Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit [*1] **Fredrik D. Johansson** [*2] **David Sontag** [2 3]

Abstract

There is intense interest in applying machine learning to problems of causal inference in fields such as healthcare, economics and education. In particular, individual-level causal inference has important applications such as precision medicine. We give a new theoretical analysis and family of algorithms for predicting individual treatment effect (ITE) from observational data, under the assumption known as strong ignorability. The algorithms learn a "balanced" representation such that the induced treated and control distributions look similar, and we give a novel and intuitive generalization-error bound showing the expected ITE estimation error of a representation is bounded by a sum of the standard generalization-error of that representation and the distance between the treated and control distributions induced by the representation. We use Integral Probability Metrics to measure distances between distributions, deriving explicit bounds for the Wasserstein and Maximum Mean Discrepancy (MMD) distances. Experiments on real and simulated data show the new algorithms match or outperform the state-of-the-art.

1. Introduction

Making predictions about causal effects of actions is a central problem in many domains. For example, a doctor deciding which medication will cause better outcomes for a patient; a government deciding who would benefit most from subsidized job training; or a teacher deciding which study program would most benefit a specific student. In this paper we focus on the problem of making these predictions based on *observational data*. Observational data is data which contains past actions, their outcomes, and possibly more context, but without direct access to the mechanism which gave rise to the action. For example we might have access to records of patients (context), their medications (actions), and outcomes, but we do not have complete knowledge of why a specific action was applied to a patient.

The hallmark of learning from observational data is that the actions observed in the data depend on variables which might also affect the outcome, resulting in *confounding*: For example, richer patients might better afford certain medications, and job training might only be given to those motivated enough to seek it. The challenge is how to untangle these confounding factors and make valid predictions. Specifically, we work under the common simplifying assumption of "no-hidden confounding", assuming that all the factors determining which actions were taken are observed. In the examples above, it would mean that we have measured a patient's wealth or an employee's motivation.

As a learning problem, estimating causal effects from observational data is different from classic learning in that in our training data we never see the individual-level effect. For each unit, we only see their response to one of the possible actions - the one they had actually received. This is close to what is known in the machine learning literature as "learning from logged bandit feedback" (Strehl et al., 2010; Swaminathan & Joachims, 2015), with the distinction that we do not have access to the model generating the action.

Our work differs from much work in causal inference in that we focus on the individual-level causal effect ("c-specific treatment effects" Shpitser & Pearl (2006); Pearl (2015)), rather than the average or population level. Our main contribution is to give what is, to the best of our knowledge, the first generalization-error[1] bound for estimating individual-level causal effect, where each individual is identified by its features x. The bound leads naturally to a new family of representation-learning based algorithms (Bengio et al., 2013), which we show to match or outperform state-of-the-art methods on several causal effect inference tasks.

[*] Equal contribution [1] CIMS, New York University, New York, NY 10003 [2] IMES, MIT, Cambridge, MA 02142 [3] CSAIL, MIT, Cambridge, MA 02139. Correspondence to: Uri Shalit <shalit@cs.nyu.edu>, Fredrik D. Johansson <fredrikj@mit.edu>, David Sontag <dsontag@csail.mit.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] Our use of the term generalization is different from its use in the study of *transportability*, where the goal is to generalize causal conclusion across distributions (Bareinboim & Pearl, 2016).

We frame our results using the Neyman-Rubin potential outcomes framework (Rubin, 2011), as follows. We assume that for a unit with features $x \in \mathcal{X}$, and an action (also known as treatment or intervention) $t \in \{0, 1\}$, there are two potential outcomes: Y_0 and Y_1. For each unit we only observe one of the potential outcomes, according to treatment assignment: if $t = 0$ we observe $y = Y_0$, if $t = 1$, we observe $y = Y_1$; this is known as the *consistency* assumption. For example, x can denote the set of lab tests and demographic factors of a diabetic patient, $t = 0$ denote the standard medication for controlling blood sugar, $t = 1$ denotes a new medication, and Y_0 and Y_1 indicate the patient's blood sugar level if they were to be given medications $t = 0$ and $t = 1$, respectively.

We will denote $m_1(x) = \mathbb{E}[Y_1|x]$, $m_0(x) = \mathbb{E}[Y_0|x]$. We are interested in learning the function $\tau(x) := \mathbb{E}[Y_1 - Y_0|x] = m_1(x) - m_0(x)$. $\tau(x)$ is the expected *treatment effect* of $t = 1$ relative to $t = 0$ on a unit with characteristics x, or the Individual Treatment Effect (ITE)[2]. Our goal is to find an estimate $\hat{\tau}$ of τ such that some loss function, e.g. $\mathbb{E}\left[(\hat{\tau} - \tau)^2\right]$, is small. For example, for a patient with features x, we attempt to predict which of two treatments will have a better outcome. The fundamental problem of causal inference is that for any x in our data we only observe Y_1 or Y_0, but never both.

As mentioned above, we make an important "no-hidden confounders" assumption, in order to make the conditional causal effect identifiable. We formalize this assumption by using the standard *strong ignorability* condition: $(Y_1, Y_0) \perp\!\!\!\perp t|x$, and $0 < p(t = 1|x) < 1$ for all x. Strong ignorability is a sufficient condition for the ITE function $\tau(x)$ to be identifiable (Imbens & Wooldridge, 2009; Pearl, 2015; Rolling, 2014): see proof in the supplement. The validity of strong ignorability cannot be assessed from data, and must be determined by domain knowledge and understanding of the causal relationships between the variables.

One approach to the problem of estimating the function $\tau(x)$ is by learning the two functions $m_0(x)$ and $m_1(x)$ using samples from $p(Y_t|x, t)$. This is similar to a standard machine learning problem of learning from finite samples. However, there is an additional source of variance at work here: For example, if mostly rich patients received treatment $t = 1$, and mostly poor patients received treatment $t = 0$, we might have an unreliable estimation of $m_1(x)$ for poor patients. In this paper we upper bound this additional source of variance using an Integral Probability Metric (IPM) measure of distance between two distributions $p(x|t = 0)$, and $p(x|t = 1)$, also known as the *control* and *treated* distributions. In practice we use two specific IPMs: the Maximum Mean Discrepancy (Gretton et al., 2012), and the Wasserstein distance (Villani, 2008; Cuturi & Doucet, 2014). We show that the expected error in learning the individual treatment effect function $\tau(x)$ is upper bounded by the error of learning Y_1 and Y_0, plus the IPM term. In the randomized controlled trial setting, where $t \perp\!\!\!\perp x$, the IPM term is 0, and our bound naturally reduces to a standard learning problem of learning two functions.

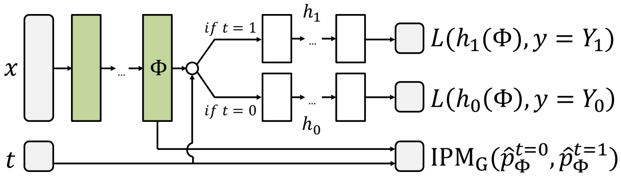

Figure 1. Neural network architecture for ITE estimation. L is a loss function, $\mathrm{IPM_G}$ is an integral probability metric. Note that only one of h_0 and h_1 is updated for each sample during training.

The bound we derive points the way to a family of algorithms based on the idea of representation learning (Bengio et al., 2013): Jointly learn hypotheses for both treated and control on top of a representation which minimizes a weighted sum of the factual loss (the standard supervised machine learning objective), and the IPM distance between the control and treated distributions induced by the representation. This can be viewed as learning the functions m_0 and m_1 under a constraint that encourages better generalization across the treated and control populations. In the Experiments section we apply algorithms based on neural nets as representations and hypotheses, along with MMD or Wasserstein distributional distances over the representation layer; see Figure 1 for the basic architecture.

In his foundational text on causality, Pearl (2009) writes: "Whereas in traditional learning tasks we attempt to generalize from one set of instances to another, the causal modeling task is to generalize from behavior under one set of conditions to [...] another set. *Causal models should therefore be chosen by a criterion that challenges their stability against changing conditions*" [emphasis ours]. We believe our work points the way to one such stability criterion, for causal inference in the strongly ignorable case.

2. Related work

Much recent work in machine learning for causal inference focuses on *causal discovery*, with the goal of discovering the underlying causal graph or causal direction from data (Hoyer et al., 2009; Maathuis et al., 2010; Triantafillou & Tsamardinos, 2015; Mooij et al., 2016). We focus on the case when the causal setup is simple and known to be of the form $(Y_1, Y_0) \perp\!\!\!\perp x|t$, with no hidden confounders. Under the causal model we assume, the most common goal of causal effect inference as used in the ap-

[2] Also known as Conditional Average Treatment Effect, CATE.

plied sciences is to obtain the average treatment effect: $ATE = \mathbb{E}_{x \sim p(x)}[\tau(x)]$. We will briefly discuss how some standard statistical causal effect inference methods relate to our proposed method. Note that most of these approaches assume some form of ignorability.

One of the most widely used approaches to estimating ATE is covariate (or back-door) adjustment, also known as the G-computation formula (Robins, 1986; Pearl, 2009). In their basic version, covariate adjustment methods aim to estimate the functions $m_1(x)$, $m_0(x)$, and are therefore natural candidates for estimating ITE as well as ATE, using the estimates of $m_t(x)$. Most previous work on this subject focused on asymptotic consistency (Belloni et al., 2014; Athey et al., 2016; Chernozhukov et al., 2016), and so far there has not been much work on the generalization-error of such a procedure. One view of our results is that we point out a previously unaccounted for source of variance when using covariate adjustment to estimate ITE. We suggest a new type of regularization, by learning representations with reduced IPM distance between treated and control, enabling a new type of bias-variance trade-off.

Another widely used family of statistical methods used in causal effect inference are weighting methods. Methods such as inverse propensity score weighting (Austin, 2011) re-weight the units in the observational data so as to make the treated and control populations more comparable, and have been used for estimating conditional effects as well (Cole et al., 2003). The major challenge, especially in high-dimensional cases, is controlling the variance of the estimates (Swaminathan & Joachims, 2015). Doubly robust methods go further and combine propensity score reweighting and covariate adjustment in clever ways to reduce model bias (Funk et al., 2011).

Adapting machine learning methods for causal effect inference, and in particular for individual level treatment effect, has gained much interest recently. For example Wager & Athey (2015); Athey & Imbens (2016) discuss how tree-based methods can be adapted to obtain a consistent estimator with semi-parametric asymptotic convergence rate. Recent work has also looked into how machine learning methods can help detect heterogeneous treatment effects when some data from randomized experiments is available (Taddy et al., 2016; Peysakhovich & Lada, 2016). Neural nets have also been used for this purpose, exemplified in early work by Beck et al. (2000), and more recently by Hartford et al. (2016)'s work on deep instrumental variables. Our work differs from all the above by focusing on the generalization-error aspects of estimating individual treatment effect, as opposed to asymptotic consistency, and by focusing solely on the observational study case, with no randomized components or instrumental variables.

Our work has strong connections with work on domain adaptation. In particular, estimating ITE requires prediction of outcomes over a different distribution from the observed one. Our ITE error upper bound has similarities with generalization bounds in domain adaptation given by Ben-David et al. (2007); Mansour et al. (2009); Ben-David et al. (2010); Cortes & Mohri (2014). These bounds employ distribution distance metrics such as the A-distance or the discrepancy metric, which are related to the IPM distance we use. Our algorithm is similar to a recent algorithm for domain adaptation by Ganin et al. (2016), and in principle other domain adaptation methods (e.g. Daumé III (2007); Pan et al. (2011); Sun et al. (2016)) could be adapted for use in ITE estimation as presented here.

Finally, our paper builds on Johansson et al. (2016), where we showed a connection between covariate shift and the task of estimating counterfactuals. We proposed learning a representation of the data that makes the treated and control distributions more similar, fitting a linear ridge-regression model on top of it. We bounded the relative error of fitting a ridge-regression using the distribution with reverse treatment assignment versus fitting a ridge-regression using the factual distribution. Unfortunately, the relative error bound is not at all informative regarding the absolute quality of the representation. In this paper we focus on a related but more substantive task: estimating the individual treatment effect, building on the counterfactual error term. We provide an informative bound on the absolute quality of the representation. We also derive a much more flexible family of algorithms, including non-linear hypotheses and much more powerful distribution metrics in the form of IPMs such as the Wasserstein and MMD distances. Finally, we conduct significantly more thorough experiments including a real-world dataset and out-of-sample performance, and show our methods outperform previously proposed ones.

3. Estimating ITE: Error bounds

In this section we prove a bound on the expected error in estimating the individual treatment effect for a given representation, and a hypothesis defined over that representation. The bound is expressed in terms of (1) the expected loss of the model when learning the observed outcomes y as a function of x and t, denoted ϵ_F, F standing for "Factual"; (2) an Integral Probability Metric (IPM) distance between the distribution of treated and control units. The term ϵ_F is the classic machine learning generalization-error, and in turn can be upper bounded using the empirical error and model complexity terms, applying standard machine learning theory (Shalev-Shwartz & Ben-David, 2014).

3.1. Problem setup

We will employ the following assumptions and notations. The most important notations are in the Notation box in the supplement. The space of covariates is a bounded subset $\mathcal{X} \subset \mathbb{R}^d$. The outcome space is $\mathcal{Y} \subset \mathbb{R}$. Treatment t is a binary variable. We assume there exists a joint distribution $p(x, t, Y_0, Y_1)$, such that $(Y_1, Y_0) \perp\!\!\!\perp t|x$ and $0 < p(t = 1|x) < 1$ for all $x \in \mathcal{X}$ (strong ignorability). The treated and control distributions are the distribution of the features x conditioned on treatment: $p^{t=1}(x) := p(x|t = 1)$, and $p^{t=0}(x) := p(x|t = 0)$, respectively.

Throughout this paper we will discuss *representation functions* of the form $\Phi : \mathcal{X} \to \mathcal{R}$, where $\mathcal{R}$ is the representation space. We make the following assumption about Φ:

Assumption 1. *The representation Φ is a twice-differentiable, one-to-one function. Without loss of generality we will assume that $\mathcal{R}$ is the image of $\mathcal{X}$ under Φ. We then have $\Psi : \mathcal{R} \to \mathcal{X}$ as the inverse of Φ, such that $\Psi(\Phi(x)) = x$ for all $x \in \mathcal{X}$.*

The representation Φ pushes forward the treated and control distributions into the new space $\mathcal{R}$; we denote the induced distribution by p_Φ.

Definition 1. *Define $p_\Phi^{t=1}(r) := p_\Phi(r|t = 1)$, $p_\Phi^{t=0}(r) := p_\Phi(r|t = 0)$, to be the treated and control distributions induced over $\mathcal{R}$. For a one-to-one Φ, the distributions $p_\Phi^{t=1}(r)$ and $p_\Phi^{t=0}(r)$ can be obtained by the standard change of variables formula, using the determinant of the Jacobian of $\Psi(r)$.*

Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a representation function, and $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ be an hypothesis defined over the representation space $\mathcal{R}$. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. We define two complimentary loss functions: one is the standard machine learning loss, which we will call the factual loss ϵ_F, as it relates to observable quantities. The other is the expected loss with respect to the distribution where the treatment assignment is flipped, which we call the counterfactual loss, ϵ_{CF}.

Definition 2. *The expected loss for the unit and treatment pair (x, t) is: $\ell_{h,\Phi}(x, t) = \int_\mathcal{Y} L(Y_t, h(\Phi(x), t))p(Y_t|x)dY_t$. The expected factual and counterfactual losses of h and Φ are:*

$$\epsilon_F(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(x, t)\, p(x, t)\, dxdt,$$
$$\epsilon_{CF}(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(x, t)\, p(x, 1 - t)\, dxdt.$$

If x denotes patients' features, t a treatment, and Y_t a potential outcome such as mortality, we think of ϵ_F as measuring how well do h and Φ predict mortality for the patients and doctors' actions sampled from the same distribution as our data sample. ϵ_{CF} measures how well our prediction with h and Φ would do in a "topsy-turvy" world where the patients are the same but the doctors are inclined to prescribe exactly the opposite treatment than the one the real-world doctors would prescribe.

Definition 3. *The expected factual* treated *and* control *losses are:*

$$\epsilon_F^{t=1}(h, \Phi) = \int_\mathcal{X} \ell_{h,\Phi}(x, 1)\, p^{t=1}(x)\, dx,$$
$$\epsilon_F^{t=0}(h, \Phi) = \int_\mathcal{X} \ell_{h,\Phi}(x, 0)\, p^{t=0}(x)\, dx.$$

For $u := p(t = 1)$, it is immediate to show that $\epsilon_F(h, \Phi) = u\epsilon_F^{t=1}(h, \Phi) + (1 - u)\epsilon_F^{t=0}(h, \Phi)$.

Definition 4. *The treatment effect (ITE) for unit x is:*

$$\tau(x) := \mathbb{E}\left[Y_1 - Y_0|x\right].$$

Let $f : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ by an hypothesis. For example, we could have that $f(x, t) = h(\Phi(x), t)$.

Definition 5. *The treatment effect estimate of the hypothesis f for unit x is:*

$$\hat{\tau}_f(x) = f(x, 1) - f(x, 0).$$

Definition 6. *The expected Precision in Estimation of Heterogeneous Effect (PEHE, Hill (2011)) loss of f is:*

$$\epsilon_{PEHE}(f) = \int_\mathcal{X} \left(\hat{\tau}_f(x) - \tau(x)\right)^2\, p(x)\, dx, \qquad (1)$$

When $f(x, t) = h(\Phi(x), t)$, we will also use the notation $\epsilon_{PEHE}(h, \Phi) = \epsilon_{PEHE}(f)$.

Our proof relies on the notion of an *Integral Probability Metric* (IPM), which is a class of metrics between probability distributions (Sriperumbudur et al., 2012; Müller, 1997). For two probability density functions p, q defined over $\mathcal{S} \subseteq \mathbb{R}^d$, and for a function family G of functions $g : \mathcal{S} \to \mathbb{R}$, we have that

$$\mathrm{IPM}_\mathrm{G}(p, q) := \sup_{g\in \mathrm{G}} \left| \int_\mathcal{S} g(s)(p(s) - q(s))\, ds \right|.$$

Integral probability metrics are always symmetric and obey the triangle inequality, and trivially satisfy $\mathrm{IPM}_\mathrm{G}(p, p) = 0$. For rich enough function families G, we also have that $\mathrm{IPM}_\mathrm{G}(p, q) = 0 \implies p = q$, and then IPM_G is a true metric. Examples of function families G for which IPM_G is a true metric are the family of bounded continuous functions, the family of 1-Lipschitz functions (Sriperumbudur et al., 2012), and the unit-ball of functions in a universal reproducing kernel Hilbert space (Gretton et al., 2012).

3.2. Bounds

We first state a Lemma bounding the counterfactual loss, a key step in obtaining the bound on the error in estimating individual treatment effect. We then give the main Theorem. The proofs and details are in the supplement.

Let $u := p(t = 1)$ be the marginal probability of treatment. By the strong ignorability assumption, $0 < u < 1$.

Lemma 1. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function, with inverse Ψ. Let $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ be an hypothesis. Let* G *be a family of functions $g : \mathcal{R} \to \mathcal{Y}$. Assume there exists a constant $B_\Phi > 0$, such that for fixed $t \in \{0, 1\}$, the per-unit expected loss functions $\ell_{h,\Phi}(\Psi(r), t)$ (Definition 2) obey $\frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r), t) \in \mathrm{G}$. We have:*

$$\begin{aligned}&\epsilon_{CF}(h, \Phi) \leq \\ &\quad (1 - u)\epsilon_F^{t=1}(h, \Phi) + u\epsilon_F^{t=0}(h, \Phi) \\ &\quad + B_\Phi \cdot \mathit{IPM}_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right),\end{aligned}$$

where ϵ_{CF}, $\epsilon_F^{t=0}$ and $\epsilon_F^{t=1}$ are as in Definitions 2 and 3.

Theorem 1. *Under the conditions of Lemma 1, and assuming the loss L used to define $\ell_{h,\Phi}$ in Definitions 2 and 3 is the squared loss, we have:*

$$\begin{aligned}&\epsilon_{PEHE}(h, \Phi) \leq \\ &2\big(\epsilon_{CF}(h, \Phi) + \epsilon_F(h, \Phi) - 2\sigma_Y^2\big) \leq \qquad (2)\\ &2\big(\epsilon_F^{t=0}(h, \Phi) + \epsilon_F^{t=1}(h, \Phi) + B_\Phi \mathit{IPM}_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - 2\sigma_Y^2\big),\end{aligned}$$

where ϵ_F and ϵ_{CF} are defined w.r.t. the squared loss, and σ_Y^2 is the variance of the outcomes Y_t (see Definition A11 in Appendix for detailed definition).

The main idea of the proof is showing that ϵ_{PEHE} is upper bounded by the sum of the expected factual loss ϵ_F and expected counterfactual loss ϵ_{CF}. However, we cannot estimate ϵ_{CF}, since we only have samples relevant to ϵ_F. We therefore bound the difference $\epsilon_{CF} - \epsilon_F$ using an IPM.

Choosing a small function family G makes the bound tighter. However, choosing too small a family could result in an incomputable bound. For example, for the minimal choice $\mathrm{G} = \{\ell_{h,\Phi}(x, 0), \ell_{h,\Phi}(x, 1)\}$, we will have to evaluate an expectation term of Y_1 over $p_\Phi^{t=0}$, and of Y_0 over $p_\Phi^{t=1}$. We cannot in general evaluate these expectations, since by assumption when $t = 0$ we only observe Y_0, and the same for $t = 1$ and Y_1. In addition, for some function families there is no known way to efficiently compute the IPM distance or its gradients. Here, we use two function families for which there are available optimization tools. The first is the family of 1-Lipschitz functions, which leads to IPM being the Wasserstein distance (Villani, 2008), denoted $\mathrm{Wass}(p, q)$. The second is the family of norm-1 reproducing kernel Hilbert space (RKHS) functions, leading to the MMD metric (Gretton et al., 2012), denoted $\mathrm{MMD}(p, q)$. Both the Wasserstein and MMD metrics have consistent estimators which can be efficiently computed for finite samples (Sriperumbudur et al., 2012), and have been used for various machine learning tasks in recent years (Gretton et al., 2009; 2012; Cuturi & Doucet, 2014).

In order to explicitly evaluate the constant B_Φ in Theorem 1, we have to make some assumptions about the elements of the problem. For the Wasserstein case these are the loss L, the Lipschitz constants of $p(Y_t|x)$ and h, and the condition number of the Jacobian of Φ. For the MMD case, we make assumptions about the RKHS representability and RKHS norms of h , Φ, and the standard deviation of $Y_t|x$. The full details are given in the supplement, with the major results stated in Theorems 2 and 3. In all cases we obtain that making Φ smaller increases the constant B_Φ precluding trivial solutions such as making Φ arbitrarily small.

For an empirical sample, and a *family* of representations and hypotheses, we can further upper bound $\epsilon_F^{t=0}$ and $\epsilon_F^{t=1}$ by their respective empirical losses and a model complexity term using standard arguments (Shalev-Shwartz & Ben-David, 2014). The IPMs we use can be consistently estimated from finite samples (Sriperumbudur et al., 2012). The negative variance term σ_Y^2 arises from the fact that, following Hill (2011); Athey & Imbens (2016), we define the error ϵ_{PEHE} in terms of the conditional mean functions $m_t(x)$, as opposed to fitting the random variables Y_t.

Our results hold for any given h and Φ obeying the Theorem conditions. This immediately suggest an algorithm in which we minimize the upper bound in Eq. (2) with respect to Φ and h and either the Wasserstein or MMD IPM, in order to minimize the error in estimating the individual treatment effect. This leads us to Algorithm 1 below.

4. Algorithm for estimating ITE

We propose a general framework called CFR (for Counterfactual Regression) for ITE estimation based on the theoretical results above. Our algorithm is an end-to-end, regularized minimization procedure which fits both a balanced representation of the data and a hypothesis for the outcome. CFR draws on the same intuition as our previous work (Johansson et al., 2016), but *overcomes* the following limitations: a) Our previous theory requires a two-step optimization procedure and is specific to *linear* hypotheses (it does not support e.g. deep neural networks), b) The treatment indicator might be washed out in the old model, if the learned representation is high-dimensional (see discussion below).

We assume there exists a distribution $p(x, t, Y_0, Y_1)$ over $\mathcal{X} \times \{0, 1\} \times \mathcal{Y} \times \mathcal{Y}$, such that strong ignorability holds. We further assume we have a sample from that distribution $(x_1, t_1, y_1), \ldots (x_n, t_n, y_n)$, where $y_i \sim p(Y_1|x_i)$ if $t_i = 1$, $y_i \sim p(Y_0|x_i)$ if $t_i = 0$. This standard assumption means that the treatment assignment determines which potential outcome we see. Our goal is to find a representation $\Phi : \mathcal{X} \to \mathcal{R}$ and hypothesis $h : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ that will minimize $\epsilon_{\text{PEHE}}(f)$ for $f(x,t) := h(\Phi(x),t)$.

We parameterize $\Phi(x)$ and $h(\Phi,t)$ by deep neural networks trained jointly, see Figure 1. This allows for learning complex non-linear representations and hypotheses with large flexibility. In Johansson et al. (2016), we parameterized $h(\Phi,t)$ with a single network, concatenating Φ and t as input. In this case, if Φ is high-dimensional, the influence of t on h might be lost during training. To combat this, we parameterize $h_1(\Phi) = h(\Phi,1)$ and $h_0(\Phi) = h(\Phi,0)$ as two separate "heads" of the joint network, the former used to estimate the outcome under treatment, and the latter under control. This way, statistical power is shared in representation layers, while the effect of treatment is preserved in the separate heads. Note that each sample is used to update only the head corresponding to the observed treatment.

Our second contribution is to explicitly adjust for the bias induced by treatment group imbalance. To this end, we seek a representation Φ and hypothesis h that minimizes a trade-off between predictive accuracy and imbalance in the representation space, using the following objective:

$$\begin{aligned}
\min_{\substack{h,\Phi \\ \|\Phi\|=1}} \quad & \frac{1}{n}\sum_{i=1}^{n} w_i \cdot L\left(h(\Phi(x_i),t_i),y_i\right) + \lambda \cdot \mathfrak{R}(h) \\
& +\alpha \cdot \text{IPM}_{\text{G}}\left(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}\right), \\
\text{with} \quad & w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}, \quad \text{where} \quad u = \frac{1}{n}\sum_{i=1}^{n} t_i, \\
\text{and} \quad & \mathfrak{R} \text{ is a model complexity term.} \qquad (3)
\end{aligned}$$

Note that $u = p(t=1)$ is simply the proportion of treated units in the population. The weights w_i compensate for the difference in treatment group size in our sample, see Theorem 1. $\text{IPM}_{\text{G}}(\cdot,\cdot)$ is the (empirical) integral probability metric w.r.t. G. For most IPMs, we cannot compute the factor B_ϕ in (2), but treat it as part of the hyperparameter α. This makes our objective sensitive to the scaling of Φ, even for a constant α. We therefore normalize Φ through either projection or batch-normalization with fixed scale.

We refer to the model minimizing (3) with $\alpha > 0$ as Counterfactual Regression (CFR) and the variant without balance regularization ($\alpha = 0$) as Treatment-Agnostic Representation Network (TARNet). Both models are trained by minimizing (3) using stochastic gradient descent, as described in Algorithm 1. Both the prediction loss and the penalty term $\text{IPM}_{\text{G}}(\cdot,\cdot)$ are computed for one mini-batch at a time. Details of how to obtain the gradient g_1 with respect to the empirical IPMs are in the supplement.

Algorithm 1 CFR: Counterfactual regression with integral probability metrics

```
Input: Factual sample $(x_1,t_1,y_1),\ldots,(x_n,t_n,y_n)$, scaling parameter $\alpha > 0$, loss function $L(\cdot,\cdot)$, representation network $\Phi_{\mathbf{W}}$ with initial weights $\mathbf{W}$, outcome network $h_{\mathbf{V}}$ with initial weights $\mathbf{V}$, function family G for IPM.
Compute $u = \frac{1}{n}\sum_{i=1}^{n} t_i$
Compute $w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}$ for $i = 1 \ldots n$
while not converged do
    Sample mini-batch $\{i_1, i_2, \ldots, i_m\} \subset \{1,2,\ldots,n\}$
    Calculate the gradient of the IPM term:
    $g_1 = \nabla_{\mathbf{W}} \text{IPM}_{\text{G}}(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{t_{i_j}=0}, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{t_{i_j}=1})$
    Calculate the gradients of the empirical loss:
    $g_2 = \nabla_{\mathbf{V}} \frac{1}{m}\sum_j w_{i_j} \cdot L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}\right)$
    $g_3 = \nabla_{\mathbf{W}} \frac{1}{m}\sum_j w_{i_j} \cdot L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}\right)$
    Obtain step size scalar or matrix $\eta$ with standard neural net methods e.g. Adam (Kingma & Ba, 2015)
    $[\mathbf{W}, \mathbf{V}] \leftarrow [\mathbf{W} - \eta(\alpha g_1 + g_3), \mathbf{V} - \eta(g_2 + 2\lambda\mathbf{V})]$
    Check convergence criterion
end while
```

5. Experiments

Evaluating causal inference algorithms is more difficult than many machine learning tasks, since we rarely have access to the ground truth treatment effect. Existing literature mostly deals with this in two ways. One is by using (semi-) synthetic datasets, where the outcome or treatment assignment are fully known; we use the semi-synthetic IHDP dataset from Hill (2011). The other is using real-world data from randomized controlled trials (RCT). The problem with using data from RCTs is that there is no imbalance between treatment groups, making our method redundant. We partially overcome this problem by using the Jobs dataset from LaLonde (1986), which includes both a randomized and a non-randomized component. We use both components for training, but only use the randomized component for evaluation. This alleviates, but does not solve, the issue of a completely randomized and balanced dataset being unsuited for our method.

We evaluate our framework CFR, and its variant without balancing regularization (TARNet), in the task of estimating ITE and ATE. Both versions are implemented[3] as feed-forward neural networks with 3 fully-connected exponential-linear layers (Clevert et al., 2016) for the representation and 3 for the hypothesis. Layer sizes were 200 for all layers used for Jobs and 200 and 100 for the representation and hypothesis used for IHDP. The model is trained using Adam (Kingma & Ba, 2015). The hypothesis parameters are regularized with a small ℓ_2 weight decay. For continuous data we use mean squared loss and for binary data, we use log-loss. While our theory does not immediately apply to log-loss, we were curious to see how our model performs with it. We use the Wasserstein (CFR_{WASS}) and the squared linear MMD (CFR_{MMD}) distances to penalize

[3] https://github.com/clinicalml/cfrnet

imbalance.

We compare our method to Ordinary Least Squares with treatment as a feature (OLS$_1$), OLS with separate regressors for each treatment (OLS$_2$), k-nearest neighbor (k-NN), Targeted Maximum Likelihood (TMLE), which is a doubly robust method (Gruber & van der Laan, 2011), Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Chipman & McCulloch, 2016), Random Forests (R. For.) (Breiman, 2001), Causal Forests (C. For.) (Wager & Athey, 2015) as well as the Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) from Johansson et al. (2016). For classification tasks we substitute Logistic Regression (LR) for OLS. Choosing hyperparameters for estimating PEHE is non-trivial; we detail our general procedure using a validation set, in subsection C.1 of the supplement.

We consider two different estimation tasks. One is *within-sample*, where the task is to estimate ITE for all units in a sample for which the (factual) outcome of *one* treatment is observed. This corresponds to the common scenario in which a cohort is selected once and not changed. This task is non-trivial, as we never observe the ITE for any unit. The other is *out-of-sample*, where the goal is to estimate ITE for units with *no* observed outcomes. This corresponds to the problem of selecting the best treatment for a *new* patient. Within-sample error is computed over both the training and validation sets, out-of-sample error over the test set.

5.1. Simulated outcome: IHDP

Hill (2011) compiled a dataset for causal effect estimation based on the Infant Health and Development Program (IHDP), in which the covariates come from a randomized experiment studying the effects of specialist home visits on cognitive test scores. The treatment groups have been made imbalanced by removing a biased subset of the treated population. The dataset comprises 747 units (139 treated, 608 control) and 25 covariates measuring aspects of children and their mothers. We use the simulated outcome implemented as setting "A" in the NPCI package (Dorie, 2016). Following Hill (2011), we use the *noiseless* outcome to compute the true effect. We report the estimated (finite-sample) PEHE loss ϵ_{PEHE} (Eq. 1), and the absolute error in average treatment effect $\epsilon_{\text{ATE}} = |\frac{1}{n}\sum_{i=1}^{n}(f(x_i, 1) - f(x_i, 0)) - \frac{1}{n}\sum_{i=1}^{n}(m_1(x_i) - m_0(x_i))|$. The results of the experiments on IHDP are presented in Table 1 (left). We average over 1000 realizations of the outcomes with 63/27/10 train/validation/test splits.

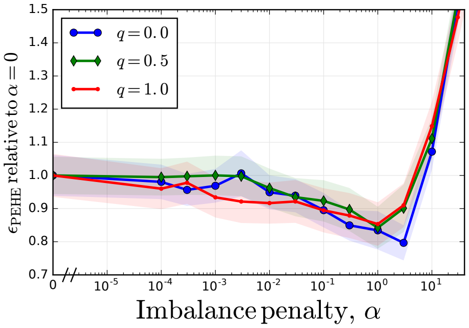

Figure 2. Out-of-sample ITE error versus IPM regularization for CFR Wass, relative to the error at $\alpha = 0$, on 500 realizations of IHDP, with high ($q = 1$), medium and low (artificial) imbalance between control and treated.

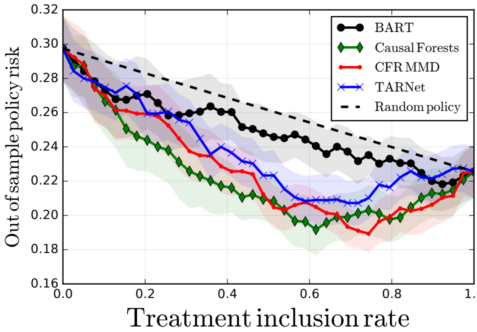

Figure 3. Policy risk on Jobs as a function of treatment inclusion rate. Lower is better. Subjects are included in treatment in order of their estimated treatment effect given by the various methods. CFR Wass is similar to TARNet and is omitted to avoid clutter.

We also investigate the effects of increasing imbalance between the original treatment groups by constructing biased subsamples of the IHDP dataset. A logistic-regression propensity score model is fit to form estimates $\hat{p}(t = 1|x)$ of the conditional treatment probability. Then, repeatedly, with probability q we remove the remaining *control* observation x that has $\hat{p}(t = 1|x)$ closest to 1, and with probability $1 - q$, we remove a random control observation. The higher q, the more imbalance. For each value of q, we remove 347 observations from each set, leaving 400.

5.2. Real-world outcome: Jobs

The study by LaLonde (1986) is a widely used benchmark in the causal inference community, where the treatment is job training and the outcomes are income and employment status after training. This dataset combines a randomized study based on the National Supported Work program with observational data to form a larger dataset (Smith & Todd, 2005). The presence of the randomized subgroup gives a way to estimate the "ground truth" causal effect. The study includes 8 covariates such as age and education, as well as previous earnings. We construct a *binary* classification task, called *Jobs*, where the goal is to predict unemployment, using the feature set of Dehejia & Wahba (2002). Following Smith & Todd (2005), we use the LaLonde experimental sample (297 treated, 425 control) and the PSID comparison group (2490 control). There were 482 (15%) subjects unemployed by the end of the study. We average

Table 1. Results on IHDP and Jobs within-sample (left) and out-of-sample (right). Lower is better. †Not applicable.

Within-sample					**Out-of-sample**				
	IHDP		JOBS			IHDP		JOBS	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}
OLS/LR$_1$	$5.8 \pm .3$	$.73 \pm .04$	$.22 \pm .00$	$.01 \pm .00$	OLS/LR$_1$	$5.8 \pm .3$	$.94 \pm .06$	$.23 \pm .02$	$.08 \pm .04$
OLS/LR$_2$	$2.4 \pm .1$	$.14 \pm .01$	$.21 \pm .00$	$.01 \pm .01$	OLS/LR$_2$	$2.5 \pm .1$	$.31 \pm .02$	$.24 \pm .01$	$.08 \pm .03$
BLR	$5.8 \pm .3$	$.72 \pm .04$	$.22 \pm .01$	$.01 \pm .01$	BLR	$5.8 \pm .3$	$.93 \pm .05$	$.25 \pm .02$	$.08 \pm .03$
k-NN	$2.1 \pm .1$	$.14 \pm .01$	$.02 \pm .00$	$.21 \pm .01$	k-NN	$4.1 \pm .2$	$.79 \pm .05$	$.26 \pm .02$	$.13 \pm .05$
TMLE	$5.0 \pm .2$	$.30 \pm .01$	$.22 \pm .00$	$.02 \pm .01$	TMLE	†	†	†	†
BART	$2.1 \pm .1$	$.23 \pm .01$	$.23 \pm .00$	$.02 \pm .00$	BART	$2.3 \pm .1$	$.34 \pm .02$	$.25 \pm .02$	$.08 \pm .03$
R.FOR.	$4.2 \pm .2$	$.73 \pm .05$	$.23 \pm .01$	$.03 \pm .01$	R.FOR.	$6.6 \pm .3$	$.96 \pm .06$	$.28 \pm .02$	$.09 \pm .04$
C.FOR.	$3.8 \pm .2$	$.18 \pm .01$	$.19 \pm .00$	$.03 \pm .01$	C.FOR.	$3.8 \pm .2$	$.40 \pm .03$	$.20 \pm .02$	$.07 \pm .03$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$.20 \pm .01$	$.04 \pm .01$	BNN	$2.1 \pm .1$	$.42 \pm .03$	$.24 \pm .02$	$.09 \pm .04$
TARNET	$.88 \pm .02$	$.26 \pm .01$	$.17 \pm .01$	$.05 \pm .02$	TARNET	$.95 \pm .02$	$.28 \pm .01$	$.21 \pm .01$	$.11 \pm .04$
CFR$_{\text{MMD}}$	$.73 \pm .01$	$.30 \pm .01$	$.18 \pm .00$	$.04 \pm .01$	CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$	$.21 \pm .01$	$.08 \pm .03$
CFR$_{\text{WASS}}$	$.71 \pm .02$	$.25 \pm .01$	$.17 \pm .01$	$.04 \pm .01$	CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$	$.21 \pm .01$	$.09 \pm .03$

over 10 train/validation/test splits with ratios 56/24/20.

Because all the treated subjects T were part of the original randomized sample E, we can compute the true average treatment effect on the treated by ATT $= |T|^{-1} \sum_{i \in T} y_i - |C \cap E|^{-1} \sum_{i \in C \cap E} y_i$, where C is the control group. We report the error $\epsilon_{\text{ATT}} = |\text{ATT} - \frac{1}{|T|} \sum_{i \in T}(f(x_i, 1) - f(x_i, 0))|$. We cannot evaluate ϵ_{PEHE} on this dataset, since there is no ground truth for the ITE. Instead, in order to evaluate the quality of ITE estimation, we use a measure we call *policy risk*. The policy risk is defined as the average loss in value when treating according to the policy implied by an ITE estimator. In our case, for a model f, we let the policy be to treat, $\pi_f(x) = 1$, if $f(x, 1) - f(x, 0) > \lambda$, and to not treat, $\pi_f(x) = 0$ otherwise. The policy risk is $R_{\text{Pol}}(\pi_f) = 1 - (\mathbb{E}[Y_1 | \pi_f(x) = 1] \cdot p(\pi_f = 1) + \mathbb{E}[Y_0 | \pi_f(x) = 0] \cdot p(\pi_f = 0))$ which we can estimate for the randomized trial subset of Jobs by $\hat{R}_{\text{Pol}}(\pi_f = 1 - (\mathbb{E}[Y_1 | \pi_f(x) = 1, t = 1] \cdot p(\pi_f = 1) + \mathbb{E}[Y_0 | \pi_f(x) = 0, t = 0] \cdot p(\pi_f = 0))$. See figure 3 for risk as a function of treatment threshold λ, aligned by proportion of treated, and Table 1 for the risk when $\lambda = 0$.

5.3. Results

We note that indeed imbalance confers an advantage to using the IPM regularization term, as our theoretical results indicate, see e.g. the results for CFR$_{\text{WASS}}$ and TARNet on IHDP in Table 1. We also see in Figure 2 that even for the harder case of increased imbalance ($q > 0$) between treated and control, the relative gain from using our method remains significant. On Jobs, our proposed methods are better than or competitive with state-of-the-art, but we don't see a significant gain from using IPM penalties. This might be because we evaluate the predictions only on a randomized subset with treatment groups distributed identically. Non-linear estimators perform significantly better than linear ones in terms of individual effect (ϵ_{PEHE}). On the Jobs dataset, straightforward logistic regression does remarkably well in estimating the ATT. However, being a linear model, LR can only ascribe a uniform policy - in this case, "treat everyone". The more nuanced policies offered by non-linear methods achieve lower policy risk in the case of Causal Forests and CFR. This emphasizes the fact that estimating average effect and individual effect can require different models. Specifically, while smoothing over many units may yield a good ATE estimate, this might significantly hurt ITE estimation. k-nearest neighbors has very good within-sample results on Jobs, because evaluation is performed over the randomized component, but suffers heavily in generalizing out of sample, as expected.

6. Conclusion

In this paper we give a meaningful and intuitive error-bound for estimating individual treatment effect. Our bound relates ITE estimation to the classic machine learning problem of learning from samples, along with methods for measuring distributional distances from samples. The bound lends itself naturally to the creation of learning algorithms; we focus on using neural nets as representations and hypotheses. We apply our theory-guided approach to both synthetic and real-world tasks, showing that in every case our method matches or outperforms the state-of-the-art. Important open questions are theoretical considerations in choosing the IPM weight α, how to best derive confidence intervals for our model's predictions, and integrating our work with more complicated causal models such as those with hidden confounding or instrumental variables.

ACKNOWLEDGMENTS

We thank Aahlad Puli for his assistance with the experiments; Sanjog Misra and Günter J. Hitsch for suggesting the policy risk evaluation; Jennifer Hill, Marco Cuturi and Esteban Tabak for fruitful conversations; and Stefan Wager for his help with the code for Causal Forests. DS and US were supported by NSF CAREER award #1350965.

References

Athey, Susan and Imbens, Guido. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, Susan, Imbens, Guido W, and Wager, Stefan. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*, 2016.

Austin, Peter C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Bareinboim, Elias and Pearl, Judea. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Beck, Nathaniel, King, Gary, and Zeng, Langche. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review*, 94(01):21–35, 2000.

Belloni, Alexandre, Chernozhukov, Victor, and Hansen, Christian. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2): 608–650, 2014.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Pereira, Fernando, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19: 137, 2007.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Vaughan, Jennifer Wortman. A theory of learning from different domains. *Machine learning*, 79(1-2): 151–175, 2010.

Bengio, Yoshua, Courville, Aaron, and Vincent, Pierre. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35 (8):1798–1828, 2013.

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

Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, et al. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

Chipman, Hugh and McCulloch, Robert. BayesTree: Bayesian Additive Regression Trees. `https://cran.r-project.org/web/packages/BayesTree`, 2016.

Chipman, Hugh A, George, Edward I, and McCulloch, Robert E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, pp. 266–298, 2010.

Clevert, Djork-Arné, Unterthiner, Thomas, and Hochreiter, Sepp. Fast and accurate deep network learning by exponential linear units (elus). *International Conference on Learning Representations*, 2016.

Cole, Stephen R, Hernán, Miguel A, Robins, James M, Anastos, Kathryn, Chmiel, Joan, Detels, Roger, Ervin, Carolyn, Feldman, Joseph, Greenblatt, Ruth, Kingsley, Lawrence, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, 158(7):687–694, 2003.

Cortes, Corinna and Mohri, Mehryar. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 685–693, 2014.

Daumé III, Hal. Frustratingly easy domain adaptation. *Conference of the Association for Computational Linguistics (ACL)*, 2007.

Dehejia, Rajeev H and Wahba, Sadek. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

Dorie, Vincent. NPCI: Non-parametrics for Causal Inference. `https://github.com/vdorie/npci`, 2016.

Funk, Michele Jonsson, Westreich, Daniel, Wiesen, Chris, Stürmer, Til, Brookhart, M Alan, and Davidian, Marie. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Ganin, Yaroslav, Ustinova, Evgeniya, Ajakan, Hana, Germain, Pascal, Larochelle, Hugo, Laviolette, François, Marchand, Mario, and Lempitsky, Victor. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL `http://jmlr.org/papers/v17/15-239.html`.

Gretton, Arthur, Smola, Alex, Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten, and Schölkopf, Bernhard. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012. ISSN 1532-4435.

Gruber, Susan and van der Laan, Mark J. tmle: An r package for targeted maximum likelihood estimation. 2011.

Hartford, Jason, Lewis, Greg, Leyton-Brown, Kevin, and Taddy, Matt. Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*, 2016.

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

Hoyer, Patrik O, Janzing, Dominik, Mooij, Joris M, Peters, Jonas, and Schölkopf, Bernhard. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

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

Johansson, Fredrik D., Shalit, Uri, and Sontag, David. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

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

Maathuis, Marloes H, Colombo, Diego, Kalisch, Markus, and Bühlmann, Peter. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

Mansour, Yishay, Mohri, Mehryar, and Rostamizadeh, Afshin. *Domain adaptation: Learning bounds and algorithms*. 2009.

Mooij, Joris M, Peters, Jonas, Janzing, Dominik, Zscheischler, Jakob, and Schölkopf, Bernhard. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Müller, Alfred. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pp. 429–443, 1997.

Pan, Sinno Jialin, Tsang, Ivor W, Kwok, James T, and Yang, Qiang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2):199–210, 2011.

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

Pearl, Judea. Detecting latent heterogeneity. *Sociological Methods & Research*, pp. 0049124115600597, 2015.

Peysakhovich, Alexander and Lada, Akos. Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385*, 2016.

Robins, James. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

Rolling, Craig Anthony. *Estimation of Conditional Average Treatment Effects*. PhD thesis, University of Minnesota, 2014.

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

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

Shpitser, Ilya and Pearl, Judea. Identification of conditional interventional distributions. In *Proceedings of the Twenty-second Conference on Uncertainty in Artificial Intelligence*, pp. 437–444. UAI Press, 2006.

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

Sriperumbudur, Bharath K, Fukumizu, Kenji, Gretton, Arthur, Schölkopf, Bernhard, Lanckriet, Gert RG, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Strehl, Alex, Langford, John, Li, Lihong, and Kakade, Sham M. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pp. 2217–2225, 2010.

Sun, Baochen, Feng, Jiashi, and Saenko, Kate. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Swaminathan, Adith and Joachims, Thorsten. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

Taddy, Matt, Gardner, Matt, Chen, Liyun, and Draper, David. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Triantafillou, Sofia and Tsamardinos, Ioannis. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16: 2147–2205, 2015.

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

Wager, Stefan and Athey, Susan. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*. `https://github.com/susanathey/causalTree`, 2015.