Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit
CIMS, New York University
shalit@cs.nyu.edu

Fredrik D. Johansson
CSE, Chalmers University
ml@fredjo.com

David Sontag
CIMS, New York University
dsontag@cs.nyu.edu

Abstract

There is intense interest in applying machine learning to problems of causal inference such as precision medicine and personalized advertising. We give a new theoretical analysis and family of algorithms for estimating individual treatment effect (ITE) from observational data, based on learning representations such that the induced treated and control distributions look similar. We give a novel and intuitive bound showing that the 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 distance. Experiments on real and simulated data show state-of-the-art performance.

1 Introduction

Causal inference questions are central to policy makers and scientists across many fields. Examples abound: in healthcare one is interested in the relative efficacy of different medications; in economics, policy makers debate the effect of job training on an individual's earnings; in marketing, companies are interested in the causal effect of an online ad on a customer's buying habits. Whereas much of the work on causal effect inference has been focused on estimating the *average* treatment effect (ATE), in this paper, we focus on the problem of estimating *individual-level* treatment effect. Specifically, we learn to predict the effect of a proposed treatment for each unit (be it a patient, employee, customer, etc.). A treatment could be medication, job training, or showing an ad. The learning is done from *observational* data: data that was collected with treatment assignment potentially dependent on the unit's characteristics. This could be past medical records, a national dataset of workers' training and earnings, or a dataset of customers' online browsing. The learning problem is different from a classic learning problem, in that in our training data we never see the individual treatment effect. For each unit, *we only see their response to one of the possible treatments* - the one they had actually received. This is close to what is known in the machine learning literature as "learning from logged bandit feedback" [33, 35].

In this paper we give a novel generalization bound on the expected error of estimating the per-unit causal effect. The bound leads naturally to a new family of representation-learning based algorithms [9], which we show match or outperform state-of-the-art methods on causal effect inference tasks.

We frame our results using the Rubin-Neyman potential outcomes framework [29], as follows. We assume that for a unit $x \in \mathcal{X}$, and a treatment or intervention $t \in \{0, 1\}$, there are two potential outcomes: Y_0 and Y_1. In our data, for each unit we only see one of the potential outcomes, depending on the treatment assignment: if $t = 0$ we observe Y_0, if $t = 1$, we observe Y_1. 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 after treatments $t = 0$ and $t = 1$, respectively. We employ the standard

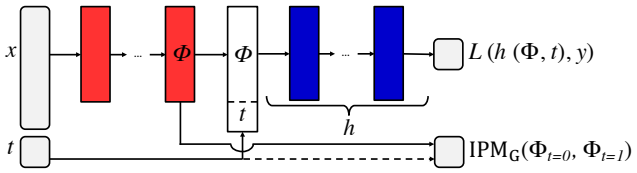

Figure 1: Architecture for ITE estimation. L is a loss function, IPM_G is an integral probability metric.

strong ignorability assumption: $(Y_1, Y_0) \perp\!\!\!\perp x \mid t$, and $0 < p(t = 0) < 1$. This assumption is sometimes stated roughly as the "no hidden confounders" assumption.

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 the individual unit x, or the Individual Treatment Effect (ITE) [1]. For example, for a patient x, we decide which of two treatments will have a better outcome. In this paper we are mainly interested in observational data, i.e. the case where the distribution of the treatment assignment t is dependent on x. For example, richer patients might better afford different medications.

The function $\tau(x)$ can be estimated by learning the two functions $m_0(x)$ and $m_1(x)$ using samples from $p(Y_t|x, t)$. Unlike in the standard machine learning problem, 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. This is similar to the phenomenon of covariate shift [24, 22]. 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. We use two specific IPMs: the Maximum Mean Discrepancy [18], and the Wasserstein distance [37, 14]. We show that the expected error of learning the individual treatment effect function $\tau(x)$ is upper bounded by the error of learning Y_1, Y_0, plus the IPM term. In the randomized trial setting, where $t \perp\!\!\!\perp x$, $p(t) = 0.5$, the IPM term is 0, and our bound naturally reduces to a standard learning problem.

The bound we derive points the way to a family of algorithms based on the idea of representation learning [9]: Jointly learn a hypothesis and a representation which minimize a weighted sum of the (supervised) factual loss, and the IPM distance between the control and treated distributions induced by the representation. In experiments, we apply algorithms based on deep neural networks as representations and hypotheses, along with MMD or Wasserstein distributional distance metrics over the representation layer; see Figure 1. A similar idea was recently proposed by [22], but the algorithms we propose are conceptually simpler, have richer and more flexible theory, and achieve better results in practice. We show that our methods achieve competitive results on a real-world causal inference benchmark: the widely used National Supposed Work survey [23, 31].

2 Related work

The most common goal of causal effect inference as used in the applied sciences is to obtain the average treatment effect: $ATE = \mathbb{E}_{x\sim p(x)}[\tau(x)]$. One of the most widely used approaches to estimating ATE is covariate adjustment, also known as back-door adjustment or the G-computation formula [28, 29]. In its basic version, covariate adjustment amounts to estimating the functions $m_1(x)$, $m_0(x)$ and is therefore a natural method for estimating ITE as well as ATE. Another widely used family of causal effect inference methods are weighting methods. Methods such as propensity score weighting [3] re-weight the units in the observational data so as to make the treated and control populations more comparable. These methods, and the related doubly robust methods [16], do not yield themselves immediately to estimating an individual level effect, however.

Adapting machine learning methods for causal effect inference, and in particular for individual level treatment effect, has gained much interest recently. For example [38, 1] discuss how tree-based methods can be adapted to obtain a consistent estimator with semi-parametric asymptotic convergence rate. Others show how to adapt high-dimensional regression methods such as Lasso to consistently estimate treatment effect, again achieving semi-parametric rates [6, 2]. Our work differs by focusing

[1] This term is sometimes known as the Conditional Average Treatment Effect, CATE.

on the generalization error aspects of estimating individual treatment effect, as opposed to asymptotic consistency. Another line of work in the causal inference community relates to bounding the estimate of the average treatment effect given an instrumental variable [4, 5], or under hidden confounding, for example when the ignorability assumption does not hold [28, 10]. Our work differs, in that we only deal with the ignorable case, and in that we bound a very different quantity: the generalization error of estimating individual level treatment effect.

Our work is connected to work on domain adaptation, as estimating ITE requires prediction over a distribution different from the observed one. Our ITE error bound has similarities with generalization bounds in domain adaptation given by [8, 24, 7, 13]. 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 [17], and in principle other domain adaptation methods (e.g. [27, 34] could be adapted for use in ITE estimation.

Finally, our paper builds upon recent work by [22], who show a connection between covariate shift and the task of estimating the counterfactual outcome. They propose learning a representation of the data that makes the treated and control distributions more similar, and fitting a linear ridge-regression model on top of it. They bound the *relative* error of fitting a ridge-regression using the counterfactual distribution versus fitting a ridge-regression using the factual distribution. Unfortunately, this not informative regarding the absolute quality of the representation. In this paper we focus on a closely related but more substantive task: estimating the individual treatment effect. We further 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.

3 Estimating ITE: Error bounds

We bound the expected error in estimating the individual treatment effect for a given representation, and a hypothesis defined over that representation. The bounds are expressed in terms of (1) the expected loss of the model over the fitting of the potential outcomes Y_0, Y_1, and (2) an Integral Probability Metric (IPM) distance measure between the distribution of treated and control units.

We will employ the following assumptions and notations. The space of covariates is a bounded subset $\mathcal{X} \subset \mathbb{R}^d$. The outcome space is $\mathcal{Y} \subset \mathbb{R}$. Treatment is a binary variable. We assume there exists a joint distribution $p(x, t, Y_0, Y_1)$, such that $(Y_1, Y_0) \perp\!\!\!\perp t|x$. We call the marginal of p over (x, t) the *factual distribution*, and denote it $p^F(x, t)$. The treated and control distributions are the factual distribution conditioned on treatment: $p^{t=1}(x) := p^F(x|t = 1)$, and $p^{t=0}(x) := p^F(x|t = 0)$, 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.*

Definition 1. *Define $\Psi : \mathcal{R} \to \mathcal{X}$ to be 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^F_Φ, defined over $\mathcal{R} \times \{0, 1\}$. We also define $p^{t=1}_\Phi(r) := p^F_\Phi(r|t = 1)$, $p^{t=0}_\Phi(r) := p^F_\Phi(r|t = 0)$, to be the treated and control distributions induced over $\mathcal{R}$. For a one-to-one Φ, the distributions p^F_Φ and p^{CF}_Φ 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, let $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ be a hypothesis defined over the representation space $\mathcal{R}$, and let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function.

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 loss of h and Φ is:*

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

$\epsilon_F(h, \Phi)$ is the generalization error for the hypothesis $h(\Phi(x), t)$ over the factual distribution.

Definition 3. *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 a hypothesis. For example, we could have that $f(x,t) = h(\Phi(x),t)$.

Definition 4. *The treatment effect estimate of the hypothesis f for unit x is:*

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

Definition 5. *The expected Precision in Estimation of Heterogeneous Effect (PEHE) loss of f is:*

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

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

Our main result relies on the notion of an *Integral Probability Metric* (IPM), which is a family of metrics between probability distributions [32, 26]. 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 $\text{IPM}_\text{G}(p,q) := \sup_{g\in\text{G}} \left|\int_\mathcal{S} g(s)(p(s) - q(s))\, ds\right|$. Integral probability metrics are always symmetric and obey the triangle inequality, and trivially satisfy $\text{IPM}_\text{G}(p,p) = 0$. For rich enough function families G, we also have that $\text{IPM}_\text{G}(p,q) = 0 \implies p = q$, and then IPM_G is a true metric over the corresponding set of probabilities. Examples of function families G for which IPM_G is a true metric are the family of all bounded countinuous functions, the family of all 1-Lipschitz functions [32], and the unit-ball of functions in a universal reproducing Hilbert kernel space [18]. Here, we will employ an extension of IPM, with probabilities scaled by positive scalars [12]. We call this *Unnormalized Integral Probability Metric*. For two probability distribution functions as above, and positive scalars u_p, u_q, we have:

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

Recall that $m_t(x) = \mathbb{E}\left[Y_t|x\right]$. The expected variance of Y_t with respect to $p(x,t)$ is $\sigma^2_Y(p) = \int_\mathcal{X} \left(Y_t - m_t(x)\right)^2 p(Y_t|x)p(x,t)\, dY_t dx dt$. We then define: $\sigma^2_Y = \min\{\sigma^2_Y(p^F(x,t)), \sigma^2_Y(p^F(x,1-t)\}$. If Y_t is a deterministic function of x, then $\sigma^2_Y = 0$. Let $u = p^F(t=1)$ be the marginal probability of treatment. By the strong ignorability assumption, $0 < u < 1$.

Theorem 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 a hypothesis. Let* G *be a family of functions $f : \mathcal{X} \to \mathcal{Y}$. Let the loss function L be the squared loss. Assume there exists a constant $B > 0$, such that for fixed $t \in \{0,1\}$, the per-unit expected loss functions $\ell_{h,\Phi}(x,t)$ (Def. 2) obey $\frac{1}{B} \cdot \ell_{h,\Phi}(x,t) \in \text{G}$. For the hypothesis $f : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ such that $f(x,t) = h(\Phi(x),t)$:*

$$\epsilon_{PEHE}(h,\Phi) \leq 4\epsilon_F(h,\Phi) + 4B \cdot \mathit{UIPM}_\text{G}\left(u \cdot p^{t=1}_\Phi, (1-u)\cdot p^{t=0}_\Phi\right) - 4\sigma^2_Y, \tag{2}$$

where ϵ_F is w.r.t. the squared loss.

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 a similar loss ϵ_{CF} defined with expectation over the so-called "counterfactual distribution" $p^{CF}(x,t) := p^F(x,1-t)$. Then, $\epsilon_{CF} - \epsilon_F$ is bounded using an IPM. For an empirical sample, and a *family* of representations and hypotheses, we can further upper bound ϵ_F by the empirical loss and a model complexity term using standard arguments [30]. In this paper we use two function families G for which there are available optimization tools. The first is the family of 1-Lipschitz functions, which leads to IPM being the Wasserstein distance [37, 32], denoted $\text{Wass}(p,q)$. The second is the family of norm-1 reproducing kernel Hilbert space (RKHS) functions, leading to the MMD metric [18, 32], denoted $\text{MMD}(p,q)$. See the full paper for a discussion on the choice of function family, how to estimate these terms and how to evaluate the constant B in Theorem 1.

4 Algorithm for estimating ITE

We propose a general framework for ITE estimation based on the theoretical results above. Our algorithm is a single regularized minimization procedure which simultaneously fits both a balanced representation of the data and a hypothesis for the outcome. This is in contrast to [22] who proposed a two-step procedure corresponding to their theoretical results based on the discrepancy distance

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 F for IPM
while not converged **do**
  Sample m control $\{(x_{i_j}, 0, y_{i_j})\}_{j=1}^{m}$ and m' treated units $\{(x_{i_k}, 1, y_{i_k})\}_{k=m+1}^{m+m'}$
  Calculate the gradient of the IPM term:
  $g_1 = \nabla_{\mathbf{W}} \text{IPM}_F(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^{m}, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m+m'})$
  Calculate the gradients of the empirical loss:
  $g_2 = \nabla_{\mathbf{V}} \frac{1}{m+m'} \sum_j L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}\right)$
  $g_3 = \nabla_{\mathbf{W}} \frac{1}{m+m'} \sum_j L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}\right)$
  Obtain step size scalar or matrix η with standard neural net methods e.g. RMSProp [36]
  Update $\mathbf{W} \leftarrow \mathbf{W} - \eta(\alpha \mathbf{g_1} + \mathbf{g_3})$, $\mathbf{V} \leftarrow \mathbf{V} - \eta \mathbf{g_2}$
  Check convergence criterion
end while

[13]. We also note that our theory supports multiple measures of balance that can can be minimized efficiently; this is only rarely true for variants of the discrepancy distance used by [22].

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 $f(x, t) := h(\Phi(x), t)$. Towards that end, we minimize the following objective:

$$\min_{\Phi, h} \frac{1}{n} \sum_{i=1}^{n} L\left(h(\Phi(x_i), t_i), y_i\right) + \quad \alpha \cdot \text{UIPM}_{\text{G}}\left(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}\right) \tag{3}$$

Here, $\text{UIPM}_{\text{G}}(\cdot, \cdot)$ is the (empirical) integral probability metric defined by the function family G.

In this work, we let $\Phi(x)$ and $h(\Phi, t)$ be parameterized by two neural networks and learn them jointly. This means that we can learn rich, non-linear representations and hypotheses with large flexibility. Our approach is visualized in Figure 1. We train our models by minimizing (3) using stochastic gradient descent, simultaneously backpropagating through the hypothesis and representation networks, see Algorithm 1. Details of how to obtain the gradient w.r.t. the empirical IPMs are in the full paper.

5 Experiments

We evaluate our framework CFR (for Counterfactual Regression) in the task of estimating ITE and ATE for all units in a sample $\{(x_i, t_i, y_i)\}_{i=1}^{n}$. CFR is implemented as a feed-forward neural network with fully-connected ReLU layers, trained using RMSProp, a small ℓ_2 weight decay, $\lambda = 10^{-3}$, and early stopping. Our architecture, dubbed CFR-2-2, consists of 2 ReLU representation layers, 2 ReLU layers after the treatment has been added, and a linear output layer (see Figure 1). For the IHDP data we use squared loss and hidden layers of 25 hidden units each. For the Jobs data, we use log-loss and layers of 50 units. The architectures were selected based on held-out factual error. For Jobs, we observed that performance varied considerably over different runs, and added batch normalization [21] to alleviate this problem.

Standard methods for hyperparameter selection, such as cross-validation, are not applicable when there are no samples of the counterfactual outcome (i.e. Y_0 when $t = 1$ and vice versa). For simulated outcomes, counterfactuals are available and we follow [22] by fitting hyperparameters on a held-out set of experiments. For real-world data, we use as surrogate the factual outcome $y_{j(i)}$ of the nearest neighbor $j(i)$ to i in the opposite treatment group, $t_{j(i)} = 1 - t_i$ (in the original space). We then choose hyperparameters based on a nearest-neighbor approximation of the PEHE loss, $\epsilon_{\text{PEHE}nn}(f) = \frac{1}{n} \sum_{i=1}^{n} \left((1 - 2t_i)(y_{j(i)} - y_i) - (f(x_i, 1) - f(x_i, 0))\right)^2$.

In regression tasks, we compare our method to Ordinary Least Squares (OLS), Targeted Maximum Likelihood, which is a doubly robust method (TMLE) [19], Bayesian Additive Regression Trees

	IHDP		JOBS, BIN.	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	$\epsilon_{\text{ATT\%}}$
OLS / LR	$5.8 \pm .3$	$.7 \pm .0$	.23	9%
TMLE	$5.0 \pm .2$	$.3 \pm .0$	.22	20%
L+R / ℓ_1-LR	$5.7 \pm .2$	$.2 \pm .0$	.23	9%
BLR	$5.7 \pm .3$	$.2 \pm .0$	†	†
BART	$1.7 \pm .2$	$.2 \pm .1$	.24	23%
C.FORESTS	$3.7 \pm .2$	$.2 \pm .0$	.17	34%
BNN-4-0	$5.6 \pm .3$	$.3 \pm .0$	†	†
BNN-2-2	$1.6 \pm .1$	$.3 \pm .0$	†	†
$\text{CFR}_{2-2\ \alpha=0}$	$1.5 \pm .1$	$.3 \pm .0$	.16	33%
$\text{CFR}_{2-2\ \text{WASS}}$	$1.4 \pm .1$	$.3 \pm .0$	.15	30%
$\text{CFR}_{2-2\ \text{MMD}}$	$1.4 \pm .1$	$.3 \pm .0$	.13	35%

Table 1: Results on IHDP over 1000 repeated experiments (left) and binary Jobs (right). MMD is squared linear MMD. $\epsilon_{\text{ATT\%}}$ is $100 \cdot \epsilon_{\text{ATT}}/\text{ATT}_{\text{true}}$. Lower is better. †Not applicable.

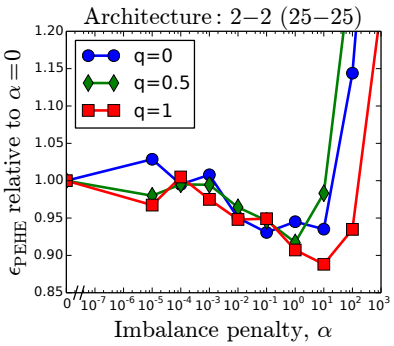

Figure 2: Error in causal effect as a function of IPM penalty on IHDP, with high ($q = 1$) and low ($q = 0$) overlap between control and treated.

(BART) [11], Causal Forests (C.Forests) [38] as well as the Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) by [22]. We also compare to a variable selection procedure dubbed LASSO + Ridge (L+R) in which a ridge regression model is fit to the variables selected by LASSO. In classification tasks we substitute Logistic Regression (LR) and ℓ_1-regularized Logistic Regression (ℓ_1-LR) for OLS and L+R respectively. The parameters of each model are fit in the same manner as for CFR (see above).

5.1 Simulated outcome: IHDP

Hill [20] compiled a semi-simulated dataset for causal effect estimation based on the Infant Health and Development Program (IHDP), in which the covariates stem from a randomized experiment studying the effects of child care and home visits on future cognitive test scores. The treatment groups have been made imbalanced by removing a biased subset of the treatment population. The dataset comprises 747 observations (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 R package. Following [20], we use the *noiseless* outcome to compute the true effect. We report the estimated (finite-sample) PEHE loss ϵ_{PEHE} (1), and the absolute error in average treatment effect $\epsilon_{\text{ATE}} = |\frac{1}{n}\sum_{i=1}^{n}(f(x_i,1) - f(x_i,0)) - \frac{1}{n}\sum_{i=1}^{n}(m_1(x_i) - m_0(x_i))|$. Results on another semi-simulated dataset called News [22] can be found in the full paper.

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

5.2 Real-world outcome: Jobs

The LaLonde study [23] is a widely used benchmark study in the causal inference literature, here referred to as *Jobs*. The outcome variable is yearly earnings and the treatment is job training. The Jobs dataset is a combination of a randomized study based on the National Supported Work program, and observational data [31]. The presence of the randomized subgroup gives a way to estimate the "ground truth" causal effect. The study includes 8 features such as age and education, as well as previous earnings. We construct a *binary* classification task of predicting unemployment, and use the augmented feature set of Dehejia & Wahba [15]. Following Smith et al. [31], we use the LaLonde experimental sample (297 treated, 425 control) and the PSID comparison group (2490 control). There were 482 subjects unemployed by the end of the study.

As all the treated T were part of the original randomized sample E, we can estimate the true average treatment effect on the treated by ATT $= \frac{1}{|T|}\sum_{i\in T} y_i - \frac{1}{|C\cap E|}\sum_{i\in C\cap E} y_i$, 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 we do not have the ITE for any of the units. Therefore, in order to evaluate the quality of ITE estimation, we use a measure we call *policy risk*. The 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 $f(x,1) - f(x,0) > 0$ 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 $\hat{R}_{\text{Pol}}(\pi_f = 1 - (\mathbb{E}[Y_1|\pi_f(x) = 1, t = 1]\cdot p(\pi_f = 1) + \mathbb{E}[Y_0|\pi_f(x) = 0, t = 0]\cdot p(\pi_f = 0))$. For results on policy risk as a function of treatment threshold, see the full paper.

5.3 Results

In Figure 2, we see that as for higher imbalance (q) between treated and control, the relative gain from using our method is higher, as well as the optimal weight α of the IPM penalty. The results on IHDP and Jobs are presented in Table 1. For IHDP, non-linear estimators do significantly better than linear ones in terms of individual effect (ϵ_{PEHE}). We see that using the IPM term ($\alpha > 0$) confers a small advantage over not using it, when estimating the individual effect (see also Figure 2). For the Jobs dataset, we see that straightforward logistic regression does remarkably well in estimating the ATE. 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. In particular, CFR with the MMD penalty achieves the lowest policy risk. 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.

References

[1] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[2] S. Athey, G. W. Imbens, and S. Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*, 2016.

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

[4] A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[5] E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *AISTATS*, 2012.

[6] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[8] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[9] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[10] Z. Cai, M. Kuroki, J. Pearl, and J. Tian. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701, 2008.

[11] H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, pages 266–298, 2010.

[12] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[13] C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[14] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of The 31st International Conference on Machine Learning*, pages 685–693, 2014.

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

[16] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, Mar. 2012.

[19] S. Gruber and M. J. van der Laan. tmle: An r package for targeted maximum likelihood estimation. 2011.

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

[21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[22] F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

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

[24] Y. Mansour, M. Mohri, and A. Rostamizadeh. *Domain adaptation: Learning bounds and algorithms*. 2009.

[25] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

[26] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2):199–210, 2011.

[27] J. Pearl. *Causality*. Cambridge university press, 2009.

[28] D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2011.

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

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

[31] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[32] A. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pages 2217–2225, 2010.

[33] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[34] A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

[35] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.

[36] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

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