Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit* SHALIT@CS.NYU.EDU
CIMS, New York University, New York, NY 10003

Fredrik D. Johansson* FREDRIKJ@MIT.EDU
IMES, MIT, Cambridge, MA 02142

David Sontag DSONTAG@CSAIL.MIT.EDU
CSAIL & IMES, MIT, Cambridge, MA 02139

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. We give a novel, simple and intuitive generalization-error bound showing that 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 (also known as "c-specific treatment effects" Shpitser & Pearl (2006); Pearl (2015)), rather that 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 indi-

* Equal contribution

[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).

vidual 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.

We frame our results using the Rubin-Neyman 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. In our data, for each unit we only see one of the potential outcomes, depending on the treatment assignment: if $t = 0$ we observe $y = Y_0$, 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 an individual unit with characteristics x, or the Individual Treatment Effect (ITE) [2]. For example, for a patient with features x, we can use this 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.

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

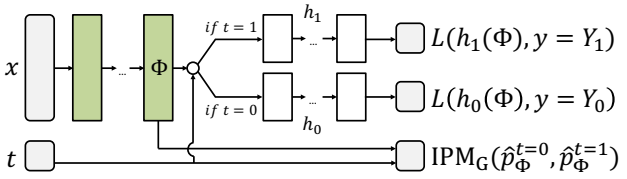


Figure 1. Neural network architecture for ITE estimation. L is a loss function, IPM_G is an integral probability metric. Note that only one of h_0 and h_1 is updated for each sample during training.

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 multilayer 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 about 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 behavior under 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 graph is simple and known to be of the form $(Y_1, Y_0) \leftarrow x \rightarrow t$, with no hidden confounders.

Under the causal model we assume, the most common goal of causal effect inference as used in the applied sciences is to obtain the average treatment effect: $ATE = \mathbb{E}_{x\sim p(x)}\left[\tau(x)\right]$. 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 adjustment, also known as back-door adjustment or the G-computation formula (Pearl, 2009; Rubin, 2011). In its basic version, covariate adjustment amounts to estimating the functions $m_1(x)$, $m_0(x)$. Therefore, covariate adjustment methods are the most natural candidates for estimating ITE as well as ATE, using the estimates of $m_t(x)$. However, 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 way to view 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 propensity score weighting (Austin, 2011) reweight the units in the observational data so as to make the treated and control populations more comparable. These methods do not yield themselves immediately to estimating an individual level effect, and adapting them for that purpose is an interesting research question. Doubly robust methods combine re-weighting the samples and covariate adjustment in clever ways to reduce model bias (Funk et al., 2011). Again, we believe that finding how to adapt the concept of double robustness to the problem of effectively estimating ITE is an interesting open question.

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 method 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.

Another line of work in the causal inference community relates to bounding the estimate of the average treatment effect given an instrumental variable (Balke & Pearl, 1997; Bareinboim & Pearl, 2012), or under hidden confounding, for example when the ignorability assumption does not hold (Pearl, 2009; Cai et al., 2008). 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 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 (2009); Pan et al. (2011); Sun et al. (2016)) could be adapted for use in ITE estimation as presented here.

Finally, our paper builds on work by Johansson et al. (2016), where the authors show a connection between covariate shift and the task of estimating the counterfactual outcome in a causal inference scenario. They proposed 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 then 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 top of the counterfactual error term. 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. 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 and denote ϵ_F. 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 $\epsilon_F(h, \Phi) = u\epsilon^{t=1}(h, \Phi) + (1 - u)\epsilon^{t=0}(h, \Phi)$.

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

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

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 bounded continuous functions, the family of 1-Lipschitz functions (Sriperumbudur et al., 2012), and the unit-ball of functions in a universal reproducing Hilbert kernel space (Gretton et al., 2012).

Definition 7. *Recall that $m_t(x) = \mathbb{E}[Y_t|x]$. The expected variance of Y_t with respect to a distribution $p(x, t)$:*

$$\sigma^2_{Y_t}(p(x,t)) = \int_{\mathcal{X}\times\mathcal{Y}} (Y_t - m_t(x))^2\, p(Y_t|x)p(x,t)\, dY_t dx.$$

We define:

$$\sigma^2_{Y_t} = \min\{\sigma^2_{Y_t}(p(x,t)), \sigma^2_{Y_t}(p(x,1-t))\},$$
$$\sigma^2_{Y} = \min\{\sigma^2_{Y_0}, \sigma^2_{Y_1}\}.$$

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 $\frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r), t) \in \text{G}$. We have:*

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

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:*

$$\epsilon_{\mathit{PEHE}}(h, \Phi) \leq$$
$$2\left(\epsilon_{CF}(h, \Phi) + \epsilon_F(h, \Phi) - 2\sigma^2_Y\right) \leq \quad (2)$$
$$2\left(\epsilon_F^{t=0}(h, \Phi) + \epsilon_F^{t=1}(h, \Phi) + B_\Phi \mathit{IPM}_{\text{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - 2\sigma^2_Y\right),$$

where ϵ_F and ϵ_{CF} are defined 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 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 will make the bound tighter. However, choosing too small a family could result in an incomputable bound. For example, for the minimal choice $\text{G} = \{\ell_{h,\Phi}(x, 0), \ell_{h,\Phi}(x, 1)\}$, 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. In this paper 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; Sriperumbudur et al., 2012), denoted $\text{Wass}(p, q)$. The second is the family of norm-1 reproducing kernel Hilbert space (RKHS) functions, leading to the MMD metric (Gretton et al., 2012; Sriperumbudur et al., 2012), denoted $\text{MMD}(p, q)$. Both the Wasserstein and MMD metrics have consistent estimators which can be efficiently computed in the finite sample case (Sriperumbudur et al., 2012). Both 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^2_Y 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 The-

orem 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 simultaneously fits both a balanced representation of the data and a hypothesis for the outcome. CFR draws on the same intuition as the approach proposed by Johansson et al. (2016), but *overcomes* the following limitations of their method: a) Their theory requires a two-step optimization procedure and is specific to *linear* hypotheses of the learned representation (and does not support e.g. deep neural networks), b) The treatment indicator might get lost 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 $(x_1, t_1, y_1), \dots (x_n, t_n, y_n)$, 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 $f(x, t) := h(\Phi(x), t)$.

In this work, we let $\Phi(x)$ and $h(\Phi, t)$ be parameterized by deep neural networks trained jointly in an end-to-end fashion, see Figure 1. This model allows for learning complex non-linear representations and hypotheses with large flexibility. Johansson et al. (2016) parameterized $h(\Phi, t)$ with a single network using the concatenation of Φ and t as input. When the dimension of Φ is high, this risks losing the influence of t on h during training. To combat this, our first contribution is to parameterize $h_1(\Phi)$ and $h_0(\Phi)$ as two separate "heads" of the joint network, the former used to estimate the outcome under treatment, and the latter under control. This means that statistical power is shared in the representation layers of the network, while the effect of treatment is retained in the separate heads. Note that each sample is used to update only the head corresponding to the observed treatment; for example, an observation $(x_i, t_i = 1, y_i)$ is only used to update h_1.

Our second contribution is to excplicitly account and 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:

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

Note that $u = p(t = 1)$ in the definition of w_i 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 defined by the function family G. For most IPMs, we cannot compute the factor B_ϕ in Equation 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).

We train our models by minimizing (3) using stochastic gradient descent, where we backpropagate the error through both the hypothesis and representation networks, 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.

5. Experiments

Evaluating causal inference algorithms is more difficult than many machine learning tasks, since for real-world data we rarely have access to the ground truth treatment effect. Existing literature mostly deals with this in two ways. One is by using synthetic or 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 in using data from RCTs is that there is no imbalance between the treated and control distributions, 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 for training, but can only use the randomized component for evaluation. This alleviates, but does not solve, the issue of a completely balanced dataset being unsuited for our method.

We evaluate our framework CFR, and its variant without

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 η with standard neural net methods e.g. Adam (Kingma & Ba, 2014)
  $[\mathbf{W}, \mathbf{V}] \leftarrow [\mathbf{W} - \eta(\alpha g_1 + g_3), \mathbf{V} - \eta(g_2 + 2\lambda \mathbf{V})]$
  Check convergence criterion
end while

balancing regularization (TARNet), in the task of estimating ITE and ATE. CFR is implemented as a feed-forward neural network with 3 fully-connected exponential-linear layers 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, 2014). For an overview, see Figure 1. Layers corresponding to the hypothesis 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 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, which is a doubly robust method (TMLE) (Gruber & van der Laan, 2011), Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Chipman & McCulloch, 2016), Random Forests (Rand. For.) (Breiman, 2001), Causal Forests (Caus. For.) (Wager & Athey, 2015) as well as the Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) by 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 selection procedure, applied to all methods, in subsection C.1 of the supplement.

We evaluate our model in two different settings. 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 the *out-of-sample* setting, where the goal is to estimate ITE for units with no observed outcomes. This corresponds to the case where a new patient arrives and the goal is to select the best possible treatment. Within-sample error is computed over both the training and validation sets, and out-of-sample error over the test set.

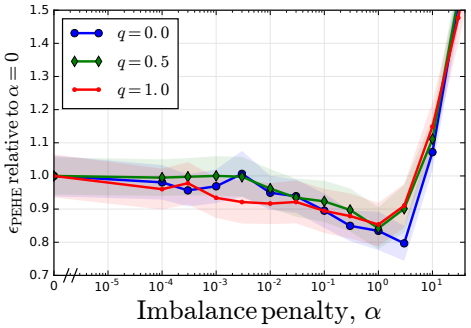


Figure 2. Out-of-sample ITE error versus IPM regularization for CFR Wass, relative to the error at $\alpha = 0$, on 500 realizations of IHDP, with high ($q = 1$), medium and low (artificial) imbalance between control and treated.

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 future 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.

We 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

Table 1. Results on IHDP (left) and Jobs (right). MMD is squared linear MMD. Lower is better.

Within-sample

	IHDP		JOBS	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}
OLS/LR-1	$5.8 \pm .3$	$.73 \pm .04$	$.22 \pm .0$	$.01 \pm .00$
OLS/LR-2	$2.4 \pm .1$	$.14 \pm .01$	$.21 \pm .0$	$.01 \pm .01$
BLR	$5.8 \pm .3$	$.72 \pm .04$	$.22 \pm .0$	$.01 \pm .01$
k-NN	$2.1 \pm .1$	$.14 \pm .01$	$.02 \pm .0$	$.21 \pm .01$
TMLE	$5.0 \pm .2$	$.30 \pm .01$	$.22 \pm .0$	$.02 \pm .01$
BART	$2.1 \pm .1$	$.23 \pm .01$	$.23 \pm .0$	$.02 \pm .00$
RAND.FOR.	$4.2 \pm .2$	$.73 \pm .05$	$.23 \pm .0$	$.03 \pm .01$
CAUS.FOR.	$3.8 \pm .2$	$.18 \pm .01$	$.19 \pm .0$	$.03 \pm .01$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$.20 \pm .0$	$.04 \pm .01$
TARNET	$.88 \pm .0$	$.26 \pm .01$	$.17 \pm .0$	$.05 \pm .02$
CFR MMD	$.73 \pm .0$	$.30 \pm .01$	$.18 \pm .0$	$.04 \pm .01$
CFR WASS	$.71 \pm .0$	$.25 \pm .01$	$.17 \pm .0$	$.04 \pm .01$

Out-of-sample

	IHDP		JOBS	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}
OLS/LR-1	$5.8 \pm .3$	$.94 \pm .06$	$.23 \pm .0$	$.08 \pm .04$
OLS/LR-2	$2.5 \pm .1$	$.31 \pm .02$	$.24 \pm .0$	$.08 \pm .03$
BLR	$5.8 \pm .3$	$.93 \pm .05$	$.25 \pm .0$	$.08 \pm .03$
k-NN	$4.1 \pm .2$	$.79 \pm .05$	$.26 \pm .0$	$.13 \pm .05$
BART	$2.3 \pm .1$	$.34 \pm .02$	$.25 \pm .0$	$.08 \pm .03$
RAND.FOR.	$6.6 \pm .3$	$.96 \pm .06$	$.28 \pm .0$	$.09 \pm .04$
CAUS.FOR.	$3.8 \pm .2$	$.40 \pm .03$	$.20 \pm .0$	$.07 \pm .03$
BNN	$2.1 \pm .1$	$.42 \pm .03$	$.24 \pm .0$	$.09 \pm .04$
TARNET	$.95 \pm .0$	$.28 \pm .01$	$.21 \pm .0$	$.11 \pm .04$
CFR MMD	$.78 \pm .0$	$.31 \pm .01$	$.21 \pm .0$	$.08 \pm .03$
CFR WASS	$.76 \pm .0$	$.27 \pm .01$	$.21 \pm .0$	$.09 \pm .03$

$1-q$, 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 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 $\text{ATT} = |T|^{-1} \sum_{i \in T} y_i - |C \cap E|^{-1} \sum_{i \in C \cap E} y_i$, 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 $\lambda = 0$.

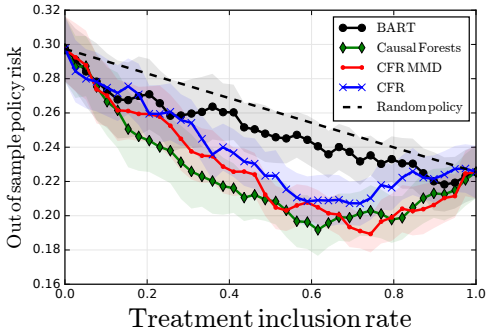


Figure 3. Policy risk on Jobs as a function of treatment inclusion rate. Lower is better. Subjects are included in treatment in order of their estimated treatment effect given by the various methods. CFR Wass is similar to CFR and is omitted to avoid clutter.

5.3. Results

We begin by noting that indeed imbalance confers an advantage to using the IPM regularization term, as our theoretical results indicate, see e.g. the results for CFR Wass ($\alpha > 0$) and TARNet ($\alpha = 0$) 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, we see a smaller gain from using IPM penalties than on IHDP. We believe this is the case because, while we are minimizing our bound over observational data and accounting for this bias, we are evaluating the predictions only on a randomized subset, where the treatment groups are distributed identically. For both IHDP, non-linear estimators do 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 the problem of estimating individual treatment effect. Our bound relates ITE estimation to the classic machine learning problem of learning from finite samples, along with methods for measuring distributional distances from finite 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 how to integrate our work with more complicated causal models such as those with hidden confounding or instrumental variables.

ACKNOWLEDGMENTS

We wish to thank Aahlad Manas for his assistance with the experiments. We also thank Jennifer Hill, Marco Cuturi, Esteban Tabak and Sanjong Misra 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

MathOverflow: functions with orthogonal Jacobian. https://mathoverflow.net/questions/228964/functions-with-orthogonal-jacobian. Accessed: 2016-05-05.

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.

Aude, Genevay, Cuturi, Marco, Peyré, Gabriel, and Bach, Francis. Stochastic optimization for large-scale optimal transport. *arXiv preprint arXiv:1605.08527*, 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.

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

Bareinboim, Elias and Pearl, Judea. Controlling selection bias in causal inference. In *AISTATS*, pp. 100–108, 2012.

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.

Ben-Israel, Adi. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.

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.

Cai, Zhihong, Kuroki, Manabu, Pearl, Judea, and Tian, Jin. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701, 2008.

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.

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. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

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. *arXiv preprint arXiv:0907.1815*, 2009.

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.

Grunewalder, Steffen, Arthur, Gretton, and Shawe-Taylor, John. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1184–1192, 2013.

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. *arXiv preprint arXiv:1412.6980*, 2014.

Kuang, Max and Tabak, Esteban. Preconditioning of optimal transport. *Preprint*, 2016.

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.

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.

Steinwart, Ingo and Christmann, Andreas. *Support vector machines*. Springer Science & Business Media, 2008.

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.

A. Proofs

A.1. Definitions, assumptions, and auxiliary lemmas

We first define the necessary distributions and prove some simple results about them. We assume a joint distribution function $p(x, t, Y_0, Y_1)$, such that $(Y_1, Y_0) \perp\!\!\!\perp t|x$, and $0 < p(t = 1|x) < 1$ for all x. Recall that we assume *Consistency*, that is we assume that we observe $y = Y_1|(t = 1)$ and $y = Y_0|(t = 0)$.

Definition A1. *The treatment effect for unit x is:*

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

We first show that under consistency and strong ignorability, the ITE function $\tau(x)$ is identifiable:

Lemma A1. *We have:*

$$\begin{aligned}
&\mathbb{E}\left[Y_1 - Y_0|x\right] = \\
&\mathbb{E}\left[Y_1|x\right] - \mathbb{E}\left[Y_0|x\right] = &(4)\\
&\mathbb{E}\left[Y_1|x, t = 1\right] - \mathbb{E}\left[Y_0|x, t = 0\right] = &(5)\\
&\mathbb{E}\left[y|x, t = 1\right] - \mathbb{E}\left[y|x, t = 0\right].
\end{aligned}$$

Equality (4) is because we assume that Y_t and t are independent conditioned on x. Equality (5) follows from the consistency assumption. Finally, the last equation is composed entirely of observable quantities and can be estimated from data since we assume $0 < p(t = 1|x) < 1$ for all x.

Definition A2. *Let $p^{t=1}(x) := p(x|t = 1)$, and $p^{t=0}(x) := p(x|t = 0)$ denote respectively the treatment and control distributions.*

Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a *representation function*. We will assume that Φ is differentiable.

Assumption A1. *The representation function Φ is one-to-one. Without loss of generality we will assume that $\mathcal{R}$ is the image of $\mathcal{X}$ under Φ, and define $\Psi : \mathcal{R} \to \mathcal{X}$ to be the inverse of Φ, such that $\Psi(\Phi(x)) = x$ for all $x \in \mathcal{X}$.*

Definition A3. *For a representation function $\Phi : \mathcal{X} \to \mathcal{R}$, and for a distribution p defined over $\mathcal{X}$, let p_Φ be the distribution induced by Φ over $\mathcal{R}$. Define $p_\Phi^{t=1}(r) := p_\Phi(r|t = 1)$, $p_\Phi^{t=0}(r) := p_\Phi(r|t = 0)$, to be the treatment and control distributions induced over $\mathcal{R}$.*

For a one-to-one Φ, the distribution p_Φ over $\mathcal{R} \times \{0, 1\}$ can be obtained by the standard change of variables formula, using the determinant of the Jacobian of $\Psi(r)$. See (Ben-Israel, 1999) for the case of a mapping Φ between spaces of different dimensions.

Lemma A2. *For all $r \in \mathcal{R}$, $t \in \{0, 1\}$:*

$$\begin{aligned}
p_\Phi(t|r) &= p(t|\Psi(r))\\
p(Y_t|r) &= p(Y_t|\Psi(r)).
\end{aligned}$$

Notation:
$p(x,t)$: distribution on $\mathcal{X} \times \{0,1\}$
$u = p(t=1)$: the marginal probability of treatment.
$p^{t=1}(x) = p(x|t=1)$: treated distribution. $p^{t=0}(x) = p(x|t=0)$: control distribution.
Φ: representation function mapping from $\mathcal{X}$ to $\mathcal{R}$.
Ψ: the inverse function of Φ, mapping from $\mathcal{R}$ to $\mathcal{X}$.
$p_\Phi(r,t)$: the distribution induced by Φ on $\mathcal{R} \times \{0,1\}$.
$p_\Phi^{t=1}(r)$, $p_\Phi^{t=0}(r)$: treated and control distributions induced by Φ on $\mathcal{R}$.
$L(\cdot,\cdot)$: loss function, from $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R}_+$.
$\ell_{h,\Phi}(x,t)$: the expected loss of $h(\Phi(x),t)$ for the unit x and treatment t.
$\epsilon_F(h,\Phi)$, $\epsilon_{CF}(h,\Phi)$: expected factual and counterfactual loss of $h(\Phi(x),t)$.
$\tau(x) := \mathbb{E}[Y_1 - Y_0|x]$, the expected treatment effect for unit x.
$\epsilon_{\text{PEHE}}(f)$: expected error in estimating the individual treatment effect of a function $f(x,t)$.
$\text{IPM}_\text{G}(p,q)$: the integral probability metric distance induced by function family G between distributions p and q.

Proof. Let $J_\Psi(r)$ be the absolute of the determinant of the Jacobian of $\Psi(r)$.

$$p_\Phi(t|r) = \frac{p_\Phi(t,r)}{p_\Phi(r)} \overset{(a)}{=} \frac{p(t,\Psi(r))J_\Psi(r)}{p(\Psi(r))J_\Psi(r)} = \frac{p(t,\Psi(r))}{p(\Psi(r))} = p(t|\Psi(r)),$$

where equality (a) is by the change of variable formula. The proof is identical for $p(Y_t|r)$. □

Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function, e.g. the absolute loss or squared loss.

Definition A4. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a representation function. Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be an hypothesis defined over the representation space $\mathcal{R}$. The expected loss for the unit and treatment pair (x,t) is:*

$$\ell_{h,\Phi}(x,t) = \int_\mathcal{Y} L(Y_t, h(\Phi(x),t)) p(Y_t|x) dY_t$$

Definition A5. *The expected factual loss and counterfactual losses of h and Φ are, respectively:*

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

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

When it is clear from the context, we will sometimes use $\epsilon_F(f)$ and $\epsilon_{CF}(f)$ for the expected factual and counterfactual losses of an arbitrary function $f : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$.

Definition A6. *The expected* treated *and* control *losses are:*

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

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

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

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

The four losses above are simply the loss conditioned on either the control or treated set. Let $u := p(t=1)$ be the proportion of treated in the population. We then have the immediate result:

Lemma A3.

$$\epsilon_F(h,\Phi) = u \cdot \epsilon_F^{t=1}(h,\Phi) + (1-u) \cdot \epsilon_F^{t=0}(h,\Phi)$$

$$\epsilon_{CF}(h,\Phi) = (1-u) \cdot \epsilon_{CF}^{t=1}(h,\Phi) + u \cdot \epsilon_{CF}^{t=0}(h,\Phi).$$

The proof is immediate, noting that $p(x,t) = u \cdot p^{t=1}(x) + (1-u) \cdot (x)$, and from the Definitions A4 and A6 of the losses.

Definition A7. *Let* G *be a function family consisting of functions $g : \mathcal{S} \to \mathbb{R}$. For a pair of distributions p_1, p_2 over $\mathcal{S}$, define the* Integral Probability Metric*:*

$$\mathit{IPM}_\text{G}(p_1,p_2) = \sup_{g\in\text{G}} \left| \int_\mathcal{S} g(s)\,(p_1(s) - p_2(s))\, ds \right|$$

$\text{IPM}_\text{G}(\cdot,\cdot)$ defines a pseudo-metric on the space of probability functions over $\mathcal{S}$, and for sufficiently large function families, $\text{IPM}_\text{G}(\cdot,\cdot)$ is a proper metric (Müller, 1997). Examples of sufficiently large functions families includes the set of bounded continuous functions, the set of 1-Lipschitz functions, and the set of unit norm functions in a universal Reproducing Norm Hilbert Space. The latter two give rise to the Wasserstein and Maximum Mean Discrepancy metrics, respectively (Gretton et al., 2012; Sriperumbudur et al., 2012). We note that for function families G such as the three mentioned above, for which $g \in \text{G} \implies -g \in \text{G}$, the absolute value can be omitted from definition A7.

A.2. General IPM bound

We now state and prove the most important technical lemma of this section.

Lemma A4 (Lemma 1, main text)**.** *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation with Ψ its inverse. Let $p_\Phi^{t=1}, p_\Phi^{t=0}$ be defined as in Definition A3. Let $u = p(t = 1)$. Let* G *be a family of functions $g : \mathcal{R} \to \mathbb{R}$, and denote by $IPM_{\mathrm{G}}(\cdot,\cdot)$ the integral probability metric induced by* G*. Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be an hypothesis. Assume there exists a constant $B_\Phi > 0$, such that for $t = 0, 1$, the function $g_{\Phi,h}(r,t) := \frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. Then we have:*

$$\begin{aligned}&\epsilon_{CF}(h,\Phi) \leq \\ &(1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + \\ &B_\Phi \cdot IPM_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right). \end{aligned} \tag{6}$$

Proof.

$$\epsilon_{CF}(h,\Phi) - \left[(1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi)\right] =$$
$$\left[(1-u)\cdot\epsilon_{CF}^{t=1}(h,\Phi) + u\cdot\epsilon_{CF}^{t=0}(h,\Phi)\right] -$$
$$\left[(1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi)\right] =$$
$$(1-u)\cdot\left[\epsilon_{CF}^{t=1}(h,\Phi) - \epsilon_F^{t=1}(h,\Phi)\right] +$$
$$u\cdot\left[\epsilon_{CF}^{t=0}(h,\Phi) - \epsilon_F^{t=0}(h,\Phi)\right] = \tag{7}$$
$$(1-u)\int_{\mathcal{X}} \ell_{h,\Phi}(x,1)\left(p^{t=0}(x) - p^{t=1}(x)\right)dx +$$
$$u\int_{\mathcal{X}} \ell_{h,\Phi}(x,0)\left(p^{t=1}(x) - p^{t=0}(x)\right)dx = \tag{8}$$
$$(1-u)\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(r),1)\left(p_\Phi^{t=0}(r) - p_\Phi^{t=1}(r)\right)dr +$$
$$u\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(r),0)\left(p_\Phi^{t=1}(r) - p_\Phi^{t=0}(r)\right)dr =$$
$$B_\Phi\cdot(1-u)\int_{\mathcal{R}} \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),1)\left(p_\Phi^{t=0}(r) - p_\Phi^{t=1}(r)\right)dr +$$
$$B_\Phi\cdot u\int_{\mathcal{R}} \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),0)\left(p_\Phi^{t=1}(r) - p_\Phi^{t=0}(r)\right)dr \leq \tag{9}$$
$$B_\Phi\cdot(1-u)\sup_{g\in\mathrm{G}}\left|\int_{\mathcal{R}} g(r)\left(p_\Phi^{t=0}(r) - p_\Phi^{t=1}(r)\right)dr\right| +$$
$$B_\Phi\cdot u\sup_{g\in\mathrm{G}}\left|\int_{\mathcal{R}} g(r)\left(p_\Phi^{t=1}(r) - p_\Phi^{t=0}(r)\right)dr\right| = \tag{10}$$
$$B_\Phi\cdot \mathrm{IPM}_{\mathrm{G}}(p_\Phi^{t=0}, p_\Phi^{t=1}). \tag{11}$$

Equality (7) is by Definition A6 of the treated and control loss, equality (8) is by the change of variables formula and Definition A3 of $p_\Phi^{t=1}$ and $p_\Phi^{t=0}$, inequality (9) is by the premise that $\frac{1}{B_\Phi}\cdot\ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$ for $t = 0, 1$, and (10) is by Definition A7 of an IPM. □

The essential point in the proof of Lemma A4 is inequality 9. Note that on the l.h.s. of the inequality, we need to evaluate the expectations of $\ell_{h,\Phi}(\Psi(r),0)$ over $p_\Phi^{t=1}$ and $\ell_{h,\Phi}(\Psi(r),1)$ over $p_\Phi^{t=0}$. Both of these expectations are in general unavailable, since they require us to evaluate treatment outcomes on the control, and control outcomes on the treated. We therefore upper bound these unknowable quantities by taking a supremum over a function family which includes $\ell_{h,\Phi}(\Psi(r),0)$ and $\ell_{h,\Phi}(\Psi(r),1)$. The upper bound ignores most of the details of the outcome, and amounts to measuring a distance between two distributions we have samples from: the control and treated distribution. Note that for a randomized trial (i.e. when $t \perp\!\!\!\perp x$) with we have that $\mathrm{IPM}(p_\Phi^{t=1}, p_\Phi^{t=0}) = 0$. Indeed, it is straightforward to show that in that case we actually have an equality: $\epsilon_{CF}(h,\Phi) = (1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi)$.

The crucial condition in Lemma A4 is that the function $g_{\Phi,h}(r) := \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t)$ is in G. In subsections A.3 and A.4 below we look into two specific function families G, and evaluate what does this inclusion condition entail, and in particular we will derive specific bounds for B_Φ.

Definition A8. *For $t = 0, 1$ define:*

$$m_t(x) := \mathbb{E}\left[Y_t|x\right].$$

Obviously for the treatment effect $\tau(x)$ we have $\tau(x) = m_1(x) - m_0(x)$.

Let $f : \mathcal{X}\times\{0,1\} \to \mathcal{Y}$ by an hypothesis, such that $f(x,t) = h(\Phi(x),t)$ for a representation Φ and hypothesis h defined over the output of Φ.

Definition A9. *The treatment effect estimate for unit x is:*

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

Definition A10. *The expected Precision in Estimation of Heterogeneous Effect (PEHE) loss of g is:*

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

Definition A11. *The expected variance of Y_t with respect to a distribution $p(x,t)$:*

$$\sigma^2_{Y_t}(p(x,t)) = \int_{\mathcal{X}\times\mathcal{Y}} \left(Y_t - m_t(x)\right)^2 p(Y_t|x)p(x,t)\,dY_t dx.$$

We define:

$$\sigma^2_{Y_t} = \min\{\sigma^2_{Y_t}(p(x,t)), \sigma^2_{Y_t}(p(x,1-t))\},$$
$$\sigma^2_Y = \min\{\sigma^2_{Y_0}, \sigma^2_{Y_1}\}.$$

If Y_t are deterministic functions of x, then $\sigma^2_Y = 0$.

We now show that $\epsilon_{\mathrm{PEHE}}(f)$ is upper bounded by $2\epsilon_F + 2\epsilon_{CF} - 2\sigma^2_Y$ where ϵ_F and ϵ_{CF} are w.r.t. to the squared loss. An analogous result can be obtained for the absolute loss, using mean absolute deviation.

Lemma A5. *For any function $f : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$, and distribution $p(x,t)$ over $\mathcal{X} \times \{0,1\}$:*

$$\int_{\mathcal{X}} \left(f(x,t) - m_t(x)\right)^2 p(x,t)\, dxdt = \epsilon_F(f) - \sigma^2_{Y_t}(p(x,t)),$$
$$\int_{\mathcal{X}} \left(f(x,t) - m_t(x)\right)^2 p(x,1-t)\, dxdt = \epsilon_{CF}(f) - \sigma^2_{Y_t}(p(x,1-t)),$$

where $\epsilon_F(f)$ and $\epsilon_{CF}(f)$ are w.r.t. to the squared loss.

Proof. For simplicity we will prove for $p(x,t)$ and $\epsilon_F(f)$. The proof for $p(x,1-t)$ and ϵ_{CF} is identical.

$$\epsilon_F(f) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f(x,t) - Y_t\right)^2 p(Y_t|x)p(x,t)\, dY_t dxdt =$$
$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f(x,t) - m_t(x)\right)^2 p(Y_t|x)p(x,t)\, dY_t dxdt +$$
$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(m_t(x) - Y_t\right)^2 p(Y_t|x)p(x,t)\, dY_t dxdt + \quad (12)$$
$$\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f(x,t) - m_t(x)\right)\left(m_t(x) - Y_t\right) p(Y_t|x)p(x,t)\, dY_t dxdt = \quad (13)$$

$$\int_{\mathcal{X}\times\{0,1\}} \left(f(x,t) - m_t(x)\right)^2 p(x,t)\, dxdt + \sigma^2_{Y_0}(p(x,t)) + \sigma^2_{Y_1}(p(x,t)) + 0,$$

where the equality (13) is by the Definition A11 of $\sigma^2_{Y_t}(p)$, and because the integral in (12) evaluates to zero, since $m_t(x) = \int_{\mathcal{X}} Y_t p(Y_t|x)\, dx$. □

Theorem 1. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function, with inverse Ψ. Let $p^{t=1}_\Phi, p^{t=0}_\Phi$ be defined as in Definition A3. Let $u = p(t=1)$. Let G be a family of functions $g : \mathcal{R} \to \mathbb{R}$, and denote by IPM$_\mathrm{G}(\cdot,\cdot)$ the integral probability metric induced by G. Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be an hypothesis. Let the loss $L(y_1,y_2) = (y_1-y_2)^2$. Assume there exists a constant $B_\Phi > 0$, such that for $t \in \{0,1\}$, the functions $g_{\Phi,h}(r,t) := \frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. We then have:*

$$\epsilon_{PEHE}(h,\Phi) \leq 2\left(\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) - 2\sigma^2_Y\right) \leq 2\left(\epsilon^{t=0}_F(h,\Phi) + \epsilon^{t=1}_F(h,\Phi) + B_\Phi IPM_\mathrm{G}\left(p^{t=1}_\Phi, p^{t=0}_\Phi\right) - 2\sigma^2_Y\right),$$

where ϵ_F and ϵ_{CF} are with respect to the squared loss.

Proof. We will prove the first inequality, $\epsilon_{\text{PEHE}}(f) \leq 2\epsilon_{CF}(h,\Phi) + 2\epsilon_F(h,\Phi) - 2\sigma^2_Y$. The second inequality is then immediate by Lemma A4. Recall that we denote $\epsilon_{\text{PEHE}}(f) = \epsilon_{\text{PEHE}}(h,\Phi)$ for $f(x,t) = h(\Phi(x),t)$.

$$\epsilon_{\text{PEHE}}(f) = \int_{\mathcal{X}} \left(\left(f(x,1) - f(x,0)\right) - \left(m_1(x) - m_0(x)\right)\right)^2 p(x)\, dx =$$
$$\int_{\mathcal{X}} \left(\left(f(x,1) - m_1(x)\right) + \left(m_0(x) - f(x,0)\right)\right)^2 p(x)\, dx \leq \quad (14)$$
$$2\int_{\mathcal{X}} \left(\left(f(x,1) - m_1(x)\right)^2 + \left(m_0(x) - f(x,0)\right)^2\right) p(x)\, dx = \quad (15)$$
$$2\int_{\mathcal{X}} \left(f(x,1) - m_1(x)\right)^2 p(x,t=1)\, dx +$$
$$2\int_{\mathcal{X}} \left(m_0(x) - f(x,0)\right)^2 p(x,t=0)\, dx +$$
$$2\int_{\mathcal{X}} \left(f(x,1) - m_1(x)\right)^2 p(x,t=0)\, dx +$$
$$2\int_{\mathcal{X}} \left(m_0(x) - f(x,0)\right)^2 p(x,t=1)\, dx =$$
$$2\int_{\mathcal{X}} \left(f(x,t) - m_t(x)\right)^2 p(x,t)\, dxdt +$$
$$2\int_{\mathcal{X}} \left(f(x,t) - m_t(x)\right)^2 p(x,1-t)\, dxdt \leq \quad (16)$$
$$2(\epsilon_F - \sigma^2_Y) + 2(\epsilon_{CF} - \sigma^2_Y).$$

where (14) is because $(x+y)^2 \leq 2(x^2+y^2)$, (15) is because $p(x) = p(x,t=0) + p(x,t=1)$ and (16) is by Lemma A5 and Definition A5 of the losses ϵ_F, ϵ_{CF} and Definition A11 of σ^2_Y. Having established the first inequality in the Theorem statement, we now show the second. We have by Lemma A4 that:

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\epsilon^{t=1}_F(h,\Phi) + u\epsilon^{t=0}_F(h,\Phi) + B_\Phi \cdot \text{IPM}_\mathrm{G}\left(p^{t=1}_\Phi, p^{t=0}_\Phi\right).$$

We further have by Lemma A3 that:

$$\epsilon_F(h,\Phi) = u\epsilon^{t=1}_F(h,\Phi) + (1-u)\epsilon^{t=0}_F(h,\Phi).$$

Therefore

$$\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) \leq \epsilon^{t=1}_F(h,\Phi) + \epsilon^{t=0}_F(h,\Phi) + B_\Phi \text{IPM}_\mathrm{G}\left(p^{t=1}_\Phi, p^{t=0}_\Phi\right).$$

□

The upper bound is in terms of the standard generalization error on the treated and control distributions separately. Note that in some cases we might have very different sample sizes for treated and control, and that will show up in the finite sample bounds of these generalization errors.

We also note that the upper bound can be easily adapted to the case of the absolute loss PEHE $|\hat{\tau}(x) - \tau(x)|$. In that case the upper bound in the Theorem will have a factor 1 instead of the 2 stated above, and the standard deviation σ_Y^2 replaced by mean absolute deviation. The proof is straightforward where one simply applies the triangle inequality in inequality (14).

We will now give specific upper bounds for the constant B_Φ in Theorem 1, using two function families G in the IPM: the family of 1-Lipschitz functions, and the family of 1-norm reproducing kernel Hilbert space functions. Each one will have different assumptions about the distribution $p(x, t, Y_0, Y_1)$ and about the representation Φ and hypothesis h.

A.3. The family of 1-Lipschitz functions

For $\mathcal{S} \subset \mathbb{R}^d$, a function $f : \mathcal{S} \to \mathbb{R}$ has Lipschitz constant K if for all $x, y \in \mathcal{S}$, $|f(x) - f(y)| \leq K\|x - y\|$. If f is differentiable, then a sufficient condition for K-Lipschitz constant is if $\|\frac{\partial f}{\partial s}\| \leq K$ for all $s \in \mathcal{S}$.

For simplicity's sake we assume throughout this subsection that the true labeling functions the densities $p(Y_t|x)$ and the loss L are differentiable. However, this assumption could be relaxed to a mere Lipschitzness assumption.

Assumption A2. *There exists a constant $K > 0$ such that for all $x \in \mathcal{X}$, $t \in \{0, 1\}$, $\|\frac{p(Y_t|x)}{\partial x}\| \leq K$.*

Assumption A2 entails that each of the potential outcomes change smoothly as a function of the covariates (context) x.

Assumption A3. *The loss function L is differentiable, and there exists a constant $K_L > 0$ such that $\left|\frac{dL(y_1,y_2)}{dy_i}\right| \leq K_L$ for $i = 1, 2$. Additionally, there exists a constant M such that for all $y_2 \in \mathcal{Y}$, $M \geq \int_{\mathcal{Y}} L(y_1, y_2)\, dy_1$.*

Assuming $\mathcal{Y}$ is compact, loss functions which obey Assumption A3 include the log-loss, hinge-loss, absolute loss, and the squared loss.

When we let G in Definition A7 be the family of 1-Lipschitz functions, we obtain the so-called 1-*Wasserstein* distance between distributions, which we denote $\text{Wass}(\cdot, \cdot)$. It is well known that $\text{Wass}(\cdot, \cdot)$ is indeed a metric between distributions (Villani, 2008).

Definition A12. *Let $\frac{\partial \Phi(x)}{\partial x}$ be the Jacobian matrix of Φ at point x, i.e. the matrix of the partial derivatives of Φ. Let $\sigma_{max}(A)$ and $\sigma_{min}(A)$ denote respectively the largest and smallest singular values of a matrix A. Define $\rho(\Phi) = \sup_{x \in \mathcal{X}} \sigma_{max}\left(\frac{\partial \Phi(x)}{\partial x}\right) / \sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)$.*

It is an immediate result that $\rho(\Phi) \geq 1$.

Definition A13. *We will call a representation function $\Phi : \mathcal{X} \to \mathcal{R}$ Jacobian-normalized if $\sup_{x \in \mathcal{X}} \sigma_{max}\left(\frac{\partial \Phi(x)}{\partial x}\right) = 1$.*

Note that any non-constant representation function Φ can be Jacobian-normalized by a simple scalar multiplication.

Lemma A6. *Assume that Φ is a Jacobian-normalized representation, and let Ψ be its inverse. For $t = 0, 1$, the Lipschitz constant of $p(Y_t|\Psi(r))$ is bounded by $\rho(\Phi)K$, where K is from Assumption A2, and $\rho(\Phi)$ as in Definition A12.*

Proof. Let $\Psi : \mathcal{R} \to \mathcal{X}$ be the inverse of Φ, which exists by the assumption that Φ is one-to-one. Let $\frac{\partial \Phi(x)}{\partial x}$ be the Jacobian matrix of Φ evaluated at x, and similarly let $\frac{\partial \Psi(r)}{\partial r}$ be the Jacobian matrix of Ψ evaluated at r. Note that $\frac{\partial \Psi(r)}{\partial r} \cdot \frac{\partial \Phi(x)}{\partial x} = I$ for $r = \Phi(x)$, since $\Psi \circ \Phi$ is the identity function on $\mathcal{X}$. Therefore for any $r \in \mathcal{R}$ and $x = \Psi(r)$:

$$\sigma_{max}\left(\frac{\partial \Psi(r)}{\partial r}\right) = \frac{1}{\sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)}, \tag{17}$$

where $\sigma_{max}(A)$ and $\sigma_{min}(A)$ are respectively the largest and smallest singular values of the matrix A, i.e. $\sigma_{max}(A)$ is the spectral norm of A.

For $x = \Psi(r)$ and $t \in \{0, 1\}$, we have by the chain rule:

$$\|\frac{\partial p(Y_t|\Psi(r))}{\partial r}\| = \|\frac{\partial p(Y_t|\Psi(r))}{\partial \Psi(r)} \frac{\partial \Psi(r)}{\partial r}\| \leq \tag{18}$$

$$\|\frac{\partial \Psi(r)}{\partial r}\| \|\frac{\partial p(Y_t|\Psi(r))}{\partial \Psi(r)}\| = \tag{19}$$

$$\frac{1}{\sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)} \|\frac{\partial p(Y_t|x)}{\partial x}\| \leq \tag{20}$$

$$\frac{K}{\sigma_{min}\left(\frac{\partial \Phi(x)}{\partial x}\right)} \leq \rho(\Phi)K, \tag{21}$$

where inequality (18) is by the matrix norm inequality, equality (19) is by (17), inequality (20) is by assumption A2 on the norms of the gradient of $p(Y_t|x)$ w.r.t x , and inequality (21) is by Definition A12 of $\rho(\Phi)$, the assumption that Φ is Jacobian-normalized, and noting that singular values are necessarily non-negative.

□

Lemma A7. *Under the conditions of Lemma A4, further assume that for $t = 0, 1$, $p(Y_t|x)$ has gradients bounded by K as in A2, that h has bounded gradient norm bK, that the loss L has bounded gradient norm K_L, and that Φ is Jacobian-normalized. Then the Lipschitz constant of $\ell_{h,\Phi}(\Psi(r), t)$ is upper bounded by $K_L \cdot K\,(M\rho(\Phi) + b)$ for $t = 0, 1$.*

Proof. Using the chain rule, we have that:

$$\|\frac{\partial \ell_{h,\Phi}(\Psi(r),t)}{\partial r}\| = \|\frac{\partial}{\partial r}\int_{\mathcal{Y}} L(Y_t,h(r,t))p(Y_t|r)dY_t\| =$$
$$\|\int_{\mathcal{Y}} \frac{\partial}{\partial r}\left[L(Y_t,h(r,t))p(Y_t|r)\right]\, dY_t\| =$$
$$\|\int_{\mathcal{Y}} p(Y_t|r)\frac{\partial}{\partial r}L(Y_t,h(r,t))+L(Y_t,h(r,t))\frac{\partial}{\partial r}p(Y_t|r)dY_t\| \leq$$
$$\int_{\mathcal{Y}} p(Y_t|r)\|\frac{\partial}{\partial r}L(Y_t,h(r,t))\|\, dY_t+$$
$$\int_{\mathcal{Y}} L(Y_t,h(r,t))\frac{\partial}{\partial r}p(Y_t|r)\, dY_t \leq \tag{22}$$
$$\int_{\mathcal{Y}} p(Y_t|r)\|\frac{\partial L(Y_t,h(r,t))}{\partial h(r,t)}\frac{\partial h(r,t)}{\partial r}\|\, dY_t+$$
$$\int_{\mathcal{Y}} L(Y_t,h(r,t))\frac{\partial}{\partial r}p(Y_t|r)\, dY_t \leq \tag{23}$$
$$\int_{\mathcal{Y}} p(Y_t|r)K_L\cdot b\cdot K + M\cdot\rho(\Phi)\cdot K, \tag{24}$$

where inequality 22 is due to Assumption A3 and inequality 23 is due to Lemma A6. □

Lemma A8. *Let $u = p(t=1)$ be the marginal probability of treatment, and assume $0 < u < 1$. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one, Jacobian-normalized representation function. Let K be the Lipschitz constant of the functions $p(Y_t|x)$ on $\mathcal{X}$. Let K_L be the Lipschitz constant of the loss function L, and M be as in Assumption A3. Let $h : \mathcal{R}\times\{0,1\} \to \mathbb{R}$ be an hypothesis with Lipschitz constant bK. Then:*

$$\epsilon_{CF}(h,\Phi) \leq$$
$$(1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi)+$$
$$2\left(M\rho(\Phi)+b\right)\cdot K\cdot K_L\cdot \mathit{Wass}(p_\Phi^{t=1}, p_\Phi^{t=0}). \tag{25}$$

Proof. We will apply Lemma A4 with $\mathrm{G} = \{g : \mathcal{R}\to\mathbb{R}$ s.t. f is 1-Lipschitz$\}$. By Lemma A7, we have that for $B_\Phi = (M\rho(\Phi)+b)\cdot K\cdot K_L$, the function $\frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. Inequality (25) then holds as a special case of Lemma A4. □

Theorem 2. *Under the assumptions of Lemma A8, using the squared loss for ϵ_F, we have:*

$$\epsilon_{\mathit{PEHE}}(h,\Phi) \leq$$
$$2\epsilon_F^{t=0}(h,\Phi) + 2\epsilon_F^{t=1}(h,\Phi) - 4\sigma_Y^2+$$
$$2\left(M\rho(\Phi)+b\right)\cdot K\cdot K_L\cdot \mathit{Wass}(p_\Phi^{t=1}, p_\Phi^{t=0}).$$

Proof. Plug in the upper bound of Lemma A8 into the upper bound of Theorem 1. □

We examine the constant $(M\rho(\Phi)+b)\cdot K\cdot K_L$ in Theorem A8. K, the Lipschitz constant of m_0 and m_1, is not under our control and measures an aspect of the complexity of the true underlying functions we wish to approximate. The terms K_L and M depend on our choice of loss function and the size of the space $\mathcal{Y}$. The term b comes from our assumption that the hypothesis h has norm bK. Note that smaller b, while reducing the bound, might force the factual loss term $\epsilon_F(h,\Phi)$ to be larger since a small b implies a less flexible h. Finally, consider the term $\rho(\Phi)$. The assumption that Φ is normalized is rather natural, as we do not expect a certain scale from a representation. Furthermore, below we show that in fact the Wasserstein distance is positively homogeneous with respect to the representation Φ. Therefore, in Lemma A8, we can indeed assume that Φ is normalized. The specific choice of *Jacobian-normalized* scaling yields what is in our opinion a more interpretable result in terms of the inverse condition number $\rho(\Phi)$. For twice-differentiable Φ, $\rho(\Phi)$ is minimized if and only if Φ is a linear orthogonal transformation (mat).

Lemma A9. *The Wasserstein distance is positive homogeneous for scalar transformations of the underlying space. Let p, q be probability density functions defined over $\mathcal{X}$. For $\alpha > 0$ and the mapping $\Phi(x) = \alpha x$, let p_α and q_α be the distributions on $\alpha\mathcal{X}$ induced by Φ. Then:*

$$\mathit{Wass}\left(p_\alpha, q_\alpha\right) = \alpha\mathit{Wass}\left(p, q\right).$$

Proof. Following (Villani, 2008; Kuang & Tabak, 2016), we use another characterization of the Wasserstein distance. Let $\mathcal{M}_{p,q}$ be the set of mass preserving maps from $\mathcal{X}$ to itself which map the distribution p to the distribution q. That is, $\mathcal{M}_{p,q} = \{M : \mathcal{X}\to\mathcal{X}$ s.t. $q(M(S)) = p(S)$ for all measurable bounded $S \subset \mathcal{X}\}$. We then have that:

$$\mathrm{Wass}(p,q) = \inf_{M\in\mathcal{M}_{p,q}}\int_{\mathcal{X}}\|M(x)-x\|p(x)\,dx. \tag{26}$$

It is known that the infimum in (26) is actually achievable (Villani, 2008, Theorem 5.2). Denote by $M^* : \mathcal{X}\to\mathcal{X}$ the map achieving the infimum for $\mathrm{Wass}(p,q)$. Define $M_\alpha^* : \alpha\mathcal{X}\to\alpha\mathcal{X}$, by $M_\alpha^*(x') = \alpha M^*(\frac{x'}{\alpha})$, where $x' = \alpha x$. M_α^* maps p_α to q_α, and we have that $\|M_\alpha^*(x') - x'\| = \alpha\|M^*(x)-x\|$. Therefore M_α^* achieves the infimum for the pair (p_α, q_α), and we have that $\mathrm{Wass}\left(p_\alpha, q_\alpha\right) = \alpha\mathrm{Wass}\left(p,q\right)$. □

A.4. Functions in the unit ball of a RKHS

Let $\mathcal{H}_x, \mathcal{H}_r$ be a reproducing kernel Hilbert space, with corresponding kernels $k_x(\cdot,\cdot)$, $k_r(\cdot,\cdot)$. We have for all $x\in\mathcal{X}$ that $k_x(\cdot,x)$ is its Hilbert space mapping, and similarly $k_r(\cdot,r)$ for all $r\in\mathcal{R}$.

Recall that the major condition in Lemma A4 is that $\frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. The function space G we use here is $\mathrm{G} = \{g\in\mathcal{H}_r$ s.t. $\|g\|_{\mathcal{H}_r}\leq 1\}$.

We will focus on the case where L is the squared loss, and we will make the following two assumptions:

Assumption A4. *There exist $f_0^Y, f_1^Y \in \mathcal{H}_x$ such that $m_t(x) = \left\langle f_t^Y, k_x(x,\cdot)\right\rangle_{\mathcal{H}_x}$, i.e. the mean potential outcome functions m_0, m_1 are in $\mathcal{H}_x$. Further assume that $\|f_t^Y\|_{\mathcal{H}_x} \leq K$.*

Definition A14. *Define $\eta_{Y_t}(x) := \sqrt{\int_{\mathcal{Y}} (Y_t - m_t(x))^2 p(Y_t|x)}$. $\eta_{Y_t}(x)$ is the standard deviation of $Y_t|x$.*

Assumption A5. *There exists $f_0^\eta, f_1^\eta \in \mathcal{H}_x$ such that $\eta_{Y_t}(x) = \langle f_t^\eta, k_x(x,\cdot)\rangle_{\mathcal{H}_x}$, i.e. the conditional standard deviation functions of $Y_t|x$ are in $\mathcal{H}_x$. Further assume that $\|f_t^\eta\|_{\mathcal{H}_x} \leq M$.*

Assumption A6. *Let $\Phi : \mathcal{X} \to \mathcal{Y}$ be an invertible representation function, and let Ψ be its inverse. We assume there exists a bounded linear operator $\Gamma_\Phi : \mathcal{H}_r \to \mathcal{H}_x$ such that $\left\langle f_t^Y, k_x(\Psi(r),\cdot)\right\rangle_{\mathcal{H}_x} = \left\langle f_t^Y, \Gamma_\Phi k_r(r,\cdot)\right\rangle_{\mathcal{H}_x}$. We further assume that the Hilbert-Schmidt norm (operator norm) $\|\Gamma_\Phi\|_{HS}$ of Γ_Φ is bounded by K_Φ.*

The two assumptions above amount to assuming that Φ can be represented as one-to-one linear map between the two Hilbert spaces $\mathcal{H}_x$ and $\mathcal{H}_r$.

Under Assumptions A4 and A6 about m_0, m_1, and Φ, we have that $m_t(\Psi(r)) = \left\langle \Gamma_\Phi^* f_t^Y, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$, where Γ_Φ^* is the adjoint operator of Γ_Φ (Grunewalder et al., 2013).

Lemma A10. *Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis, and assume that there exist $f_t^h \in \mathcal{H}_r$ such that $h(r,t) = \left\langle f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$, and such that $\|f_t^h\|_{\mathcal{H}_r} \leq b$. Under Assumption A4 about m_0, m_1, we have that $\ell_{h,\Phi}(\Psi(r),t) = \int_{\mathcal{Y}} (Y_t - h(r,t))^2 p(Y_t|r) dY_t$ is in the tensor Hilbert space $\mathcal{H}_r \otimes \mathcal{H}_r$. Moreover, the norm of $\ell_{h,\Phi}(\Psi(r),t)$ in $\mathcal{H}_r \otimes \mathcal{H}_r$ is upper bounded by $4\left(K_\Phi^2 K^2 + b^2\right)$.*

Proof. We first decompose $\int_{\mathcal{Y}} (Y_t - h(r,t))^2 p(Y_t|x) dY_t$ into a noise and mean fitting term, using $r = \Phi(x)$:

$$\ell_{h,\Phi}(\Psi(r),t) =$$
$$\int_{\mathcal{Y}} (Y_t - h(r,t))^2 p(Y_t|r)\, dY_t =$$
$$\int_{\mathcal{Y}} (Y_t - m_t(x) + m_t(x) - h(\Phi(x),t))^2 p(Y_t|x)\, dY_t =$$
$$\int_{\mathcal{Y}} (Y_t - m_t(x))^2 p(Y_t|x)\, dY_t +$$
$$(m_t(x) - h(\Phi(x),t))^2 +$$
$$2\int_{\mathcal{Y}} (Y_t - m_t(x))(m_t(x) - h(\Phi(x),t)) p(Y_t|x) dY_t = \quad (27)$$
$$\eta_{Y_t}^2(x) + (m_t(x) - h(\Phi(x),t))^2 + 0, \quad (28)$$

where equality (27) is by Definition A14 of η, and because $\int_{\mathcal{Y}} (Y_t - m_t(x))\, p(Y_t|x)\, dY_t = 0$ by definition of $m_t(x)$.

Moving to $\mathcal{R}$, recall that $r = \Phi(x)$, $x = \Psi(r)$. By linearity of the Hilbert space, we have that $m_t(\Psi(r)) - h(r,t) = \left\langle \Gamma_\Phi^* f_t^Y, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r} - \left\langle f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r} = \left\langle \Gamma_\Phi^* f_t^Y - f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$. By a well known result (Steinwart & Christmann, 2008, Theorem 7.25), the product $(Y_t(\Psi(r)) - h(r,t)) \cdot (Y_t(\Psi(r)) - h(r,t))$ lies in the tensor product space $\mathcal{H}_r \otimes \mathcal{H}_r$, and is equal to $\left\langle (\Gamma_\Phi^* f_t^Y - f_t^h) \otimes (\Gamma_\Phi^* f_t^Y - f_t^h), k_r(r,\cdot) \otimes k_r(r,\cdot)\right\rangle_{\mathcal{H}_r \otimes \mathcal{H}_r}$. The norm of this function in $\mathcal{H}_r \otimes \mathcal{H}_r$ is $\|\Gamma_\Phi^* f_t^Y - f_t^h\|_{\mathcal{H}_r}^2$. This is the general Hilbert space version of the fact that for a vector $w \in \mathbb{R}^d$ one has that $\|ww^\top\|_F = \|w\|_2^2$, where $\|\cdot\|_F$ is the matrix Frobenius norm, and $\|\cdot\|_2^2$ is the square of the standard Euclidean norm. We therefore have a similar result for $\eta_{Y_t}^2$, using Assumption A5: $\eta_{Y_t}^2(x) = \eta_{Y_t}^2(\Psi(r)) = \left\langle \Gamma_\Phi^* f_t^\eta \otimes \Gamma_\Phi^* f_t^\eta, k_r(r,\cdot) \otimes k_r(r,\cdot)\right\rangle_{\mathcal{H}_r \otimes \mathcal{H}_r}$. The norm of this function in $\mathcal{H}_r \otimes \mathcal{H}_r$ is $\|\Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r}^2$. Overall this leads us to conclude, using Equation (28) that $\ell_{h,\Phi}(\Psi(r),t) \in \mathcal{H}_r \otimes \mathcal{H}_r$. Now we have, using (28):

$$\|\ell_{h,\Phi}(\Psi(r),t)\|_{\mathcal{H}_r \otimes \mathcal{H}_r} =$$
$$\|(\Gamma_\Phi^* f_t^Y - f_t^h) \otimes (\Gamma_\Phi^* f_t^Y - f_t^h) + \Gamma_\Phi^* f_t^\eta \otimes \Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r \otimes \mathcal{H}_r} \leq \quad (29)$$
$$\|\Gamma_\Phi^* f_t^Y - f_t^h\|_{\mathcal{H}_r}^2 + \|\Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r}^2 \leq \quad (30)$$
$$2\|\Gamma_\Phi^* f_t^Y\|_{\mathcal{H}_r}^2 + 2\|f_t^h\|_{\mathcal{H}_r}^2 + \|\Gamma_\Phi^* f_t^\eta\|_{\mathcal{H}_r}^2 \leq \quad (31)$$
$$\|\Gamma_\Phi^*\|_{HS}^2 \left(2\|f_t^Y\|_{\mathcal{H}_x}^2 + \|f_t^\eta\|_{\mathcal{H}_x}^2\right) + 2\|f_t^h\|_{\mathcal{H}_r}^2 = \quad (32)$$
$$\|\Gamma_\Phi\|_{HS}^2 \left(2\|f_t^Y\|_{\mathcal{H}_x}^2 + \|f_t^\eta\|_{\mathcal{H}_x}^2\right) + 2\|f_t^h\|_{\mathcal{H}_r}^2 \leq \quad (33)$$
$$2K_\Phi^2(K^2 + M^2) + 2b^2.$$

Inequality (29) is by the norms given above and the triangle inequality. Inequality (30) is because for any Hilbert space $\mathcal{H}$, $\|a-b\|_{\mathcal{H}}^2 \leq 2\|a\|_{\mathcal{H}}^2 + 2\|b\|_{\mathcal{H}}^2$. Inequality (31) is by the definition of the operator norm. Equality (32) is because the norm of the adjoint operator is equal to the norm of the original operator, where we abused the notation $\|\cdot\|_{HS}$ to mean both the norm of operators from $\mathcal{H}_x$ to $\mathcal{H}_r$ and vice-versa. Finally, inequality (33) is by Assumptions A4, A5 and A6, and by the Lemma's premise on the norm of f_T^h. □

Lemma A11. *Let $u = p(t=1)$ be the marginal probability of treatment, and assume $0 < u < 1$. Assume the distribution of Y_t conditioned on x follows Assumptions A5 with constant M. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function which obeys Assumption A6 with corresponding operator Γ_Φ with operator norm K_Φ. Let the functions Y_0, Y_1 obey Assumption A4, with bounded Hilbert space norm K . Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis, and assume that there exist $f_t^h \in \mathcal{H}_r$ such that $h(r,t) = \left\langle f_t^h, k_r(r,\cdot)\right\rangle_{\mathcal{H}_r}$, such that $\|f_t^h\|_{\mathcal{H}_r} \leq b$. Assume that ϵ_F and ϵ_{CF} are defined with respect to L being the*

squared loss. Then:

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + 2\left(K_\Phi^2(K^2+M^2)+b^2\right)\cdot \mathit{MMD}(p_\Phi^{t=1}, p_\Phi^{t=0}), \tag{34}$$

where ϵ_{CF} and ϵ_F use the squared loss.

Proof. We will apply Lemma A4 with $\mathrm{G} = f \in \mathcal{H}_r \otimes \mathcal{H}_r$ s.t. $\|f\|_{\mathcal{H}_r\otimes\mathcal{H}_r} \leq 1$. By Lemma A10, we have that for $B_\Phi = 2\left(K_\Phi^2(K^2+M^2)+b^2\right)$ and L being the squared loss, $\frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. Inequality (34) then holds as a special case of Lemma A4. □

Theorem 3. *Under the assumptions of Lemma A11, using the squared loss for ϵ_F, we have:*

$$\epsilon_{PEHE}(h,\Phi) \leq 2\epsilon_F^{t=0}(h,\Phi) + 2\epsilon_F^{t=1}(h,\Phi) - 4\sigma_Y^2 + 4\left(K_\Phi^2(K^2+M^2)+b^2\right)\cdot \mathit{MMD}(p_\Phi^{t=1}, p_\Phi^{t=0}).$$

Proof. Plug in the upper bound of Lemma A11 into the upper bound of Theorem 1. □

B. Algorithmic details

We give details about the algorithms used in our framework.

B.1. Minimizing the Wasserstein distance

In general, computing (and minimizing) the Wasserstein distance involves solving a linear program, which may be prohibitively expensive for many practical applications. Cuturi (2013) showed that an approximation based on entropic regularization can be obtained through the Sinkhorn-Knopp matrix scaling algorithm, at orders of magnitude faster speed. Dubbed Sinkhorn distances, the approximation is computed using a fixed-point iteration involving repeated multiplication with a kernel matrix K. We can use the algorithm of Cuturi (2013) in our framework. See Algorithm 2 for an overview of how to compute the gradient g_1 in Algorithm 1. When computing g_1, disregarding the gradient $\nabla_{\mathbf{W}} T^*$ amounts to minimizing an upper bound on the Sinkhorn transport. More advanced ideas for stochastic optimization of this distance have recently proposed by Aude et al. (2016), and might be used in future work.

While our framework is agnostic to the parameterization of Φ, our experiments focus on the case where Φ is a neural network. For convenience of implementation, we may represent the fixed-point iterations of the Sinkhorn algorithm as a recurrent neural network, where the states u_t evolve according to

$$u_{t+1} = n_t./(n_c K(1./(u_t^\top K)^\top)) .$$

Here, K is a kernel matrix corresponding to a metric such as the euclidean distance, $K_{ij} = e^{-\lambda\|\Phi(x_i)-\Phi(x_j)\|_2}$, and n_c, n_t are the sizes of the control and treatment groups. In this way, we can minimize our entire objective with most of the frameworks commonly used for training neural networks, out of the box.

Algorithm 2 Computing the stochastic gradient of the Wasserstein distance

1: **Input:** Factual $(x_1,t_1,y_1),\ldots,(x_n,t_n,y_n)$, representation network $\Phi_{\mathbf{W}}$ with current weights by $\mathbf{W}$
2: Randomly sample a mini-batch with m treated and m' control units $(x_{i_1},0,y_{i_1}),\ldots,(x_{i_m},0,y_{i_m}),(x_{i_{m+1}},1,y_{i_{m+1}}),\ldots,(x_{i_{2m}},1,y_{i_{2m}})$
3: Calculate the $m\times m$ pairwise distance matrix between all treatment and control pairs $M(\Phi_{\mathbf{W}})$: $M_{kl}(\Phi) = \|\Phi_{\mathbf{W}}(x_{i_k}) - \Phi_{\mathbf{W}}(x_{i_{m+l}})\|$
4: Calculate the approximate optimal transport matrix T^* using Algorithm 3 of Cuturi & Doucet (2014), with input $M(\Phi_{\mathbf{W}})$
5: Calculate the gradient: $g_1 = \nabla_{\mathbf{W}} \langle T^*, M(\Phi_{\mathbf{W}})\rangle$

B.2. Minimizing the maximum mean discrepancy

The MMD of treatment populations in the representation Φ, for a kernel $k(\cdot,\cdot)$ can be written as,

$$\mathrm{MMD}_k(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^m, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m'}) = \tag{35}$$

$$\frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{k=1,k\neq j}^{m} k(\Phi_{\mathbf{W}}(x_{i_j}),\Phi_{\mathbf{W}}(x_{i_k})) \tag{36}$$

$$+\frac{2}{mm'}\sum_{j=1}^{m}\sum_{k=m}^{m+m'} k(\Phi_{\mathbf{W}}(x_{i_j}),\Phi_{\mathbf{W}}(x_{i_k})) \tag{37}$$

$$+\frac{1}{m'(1-m')}\sum_{j=1}^{m}\sum_{k=m,k\neq j}^{m'} k(\Phi_{\mathbf{W}}(x_{i_j}),\Phi_{\mathbf{W}}(x_{i_k})) \tag{38}$$

The linear maximum-mean discrepancy can be written as a distance between means. In the notation of Algorithm 1,

$$\mathrm{MMD} = 2\left\|\frac{1}{m}\sum_{j=1}^{m}\Phi_{\mathbf{W}}(x_{i_j}) - \frac{1}{m'}\sum_{k=m+1}^{m'}\Phi_{\mathbf{W}}(x_{i_k})\right\|_2$$

Let

$$\mathbf{f}(\mathbf{W}) = \frac{1}{m}\sum_{j=1}^{m}\Phi_{\mathbf{W}}(x_{i_j}) - \frac{1}{m'}\sum_{k=m+1}^{m+m'}\Phi_{\mathbf{W}}(x_{i_k})$$

Table 2. Hyperparameters and ranges.

Parameter	Range
Imbalance parameter, α	$\{10^{k/2}\}_{k=-10}^{6}$
Num. of representation layers	$\{1, 2, 3\}$
Num. of hypothesis layers	$\{1, 2, 3\}$
Dim. of representation layers	$\{20, 50, 100, 200\}$
Dim. of hypothesis layers	$\{20, 50, 100, 200\}$
Batch size	$\{100, 200, 500, 700\}$

Then the gradient of the MMD with respect to $\mathbf{W}$ is,

$$g_1 = 2\frac{d\mathbf{f}(\mathbf{W})}{d\mathbf{W}}\frac{\mathbf{f}(\mathbf{W})}{\|\mathbf{f}(\mathbf{W})\|_{\mathbf{2}}} .$$

C. Experimental details

C.1. Hyperparameter selection

Standard methods for hyperparameter selection, such as cross-validation, are not generally applicable for estimating the PEHE loss since only one potential outcome is observed (unless the outcome is simulated). For real-world data, we may use the observed outcome $y_{j(i)}$ of the nearest neighbor $j(i)$ to i in the opposite treatment group, $t_{j(i)} = 1 - t_i$ as surrogate for the counterfactual outcome. We use this to define a nearest-neighbor approximation of the PEHE loss, $\epsilon_{\text{PEHE}nn}(f) = \frac{1}{n}\sum_{i=1}^{n}\left((1-2t_i)(y_{j(i)} - y_i) - (f(x_i, 1) - f(x_i, 0))\right)^2$. On IHDP, we use the objective value on the validation set for early stopping in CFR, and $\epsilon_{\text{PEHE}nn}(f)$ for hyperparameter selection. On the Jobs dataset, we use the policy risk on the validation set.

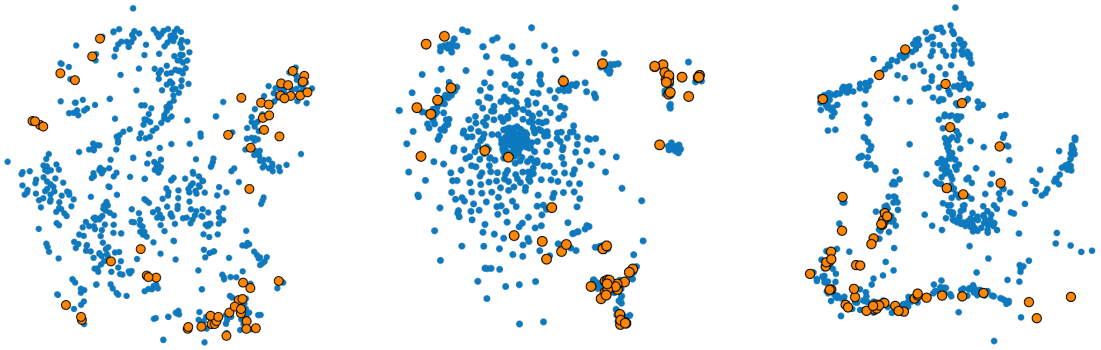
See Table 2 for a description of hyperparameters and search ranges.

C.2. Learned representations

Figure 4 show the representations learned by our CFR algorithm.

C.3. Absolute error for increasingly imbalanced data

Figure 5 shows the results of the same experiment as Figure 2 of the main paper, but in absolute terms.



(a) Original data (b) Linear MMD (c) Wasserstein

Figure 4. t-SNE visualizations of the balanced representations of IHDP learned by our algorithms CFR, CFR MMD and CFR Wass. We note that the nearest-neighbor like quality of the Wasserstein distance results in a strip-like representation, whereas the linear MMD results in a ball-like shape in regions where overlap is small.

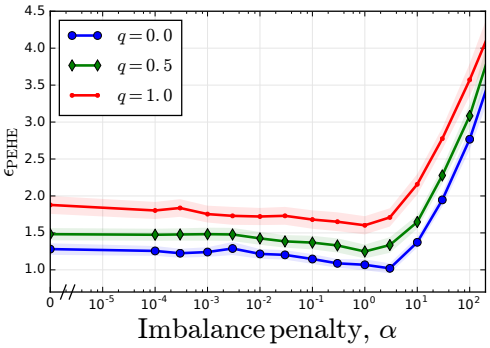


Figure 5. Out-of-sample error in estimated ITE, as a function of IPM regularization parameter for CFR Wass, on 500 realizations of IHDP, with high ($q = 1$), medium and low (artificial) imbalance between control and treated.