DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training

Nathan Kallus [1]

Abstract

We study optimal covariate balance for causal inferences from observational data when rich covariates and complex relationships necessitate flexible modeling with neural networks. Standard approaches such as propensity weighting and matching/balancing fail in such settings due to miscalibrated propensity nets and inappropriate covariate representations, respectively. We propose a new method based on adversarial training of a weighting and a discriminator network that effectively addresses this methodological gap. This is demonstrated through new theoretical characterizations and empirical results on both synthetic and clinical data showing how causal analyses can be salvaged in such challenging settings.

1. Introduction

Drawing causal inferences from observational data often relies on a careful accounting of relevant and systematic differences between treatment and control groups, or else any observed variation in response can be dismissed as spurious correlation rather than a bona-fide causal relationship. For example, hypothetically, differences in lung cancer incidence in coffee drinkers and non-drinkers might be explained away by differing rates of cigarette smoking. To eliminate that possibility, we must contrast groups that are comparable in their smoking rates, *i.e.*, control for that variable. The approach of controlling for confounders relies of course on the assumption that no other unobserved confounders exist, or, more generally and in terms of a causal diagram, that these covariates satisfy the back door criterion (Pearl, 2009).

But controlling for confounders also requires that we understand the way in which they affect treatment, outcome, or both. The rapid development in training neural networks (NNs) holds promise in enabling us to control for potentially complex relationships and rich covariates. Standard ways of finding weights that balance covariates for causal inferences rely either on estimating propensity scores (Rosenbaum & Rubin, 1983) or on directly minimizing imbalance metrics, such as kernel maximum mean discrepancy (MMD, Gretton et al., 2009; Kallus, 2016; 2017). But both of these approaches break down when dealing with rich covariates and complex relationships that necessitate flexible modeling with NNs. Propensity scores estimated by deep NNs tend to be highly volatile and often miscalibrated as probability estimates. This issue similarly plagues doubly robust approaches (Kang & Schafer, 2007; Robins et al., 1994). At the same time, optimal balancing weights rely very crucially on already having an appropriate representation of the data to balance.

In this paper, we develop a new approach to the problem of balancing covariates in such situations. The approach, termed DeepMatch, solves a game between a weighting and a discriminator network using adversarial training. Underlying this is a new discriminative discrepancy metric that we theoretically characterize and relate to existing metrics used for causal inference. To use it in the context of NNs requires a few further developments that in the end enable the use of alternating gradient approaches similar to Goodfellow et al. (2016). The method is shown to be statistically consistent in estimating both average and conditional effects. Studying a case using fully connected networks to learn complex relationships, a case using convolutional networks to deal with image confounders, as well as a case with real clinical outcomes, we demonstrate how DeepMatch can enable strong causal analyses in these challenging settings.

1.1. Related Literature

There has been intense interest in using machine learning, and NNs in particular, to estimate causal effects. For the problem of directly regressing individual effects under unconfoundedness, Athey & Imbens (2016); Wager & Athey (2017) study adapting tree-based methods and Johansson et al. (2016); Shalit et al. (2017) study more effective regularization techniques for NNs. These refine regression-based approaches that would ignore covariate shifts. They do not

[1]Cornell Tech, Cornell University, New York, NY, USA. Correspondence to: Nathan Kallus <kallus@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

provide covariate balancing weights that can be used for estimating other conditional effects or for doubly robust estimation. Using another identification condition altogether, Hartford et al. (2017) use NNs for instrumental-variable analysis, relying on identifying latent natural experiments rather than identifying and controlling confounders.

Another strand of work has focused on wrappers that can leverage machine learning predictors as subroutines. Chernozhukov et al. (2018); Hahn (1998) develop doubly robust estimators that use flexible estimators such as NNs to enable efficient estimation of simpler parameters like average effects. Künzel et al. (2017); Nie & Wager (2017) develop meta-learners that combine base learners like NNs to learn individual causal effects. All of the above rely on having access to weights that balance covariates and generally use estimated propensities. But, as discussed below, deep nets, while good classifiers, yield unwieldy inverse propensity weights. DeepMatch weights may provide a more stable alternative for these wrappers.

A variety of work has recently taken the approach of directly optimizing imbalance metrics for various causal inference tasks (Athey et al., 2018; Bertsimas et al., 2015; Kallus, 2016; 2017; 2018; Zubizarreta, 2012; 2015). For causal estimation from observational data, these take the form of minimizing metrics like Mahalanobis, Wasserstein, and MMD, but, as discussed below, this relies on already having an appropriate representation of the data. DeepMatch proceeds in the same spirit as these, optimizing directly for balance, but uses flexible NNs to allow for complex relationships and balancing deeper, learned representations.

2. Counterfactual Errors, Covariate Balance, and Representations

We consider an observational study consisting of n observations $\{(X_i, T_i, Y_i^{\text{obs}}) : i = 1, \dots, n\}$ of the variables (X, T, Y^{obs}), where $X \in \mathcal{X}$ denotes baseline covariates, $T \in \{0, 1\}$ treatment assignment, and $Y^{\text{obs}} \in \mathbb{R}$ observed outcome. For $t = 0, 1$, we let $\mathcal{T}_t = \{i : T_i = t\}$ and $n_t = |\mathcal{T}_t|$. We also let $T_{1:n} = (T_1, \dots, T_n)$ and $X_{1:n} = (X_1, \dots, X_n)$ denote all the observed treatment assignments and baseline covariates, respectively. We let $Y_i(0), Y_i(1) \in \mathbb{R}$ be the potential outcomes for unit i so that $Y_i^{\text{obs}} = Y_i(T_i)$. We further assume unconfoundedness:

Assumption 1. $\mathbb{E}[Y(0) \mid T, X] = \mathbb{E}[Y(0) \mid X]$.

2.1. Estimating Causal Effects

The simplest learning task we consider is to estimate the average treatment on the treated (ATT):

$$\tau = \mathbb{E}[Y(1) - Y(0) \mid T = 1] = \mathbb{E}\tau_{n_1},$$
$$\tau_{n_1} = \tfrac{1}{n_1} \textstyle\sum_{i \in \mathcal{T}_1} (Y_i(1) - Y_i(0))$$

The task is nontrivial because τ_{n_1} sums over treated units but only $Y_i(1)$ is observed for treated units and never $Y_i(0)$. This is termed *the fundamental problem of causal inference*: counterfactuals are unobserved. Under Asn. 1, this is solved by comparing treated and control units with similar covariates, using regression, matching/weighting, or both.

A related task is learning the best-in-$\mathcal{C}$-class conditional ATT (CATT):

$$\operatorname*{argmin}_{\tau \in \mathcal{C}} \mathbb{E}\left[(Y(1) - Y(0) - \tau(X^{\text{H}}))^2 \mid T = 1\right], \quad (2.1)$$

where X^{H} is some small subset of interest of the covariates. Eq. (2.1) is equivalent to the best model in $\mathcal{C}$ to represent the conditional average $\mathbb{E}\left[Y(1) - Y(0) \mid X^{\text{H}} = x^{\text{H}}, T = 1\right]$. CATT is of interest when our observational data contains a lot of very rich data X that can help us justify Asn. 1, but only much fewer variables are actually available when we need to make a treatment decision for incoming units, the heterogeneity of the causal effect is only interesting/reasonable along a particular direction of unit difference, and/or we need a simple and interpretable model to understand the heterogeneity. One can also consider the ATE and CATE – the analogous quantities on the general population without conditioning on $T = 1$ – and all of the methods we discuss easily extend. But, for clarity and brevity, we will focus on ATT and CATT.

Under Asn. 1, many estimators for ATT and CATT use the *propensity score* (Rosenbaum & Rubin, 1983). The propensity score of unit i is $e(X_i)$ where $e(x) = \mathbb{P}(T = 1 \mid X = x)$. The propensity score is unknown but it can be estimated by fitting $\hat{e}(x)$ as a probabilistic classifier trained to predict T from X, such as logistic regression or a NN. Given such a $\hat{e}(x)$, one can estimate the inverse probability weights (IPW) $W_i^{\text{IPW}} = T_i + (1 - T_i)(1 - \hat{e}(X_i))^{-1}\hat{e}(X_i)$. This gives rise to two popular estimators for ATT: the IPW estimator $\hat{\tau}_{W^{\text{IPW}}}$ (IPW, Horvitz & Thompson, 1952) and, given also an estimate $\hat{\mu}_0(x)$ of $\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]$, the doubly robust AIPW $\hat{\tau}_{W^{\text{IPW}}, \hat{\mu}_0}$ (Hahn, 1998; Robins et al., 1994), which we get by using $W_i = W_i^{\text{IPW}}$ in

$$\hat{\tau}_W = \frac{1}{n_1} \sum_{i=1}^{n} (-1)^{1+T_i} W_i Y_i^{\text{obs}}, \quad (2.2)$$

$$\hat{\tau}_{W, \hat{\mu}_0} = \frac{1}{n_1} \sum_{i=1}^{n} (-1)^{1+T_i} W_i (Y_i^{\text{obs}} - \hat{\mu}_0(X_i)). \quad (2.3)$$

CATT can similarly be estimated by a reweighted loss minimization problem, using $W_i = W_i^{\text{IPW}}$ in

$$\hat{\tau}_{W,\mathcal{C}}(\cdot) = \operatorname*{argmin}_{\tau \in \mathcal{C}} \sum_{i=1}^{n} W_i \left(Y_i^{\text{obs}} - \frac{2T_i - 1}{2} \tau(X_i^{\text{H}})\right)^2 \quad (2.4)$$

Unfortunately, by naïvely plugging in the estimate $\hat{e}(x)$ into the denominator, all these estimators are very susceptible to errors in $\hat{e}(x)$: whenever $1 - e(x)$ is small, even minute errors in $\hat{e}(x)$ can translate to outsize errors in effect estimates. This leads to very volatile estimates with high variance and hence large errors. Practical use requires heuristic stopgaps like trimming and normalizing the weights. The problem is not alleviated either by double robustness alone, as observed by Kang & Schafer (2007) who note that estimators using inverse propensity weights, whether doubly robust or not, are unreliable.

This issue is further exacerbated when X is rich so that reliably predicting T, and hence estimating $e(x)$, requires a deep NN. Such models are notorious for overconfidence and miscalibration (Guo et al., 2017). That is, while $\operatorname{sign}(\hat{e}(x) - 1/2)$ may be a good estimate of $\operatorname{sign}(e(x) - 1/2)$, the value of $\hat{e}(x)$ can be far off from $e(x)$ and will generally be close to 0 or 1, leading to huge errors in $1/\hat{e}(x)$. With simple, quantitative covariates, in order to avoid the instability of estimated propensity weights, the standard practice is to use matching (Iacus et al., 2012; Rosenbaum, 1989), or more generalized forms of matching and weighting that leverage modern optimization tools to directly optimize covariate balance (Athey et al., 2018; Hainmueller, 2012; Imai & Ratkovic, 2014; Kallus, 2016; 2017; Zubizarreta, 2012; 2015). But as explained below, these methods cannot address rich covariates and complex relationships.

2.2. Optimal Weighting for Covariate Balance

Since estimated propensity weights W_i^{IPW} are very volatile as discussed above, we consider replacing them by other data-driven weights W in eq. (2.2)–(2.4). We focus on weights that maintain a distribution on control units, preserving the unit of analysis: $W \in \mathcal{W} = \left\{W \in \mathbb{R}_+^{n_0} : \sum_{i\in\mathcal{T}_0} W_i = n_1\right\}$. Like estimated propensity weights, we consider choosing the weights W based only on the covariate and treatment data, $W = W(X_{1:n}, T_{1:n})$. Letting $\sigma_0^2(x) = \operatorname{Var}(Y(0) \mid X = x)$, a bias-variance decomposition shows that, under Asn. 1, the risk of $\hat{\tau}_W$ for a particular choice of weights is

$$\mathbb{E}[(\hat{\tau}_W - \tau_{n_1})^2 \mid X_{1:n}, T_{1:n}] = E^2(W; \mu_0, \sigma_0^2) \tag{2.5}$$
$$= \frac{1}{n_1^2}\left(\sum_i (-1)^{T_i} W_i \mu_0(X_i)\right)^2 + \frac{1}{n_1^2}\sum_i W_i^2 \sigma_0^2(X_i),$$

where for simplicity we let $W_i = 1$ for $i \in \mathcal{T}_1$. Similar expansions are possible for $\hat{\tau}_{W,\hat{\mu}_0}$ (Kallus, 2016). The error is decomposed into the squared bias given by covariate imbalance in weighted μ_0-moments plus conditional variance given by a σ_0^2-weighted norm of weights. That neither of these are known suggests a minimax approach over a function class $\mathcal{F} \subset [\mathcal{X} \to \mathbb{R}]$ for μ_0. Define the integral probability metric (IPM, Müller, 1997) between two weighted sets as

$$\operatorname{IPM}_{\mathcal{F}}(S^+, S^-) = \sup_{f\in\mathcal{F}} \left|\sum_{\pm}\sum_{i=1}^{n^\pm} \pm w_i^\pm f(x_i^\pm)\right|,$$
$$\text{where } S^\pm = \{(w_1^\pm, x_1^\pm), \dots, (w_{n^\pm}^\pm, x_{n^\pm}^\pm)\}$$

Then for an exchange rate λ, choosing W to minimize worst-case error can be written as (for any $\overline{\sigma}^2$)

$$\operatorname*{argmin}_{W\in\mathcal{W}} \sup_{(\lambda f/\overline{\sigma})\in\mathcal{F}, \sigma^2\le\overline{\sigma}^2} E^2(W; f, \sigma^2) = \tag{2.6}$$
$$\operatorname*{argmin}_{W\in\mathcal{W}} \left(\operatorname{IPM}_{\mathcal{F}}^2(\{(\tfrac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\tfrac{W_i}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) + \frac{\lambda\|W\|_2^2}{n_1^2}\right)$$

2.3. The Role of a Representation

Popular examples of IPMs include the total variation (TV) distance, given by the IPM for all functions point-wise bounded by 1; the MMD (Gretton et al., 2006), given by the IPM for the unit ball of a reproducing kernel Hilbert space (RKHS); and the Wasserstein distance, given by the IPM for all 1-Lipschitz functions. The MMD and Wasserstein distances have been popular for handling both matching for causal inference and for covariate shift (Gretton et al., 2009; Kallus, 2016). Specifically, Kallus (2016) proposes eq. (2.6) as a weighting objective for estimators of the form in eqs. (2.2)–(2.3) and shows that solving eq. (2.6) with the Wasserstein metric and $\lambda \to 0$ corresponds exactly to the classic pairwise matching approach to causal estimation (Rosenbaum, 1989). The TV distance, on the other hand, is uninformative for measuring covariate balance for non-discrete data: since it corresponds to almost-everywhere pointwise differences between measures, it is always equal 2 regardless of w_i^s whenever $\{x_i^+\} \cap \{x_i^-\} = \varnothing$ and is therefore not useful for finding covariate-balancing weights.

Balancing covariates based on MMD and Wasserstein IPMs delicately relies on having an appropriate representation of X for the task. For example, the MMD measures distances as the 2-norm distance between the weighted sample means in a *known* representation $X \mapsto K(X, \cdot)$. But with rich covariates like images, it is not clear that a sufficiently structured such representation is known a priori. Similarly, minimizing Wasserstein IPM (*without* further restricting $\mathcal{F}$ using a NN architecture as in (Arjovsky et al., 2017)) leads to pairwise matching on a prespecified metric such as (potentially kernelized) Euclidean in image pixels. Both of these are wholly inappropriate for such rich data because they rely on a functional family $\mathcal{F}$ without the right structure – loosely including all Lipschitz functions or misguidedly relying on smoothness in raw pixels, or other very rich data structure.

A solution might be to first put the data into an appropriate representation. But this requires learning a representation

and relies on its sufficiency for Assumption 1. For example, for image data, one would first use auxiliary labeled data such as ImageNet to learn a convolutional neural network (CNN) and extract from it an appropriate hidden representation. But even if such auxiliary data is available, there is no guarantee that the learned representation is necessarily appropriate for the causal estimation task and it is likely to break Assumption 1 by omitting predictors of both outcome and treatment.

Instead, in this paper, we will *directly* imbue $\mathcal{F}$ with the structure of a (potentially deep) NN, such as a CNN. This will induce a less loose structure more appropriate for the task and lead to the simultaneous learning and balancing of a sufficient representation of the covariate data. However, optimizing eq. (2.6) with respect to such functional families is problematic – the corresponding IPM is not closed form and using alternating saddle-point-finding methods directly on eq. (2.6) fail. We will instead develop a new, discriminative model for covariate balance, relate it the one based on IPMs, and use adversarial training in order to optimize it. This will lead to an alternative way to find weights that more resembles the generator-discriminator game of Goodfellow et al. (2014).

3. A Discriminative Model of Covariate Balance

In this section we develop the *discriminative distance* as a measure of distance between distributions (for concreteness we focus on finite weighted samples). This serves as an alternative to IPM distributional distance measures and is defined in terms of the minimal cross-entropy loss acquirable by a discriminator attempting to classify the two distributions. After defining the discriminative distance, we will theoretically characterize it, showing that it is not too different from the IPM with the same function class. We will then discuss finding weights that optimize it.

4. Defining the Discriminative Distance

For $f : \mathcal{X} \to \mathbb{R}$, consider the binary classifier that predicts the probability of a positive given x as $\operatorname{expit}(f(x))$ where $\operatorname{expit}(v) = 1/(1 + e^{-v})$. Letting $\ell(v) = \log(\operatorname{expit}(v)) + \log 2$, and given weighted positive and negative samples, S^+ and S^-, the log likelihood of the classifier relative to the random classifier is then

$$L(f; S^+, S^-) = \sum_{\pm} \sum_{i=1}^{n^\pm} w_i^\pm \ell(\pm f(x_i^\pm)).$$

$L(f; S^+, S^-)$ with $w_i^\pm = 1$ is the training objective for both logistic regression and NN classifiers. Similarly, training generative adversarial networks corresponds to having $w_i^\pm = 1$ and letting x_i^+ be given data and x_i^- be generated by a generator network with random inputs. Instead, we are here interested in training the weights $w_i^\pm$, leaving *both* x_i^+ and x_i^- as fixed, given data. Maximizing $L(f; S^+, S^-)$ over *all* $f(x)$ yields the Jensen-Shannon (JS) divergence, as observed by Goodfellow et al. (2014). But, like the TV distance, the JS divergence is uninformative for balancing when $\{x_i^+\} \cap \{x_i^-\} = \varnothing$ (always 1, regardless of $w_i^\pm$). Instead, we will be interested in restricting $\mathcal{F}$ in a meaningful way in order to come up with a balancing metric more similar to IPM but based on a discriminative objective.

Toward that end, define more generally the squared ψ-*discriminative distance* (ψ-DD) for $\psi \geq 0$ with respect to $\mathcal{F}$ between the weighted samples S^+, S^- as

$$\mathrm{DD}^2_{\mathcal{F},\psi}(S^+, S^-) = \sup_{f \in \mathcal{F}, t \in \mathbb{R}} L_\psi(f, t; S^+, S^-), \quad (4.1)$$
$$L_\psi(f, t; S^+, S^-) = L(tf; S^+, S^-) - \frac{\psi}{2} t^2$$

If $\mathcal{F}$ are all NNs of a fixed architecture with variable weights with sum of squared weights less than 1, eq. (4.1) corresponds to training a NN classifier with ψ weight decay. If $\mathcal{F}$ is an RKHS unit ball then eq. (4.1) corresponds to kernelized logistic regression (Jaakkola & Haussler, 1999). Standard or regularized logistic regression is a special case of either.

Theorem 1. $\mathrm{DD}^2_{\mathcal{F},\psi}(S^+, S^-)$ *is finite nonnegative. Hence,* $\mathrm{DD}_{\mathcal{F},\psi}(S^+, S^-)$ *is well-defined.*

Unlike IPM, DD is *not* a pseudo-metric because it does not satisfy the triangle inequality:

Example 1. Let $S^\pm = \{(1, \pm 1)\}$, $S' = \{(\frac{1}{2}, 1), (\frac{1}{2}, -1)\}$, and $\mathcal{F} = \{x \mapsto \pm x\}$. Then $\mathrm{DD}_{\mathcal{F},0}(S^+, S^-) = \sqrt{2\log(2)} > \sqrt{2\log(27/16)} = \sum_\pm \mathrm{DD}_{\mathcal{F},0}(S^\pm, S')$. In fact, the inequality remains strict for ψ-DD with any $\psi \geq 0$.

4.1. Characterizing the Discriminative Distance

Next, we argue that DD, while not a metric, is related to, and sometimes similar to, the IPM. We first give a dual characterization of DD that relates it directly to the IPM over the same $\mathcal{F}$.

Theorem 2. *Let* $h(p) = p\log(p) + (1-p)\log(1-p) + \log 2$. *Suppose* $\mathcal{F}$ *is a symmetric convex set. (*i.e.*, for all* $f, f' \in \mathcal{F}$, $p \in [0, 1]$, *and sign* $\pm$, *we have* $pf \pm (1-p)f' \in \mathcal{F}$*).*

If $\psi > 0$ *then* $\mathrm{DD}^2_{\mathcal{F},\psi}(S^+, S^-)$ *is equal to*

$$\inf_{0 \leq \mathbf{p} \leq 1} \left(\sum_{i=1}^{n_+} w_i^+ h(p_i^+) + \sum_{i=1}^{n_-} w_i^- h(p_i^-) + \frac{\psi^{-1}}{2} \mathrm{IPM}^2_{\mathcal{F}}(\{(p_i^+ w_i^+, x_i^+)\}_{i=1}^{n^+}, \{(p_i^- w_i^-, x_i^-)\}_{i=1}^{n^-}) \right) \quad (4.2)$$

And, if $\psi = 0$ *then* $\mathrm{DD}^2_{\mathcal{F},\psi}(S^+, S^-)$ *is equal to*

$$\inf_{0 \leq \mathbf{p} \leq 1} \sum_{i=1}^{n_+} w_i^+ h(p_i^+) + \sum_{i=1}^{n_-} w_i^- h(p_i^-) \quad (4.3)$$

$$\text{s.t. } \text{IPM}_{\mathcal{F}}(\{(p_i^+ w_i^+, x_i^+)\}_{i=1}^{n^+}, \{(p_i^- w_i^-, x_i^-)\}_{i=1}^{n^-}) = 0$$

This result can be used to show that DD is zero precisely when the corresponding IPM is zero.

Theorem 3. $\text{DD}_{\mathcal{F},\psi}(S^+, S^-) = 0 \iff \text{IPM}_{\mathcal{F}}(S^+, S^-) = 0.$

However, when $\psi = 0$, they need not induce the same topology. The next example shows that the IPM can approach zero while the discriminative distance remains constant:

Example 2. Fix $\delta > 0$. Let $S^+ = \{(1, \delta/2)\}$, $S^- = \{(1, -\delta/2)\}$, and $\mathcal{F} = \{x \mapsto \pm x\}$. Then we have that $\text{IPM}_{\mathcal{F}}(S^+, S^-) = \delta$ is arbitrarily small while $\text{DD}_{\mathcal{F},0}(S^+, S^-) = \sqrt{2\log(2)}$ is bounded away from zero.

But when $\psi > 0$, we can show that they *will* in fact induce the *same topology* on any space of weighted sets such that

$$M = \sup_{\pm, i \le n^\pm, f \in \mathcal{F}} |f(x_i^\pm)|,$$
$$\overline{w} = \sum_\pm \sum_{i=1}^{n^\pm} w_i^\pm$$

are bounded. Usually, we have $\overline{w} = 2$ as each set of weights sums to 1. If $\mathcal{F}$ is all NNs of a given architecture with sum squared weights no more than 1, then M is bounded over points x of bounded norm. M is also bounded if $\mathcal{F}$ is the unit ball of an RKHS with a bounded kernel (*e.g.*, RBF). The next result bounds the ratio between IPM and DD, showing equivalence if M and $\overline{w}$ are bounded.

Theorem 4.
$$2\sqrt{2\psi} \le \frac{\text{IPM}_{\mathcal{F}}(S^+, S^-)}{DD_{\mathcal{F},\psi}(S^+, S^-)} \le \max(2M\sqrt{\overline{w}}, 4\sqrt{\psi}).$$

Moreover, we can show that, as ψ grows, IPM and DD become the same up to scaling.

Theorem 5. *For any weighted sets S^+, S^-,*

$$\lim_{\psi \to \infty} 2\sqrt{2\psi}\, \text{DD}_{\mathcal{F},\psi}(S^+, S^-) = \text{IPM}_{\mathcal{F}}(S^+, S^-).$$

The limit holds uniformly over S^+, S^- with bounded $\overline{w}$, M.

4.2. Optimizing the Discriminative Distance

The last section suggests that DD with respect to $\mathcal{F}$ can be used as a surrogate for covariate balance. In particular, measuring covariate imbalances using DD, we have shown that eliminating the DD when $\mathcal{F}$ consists of NNs of a given architecture corresponds exactly to eliminating *any* estimation bias due to imbalances for any outcomes that can be well-approximated by such NNs. Therefore, an alternative criterion for choosing weights is to minimize imbalance as measured by the DD plus a potential variance regularizer:

$$W^* \in \operatorname*{argmin}_{W \in \mathcal{W}} \left(I^2(W) + \frac{\lambda}{n_1^2} \|W\|_2^2 \right) \tag{4.4}$$

Algorithm 1 Conditional Gradient for Eq. (4.4)

input: $X_{1:n}, T_{1:n}, \lambda, K$, and an oracle as in eq. (4.5)
Set $W_i = 1/n_0$ for all $i \in \mathcal{T}_0$
for $k = 1, \ldots, K$ **do**
  Set $S^+ = \{(\frac{1}{n_1}, X_i)\}_{i \in \mathcal{T}_1}$ and $S^- = \{(\frac{W_i}{n_1}, X_i)\}_{i \in \mathcal{T}_0}$
  Get the corresponding $f^{\text{oracle}}, t^{\text{oracle}}$ in eq. (4.5)
  Let $i = \text{argmin}_{i \in \mathcal{T}_0}(\ell(-t^{\text{oracle}} f^{\text{oracle}}(X_i)) + \frac{2\lambda}{n_1^2} W_i)$
  $W \leftarrow (k-1)/(1+k) W, W_i \leftarrow W_i + 2/(1+k)$
end for
output: W

$$I^2(W) = \text{DD}^2_{\mathcal{F},\psi}(\{(\tfrac{1}{n_1}, X_i)\}_{i \in \mathcal{T}_1}, \{(\tfrac{W_i}{n_1}, X_i)\}_{i \in \mathcal{T}_0})$$

The question then is how to find such weights W that optimize eq. (4.4). We first show the problem is convex.

Theorem 6. *The optimization problem eq.* (4.4) *is convex for any $\mathcal{F}$, ψ, and $\lambda \ge 0$.*

To optimize eq. (4.4), we first consider the simple case where we have an oracle for the optimization problem eq. (4.1). Specifically, suppose that, given S^+, S^- we could easily find $f^{\text{oracle}} \in \text{cl}(\mathcal{F})$, $t^{\text{oracle}} \in \mathbb{R}$ such that

$$L_\psi(f^{\text{oracle}}, t^{\text{oracle}}; S^+, S^-) = \text{DD}^2_{\mathcal{F},\psi}(S^+, S^-). \tag{4.5}$$

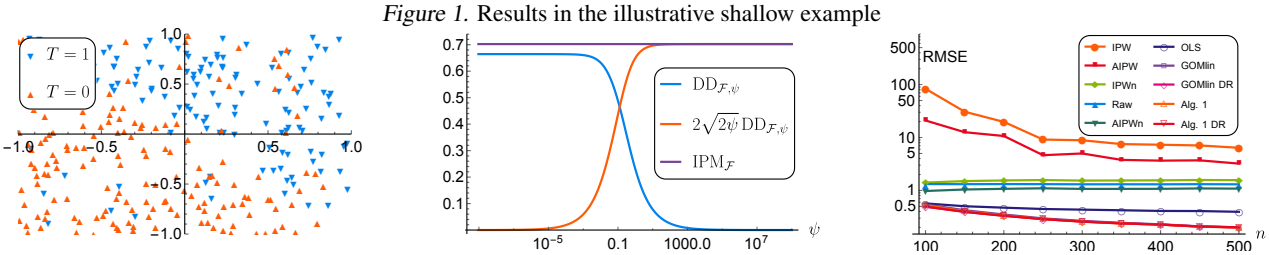
For example, in the case of $\mathcal{F}$ being linear functions or being an RKHS, this amounts to solving logistic regression (possibly kernelized), which can be done quickly and easily for even large (but not huge) problems (Jaakkola & Haussler, 1999). Given such an oracle, we can employ the conditional gradient algorithm (Frank & Wolfe, 1956; Jaggi, 2013) to solve eq. (4.4), producing Algo. 1. Note that we can employ Thm. 2 to show that eq. (4.4) can also be written as a single convex minimization problem in W and $\mathbf{p}$ with an IPM in the objective (or, in a constraint for $\psi = 0$). Therefore, we can also solve eq. (4.4) using an oracle for IPM instead, albeit with slightly more complicated gradients and twice the number of optimization variables.

4.3. A Shallow Example

Before developing DeepMatch, we first consider a shallow example, that is, one with few dimensions and simple treatment and outcome models, in order to illustrate DD and Alg. 1. The example is particularly simplistic and is meant only for illustration and not as a comparison, which we do instead in Sec. 6. We consider bivariate covariates $X_i \sim \text{Uniform}[-1, 1]^2$, let $e(X_i) = 0.1$ if $X_{i1} + X_{i2} < 0$ and otherwise $e(X_i) = 0.9$, and let outcomes be exponential in X_i: $Y_i(0) = Y_i(1) = e^{X_{i1} + X_{i2}} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$.

Fixing a particular draw of $n = 300$ units, we plot the covariates in Fig. 1(a). Letting $\mathcal{F} = \{x \mapsto \alpha + \beta^T x : \|\beta\|_2 \le 1\}$, we plot the IPM (Euclidean distance between means) and

Figure 1. Results in the illustrative shallow example



(a) 2-dimensional covariates $X_{1:n}$ (b) The discriminative distance and IPM (c) RMSE of different ATT estimators

DD (given by regularized logistic regression) between the raw (not reweighted) control and treated samples in Fig. 1(b) for varying $\psi \geq 0$. We see that, as Thms. 4 and 5 promise, the scaled DD converges to IPM from below.

Next, we consider how well Alg. 1 balances covariates for estimating causal effects. For each of $n = 100, 150, \ldots, 500$, we compute the RMSE of 8 different estimators over 2000 replications and plot the results in Fig. 1(c). IPW is given by propensities estimated by logistic regression and IPWn normalizes IPW to sum to 1. AIPW and AIPWn are the corresponding doubly robust estimators with $\hat{\mu}_0$ given by OLS, which respectively improve on IPW, IPWn. The raw (unweighted) mean difference, which beats the IPW, IPWn, and AIPW estimates. An OLS estimate, which is also equivalent to using unconstrained weights that minimize the IPM (Kallus, 2016, Thm. 13), beats all propensity-based estimates. Finally, we compute W^* to minimize DD ($\psi = 0$) using Alg. 1 ($\lambda = 1$) and consider both $\hat{\tau}_W$ and $\hat{\tau}_{W,\hat{\mu}_0}$. We also consider the same when replacing DD with the corresponding IPM with the same $\mathcal{F}$, which amounts to using generalized optimal matching with a linear kernel (GOMlin; Kallus, 2016). These last two match almost exactly and yield the least error, with double robustness giving marginal improvements for small n.

5. Going Deep

We are now prepared to develop the DeepMatch method, which seeks balance complex and deep representation of the covariate data. DeepMatch does this by seeking weights that minimize the discriminative distance with respect to a deep discriminator, *i.e.*, when $\mathcal{F}$ is given by all NNs of a given architecture. There are two barriers to doing this with the tools we developed so far (Alg. 1). First, as n grows, additional runs of the conditional gradient algorithm are necessary to generate a dense set of weights, since the number of nonzero weights is bounded above by the number of iterations. Second, each of these runs becomes increasingly difficult to compute when $\mathcal{F}$ is complex, as in the case of a deep neural network. Computing the corresponding oracle can be exceedingly difficult because, not only is optimizing a NN taxing, one usually does not optimize it fully, even to mere local optimality, and so the weight gradient recovered may be highly inaccurate and usable.

Instead, DeepMatch relies on alternating descent methods as are used in adversarial training of GANs (Goodfellow et al., 2014) to solve minimax problems with NNs. That is, we will address problem eq. (4.4) by taking alternating (or, simultaneous) gradient steps in $W \in \mathcal{W}$ and $f \in \mathcal{F}$, each while treating the other as fixed. However, to apply this approach here, we will need to make further developments to make our problem amenable to the standard solution methods that use stochastic optimization in order to deal with large datasets and complex objectives. In particular, we need to be able to solve the problem in *mini-batches*. To do this, we first show we can relax constraints that link the optimization variables across data points and we then show how we can parametrize the weights using their own NN.

5.1. Relaxing Weight Normalization

One impediment to applying mini-batched descent to problem eq. (4.4) is the normalization constraint, that weights have to sum to a certain fixed value, which links the optimization variables across data points i. Not only does this make stochastic mini-batching difficult, the constraint also becomes non-convex if we parametrize the weights as we are going to do in the next section. Therefore, the first step is to find an alternative way to enforce the normalization constraint. The next theorem will guide the approach.

Theorem 7. *Problem eq.* (4.4) *is equivalent to*

$$\min_{\phi \geq 0, W \in \mathcal{W}(\phi)} I^2\left(\frac{n_1 W}{\sum_{i \in \mathcal{T}_0} W_i}\right) + \lambda \frac{\sum_{i \in \mathcal{T}_0} W_i^2}{(\sum_{i \in \mathcal{T}_0} W_i)^2}, \quad (5.1)$$

$$\mathcal{W}(\phi) = \operatorname*{argmin}_{W \geq 0} I^2(W) + \sum_{i \in \mathcal{T}_0} \left(\lambda \frac{W_i^2}{n_1^2} + \phi \frac{W_i}{n_1}\right). \quad (5.2)$$

Thm. 7 suggests the following approach to optimizing eq. (4.4): select a grid of ϕ values, $\phi_1, \ldots, \phi_K$; solve eq. (5.2) for each ϕ value; normalize the resulting optimal set of weights; and chose the best set of weights by plugging each into the objective of eq. (4.4). We discuss the details of optimizing eq. (5.2) and choosing the ϕ grid in Sec. 5.3.

5.2. Parameterizing W and Choosing an Activation

Next, we parameterize the weights W using their own NN. That is, we replace the n_0 optimization variables given by W by a NN that produces their values. A question that arises is what activation to choose for the network's output. Choosing a nonnegative activation, such as a ReLU, easily accommodates the constraint that the weights be nonnegative. However, it may produce weights that are too sparse, outputting many zeros. Other activations such as logistic or $1 + \tanh$ are everywhere positive but have limited range, while the weight variables in eq. (5.2) are unrestricted and unnormalized.

Instead, we take inspiration from the weight function that would arise from a direct application of inverse propensity weighting with propensities that would be estimated by logistic regression or a NN. Estimating the propensity function as $\hat{e}(x) = \mathrm{expit}(g(x))$, for any function g, leads to the plug-in importance weight $W_i^{\mathrm{IPW}} = \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} = \exp(g(x))$. As previously discussed and illustrated, such plug in weights lead to unwieldy and inaccurate estimates, but, as a functional form, this suggests $\exp$ as the activation on the output layer for any weight-producing network, which can then be trained to produce weights directly optimized for accuracy. This is only a design choice that seems to work – similar to using softmax for classification NNs – and other choices exist. Considering some parametric family $g(x; \theta_g)$, such as NNs of a given architecture, we will solve eq. (5.2) by transforming $W_i = \exp(g(X_i; \theta_g))$ and optimizing over θ_g instead of over $W \in \mathcal{W}$.

5.3. Putting It Together

We now use the insights of the last two sections to develop DeepMatch. Let us express $\mathcal{F}$ parametrically and assume it has the form $\mathcal{F} = \{f(\cdot; \theta_f) : R(\theta_f) \leq 1\}$ for some degree-2 homogeneous regularizer R (can be zero). And let us write $\ell_i(\theta_f) = \ell((-1)^{T_i+1} f(X_i; \theta_f))$. Then solving the parametrized version of eq. (5.2), where we set $W_i = \exp(g(X_i; \theta_g))$, amounts to solving the following zero-sum game in θ_g and θ_f:

$$\min_{\theta_g} \max_{\theta_f} L_\phi(\theta_g, \theta_f), \quad \text{where} \tag{5.3}$$

$$L_\phi(\theta_g, \theta_f) = \frac{1}{n_1} \sum_{i=1}^{n} u_i(\theta_f, \theta_g; \psi, \lambda, \phi) - \frac{\psi}{2} R(\theta_f),$$

$$u_i(\theta_f, \theta_g; \psi, \lambda, \phi) = T_i \ell_i(\theta_f) + (1 - T_i) \Bigg(e^{g(X_i;\theta_g)} \ell_i(\theta_f) + \frac{\lambda}{n_1} e^{2g(X_i;\theta_g)} + \phi e^{g(X_i;\theta_g)} \Bigg).$$

Note each u_i depends only on the data X_i, T_i. Note that if $R(\theta_f)$ is convex in $f(\cdot; \theta_f)$ then $L(\theta_g, \theta_f)$ is convex in $g(\cdot; \theta_g)$ and concave in $f(\cdot; \theta_f)$. Consequently, by the von Neumann minimax theorem, the min and max can be exchanged (under some regularity). Note also that $L(\theta_g, \theta_f)$ is of the form of a sum of loss functions separable over the data plus some regularization (we can also easily include regularization in θ_g).

Using this formulation, we can develop Alg. 2 to find optimal weights for balancing covariates with respect to DD over NN discriminators. For each value of ϕ, the algorithm proceeds in two stages. In the first stage we address eq. (5.2) by seeking an equilibrium to eq. (5.3). To do this, over K_1 epochs, we cycle through mini-batches of size B and, for each, we let θ_f ascend its gradient and θ_g descend its (we use simultaneous updates; alternating is another option). We can use any stochastic gradient update rule. In our experiments, we use Adam (Kingma & Ba, 2014) with a global learning rate of 10^{-4} (for other options see Goodfellow et al., 2016, Ch. 8). In the second stage, we address eq. (5.1) by trying to evaluate DD of the resulting weights after normalization. To do this, we simply fix the weights and train *only* the discriminator in this stage, over K_2 epochs with mini-batches of size B. Because even for a single ϕ the first stage can end up at different weights, we do several (M) runs of each ϕ. Finally, we make sure to keep the weights over all these runs with the best objective so far in eq. (5.1).

A remaining detail is how to choose the grid Φ of ϕ values. Due to convexity, Thm. 7 also gives that $\sum_i \mathcal{W}_i(\phi)$ is monotonic in ϕ and the optimizer of eq. (4.4) is given by the root where the sum is exactly n_1. Above, we directly compare objectives rather than use a line search over Lagrange multipliers because our optimization for each ϕ is inexact. Nonetheless, we can use this to find an appropriate range for ϕ. Picking a tolerance $\eta \in (0, 1)$, we use binary search to find values $\underline{\phi}, \overline{\phi}$ that give weights that sum $\geq \eta n_1$ and $\leq n_1/\eta$, respectively. Then, we make a linear grid of M values in between.

5.4. Theoretical Characterization

We next prove that if we were to do the optimization exactly, then we indeed estimate the causal effects of interest. The rates will depend on the Rademacher complexity of $\mathcal{F}$:

$$\mathfrak{R}_n(\mathcal{F}) = \frac{1}{2^n} \sum_{\xi \in \{\pm 1\}^n} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i; \theta_f) \right|.$$

For both linear models and NNs, $\mathfrak{R}_n(\mathcal{F}) = O_p(1/\sqrt{n})$ (Bartlett et al., 2017; Golowich et al., 2018; Kakade et al., 2009).

Theorem 8. *Let*

$$\Theta_g = \bigcup_{\phi > 0} \operatorname*{argmin}_{\theta_g} \max_{\theta_f} L_\phi(\theta_g, \theta_f)$$

$$W_i(\theta_g) = T_i + (1 - T_i) \frac{n_1 e^{g(X_i;\theta_g)}}{\sum_{j \in \mathcal{T}_0} e^{g(X_j;\theta_g)}},$$

Algorithm 2 DeepMatch

input: $X_{1:n}$, $T_{1:n}$, ψ, λ, K_1, K_2, B, M, Φ, regularizer R, network architectures, and a gradient update rule
$v^* \leftarrow \infty$
for $\phi \in \Phi, m = 1, \ldots, M$ **do**
  Randomly initialize the parameters θ_f, θ_g of the discriminator and weight networks
  for $k = 1, \ldots, K_1$ **do**
    Shuffle the data into mini-batches $I_1, \ldots, I_{\lceil n/B \rceil}$ of sizes $|I_j| \in \{B, B-1\}$
    for $j = 1, \ldots, \lceil n/B \rceil$ **do**
      $\delta_g \leftarrow \frac{n/n_1}{|I_j|} \sum_{i \in I_j} \nabla_{\theta_g} u_i(\theta_f, \theta_g; \psi, \lambda, \phi)$
      $\delta_f \leftarrow \frac{n/n_1}{|I_j|} \sum_{i \in I_j} \nabla_{\theta_f} u_i(\theta_f, \theta_g; \psi, \lambda, \phi) - \frac{\psi}{2} \nabla R(\theta_f)$
      Move θ_f in direction δ_f and θ_g in direction $-\delta_g$
    end for
  end for
  for $i = 1, \ldots, n$ **do** $W_i \leftarrow T_i + (1 - T_i) \frac{n_1 e^{g(X_i; \theta_g)}}{\sum_{j \in \mathcal{T}_0} e^{g(X_j; \theta_g)}}$
  for $k = 1, \ldots, K_2$ **do**
    Shuffle the data into mini-batches $I_1, \ldots, I_{\lceil n/B \rceil}$
    for $j = 1, \ldots, \lceil n/B \rceil$ **do**
      $\delta_f \leftarrow \frac{1}{n_1} \sum_{i \in I_j} W_i \nabla \ell_i(\theta_f) - \frac{\psi}{2} \nabla R(\theta_f)$
      Move θ_f in direction δ_f
    end for
  end for
  $v \leftarrow \frac{1}{n_1} \sum_{i \in I_j} W_i \ell_i(\theta_f) - \frac{\psi}{2} R(\theta_f) + \lambda \sum_{i \in \mathcal{T}_0} W_i^2$
  if $v^* > v$ **then** $v^* \leftarrow v, W^* \leftarrow W$
end for
output: W^*

$$I^2(W) = \sup_{\theta_f} \frac{1}{n_1} \sum_{i=1}^{n} W_i \ell_i(\theta_f) - \frac{\psi}{2} R(\theta_f), \text{ and}$$
$$W^* \in \operatorname*{argmin}_{W \in W(\Theta_g)} I^2(W) + \lambda \sum_{i \in \mathcal{T}_0} W_i^2.$$

Suppose that Asn. 1 holds, $\psi > 0$, $\lambda > 0$, $\sigma_0^2(X)$ is a.s. bounded, $e(X)$ is a.s. bounded away from 1, $\exists \theta_f : \mu_0(x) = f(x; \theta_f)$, and $\exists \theta_g : e(x) = \text{expit}(g(x; \theta_g))$. Then,

$$\hat{\tau}_{W^*} - \tau = O_p(\mathfrak{R}_n(\mathcal{F}) + 1/\sqrt{n})$$

6. Experiments

We next proceed to evaluate DeepMatch empirically in three examples, one fully synthetic, one with synthetic outcomes and treatment, and one with real covariate and outcome data. In each we consider two tasks: learning ATT and CATT. We consider several NN-based strategies: a regression approach based on estimating $\hat{\mu}_0(x)$ using a NN, an IPW approach based on estimating $\hat{e}(x)$ using a NN, an AIPW approach combining these (only for ATT), and the same after *normalized* propensities (A/IPWn). Each network has a different and appropriate output activation and loss. The architectures change in each example. We also include three straw-man comparisons: the raw unweighted control sample, generalized optimal matching with a linear kernel (GOMlin; Kallus, 2016), and entropy balancing (E-bal; Hainmueller, 2012). The latter two represent optimal weighting methods with the *wrong* representation as both rely essentially on a linear representation. These stand mostly to highlight the complex relationships in the experiments given below, which are particularly chosen to highlight the need for NN models, as these methods give state-of-the-art performance in simpler settings with known or easily kernelizable representations.

We compare all of these to DeepMatch with the same network architectures as the NN-based methods, $\psi = 0$, and λ either 0 or 1. All NN-based methods use the *same* architectures, with possibly a different final activation, as detailed in each of the following sections. We use $K_1 = 20$ epochs with mini-batches of 100 to train all networks and for the first stage of DeepMatch and we use $K_2 = 10$ epochs for the second stage. We use $M = 5$ and a grid of 50 ϕ values based on $\eta = 0.01$. For CATT, we let $\mathcal{C}$ be linear functions on a univariate X^{H} and ask how well we estimate its coefficient. For weighting methods, we solve eq. (2.4). For regression, we regress on imputed effect. All results are based on 100 replications with $n = 1000$. DeepMatch took on average 17x longer to train than the basic propensity and regression networks, for which we also used more epochs to train fully as they only required one training run.

6.1. A Fully Connected Example

We first consider a fully synthetic example specifically geared to highlight the issues with standard approaches. In the next examples, we will consider real covariate and outcome data. We consider 6-dimensional covariates $X_i \sim \text{Uniform}[-2, 2]^6$ and a treatment model involving the XOR of signs: $e(X_i) = 0.05$ if $\bigoplus_{j=1}^{6}(X_{ij} > 0)$ and otherwise $e(X_i) = 0.95$. We let outcomes be $Y_i(t) = \exp(\sum_{j=1}^{6} X_{ij}) + T_i \sum_{j=1}^{6} X_{ij} - T_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 1)$. We let $X^{\text{H}} = \sum_{j=1}^{6} X_{ij}$. All NNs are fully connected with 5 hidden layers of 6, 3, 3, 2, and 2 neurons and ReLU activations (enough to model the XOR). (Simpler 3-layer NNs got similar results but the above is fairer to A/IPW, which needs more layers specify the XOR.)

We see (Tab. 1) that inverse-propensity-based approaches yield very volatile estimates, where normalizing or augmenting does help but only marginally. Using DeepMatch (DM) reduces MSE significantly. Combining it with the NN regression in a doubly robust estimator (DM-DR) reduces error slightly further. Using $\lambda > 0$ provides somewhat better results by controlling variance, but does not make a substantial difference in MSE.

6.2. A Convolutional Example

We next consider an example with confounding image data using the MNIST dataset (LeCun, 1998) and solutions based

Table 1: Fully connected

Method	(a) ATT Bias	(a) ATT SE	(a) ATT RMSE	(b) CATT Bias	(b) CATT SE	(b) CATT RMSE
Raw	-8.33	6.44	10.53	-10.22	8.35	13.20
GOMlin	-9.60	6.49	11.59	-11.36	7.44	13.58
E-bal	-7.80	6.11	9.90	-9.67	6.88	11.87
IPW	-8.25	6.29	10.38	-10.19	7.77	12.82
IPWn	-8.34	6.37	10.50	-10.26	7.84	12.91
Regress	9.42	4.00	10.24	8.37	2.40	8.71
AIPW	-8.07	6.22	10.19	–	–	–
AIPWn	-8.14	6.30	10.30	–	–	–
DM_0	-0.45	4.26	4.28	-2.70	4.54	5.29
$DM_{\frac{1}{2}}$	0.85	3.89	3.98	-1.25	3.68	**3.89**
DM_0 DR	-0.55	4.03	4.06	–	–	–
$DM_{\frac{1}{2}}$ DR	0.62	3.76	**3.82**	–	–	–

Table 2: Convolutional

Method	(a) ATT Bias	(a) ATT SE	(a) ATT RMSE	(b) CATT Bias	(b) CATT SE	(b) CATT RMSE
Raw	3.90	0.06	3.90	-0.44	0.03	0.44
GOMlin	3.24	0.14	3.25	-0.36	0.05	0.36
E-bal	3.75	0.15	3.76	-0.37	0.05	0.37
IPW	6.43	0.06	6.43	1.05	0.01	1.05
IPWn	3.32	0.41	3.35	-0.39	0.14	0.42
Regress	2.61	0.17	2.62	-0.33	0.06	0.34
AIPW	2.61	0.17	2.62	–	–	–
AIPWn	2.61	0.13	2.61	–	–	–
DM_0	2.24	0.96	**2.44**	-0.30	0.12	0.32
$DM_{\frac{1}{2}}$	2.49	0.65	2.57	-0.28	0.08	**0.30**
DM_0 DR	2.59	0.14	2.59	–	–	–
$DM_{\frac{1}{2}}$ DR	2.59	0.14	2.60	–	–	–

Table 3: Clinical (Twins)

Method	(a) ATT (in 0.01s) Bias	(a) ATT (in 0.01s) SE	(a) ATT (in 0.01s) RMSE	(b) CATT Bias	(b) CATT SE	(b) CATT RMSE
Raw	3.36	0.97	3.50	0.31	0.22	0.38
GOMlin	0.57	2.68	2.74	0.27	0.48	0.56
E-bal	-3.22	5.14	6.06	0.21	0.48	0.53
IPW	-0.16	4.87	4.87	0.14	0.55	0.56
IPWn	-0.09	4.44	4.44	0.13	0.54	0.56
Regress	2.12	6.50	6.84	0.24	0.15	**0.28**
AIPW	0.09	4.74	4.74	–	–	–
AIPWn	0.52	4.00	4.03	–	–	–
DM_0	1.64	1.26	2.06	0.21	0.28	0.35
$DM_{\frac{1}{2}}$	1.58	1.31	**2.05**	0.22	0.28	0.36
DM_0 DR	2.15	1.09	2.41	–	–	–
$DM_{\frac{1}{2}}$ DR	2.06	1.13	2.35	–	–	–

on CNNs. First we draw n digits uniformly at random. For each one we uniformly draw a random image labeled with that digit from the MNIST dataset and let the pixels $X_i \in \mathbb{R}^{28\times 28}$ be the covariates. Let $N(X) \in \{0, \ldots, 9\}$ be the digit corresponding to the image and let X^{H} be the sum of pixel brightnesses. To introduce a complex form of confounding, for each digit 0–4 separately, we take the 10% lightest images (smallest X^{H}) and label them treated $T_i = 1$ and the rest as untreated $T_i = 0$. We do the same for 5–9 but take 90% lightest images as treated. We generate outcomes as $Y_i(t) = \mathrm{clip}_{[0,9]}(N(X) + \epsilon_i) + T_i X^{\mathrm{H}} - T_i$ where $\epsilon_i \sim \mathrm{Uniform}\{-1, 0, 1\}$. All NNs have two 5×5 conv layers with volume depths 32 and 64 followed by a fully connected layer of 1024 neurons. The results (Tab. 2) show similar trends as Tab. 1 and demonstrate that DeepMatch can also improve causal estimates in settings with rich data and complex confounding.

6.3. Twins: Real Outcome Data

We next consider an example with real rather than simulated outcomes. We use the Twins dataset of 71345 twin births in the US between 1989–1991 as used by (Louizos et al., 2017). We let each birth be a unit, $Y(0), Y(1) \in \{0, 1\}$ be the mortality of the lighter and heavier twin, respectively, and $T_i = 1$ indicate being the heavier twin. X_i has "46 covariates relating to the parents, the pregnancy and birth: mother and father education, marital status, race and residence; number of previous births; pregnancy risk factors such as diabetes, renal disease, smoking and alcohol use; quality of care during pregnancy; whether the birth was at a hospital, clinic or home; and number of gestation weeks prior to birth" (Louizos et al., 2017). After encoding categoricals using dummies we have 161 covariates. We let $e(X_i) = 0.95$ if the pregnancy gestated less than the median period or the XOR over 10 variables with high impact on mortality (see appendix) being less than their mean; otherwise $e(X_i) = 0.05$. We let X^{H} be the risk factor oligo/hydramnios. We use the same architectures as in the fully-connected example but with logistic output for the outcome NN. Again, all methods share the same architecture, aside from the final activation.

For estimating the ATT, we observe (Tab. 3) similar trends, where DeepMatch provides much more accurate estimates, even with *real* outcomes. Again, increasing λ reduces SE and increases bias, but does not materially affect total RMSE. For estimating CATT, in this particular example, eq. (2.4) with linear $\mathcal{C}$ is not a good estimator, perhaps because of the binary outcomes. Nonetheless, DeepMatch provides the best results among the weighting methods that use eq. (2.4).

7. Conclusions

Training balancing weights against an adversarial discriminator, each represented by a NN, provided a way to balance covariates that may be rich and/or have complex relationships with outcomes and treatments – a challenging setting for standard approaches to causal inference. Both theory and empirical results uniformly demonstrated that DeepMatch yields stable balancing weights that lead to improved causal estimates even when an appropriate representation is unavailable a priori. A limitation is the heavy computational burden of searching over the Lagrange multiplier for weight normalization, alleviating which with new algorithms or approximations remains future work.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

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

Athey, S., Imbens, G., and Wager, S. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B*, 80(4):597–623, 2018.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *arXiv preprint arXiv:1703.02930*, 2017.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific, Belmont, 1999.

Bertsimas, D., Johnson, M., and Kallus, N. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4): 868–876, 2015.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*. MIT press Cambridge, 2016.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2006.

Gretton, A., Smola, A. J., Huang, J., Schmittfull, M., Borgwardt, K. M., and Schölkopf, B. Covariate shift by kernel mean matching. 2009.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pp. 315–331, 1998.

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

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423, 2017.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Iacus, S. M., King, G., and Porro, G. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24, 2012.

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

Jaakkola, T. S. and Haussler, D. Probabilistic kernel regression models. In *Artificial Intelligence and Statistics (AISTATS)*, 1999.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, pp. 427–435, 2013.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pp. 793–800, 2009.

Kallus, N. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

Kallus, N. A framework for optimal matching for causal inference. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 372–381, 2017.

Kallus, N. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

Kang, J. D. and Schafer, J. L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Künzel, S., Sekhon, J., Bickel, P., and Yu, B. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

LeCun, Y. The mnist database of handwritten digits. 1998.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6449–6459, 2017.

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

Nie, X. and Wager, S. Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*, 2017.

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

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Rockafellar, R. T. *Convex analysis*. Princeton university press, 1997.

Rosenbaum, P. R. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84 (408):1024–1032, 1989.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085, 2017.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017.

Zubizarreta, J. R. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.

Zubizarreta, J. R. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A. Proofs

Proof of Thm. 1. Note that $t = 0$ is feasible in eq. (4.1) and, since $\ell(0) = \log(1/2) + \log(2) = 0$, it gives objective value 0, so $\mathrm{DD}^2_{\mathcal{F},\psi}(S^+, S^-) \geq 0$. On the other hand, $\ell(\cdot) \leq \log(2)$ so that the objective in eq. (4.1) is bounded above: $\mathrm{DD}^2_{\mathcal{F},\psi}(S^+, S^-) \leq \sum_{s=\pm} \sum_{i=1}^{n^s} w_i^s \log(2) < \infty$. □

Proof of Thm. 2. Letting

$$\mathcal{F}_{|\mathbf{x}} = \{(f(x_i^s))_{s=\pm, i\leq n^s} : f \in \mathcal{F} \cup -\mathcal{F}\}, \quad \overline{\mathcal{F}}_{|\mathbf{x}} = \{(t, t'\mathbf{f}) : f \in \mathcal{F}_{|\mathbf{x}}, t \geq t' \geq 0\},$$

we can rewrite $\mathrm{DD}^2_{\mathcal{F},\psi}(\{(w_i^+, x_i^+)\}_{i=1}^{n^+}, \{(w_i^-, x_i^-)\}_{i=1}^{n^-})$ as

$$\sup_{(t,\mathbf{f})\in\overline{\mathcal{F}}_{|\mathbf{x}}, \mathbf{z}\in\mathbb{R}^{n^+ + n^-} : z_i^s = s f_i^s} \sum_{s=\pm} \sum_{i=1}^{n_+} w_i^s \ell(z_i^s) - \frac{\psi}{2} t^2. \tag{A.1}$$

First, we show that $\overline{\mathcal{F}}_{|\mathbf{x}} \subset \mathbb{R}^{n^+ + n^-}$ is convex so that eq. (A.1) is a convex program. Let $(t_0, (t_0'\mathbf{f}_0)), (t_1, (t_1'\mathbf{f}_1)) \in \mathcal{F}_{|\mathbf{x}}$. Letting $t = \lambda t_0 + (1-\lambda)t_1$, $t' = \lambda t_0' + (1-\lambda)t_1'$, and $\mathbf{f} = (\lambda t_0'/t')\mathbf{f}_0 + (1 - \lambda t_0'/t')\mathbf{f}_1$, we have $t \geq t'$, $\mathbf{f} \in \mathcal{F}_{|\mathbf{x}}$ due to convexity of $\mathcal{F}$, and $t'\mathbf{f} = \lambda t_0'\mathbf{f}_0 + (1-\lambda)t_1'\mathbf{f}_1$. Next we show that $(1, \mathbf{0}) \in \mathrm{relint}(\overline{\mathcal{F}}_{|\mathbf{x}})$. Let any $(t, t'\mathbf{f}) \in \mathcal{F}_{|\mathbf{x}}$ be given. Let $\mu = 2t/(2t-1)$ if $t > 1$ and otherwise $\mu = 2$. Then $\mu \in (1, 2]$ and $\mu/(2\mu - 1) \geq t$ so that $(1-\mu)t + \mu \geq (\mu - 1)t \geq (\mu - 1)t' \geq 0$. Note that $-\mathbf{f} \in \mathcal{F}_{|\mathbf{x}}$ by symmetry of $\mathcal{F}$. Therefore,

$$(1-\mu)(t, t'\mathbf{f}) + \mu(1, \mathbf{0}) = ((1-\mu)t + \mu, (\mu - 1)t'(-\mathbf{f})) \in \overline{\mathcal{F}}_{|\mathbf{x}}$$

So, by Thm. 6.4 of Rockafellar (1997), $(1, \mathbf{0}) \in \mathrm{relint}(\overline{\mathcal{F}}_{|\mathbf{x}})$. Letting $\mathbf{z}$ be defined by its constraint, we then have a Slater point. Note that $\ell_*(p) = \inf_z(pz - \ell(z)) = -h(p) - \log(2)$ for $p \in [0, 1]$ and otherwise $\ell_*(p) = -\infty$. By Prop. 5.3.2 (strong duality) of Bertsekas (1999), we have that eq. (A.1) is equal to

$$\begin{aligned}
&\inf_{\gamma\in\mathbb{R}^{n^+ + n^-}} \sup_{(t,\mathbf{f})\in\overline{\mathcal{F}}_{|\mathbf{x}}, \mathbf{z}\in\mathbb{R}^{n^+ + n^-}} \sum_{s=\pm} \sum_{i=1}^{n_+} (w_i^s \ell(z_i^s) - \gamma_i^s z_i^s) + \sum_{s=\pm} \sum_{i=1}^{n_+} s\gamma_i^s f_i^s - \frac{\psi}{2} t^2 \\
&= \inf_{\mathbf{p}\in\mathbb{R}^{n^+ + n^-}} \sum_{s=\pm} \sum_{i=1}^{n_+} \sup_{z_i^s} w_i^s (\ell(z_i^s) - p_i^s z_i^s) + \sup_{t\geq t'\geq 0} \left(t' \sup_{f\in\mathcal{F}_{|\mathbf{x}}} \sum_{s=\pm} \sum_{i=1}^{n_+} s w_i^s p_i^s f_i^s - \frac{\psi}{2} t^2 \right) \\
&= \inf_{0\leq\mathbf{p}\leq 1} \sum_{s=\pm} \sum_{i=1}^{n_+} w_i^s h(p_i^s) + \sup_{t\geq 0} \left(t\, \mathrm{IPM}_{\mathcal{F}}(\{(p_i^+ w_i^+, x_i^+)\}_{i=1}^{n^+}, \{(p_i^- w_i^-, x_i^-)\}_{i=1}^{n^-}) - \frac{\psi}{2} t^2 \right),
\end{aligned}$$

which yields the stated result for the two cases, $\psi > 0$ and $\psi = 0$, after taking sup over $t \geq 0$. □

Proof of Thm. 3. Fix $\mathbf{x}$ and $\mathbf{w}$. Write $\mathrm{DD}_{\mathcal{F},\psi}$ and $\mathrm{IPM}^2_{\mathcal{F}}$ for shorthand for the distances between the two weighted samples and let

$$H^2(\mathbf{p}) = \sum_{i=1}^{n_+} w_i^+ h(p_i^+) + \sum_{i=1}^{n_-} w_i^- h(p_i^-), \quad D^2(\mathbf{p}) = \mathrm{IPM}^2_{\mathcal{F}}(\{(p_i^+ w_i^+, x_i^+)\}_{i=1}^{n^+}, \{(p_i^- w_i^-, x_i^-)\}_{i=1}^{n^-}),$$

so that $\mathrm{DD}_{\mathcal{F},\psi} = \inf_{0\leq\mathbf{p}\leq 1} H^2(\mathbf{p}) + (\psi^{-1}/2) D^2(\mathbf{p})$. Note that for $p \in [0, 1]$, $D^2(p\mathbf{1}) = p^2\, \mathrm{IPM}^2_{\mathcal{F}}$. Note also that $h(p)$ is 4-strongly convex with a unique minimum at $p = 1/2$ so that $h(p) \geq 2(p - 1/2)^2 \geq 0$ and $h(p) = 0 \iff p = 1/2$. Hence, the objective in eq. (4.2) is nonnegative for all $\mathbf{p}$ and therefore, by Thm. 2, $\mathrm{DD}^2_{\mathcal{F},\psi} \geq 0$ and so $\mathrm{DD}_{\mathcal{F},\psi}$ is its well-defined square-root or infinite when $\mathrm{DD}^2_{\mathcal{F},\psi}$ is infinite.

Now suppose $\mathrm{IPM}_{\mathcal{F}} = 0$. Then $D^2(\mathbf{1}/2) = \mathrm{IPM}^2_{\mathcal{F}}/4 = 0$ and $H^2(\mathbf{1}/2) = 0$. Therefore, since $\mathbf{p} = \mathbf{1}/2$ is feasible in eqs. (4.2) and (4.3), we have that $\mathrm{DD}^2_{\mathcal{F},\psi} = 0$.

Now suppose $\mathrm{DD}^2_{\mathcal{F},\psi} = 0$. Then there exists $\mathbf{p}$ such that $D^2(\mathbf{p}) = H^2(\mathbf{p}) = 0$ since both functions are nonnegative. Since $H^2(\mathbf{p}) = 0$ and $w_i^s h(p_i^2)$ are nonnegative functions, we have $w_i^s h(p_i^s) = 0$ and hence $w_i^s > 0 \implies p_i^s = 1/2$ and consequently $w_i^s p_i^s = w_i^s/2$. Therefore, $0 = D^2(\mathbf{p}) = D^2(\mathbf{1}/2) = \mathrm{IPM}^2_{\mathcal{F}}/4$. □

Proof of Thm. 4. Reusing the notation from the proof of Thm. 3, for any $\mathbf{p}$, we have

$$\begin{aligned}
\left| D(\mathbf{p}) - \frac{1}{2}\operatorname{IPM}_{\mathcal{F}} \right| &= \sup_{f\in\mathcal{F}} \left| \sum_{s=\pm}\sum_{i=1}^{n^s} s w_i^s p_i^s f(x_i^s) \right| - \frac{1}{2}\sup_{f\in\mathcal{F}} \left| \sum_{s=\pm}\sum_{i=1}^{n^s} s w_i^s f(x_i^s) \right| \\
&\le \sup_{f\in\mathcal{F}} \left| \sum_{s=\pm}\sum_{i=1}^{n^s} s w_i^s \left(p_i^s - \frac{1}{2}\right) f(x_i^s) \right| \\
&\le M \sum_{s=\pm}\sum_{i=1}^{n^s} w_i^s \left| p_i^s - \frac{1}{2} \right| \\
&\le M \sqrt{\sum_{s=\pm}\sum_{i=1}^{n^s} w_i^s} \sqrt{\sum_{s=\pm}\sum_{i=1}^{n^s} w_i^s \left(p_i^s - \frac{1}{2}\right)^2} \\
&\le M \sqrt{\sum_{s=\pm}\sum_{i=1}^{n^s} w_i^s} \sqrt{\frac{1}{2}\sum_{s=\pm}\sum_{i=1}^{n^s} w_i^s h(p_i^s)} \\
&\le \frac{M\sqrt{w}}{\sqrt{2}} H(\mathbf{p}).
\end{aligned}$$

First, consider $\psi = 0$. Let $\mathbf{p}$ be such that $\operatorname{DD}_{\mathcal{F},\psi} = H(\mathbf{p})$ and $D(\mathbf{p}) = 0$. Applying the above bound, we get

$$\begin{aligned}
\operatorname{IPM}_{\mathcal{F}} &\le \sqrt{2} M \sqrt{w} \operatorname{DD}_{\mathcal{F},\psi} \\
&\le 2M\sqrt{w} \operatorname{DD}_{\mathcal{F},\psi} \\
&\le \max(2M\sqrt{w}, 4\sqrt{\psi}) \operatorname{DD}_{\mathcal{F},\psi}.
\end{aligned}$$

Next, consider $\psi > 0$. Let $\mathbf{p}$ be such that $\operatorname{DD}^2_{\mathcal{F},\psi} = H(\mathbf{p}) + \psi^{-1} D^2(\mathbf{p})/2$. Applying Jensen's inequality to the above bound, we get

$$\begin{aligned}
\operatorname{IPM}_{\mathcal{F}} &\le M\sqrt{2w} H(\mathbf{p}) + 2D(\mathbf{p}) \\
&\le \sqrt{4M^2 w H^2(\mathbf{p}) + 8D^2(\mathbf{p})} \\
&= \sqrt{4M^2 w H^2(\mathbf{p}) + 16\psi\psi^{-1} D^2(\mathbf{p})/2} \\
&\le \max(2M\sqrt{w}, 4\sqrt{\psi}) \operatorname{DD}_{\mathcal{F},\psi}.
\end{aligned}$$

On the other hand, because $\mathbf{1}/2$ is feasible in eq. (4.2),

$$8\psi \operatorname{DD}^2_{\mathcal{F},\psi} \le 8\psi H^2(\mathbf{1}/2) + 4D^2(\mathbf{1}/2) = \operatorname{IPM}^2_{\mathcal{F}}$$

and so $2\sqrt{2\psi}\operatorname{DD}_{\mathcal{F},\psi} \le \operatorname{IPM}_{\mathcal{F}}$. □

Proof of Thm. 5. We reuse the notation from the proof of Thm. 3. Thm. 4 showed that $2\sqrt{2\psi}\operatorname{DD}_{\mathcal{F},\psi} \le \operatorname{IPM}_{\mathcal{F}}$. To prove the present result, we further show that whenever $\psi \ge \overline{w}M^2$, we have $\sqrt{1 - \psi^{-1}\overline{w}M^2}\operatorname{IPM}_{\mathcal{F}} \le 2\sqrt{2\psi}\operatorname{DD}_{\mathcal{F},\psi}$. Toward that end, let $D^2(\mathbf{p}) = H^2(\mathbf{p}) + \psi^{-1}D^2(\mathbf{p})/2$ and let $\mathbf{p}^*$ be such that $\operatorname{DD}^2_{\mathcal{F},\psi} = D^2(\mathbf{p}^*)$. Let $W \in \mathbb{R}^{(n^+ + n^-)\times(n^+ + n^-)}$ be the diagonal matrix with w_i^s on its diagonal. WLOG, $W \succ 0$ because otherwise we can consider the equivalent problem after removing all the points with zero weight. Note that H^2 is a convex twice-differentiable function, that $\frac{\partial^2}{\partial\gamma^2}h(p) = 1/p + 1/(1-p) \ge 4$, and therefore $\nabla^2 H^2(\mathbf{p})$ is diagonal with entries bounded below by $4w_i^s$. Since D^2 is also convex, we have that $D^2(\mathbf{p}) - 2(\mathbf{p} - \mathbf{p}^*)^T W(\mathbf{p} - \mathbf{p}^*)$ is convex and therefore also has an optimum at $\mathbf{p}^*$. Therefore,

$$D^2(\mathbf{p}^*) \le D^2(\mathbf{1}/2) - 2(\mathbf{p} - \mathbf{1}/2)^T W(\mathbf{p} - \mathbf{1}/2).$$

Let $g \in \partial D^2(\mathbf{1}/2)$ be any subderivative of D^2 at $\mathbf{1}/2$. Then, by Cauchy-Schwartz and the above,

$$D^2(\mathbf{1}/2) - D^2(\mathbf{p}^2) \le g^T(\mathbf{p}^* - \mathbf{1}/2)$$

$$
\begin{aligned}
&= g^T W^{-1/2} W^{1/2} (\mathbf{p}^* - \mathbf{1}/2) \\
&\leq \sqrt{g^T W^{-1} g} \sqrt{(\mathbf{p}^* - \mathbf{1}/2) W (\mathbf{p}^* - \mathbf{1}/2)} \\
&\leq \sqrt{g^T W^{-1} g} \sqrt{D^2(\mathbf{1}/2) - D^2(\mathbf{p}^*)} / \sqrt{2}.
\end{aligned}
$$

Therefore,

$$
D^2(\mathbf{1}/2) - D^2(\mathbf{p}^*) \leq g^T W^{-1} g / 2.
$$

Note that $\partial H^2(\mathbf{1}/2) = 0$ so that $\partial D^2(\mathbf{1}/2) = (2\psi)^{-1} \partial D^2(\mathbf{1}/2) = \psi^{-1} D(\mathbf{1}/2) \partial D(\mathbf{1}/2) = (2\psi)^{-1} \operatorname{IPM}_{\mathcal{F}} \partial D(\mathbf{1}/2)$. Moreover,

$$
\begin{aligned}
\left| (\partial D(\mathbf{p}))_{s,i} \right| &= \left| \left(\partial_{\mathbf{p}} \sup_{f \in \mathcal{F}_{|\mathbf{x}}} \left| \sum_{s,i} s w_i^s p_i^s f_i^s \right| \right)_{s,i} \right| \\
&= \left| \left(\partial_{\mathbf{p}} \sup_{f \in \mathcal{F}_{|\mathbf{x}} \cup -\mathcal{F}_{|\mathbf{x}}} \sum_{s,i} s w_i^s p_i^s f_i^s \right)_{s,i} \right| \\
&\leq w_i^s \sup_{f \in \mathcal{F}_{|\mathbf{x}} \cup -\mathcal{F}_{|\mathbf{x}}} |f_i^s| \leq w_i^s M.
\end{aligned}
$$

Therefore,

$$
\frac{\operatorname{IPM}_{\mathcal{F}}^2}{8\psi} - \operatorname{DD}_{\mathcal{F},\psi}^2 = D^2(\mathbf{1}/2) - D^2(\mathbf{p}^*) \leq g^T W^{-1} g / 2 \leq \frac{\operatorname{IPM}_{\mathcal{F}}^2}{8\psi^2} \overline{w} M^2,
$$

which, when $\psi \geq \overline{w} M^2$, yields $2\sqrt{2\psi} \operatorname{DD}_{\mathcal{F},\psi}^2 \geq \sqrt{1 - \psi^{-1} \overline{w} M^2} \operatorname{IPM}_{\mathcal{F}}$ and hence the result. □

Proof of Thm. 6. Since $\mathcal{W}$ is a convex set, it remains only to be shown that the objective is convex. The second term, $\frac{\lambda}{n_1} \|W\|_2^2$, is clearly convex when $\lambda \geq 0$. The first term is the supremum over linear forms in W and is therefore also convex. □

Proof of Thm. 7. By Thm. 6, eq. (4.4) is convex. Since $W_i = n_1/n_0$ gives a Slater point, strong duality holds and there exists an optimal Lagrange multiplier ϕ^* such that eq. (4.4) is equivalent to $\mathcal{W}(\phi^*)$ and any optimizer $W^* \in \mathcal{W}(\phi^*)$ also optimizes eq. (4.4). It must therefore already satisfy the constraint $\sum_{i \in \mathcal{T}_0} W_i^* = n_1$ and hence its objective in eq. (5.1) is exactly the same as in eq. (4.4). For any other $W \in \bigcup_{\phi \geq 0} \mathcal{W}(\phi)$, the objective in eq. (5.1) is the same as the objective of $\frac{n_1 W}{\sum_{i \in \mathcal{T}_0} W_i}$ in eq. (4.4), where the latter is feasible. So. by its optimality in eq. (4.4), W^* also optimizes eq. (5.1). □

We use the following lemma in the proof of Thm. 8.

Lemma 9. *For nonnegative random variables $Z_n \geq 0$ and any sub-sigma algebra $\mathcal{G}$,*

$$
\mathbb{E}[Z_n \mid \mathcal{G}] = O_p(1) \implies Z_n = O_p(1).
$$

Note that when $\mathcal{G} = \sigma(Z_1, \dots)$ is the complete sigma algebra then $\mathbb{E}[Z_n \mid \mathcal{G}] = Z_n$ and the result is trivial and when $\mathcal{G} = \{\varnothing, \Omega\}$ is the trivial sigma algebra then $\mathbb{E}[Z_n \mid \mathcal{G}] = \mathbb{E}[Z_n]$ and the result is direct from Markov's inequality. The lemma provides a proof for the in-between cases.

Proof of Lemma 9. Suppose $\mathbb{E}[Z_n \mid \mathcal{G}] = O_p(1)$. Let $\nu > 0$ be given. Then $\mathbb{E}[Z_n \mid \mathcal{G}] = O_p(1)$ says that there exist N, M such that $\mathbb{P}(\mathbb{E}[Z_n \mid \mathcal{G}] > M) \leq \nu/2$ for all $n \geq N$. Let $M_0 = \max\{M, 2/\nu\}$. Then, for all $n \geq N$,

$$
\begin{aligned}
\mathbb{P}(Z_n > M_0^2) &= \mathbb{P}(Z_n > M_0^2, \mathbb{E}[Z_n \mid \mathcal{G}] > M_0) + \mathbb{P}(Z_n > M_0^2, \mathbb{E}[Z_n \mid \mathcal{G}] \leq M_0) \\
&= \mathbb{P}(Z_n > M_0^2, \mathbb{E}[Z_n \mid \mathcal{G}] > M_0) + \mathbb{E}[\mathbb{P}(Z_n > M_0^2 \mid \mathcal{G}) \mathbb{I}[\mathbb{E}[Z_n \mid \mathcal{G}] \leq M_0]] \\
&\leq \nu/2 + \mathbb{E}\left[\frac{\mathbb{E}[Z_n \mid \mathcal{G}]}{M_0^2} \mathbb{I}[\mathbb{E}[Z_n \mid \mathcal{G}] \leq M_0] \right] \leq \nu/2 + 1/M_0 \leq \nu.
\end{aligned}
$$

□

Proof of Thm. 8. Let $Z = \sum_{i\in\mathcal{T}_0} \frac{e(X_i)}{1-e(X_i)}$ and $\tilde{W}_i = \frac{n_1}{Z}\frac{e(X_i)}{1-e(X_i)}$. First, note that by assumption $\tilde{W} \in \mathcal{W}(\Theta_g)$ is feasible. Moreover,

$$\begin{aligned}
\mathrm{IPM}_{\mathcal{F}}(\{(\tfrac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\tfrac{\tilde{W}_i}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) &= \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n \left(\frac{T_i}{n_1} - \frac{(1-T_i)e(X_i)}{Z(1-e(X_i))} \right) f(X_i) \right| \\
&= \frac{n_1}{Z} \sup_{f\in\mathcal{F}} \left| \frac{1}{n_1} \sum_{i=1}^n \left(\frac{ZT_i}{n_1} - \frac{(1-T_i)e(X_i)}{(1-e(X_i))} \right) f(X_i) \right| \\
&\le \frac{n_1}{Z} \sup_{f\in\mathcal{F}} \left| \frac{1}{n_1} \sum_{i=1}^n \left(\frac{ZT_i}{n_1} - T_i \right) f(X_i) \right| \\
&\quad + \frac{n_1}{Z} \sup_{f\in\mathcal{F}} \left| \frac{1}{n_1} \sum_{i=1}^n \left(T_i - \frac{(1-T_i)e(X_i)}{(1-e(X_i))} \right) f(X_i) \right| \\
&\le \underbrace{\frac{n_1}{Z} M \left| \frac{Z}{n_1} - 1 \right|}_{A} + \underbrace{\frac{n_1}{Z}\frac{n}{n_1} \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \left(T_i - \frac{(1-T_i)e(X_i)}{(1-e(X_i))} \right) f(X_i) \right|}_{B}
\end{aligned}$$

We will next bound each of the two terms, A and B.

Note that $\mathbb{E}[Z] = n_1$. By assumption, there exists η with $1 - e(x) > \eta > 0$. Therefore, since $\mathbb{E}\left[e^2(X_i)/(1-e(X_i))^2\right] < 1/\eta^2$, Chebychev's inequality yields $|Z - n_1| = O_p(\sqrt{n})$ and hence $\left|\frac{Z}{n_1} - 1\right| = O_p(1/\sqrt{n})$. Since this also means that $n_1/Z \to_p 1$, Slutsky's theorem yields that $A = O_p(1/\sqrt{n})$.

Next, let $\Delta_n = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \left(T_i - \frac{(1-T_i)e(X_i)}{(1-e(X_i))} \right) f(X_i) \right|$. Note that Δ_n satisfies bounded differences with $c_i = 2M/(n\eta)$. Therefore, by McDiarmid's inequality, $\Delta_n \le \mathbb{E}\Delta_n + O_p(1/\sqrt{n})$. Since $\mathbb{E}\left[T_i - \frac{(1-T_i)e(X_i)}{(1-e(X_i))}\right] = 0$, symmetrization yields that

$$\mathbb{E}\Delta_n \le 2\mathbb{E}_{T_{1:n},X_{1:n}}\mathbb{E}_{\xi_{1:n}} \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \left(T_i - \frac{(1-T_i)e(X_i)}{(1-e(X_i))} \right) f(X_i) \right|,$$

where ξ_i are iid Rademacher random variables. Finally, by the Ledoux-Talagrand comparison lemma (Ledoux & Talagrand, 1991, Thm. 4.12), since $\left| T_i - \frac{(1-T_i)e(X_i)}{(1-e(X_i))} \right| \le 1/\eta$ we have from the above that

$$\mathbb{E}\Delta_n \le \frac{2}{\eta}\mathbb{E}\mathfrak{R}_n(\mathcal{F}),$$

Finally, since $\mathfrak{R}_n(\mathcal{F})$ satisfies bounded differences itself with $c_i = 2M/n$, McDiarmid's inequality yields that $\mathbb{E}\mathfrak{R}_n(\mathcal{F}) \le \mathfrak{R}_n(\mathcal{F}) + O_p(1/\sqrt{n})$. Since $n_1/Z \to_p 1$ and $n/n_1 \to_p 1/\mathbb{P}(T_1 = 1)$, Slutsky's theorem yields $B = O_p(\mathfrak{R}_n(\mathcal{F}) + 1/\sqrt{n})$.

All together, we see that $\mathrm{IPM}_{\mathcal{F}}(\{(\frac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\frac{\tilde{W}_i}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) = O_p(\mathfrak{R}_n(\mathcal{F}) + 1/\sqrt{n})$. Consequently, for any $\psi > 0$, by Thm. 4, we have that $\mathrm{DD}_{\mathcal{F},\psi}(\{(\frac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\frac{\tilde{W}_i}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) = O_p(\mathfrak{R}_n(\mathcal{F}) + 1/\sqrt{n})$.

Next, note that since $\mathbb{E}\left[e^4(X_i)/(1-e(X_i))^4\right] < 1/\eta^4$, Chebychev's inequality yields $\frac{1}{n}\sum_{i=1}^n e^2(X_i)/(1-e(X_i))^2 = O_p(1)$, which, since $n_1^2/Z^2 \to_p 1$, yields by Slutsky's theorem that $\|\tilde{W}\|_2^2 = O_p(n)$. We conclude that the objective of $\tilde{W}$ in eq. (4.4) satisfies

$$\mathrm{DD}^2_{\mathcal{F},\psi}(\{(\tfrac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\tfrac{\tilde{W}_i}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) + \tfrac{\lambda}{n_1^2}\|\tilde{W}\|_2^2 = O_p(\mathfrak{R}_n^2(\mathcal{F}) + 1/n).$$

Since W^* is optimal and $\tilde{W}$ is feasible, we must therefore also have

$$\mathrm{DD}^2_{\mathcal{F},\psi}(\{(\tfrac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\tfrac{W_i^*}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) + \tfrac{\lambda}{n_1^2}\|W^*\|_2^2 = O_p(\mathfrak{R}_n^2(\mathcal{F}) + 1/n),$$

and, since each term is nonnegative and by Thm. 4 for $\psi > 0$, consequently

$$\mathrm{IPM}_{\mathcal{F}}(\{(\tfrac{1}{n_1}, X_i)\}_{i\in\mathcal{T}_1}, \{(\tfrac{W_i^*}{n_1}, X_i)\}_{i\in\mathcal{T}_0}) = O_p(\mathfrak{R}_n(\mathcal{F}) + 1/\sqrt{n}), \quad \frac{1}{n_1^2}\|W^*\|_2^2 = O_p(\mathfrak{R}_n^2(\mathcal{F}) + 1/n).$$

Because $\mu_0 = f(x;\theta_f)$ and $R(\cdot)$ is degree-2 homogeneous, we have that $|B(W^*,\mu_0)| \leq R(\theta_f)\text{IPM}_{\mathcal{F}}(\{(\frac{1}{n_1},X_i)\}_{i\in\mathcal{T}_1},\{(\frac{W_i^*}{n_1},X_i)\}_{i\in\mathcal{T}_0})$ and since $\sigma^2(X)\leq\overline{\sigma}^2$ is almost surely bounded we have $V^2(W;\sigma^2)\leq \overline{\sigma}\left(\frac{1}{n_1}+\frac{1}{n_1^2}\|W^*\|\right)$. By eq. (2.5) (see Kallus, 2016, Thm. 1), $\mathbb{E}\left[(\hat{\tau}_W-\tau_{n_1})^2 \mid X_{1:n},T_{1:n}\right]=O_p(\mathfrak{R}_n^2(\mathcal{F})+1/n)$. Hence, by Jensen's, inequality $\mathbb{E}\left[|\hat{\tau}_W-\tau_{n_1}| \mid X_{1:n},T_{1:n}\right]=O_p(\mathfrak{R}_n(\mathcal{F})+1/\sqrt{n})$. Therefore, by Lemma 9, we then must also have that $\hat{\tau}_{W^*}-\tau_{n_1}=O_p(\mathfrak{R}_n(\mathcal{F})+1/\sqrt{n})$. Noting that, by LLN, $\tau_{n_1}-\tau=O_p(1/\sqrt{n})$ completes the proof. □

A.1. Detail on twins example

The 10 high-impact variables included in the treatment model are

1. gestat10: gestational age
2. hydra: risk factor for hydramnios/oligohydramnios
3. incervix: risk factor for incompetent cervix
4. nprevistq: quintile number of prenatal visits
5. csex: sex of child
6. anemia: risk factor for Anemia
7. uterine: risk factor for uterine bleeding
8. dfageq: octile age of father
9. mager8: mom age
10. adequacy: adequacy of care