuting by treatment assignment or specifying the full cross-product), the resulting importance sampler is $\frac{p(\mathbf{a}_i)p(\mathbf{x}_i)}{p(\mathbf{a}_i, \mathbf{x}_i)}$. The use of a probabilistic classifier for density ratio estimation has a growing literature (Sugiyama et al., 2012; Menon and Ong, 2016; Mohamed and Lakshminarayanan, 2016), and was used by Yamada and Sugiyama (2010) for inferring the causal direction between two random variables, but it has yet to be employed in the context of observational causal inference.

We define the classifier loss as some function $\lambda : \{-1, 1\} \times [0, 1] \mapsto \mathbb{R}_+$. Throughout we use λ_1 and λ_{-1} to refer to the losses for the permuted and unpermuted classes, respectively. Using $\mathcal{D}$ to denote the distribution of the stacked dataset over which the classifier is trained on (an equal mixture of P and Q), the risk for the classifier $\hat{\eta}$ under loss λ is then defined as

$$\mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda) = \mathbb{E}_P \left[\lambda_1(\hat{\eta}(a, \mathbf{x}))\right] + \mathbb{E}_Q \left[\lambda_{-1}(\hat{\eta}(a, \mathbf{x}))\right]$$

The Bayes risk is given as $\mathbb{L}^*(\mathcal{D}, \lambda) = \min_{\hat{\eta}} \mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda)$ (Reid and Williamson, 2011; Menon and Ong, 2016). The regret is defined as the difference between risk of a classifier, $\hat{\eta}$ and the Bayes risk, $\mathrm{reg}(\hat{\eta}; \mathcal{D}, \lambda) = \mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda) - \mathbb{L}^*(\mathcal{D}, \lambda)$.

In order to ensure that the probabilities produced by the classifier are well calibrated, we introduce the following assumption:

A4. *The classifier, $\hat{\eta}$, is trained using a twice differentiable strictly proper scoring rule, i.e., $\hat{\eta} \neq \eta \implies \mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda) > \mathbb{L}(\eta; \mathcal{D}, \lambda)$ (Buja et al., 2005; Gneiting and Raftery, 2007).*

More intuitively, strictly proper scoring rules define functions which, when minimized, provide calibrated forecasts. The most common examples of strictly proper scoring rules are logistic, exponential, and quadratic losses (Gneiting and Raftery, 2007). Strictly proper scoring rules also provide a natural connection to statistical divergences: every proper scoring rule is associated with a divergence between the estimated and true forecasting distribution (Reid and Williamson, 2011; Huszar, 2013). Finally, we also assume consistency of the classifier under the data-generating process.

A5. *The classifier error, $\hat{\eta} - \eta$, scales uniformly as $O(n^{-\epsilon}), \epsilon \in (0, 1)$.*

If the permuted dataset obeys the balance condition, then the weights will target balance. In finite samples, this dataset may not have perfect balance, so we perform multiple permutations where the classification procedure is carried out to obtain weights, which are averaged to provide the final estimate of the weight. Justification for this procedure is provided in the proof of Corollary A.2 in Appendix A, which relies on the fact that each permutation is a random sample from the ideal balanced distribution. For low cardinality treatments, the permuted dataset can be constructed as the cross product of the unique values of treatment with $\mathbf{X}$, and no iteration is necessary.

Inferring weights using a classifier confers three important advantages:

1. Regardless of the type of the treatment (binary, continuous, multinomial, etc.), the problem reduces to the same binary classification task. In contrast to many existing methods, this means that it is not necessary to explicitly assume a parametric form for the treatment conditional on covariates (for instance, generalized propensity scores often assume that dosage is conditionally normal). As such, the use of binary classification to directly estimate weights requires weaker assumptions in environments with complicated treatments.

2. As we discuss in section 4, minimizing the error of the binary classifier directly results in minimizing both imbalance (proposition 4.4) and the error of the causal estimate itself (propositions 4.1 and 4.2). As a result, both the hyperparameters and the measure of balance itself (via the choice of feature representation and loss) can be optimized directly by considering the cross-validated error of the binary classifier. In addition to theory, we demonstrate the empirical efficacy of hyperparameter tuning and model selection for estimating causal effects in section 5.3.

3. There is a deep connection between binary classification and two sample testing (Friedman, 2004; Reid and Williamson, 2011). Through this lens, the choice of feature representation and classification loss is equivalent to choosing a balance condition.

4. Properties

We now examine the finite sample and asymptotic behavior of permutation weighting. To do so, we will first consider a slightly more general setting than the procedure outlined in the previous section. Specifically, propositions 4.1, 4.2, 4.3 and 4.4 examine the behavior of importance sampling from the observed distribution P to an arbitrary distribution Q (under positivity, assumption A2), using a classifier trained with a strictly proper scoring rule (assumption A4). These may be of independent interest as they admit reasoning over a broad class of estimands (e.g. Bickel et al., 2007; Sugiyama et al., 2012; Menon and Ong, 2016), including common causal estimands like the average treatment effect on the treated. Indeed, any distribution of $\mathbf{A}$ and $\mathbf{X}$ which conforms to the overlap assumption can be used as a target distribution under this framework. Before presenting our results, we first introduce Bregman divergences, a class of statistical distances.

Definition 1 (Bregman divergence (Bregman, 1967)). *Define the Bregman generator, $g : S \to \mathbb{R}$, to be a convex, differentiable function. The difference between the value of g at point s and the value of the first-order Taylor expansion of g around point s_0 evaluated at point s is given by $B_g(s, s_0) \equiv g(s) - g(s_0) - \langle s - s_0, \nabla g(s_0)\rangle$.*

Minimizing many commonly used classification losses correspond to minimizing a Bregman divergence, e.g., accuracy (0-1) loss corresponds to total variation distance, squared loss corresponds to triangular discrimination distance, log loss corresponds to the Jensen-Shannon divergence, and exponential loss corresponds to the Hellinger distance (Reid and Williamson, 2011). The latter three losses are proper scoring rules and conform to our assumption A4. All strictly proper scoring rules have a corresponding Bregman divergence (Dawid, 2007). We use this correspondence in the proofs of our theoretical results, which are generally deferred to the supplement.

4.1. Estimation

Throughout this section we will denote the target weights for the permutation weighting importance sampler as w and the estimated weights $\hat{w}$. We now begin by deriving bounds on the bias for weighting estimators.

Proposition 4.1 (Bias of PW). *Let $\mathbb{E}_P$ and $\mathbb{E}_Q$ denote the expectation under the distributions P and Q, respectively. The bias of the dose response function $\mathbb{E}_P[y\hat{w}]$ with respect to $\mathbb{E}_Q[y]$ is bounded by*

$$|\mathbb{E}_Q[y] - \mathbb{E}_P[y\hat{w}]| \leq \mathbb{E}_P\left[\frac{2|y|}{\sqrt{g''(1)}}\kappa_r\right],$$

where $g(\cdot)$ is a Bregman generator and $\kappa_r = \sqrt{\text{reg}(\hat{\eta}; \mathcal{D}, \lambda)}$

Minimizing this bound corresponds to minimizing the regret of the classsifier. We next bound the variance of the permutation weighting dose-response estimator.

Proposition 4.2 (Variance). *Let $\mathbb{V}_Q[y]$ denote the variance of Y under the distribution q. $\mathbb{V}_Q[y]$ is bounded by*

$$\mathbb{V}_Q[y] \leq \frac{1}{n}\mathbb{E}_Q[y^2] + \frac{4\kappa_r}{n\sqrt{g''(1)}}\mathbb{E}_P\left[y^2 w + \frac{y^2}{\sqrt{g''(1)}}\kappa_r\right]$$

where $\kappa_r = \sqrt{\text{reg}(\hat{\eta}; \mathcal{D}, \lambda)}$

The bounds given by propositions 4.1 and 4.2 demonstrate that the quality of importance sampling weights is governed by the regret of the classifier used. For KL divergence, the bound given by proposition 4.1 is essentially Pinsker's inequality (Reid and Williamson, 2010).

Finally, consistency of the importance sampler used by permutation weighting is given by the following proposition, which follows as a consequence of propositions 4.1 and 4.2:

Proposition 4.3 (Consistency). *Under Assumptions A1-A5, and bounded outcomes y, the permutation weighting dose-response estimator is consistent, i.e., as $n \longrightarrow \infty$, $\mathbb{E}_P[y\hat{w}] \longrightarrow \mathbb{E}_Q[y]$.*

4.2. Balance

We next show how the importance sampler provided by permutation weighting provides balance. We preface this with a general definition of balance:

Definition 2 (Functional discrepancy). *The L_p functional discrepancy for functions ϕ and ψ, under a weighting estimator $\hat{w}$ is*

$$\|\mathbb{E}_P\left[\phi(\mathbf{a}_i) \otimes \psi(\mathbf{x}_i)\hat{w}(\mathbf{a}_i, \mathbf{x}_i)\right] - \mathbb{E}_Q\left[\phi(\mathbf{a}_i) \otimes \psi(\mathbf{x}_i)\right]\|_p .$$

When the target distribution can be factored as $p(\mathbf{A})p(\mathbf{X})$, this is a measure of imbalance:

$$\|\mathbb{E}_P\left[\phi(\mathbf{a}_i) \otimes \psi(\mathbf{x}_i)\hat{w}(\mathbf{a}_i, \mathbf{x}_i)\right] - \mathbb{E}_P\left[\phi(\mathbf{a}_i)\right] \otimes \mathbb{E}_P\left[\psi(\mathbf{x}_i)\right]\|_p .$$

This quantity is the extent to which the reweighted expectation differs from the expectation under the product distribution. A functional discrepancy of zero *for all* ϕ and ψ implies independence between $\mathbf{A}$ and $\mathbf{X}$. When both ϕ

and ψ are the identity function, then a discrepancy of zero is synonymous with linear balance. With this definition in hand, we can provide an explicit expression for the functional imbalance attained by permutation weighting:

Proposition 4.4 (Minimizing Imbalance)**.** *The L_p functional discrepancy between the observed data drawn from $p(\mathbf{a}, \mathbf{x})$ and the proposed distribution $q(\mathbf{a}, \mathbf{x})$ under permutation weighting is*

$$\begin{aligned}&\left\|\mathbb{E}_{p(\mathbf{a},\mathbf{x})}\left[\phi(\mathbf{a}_i) \otimes \psi(\mathbf{x}_i)(\hat{w}(\mathbf{a}_i, \mathbf{x}_i) - w(\mathbf{a}_i, \mathbf{x}_i))\right]\right\|_p \\ &\leq \frac{2}{\sqrt{g''(1)}}\kappa_r \left\|\mathbb{E}_{p(\mathbf{a},\mathbf{x})}\left[\phi(\mathbf{a}_i) \otimes \psi(\mathbf{x}_i)\right]\right\|_p\end{aligned}$$

where $p \geq 0$ and $\kappa_r = \sqrt{\mathrm{reg}(\hat{\eta}; \mathcal{D}, \lambda)}$.

Proposition 4.4 demonstrates the balancing behavior of the importance sampler employed by permutation weighting. The importance sampler defined by permutation weighting has a *linear* dependence on the error of the density ratio estimate which is minimized as classifier regret is minimized.

The above demonstrates properties for estimating importance weights to any pre-specified joint distribution, Q, of $\mathbf{A}$ and $\mathbf{X}$. We now focus our attention on the distribution $p(\mathbf{A})p(\mathbf{X})$ – that of marginal-preserving independence between treatment and covariates. With a low cardinality treatment (such as binary), it's possible to directly construct a balanced dataset to satisfy $p(\mathbf{A})p(\mathbf{X})$. This is done by taking the cross-product of the unique values of $\mathbf{A}$ and of $\mathbf{X}$. Rows can then be weighted to match the marginals from the original data. Each row in the pseudo-dataset, i, would receive a weight of $\frac{1}{n^2}n(\mathbf{a}_i)n(\mathbf{x}_i)$, where $n(\mathbf{a}_i)$ denotes the number of rows in the original dataset with treatment level $\mathbf{a}_i$, and likewise for $n(\mathbf{x}_i)$. However, if $\mathbf{A}$ has larger cardinality, it may be computationally difficult or (in the case of continuous treatments) impossible to construct such a dataset. Instead, the easiest way to target this distribution is through a simple permutation, in which the treatment vector is reshuffled. By design, this permuted dataset will have no systematic relationship between $\mathbf{A}$ and $\mathbf{X}$ except for that which occurs by chance. This is true for the same reason that fixed-margins randomization in RCTs attains balance at expectation. This easy permutation construction allows the application of all the properties of density ratio estimation discussed above. Asymptotically in n, a single permutation will (like data observed from an RCT) converge to the appropriate balanced target distribution. In finite samples, there may remain minor imbalances from a single permutation. For this reason, we propose averaging across multiple permutations to attain an effective balancing weight. Theoretical justification for this procedure is given in the supplement.

4.3. Choosing among scoring rules

In contrast to existing work which requires a priori specification of the balance condition, the choice of the balance condition can be performed by considering the out of sample performance of the classifier with respect to the receiver operator characteristic (ROC) curve, a common measure of classifier performance. From the bounds given in propositions 4.1 and 4.2, we can select the condition which minimizes the error of the causal estimate. By noting the connection between the choice of classifier loss and two sample discrepancies provided by Reid and Williamson (2011), this procedure also corresponds to choosing a balance condition. This brings us to Proposition 13 of (Menon and Williamson, 2016), that stochastic dominance of the ROC curve for one classifier over another implies dominance with respect to *any* strictly proper scoring rule.

That is, when two ROC curves do not cross, the one with higher true positive rates across all false negative rates will also be superior according to *all* proper scoring rules. This property shows how the AUC is an effective diagnostic to choose between classifiers. When the ROC curves for two classifiers cross one another, it may be the case that different choices of loss would suggest different “optimal” classifiers. Within the context of this paper, these results imply that the choice of a balance criterion can be made by examining the ROC curves produced by different modeling choices. These modeling choices correspond to the estimation of different balancing weights.

5. Experiments

For the following simulation studies, we will examine only performance of simple weighting estimators for scalar-valued treatments, $E[Y(a)] \approx \sum_{i=1}^n y\hat{w}(a_i, \mathbf{x}_i)K(a_i, a)$. For binary treatments, $K(\cdot, \cdot)$ is an indicator for treatment status, while for continuous treatments, it is a kernel weighting term as analyzed in the context of doubly-robust estimators by Kennedy et al. (2016). This simple estimator is used in our evaluation to provide the most direct test of the efficacy of the various estimators of the weights. Appendix D provides results for the doubly-robust estimators of Kennedy et al. (2016) as well as the estimation of weighted outcome regressions. These more complex evaluations do not differ in their substantive conclusions (i.e. rank-order and relative performance). Error is measured via integrated root mean squared error (IRMSE) as in Kennedy et al. (2016), with s indexing S simulations and $\theta_s(a)$ being the unconditional expectation of a given potential outcome in a single simulation, $\mathbb{E}_s[Y(a)]$, i.e.,

$$\widehat{\mathrm{IRMSE}} = \int_{\mathcal{A}^*}\left[\frac{1}{S}\sum_{s=1}^{S}\{\hat{\theta}_s(a) - \theta_s(a)\}^2\right]^{\frac{1}{2}} p(a)da$$

That is, we take an average of RMSE weighted over the marginal probability of treatment. Following Kennedy et al. (2016), we perform this evaluation over $\mathcal{A}^*$, the central 90% of the distribution of A (in the case of binary treatments, we evaluate over the entire support of A). We also evaluate the Integrated Mean Absolute Bias, which replaces the inner average with $\left|\frac{1}{S}\sum_{s=1}^{S}\{\hat{\theta}_s(a) - \theta_s(a)\}\right|$. When permutation weighting is used, we perform 100 independent iterations of the permutation procedure to generate weights. Our evaluations center around two main classifiers: logistic regression and gradient boosted decision trees. The former focuses on minimizing a log-loss and therefore the balance condition corresponds to minimization of the Jensen-Shannon divergence. The boosting model corresponds to an exponential loss (Lebanon and Lafferty, 2002) which implies the minimization of the Hellinger divergence. Achieving equivalence to linear balancing methods using the permutation weighting framework entails the addition of an interaction term between A and $\mathbf{X}$ due to the different setup of the classification problem; otherwise, the linear classifier would only be able to account for differing marginal distributions of A and $\mathbf{X}$ (asymptotically, there are no such differences). We include this interaction in all of the models we evaluate.

5.1. Binary treatment simulation

Our first simulation study follows the design of Kang and Schafer (2007). In this simulation, four independent, standard normal covariates are drawn. A linear combination of them is taken to form the outcome and treatment process (the latter passed through an inverse-logistic function). In this simulation, we induce misspecification by observing only four non-linear and interactive real-valued functions of the covariates. See Appendix D for more details and results on a correctly-specified model.

Figure 1 shows results for the realistic case in which the researcher does not know the correct specifications of the confounding relationships of the covariate set with treatment. In these results, PW with boosting drastically improves on the existing weighting estimators, reducing by around 30% the IRMSE relative to balancing propensity scores at $n = 2000$. At smaller sample sizes, the improvements are less substantial, but even by $n = 500$, PW with boosting provides superior performance. This is unsurprising, given that boosting is able to learn a more expressive balancing model (and, therefore, reduce bias) more effectively than other balancing methods. A standard propensity score estimated by gradient boosted decision trees does not solve the issues faced by propensity scores, leading to large biases in estimation and subsequently large IRMSE across all sample sizes. Detailed results in tabular format are available in Appendix D. A similar simulation study on a continuous treatment is detailed in Appendix C.

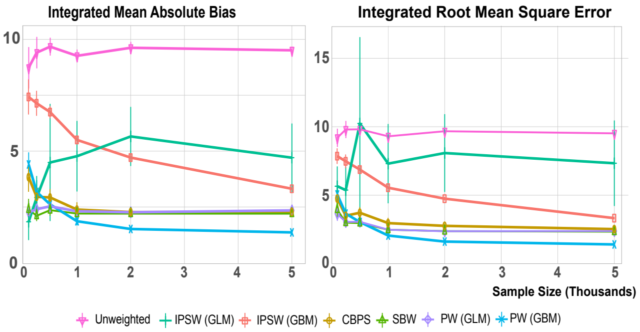

Figure 1: On the Kang and Schafer (2007) simulation under misspecification of confounding variables, PW attains state-of-the-art performance in both bias and RMSE by $n = 500$ and reduces RMSE by 30% at $n = 2000$. `Unweighted` uses no weighting. `IPSW (GLM)` is a logistic-regression-based propensity score model. `IPSW (GBM)` is a propensity score model trained with a gradient boosted decision tree. `CBPS` is covariate balancing propensity scores (Imai and Ratkovic, 2014). `SBW` is stable balancing weights (Zubizarreta, 2015). `PW (GLM)` is a permutation weighting model using a logistic regression. `PW (GBM)` is a permutation weighting model using a gradient boosted decision tree.

5.2. Lalonde evaluation with continuous treatment

To explore the behavior of permutation weighting under continuous treatment regimes with irregularly distributed data, we turn to the data of LaLonde (1986), and in particular, the Panel Study of Income Dynamics observational sample of 2915 units (discarding all treated units from the sample and retaining only the experimental control units). Our evaluation is based around the differences between the experimental control group and the observational control group, which are known to differ greatly based on observed covariates (Smith and Todd, 2005). The covariates in this data are highly non-ellipsoidal, consisting of point-masses and otherwise irregular distributions. Following the simulation study in Diamond and Sekhon (2013), we simulate a nonlinear process determining assignment of units to dosage level and, then, to outcome based on observed covariates (full details in Appendix F). The treatment process is made to behave similarly to real-world data by estimating a random forest to predict presence in the experimental / observational sample as a function of observed covariates. Dosage is then a quartic function of that predicted score as well as the nonlinear function determining treatment assignment in Diamond and Sekhon (2013). The shape of the true dose-response function is similarly a quartic function of dose.

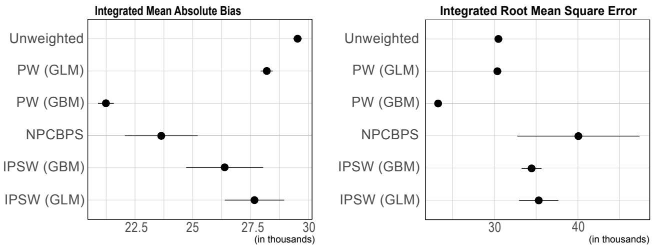

Figure 2: On the continuous-treatment simulation based on the LaLonde data, PW greatly reduces both bias and variance relative to existing methods. `Unweighted` uses no weighting. `PW (GLM)` is a permutation weighting model using a logistic regression. `PW (GBM)` is a permutation weighting model using a gradient boosted decision tree. `NPCBPS` is non-parametric covariate balancing propensity scores (Fong et al., 2018). `IPSW (GLM)` is a normal-linear regression based propensity score model. `IPSW (GBM)` is a gradient boosted regression generalized propensity score model with homoskedastic normal conditional densities.

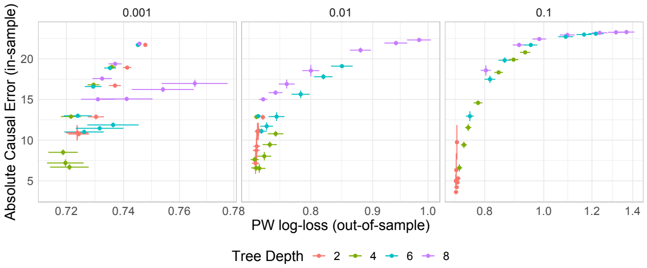

Figure 3: The estimated GBDT classifier error out-of-sample correlates strongly with the error of the causal estimate over a grid of hyperparameter values. The y-axis represents in-sample causal error, while the x-axis is the out-of-sample PW loss (i.e. out-of-sample imbalance). Hyperparameters tuned were the tree-depth of each decision tree (color), the learning rate, ν (columns) and the number of trees (not annotated, from 100 to 5000).

Figure 2 shows the IRMSE of a variety of weighting estimators on this simulated benchmark. Only weights generated by permutation weighting out-perform the raw, unweighted data in terms of IRMSE. All models reduce bias relative to the raw, unweighted dose-response, but induce unacceptably large variance as they do so. When a logistic regression is used as the classifier, PW performs better than no weighting by just half a percent in terms of IRMSE (as it does not greatly reduce bias). When a boosted model is used, however, this gap grows substantially, with PW out-performing the raw estimates by around 25% – the only substantial improvement in accuracy among these estimators. As the earlier simulations have shown, using a boosting model to estimate a standard propensity score does not perform well, increasing IRMSE relative to the unweighted estimate. It's also worth noting the much reduced variability around the estimates of IRMSE from the permutation weighting models relative to other methods which often have very unstable performance characteristics. Rank ordering among methods remains largely unchanged when an outcome model is incorporated (see appendix F).

5.3. Cross-validation

In this section, we demonstrate how cross-validation may be used to effectively tune the permutation weight model. For this experiment, we took one instance of the Kang-Schafer simulation (with a correctly specified linear model) described in section 5.1 with a sample size of 2000 and performed 10-fold cross-validation on this data, measuring both in and out-of-sample errors. Presented in figure 3 are the results of this exercise for out-of-sample PW error and in-sample PW error, respectively. Appendix E shows the ROC curve for three models and demonstrates weighting-model selection on that basis.

In the traditional causal inference environment, practitioners care about the in-sample causal error, rather than generalization error to other potential samples. Minimizing classifier error, through proposition 4.4, minimizes imbalance. Minimization of in-sample imbalance may seem desirable since any residual imbalance can lead to bias in a purely weighting estimator. Our results are shown in figure 3 (with a similar figure for in-sample error in Appendix D). The primary takeaway from these results is that generalization performance of weights *does* matter. Simply minimizing in-sample imbalance is not necessarily the way to best optimize estimation accuracy. More important than in-sample imbalance is *out-of-sample* imbalance, which can be consistently measured through the PW error. We can see clearly that while improving the PW loss out-of-sample brings with it corresponding improvements in in-sample causal error, this is not true for in-sample PW loss. This represents classic over-fitting to the sample. Importantly, looking at the error of permutation-weighting out-of-sample gives a reliable way to assess the quality of fit: choosing a permutation weighting model which generalizes well will, in turn, ensure that in-sample causal error is minimized. In short, permutation weighting provides a reliable framework through which to use cross-validation for model selection and hyperparameter tuning for weighting.

6. Conclusion

Weighting is one of the most commonly applied estimators for causal inference. This work provides a new lens on weighting by framing the problem in terms of importance sampling towards the distribution of treatment and covariates that would be observed under randomization. Through this lens we introduced permutation weighting, which casts the balancing weights problem into generic binary classification and allows the standard machine learning toolkit to be applied to the problem. We show that regret in this classification problem bounds the bias and variance of the causal estimation problem. Thus, methods for regularization and model selection from the supervised learning literature can be used directly to manage the bias-variance tradeoff of this causal effect estimation problem. Permutation weighting generalizes existing balancing schemes, admits selection via cross-validation, and provides a framework to sensibly integrate generic treatment types within the same weighting estimator. Simulations show that permutation weighting outperforms existing estimation methods even in conditions unfavorable to the assumptions underlying the model.

References

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7. URL http://www.sciencedirect.com/science/article/pii/0041555367900407.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical Report*, 2005.

Kuang Fu Cheng, Chih-Kang Chu, et al. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

Stephen R. Cole and Miguel A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008. doi: 10.1093/aje/kwn164. URL http://dx.doi.org/10.1093/aje/kwn164.

A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1): 77–93, 2007.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

Christian Fong, Chad Hazlett, Kosuke Imai, et al. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

Jerome Friedman. On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA (US), 2004.

Gustavo L Gilardoni. On pinsker's type inequalities and csiszár's f-divergences. part i: Second and fourth-order inequalities. *arXiv preprint cs/0603097*, 2008.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schoelkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Scholkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, pages 131–160, 2009.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20 (1):25–46, 2012.

Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.*, 35(4):1491–1523, 12 1964. doi: 10.1214/aoms/1177700375. URL https://doi.org/10.1214/aoms/1177700375.

Valerie S Harder, Elizabeth A Stuart, and James C Anthony. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3):234, 2010.

Chad Hazlett. Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *arXiv preprint arXiv:1605.00155*, 2016.

Miguel A Hernán and James M Robins. Causal inference, 2010.

Miguel A Hernán and Sarah L Taubman. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *International journal of obesity*, 32(S3):S8, 2008.

Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

Ferenc Huszar. *Scoring rules, divergences and information in Bayesian machine learning*. PhD thesis, University of Cambridge, 2013.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 143–151. Morgan Kaufmann Publishers Inc., 1997.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Edward H. Kennedy, Zongming Ma, Matthew D. McHugh, and Dylan S. Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2016. doi: 10.1111/rssb.12212. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12212.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Guy Lebanon and John D Lafferty. Boosting and maximum likelihood for exponential models. In *Advances in neural information processing systems*, pages 447–454, 2002.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313, 2016.

Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *The Journal of Machine Learning Research*, 17(1):6766–6867, 2016.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.

Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(Mar):731–817, 2011.

James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2011.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

Le Song. *Learning via Hilbert space embedding of distributions*. University of Sydney, 2008.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert RG Lanckriet, and Bernhard Schölkopf. Injective hilbert space embeddings of probability measures. In *COLT*, volume 21, pages 111–122, 2008.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Raymond KW Wong and Kwun Chuen Gary Chan. Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213, 2017.

Makoto Yamada and Masashi Sugiyama. Dependence minimizing regression with model selection for non-linear causal inference under non-gaussian noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2): 965–993, 2019.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.