Counterfactual Representation Learning with Balancing Weights

Serge Assaad[1] **Shuxi Zeng**[2] **Chenyang Tao**[1] **Shounak Datta**[1]
Nikhil Mehta[1] **Ricardo Henao**[1] **Fan Li**[2] **Lawrence Carin**[1]
[1]Department of ECE, Duke University [2]Department of Statistical Science, Duke University

Abstract

A key to causal inference with observational data is achieving balance in predictive features associated with each treatment type. Recent literature has explored representation learning to achieve this goal. In this work, we discuss the pitfalls of these strategies – such as a steep trade-off between achieving balance and predictive power – and present a remedy via the integration of balancing weights in causal learning. Specifically, we theoretically link balance to the quality of propensity estimation, emphasize the importance of identifying a proper target population, and elaborate on the complementary roles of feature balancing and weight adjustments. Using these concepts, we then develop an algorithm for flexible, scalable and accurate estimation of causal effects. Finally, we show how the learned weighted representations may serve to facilitate alternative causal learning procedures with appealing statistical features. We conduct an extensive set of experiments on both synthetic examples and standard benchmarks, and report encouraging results relative to state-of-the-art baselines.

1 INTRODUCTION

Solving many scientific, engineering, and socioeconomic problems – *e.g.*, personalized healthcare (Glass et al., 2013; Johnson et al., 2018), computational advertising (Chan et al., 2010), complex systems (Chen et al., 2020), and policymaking (Athey, 2015) – requires an understanding of cause and effect beyond observed associations. Consequently, the study of *causal inference* (Pearl, 2009; Rubin, 2005) is central to various disciplines and has received growing attention in the machine learning community. To exploit the new opportunities and cope with the challenges brought by modern datasets, various new causal inference methods have been proposed (Shalit et al., 2017; Yoon et al., 2018; Louizos et al., 2017; Hassanpour & Greiner, 2019; Johansson et al., 2018, 2020; Li & Fu, 2017; Alaa & van der Schaar, 2018, 2017).

This paper focuses on predicting conditional average treatment effects (CATE) from observational data, defined as the difference between an individual's expected potential outcomes for different treatment conditions. This problem differs fundamentally from standard supervised learning (Pearl, 2009; Rubin, 2005), because for each unit only the potential outcome corresponding to the assigned treatment is observed and the other potential outcome is missing. The absence of the "counterfactual" outcome prohibits the direct learning and validation of causal effects. Further, observational studies are subject to selection bias due to confounders (Heckman, 1979) – variables that affect both the treatment assignment and the outcomes. Within the associated data this is typically manifested as covariate imbalance (Shalit et al., 2017), *i.e.*, treatment-dependent distributions of covariates. Without careful adjustment, this leads to a biased estimate of the causal effect (Zubizarreta, 2015).

Mitigation of covariate imbalance in high-dimensional spaces has motivated representation learning schemes for causal inference that seek balance in the learned feature space (Shalit et al., 2017; Johansson et al., 2016). Despite the empirical success of such methods, it has been recognized that over-enforcing balance can be harmful, as it may inadvertently remove information that is predictive of outcomes (Alaa & van der Schaar, 2018). To see this, one may consider an example where a moderately predictive feature might get erased in the learned representation for being highly imbalanced. As such, representation learning-based schemes are sensitive to the hyperparameter that tunes the desired level of imbalance mitigation.

More classical causal inference approaches seek to

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

match the statistics of the covariates associated with both treatment types (Pearl, 2009; Lunceford & Davidian, 2004; Rubin, 2005; Holland, 1986). Matching methods create a balanced sample by searching for "similar" units from the opposite treatment group (Stuart, 2010). Matching unfortunately does not scale well to higher dimensions (Abadie & Imbens, 2006), and will often improve balance for some covariates at the expense of balance for others. Weighting methods assign to each unit a different importance weight so as to match the covariate distributions in different treatment arms after reweighting (Li et al., 2018; Lunceford & Davidian, 2004). In much of the causal inference literature, weighting is employed for *average* treatment effect (ATE) estimation over a population.

In this paper, we employ weighting for *conditional* average treatment effect (CATE) estimation. In this context we demonstrate the advantages of learning from regions of good overlap, achieved by employing weighting prior to representation learning. We investigate the coupling of weighting methods (Li et al., 2018; Zubizarreta, 2015; Hassanpour & Greiner, 2019; Johansson et al., 2018) with representation-based causal inference, and demonstrate how the use of properly designed weights alleviates the aforementioned difficulties of representation learning applied to causal inference. We show how targeting an alternative population for empirical loss minimization (Li et al., 2018) benefits CATE estimation. As discussed below, if appropriately designed weights are learned perfectly, then balance is achieved for *any* features constituted from the covariates (since balance is achieved in the covariates themselves). However, most weighting methods are computed from the propensity score (D'Agostino, 1998), which must be *approximated* numerically. Because in practice the weights are always imperfect, exact balance is rarely achieved based on weighting alone, motivating our augmentation of weighting with representation learning.

This paper makes the following contributions: (*i*) demonstration that the integration of balancing weights alleviates the trade-off between feature balance and predictive power for representation learning; (*ii*) derivation of theoretical results bounding the degree of imbalance as a function of the quality of the propensity model; (*iii*) exploration of the benefits of the learned weights and representations as inputs to other learning procedures such as causal forests. We demonstrate that our method, *Balancing Weights Counterfactual Regression (BWCFR)*, mitigates the weaknesses of propensity-weighting and representation learning. In this approach, we do not impose that the features themselves be balanced, as this would likely result in loss of information. Instead, we promote balance for *reweighted* feature distributions, with weights targeting regions for which there is already good overlap.

2 RELATED WORK

Representation learning has been used to achieve balance between treatment group distributions, seeking representations that are both predictive of potential outcomes, and balanced across treatment groups (Kallus, 2018; Shalit et al., 2017). Zhang et al. (2020) argue that there is often a tradeoff between these objectives, and that over-enforcing balance leads to representations that are less useful for outcome prediction – our proposal mitigates this tradeoff by enforcing balance between *weighted* feature distributions. Our theory on the discrepancy between the treatment arm distributions (Propositions 2 and 3) is also conceptually related to sensitivity modeling in causal analysis (Kallus et al., 2019).

Weighting-based methods typically construct weights as a function of the propensity score to balance covariates (Rosenbaum & Rubin, 1983; Lunceford & Davidian, 2004), such as inverse probability weighting (IPW). The performance of these methods critically depends on the quality of the propensity score model and is highly sensitive to the extreme weights (Hainmueller, 2012). To overcome these limitations, alternative weighting schemes such as Matching Weights (Li & Greene, 2013), Truncated IPW (Crump et al., 2009) or Overlap Weights (Li et al., 2018) seek to change the target population, thereby eliminating extreme weights. Another popular line of solutions directly incorporates covariate balance in constructing the weights (Graham et al., 2012; Diamond & Sekhon, 2013), and usually calculate weights via an optimization program with moment matching conditions as the hard (Li & Fu, 2017; Hainmueller, 2012; Imai & Ratkovic, 2014) or soft constraints (Zubizarreta, 2015). While these bypass propensity score modeling and hence are no longer afflicted by extreme weights, they struggle to scale in high-dimensional settings.

Combining weighting with representation learning is appealing, as it avoids over-enforcing covariate balance at the expense of predictive power. Hassanpour & Greiner (2019) reweight regression terms with inverse probability weights (IPW) estimated from the representations. Our solution differs in a few ways: First, we do not recommend the use of IPW weights since they often take on extreme values, especially in high dimensions (Li & Fu, 2017). Second, Hassanpour & Greiner (2019) do not state the theoretical benefits of using weights in the first place – that is, that weights including (but not limited to) the IPW achieve balance between treatment group distributions, given the true propensity. Finally, Hassanpour & Greiner (2019) learn the propensity score from the learned representations – this leads to an optimization procedure where one is required to alternate between learning weights and learning regressors. In contrast, we propose to train a propensity score estimator in the design stage (before any representation learning), then use it to train the regressors to estimate causal effects.

Also related to our setup is the work of Johansson et al. (2018), which tackles the slightly different problem of model generalization under design shift, for which they alternately optimize a weighting function and outcome models for prediction. Importantly, our work differs from that of Johansson et al. (2018) in that we learn a propensity score model, and use it to compute the weights, inspired by Crump et al. (2008); Li et al. (2018) – we argue that this constitutes a more principled approach to learning weights, since we benefit from the so-called *balancing property*, that is: given the true propensity, the reweighted treatment and control arms are guaranteed to be balanced, a desirable property for the estimation of causal effects. The work of Johansson et al. (2018) does not provide a similar guarantee about the weights allowing achievement of balance, and their learned weights are harder to interpret.

Empowering other causal estimators with the learned balanced representations is an appealing proposal, motivated by several considerations: (*i*) empirical evidence suggests that there is no "silver bullet" causal estimator given the diversity of causal mechanisms investigators might encounter (Alaa & Van Der Schaar, 2019); (*ii*) many classical solutions (*e.g.*, BART [Chipman et al., 2010], causal forests [Wager & Athey, 2018]) that do not have the luxury of automated representation engineering may possess appealing statistical properties (*e.g.*, built-in CATE uncertainty quantification). Repurposing the learned balanced representations and associated weights can help to free other causal inference procedures from the struggle of resolving the complexity of high-dimensional inputs, thereby boosting both performance and scalability.

3 METHODOLOGY

3.1 Basic setup

Assumptions, Identifiability of CATE Suppose we have $N = N_0 + N_1$ units, with N_0 and N_1 units in the control and treatment group, respectively. For each unit i, we have a binary treatment indicator T_i ($T_i = 1$ for treated and $T_i = 0$ for control), covariates $X_i \in \mathcal{X} \subset \mathbb{R}^p$, and two potential outcomes $\{Y_i(0), Y_i(1)\} \in \mathcal{Y} \subset \mathbb{R}$ corresponding to the control and treatment conditions, respectively. We refer to $Y_i = Y_i(T_i)$ as the factual outcome, and $Y_i^{CF} = Y_i(1 - T_i)$ as the counterfactual/unobserved outcome. The observed dataset is denoted $\mathcal{D}_F = \{X_i, T_i, Y_i\}_{i=1}^N$. The *propensity score* is $e(x) = \Pr(T_i = 1|X_i = x)$, and in practice it is estimated from $\{X_i, T_i\}_{i=1}^N$ (Rosenbaum & Rubin, 1983).

We are interested in predicting the *conditional average treatment effect* (CATE) for a given unit with covariates x: $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$. As is typical in causal inference, we make the strong ignorability assumptions: (*i*) ignorabililty, which states $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i$; and (*ii*) positivity, represented as $0 < e(x) < 1, \quad \forall x \in \mathcal{X}$. Under these assumptions, we can show that $\tau(x)$ is *identifiable* from observed data (Imbens & Wooldridge, 2009; Pearl, 2009), and $\tau(x) = \mathbb{E}[Y_i|X_i = x, T_i = 1] - \mathbb{E}[Y_i|X_i = x, T_i = 0]$.

Target populations Often causal comparisons are not for a single unit but rather on a *target distribution* of the covariates. Denote $p(x) \triangleq \Pr(X_i = x)$ as the density of the covariates, and the densities in the treated and control arms as $p(x|T = 1) \triangleq \Pr(X_i = x|T_i = 1)$ and $p(x|T = 0) \triangleq \Pr(X_i = x|T_i = 0)$, respectively. We are interested in performing inference w.r.t. some *target* population density $g(x) \overset{\triangle}{\propto} f(x)p(x)$, where $f(x)$ is a pre-specified *tilting function* (Li et al., 2018). Different choices of target densities $g(x)$ give rise to a class of average causal estimands

$$\tau_{\text{ATE},g} \triangleq \mathbb{E}_{g(x)}[\tau(x)] = \int_{\mathcal{X}} \tau(x)g(x)dx, \tag{1}$$

which includes popular estimands such as the *average treatment effect* (ATE) (with $g(x) = p(x)$) and the *average treatment effect on the treated* (ATT) (with $g(x) = p(x|T = 1)$). Table 1 details popular target populations defined by their tilting functions. Intuitively, the tilting functions in Table 1 (with the exception of IPW) place an emphasis on regions of covariate space that are *balanced* in both treatments, *i.e.* regions of overlap, where $e(x) \approx 0.5$ – this is shown in Figure 1.

Metrics for effect estimation Suppose we have a model $h(x, t)$ for the expected outcome $\mathbb{E}[Y_i|X_i = x, T_i = t]$ with covariates x under treatment t. We can estimate $\tau(x)$ and $\tau_{\text{ATE},g}$ with

$$\hat{\tau}(x) \triangleq h(x, 1) - h(x, 0), \tag{2}$$

$$\hat{\tau}_{\text{ATE},g} \triangleq \mathbb{E}_{g(x)}[\hat{\tau}(x)] \approx \frac{1}{\sum_{i=1}^N f(X_i)} \sum_{i=1}^N f(X_i)\hat{\tau}(X_i). \tag{3}$$

To evaluate the quality of estimation of the treatment effect on average, we use a metric $\epsilon_{\text{ATE},g} \triangleq |\tau_{\text{ATE},g} - \hat{\tau}_{\text{ATE},g}|$. To quantify the prediction accuracy of a CATE model $\hat{\tau}$, we use the *Precision in Estimation of Heterogeneous Effects* (PEHE) (Hill, 2011) with

Table 1: Choices of tilting function $f(x)$ and associated weight schemes $w(x,t)$ in (6). Note $\mathbb{1}(\cdot)$ is the indicator function. We set $\xi = 0.1$ as in Crump et al. (2009).

Tilting function $f(x)$	**Weight scheme** $w(x,t)$
1	Inverse Prob. Weights (IPW)
$\mathbb{1}(\xi < e(x) < 1-\xi)$	Truncated IPW (TruncIPW)
$\min(e(x), 1-e(x))$	Matching Weights (MW)
$e(x)(1-e(x))$	Overlap Weights (OW)

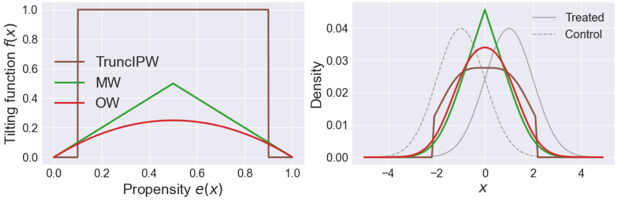


Figure 1: (Left) Tilting functions $f(x)$ used. (Right) Illustrative treatment group densities $p(x|T = t)$, and reweighted densities $g(x) \propto f(x)p(x)$ for different $f(x)$, which emphasize regions of good overlap between the treatment and control groups.

target density $g(x)$:

$$\epsilon_{\text{PEHE},g} \triangleq \mathbb{E}_{g(x)}[\{\tau(x) - \hat{\tau}(x)\}^2] \tag{4}$$

$$\approx \frac{1}{\sum_{i=1}^{N} f(X_i)} \sum_{i=1}^{N} f(X_i)\{\tau(X_i) - \hat{\tau}(X_i)\}^2. \tag{5}$$

Equation (4) is a generalization of the PEHE used in previous work (Shalit et al., 2017; Yoon et al., 2018; Louizos et al., 2017) to target populations $g(x)$. In the next section, we propose different weighting schemes and discuss how to reweight the units in the treated and control group to match the same target distribution $g(x)$.

3.2 Balancing weights

Balancing with true propensity For observational studies, typically $p(x|T = 1) \neq p(x|T = 0)$ due to selection bias resulting from confounding. To achieve balance in the statistics of covariates between the two treatment types, we would like to weight each unit in respective treatment arms towards a common target density $g(x)$. In this study we are particularly interested in a family of target distributions defined by the *balancing weights* (Li et al., 2018),

$$w(x,t) = f(x)/[t \cdot e(x) + (1-t) \cdot (1-e(x))]. \tag{6}$$

Table 1 details popular choices of balancing weights and their corresponding tilting functions. For example, when $f(x) = 1$, the weights are the inverse probability weights (IPW) $w(x,1) = 1/e(x), w(x,0) = 1/(1-e(x))$. Using balancing weights, we define the reweighted conditional distributions as $g(x|T = 1) \overset{\Delta}{\propto} w(x,1)p(x|T = 1)$ and $g(x|T = 0) \overset{\Delta}{\propto} w(x,0)p(x|T = 0)$. Due to space limitations, all proofs are relegated to the Supplementary Material (SM).

Proposition 1 (Balancing Property; Li et al., 2018)**.** *Given the true propensity $e(x)$, the reweighted treatment and control arms both equal the target distribution. In other words, $g(x|T = 1) = g(x|T = 0) = g(x)$.*

Per Proposition 1, we can balance the treatment and control distributions for estimation of treatment effects prior to any representation learning: the use of balancing weights thus complements the use of representation learning (addressed in Section 3.3) in seeking balance between treatment group distributions – Figure 1 shows the emphasis that balancing weights place on regions with good overlap between treated and control distributions.

Balancing with model propensity In practice, we do not have access to the true propensity $e(x)$, and we need to estimate it using a model $e_\eta(x)$ with parameters η (Robins et al., 1994). We plug in the estimated propensity score $e_\eta(x)$ in (6) to obtain the approximated balancing weights $w_\eta(x,t)$. With the estimated propensity score, Proposition 1 no longer holds in general, unless $e_\eta(x) = e(x)$. Given this, we may define the approximate reweighted conditional distributions $g_\eta(x|T = 1) \overset{\Delta}{\propto} w_\eta(x,1)p(x|T = 1)$ and $g_\eta(x|T = 0) \overset{\Delta}{\propto} w_\eta(x,0)p(x|T = 0)$. Though they are not equal in general, we can intuit that, the better the propensity score model, the closer we are to achieving balance between the reweighted treatment arms – this intuition is supported by Proposition 2 below.

Assumption 1. *The odds ratio between the model propensity and true propensity is bounded, namely:*

$$\exists\, \Gamma \geq 1 \;\; s.t.\; \forall x \in \mathcal{X}, \;\; \frac{1}{\Gamma} \leq \frac{e(x)(1-e_\eta(x))}{e_\eta(x)(1-e(x))} \leq \Gamma$$

Proposition 2 (Generalized Balancing)**.** *Under Assumption 1, and further assuming that all tilting functions f satisfy $f(x) > 0 \;\; \forall x \in \mathcal{X}$, we have:*

$$D_{KL}(g_\eta(x|T = 1)||g_\eta(x|T = 0)) \leq 2 \cdot \log \Gamma,$$

where D_{KL} is the KL-divergence.

Proposition 2 links the (im)balance between reweighted treatment groups to the quality of estimation of the propensity score, quantified by Γ: the closer Γ is to 1, the better the propensity score model. It can be shown immediately that this bound is tight when $\Gamma = 1$ – indeed, perfect estimation of the propensity score yields balance between reweighted treatment and control arms (Proposition 1), so the KL-divergence vanishes.

To estimate treatment effects, we learn a model $h(x, t)$. Such a model is less needed in regions of covariate space that are highly imbalanced (*i.e.*, where $e(x)$ is close to 0 or 1), as for such covariates domain experts generally have a good sense of the appropriate treatment to assign. The MW, OW and TruncIPW weights emphasize regions of covariate space where $e(x)(1 - e(x))$ is *not* close to zero, and it is this region for which causal predictions are often of most practical utility (the characteristics of $e(x)$ here imply that practitioners are less clear on what the best treatment is). Further, MW, OW and TruncIPW weightings have the advantage of de-emphasizing extreme propensity scores, concentrating on where $e_\eta(x)$ is expected to be most accurate.

MW are a weighting analogue to pair matching (Li & Greene, 2013), and OW (Li et al., 2018) target the units who are at equipoise (*i.e.*, who are likely to appear in either treatment group). In general, we recommend using OW since there is no cutoff hyperparameter (as in TruncIPW). Further, Li et al. (2018) showed that OW is the minimal asymptotic variance balancing weight for the weighted ATE (though we have yet to show an analogous result for CATE estimation). For a more exhaustive treatment of the different weighting schemes and their interpretation, we refer the readers to Li et al. (2018) and Li & Greene (2013).

Figure 2(a) illustrates the effect of the Overlap Weights in covariate space – namely, the emphasis on balanced regions of covariate space.

3.3 Representation learning with weighting

Representation learning makes use of an encoder $\Phi : \mathcal{X} \to \mathcal{R} \subset \mathbb{R}^{p'}$ to transform the original covariates to a representation space for CATE prediction using the outcome model $h(\cdot, \cdot) : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$, where $h(\Phi(x), t)$ is the predicted mean potential outcome given covariates x under treatment t. The overall model consists of the parameters for $\Phi(x)$ (typically a deep neural network) and the parameters associated with $h(\cdot, t)$, with the latter consisting of two fully-connected neural networks, one for $t = 1$ and the other for $t = 0$.

Our development is motivated by a generalization bound modified from Shalit et al. (2017), which states that under mild technical assumptions the counterfactual prediction error, and consequently, the causal effect prediction error can be upper bounded by a sum of the factual prediction error and a representation discrepancy (*i.e.*, quantified imbalance) between the treatment groups. More formally, let

$$\ell_{h,\Phi}(x, t) \triangleq \int_{\mathcal{Y}} L(y, h(\Phi(x), t)) \Pr(Y(t) = y | X = x) dy,$$

be the unit loss, where $L(y, y') : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function (*e.g.*, squared loss $(y - y')^2$). We can further define the expected factual loss w.r.t. the target density under treatment $t \in \{0, 1\}$:

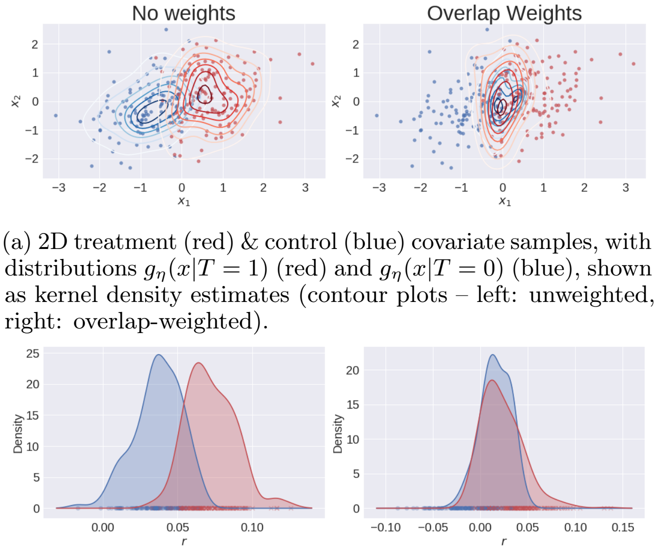
$$\epsilon_{F,g}^{T=t} \triangleq \int_{\mathcal{X}} \ell_{h,\Phi}(x, t) g(x | T = t) dx. \tag{7}$$

Remark: This differs from the original setup in Shalit et al. (2017) in that our expectation is taken w.r.t. the *target densities* $g(x|T = t)$ rather than the observational densities $p(x|T = t)$. Under the technical conditions listed in the SM, the following generalization bound holds:

$$\begin{aligned} \epsilon_{\text{PEHE},g} \leq & 2 \cdot (\epsilon_{F,g}^{T=1} + \epsilon_{F,g}^{T=0}) + C \qquad (8) \\ & + \alpha \cdot \text{IPM}_G(g_\Phi(r|T = 1), g_\Phi(r|T = 0)) \triangleq B, \end{aligned}$$

where C is a constant w.r.t. model parameters, $r = \Phi(x)$ is the representation for a unit with covariates x, and $g_\Phi(r|T = 1), g_\Phi(r|T = 0)$ are the distributions induced by the map Φ (which is invertible by assumption) from $g(x|T = 1), g(x|T = 0)$, respectively. The *integral probability metric* is defined as $\text{IPM}_G(u, v) \triangleq \sup_{m \in G} \int_{\mathcal{R}} m(r)[u(r) - v(r)]dr$ (Müller, 1997), and measures the discrepancy between two distributions u and v by identifying the maximal expected contrast w.r.t. function class G. Prominent examples of IPMs include the Wasserstein distance (Villani, 2008) and the Maximum Mean Discrepancy (MMD; Gretton et al., 2012). With stronger technical assumptions (such as G being the space of all Lipschitz-1 functions, being dense in L^2, or derived from a characteristic kernel), the IPM becomes a formal distance metric for distributions.

Standard decomposition of generalization error typically consists of two parts: the training error and model complexity, where the latter is often formally characterized by measures like Rademacher complexity or VC dimension (Shalev-Shwartz & Ben-David, 2014). The latter term usually encourages models from a simpler hypothesis space, to avoid overfitting. Compared with the bound in Shalit et al. (2017), we can reduce the bound in (8) through proper weighting in the design stage without restricting the representations themselves to be exactly balanced across treatment groups, but rather enforcing that the *reweighted* representations are balanced. Equivalently, with the proper weighting, we can improve the overall generalization bound by reducing the factual training error without sacrificing the counterfactual generalization. This reconciles the conflict that the IPM and prediction error are at odds.



(a) 2D treatment (red) & control (blue) covariate samples, with distributions $g_\eta(x|T=1)$ (red) and $g_\eta(x|T=0)$ (blue), shown as kernel density estimates (contour plots – left: unweighted, right: overlap-weighted).

(b) Learned representation distributions $g_{\Phi,\eta}(r|T=1)$ (red) and $g_{\Phi,\eta}(r|T=0)$ (blue).

Figure 2: Illustrative example with highly imbalanced treatment arms. The columns are the weight schemes used for training the outcome models and weighting the representations. (a) shows that the overlap weights (OW) focus the learning on regions of overlap in covariate space. (b) illustrates that the weighting schemes can help achieve balance in representation space under severe selection bias.

If the weights $w_\eta(x,t)$ are computed perfectly (*i.e.*, if the propensity-score model satisfies $e_\eta(x)=e(x), \forall x \in \mathcal{X}$), the IPM term in (8) vanishes – a direct consequence of Proposition 1. However, as we do not know the true propensity in practice, we approximate the bound as

$$\begin{aligned} B \approx & 2\cdot(\epsilon_{F,g_\eta}^{T=1}+\epsilon_{F,g_\eta}^{T=0})+C \\ & +\alpha\cdot \mathrm{IPM}_G(g_{\Phi,\eta}(r|T=1), g_{\Phi,\eta}(r|T=0)), \end{aligned} \tag{9}$$

where $g_{\Phi,\eta}(r|T=1), g_{\Phi,\eta}(r|T=0)$ are the distributions induced by the map Φ from the reweighted distributions $g_\eta(x|T=1)$ and $g_\eta(x|T=0)$. In practice, we use the Wasserstein distance and the MMD as the IPM in equation (9).

Proposition 3. *Under Assumption 1, assuming the representation space $\mathcal{R}$ is bounded, and assuming the tilting functions satisfy $f(x)>0\ \forall x\in\mathcal{X}$, the following bounds hold:*

$$\mathcal{W}(g_{\Phi,\eta}(r|T=1), g_{\Phi,\eta}(r|T=0)) \leq \mathrm{diam}(\mathcal{R})\sqrt{\log\Gamma};$$
$$\mathrm{MMD}_k(g_{\Phi,\eta}(r|T=1), g_{\Phi,\eta}(r|T=0)) \leq 2\sqrt{C_k\log\Gamma},$$

where $\mathcal{W}$ is the Wasserstein distance, $\mathrm{diam}(\mathcal{R}) \triangleq \sup_{r,r'\in\mathcal{R}}||r-r'||_2$, MMD_k *is the MMD with kernel k, and* $C_k \triangleq \sup_{r\in\mathcal{R}} k(r,r)$.

Proposition 3 bounds the IPM by the factor Γ which quantifies the quality of the propensity score model as in Assumption 1 – it is again easy to show that the bounds are tight when $\Gamma=1$. This result is intuitive, and shows that, the better the propensity model, the more balanced the reweighted feature distributions. Here the IPM term may be seen as a *correction* to the weights, addressing errors manifested by imperfections in the estimated propensity score. However, since much of the balance is achieved by the weights, it is less likely that the weighted IPM term will remove predictive features. Figure 2(b) illustrates how weighting can achieve balance in representation space. The weighted density plots show that the learned weighted representations become more balanced compared with the unweighted one. Weighting achieves a similar effect as the IPM term in balancing the representations, but it does not enforce that the (unweighted) empirical distributions of the representations need to be matched across treatments.

3.4 Implementation

We train a propensity score model $e_\eta(x)$ by minimizing $\mathcal{L}_{prop}(\eta)$ w.r.t. η, where:

$$\begin{aligned} \mathcal{L}_{prop}(\eta) = -\sum_{i=1}^{N}\{ & \frac{T_i}{N_1}\cdot\log[\sigma(s_\eta(X_i))] \\ & + \frac{1-T_i}{N_0}\cdot\log[1-\sigma(s_\eta(X_i))]\}. \end{aligned} \tag{10}$$

$\sigma(z) \triangleq 1/[1+\exp(-z)]$, and $s_\eta(x)$ is a fully-connected neural network with $e_\eta(x)\triangleq\sigma(s_\eta(x))$. Once $e_\eta(x)$ is trained, we learn the parameters of the encoder $\Phi(x)$ and the outcome models $h(\Phi(x),1)$ and $h(\Phi(x),0)$. We can show that the approximation in (9) leads to the following finite-sample objective, which we minimize w.r.t. h,Φ:

$$\begin{aligned} \mathcal{L}(h,\Phi) \triangleq & \mathcal{L}_F(h,\Phi) \\ & +\alpha\cdot\mathrm{IPM}_G\left(\hat{g}_{\Phi,\eta}(r|T=1), \hat{g}_{\Phi,\eta}(r|T=0)\right) \end{aligned} \tag{11}$$

where $\mathcal{L}_F(h,\Phi)$ is a Monte Carlo approximation of $\epsilon_{F,g_\eta}^{T=1}+\epsilon_{F,g_\eta}^{T=0}$ and $\hat{g}_{\Phi,\eta}(r|T=t)$ is the empirical approximation of $g_{\Phi,\eta}(r|T=t)$ $(t\in\{0,1\})$, defined as:

$$\mathcal{L}_F(h,\Phi) \triangleq \frac{1}{N}\sum_{i=1}^{N} w_\eta(X_i,T_i)\ \left(Y_i - h(\Phi(X_i),T_i)\right)^2,$$
$$\hat{g}_{\Phi,\eta}(r|T=t) \triangleq \sum_{i:T_i=t}\frac{w_\eta(X_i,t)}{\sum_{j:T_j=t} w_\eta(X_j,t)}\delta(r-\Phi(X_i)).$$

$\delta(r-z)$ is a point mass centered at z and $\Phi(\cdot), h(\cdot,1), h(\cdot,0)$ are fully-connected neural networks. More details on how to obtain the finite-sample approximation in (11) and how to compute the weighted IPM term in practice are provided in the SM.

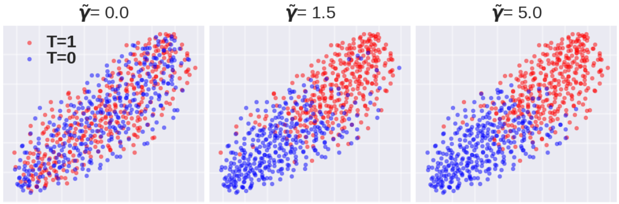


Figure 3: Higher $\tilde{\gamma}$ leads to larger imbalance between treatment groups.

4 EXPERIMENTS

4.1 Synthetic data

Data generating process We wish to understand the effect of distribution imbalance (the extent to which the treatment and control distributions differ) on the performance of our methods for CATE estimation. Specifically, we construct datasets for which we vary the distribution imbalance and the amount of confounding. Consider the following data-generating mechanism:

- Fix $\sigma_X, \sigma_Y, \rho, \theta \in \mathbb{R}$. Set $\beta_0, \beta_\tau, \gamma \in \mathbb{R}^p$ to be p^*-sparse vectors (*i.e.*, $||\beta_0||_0 = ||\beta_\tau||_0 = ||\gamma||_0 = p^*$), and further set $supp(\beta_0) = supp(\beta_\tau) \triangleq \mathcal{B}$, $\mathcal{G} \triangleq supp(\gamma)$, and the *confounding parameter* $\Omega \triangleq |\mathcal{B} \cap \mathcal{G}|$.
- For simplicity, set $\gamma = \tilde{\gamma} \cdot \mathbb{1}_{\mathcal{G}}$, where $\mathbb{1}_{\mathcal{G}} \in \{0,1\}^p$ is a binary vector with ones at elements of $\mathcal{G}$, and $\tilde{\gamma} \geq 0$ is the *imbalance parameter.* Note $||\gamma||_2 = \tilde{\gamma} \cdot p^*$.
- Draw $X_i, T_i, Y_i(1), Y_i(0)$ as follows:

$$X_i \sim \mathcal{MVN}(0, \sigma_X^2[(1-\rho)I_p + \rho 1_p 1_p^T]), \tag{12}$$

$$T_i | X_i \sim \text{Bernoulli}(\sigma(X_i^T \gamma)), \tag{13}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_Y^2), \quad Y_i(0) = X_i^T \beta_0 + \epsilon_i, \tag{14}$$

$$Y_i(1) = X_i^T \beta_0 + X_i^T \beta_\tau + \theta + \epsilon_i. \tag{15}$$

This data-generating process satisfies the assumptions of ignorability and overlap. We construct multiple such datasets by varying the distribution imbalance and amount of confounding, as follows:

Distribution imbalance: We increase the distribution imbalance by increasing $\tilde{\gamma}$ in the range [0,5]. Figure 3 shows a t-SNE plot of the 50-dimensional covariates, for different values of $\tilde{\gamma}$.
Level of confounding: We increase the level of confounding by increasing Ω, *i.e.*, the extent to which the same covariates are predictive of treatment and potential outcomes. We vary Ω to be equal to p^* ("high confounding"), $\frac{p^*}{2}$ ("moderate confounding"), and 0 ("low confounding"). In the "low confounding" setting, $(\Omega = 0)$, there is still some confounding by way of the correlation ρ between the covariates.

In total we generate 33 datasets (3 values of Ω × 11 values of $||\gamma||_2$). For more details on the data-generating process, see the SM.

Results We compare the weighted-model performance across the 33 datasets generated as discussed above, and we compare against using no weights. For a fair comparison, we fix all hyperparameters with the exception of the IPM regularization strength α (for details on hyperparameters, see the SM). Figure 4 shows the performance of each method for all datasets. For a given dataset and weight scheme, we select the α that minimizes $\epsilon_{\text{PEHE},p}$. We picked the α minimizing the true $\epsilon_{\text{PEHE},p}$ (which includes knowledge of counterfactual outcomes) to avoid introducing any noise in the comparisons via a proxy such as a 1-nearest-neighbor imputation (1NNI) of missing potential outcomes. For the remaining experiments on real data (Sections 4.2 and 4.3), we use 1NNI, which makes no use of counterfactual outcomes, for model selection in order to compare with existing work.

From Figure 4, one can immediately see the benefit of using a weighted objective (weighted regression + weighted IPM) over its unweighted counterpart. More specifically, the MW, OW, and TruncIPW weights do well in comparison with the other weight schemes, especially in settings of high imbalance (*i.e.*, high values of $\tilde{\gamma}$). On the other hand, IPW is numerically unstable (Li & Fu, 2017) and yields only marginally better results than its unweighted counterpart, so we do not recommend its use as a weighting scheme. This provides empirical evidence for the fact that weighted CATE models, though trained to perform well on a *target* population $g(x)$ (namely, for the non-IPW weights, regions of good overlap), vastly improve CATE estimation on the *observed* population $p(x)$. We also compared the performance of our models on the target populations (*i.e.*, as measured by $\sqrt{\epsilon_{\text{PEHE},g}}$), and found that the weight schemes perform well on the respective populations they target. For details about the performance on target populations, see the SM.

Benefit of weighted IPM regularization We seek to understand the benefit of the weighted IPM term in our objective formulation. We make this comparison by taking the difference between the best $\sqrt{\epsilon_{\text{PEHE},p}}$ across all values of α (*i.e.*, the plots shown in Figure 4), and the $\sqrt{\epsilon_{\text{PEHE},p}}$ for $\alpha = 0$ (*i.e.*, without the IPM term). The benefit of using the weighted IPM term vs. an unweighted IPM term is immediately visible from Figure 5, especially in cases of moderate to high imbalance. A likely explanation for this is that the IPM term is attempting to match the *weighted* distributions in representation space rather than the unweighted ones, which means it is less prone to "erasing" informa-

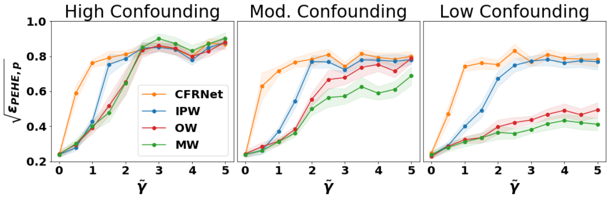


Figure 4: $\sqrt{\epsilon_{\text{PEHE},p}}$ *vs.* dataset imbalance parameter $\tilde{\gamma}$ for different confounding settings (high, moderate, low). The colored bands are standard errors over 20 realizations. TruncIPW was omitted to avoid clutter, because it was similar to OW and MW. CFRNet is from Shalit et al. (2017), which uses propensity-independent weights $(1 - T_i)/N_0 + T_i/N_1$. Lower is better.

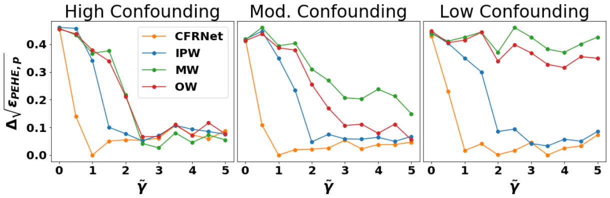


Figure 5: Improvement in $\sqrt{\epsilon_{\text{PEHE},p}}$ between the model with and without the IPM term, denoted $\Delta\sqrt{\epsilon_{\text{PEHE},p}}$, *vs.* imbalance parameter $\tilde{\gamma}$. The colored bands are standard errors over 20 realizations. Higher is better.

tion from confounders. Finally, the IPM term when $\tilde{\gamma}$=0 should theoretically be 0, since the treatment and control covariate distributions are the same. However, there still may be imbalance between the drawn *samples*, but the Wasserstein distance would vanish with increasing sample size (Sriperumbudur et al., 2009).

4.2 Infant Health and Development Program

The Infant Health and Development Program (IHDP) dataset (Hill, 2011) is semi-simulated (real covariates with simulated outcomes) measuring the effect of home visits from a trained provider on children's cognitive test scores. This dataset has a more realistic covariate distribution than the above synthetic data, but we cannot control the degree of imbalance. We report out-of-sample results on the IHDP1000 dataset from Shalit et al. (2017) in Table 2, showing competitive performance both in terms of CATE prediction ($\sqrt{\epsilon_{\text{PEHE},p}}$) and ATE prediction ($\epsilon_{\text{ATE},p}$). For details on model selection and training, see the SM. For results with an IPW-based solution, we point the readers. We note from Table 2 that our method (BWCFR) outperforms many classical causal inference methods, such as the causal forest. This is in part because our method benefits from automated representation learning (via the map $\Phi(x)$) upstream of the outcome models $h(\Phi(x), 1), h(\Phi(x), 0)$. In the next section, we explore what happens when we leverage the learned features and weights to benefit classical methods with appealing statistical features (*e.g.*, CATE uncertainty estimates).

Table 2: Results on IHDP1000 test set. The top block consists of baselines from recent work. The bottom block consists of our proposed methods. Lower is better.

Model	$\sqrt{\epsilon_{\text{PEHE},p}}$	$\epsilon_{\text{ATE},p}$
OLS-1 (Johansson et al., 2016)	5.8 ± .3	.94 ± .06
OLS-2 (Johansson et al., 2016)	2.5 ± .1	.31 ± .02
BLR (Johansson et al., 2016)	5.8 ± .3	.93 ± .05
k-NN (Crump et al., 2008)	4.1 ± .2	.79 ± .05
BART (Chipman et al., 2010)	2.3 ± .1	.34 ± .02
Random Forest (Breiman, 2001)	6.6 ± .3	.96 ± .06
Causal Forest (Wager & Athey, 2018)	3.8 ± .2	.40 ± .03
BNN (Johansson et al., 2016)	2.1 ± .1	.42 ± .03
TARNET (Shalit et al., 2017)	.88 ± .02	.26 ± .01
CFRNet-Wass (Shalit et al., 2017)	.76 ± .02	.27 ± .01
CFR-ISW[1] (Hassanpour & Greiner, 2019)	.70 ± .1	.19 ± .03
RCFR[1] (Johansson et al., 2018)	.67 ± .05	-
CMGP (Alaa & van der Schaar, 2017)	.74 ± .11	-
DKLITE (Zhang et al., 2020)	.65 ± .03	-
BWCFR-MW (Ours)	.66 ± .02	**.18 ± .01**
BWCFR-OW (Ours)	.65 ± .02	**.18 ± .01**
BWCFR-TruncIPW (Ours)	**.63 ± .01**	.19 ± .01

4.3 Improving causal forest with the balanced representations learned

We further examine the extent to which the learned balanced representations of our proposal can facilitate other causal learning algorithms. In particular, we quantitatively assess the potential gains for causal forests (CF; Wager & Athey, 2018), and report our findings in Table 3. In the first experiment we evaluate the performance difference with and without the learned balanced representation and weights on the IHDP100 dataset (Hill, 2011; Shalit et al., 2017) w.r.t. both the individual and population level metrics (*i.e.*, $\sqrt{\epsilon_{\text{PEHE},p}}$, $\epsilon_{\text{ATE},p}$). Also, we examine the proportion of datasets (out of 77) in the ACIC2016 benchmark (Dorie et al., 2019; Alaa & Van Der Schaar, 2019) for which the learned representations and weights improve (*i.e.*, % ↓ $\sqrt{\epsilon_{\text{PEHE},p}}$ and % ↓ $\epsilon_{\text{ATE},p}$, respectively), compared to a "vanilla" CF model. For both datasets we observe substantial gains in both CATE and ATE estimation (relative to the vanilla CF trained on the original covariates), which demonstrates the effectiveness of using pre-balanced representations and weights, in this case learned by our model, to augment other causal models.

To note, our methods on IHDP (bottom-right of Table 2) still outperform that of the causal forest (in terms of

[1] These methods reported their results on the IHDP100 dataset (equivalent to IHDP1000, but with 100 repetitions instead of 1000). For a comparison on IHDP100, see SM.

$\sqrt{\epsilon_{\text{PEHE},p}}$), even when the causal forest has access to the same balanced representations and weights learned (bottom-left of Table 3). One potential explanation is that tree-based learner lacks the sophistication to decode the rich representation encoded by a more flexible neural net. It would be interesting to explore an end-to-end optimization strategy that combines our proposed representation engineering and the causal forest model; we leave this for future work. For details about hyperparameter tuning and additional analyses, see the SM.

Table 3: Causal forest (CF) results. The top block is a vanilla CF model. The bottom block consists of causal forest models using the learned representations and weights. The bottom block rows are the weights used in the objective (11) and as the per-sample weights to train the CF. The left block shows $\sqrt{\epsilon_{\text{PEHE},p}}$ and $\epsilon_{\text{ATE},p}$ results on the IHDP dataset (lower is better), and the right block shows % ↓ $\sqrt{\epsilon_{\text{PEHE},p}}$ and % ↓ $\epsilon_{\text{ATE},p}$ results on the ACIC2016 dataset (higher is better).

	IHDP100		ACIC2016	
	$\sqrt{\epsilon_{\text{PEHE},p}}$	$\epsilon_{\text{ATE},p}$	↓ $\sqrt{\epsilon_{\text{PEHE},p}}$	↓ $\epsilon_{\text{ATE},p}$
CF	3.54 ± .58	.47 ± .06	-	-
CF + MW	1.51 ± .31	.20 ± .03	92.2%	89.6%
CF + OW	1.59 ± .31	.19 ± .03	93.5%	85.7%
CF + TruncIPW	1.55 ± .35	.22 ± .03	87.0%	71.4%

5 CONCLUSIONS

We show that the use of balancing weights complements representation learning in mitigating covariate imbalance. Our claims are supported with theoretical results and evaluations on synthetic datasets and realistic test benchmarks, reporting better or competitive performance throughout. Further, we demonstrated how our learned balanced features can augment other causal inference procedures, towards the goal of building more reliable and accurate hybrid solutions. Directions for future work include *learning* the tilting function f rather than selecting it in order to determine an "optimal" target population, as well as exploring more advanced weighting approaches (Hainmueller, 2012; Zubizarreta, 2015; Ozery-Flato et al., 2018). Additionally, the current form of Assumption 1 is somewhat restrictive (it posits the existence of Γ such that odds ratio is bounded *for all* x). Softening it (*e.g.*, by bounding the odds ratio, integrated over the population of interest) would provide a more realistic form of Assumption 1, which would make Proposition 2 more useful. Finally, an end-to-end approach to training the propensity and regression models (as in Hassanpour & Greiner, 2020) seems like a promising alternative to our current two-step training procedure.

Acknowledgements

The research reported here was supported in part by the DOE, NSF and ONR.

Bibliography

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 129–138, Stockholmsmässan, Stockholm Sweden, 7 2018. PMLR.

Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 191–201, Long Beach, California, USA, 6 2019. PMLR.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes, 2017.

Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–6, 2015.

Leo Breiman. Random forests. *Machine learning*, 45 (1):5–32, 2001.

David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. Evaluating online ad campaigns in a pipeline: Causal models at scale. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pp. 7–16, New York, NY, USA, 2010. Association for Computing Machinery.

Junya Chen, Jianfeng Feng, and Wenlian Lu. A wiener causality defined by divergence. *Neural Processing Letters*, pp. 1–22, 2020.

Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 01 2009.

Ralph B. D'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 1998.

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.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34 (1):43–68, 2019.

Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34 (1):61–75, 2013.

Bryan S Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79(3):1053–1079, 2012.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, pp. 25–46, 2012.

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5880–5887, 2019.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxBJT4YvB.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960, 1986.

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

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 3 2009.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. 2018.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. 2020.

Kipp W Johnson, Benjamin S Glicksberg, Rachel Hodos, Khader Shameer, and Joel T Dudley. Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula. In *PSB*, pp. 180–191. World Scientific, 2018.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training, 2018.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. volume 89 of *Proceedings of Machine Learning Research*, pp. 2281–2290. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/kallus19a.html.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113 (521):390–400, 2018.

Liang Li and Tom Greene. A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*, 9(2):215 – 234, 2013.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 929–939. Curran Associates, Inc., 2017.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models, 2017.

Jared K. Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, August 2004.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

Michal Ozery-Flato, Pierre Thodoroff, Matan Ninio, Michal Rosen-Zvi, and Tal El-Hay. Adversarial balancing for causal inference. 2018.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

PR Rosenbaum and DB 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*, 100(469):322–331, 2005.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

U. Shalit, F.D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. 2017.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification, 2009.

E.A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 2010.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. 2020.

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.