Bounding and Minimizing Counterfactual Error

Uri Shalit*
Courant Institute for Mathematical Sciences
New York University
New York, NY 10003
shalit@cs.nyu.edu

Fredrik D. Johansson*
Computer Science and Engineering Department
Chalmers University of Technology
Göteborg, SE-412 96, Sweden
frejohk@chalmers.se

David Sontag
Courant Institute for Mathematical Sciences
New York University
New York, NY 10003
dsontag@cs.nyu.edu

Abstract

There is intense interest in applying machine learning methods to problems of causal inference which arise in applications such as healthcare, economic policy, and education. In this paper we use the counterfactual inference approach to causal inference, and propose new theoretical results and new algorithms for performing counterfactual inference. Building on an idea recently proposed by Johansson et al. [20], our results and methods rely on learning so-called "balanced" representations: representations that are similar between the factual and counterfactual distributions. We give a novel, simple and intuitive bound, showing that the expected counterfactual error of a representation is bounded by a sum of the factual error of that representation and the distance between the factual and counterfactual distributions induced by the representation. We use Integral Probability Metrics to measure distances between distributions, and focus on two special cases: the Wasserstein distance and the Maximum Mean Discrepancy (MMD) distance. Our bound leads directly to new algorithms, which are simpler and easier to employ compared to those suggested in [20]. Experiments on real and simulated data show the new algorithms match or outperform state-of-the-art methods.

1 Introduction

Causal inference questions are central to policy makers and scientists across many fields. Examples abound: in healthcare one is often interested in the relative efficacy of different medications; in economic policy, policy makers debate the effect of job training; in marketing, companies are interested in the causal effect of an online ad on a customer's buying habits. However, traditional causal inference methods are sometimes ill-suited to take advantage of the vast increase in the size and scope of data available across these fields. Here, we build on earlier work by [20], and show how modern machine learning techniques of representation learning [4] along with using integral probability metrics [30] can be fruitfully applied to problems of causal inference.

Our work relies on the framework of counterfactual inference as a means towards causal inference, associated with the Rubin-Neyman potential outcomes framework [27]. It assumes that for a context $x \in \mathcal{X}$, and a treatment or intervention $t \in \{0, 1\}$, there are two potential outcomes: $Y_0(x)$ for $t = 0$, and $Y_1(x)$ for $t = 1$. For example, x can denote the set of lab tests and demographic factors of a diabetic patient, $t = 0$ denote the standard medication for controlling blood sugar, $t = 1$ denotes a

*Equal contribution

new medication, and $Y_0(x)$ and $Y_1(x)$ indicate the patient's blood sugar level after treatments $t = 0$ and $t = 1$, respectively. We are usually interested in knowing $\tau(x) = Y_1(x) - Y_0(x)$, i.e. what is the *treatment effect* of $t = 1$. The fundamental problem of causal inference is, that, for a given context x, we only ever observe $Y_1(x)$ or $Y_0(x)$, but never both of them. Moreover, in many cases of interest, the distribution of the treatment assignment t is dependent on the context x, as is often the case in observational studies [26]. For example, medications might be prescribed based on specific coverage policies, and job training might only be given to those motivated enough to seek it.

We call the joint distribution over contexts and treatment the *factual* distribution, $p^F(x, t)$. The counterfactual distribution is $p^{CF}(x, t) = p^F(x, 1 - t)$. The factual (observed) outcomes for unit $x_i \in \mathcal{X}$ with treatment $t_i \in \{0, 1\}$ are: $y^F(x_i) = t_i Y_1(x_i) + (1 - t_i)Y_0(x_i)$. Similarly, the counterfactual outcomes are $y^F(x_i) = (1 - t_i)Y_1(x_i) + t_i Y_0(x_i)$.

Although we never have access to $y^{CF}(x_i)$, we do have access to $y^F(x_i)$. Therefore, one approach would to this problem would be to estimate Y_0 and Y_1 directly from the factual sample, using ordinary machine learning techniques. However, when generalizing to the counterfactual sample we have another source of variance beyond those considered in the classic supervised learning formulation. This is the variance incurred by learning from one distribution and performing inference on another [2, 23]. This problem is strongly related to the problem of covariate shift [20], and indeed our bound and algorithm could be easily adapted for use in covariate shift problems, in line with recently proposed methods [13].

Recently, [20] have argued that the covariate shift variance can be controlled by learning a so-called "balanced representation": a representation of the data that makes the factual and counterfactual distributions more similar. In their work, [20] study the performance of learning a balanced representation and then fitting a linear ridge-regression model on top of it. They bound the relative error of fitting a ridge-regression using the counterfactual distribution versus fitting a ridge-regression using the factual distribution. Unfortunately, the relative error term they bound is not informative regarding the absolute quality of the representation, neither on the factual nor on the counterfactual distribution.

In this paper we prove what we consider a much more informative bound. We show that under mild assumptions about the invertability of the learned representation, the expected counterfactual loss is upper-bounded by a sum of two terms: the expected factual loss, and a constant C times the distance between the distributions of the treated ($t = 1$) and control ($t = 0$) populations in the representation space. The intuition behind this bound is clear: generalizing to the counterfactual distribution is no harder than generalizing to the factual distribution from which we can sample, plus a term measuring the difference between the two distributions. Here, we use a general family of distance functions between distributions called *Intergral Probability Metrics* (IPM) [24, 30]. In particular we concentrate on two IPMs: The Wasserstein metric [32, 8]and the Maximum Mean Discrepancy metric (MMD, [15]). Both have been successfully applied in recent years to a wide variety of applications [16, 25, 11, 9, 12]. For the Wasserstein metric, we give an explicit term upper-bounding the constant C. Our bound is not limited to linear hypotheses, and holds for a wide class of loss functions, including the squared loss over a compact domain, the absolute loss, and the logistic loss, making our approach applicable to both classification and regression tasks.

The bound we derive points the way to family of algorithms: jointly learn a hypothesis and a representation which minimize a weighted sum of the factual loss (the standard supervised machine learning objective), and the distance between the control and treated distributions induced by the representation. In the Experiments section 4 we apply algorithms based on multi-layer neural nets as representations and hypotheses, along with MMD or Wasserstein distributional distance metrics; see Figure 1 for the basic architercture. These algorithms are conceptually simpler than the ones proposed by [20], since they do not require a two-stage fitting procedure, and avoid the need to compute nearest neighbors. We show that our methods achieve better performance on the synthetic tasks presented by [20]. We also show that our methods can achieve competitive results on a real world causal inference benchmark: the widely used National Supposed Work survey [22, 10, 29].

Our contributions in this paper are as follows: (1) We prove a novel, simple and meaningful upper bound on the counterfactual loss. (2) We show the relation between minimizing counterfactual loss and minimizing the error in estimating treatment effect, which is often the quantity we are truly interested in. (3) Our bound and algorithm are applicable to a very wide family of loss functions and models for classification and regression, compared with the previous paper's restriction to linear

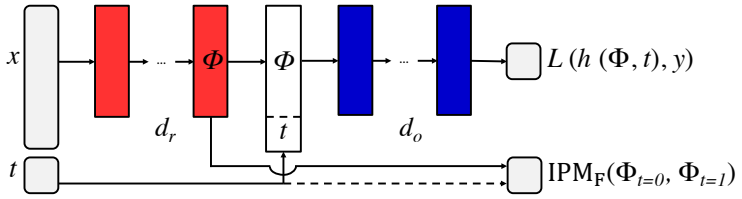


Figure 1: Neural network architecture for counterfactual regression. L is a loss function, IPM_F is an Integral Probability Metric.

models and absolute loss. (4) Our experiments show the benefit of our method, and we show results on a real-world causal inference benchmark, the National Work Survey dataset [22, 29].

2 Counterfactual learning bound

In this section we will prove several bounds relating the learning of representations, factual and counterfactual loss, and error in estimating the treatment effect for unit x. The most important quantity we bound is ℓ_{CF}: the expected loss of a model, with expectation taken over the *counterfactual distribution*, defined below. The bounds are expressed in terms of ℓ_F: the expected loss of the same model over the *factual distribution*, along with a distance measure between the distribution of treated and control units. The term ℓ_F is the classic machine learning generalization error, and in turn can be upper bounded using the empirical error over the observations and model complexity terms, applying well-known machine learning theory [28]. All proofs are in the supplement.

2.1 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 is a binary variable. There exists a *factual distribution* $p^F(x,t)$ defined jointly over the covariate and treatment space $\mathcal{X} \times \{0,1\}$. The treated and control distributions are the factual distribution conditioned on treatment: $p^{t=1}(x) := p^F(x|t=1)$, and $p^{t=0}(x) := p^F(x|t=0)$, respectively. The counterfactual distribution is the factual distribution with the treatment assignment flipped: $p^{CF}(x,t) := p^F(x,1-t) = p^F(1-t|x)p^F(x)$. Note that if treatment assignment is independent of the context x, and $p^F(t|x) = \frac{1}{2}$, then the counterfactual and factual distributions are identical; this is exactly the case of a randomized-control trial, where treatment is assigned randomly with equal probability to units.

In this paper we employ the Rubin-Neyman potential outcomes framework [27]: We assume there exist two deterministic functions which are the true, unknown labeling functions $Y_t : \mathcal{X} \to \mathcal{Y}$, for the two treatment assignments $t = 0, 1$. For a labeled sample $(x_1, t_1, y_1^F), \ldots (x_n, t_n, y_n^F) \subset \mathcal{X} \times \{0,1\} \times \mathcal{Y}$ from the factual distribution, we have that the factual labels are $y_i^F = t_i Y_1(x_i) + (1-t_i)Y_0(x_i)$.

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

Assumption 1. *The representation Φ is a twice-differentiable, one-to-one function. Without loss of generality we will assume that $\mathcal{R}$ is the image of $\mathcal{X}$ under Φ.*

Definition 1. *Define $\Psi : \mathcal{R} \to \mathcal{X}$ to be the inverse of Φ, such that $\Psi(\Phi(x)) = x$ for all $x \in \mathcal{X}$.*

The representation Φ pushes forward the treated and control distributions into the new space $\mathcal{R}$: we denote the induced distribution by p_Φ^F, defined over $\mathcal{R} \times \{0,1\}$. Define p_Φ^{CF} analogously. We also define $p_\Phi^{t=1}(r) := p_\Phi^F(r|t=1)$, $p_\Phi^{t=0}(r) := p_\Phi^F(r|t=0)$, to be the treated and control distributions induced over $\mathcal{R}$. For a one-to-one Φ, the distributions p_Φ^F and p_Φ^{CF} can be obtained by the standard change of variables formula, using the determinant of the Jacobian of $\Psi(r)$.

Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function.

Definition 2. *Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis defined over the representation space $\mathcal{R}$. The expected factual loss and counterfactual losses of h and Φ are, respectively:*

$$\ell_F(h,\Phi) = \int_{\mathcal{X}\times\{0,1\}} L\left(Y_t(x),\, h(\Phi(x),t)\right) p^F(x,t)\, dxdt = \int_{\mathcal{R}\times\{0,1\}} L\left(Y_t(\Psi(r)),\, h(r,t)\right) p^F_\Phi(r,t)\, drdt,$$

$$\ell_{CF}(h,\Phi) = \int_{\mathcal{X}\times\{0,1\}} L\left(Y_t(x),\, h(\Phi(x),t)\right) p^{CF}(x,t)\, dxdt = \int_{\mathcal{R}\times\{0,1\}} L\left(Y_t(\Psi(r)),\, h(r,t)\right) p^{CF}_\Phi(r,t)\, drdt,$$

where the equalities follow from change of variables in integration and the definition of $p^F_\Phi(r,t)$, $p^{CF}_\Phi(r,t)$. As noted above, the term ℓ_F is simply the generalization error for the hypothesis $h(\Phi,t)$ over the factual distribution. The term ℓ_{CF} is the generalization of the same hypothesis over the counterfactual distribution.

Our proof relies on the notion of an *Integral Probability Metric*, which is a family of metrics between probability distributions [30, 24]. For two probability density functions p, q defined over $\mathcal{S} \subseteq \mathbb{R}^d$, and for a function family F of functions $g : \mathcal{S} \to \mathbb{R}$, we have that

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

Integral probability metrics are always symmetric and obey the triangle inequality, and trivially satisfy $\text{IPM}_\text{F}(p,p) = 0$. For rich enough function families F, we also have that $\text{IPM}_\text{F}(p,q) = 0 \implies p = q$, and then IPM_F is a true metric. Examples of function families F for which IPM_F is a true metric are the family of all bounded countinuous functions, the family of all 1-Lipschitz functions [30], and the unit-ball of functions in a universal reprodcuing Hilbert kernel space [15]. Here, we will employ an extended notion of IPM, where the probabilities are not necessarily normalized. We call this *Unnormalized Integral Probability Metric*. For two probability distribution functions as above, and positive scalars u_p, u_q, we have:

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

2.2 Bounds

Assumption 2. *Let $u = p^F(t=1)$ be the marginal probability of treatment, and assume $0 < u < 1$.*

Theorem 1 (General bound)**.** *Let* F *be a family of functions $f : \mathcal{R} \to \mathbb{R}$, and UIPM$_\text{F}(\cdot,\cdot)$ the unnormalized integral probability metric induced by* F*. Let Y_0, Y_1 be the true labeling functions. Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be an hypothesis. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function, with inverse Ψ. Assume that for $t = 0, 1$, we have that $f_{\Phi,h}(r) := L(Y_t(\Psi(r)), h(r,t)) \in \text{F}$. Then:*

$$\ell_{CF}(h,\Phi) \le \ell_F(h,\Phi) + 2\textit{UIPM}\left(u \cdot p^{t=1}_\Phi, (1-u)\cdot p^{t=0}_\Phi\right). \tag{1}$$

The upper bound in Eq (1) is itself a sum of two expectation terms. For an empirical sample from $p^F(x,t)$, and a family of representations and hypotheses, we can further upper bound ℓ_F by the empirical loss and a model complexity term using standard arguments [28]. The UIPM term can be consistently estimated from finite samples for the function families we use below [30].

Choosing a small function family F will make the bound tighter.However, choosing too small a family could result in an imcomputable bound. For example, for $\text{F} = \{f_{\Phi,h}\}$ we will have that $\text{IPM}_\text{F} = \ell_{CF}(\Phi,h) - \ell_F(\Phi,h)$, which we cannot compute since we do not know the counterfactual outcomes. In addition, for some function families there is no known way to efficiently compute the IPM distance or to take its gradients. In this paper we employ 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 [32, 30], denoted $\text{Wass}(p,q)$. The second is the family of norm-1 reproducing kernel Hilbert space (RKHS) functions, leading to the MMD metric [15, 30], denoted $\text{MMD}(p,q)$. Both the Wasserstein and MMD metrics have efficient algorithms which are consistent and can be applied in the finite sample case [30], and have been used for various machine learning tasks in recent years [16, 15, 8, 9]. In supplement C we show how to reduce the calculation of unnormalized Wasserstein and MMD to calculating standard Wasserstein and MMD distances.

The Wasserstein metric

Assumption 3. *For the loss L, for all $y \in \mathcal{Y}$, both $L(\cdot, y)$ and $L(y, \cdot)$ are Lipschitz functions with Lipschitz constant upper bounded by K_L.*

Examples of loss function which follow Assumption 3 are the absolute loss $|y_1 - y_2|$, the logistic loss for binary labels $\log(1 + e^{-y_1 y_2})$, and the squared loss $(y_1 - y_2)^2$ if $\mathcal{Y}$ is bounded. We now employ a measure bounding the degree to which Φ is information-preserving. For this we use the reciprocal condition number of the Jacobian matrix of Φ.

Definition 3. *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 the 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 $\rho(\Phi) \geq 1$. 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 scaling it with $1/\sup_{x \in \mathcal{X}} \sigma_{max}\left(\frac{\partial \Phi(x)}{\partial x}\right)$.

Theorem 2 (Wasserstein bound)**.** *Let Φ be a one-to-one, Jacobian-normalized representation function. Let K be the Lipschitz constant of the functions Y_0, Y_1 on $\mathcal{X}$. Let K_L be the Lipschitz constant of the loss function L. Let $h : \mathcal{R} \times \{0, 1\} \to \mathbb{R}$ be an hypothesis with Lipschitz constant bK. Then:*

$$\ell_{CF}(h, \Phi) \leq \ell_F(h, \Phi) + 2\,(\rho(\Phi) + b) \cdot K \cdot K_L \cdot \mathit{Wass}(u \cdot p_\Phi^{t=1}, (1 - u) \cdot p_\Phi^{t=0}). \tag{2}$$

We examine the constant $(\rho(\Phi) + b) \cdot K$ in Theorem 2. K, the Lipschitz constant of Y_0 and Y_1, is not under our control and measures an aspect of the complexity of the true underlying functions we wish to approximate. The term K_L depends on our choice of loss function. 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 $\ell_F(h, \Phi)$ to be larger since a small b implies a less flexible h. Finally, consider the term $\rho(\Phi)$. For specific families of representation Φ, $\rho(\Phi)$ can be upper bounded. For example, if Φ is a single-layer neural net with weight matrix W and $\tanh$ non-linearity, then one can show that if the norm and the inverse condition number of W are bounded, then so is $\rho(\Phi)$.

The MMD metric We prove a Theorem similar to 2, for the case of Reproducing Kernel Hilbert Space function spaces, with the corresponding Maximum Mean Discrepancy metric. For lack of space we state the results in the supplement. Unlike in Theorem 2, the MMD case has a constant which is harder to explicitly bound; we defer a deeper look into this issue to future work.

The above Theorems hold for any given h and Φ obeying the theorems' conditions. Therefore, we can attempt to minimize the upper bounds in Eqs. (1) and (2) with respect to Φ and h, as we suggest in Algorithm 1, in order to minimize the counterfactual loss.

Proof idea The full proof of the Theorems and Lemmas above is given in the supplement. The main idea is bounding the difference $\ell_{CF} - \ell_F$ in terms of an IPM between the counterfactual and factual distributions. Once we have that, we have that

$$\ell_{CF} = \ell_F + (\ell_{CF} - \ell_F) \leq \ell_F + \text{IPM}_\text{F}(P^{CF}, p^F).$$

We then show that we can in fact calculate $\text{IPM}_\text{F}(P^{CF}, p^F)$ in terms of an IPM between the treatment and control distributions $\text{IPM}_\text{F}(p^{t=1}, p^{t=0})$.

Counterfactual loss and treatment effect estimation In section B of the supplement, we show how the counterfactual loss relates to the error in estimating the true treatment effect $Y_1(x) - Y_0(x)$. We show that in the transductive setting where one of the potential outcomes is known, minimizing counterfactual loss is equivalent to minimizing the error in treatment effect estimation. For the inductive setting we show that the error can be bounded by the sum of the factual and counterfactual errors. Putting this together with our bounds above, we see that the error in estimating treatment effect is also upper bounded by a weighted sum of the factual error and an IPM distance term, justifying our theoretical and empirical approach.

Algorithm 1 Counterfactual balanced regression with integral probability metrics

Input: Factual sample $(x_1, t_1, y_1^F), \ldots, (x_n, t_n, y_n^F)$, scaling parameter $\alpha > 0$, loss function $L(\cdot, \cdot)$, representation network $\Phi_{\mathbf{W}}$ with initial weights by $\mathbf{W}$, outcome network $h_{\mathbf{V}}$ with intial weights $\mathbf{V}$, function family F for IPM loss
while not converged **do**
  Sample m control $\{(x_{i_j}, 0, y_{i_j}^F)\}_{j=1}^m$ and m' treated units $\{(x_{i_k}, 1, y_{i_k}^F)\}_{k=m}^{m+m'}$
  Calculate the gradient of the imbalance penalty:
  $g_1 = \nabla_{\mathbf{W}} \text{ IPM}_F \left(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^m, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m'} \right)$
  Calculate the gradients of the empirical loss:
  $g_2 = \nabla_{\mathbf{V}} \sum_j \frac{L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}^F\right)}{m+m'}$, $g_3 = \nabla_{\mathbf{W}} \frac{1}{m+m'} \sum_j \frac{L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}^F\right)}{m+m'}$
  Obtain step size scalar or matrix η with standard neural net methods e.g. RMSProp
  update $\mathbf{W} \leftarrow \mathbf{W} - \eta(\alpha \mathbf{g_1} + \mathbf{g_3})$ and $\mathbf{V} \leftarrow \mathbf{V} - \eta(\mathbf{g_2})$
  check convergence criterion
end while

3 Our approach

We propose a general framework for counterfactual estimation, based on the theoretical results of Section 2. Our algorithm is a single regularized minimization procedure which simultaneously fits both a balanced representation of the data, and a hypothesis for the outcome. This is in contrast to [20] who proposed a two-step procedure corresponding to their theoretical results based on the discrepancy distance [7]. We note that our framework is also more flexible in practice, as our theory supports multiple measures of balance that can can be minimized efficiently; this is only rarely true for variants of the discrepancy distance We minimize the following objective.

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

Here, $\text{IPM}_{\text{F}}(\cdot, \cdot)$ is the (empirical) integral probability metric defined by the function family F. In Section 2, we show that under certain conditions when IPM_{F} is the maximum mean discrepancy or the Wasserstein distance, (3) is an upper bound on the counterfactual error.

In this work, we let $\Phi(x)$ and $h(\Phi, t)$ be parameterized by a single neural network. This means that we can learn rich, non-linear representations and hypotheses with large flexibility. In [20], the authors considered using a linear variable selection model, and while our model allows for a sparse diagonal first layer which generalizes their framework, in practice, we have observed no gain from the variable selection. Our approach, in the neural network parameterization, is visualized in Figure 1. We train our models by minimizing (3) using stochastic gradient descent, as described in Algorithm 1. The details of how to obtain the gradient g_1 with respect to the IPM can be found in the supplement.

4 Experiments

We evaluate our framework in counterfactual regression and classification tasks, using different functions to measure imbalance, including the Wasserstein distance and the MMD, and compare to established methods. We report the absolute error or bias in estimating the average and individual treatment effects, ϵ_{ATE} and ϵ_{ITE}, as well as the *Precision in Estimation of Heterogeneous Effect* (PEHE)[19], $\text{PEHE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [(Y_1(x_i) - Y_0(x_i)) - (g(x_i, 1) - g(x_i, 0))]^2}$, where $g(x, t)$ is the predicted outcome for individual x under treatment t. We also report the RMSE of the predicted factual outcome (RMSE_{fact}) or the binary classification error (Err_{fact}). Standard methods for hyperparameter selection such as cross-validation are not applicable, as there are no samples of the counterfactual outcome. In simulated experiments, counterfactuals are available and we follow [20] by fitting hyperparameters on held-out set of repeated experiments. In experiments with real outcome, we use a surrogate for the counterfactual outcome to estimate the error, namely $y_{j(i)}^F$ – the factual outcome of the nearest neighbor $j(i)$ to i, that is in the opposite treatment group.

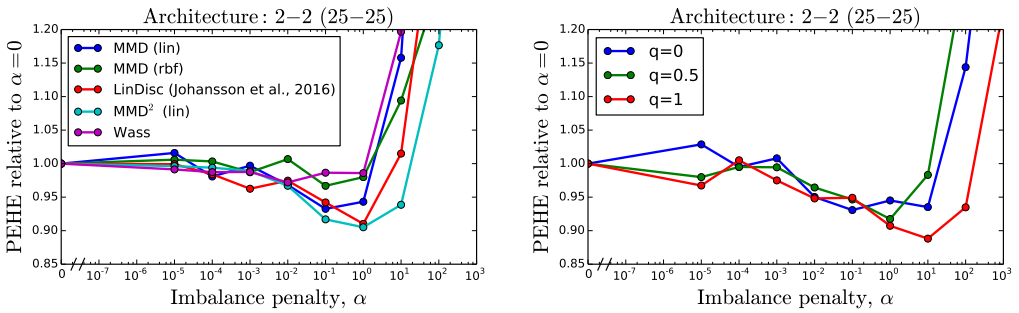


Figure 2: Relative gain of imbalance regularization on 500 repeated experiments of the IHDP, using different distance measures on (left), and with varying amount of imbalance (right).

Our approach is implemented as a feed-forward neural network with fully-connected ReLU layers, trained using RMSProp, with a small l2 weight decay, $\lambda = 10^{-3}$. Two architectures are evaluated. CFR-4-0 consists of 4 ReLU layers used for representation, and a single linear output layer. Following the notation of Fiugre 1, $dr = 4$, $do = 0$. CFR-2-2 consists of 2 ReLU representation layers, 2 ReLU layers after the treatment has been added, and a single linear output layer, $dr = 2, do = 2$, see Figure 1. For the IHDP data we use layers of 25 hidden units each. For Jobs, layers have 50 units. These architectures were selected based on factual validation error alone, by splitting the data.

In regression tasks, we compare our method to Ordinary Least Squares (OLS), Doubly Robust Regression (DR) [1], Bayesian Additive Regression Trees[2] (BART) [5], Causal Forests[3] (C.Forests) [33] as well as the Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) methods proposed by [20]. For DR, we estimate propensity scores using logistic regression. We also compare to a variable selection procedure dubbed LASSO + Ridge (L+R) in which a ridge regression model is fit to the variables selected by LASSO. In classification tasks we compare to Logistic Regression (LR), ℓ_1-regularized Logistic Regression (ℓ_1-LR) instead of OLS and L+R.

4.1 Simulated outcome - IHDP

Hill [19] compiled a semi-simulated dataset for counterfactual inference based on the Infant Health and Development Program (IHDP), in which the outcome is simulated. The covariates stem from a randomized experiment studying the effects of child care and home visits on (future) cognitive test scores. Imbalance in the covariates has been artificially introduced by removing a biased subset of the treatment population. The dataset comprises 747 observations (139 treated, 608 control) and 25 covariates measuring aspects of children and their mothers, see Hill (2011) [19]. We use the log-linear outcome model implemented as setting "B" in the NPCI package[4]. We also evaluate our method on another semi-synthetic dataset called News [20], the results on which are in the supplement.

The results of the experiments on IHDP are presented in Table 1. We see that in general, non-linear estimators such as BART, CFR, BNN and Causal Forests fair the best on both datasets. On the whole, using neural networks, even without balancing, give very good results. For a comparison of different distance measures, the MMD and its square (MMD^2) with linear and Gaussian-RBF kernels, the Wasserstein distance (Wass), and the linear discrepancy (LinDisc), see Figure 2. We see that MMD^2 (lin) performs the best, and the Wasserstein and MMD (rbf) the worst, although this might improve by chosing the variance parameter σ.

We investigate the effects of varying imbalance by constructing subsampled variants of the IHDP dataset. First, we fit a propensity score model, to form estimates $\hat{p}(t)$ of the treatment probability $p(t)$. Then, repeatedly with probability q we remove the remaining *control* observation that has propensity score $\hat{p}(t)$ closest to 1 and with probability $1 - q$, we remove a control observation uniformly at random. For high values of q, this gives a more imbalanced dataset, and for small q a more balanced. In our experiments, we consider $q = 0, 0.5, 1.0$. In total, we remove $m = 347$ observations from

[2] `https://cran.r-project.org/web/packages/BayesTree`

[3] `https://github.com/susanathey/causalTree/tree/forestCode`

[4] `https://github.com/vdorie/npci`

Table 1: Results and standard errors on IHDP over 1000 repeated experiments and Jobs (binary). MMD is the squared linear MMD. ITE and ATE are transductive. Lower is better. *Not applicable.

	IHDP			JOBS, BINARY			
	ϵ_{ITE}	ϵ_{ATE}	PEHE		ϵ_{ATT}	(%)	ERR_{fact}
OLS	4.6 ± 0.2	0.7 ± 0.0	5.2 ± 0.3	LR	0.043	56%	12.6%
DR	3.0 ± 0.1	0.2 ± 0.0	5.7 ± 0.3		0.065	84%	15.8%
L + R	2.8 ± 0.1	0.2 ± 0.0	5.7 ± 0.2	ℓ_1-LR	0.044	57%	12.4%
BLR	2.8 ± 0.1	0.2 ± 0.0	5.7 ± 0.3		*	*	*
BART [5]	2.1 ± 0.2	0.2 ± 0.0	1.7 ± 0.2		0.023	30%	11.6%
C.FORESTS [33]	2.2 ± 0.1	0.2 ± 0.0	3.7 ± 0.2		0.019	28%	11.7%
BNN-4-0 [20]	3.0 ± 0.0	0.3 ± 0.0	5.6 ± 0.3		*	*	*
BNN-2-2 [20]	1.7 ± 0.0	0.3 ± 0.0	1.6 ± 0.1		*	*	*
CFR-4-0 $\alpha = 0$	5.3 ± 0.1	0.8 ± 0.0	5.8 ± 0.3		0.059	75%	11.2%
CFR-4-0 WASS	3.0 ± 0.2	0.3 ± 0.0	5.7 ± 0.3		**0.012**	**16%**	11.0%
CFR-4-0 MMD	3.0 ± 0.1	0.3 ± 0.0	5.7 ± 0.3		0.041	52%	10.7%
CFR-2-2 $\alpha = 0$	1.8 ± 0.0	0.3 ± 0.0	1.5 ± 0.1		0.090	116%	10.9%
CFR-2-2 WASS	1.7 ± 0.1	0.3 ± 0.0	$\mathbf{1.4 \pm 0.1}$		0.040	51%	11.0%
CFR-2-2 MMD	$\mathbf{1.6 \pm 0.0}$	$\mathbf{0.2 \pm 0.0}$	$\mathbf{1.4 \pm 0.1}$		0.061	78%	10.6%

each set, leaving 400. In Figure 2, we see that as imbalance (q) increases, the relative gain from using our method, and the penalty needed, increase as well. This is expected as for data with no overlap at all, the best you can do is to disregard the covariates completely, corresponding to $\alpha = \infty$.

4.2 Real-world outcome - Jobs

LaLonde [22] conducted a well-known observational study based on the National Supported Work (NSW) program. It combines a randomized study based on the NSW with observational data to form a larger, observational dataset [29]. We refer to this dataset as *Jobs*. The original outcome to predict is the 1978 earnings and the 8 original covariates include age, education, ethnicity, as well as earnings in 1974 and 1975. To evaluate our framework for classification, we construct an alternative *binary* task of predicting unemployment in the Jobs study, i.e. the event that $y = 0$. We use the augmented feature set of Dehejia & Wahba [10]. In the notation of Smith et al. [29], we use the LaLonde experimental sample (297 treated, 425 control) and the PSID comparison group (2490 control). We can compute the average treatment effect on the treated (ATT = $886), as all treated were part of the original randomized experiment. In total, there were 482 (15%) people unemployed by the end of the study. We select parameters based on the PEHE_{nn} criteria defined above.

The results of the binary Jobs experiments are presented in Table 1. The results for the original, continuous task is in the supplement. On the whole, our proposed methods CFR-2-2 and CFR-4-0 perform well. The results only show error in the average effect, something that even linear methods can estimate well. However, since we achieve comparable average effect error, as well as better individual factual error, we expect that the counterfactual error is lower for our method.

5 Conclusion

In this paper we give a meaningful and intuitive bound for the problem of learning representations for counterfactual inference. Our bound relates counterfactual inference 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 leads to both synthetic and real-world tasks, showing that in every case our method matches or outperforms the most recent methods proposed for these tasks. Important open questions are theoretical considerations in choosing the imbalance weight α, and how to best derive confidence intervals for our model's predictions, as well as exploring in more depths the connection to domain adaptation problems.

Acknowledgments

We wish to thank Esteban Tabak and Marco Cuturi for fruitful conversations. We also wish to thank Stefan Wager for his help with the code for Causal Forests. DS and US were supported by NSF CAREER award #1350965.

References

[1] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

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

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

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

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

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

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

[8] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

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

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

[11] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[12] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2044–2052, 2015.

[13] Y. Gani, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *arXiv preprint arXiv:1505.07818*, 2015.

[14] A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

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

[16] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[17] S. Grunewalder, G. Arthur, and J. Shawe-Taylor. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1184–1192, 2013.

[18] K. Guittet. Extended Kantorovich norms: a tool for optimization. 2002.

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

[20] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. *arXiv preprint arXiv:1605.03661*, 2016.

[21] M. Kuang and E. Tabak. Preconditioning of optimal transport. *Preprint*, 2016.

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

[23] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

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

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

[26] P. R. Rosenbaum. *Observational studies*. Springer, 2002.

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

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

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

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

[31] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

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

[33] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2015.

A Proofs

A.1 Definitions, assumptions, and auxiliary lemmas

Notation:
$p^F(x,t)$, $p^{CF}(x,t)$: factual and counterfactual distributions on $\mathcal{X} \times \{0,1\}$
$u = p^F(t=1)$: the marginal probability of treatment.
$p^{t=1}(x) = p^F(x|t=1)$: treated distribution. $p^{t=0}(x) = p^F(x|t=0)$: control distribution.
Φ: representation function mapping from $\mathcal{X}$ to $\mathcal{R}$.
Ψ: the inverse function of Φ, mapping from $\mathcal{R}$ to $\mathcal{X}$.
$p^F_\Phi(r,t)$, $p^{CF}_\Phi(r,t)$: factual and counterfactual distributions induced by Φ on $\mathcal{R} \times \{0,1\}$.
$p^{t=1}_\Phi(r)$, $p^{t=0}_\Phi(r)$: treated and control distributions induced by Φ on $\mathcal{R}$.
$Y_0(x)$, $Y_1(x)$: true labeling functions for $t=0$, $t=1$.
$L(\cdot,\cdot)$: loss function, from $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R}_+$.
$\ell_F(h,\Phi)$, $\ell_{CF}(h,\Phi)$: expected loss of $h(\Phi(x),t)$ with respect to the factual, counterfactual distributions.
$\mathrm{IPM}_\mathrm{F}(p,q)$: the integral probability metric distance induced by function family F between distributions p and q.
$\mathrm{UIPM}_\mathrm{F}(u_p \cdot p, u_q \cdot q)$: the unnormalized integral probability metric distance induced by function family F between distributions p and q scaled respectively by $u_p, u_q \in \mathbb{R}_+$

We first define the necessary distributions and prove some simple results about them.

Definition 4. *Let the factual and counterfactual distributions $p^F(x,t)$ and $p^{CF}(x,t)$ be distributions over $\mathcal{X} \times \{0,1\}$ such that $p^F(t|x) = p^{CF}(1-t|x)$, where $\mathcal{X} \subset \mathbb{R}^d$.*

Assumption 4. *The marginal distribution $0 < p^F(t=1) < 1$.*

Assumption 5. *The marginal distribution of the covariates x is equal for the factual and counterfactual distributions: $p^F(x) = p^{CF}(x)$.*

Definition 5. *Let $p^{t=1}(x) \equiv p^F(x|t=1)$, and $p^{t=0}(x) \equiv p^F(x|t=0)$ 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 6. *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 6. *For a representation function $\Phi : \mathcal{X} \to \mathcal{R}$, and for a distribution p^F defined over $\mathcal{X}$, let p^F_Φ be the distribution induced by Φ over $\mathcal{R}$. Define p^{CF}_Φ analogously. Also define $p^{t=1}_\Phi(r) \equiv p^F_\Phi(r|t=1)$, $p^{t=0}_\Phi(r) \equiv p^F_\Phi(r|t=0)$, to be the treatment and control distributions induced over $\mathcal{R}$.*

For a one-to-one Φ, the distributions p^F_Φ and p^{CF}_Φ can be obtained by the standard change of variables formula, using the determinant of the Jacobian of $\Psi(r)$. See [3] for the case of a mapping Φ between spaces of different dimensions.

Lemma 1. $p^F_\Phi(r) = p^{CF}_\Phi(r)$

Proof. Let $J_\Psi(r)$ be the absolute of the determinant of the Jacobian of $\Psi(r)$. Then by the change of variable formula we have:

$$p^F_\Phi(r) = J_\Psi(r) \cdot p^F(\Psi(r)) = J_\Psi(r) \cdot p^{CF}(\Psi(r)) = p^{CF}_\Phi(r),$$

since by Assumption 5 $p^F(\Psi(r)) = p^{CF}(\Psi(r))$. □

Lemma 2. *For all $r \in \mathcal{R}$, $t \in \{0, 1\}$:*

$$p^F_\Phi(t|r) = p^F(t|\Psi(r)), \quad p^{CF}_\Phi(t|r) = p^{CF}(t|\Psi(r)$$

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

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

where equality (a) is by the change of variable formula. The proof is identical for p^{CF}_Φ. □

Lemma 3. *For all $r \in \mathcal{R}$:*

$$p^{CF}_\Phi(r,t) = p^F_\Phi(r|1-t)p^F_\Phi(1-t)$$

Proof.

$$p^{CF}_\Phi(r,t) = p^{CF}_\Phi(r)p^{CF}_\Phi(t|r) = p^F_\Phi(r)p^F_\Phi(1-t|r),$$

since by Lemma 1 we have $p^{CF}_\Phi(r) = p^F_\Phi(r)$, and by Definition 4 and Lemma 2 □

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

Definition 7. *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 $\mathcal{R}$. Let $Y_t : \mathcal{X} \to \mathbb{R}$ be the true labeling functions, for $t = 0, 1$. The expected factual loss and counterfactual losses of h and Φ are, respectively:*

$$\ell_F(h,\Phi) = \int_{\mathcal{R}\times\{0,1\}} L\left(Y_t(\Psi(r)), h(r,t)\right) p^F_\Phi(r,t)\, dr dt$$

$$\ell_{CF}(h,\Phi) = \int_{\mathcal{R}\times\{0,1\}} L\left(Y_t(\Psi(r)), h(r,t)\right) p^{CF}_\Phi(r,t)\, dr dt,$$

where Ψ is the inverse function of Φ.

Definition 8. *Let F 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*:*

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

$\mathrm{IPM}_\mathrm{F}(\cdot,\cdot)$ defines a pseudo-metric on the space of probability functions over $\mathcal{S}$, and for sufficiently large function families, $\mathrm{IPM}_\mathrm{F}(\cdot,\cdot)$ is a proper metric [24]. 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 [15, 30]. We note that for function families F such as the three mentioned above, for which $g \in \mathrm{F} \implies -g \in \mathrm{F}$, the absolute value can be omitted from definition 8.

Definition 9. *Let* F *be a function family consisting of functions* $g : \mathcal{S} \to \mathbb{R}$*. For a pair of distributions* p_1*,* p_2 *over* $\mathcal{S}$*, and a pair of positive scalar* u_1, u_2*, define the* Unnormalized Integral Probability Metric*:*

$$\textit{UIPM}_{\mathrm{F}}(u_1 p_1, u_2 p_2) = \sup_{g \in \mathrm{F}} \left| \int_{\mathcal{S}} g(s) \left(u_1 p_1(s) - u_2 p_2(s)\right) \, ds \right|$$

$\mathrm{UIPM}_{\mathrm{F}}$ can be considered as a distance function between positive measures on $\mathcal{S}$, see also [18, 14, 6].

Lemma 4. *For constant* $v \geq 0$*, we have that* $\textit{UIPM}_{\mathrm{F}}(v \cdot u_1 p_1, v \cdot u_2 p_2) = v \cdot \textit{UIPM}_{\mathrm{F}}(u_1 p_1, u_2 p_2)$*. In particular,* $\textit{UIPM}_{\mathrm{F}}(v p_1, v p_2) = v \cdot \textit{IPM}_{\mathrm{F}}(p_1, p_2)$*.*

The proof is immediate from Definitions 8 and 9.

A.2 General IPM bound

Lemma 5. *Let* $u = p^F(t=1)$*. For a function* $g : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ *and the distributions* p^{CF}_Φ*,* p^F_Φ*:*

$$\int_{\mathcal{R}\times\{0,1\}} g(r,t) \left(p^{CF}_\Phi(r,t) - p^F_\Phi(r,t)\right) \, dr dt =$$
$$\int_{\mathcal{R}} g(r,0) \left(u \cdot p^{t=1}_\Phi(r) - (1-u) \cdot p^{t=0}_\Phi(r)\right) \, dr + \int_{\mathcal{R}} g(r,1) \left((1-u) \cdot p^{t=0}_\Phi(r) - u \cdot p^{t=1}_\Phi(r)\right) \, dr$$

Proof.

$$\int_{\mathcal{R}\times\{0,1\}} g(r,t) \left(p^{CF}_\Phi(r,t) - p^F_\Phi(r,t)\right) \, dr dt =$$
$$\int_{\mathcal{R}} g(r,0) \left(p^{CF}_\Phi(r,0) - p^F_\Phi(r,0)\right) \, dr + \int_{\mathcal{R}} g(r,1) \left(p^{CF}_\Phi(r,1) - p^F_\Phi(r,1)\right) \, dr \overset{(a)}{=} \tag{4}$$
$$\int_{\mathcal{R}} g(r,0) \left(u \cdot p^F_\Phi(r|t=1) - (1-u) \cdot p^F_\Phi(r|t=0)\right) \, dr +$$
$$\int_{\mathcal{R}} g(r,1) \left((1-u) \cdot p^F_\Phi(r|t=0) - u \cdot p^F_\Phi(r|t=1)\right) \, dr =$$
$$\int_{\mathcal{R}} g(r,0) \left(u \cdot p^{t=1}_\Phi(r) - (1-u) \cdot p^{t=0}_\Phi(r)\right) \, dr + \int_{\mathcal{R}} g(r,1) \left((1-u) \cdot p^{t=0}_\Phi(r) - u \cdot p^{t=1}_\Phi(r)\right) \, dr,$$

where equality (a) is due to Lemma 3 and since $p^F_\Phi(t) = p^F(t)$. □

We prove a very slightly more general version of Theorem 1 from the main paper. The only difference being that we allow a scaled version of $L(Y_t(\Psi(r)), h(r,t))$ to be in the function family F. This modification simplifies the following proofs.

Theorem 3. *Let* $p^{t=1}_\Phi, p^{t=0}_\Phi$ *be defined as in Definition 6. Let* $u = p^F(t=1)$*. Let* F *be a family of functions* $f : \mathcal{R} \to \mathbb{R}$*, and denote by* $\textit{UIPM}_{\mathrm{F}}(\cdot,\cdot)$ *the unnormalized integral probability metric induced by* F*. Let* Y_0, Y_1 *be the true labeling functions. Let* $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ *be an hypothesis. Let* $\Phi : \mathcal{X} \to \mathcal{R}$ *be a one-to-one representation function, with inverse* Ψ*. Assume there exists a constant* $B > 0$*, such that for* $t = 0, 1$*, we have that* $f_{\Phi,h}(r) := \frac{1}{B} \cdot L(Y_t(\Psi(r)), h(r,t)) \in \mathrm{F}$*. Then we have:*

$$\ell_{CF}(h, \Phi) \leq \ell_F(h, \Phi) + 2B \cdot \textit{UIPM}\left(u \cdot p^{t=1}_\Phi, (1-u) \cdot p^{t=0}_\Phi\right). \tag{5}$$

Proof.

$$
\begin{aligned}
&\ell_{CF}(h,\Phi) = \\
&\ell_F(h,\Phi) + (\ell_{CF}(h,\Phi) - \ell_F(h,\Phi)) = \\
&\ell_F(h,\Phi) + \int_{\mathcal{R}\times\{0,1\}} L\left(Y_t(\Psi(r)), h(r,t)\right)\left(p_\Phi^{CF}(r,t) - p_\Phi^F(r,t)\right)\, dr dt = \qquad (6)\\
&\ell_F(h,\Phi) + \int_{\mathcal{R}} L\left(Y_0(\Psi(r)), h(r,0)\right)\left(u\cdot p_\Phi^{t=1}(r) - (1-u)\cdot p_\Phi^{t=0}(r)\right)\, dr \\
&\qquad + \int_{\mathcal{R}} L\left(Y_1(\Psi(r)), h(r,1)\right)\left((1-u)\cdot p_\Phi^{t=0}(r) - u\cdot p_\Phi^{t=1}(r)\right)\, dr = \\
&\ell_F(h,\Phi) + B\cdot\int_{\mathcal{R}} \frac{1}{B} L\left(Y_0(\Psi(r)), h(r,0)\right)\left(u\cdot p_\Phi^{t=1}(r) - (1-u)\cdot p_\Phi^{t=0}(r)\right)\, dr \\
&\qquad + B\cdot\int_{\mathcal{R}} \frac{1}{B} L\left(Y_1(\Psi(r)), h(r,1)\right)\left((1-u)\cdot p_\Phi^{t=0}(r) - u\cdot p_\Phi^{t=1}(r)\right)\, dr \leq \qquad (7)\\
&\ell_F(h,\Phi) + B\cdot\sup_{g\in \mathrm{F}}\left|\int_{\mathcal{R}} g(r)\left(u\cdot p_\Phi^{t=1}(r) - (1-u)\cdot p_\Phi^{t=0}(r)\right)\, dr\right| \\
&\qquad + B\cdot\sup_{g\in \mathrm{F}}\left|\int_{\mathcal{R}} g(r)\left((1-u)\cdot p_\Phi^{t=0}(r) - u\cdot p_\Phi^{t=1}(r)\right)\, dr\right| = \\
&\ell_F(h,\Phi) + 2B\cdot \mathrm{UIPM}\left(u\cdot p_\Phi^{t=1}, (1-u)\cdot p_\Phi^{t=0}\right),
\end{aligned}
$$

where equality (6) is by Lemma 5, and inequality (7) is by the assumption that $L(Y_t(\Psi(r)), h(r,t)) \in \mathrm{F}$ for $t = 0, 1$. □

We also have an immediate corollary:

Corollary 1. *Under the conditions of Theorem 3 above, if $p^F(t=1) = \frac{1}{2}$, then:*

$$\ell_{CF}(h,\Phi) \leq \ell_F(h,\Phi) + \mathit{IPM}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right). \qquad (8)$$

Proof. The result follows using Lemma 4. □

The essential point in the proof of Theorem 3 is inequality 7. Note that on the l.h.s. of the inequality, we need to evaluate the expectations of $L\left(Y_0(\Psi(r)), h(r,0)\right)$ over $p^{t=1}$ and $L\left(Y_1(\Psi(r)), h(r,1)\right)$ over $p^{t=0}$. 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 $L\left(Y_0(\Psi(r)), h(r,0)\right)$ and $L\left(Y_1(\Psi(r)), h(r,1)\right)$. The upper bound ignores the outcome information, and amounts to measuring a distance between two distributions we have samples from: the control and treated distribution. Note that for a randomized trial with $p(t=1) = \frac{1}{2}$, we have that $\mathrm{IPM}(p_\Phi^{t=1}, p_\Phi^{t=0}) = 0$, and indeed in that case $\ell_{CF}(h,\Phi) = \ell_F(h,\Phi)$.

The crucial condition in Theorem 3 is that the function $f_{\Phi,h}(r) := L\left(Y_t(\Psi(r)), h(r,t)\right)$ is in F. We now look into two specific function families F, and evaluate what does this inclusion condition entail.

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 Y_0, Y_1 and the loss L are differentiable. However, this assumption could be relaxed to a mere Lipschitzness assumption.

Assumption 7. *There exists a constant $K > 0$ such that for all $x \in \mathcal{X}$, $t \in \{0,1\}$, $\|\frac{\partial Y_t(x)}{\partial x}\| \leq K$.*

Assumption 8. *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 $i = 1, 2$.*

Loss functions which obey Assumption 8 include the log-loss, hinge-loss, absolute loss, and for compact $\mathcal{Y}$ also the squared loss.

When we let F in Definition 8 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 [32].

The Wasserstein distance has been extended to unnormalized distributions (i.e. positive measures) by several authors. We use the formulation of [18], who has shown how to reduce the problem of Wasserstein distance between unnormalized distributions into the ordinary Wasserstein distance, by adding a point "at infinity". This has been followed through by [14], and we use the latter's algorithmic framework in the experimental section of our paper. For a much deeper mathematical treatment of the subject, we refer the reader to [6].

Definition 10. *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 11. *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 6. *Assume that Φ is a Jacobian-normalized representation, and let Ψ be its inverse. Define $\tilde{Y}_t : \mathcal{R} \to \mathbb{R}$ by $\tilde{Y}_t(r) \equiv Y_t(\Psi(r))$. The Lipschitz constant of $\tilde{Y}_t$ is bounded by $\rho(\Phi)K$, where K is from Assumption 7, and $\rho(\Phi)$ as in Definition 10.*

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 $x = \Psi(r)$:

$$\sigma_{max}\left(\frac{\partial\Psi(r)}{\partial r}\right) = \frac{1}{\sigma_{min}\left(\frac{\partial\Phi(x)}{\partial x}\right)}, \tag{9}$$

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:

$$\|\frac{\partial\tilde{Y}_t(r)}{\partial r}\| = \|\frac{\partial Y_t(\Psi(r))}{\partial r}\| = \|\frac{\partial Y_t(\Psi(r))}{\partial\Psi(r)}\frac{\partial\Psi(r)}{\partial r}\| \leq \tag{10}$$

$$\|\frac{\partial\Psi(r)}{\partial r}\|\|\frac{\partial Y_t(\Psi(r))}{\partial\Psi(r)}\| = \tag{11}$$

$$\frac{1}{\sigma_{min}\left(\frac{\partial\Phi(x)}{\partial x}\right)}\|\frac{\partial Y_t(x)}{\partial x}\| \leq \tag{12}$$

$$\frac{K}{\sigma_{min}\left(\frac{\partial\Phi(x)}{\partial x}\right)} \leq \rho(\Phi)K, \tag{13}$$

where inequality (10) is by the matrix norm inequality, equality (11) is by (9), inequality (12) is by assumption 7 on the norms of the gradient of $Y_t(x)$ w.r.t x , and inequality (13) is by Definition 10 of $\rho(\Phi)$, the assumption that Φ is Jacobian-normalized, and noting that singular values are necessarily non-negative.

□

Lemma 7. *Under the conditions of Theorem 3, further assume that Y_0, Y_1 have gradients bounded by K as in 7, 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 $L\left(Y_t(\Psi(r)), h(r,t)\right)$ is upper bounded by $K_L \cdot K\left(\rho(\Phi)+b\right)$ for $t=0,1$.*

Proof. Using the chain rule, we have that:

$$\frac{\partial L}{\partial r} = \frac{\partial L}{\partial Y_t(\Psi(r))}\frac{\partial Y_t(\Psi(r))}{\partial r} + \frac{\partial L}{\partial h(r,t)}\frac{\partial h(r,t)}{\partial r} \leq$$
$$K_l K \rho(\Phi) + K_l \cdot bK = K_L \cdot K\left(\rho(\Phi)+b\right),$$

where the inequality is due to Lemma 6. □

Theorem 4. *Let $u = p^F(t=1)$ 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 Y_0, Y_1 on $\mathcal{X}$. Let K_L be the Lipschitz constant of the loss function L. Let $h : \mathcal{R} \times \{0,1\} \to \mathbb{R}$ be an hypothesis with Lipschitz constant bK. Then:*

$$\ell_{CF}(h,\Phi) \leq \ell_F(h,\Phi) + 2\left(\rho(\Phi)+b\right) \cdot K \cdot K_L \cdot \textit{Wass}(u \cdot p_\Phi^{t=1}, (1-u)\cdot p_\Phi^{t=0}). \tag{14}$$

Proof. We will apply Theorem 3 with $\mathrm{F} = \{f : \mathcal{R} \to \mathbb{R} \text{ s.t. } f \text{ is 1-Lipschitz}\}$. By Lemma 7, we have that for $B = \left(\rho(\Phi)+b\right) \cdot K \cdot K_L$, the function $\frac{1}{B} L\left(Y_t(\Psi(r)), h(r,t)\right) \in \mathrm{F}$. Inequality (14) then holds as a special case of Theorem 3. □

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 Theorem 4, 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 $\rho(\Phi)$.

Lemma 8. *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:*

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

Proof. Following [32, 21], 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, $\mathcal{M}_{p,q} = \{M : \mathcal{X} \to \mathcal{X} \text{ s.t. } q(M(S)) = p(S) \text{ for all measurable bounded } S \subset \mathcal{X}\}$. We then have that:

$$\mathrm{Wass}(p,q) = \inf_{M \in \mathcal{M}_{p,q}} \int_{\mathcal{X}} \|M(x) - x\| p(x)\, dx. \tag{15}$$

It is known that the infimum in (15) is actually achievable [32, 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 Theorem 3 is that $L\left(Y_t(\Psi(r)), h(r,t)\right) \in \mathrm{F}$. The function space F we use here is $\mathrm{F} = \{g \in \mathcal{H}_r \text{ s.t. } \|g\|_{\mathcal{H}_r} \leq 1\}$.

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

Assumption 9. *There exist $f_1^Y, f_2^Y \in \mathcal{H}_x$ such that $Y_t(x) = \left\langle f_t^Y, k_x(x,\cdot)\right\rangle_{\mathcal{H}_x}$, i.e. the true labeling functions Y_0, Y_1 are in $\mathcal{H}_x$. Further assume that $\|f_t^Y\|_{\mathcal{H}_x} \leq K$.*

Assumption 10. *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 9 and 10 about Y_0, Y_1, and Φ, we have that $Y_t(\Psi(r)) = \left\langle \Gamma_\Phi^* f_t^Y, k_r(r, \cdot)\right\rangle_{\mathcal{H}_r}$, where Γ_Φ^* is the adjoint operator of Γ_Φ [17].

Lemma 9. *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 $\|f_t^h\|_{\mathcal{H}_r} \leq b$. Under Assumption 9 about Y_0, Y_1, we have that $L\left(Y_t(\Psi(r)), h(r, t)\right) = \left(Y_t(\Psi(r)) - h(r, t)\right)^2$ is in the tensor Hilbert space $\mathcal{H}_r \otimes \mathcal{H}_r$. Moreover, the norm of $\left(Y_t(\Psi(r)) - h(r, t)\right)^2$ in $\mathcal{H}_r \otimes \mathcal{H}_r$ is upper bounded by $4\left(K_\Phi^2 K^2 + b^2\right)$.*

Proof. By linearity of the Hilbert space, we have that $Y_t(\Psi(r)) - h(r, t) = \left\langle \Gamma_\Phi^* f_t^Y, k_r(r, \cdot)\right\rangle_{\mathcal{H}_r} - \left\langle f_t^h, k_r(r, \cdot)\right\rangle_{\mathcal{H}_r} = \left\langle \Gamma_\Phi^* f_t^Y - f_t^h, k_r(r, \cdot)\right\rangle_{\mathcal{H}_r}$. By a well known result [31, Theorem 7.25], the product $\left(Y_t(\Psi(r)) - h(r, t)\right) \cdot \left(Y_t(\Psi(r)) - h(r, t)\right)$ 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. Now we have that:

$$\|\Gamma_\Phi^* f_t^Y - f_t^h\|_{\mathcal{H}_r}^2 \leq \tag{16}$$

$$2\|\Gamma_\Phi^* f_t^Y\|_{\mathcal{H}_r}^2 + 2\|f_t^h\|_{\mathcal{H}_r}^2 \leq \tag{17}$$

$$2\|\Gamma_\Phi^*\|_{HS}^2 \|f_t^Y\|_{\mathcal{H}_x}^2 + 2\|f_t^h\|_{\mathcal{H}_r}^2 = \tag{18}$$

$$2\|\Gamma_\Phi\|_{HS}^2 \|f_t^Y\|_{\mathcal{H}_x}^2 + 2\|f_t^h\|_{\mathcal{H}_r}^2 leq \tag{19}$$

$$2K_\Phi^2 K^2 + 2b^2. \tag{20}$$

Inequality (16) 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 (17) is by the definition of the operator norm. Equality (18) 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 (19) is by Assumptions 9 and 10, and by the premise on the norm of f_t^h. □

Theorem 5. *Let $u = p^F(t = 1)$ be the marginal probability of treatment, and assume $0 < u < 1$. Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one representation function which obeys Assumption 10 with corresponding operator Γ_Φ with operator norm K_Φ. Let the functions Y_0, Y_1 obey Assumption 9, 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 ℓ_F and ℓ_{CF} are defined with respect to L being the squared loss. Then:*

$$\ell_{CF}(h, \Phi) \leq \ell_F(h, \Phi) + 4\left(K_\Phi^2 K^2 + b^2\right) \cdot \mathit{MMD}(u \cdot p_\Phi^{t=1}, (1 - u) \cdot p_\Phi^{t=0}), \tag{21}$$

where ℓ_{CF} and ℓ_F use the squared loss.

Proof. We will apply Theorem 3 with $\mathrm{F} = f \in \mathcal{H}_r \otimes \mathcal{H}_r$ s.t. $\|f\|_{\mathcal{H}_r \otimes \mathcal{H}_r} \leq 1$. By Lemma 9, we have that for $B = 2\left(K_\Phi^2 K^2 + b^2\right)$ and L being the squared loss, $\frac{1}{B} L\left(Y_t(\Psi(r)), h(r, t)\right) \in \mathrm{F}$. Inequality (21) then holds as a special case of Theorem 3. □

B Treatment effect estimation loss

Definition 12. *The treatment effect for unit x is:*

$$\tau(x) = Y_1(x) - Y_0(x).$$

We now show that minimizing the counterfactual loss ℓ_{CF} is closely tied to minimizing the error in estimating the individualized treatment effect $\tau(x) = Y_1(x) - Y_0(x)$. We note that the statements in this subsection do not specifically rely on the representation learning aspect of our work, and could also be phrased without using the representation function Φ.

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

Definition 13. *The inductive treatment effect estimate for unit x is:*

$$\hat{\tau}_g(x) = g(x, 1) - g(x, 0).$$

Definition 14. *The transductive treatment effect estimate for unit x_i with treatment assignment t_i is:*

$$\tau'_g(x_i, t_i) = \begin{cases} Y_1(x_i) - g(x_1, 0) & \text{if } t_i = 1 \\ g(x_i, 1) - Y_0(x_i) & \text{if } t_i = 0 \end{cases}$$

Definition 15. *Let $Y_t : \mathcal{X} \to \mathbb{R}$ be the true labeling functions for $t = 0, 1$. The expected Precision in Estimation of Heterogeneous Effect (PEHE) loss of g is:*

$$\ell_{\textit{PEHE}}(g) = \int_{\mathcal{X}} |\hat{\tau}_g(x) - \tau(x)|^a \; p(x)\, dx$$

The expected Individualized Treatment Effect (ITE) loss of g is:

$$\ell_{\textit{ITE}}(g) = \int_{\mathcal{X} \times \{0,1\}} \left|\tau'_g(x, t) - \tau(x)\right|^a \; p^F(x, t)\, dxdt$$

where $a = 1$ corresponds to the absolute loss, and $a = 2$ corresponds to the squared loss.

Note that the transductive treatment effect and the corresponding ℓ_{ITE} loss are defined with respect to the treatment assignment, whereas the inductive treatment effect and the corresponding ℓ_{PEHE} loss are defined irrespective of the treatment assignment.

We show that for the transductive case where we assume that for a unit x_i either $Y_1(x_i)$ or $Y_0(x_i)$ is known, minimizing ℓ_{CF} with a squared loss is equivalent to minimizing $\mathbb{E}_{x \sim \hat{p}(x)} \left[(\tau'(x) - \tau(x))^2\right]$. For the inductive case, we show that $\mathbb{E}_{x \sim p(x)} \left[(\hat{\tau}(x) - \tau(x))^2\right]$ is upper bounded by $2\ell_F + 2\ell_{CF}$ where ℓ_F and ℓ_{CF} are w.r.t. to the squared loss, while $\mathbb{E}_{x \sim p(x)} \left[|\hat{\tau}(x) - \tau(x)|\right]$ is upper bounded by $\ell_F + \ell_{CF}$ where ℓ_F and ℓ_{CF} are w.r.t. to the absolute loss.

The intuition for both results is simple: for the transductive case, the only unknown quantity are the counterfactual outcomes, which we seek to know. For the inductive case, we incur an additional error on top of the transductive case. This error comes from inferring the unknown factual outcomes for new samples, which is the classic machine learning inductive learning scenario.

Lemma 10. *For an hypothesis $g : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ such that $g(x, t) = h(\Phi(x), t)$, and using the squared loss or absolute loss, we have:*

$$\ell_{\textit{ITE}}(g) = \ell_{CF}(h, \Phi).$$

Proof.

$$\ell_{\text{ITE}}(g) = \tag{22}$$

$$\int_{\mathcal{X}} |(Y_1(x) - g(x, 0)) - (Y_1(x) - Y_0(x))|^a \; p^F(x, t = 1)\, dx$$

$$+ \int_{\mathcal{X}} |(g(x, 1) - Y_0(x)) - (Y_1(x) - Y_0(x))|^a \; p^F(x, t = 0)\, dx =$$

$$\int_{\mathcal{X}} |Y_0(x) - g(x, 0)|^a \; p^F(x, t = 1) dx + \int_{\mathcal{X}} |g(x, 1) - (Y_1(x)|^a \; p^F(x, t = 0)\, dx = \tag{23}$$

$$\int_{\mathcal{X}} |Y_0(\Psi(r)) - h(r, 0)|^a \; p^F_\Phi(r, t = 1)\, dr + \int_{\mathcal{R}} |h(r, 1) - (Y_1(\Psi(r))|^a \; p^F_\Phi(r, t = 0)\, dx =$$

$$\int_{\mathcal{X}} |Y_0(\Psi(r)) - h(r, 0)|^a \; p^{CF}_\Phi(r, t = 0)\, dr + \int_{\mathcal{R}} |h(r, 1) - (Y_1(\Psi(r))|^a \; p^F_\Phi(r, t = 0)\, dx, =$$

$$\ell_{CF}(h, \Phi)$$

Algorithm 2 Counterfactual balanced regression with integral probability metrics

Input: Factual sample $(x_1, t_1, y_1^F), \ldots, (x_n, t_n, y_n^F)$, scaling parameter $\alpha > 0$, loss function $L(\cdot, \cdot)$, representation network $\Phi_{\mathbf{W}}$ with initial weights by $\mathbf{W}$, outcome network $h_{\mathbf{V}}$ with intial weights $\mathbf{V}$, function family F for IPM loss
while not converged **do**
  Sample a mini-batch with m' control and m treated units $(x_{i_1}, 0, y_{i_1}^F), \ldots, (x_{i_m}, 0, y_{i_m}^F), (x_{i_{m+1}}, 1, y_{i_{m+1}}^F), \ldots, (x_{i_{m+m'}}, 1, y_{i_{m+m'}}^F)$
  Calculate the gradient of the imbalance penalty: $g_1 = \nabla_{\mathbf{W}} \text{ IPM}_F \left(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^{m}, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m'} \right)$
  Calculate the gradients of the empirical loss: $g_2 = \nabla_{\mathbf{V}} \frac{1}{m+m'} \sum_{j=1}^{m+m'} L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}^F \right)$ $g_3 = \nabla_{\mathbf{W}} \frac{1}{m+m'} \sum_{j=1}^{m+m'} L\left(h_{\mathbf{V}}(\Phi_{\mathbf{W}}(x_{i_j}), t_{i_j}), y_{i_j}^F \right)$
  Obtain step size scalar or matrix η with standard neural net methods e.g. RMSProp
  update $\mathbf{W} \leftarrow \mathbf{W} - \eta(\alpha \mathbf{g_1} + \mathbf{g_3})$ and $\mathbf{V} \leftarrow \mathbf{V} - \eta(\mathbf{g_2})$
  check convergence criterion
end while

where (22) is by decomposing τ_g' over $t = 0, 1$, (23) is by the change of variable formula and definition of p_Φ^F, and Ψ is the inverse function of Φ. □

Lemma 11. *For an hypothesis $g : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ such that $g(x, t) = h(\Phi(x), t)$, and using the loss $L(y_1, y_2) = |y_1 - y_2|^a$ for $a = 1, 2$, we have:*

$$\ell_{PEHE}(g) \leq a \cdot \left(\ell_{CF}(h, \Phi) + \ell_F(h, \Phi) \right).$$

Proof. For $a = 1, 2$:

$$\begin{aligned}
&\ell_{\text{PEHE}}(g) = \\
&\int_{\mathcal{X}} |(g(x,1) - g(x,0)) - (Y_1(x) - Y_0(x))|^a \, p(x) \, dx = \\
&\int_{\mathcal{X}} |(g(x,1) - Y_1(x)) + (Y_0(x) - g(x,0))|^a \, p(x) \, dx \leq && (24) \\
&a \cdot \int_{\mathcal{X}} |(g(x,1) - Y_1(x))|^a + |(Y_0(x) - g(x,0))|^a \, p(x) \, dx = && (25) \\
&a \int_{\mathcal{X}} |(g(x,1) - Y_1(x))|^a + |(Y_0(x) - g(x,0))|^a \, p^F(x, t=0) \, dx \\
&\quad + a \int_{\mathcal{X}} |(g(x,1) - Y_1(x))|^a + |(Y_0(x) - g(x,0))|^a \, p^F(x, t=1) \, dx = \\
&a \int_{\mathcal{X}} |g(x,1) - Y_1(x)|^a + |(Y_0(x) - g(x,0))|^a \, p^{CF}(x, t=1) \, dx \\
&\int_{\mathcal{X}} |g(x,1) - Y_1(x)|^a + |(Y_0(x) - g(x,0))|^a \, p^F(x, t=1) \, dx = && (26) \\
&a \left(\ell_{CF}(h, \Phi) + \ell_F(h, \Phi) \right),
\end{aligned}$$

where (24) is by the inequality $|x + y|^a \leq a(|x|^a + |y|^a)$ for $a = 1, 2$, (25) is because $p(x) = p(x, t = 0) + p(x, t = 1)$, and (26) follows from the same change of variable substitution made in (23) in Lemma 10 above. □

C Algorithmic details

We give details about the algorithms used in our framework. First, we restate Algorithm 1.

Algorithm 3 Computing the stochastic gradient of the Wasserstein distance

1: **Input:** Factual $(x_1, t_1, y_1^F), \ldots, (x_n, t_n, y_n^F)$, representation network $\Phi_{\mathbf{W}}$ with current weights by $\mathbf{W}$
2: Randomly sample a mini-batch with m treated and m' control units $(x_{i_1}, 0, y_{i_1}^F), \ldots, (x_{i_m}, 0, y_{i_m}^F), (x_{i_{m+1}}, 1, y_{i_{m+1}}^F), \ldots, (x_{i_{2m}}, 1, y_{i_{2m}}^F)$
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 [9], with input $M(\Phi_{\mathbf{W}})$
5: Calculate the gradient: $g_1 = \nabla_{\mathbf{W}} \langle T^*, M(\Phi_{\mathbf{W}}) \rangle$

C.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 citecuturi2013sinkhorn 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 [8], the approximation is computed using a fixed-point iteration involving repeated multiplication with a kernel matrix K. We can use the algorithm of [8] in our framework. See Algorithm 3 for an overview of how to compute the gradient g_1 in Algorithm 2. When computing g_1, disregarding the gradient $\nabla_{\mathbf{W}} T^*$ amounts to minimizing an upper bound on the Sinkhorn transport.

For non-uniform populations $p(t) \neq 1/2$, our method can still be applied using the unnormalized version of the Wasserstein distance, using ideas presented by [18, 14], and explored much more deeply by [6]. The Sinkhorn transport can still be used to approximate the Wasserstein distance, but with small modifications. Let λ and δ be parameters and $\tilde{M}$ the matrix

$$\tilde{M} = \begin{bmatrix} M & \delta \\ \delta & 0 \end{bmatrix} .$$

Then, define an $n_t + 1$-dimensional vector a

$$a = [p(t), ..., p(t), 1 - p(t)]^\top$$

and an $n_c + 1$-dimensional vector b

$$b = [1 - p(t), ..., 1 - p(t), p(t)]^\top$$

where n_t and n_c are the number of treated and controls. Then, to get T^*, apply Algorithm 3 of [9] on $\tilde{M}, a$ and b.

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.

C.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,

$$\text{MMD}_k(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^{m}, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m'}) = \tag{27}$$

$$\frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{k=1,k\neq j}^{m} k(\Phi_{\mathbf{W}}(x_{i_j}), \Phi_{\mathbf{W}}(x_{i_k})) \tag{28}$$

$$+\frac{2}{mm'} \sum_{j=1}^{m} \sum_{k=m}^{m+m'} k(\Phi_{\mathbf{W}}(x_{i_j}), \Phi_{\mathbf{W}}(x_{i_k})) \tag{29}$$

$$+\frac{1}{m'(1-m')} \sum_{j=1}^{m} \sum_{k=m,k\neq j}^{m'} k(\Phi_{\mathbf{W}}(x_{i_j}), \Phi_{\mathbf{W}}(x_{i_k})) \tag{30}$$

The unnormalized version, where the marginal probability of treatment $u \neq 1/2$, is

$$\text{MMD}_k(\{\Phi_{\mathbf{W}}(x_{i_j})\}_{j=1}^{m}, \{\Phi_{\mathbf{W}}(x_{i_k})\}_{k=m+1}^{m'}) = \tag{31}$$

$$\frac{2p(t)^2}{m(m-1)} \sum_{j=1}^{m} \sum_{k=1,k\neq j}^{m} k(\Phi_{\mathbf{W}}(x_{i_j}), \Phi_{\mathbf{W}}(x_{i_k})) \tag{32}$$

$$+\frac{4p(t)(1-p(t))}{mm'} \sum_{j=1}^{m} \sum_{k=m}^{m+m'} k(\Phi_{\mathbf{W}}(x_{i_j}), \Phi_{\mathbf{W}}(x_{i_k})) \tag{33}$$

$$+\frac{2(1-p(t))^2}{m'(1-m')} \sum_{j=1}^{m} \sum_{k=m,k\neq j}^{m'} k(\Phi_{\mathbf{W}}(x_{i_j}), \Phi_{\mathbf{W}}(x_{i_k})) \tag{34}$$

The (unnormalized) linear maximum-mean discrepancy can be written as a distance between means. In the notation of Algorithm 2,

$$\text{MMD} = 2 \left\| \frac{1}{m} \sum_{j=1}^{m} p(t)\Phi_{\mathbf{W}}(x_{i_j}) - (1-p(t))\frac{1}{m'} \sum_{k=m+1}^{m'} \Phi_{\mathbf{W}}(x_{i_k}) \right\|_2$$

Let

$$\mathbf{f}(\mathbf{W}) = p(t)\frac{1}{m} \sum_{j=1}^{m} \Phi_{\mathbf{W}}(x_{i_j}) - (1-p(t))\frac{1}{m'} \sum_{k=m+1}^{m'} \Phi_{\mathbf{W}}(x_{i_k})$$

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

$$g_1 = 2\frac{d\mathbf{f}(\mathbf{W})}{d\mathbf{W}} \frac{\mathbf{f}(\mathbf{W})}{\|\mathbf{f}(\mathbf{W})\|_2} .$$

D Empirical results

D.1 General observation

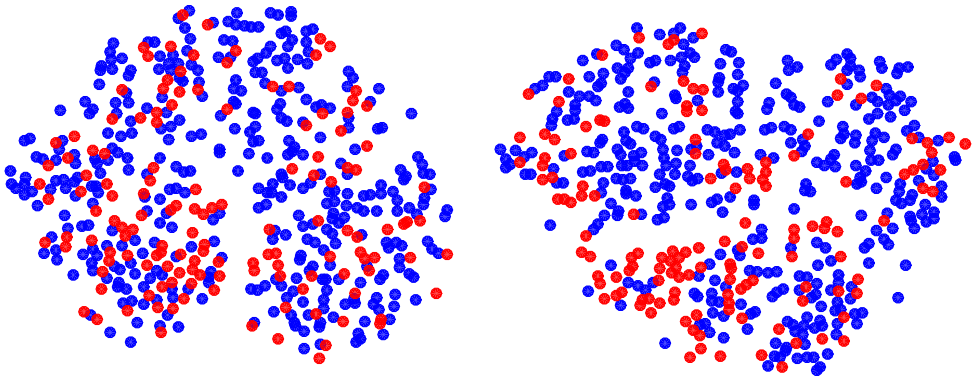
While not visible in the tables due to averaging, for some of the 1000 realizations of the IHDP dataset the imbalance penalty does not have an effect (neither positive nor negative). We believe this is due to these sets already being fairly balanced. We also note that in some cases, methods with linear hypothesis in the treatment, perform the best by removing any influence of the covariates on the model. This is indeed what happens for L+R and BLR on IHDP.

D.2 IHDP and News

Johansson et al. [20] introduced another semi-simulated dataset dubbed News, based on topic-modeling of a collection of news articles. Covariates represent counts of words in a pre-specified

Table 2: Results and standard errors on IHDP and News over 1000 and 50 repeated experiments respectively. MMD is the squared linear MMD. ITE and ATE are transductive. Lower is better.

	IHDP			NEWS		
	ϵ_{ITE}	ϵ_{ATE}	PEHE	ϵ_{ITE}	ϵ_{ATE}	PEHE
OLS	4.6 ± 0.2	0.7 ± 0.0	5.2 ± 0.3	3.1 ± 0.2	0.2 ± 0.0	3.3 ± 0.2
DR	3.0 ± 0.1	0.2 ± 0.0	5.7 ± 0.3	3.1 ± 0.2	0.2 ± 0.0	3.3 ± 0.2
L + R	2.8 ± 0.1	0.2 ± 0.0	5.7 ± 0.2	2.2 ± 0.1	0.6 ± 0.0	3.4 ± 0.2
BLR	2.8 ± 0.1	0.2 ± 0.0	5.7 ± 0.3	2.2 ± 0.1	0.6 ± 0.0	3.3 ± 0.2
BART [5]	2.1 ± 0.2	0.2 ± 0.0	1.7 ± 0.2	5.8 ± 0.2	0.2 ± 0.0	3.2 ± 0.2
C.FORESTS [33]	2.2 ± 0.1	0.2 ± 0.0	3.7 ± 0.2	2.1 ± 0.1	0.6 ± 0.0	2.6 ± 0.1
BNN-4-0 [20]	3.0 ± 0.0	0.3 ± 0.0	5.6 ± 0.3	2.7 ± 0.0	0.7 ± 0.0	3.6 ± 0.2
BNN-2-2 [20]	1.7 ± 0.0	0.3 ± 0.0	1.6 ± 0.1	2.0 ± 0.0	0.3 ± 0.0	$\mathbf{2.0 \pm 0.1}$
CFR-4-0 $\alpha = 0$	5.3 ± 0.1	0.8 ± 0.0	5.8 ± 0.3	2.9 ± 0.2	0.7 ± 0.1	3.4 ± 0.2
CFR-4-0 WASS	3.0 ± 0.2	0.3 ± 0.0	5.7 ± 0.3	2.8 ± 0.2	0.7 ± 0.1	3.4 ± 0.2
CFR-4-0 MMD	3.0 ± 0.1	0.3 ± 0.0	5.7 ± 0.3	2.9 ± 0.2	0.6 ± 0.1	3.4 ± 0.2
CFR-2-2 $\alpha = 0$	1.8 ± 0.0	0.3 ± 0.0	1.5 ± 0.1	2.1 ± 0.1	0.3 ± 0.0	2.2 ± 0.1
CFR-2-2 WASS	1.7 ± 0.1	0.3 ± 0.0	$\mathbf{1.4 \pm 0.1}$	2.1 ± 0.2	0.3 ± 0.2	$\mathbf{2.0 \pm 0.4}$
CFR-2-2 MMD	$\mathbf{1.6 \pm 0.0}$	$\mathbf{0.2 \pm 0.0}$	$\mathbf{1.4 \pm 0.1}$	2.1 ± 0.1	0.5 ± 0.1	$\mathbf{2.0 \pm 0.1}$



(a) $q = 0$ (b) $q = 1$

Figure 3: PEHE for the CFR-2-2 model on 200 subsampled versions of the IHDP with increasing imbalance. Right: subsampled datasets visualized using t-SNE. $q = 0$ corresponds to minimum imbalance, and $q = 1$ to maximum.

vocabulary representing words that are significant to at least one of the topics. The outcome and treatment models are based on the topic-distribution of documents. The set comprises 5000 observations (articles) with 3477 covariates (words). The results of both the IHDP and News experiments are presented in Table 2. For the News data, we train representation layers with 400 units and output layers with 200 units. For DR on News, we perform the logistic regression on the first 100 principal components of the data.

On News, we don't see a significant gain from using the proposed imbalance penalty (compare e.g. MMD and $\alpha = 0$), but neither does the penalty hurt the performance. A possible explanation is that the dataset is not very imbalanced to begin with.

D.3 IHDP - Increasing imbalance

T-SNE visualizations of the subsampled IHDP for $q = 0$ and $q = 1$ respectively can be seen in Figure 3.

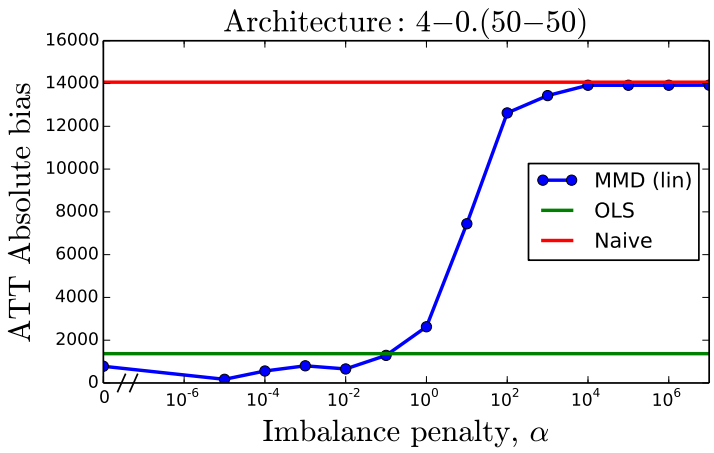


Figure 4: Absolute error in treatment effect on the treated on Jobs.

Table 3: Results on the continuous (left) and binary (right) versions of Jobs. BIAS_{ATT} is the (signed) bias in estimated average treatment effect on the treated. $(\%)$ is the bias in percent of the true effect. ERR_F is the factual classification error. *Not included in [29]. †Not applicable.

	Continuous outcome				Binary outcome		
	BIAS_{ATT}	$\%$	RMSE_{fact}		ϵ_{ATE}	$(\%)$	ERR_{fact}
Naïve	-14062	-1587%	/		0.160	206%	/
OLS	-1371	-154%	9537	LR	0.043	56%	12.6%
DR	-4973	-561%	12709		0.065	84%	15.8%
L+R	-886	-100%	9570	ℓ_1-LR	0.044	57%	12.4%
DID-Reg [29]	-522	-60%	*		†	†	†
DID-Match [29]	$\mathbf{-270}$	$\mathbf{-30}\%$	*		*	*	*
BART [5]	-1187	-133%	8618		0.023	30%	11.6%
C.Forests [33]	-2803	-316%	11576		0.019	28%	11.7%
CFR-4-0 $\alpha = 0$	-845	-95%	7069		0.059	75%	11.2%
CFR-4-0 Wass	-691	-78%	6997		**0.012**	**16**%	11.0%
CFR-4-0 MMD	-638	-72%	7047		0.041	52%	10.7%
CFR-2-2 $\alpha = 0$	-980	-110%	6832		0.090	116%	10.9%
CFR-2-2 Wass	-907	-102%	7026		0.040	51%	11.0%
CFR-2-2 MMD	-913	-103%	8150		0.061	78%	10.6%

D.4 Jobs

The results of both the continuous and binary task are presented in Table 3.

Figure 4 shows the bias in the estimated ATE for our CFR-4-0 method, OLS, and a naive estimator. We can see that when we don't impose balance penalties, our neural network method is slightly better than OLS, in terms of estimating the ATE. This is expected as we are free to use non-linear interactions between the covariates, but still restricted to hypotheses that are linear functions of the treatment variable. As we impose balance penalties, increasing α, the performance improves until it dominates the behavior and the method reduces to the naive estimator. This is also expected as the model disregards the covariates completely when $\alpha \to \infty$.