Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects

N.B. When citing this work, cite the original published paper.

(article starts on next page)

Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects

Fredrik D. Johansson   FREDRIK.JOHANSSON@CHALMERS.SE
Chalmers University of Technology
Göteborg, 412 96, Sweden

Uri Shalit   URISHALIT@TECHNION.AC.IL
Technion - Israel Institute of Technology
Haifa, 3200003, Israel

Nathan Kallus   KALLUS@CORNELL.EDU
Cornell University
New York, NY 10044, USA

David Sontag   DSONTAG@CSAIL.MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract

Practitioners in diverse fields such as healthcare, economics and education are eager to apply machine learning to improve decision making. The cost and impracticality of performing experiments and a recent monumental increase in electronic record keeping has brought attention to the problem of evaluating decisions based on non-experimental observational data. This is the setting of this work. In particular, we study estimation of individual-level potential outcomes and causal effects—such as a single patient's response to alternative medication—from recorded contexts, decisions and outcomes. We give generalization bounds on the error in estimated outcomes based on distributional distance measures between re-weighted samples of groups receiving different treatments. We provide conditions under which our bounds are tight and show how they relate to results for unsupervised domain adaptation. Led by our theoretical results, we devise algorithms which learn representations and weighting functions that minimize our bounds by regularizing the representation's induced treatment group distance, and encourage sharing of information between treatment groups. Finally, an experimental evaluation on real and synthetic data shows the value of our proposed representation architecture and regularization scheme.

Keywords: Causal effects, overlap, generalization theory, domain adaptation.

1. Introduction

Evaluating intervention decisions is a key question in many diverse fields including medicine, economics, and education. In medicine, an optimal choice of treatment for a patient in the intensive care unit may mean the difference between life and death. In public policy, job reforms have impact on the unemployment rate and the economy of a nation. To evaluate such interventions, we must study their *causal effect*—the difference in an outcome of inter-

est under alternative choices of intervention. Since only one option may be carried out at a time, any data to support such evaluations only reveals the outcome of the action taken and never the outcome of the action not taken, which remains an unknown *counterfactual*. To estimate causal effects, we must therefore infer what would have happened had we made another decision, predicting the *potential outcomes* (Rubin, 2005) of unexplored interventions. Furthermore, to decide on personalized interventions, such as tailoring treatments to patients, we must understand individual-level causal effects, conditioned on the available information on an individual recorded prior to intervention.

In this work, we study estimation of individual-level potential outcomes and causal effects from non-experimental, *observational* data under a machine learning, or risk minimization, perspective. An observational data set consists of historical records of interventions, the contexts in which they were made, and the observed outcomes. Our running example is that of patients represented by their medical history, the medication they were prescribed and the outcome of treatment, such as mortality. An individual-level effect measures the causal effect of medication choice, conditioned on what is known about the patient. Finally, though we know which interventions took place, the policy by which interventions were chosen in this data is typically unknown to us.

Working with observational data is our best bet when experiments such as randomized controlled trials (RCT) are infeasible, impractical or prohibitively expensive. While cheaper and easier to implement, observational studies come with new, fundamental difficulties. Perhaps the most challenging of these is *confounding*—influence of variables that are causal of both the intervention and the outcome, and may introduce spurious, non-causal correlations between the two (Pearl, 2009). For example, richer patients might both have more access to certain medications and have better outcomes regardless of medication, making such medications appear better than they might be. Similarly, job training might only be given to those motivated enough to seek it. Naïve estimates of causal effects may therefore be biased by subsuming the effect of confounding variables on the outcome. Here, we make the common assumption that confounding variables, such as wealth or motivation in the examples above, have been measured and can be adjusted for in our estimation. This, however, introduces another difficulty, which is contending with the systematic differences in such variables between different treatment groups. Moreover, if these groups only partially overlap in terms of variables causal of the outcome, consistent estimation of causal effects (estimates that converge asymptotically to the true effect) may not always be guaranteed.

Causal estimation from observational data has been studied extensively in the statistics, econometrics, and computer science literature, often focused on *average* effects in a population or on simple models of effect heterogeneity. With sights set on personalization based on rich data, more flexible models are required, and machine learning is more often considered for the task. When we can no longer make strong assumptions such as linearity or low dimensionality, new questions arise: How well will our models generalize? What should be our criteria for fitting them? How should we regularize them? What assumptions are necessary for good performance guarantees or asymptotic consistency? What can be said when these assumptions are not met? With this paper, we begin to answer these questions.

We study estimation of both (i) potential outcomes under interventions and (ii) conditional average treatment effects (CATE), by risk minimization with flexible model classes. We show that these problems involve generalization of predictions under distributional shift

2

and how such predictions can be improved using sample re-weighting. In doing so, we draw connections to unsupervised domain adaptation and show that one solution to CATE estimation is to solve two (possibly dependent) domain adaptation problems. In contrast to most theoretical results in observational causal estimation, bounds obtained using the risk minimization approach are applicable also in finite samples and for misspecified classes of hypotheses. We use distributional measures of distance between treatment groups to give upper bounds on the marginal risk of hypotheses in a given class—risk caused by model bias, sub-optimal sample weighting or lack of treatment group overlap.

Sample weighting lies at the heart of many classical approaches to causal estimation, particularly using the *propensity score*—the probability for a subject to receive treatment under the observed policy, given their characteristics. In practice, the propensity score is typically unknown[1] and replacing it with an estimate is prone to introduce bias. However, biased but well-chosen weights may help control variance in small samples (Swaminathan and Joachims, 2015; Hirano et al., 2003). Our bounds make this trade-off explicit through the measure of treatment group distance: the more uniform (and possibly more biased) the weights, the larger the treatment group distance, but the smaller the variance.

Our theoretical results are then extended to give generalization bounds for learned invertible representations of the input space, representations in which the distributional distance may be smaller than the original space; see illustration in Figure 5. Combining representation learning with re-weighing allows both for tight bounds when the sample size is large and treatment groups overlap, making the treatment group density ratio an admissible weight, and for a principled bias-variance trade-off for smaller samples. Notably, these guarantees are valid under appropriate assumptions also when treatment and control groups do not fully overlap. In line with our theoretical results, we devise representation learning algorithms that minimize the weighted empirical risk of predicted potential outcomes, regularized by the distributional distance between treatment groups in the representation space. We give conditions under which our algorithms are consistent estimators of the causal effect. Finally, we evaluate our framework on synthetic and real-world benchmark tasks and demonstrate the value of representation learning, sample weighting and our proposed regularization scheme.

Parts of this work have been published in conference proceedings (Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2019). This manuscript extends these works considerably, foremost by developing generalization bounds and accompanying algorithms that support re-weighting of the relevant risk and distributional distance. This is a significant conceptual change since it allows to give conditions under which our algorithms lead to consistent estimation. In addition, we provide a longer, more self-contained theoretical exposition and compare it to both older results and recent developments.

1.1 Notation and Terminology

Random variables are denoted with capital roman letters $A, B, C, \ldots$ and observations thereof with a corresponding indexed lower-case letter $a_i, b_i, c_i, \ldots$. The empirical density of a draw of m samples from a density p is denoted $\hat{p}^m$. Unless stated otherwise, all random

---

1. This has some notable exceptions such as in advertising, in which an existing policy for serving ads was designed and known to the advertiser (Swaminathan and Joachims, 2015; Lefortier et al., 2016).
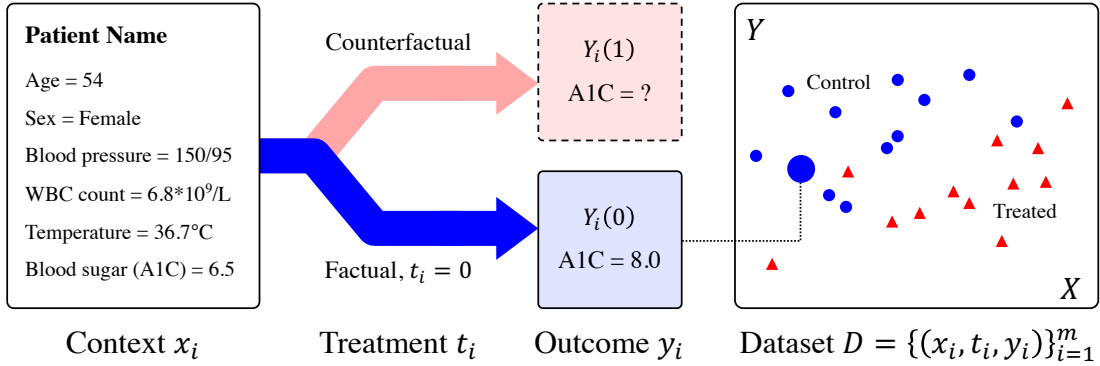
Figure 1: Formation of an observational data set of patient contexts X, treatments T and outcomes Y in an imagined application to diabetes management. Treatments are assigned based on the context and determines which of the two potential outcomes, $Y(0)$ and $Y(1)$, is observed for the patient, here in terms of blood sugar (A1C). The observations are then recorded in the data set D. The multivariate context X is represented in one dimension for illustration purposes. In the figure on the right, red and blue colors denote which treatment the patient received.

variables are distributed according to a fixed distribution $p(A, B, ...)$. Expectations over a variable X distributed according to $p(X)$ are denoted $\mathbb{E}_X[\cdot]$, and conditional expectations over X given Y distributed according to $p(X \mid Y)$, $\mathbb{E}_{X|Y}[\,\cdot \mid Y = t]$. When expectations are defined w.r.t. a density q, different from p, the notation $\mathbb{E}_{X \sim q}[\cdot]$ is used.

2. Potential Outcomes and Effects of Interventions

We introduce the problem of estimating potential outcomes and effects of interventions from observational data. Throughout the paper, we adopt the running example of estimating the effect of a medical treatment on a patient. This informs our choice of terminology and notation and serves to give intuition for the mathematical quantities involved. However, the applicability of the theory and algorithms described are in no way limited to this domain.

Using the Neyman-Rubin potential outcome framework (Imbens and Rubin, 2015), we associate each unit (e.g., patient) $i = 1, \ldots, m$ with the following partially observed random variables, illustrated by samples in Figure 1.

- A *context* $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, defined by all information observed about a patient before the choice of treatment is determined. These covariates may influence both the treatment choice and the outcome of an experiment. The context is represented as a d-dimensional real-valued vector throughout this paper.

- A *treatment* $T_i \in \mathcal{T} = \{0, 1\}$, which is an intervention performed in an observed context. Treatments are assumed to be binary variables throughout this paper,[2] where $T_i = 1$ is usually referred to as "treatment" and $T_i = 0$ as "control."

---

2. We discuss generalizations of our results to non-binary settings in Section 8.

- An *outcome* $Y_i \in \mathcal{Y} \subset \mathbb{R}$, measuring an aspect of interest of a patient, such as blood pressure or mortality, after the administration of a treatment, represented by a real-valued variable.

- *Partially observed potential outcomes* $Y_i(0), Y_i(1) \in \mathcal{Y} \subseteq \mathbb{R}$ that correspond to the outcomes that *would* be observed for unit i under treatments $T = 0$ and $T = 1$, respectively. We assume that $Y_i = Y_i(T_i)$ throughout the paper, capturing both *consistency* and *non-interference*, also known in conjunction as the stable unit treatment value assumption (SUTVA), which is key to the existence of potential outcomes and the relevance of this hypothetical construct to the actual data (Rubin, 2005).

Access to both potential outcomes $Y_i(0)$ and $Y_i(1)$ would allow us to retroactively determine which treatment $t \in \{0, 1\}$ *would have been* preferable for unit i, or proactively predict which one *will be*. If the individual treatment effect (ITE), $Y_i(1) - Y_i(0)$ is positive and higher outcomes are preferred, $t = 1$ is the better choice for i. However, the potential outcome of the treatment not administered, $Y_i(1 - T_i)$, is an *unobserved counterfactual*, which is the key impediment to assessing the ITE. That we do not observe what would have happened if we did something differently is often termed *the fundamental problem of causal inference*. Without strong assumptions, the best we can hope for is instead to infer or predict the following quantities (Künzel et al., 2019).

Definition 1 (Main estimands)**.** The expected potential outcomes μ_t, for treatments $t \in \{0, 1\}$, conditioned on a context $X = x$, are

$$\mu_t(x) := \mathbb{E}_{Y(t)|X}[Y(t) \mid X = x], \quad \text{for } t \in \{0, 1\} , \tag{1}$$

and the conditional average treatment effect (CATE) given a context x is

$$\tau(x) := \mathbb{E}_{Y(0),Y(1)|X}[Y(1) - Y(0) \mid X = x] = \mu_1(x) - \mu_0(x) . \tag{2}$$

The CATE is an object of key interest as it tells us what is the best prediction of the effect on an individual given only their context variables. This has a variety of uses. One important use is the personalization of treatment to make sure that the treatment is effective for the target. We illustrate μ_0, μ_1 and τ in Figure 3 in Section 4.

We consider learning from a simple random sample of size m from a population distributed according to a population distribution $p(X, T, Y)$. The observations of a unit i are denoted (x_i, t_i, y_i). These data are assumed to be sampled i.i.d. from p. For this reason, we drop the subscript i when dealing with the distribution of any single such random draw from this population. Note that since counterfactual outcomes are unobserved, our data consists just of $(X_1, T_1, Y_1), \ldots, (X_m, T_m, Y_m)$, and $Y(1 - T)$ remains unobserved. Following a long tradition, observed samples from $p(X \mid T = 0)$ and $p(X \mid T = 1)$ are referred to as the control and treated groups, respectively. With slight abuse of terminology, we use these labels also in reference to the conditional densities themselves. For convenience, for $t \in \{0, 1\}$, we introduce the short-hands $p_t(X) := p(X \mid T = t)$ and $\hat{p}_t(X)$ analogously. We assume that $p(X), p_1(X), p_0(X)$ are all probability measures on $\mathcal{X}$, absolutely continuous w.r.t. the Lebesgue measure.
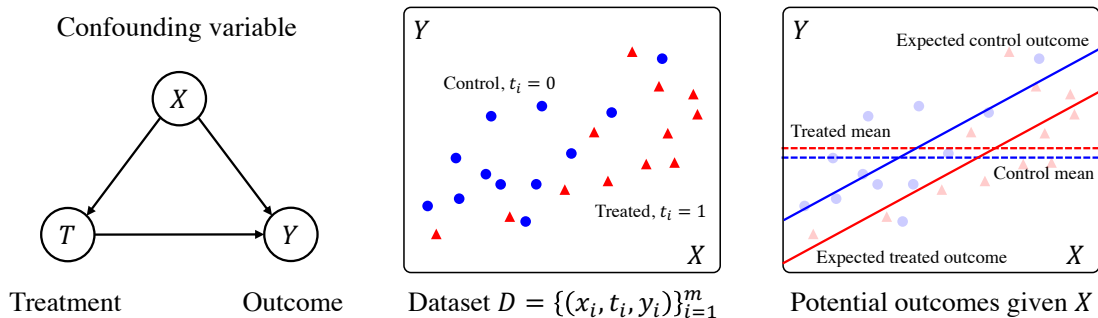
Figure 2: Effects of confounding in observational data. A variable X influences both treatment assignments and outcomes (left) observed in the data set D (middle). This manifests as a distribution difference in X between control (blue) and tread (red) observations. Despite the causal effect being negative for all X (illustrated by the red solid line being below the blue solid line, right), the average outcome in the control group (dashed, blue) is lower than the average outcome in the treated group (dashed, red).

Treatment group imbalance and confounding bias. In practical applications, different treatments are given in different contexts. For example, more expensive drugs may only be available to patients with sufficient insurance coverage. As a result, a central difficulty in working with observational data, unlike in randomized experiments, is that treatment and control groups can not be assumed equivalent. When treatments T are selected based on the observed context X, this presents as a distributional difference, $p_0(X) \neq p_1(X)$. We illustrate treatment group imbalance and the effect it may have in Figure 2. If not adjusted for, this imbalance may result in confounding bias and inflated variance in estimates of potential outcomes and causal effects. Mitigating this bias and inflated variance in risk minimization is one of the main challenges addressed in this work.

Aside: Individual treatment effect. The conditional average treatment effect conditioned on everything that is known about a subject captures a causal effect which is as specific to an individual as the data supports, rather than a population-level causal effect. It is therefore sometimes called *individual treatment effect* (ITE) (Johansson et al., 2016; Shalit et al., 2017). While this definition aligns with concepts used in machine learning, it overloads existing terminology. In causal inference, the ITE has been previously defined as the difference $Y_i(1) - Y_i(0)$. The distinction between this ITE and CATE is that ITE is unique to an individual and may not be described exactly by any set of features X. For this reason, we adopt the more precise label CATE for the feature-conditional treatment effect function $\tau(x)$.

2.1 Identifying Assumptions

Expected potential outcomes μ_t and conditional effects τ are said to be *identifiable* if they can be uniquely computed from the distribution $p(X, T, Y)$ of the observed data. This is important because that distribution is the most one could ever hope to learn from an i.i.d. sample from p, and so anything that cannot be learned from the distribution cannot be learned from i.i.d. samples from it. Without additional assumptions on p, the functions μ_t

and τ may not be identifiable. To see this, consider an observational study where treatment was given only to subjects over the age of 30, and the control group consists only of subjects under the age of 30. If age has an effect on a potential outcome of interest, there is no guarantee that it can be estimated for all subjects from such data.

Sufficient conditions for identification have been studied in both very general settings (Pearl, 2009; Rubin, 2005) and in special cases that are commonly accepted in real-world applications. In most of this work, we adopt and refer to the following assumptions.

Assumption 1 (Ignorability). *The potential outcomes $Y(0), Y(1)$ and the treatment T are conditionally independent given X,*

$$Y(0), Y(1) \perp\!\!\!\perp T \mid X .$$

Ignorability is often imprecisely (see remark below) called the *no unmeasured confounders* assumption as it holds if all variables which affect both treatment and potential outcomes are included in X *and* if these satisfy additional conditions.

Assumption 2 (Overlap). *In any context $x \in \mathcal{X}$, any treatment $t \in \{0, 1\}$ has a non-zero probability of being observed in the data*

$$\forall x \in \mathcal{X}, t \in \{0, 1\} : p(T = t \mid X = x) > 0 .$$

Equivalently, the population $p(X)$ is absolutely continuous w.r.t. $p_0(X), p_1(X)$.

Overlap is sufficient to ensure that knowledge of the outcomes in one treatment group may be generalized to the opposite group given access to a large enough sample size. Note that overlap only requires that the supports of the treatment groups are equal, not that they have similar densities. The degree to which treatment group densities are equal on this support is sometimes referred to as *balance*.

Under Assumptions 1, 2 and SUTVA ($Y = Y(T)$), expected potential outcomes and the conditional average treatment effect are identifiable as can be seen by the simple identity

$$
\begin{aligned}
\tau(x) &= \mathbb{E}_{Y(0),Y(1)|X} \left[Y(1) - Y(0) \mid X = x \right] \\
&= \mathbb{E}_{Y|X,T} \left[Y \mid X = x, T = 1 \right] - \mathbb{E}_{Y|X,T} \left[Y \mid X = x, T = 0 \right] .
\end{aligned}
$$

Plausibility of the identifying assumptions. Both ignorability and SUTVA are assumptions that are fundamentally untestable given observational data alone (Pearl, 2009). Despite this, they are often made in practice to justify subsequent analysis, or make clear its potential limitations. A common heuristic motivation for ignorability is related to the richness of the variable X. The richer the data, the more likely are they to cover all confounding variables. It should be noted, that even if all confounders are measured, adjusting for some of them may introduce additional estimation bias or variance nonetheless (Ding et al., 2017). Furthermore, the overlap assumption becomes increasingly difficult to both satisfy and check as the dimensionality of X grows (D'Amour et al., 2021).

3. Related Work

Observational causal inference has received considerable renewed attention with the recent growth of research into machine learning and artificial intelligence. We survey recent results and briefly cover relevant fundamental work below. The field may be broadly divided into two categories, dedicated to causal discovery and causal effect estimation, respectively. In the former, the direction and presence of causal relationships between observed variables is unknown, and the task is to infer them from data (Geiger et al., 2015; Spirtes and Glymour, 1991; Hoyer et al., 2009; Eberhardt, 2008; Hyttinen et al., 2014; Silva et al., 2006). In the latter, which is the setting of this work, the structure of causal relationships is assumed to be known: confounders X are causal of treatment T and outcome Y; treatment T is causal of Y (unless the effect is 0); any unmeasured variable is causal only of either X, T or Y. We are primarily interested in estimating the *conditional* average treatment effect of the treatment T on the outcome Y conditioned on the context X (Johansson et al., 2016; Shalit et al., 2017; Athey and Imbens, 2016; Wager and Athey, 2018; Pearl, 2017; Abrevaya et al., 2015; Bertsimas et al., 2017; Green and Kern, 2010; Alaa and Schaar, 2018).

Outcome regression estimators. Under ignorability, CATE is identified by the difference of two conditional expectations: the expected outcome under treatment conditioned on covariates minus the expected outcome under control conditioned on covariates. In an early influential paper, Hill (2011) demonstrated the power in separately estimating each conditional expectation, utilizing Bayesian Additive Regression Trees (Chipman et al., 2010) for this task. The approach of fitting treatment-specific outcome regressions was later referred to as a "T-learner" (Künzel et al., 2019), where T stands for *two*, see Figure 6a. Another approach, classically used with linear models, is to fit a single regression model from the covariates X *and* the treatment T to the outcome Y. CATE is then estimated by evaluating the difference between the prediction for $X, T = 1$ and $X, T = 0$. This approach is known as "S-learner", where S stands for *single*. It has been argued that both methods are prone to *compounding bias* when applied in high-dimensional, strongly confounded, small-sample settings which may require substantial regularization (Nie and Wager, 2021; Hahn et al., 2020). Rather than regressing on potential outcomes jointly or separately, a variety of work has studied directly estimating their difference, e.g. using trees (Athey and Imbens, 2016) and forests (Wager and Athey, 2018). Other work has studied meta-learners that combine base learners of the underlying regression functions using methods which are different from simple differencing in efforts to be more sample efficient (Robins et al., 2000; Künzel et al., 2019). Double/debiased machine learning (Chernozhukov et al., 2017), pseudo-outcome regression (Kennedy et al., 2017; Kennedy, 2020) and the R-learner (Nie and Wager, 2021) also attempt to mitigate bias by separately fitting non-parametric nuisance functions and effect estimates based on these. Multiple works have shown that when having a well-specified or non-parametric estimator for each regression, CATE estimation can be asymptotically consistent and/or asymptotically normal with optimal convergence rates, under appropriate assumptions (Chernozhukov et al., 2017; Belloni et al., 2014; Nie and Wager, 2021).

Representation learning. Representation learning has become a cornerstone of machine learning (Bengio et al., 2013), and is a natural component of causal effect estimators. In Johansson et al. (2016); Shalit et al. (2017), we used representation learning to share in-

formation about the outcome across treated and controls, while regularizing representation distributional distance between the groups. Zhang et al. (2020); Johansson et al. (2019) recognized that such regularization may result in loss of ignorability in the representation and proposed to learn representations in which context information is preserved but where treatment groups overlap. Hassanpour and Greiner (2019) identified the potential value in learning representations which disentangle parts of the contexts which are causes of a) treatment assignment, b) potential outcomes or c) both. By fitting a generative adversarial network, Yoon et al. (2018) estimate the uncertainty in predicted counterfactuals using their GANITE model—typical regression estimators, including representation learning methods, model only the conditional expectation of the outcomes given the input. Chen et al. (2019) used multi-task representation learning to estimate treatment effects from electronic healthcare records. In this work, we show that simultaneous representation learning and re-weighting may overcome the issue of destructive representations encountered in early deep learning approaches to CATE estimation. A related idea was used by Kallus (2020a) to balance treatment groups using weights defined in a representation space learned using a deep network adversary.

Sample weighting estimators. A complement to regression estimation of CATE is *sample weighting*, where the goal is to alleviate systematic differences in baseline covariates across treatment groups. This idea is used in propensity weighting methods (Austin, 2011; Rosenbaum and Rubin, 1983; Jung et al., 2020a) which use the observed treatment policy to re-weight samples for causal effect estimation, and more generally in re-weighted regression, see e.g. (Freedman and Berk, 2008; Jung et al., 2020b). Two challenges with propensity weighting methods are 1) the need to estimate the propensity score when it is unknown, and 2) the high variance introduced when the propensities are small, so that minute estimation errors lead to dividing by near-zeros. Interestingly, solving one of the above problems may mitigate also the other; Hirano et al. (2003); Abadie and Imbens (2016) showed that weighting or matching using the estimated propensity score may be more favorable than using the true propensity score when the increase in bias is smaller than the reduction in variance. We observe a similar phenomenon in our proposed regularization scheme. Gretton et al. (2009); Kallus (2020b, 2017) proposed learning sample weights by minimizing a distributional distance between weighted samples from different treatment groups. These rely on specifying the data representation a priori, without regard for which aspects of the data matter for outcome prediction. We build on the sampling weighting (importance sampling) literature by developing theory and algorithms for weighted risk minimization for potential outcomes and CATE, both in a fixed representation and one learned from data. A similar idea was also used in concurrent work by Kallus (2020a).

Learned prognostic scores. Our work on representation learning has conceptual ties to the idea of the *prognostic score* (Hansen, 2008). A prognostic score is any function $\Phi(X)$ of the context X that Markov separates $Y(0)$ and X, such that $Y(0) \perp\!\!\!\perp X \mid \Phi(X)$. An extreme example is $\Phi(X) = X$. If $Y(0)$ follows a generalized linear model, then $\Phi(X) = \mathbb{E}\left[Y(0)|X\right]$ is also a prognostic score. The prognostic score is a form of dimension reduction which under certain assumptions is sufficient for causal inference. Note that unlike the propensity score, the prognostic score might very well be vector valued. One can view our approach as attempting to find approximate non-linear prognostic functions for both $Y(0)$ and $Y(1)$.

We stress the approximate, because in fact we trade off how well our learned representation $\Phi(\cdot)$ is sufficient to explain the potential outcomes with a balancing objective which we show is important for good *finite-sample* estimation of CATE.

Assumptions on observational data. Estimation of causal effects from observational data is mostly performed under the assumptions of *ignorability*, treatment group *overlap* and *consistency*, as they are otherwise generally *unidentifiable*. In this work, we are motivated by the ignorability assumption throughout, but give several results that hold in its absence. For work on CATE estimation without ignorability, see e.g., Kallus and Zhou (2018); Kallus et al. (2018, 2019); Louizos et al. (2017); Rosenbaum (2002); Jesson et al. (2021). In contrast, we focus on the case where overlap is only partially satisfied. Lack of overlap is widely acknowledged as a problem (D'Amour et al., 2021) but estimation in this setting has received considerably less attention in the literature.

Learning guarantees. In this work, we provide generalization bounds for the estimation of potential outcomes as functions of an observed context, based on the distributional difference between treated and controls. Alaa and Schaar (2018) showed that the best achievable minimax rate for non-parametric estimation in this task is independent of treatment group differences (selection bias). However, this main result is limited to the asymptotic regime. In low-sample settings, the error in estimating CATE is still affected by selection bias, and the best estimator may be one that adjusts for it. Künzel et al. (2019); Nie and Wager (2021); Kennedy (2020) also studied the minimax rate of CATE estimators and proposed methods which combine estimates of nuisance functions which achieve the optimal rate under smoothness conditions. In contrast, estimates based on parametric or otherwise restricted hypothesis classes may not approach the optimal minimax rate (see paragraph below). Non-parametric estimation was also studied by Curth and Schaar (2021), in the context of different meta-learners (e.g., T-learner, S-learner). They find, similarly, that the best asymptotic estimators and the best low-sample estimators are not the same in general. Athey and Imbens (2016) and Wager and Athey (2018) gave early results on valid confidence intervals for non-parametric estimation of heterogeneous causal effects.

Best-in-class estimation. In this work, we are primarily interested in guarantees which hold also for misspecified models, functions that may not be able to exactly fit the outcome in terms of high-dimensional baseline covariates and treatment. *Agnostic machine learning* focuses on finding best-in-class models and bounding the generalization error of any model, whether well-specified or not (Vapnik, 2013; Cortes et al., 2010). However, in the causal inference setting, under model misspecification, regression methods may suffer from additional bias when generalizing across populations subject to different treatments. Our work addresses this issue by extending specification-agnostic generalization bounds to the CATE estimation problem. These bounds motivate our algorithms in the same way that standard supervised learning generalization bounds motivate structural risk minimization (Vapnik, 1998). As we will discuss later, the setting closest to ours is unsupervised domain adaptation (UDA) (Blitzer et al., 2008), in which models must generalize from a source domain to a target domain, having observed labeled samples only from the former. Agnostic machine learning bounds for UDA have a rich literature, spanning results for regression (Cortes and Mohri, 2011) and classification (Blitzer et al., 2008), using the PAC (Cortes et al., 2010;

Ben-David and Urner, 2012; Mansour et al., 2009) and PAC-Bayes (Germain et al., 2020) frameworks, and in combination with representation learning (Ben-David et al., 2007; Long et al., 2015). We cannot provide a complete survey here but return briefly to discuss the relation between our work and domain adaptation in Section 5.1.

4. Generalization Bounds for Estimating Potential Outcomes and CATE

Our goal is to accurately estimate the conditional means of potential outcomes $Y(t)$, and the CATE τ without making parametric assumptions on the functional form of the true potential outcomes. For this reason, we adopt the risk minimization approach to learning and search for best-in-class hypotheses. This requires generalization from the treated and control groups to the general population. Next, we derive bounds on the risk of such hypotheses in the following steps:

1. We define prediction *risk* with respect to potential outcomes μ_0, μ_1 and relate it to the expected error in estimates of the CATE, τ

2. We show how the risk on the observed distribution is a *biased* estimate of the desired population risk and give sample re-weighting schemes that removes this bias.

3. We give bounds on the expected risk under imperfect re-weighting schemes by placing assumptions on the loss with respect to the true outcome.

4. We derive finite-sample versions of these bounds and combine them to form a single fully observable bound on the risk in estimates of potential outcomes and CATE.

Our main generalization bound does not depend on treatment group overlap (Assumption 2). This diverges from most theoretical results for treatment effect estimation and provides intuition for when we can expect extrapolation to succeed approximately. Consistent non-parametric estimation, however, still requires overlap. In Section 5, we extend these results to estimators which make use of representation learning.

4.1 Risk for Hypotheses of Potential Outcomes and CATE

We study prediction of potential outcomes $Y(t) \in \mathcal{Y}$, for binary treatments $t \in \{0, 1\}$. We assume that outcomes are bounded, $\mathcal{Y} \subset [-\omega, \omega]$ for finite $\omega > 0$. Predictions are made using hypotheses $f_t \in \mathcal{H}$, from some hypothesis class $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$. These can then be combined to form estimates of CATE,

$$\hat{\tau}(x) \coloneqq f_1(x) - f_0(x) .$$

We note that while this is not the only way to estimate τ (see e.g., (Robins et al., 2000; Künzel et al., 2019; Nie and Wager, 2021) for alternatives), it does allow us to leverage separate bounds on the risk of hypotheses f_0, f_1 with respect to the potential outcomes, to then bound the risk of $\hat{\tau}$. We define the risk of hypotheses f_0, f_1 below.

Definition 2. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function, such as the squared loss $L(y, y') = (y - y')^2$. The expected *pointwise loss* of a hypothesis f_t at a point x is:

$$\ell_{f_t}(x) := \mathbb{E}_{Y(t)|X}[L(Y(t), f_t(x)) \mid X = x] . \tag{3}$$

The *marginal risk* of a hypothesis f_t w.r.t. a population $p(X)$ is

$$R(f_t) := \mathbb{E}_X [\ell_{f_t}(X)] , \tag{4}$$

and the *factual risk* of f_t w.r.t. treatment group $p(X \mid T = t)$ is

$$R_t(f_t) := \mathbb{E}_{X|T} [\ell_{f_t}(X) \mid T = t] . \tag{5}$$

The *counterfactual risk* is $R_{1-t}(f_t) := \mathbb{E}_{X|T} [\ell_{f_t}(X) \mid T = 1 - t]$. The subscript on the risk R_t indicates the treatment group over which it is evaluated. Note that the potential outcome against which the risk is evaluated is implicit in this notation—we only consider evaluating f_t against $Y(t)$ or μ_t, $\hat{\tau}$ against τ, et cetera.

In most of this work, we restrict our attention to the squared error loss but note that our analysis generalizes to other convex loss functions, such as the mean absolute deviation.

Similar to potential outcomes, we assess the quality of an estimate $\hat{\tau}$ of τ based on the expectation of the loss function L over the marginal density of covariates, $p(X)$,

$$R(\hat{\tau}) := \mathbb{E}_X [L(\tau(X), \hat{\tau}(X))] . \tag{6}$$

The marginal risk $R(\hat{\tau})$ is the overall expected error in estimating CATE, taken over the entire population. However, $R(\hat{\tau})$ is not readily computable from data because neither $\tau(X)$ nor $p(X)$ are known. Moreover, we cannot make an empirical average estimate of it because, again, neither $\tau(X_i)$ nor $Y_i(1) - Y_i(0)$ are known. Instead, we will bound $R(\hat{\tau})$ from above. The main challenge of computing the marginal risk for hypotheses of potential outcomes is to quantify the counterfactual risk, and this is the primary concern of this work.

Unlike $R(f_t)$, $R(\hat{\tau})$ does not depend on the noise (conditional variance) in $Y(t)$.[3] We adopt this convention for $R(\hat{\tau})$ as it coincides with the Precision in Estimation of Heterogeneous Effects (PEHE) (Hill, 2011). However, similarly to $L(Y(t), f_t)$, $L(\tau, \hat{\tau})$, is not observed over p, as $Y(1)$ is only observed for the treated group p_1, and $Y(0)$ only for the control group p_0. We return to this issue later, and begin instead by stating the following result relating the risk of $\hat{\tau}$ to those of f_0 and f_1, in the case of L the squared loss.

Lemma 1. *Let $L(y, y') = (y - y')^2$ be the squared loss function. For hypotheses f_0, f_1 of expected potential outcomes μ_0, μ_1, with marginal risks $R(f_0), R(f_1)$, and $\hat{\tau} = f_1 - f_0$, there is a constant σ_Y (defined in the proof below), such that*

$$R(\hat{\tau}) \leq 2 \left(R(f_0) + R(f_1) \right) - 4\sigma_Y^2 . \tag{7}$$

---

3. This is because that τ is defined in terms of expectations over the potential outcomes.

> *Similar results hold also for other losses that satisfy the (relaxed) triangle inequality,*
> *e.g., the absolute loss, $L(y, y') = |y - y'|$, with corresponding noise decompositions.*

Proof. Due to the relaxed triangle inequality for squared differences,

$$\mathbb{E}_X[(\tau(X) - \hat{\tau}(X))^2] = \mathbb{E}_X[(\mu_1(X) - \mu_0(X) - f_1(X) + f_0(X))^2]$$
$$\leq 2\left(\mathbb{E}_X[(\mu_1(X) - f_1(X))^2] + \mathbb{E}_X[(\mu_0(X) - f_0(X))^2]\right)$$

Now, by the standard bias-noise decomposition,

$$R(f_t) = \mathbb{E}_{X,Y(t)}\left[((Y(t) - \mu_t(X))^2] + \mathbb{E}_X\left[(\mu_t(X) - f_t(X)))^2\right]\right.$$

Hence, for $t \in \{0, 1\}$ with $\sigma^2_{Y(t)} := \mathbb{E}_{X,Y(t)}\left[((Y(t) - \mu_t(X))^2]\right.$,

$$\mathbb{E}_X[(\tau(X) - \hat{\tau}(X))^2] \leq 2\left(R(f_0) - \sigma^2_{Y(0)} + R(f_1) - \sigma^2_{Y(1)}\right)$$

and with $\sigma^2_Y := \max(\sigma^2_{Y(0)}, \sigma^2_{Y(1)})$, we have our result. $\qquad\square$

Lemma 1 implies that small errors in hypotheses of potential outcomes guarantee small errors in CATE. However, it is worth noting that this decomposition need not lead to the best achievable bound in all cases. Even when $Y(0)$ and $Y(1)$ are complex functions, $\tau(x)$ may be a simple function. In this work, we do not address this in our theoretical treatment but find that sharing parameters in estimation of $Y(0)$ and $Y(1)$ lead to better results empirically. Next, we study $R(f_0)$ and $R(f_1)$ separately, in terms of observable quantities, to later give a self-contained result for $R(\hat{\tau})$.

4.2 Importance-weighting Hypotheses & Propensity Scores

We proceed to show how the marginal risk R in potential outcomes and CATE may be computed by re-weighting the factual risk R_t. This approach is widely used within machine learning (Shimodaira, 2000; Cortes et al., 2010; Rosenbaum and Rubin, 1983). We note in passing that R_t is not observed directly but can be readily estimated from an empirical sample. We return to this issue in later sections.

Due to the fundamental impossibility of observing counterfactual outcomes, each potential outcome $Y(t)$ is only observed for subjects who were given treatment $T = t$, distributed according to $p(X \mid T = t)$. As a result, unless treatment is assigned randomly (independently of X), $R_t(f_t)$ is different from $R(f_t)$ in general. In particular, a minimizer f_t^* of $R_t(f_t)$ can be arbitrarily different from a minimizer of $R(f_t)$, depending on the difference between p and p_t. This bias can have large impact on treatment policies derived from f_0, f_1 and τ. We illustrate this problem in Figure 3.

To reduce the bias described above, a classical solution is to re-weight the observable risk (Shimodaira, 2000) using a function $w : \mathcal{X} \to \mathbb{R}_+$, such that $\mathbb{E}_{X|T}[w(X) \mid T = t] = 1$. We define the re-weighted population and treatment group risks as follows,

$$R^w(f_t) := \mathbb{E}_X\left[w(X)\ell_{f_t}(X)\right] \tag{8}$$
$$R_t^w(f_t) := \mathbb{E}_{X|T}\left[w(X)\ell_{f_t}(X) \mid T = t\right], \tag{9}$$

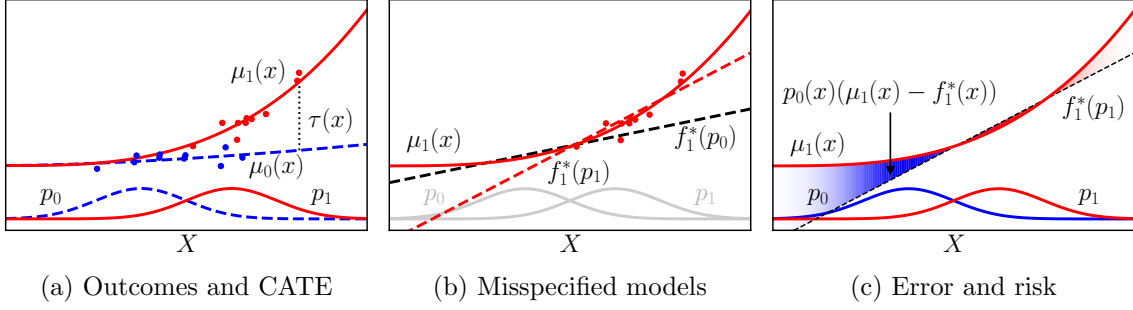(a) Outcomes and CATE   (b) Misspecified models   (c) Error and risk

Figure 3: Illustrative example of bias in regression adjustment of expected potential outcomes $\mu_0(x), \mu_1(x)$ and CATE $\tau(x)$. In (a), we show the two potential outcomes and the two treatment groups p_0, p_1 in dashed blue and solid red lines, respectively, as well as samples of each group. In (b), we show the best *linear models* $f_1^*(p_0), f_1^*(p_1)$ of the potential outcome under treatment $\mu_1(x)$ fit to the potential outcome of the control group and treated group respectively. In (c), we illustrate the difference in weighted error (bias) for the model fit to the treated group $f_1^*(p_1)$ evaluated in the control group and treated group.

where w is chosen to skew the sample to mimic the distribution of the full population p. Many common choices of weights are based on the family of *balancing scores* (Rosenbaum and Rubin, 1983), of which the best known is the *propensity score* $\eta(X)$ with respect to X,

$$\eta(x) := p(T = 1 \mid X = x) \ . \tag{10}$$

We can now state the following result.

Lemma 2. *For fixed $t \in \{0, 1\}$, under Assumption 2 (overlap), there exists a weighting function $w : \mathcal{X} \to \mathbb{R}_+$, such that $R_t^w(f_t) = R(f_t)$. In particular, this holds for*

$$w(x) := \frac{p(T = t)}{(2t - 1)\eta(x)} + 1 - t \ . \tag{11}$$

More generally, it holds for (11) with $\eta(\phi(x))$ for any ϕ such that $\ell_f \perp\!\!\!\perp X \mid \phi(X)$. We refer to weights that satisfy (11) as balancing weights.

Proof. For any weighting function w,

$$R_t^w(f_t) = \int_{x \in \mathcal{X}} w(x) \ell_{f_t}(x) p_t(x) dx = \int_{x \in \mathcal{X}} \frac{p_t(x)}{p(x)} w(x) \ell_{f_t}(x) p(x) dx = R^{\frac{p_t}{p}w}(f_t) \ .$$

With $w(x) = p(x)/p_t(x)$, the special case in (11) follows from Bayes theorem and the definition of $\eta(x)$. The second step uses Assumption 2 to ensure that $p(x)/p_t x(x)$ is defined. The more general statement follows from integration over ϕ and a change of variables. $\square$

Remark 1 (Violation of Assumptions 1 & 2)**.** If overlap is only partially satisfied, Lemma 2 may still be applied to the expected risk over the subset of $\mathcal{X}$ for which overlap holds. More generally, weights may be chosen to emphasize regions where treatment groups are more

similar (Li et al., 2018). If ignorability is violated, such as when unobserved confounders exist, a consistent estimator could in theory be obtained by letting the weights w depend also on Y. However, such weights are not identified from observed data. Instead, a worst-case bound may be obtained by searching over a family of weighting functions in which these optimal weights are members. This is the topic of sensitivity analysis (see e.g., Rosenbaum (2002) for a comprehensive overview).

In practice, $\eta(X)$ and balancing weights w are typically unknown and have to be estimated from data. Moreover, even though weights based on η are optimal in expectation, they can lead to poor finite-sample behavior (Swaminathan and Joachims, 2015). For these reasons, *even if we had knowledge of $\eta(X)$*, we are often interested in weighting functions that do not satisfy Lemma 2. Next, we give bounds on the difference between the re-weighted (factual) empirical risk, under *arbitrary* weightings, and the marginal risk.

4.3 Bounds on the Risk of a Re-weighted Estimator

When overlap is not satisfied everywhere or the chosen weighting function w is not perfectly balancing, the difference between the weighted factual risk $R_t^w(f_t)$ and the marginal risk $R(f_t)$ may be arbitrarily large, without further assumptions on the potential outcomes or the hypothesis class $\mathcal{H}$. However, in many cases we have reason to make assumptions about the worst-case loss in generalization, as is typical in statistical learning theory. In this section, we give bounds on $R(f_t)$ under such assumptions.

Let $\mathcal{L} \subset \{\mathcal{X} \to \mathbb{R}_+\}$ be a space of pointwise loss functions with respect to the covariates X endowed with a norm $\|\cdot\|_{\mathcal{L}}$. In this work, we assume that the expected conditional loss ℓ_{f_t} for each potential outcome belongs to such a family, i.e., that $\ell_{f_t} \in \mathcal{L}$. A simple example of such a family is the set of loss functions with bounded maximum value, $\mathcal{L}_M = \{\ell :\to \mathbb{R}_+ \; ; \; \sup_{x \in \mathcal{X}} \ell(x) \leq M\}$. This assumption is satisfied without loss of generality as long as the outcome Y is bounded. However, it is not very informative and will lead to loose bounds in general. Instead, we may make assumptions about the functional properties of ℓ_{f_t}. Such assumptions include that ℓ_{f_t} is C-Lipschitz or belongs to a reproducing-kernel Hilbert space (RKHS). We illustrate the former with an example in Figures 4a–4b. Note that these assumptions are not testable without further assumptions such as smoothness.

Now, consider the marginal distribution p and a re-weighted treatment group p_t^w on $\mathcal{X}$, with $p_t^w(x) = p_t(x)w(x)$ for a weighting function w with mean 1 over p_t as in (8). Let $\ell \in \mathcal{L}$ be a pointwise loss on $\mathcal{X}$. Recall that $R(f_t)$ and $R_t^w(f_t)$ denote the marginal and re-weighted factual risks respectively. By definition,

$$
\begin{aligned}
R(\ell) := \mathbb{E}_{X \sim p}[\ell] &= \mathbb{E}_{X \sim p_t^w}[\ell] + \int_{x \in \mathcal{X}} \ell(x)(p(x) - p_t^w(x))dx \\
&\leq \mathbb{E}_{X \sim p_t^w}[\ell] + \sup_{\ell' \in \mathcal{L}} \left| \int_{x \in \mathcal{X}} \ell'(x)(p(x) - p_t^w(x))dx \right| .
\end{aligned} \tag{12}
$$

The second term on the right-hand side in (12) is known as the *integral probability metric distance* (IPM) between p and p_t^w w.r.t. $\mathcal{L}$, defined as follows (Müller, 1997) :

$$
\mathrm{IPM}_{\mathcal{L}}(p, q) := \sup_{\ell \in \mathcal{L}} |\mathbb{E}_p[\ell(x)] - \mathbb{E}_q[\ell(x)]| . \tag{13}
$$

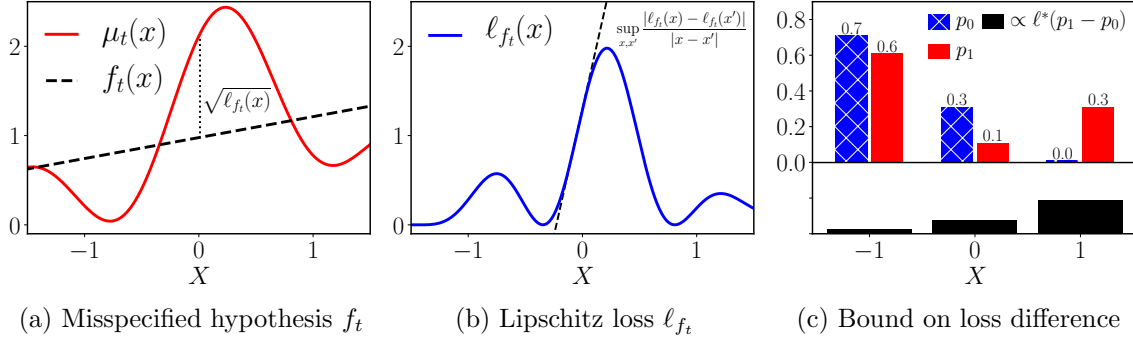(a) Misspecified hypothesis f_t (b) Lipschitz loss ℓ_{f_t} (c) Bound on loss difference

Figure 4: Example illustrating assumptions on the pointwise loss ℓ_{f_t} in a noiseless setting. In (a) we see the true potential outcome μ_t and a hypothesis f_t. The pointwise loss between them is plotted in (b). In (c), we illustrate the difference between two densities p_0 and p_1 on $\{-1, 0, 1\}$. The bottom panel shows the worst-case contribution of any loss function in an RBF-kernel RKHS $\mathcal{L}$ to the *difference* in risk $R_0(f_t) - R_1(f_t)$. The more similar p_0, p_1, or the smoother the functions in $\mathcal{L}$, the smaller the overall contribution.

Particular choices of $\mathcal{L}$ make the IPM equivalent to different well-known distances on distributions: With $\mathcal{L}$ the family of functions in the norm-1 ball in a reproducing kernel Hilbert space (RKHS), IPM$_{\mathcal{L}}$ is the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012); When $\mathcal{L}$ is the family of functions with Lipschitz constant at most 1, we obtain the Wasserstein distance (Villani, 2008). As pointed out by e.g., Long et al. (2015), the IPM may be viewed as the loss of a treatment group classifier, and adversarial losses may be considered in its place (Ganin et al., 2016). In Figure 4c, we illustrate the maximizer ℓ^* of the supremum, in terms of its contribution to the expected difference in risk in the MMD case.

Before stating the final form of our bounds on $R(f_t) - \hat{R}_t^w(f_t)$, we note that for $t \in \{0, 1\}$, with $\pi_t = p(T = t)$, we may decompose the population risk R as follows.

$$R(f_t) \;=\; \pi_t \underbrace{R_t(f_t)}_{\text{Observable}} + (1 - \pi_t) \underbrace{R_{1-t}(f_t)}_{\text{Unobserved}} . \tag{14}$$

The factual risk $R_t(f_t)$ is identifiable under ignorability, as

$$\ell_{f_t}(X) = \mathbb{E}_{Y(t)|X}[L(f_t(X), Y(t)) \mid X] = \mathbb{E}_{Y|X,T}[L(f_t(X), Y) \mid X, T = t]$$

For this reason, to bound the risk of f_t on the whole population p it is sufficient for us to bound the *counterfactual risk* $R_{1-t}(f_t)$, and estimate $R_t(f_t)$ empirically.

Lemma 3. *For a hypothesis f with expected point-wise loss $\ell_f(x)$ such that $\ell_f / \|\ell_f\|_{\mathcal{L}} \in \mathcal{L}$ and any re-weighting $w \in \{w' : w'(x) \geq 0, \mathbb{E}_X[w'(X) \mid T = t] = 1\}$, with $t \in \{0, 1\}$,*

$$R_{1-t}(f) - R_t^w(f) \;\leq\; \|\ell_f\|_{\mathcal{L}} \mathrm{IPM}_{\mathcal{L}}(p_{1-t}, p_t^w) ,$$

where $p_t(x) = p(X = x \mid T = t)$. Further, there exist re-weightings w^* such that,

$$R_{1-t}(f) - R_t^{w^*}(f) \;\leq\; \|\ell_f\|_{\mathcal{L}} \mathrm{IPM}_{\mathcal{L}}(p_{1-t}, p_t^{w^*}) \;\leq\; \|\ell_f\|_{\mathcal{L}} \mathrm{IPM}_{\mathcal{L}}(p_{1-t}, p_t) . \tag{15}$$

The first inequality is tight under Assumption 2 for weights $w(x) = p_1(x)/p_0(x)$. The second is not tight for general f unless $p_0 = p_1$.

Proof. The result follows from the definition of IPMs. We prove it for $t = 0$.

$$
\begin{aligned}
R_1(f) - R_0^w(f) &= \mathbb{E}_{X|T}[\ell_f(X) \mid T = 1] - \mathbb{E}_{X|T}[w(X)\ell_f(X) \mid T = 0] \\
&\leq \left| \mathbb{E}_{X|T}[\ell_f(X) \mid T = 1] - \mathbb{E}_{X|T}[w(X)\ell_f(X) \mid T = 0] \right| \\
&\leq \|\ell_f\|_{\mathcal{L}} \sup_{h \in \mathcal{L}} \left| \mathbb{E}_{X|T}[h(X) \mid T = 1] - \mathbb{E}_{X|T}[w(X)h(X) \mid T = 0] \right| \qquad (16) \\
&= \|\ell_f\|_{\mathcal{L}} \mathrm{IPM}_{\mathcal{L}}(p_1, p_0^w) \ .
\end{aligned}
$$

Step (16) relies on that $\ell_f/\|\ell_f\|_{\mathcal{L}} \in \mathcal{L}$. Further, for importance weights $w^*(x) = p_1(x)/p_0(x)$, for any $\ell \in \mathcal{L}$, under Assumption 2 (overlap),

$$
\begin{aligned}
&\mathbb{E}_{X|T}[\ell(X) \mid T = 1] - \mathbb{E}_{X|T}[w^*(x)\ell(x) \mid T = 0] \\
&= \mathbb{E}_{X|T}[\ell(X) \mid T = 1] - \mathbb{E}_{X|T}\left[\frac{p_1(x)}{p_0(x)} \ell(x) \mid T = 0 \right] = 0
\end{aligned}
$$

and the first inequality in (15) is tight, as $\mathrm{IPM}_{\mathcal{L}}(p_1, p_0^{w^*}) = 0$. Given that $\mathrm{IPM} \geq 0$ in general, the second inequality holds in this case as well. If overlap does not hold, the ratio $\frac{p_1(x)}{p_0(x)}$ is not defined on the support of $p_1(x)$. The result for $t = 1$ follows analogously. $\qquad \square$

Corollary 1. *Under the conditions of Lemma 3, with $\tilde{w}(x) := \pi_t + (1 - \pi_t)w(x)$,*

$$
R(f_t) \leq R_t^{\tilde{w}}(f_t) + (1 - \pi_t)\|\ell_{f_t}\|_{\mathcal{L}} \mathrm{IPM}_{\mathcal{L}}(p_{1-t}, p_t^w) \qquad (17)
$$

Proof. The result follows immediately from Lemma 3. $\qquad \square$

Remark 2 (Necessity of assumptions)**.** Lemma 3 and Corollary 1 do not strictly speaking depend on Assumption 1 (ignorability) due to the definitions of $R(f_t)$ and ℓ_{f_t} being made w.r.t. the potential outcomes $Y(t)$. However, to estimate the right-hand side of (17) from observational data, ignorability is required. Moreover, neither result depend on Assumption 2 (overlap) as long as $w(x)$ is defined everywhere on $p_t(x)$. For particular losses, we can avoid making assumptions about ℓ_{f_t}, by making assumptions on f_t and the hypothesis class $\mathcal{H}$ instead. This approach was taken by Ben-David et al. (2007), who used the so-called "triangle inequality for loss functions" to give bounds on the risk in unsupervised domain adaptation under assumptions on $\mathcal{H}$. However, this leads to the rather unattractive property that the resulting bounds are not tight even in the special case that $p_0 = p_1$.

4.4 Bounds Based on Finite Samples

Adopting results from statistical learning theory, we bound the difference between empirical estimates of $R_t^w(f_t)$ and $\mathrm{IPM}_{\mathcal{L}}(p_{1-t}, p_t^w)$ and their expected counterparts. These results are then combined to form a bound on $R(f_t)$.

The re-weighted risk R_t^w may be estimated, for a fixed weighting function w by the standard Monte-Carlo method. Consider a sample $\mathcal{D} = \{(x_1, t_1, y_1), ..., (x_n, t_n, y_n)\} \sim$

$p^n(X, T, Y)$, with $n_t = \sum_{i=1}^{n} \mathbb{1}[t_i = t]$ and define the empirical weighted factual risk,

$$\hat{R}_t^w(f_t) := \frac{1}{n_t} \sum_{i:t_i=t} w(x_i) L(f_t(x_i), y_i) .$$

We aim to bound the difference

$$\Delta(f_t) := R_t^w(f_t) - \hat{R}_t^w(f_t) .$$

To achieve this, we use a result from the literature which builds on the concept of pseudo-dimension $\mathrm{Pdim}(\mathcal{H})$ of a function class $\mathcal{H}$, originally due to Pollard (1984).

Definition 3 (Anthony and Bartlett (2009), Def. 11.1–11.2)**.** Let $\mathcal{H}$ be a set of functions mapping from a domain $\mathcal{X}$ to $\mathbb{R}$ and suppose that $\mathcal{S} = \{x_1, x_2, \ldots, x_m\} \subseteq \mathcal{X}$. Then, $\mathcal{S}$ is pseudo-shattered by $\mathcal{H}$ if there are real numbers $y_1, y_2, \ldots, y_m$ such that for each $b \in \{0, 1\}^m$, there is a function $h_b \in \mathcal{H}$ with $\mathrm{sgn}(h_b(x_i) - y_i) = b_i$ for $1 \le i \le m$. $\mathcal{H}$ has pseudo-dimension d if d is the maximum cardinality of a subset $\mathcal{S}$ of $\mathcal{X}$ that is pseudo-shattered by $\mathcal{H}$. If no such maximum exists, we say that $\mathcal{H}$ has infinite pseudo-dimension. The pseudo-dimension of $\mathcal{H}$ is denoted $\mathrm{Pdim}(\mathcal{H})$.

We have modified the result below to account for the variance in the outcome Y.

Lemma 4 (Cortes et al. (2010), Theorem 4)**.** *Let $\ell_h = \mathbb{E}_{Y|X}[L(h(x), Y) \mid X = x]$ be the expectation of the squared loss $L(y, y') = (y - y)^2$ of a hypothesis $h \in \mathcal{H} \subseteq \{h' : \mathcal{X} \to \mathbb{R}\}$, let $d = \mathrm{Pdim}(\{\ell_h : h \in \mathcal{H}\})$ and let $\sigma_Y^2 = \mathbb{E}_{X,Y}[L(Y, \mathbb{E}_{Y|X}[Y \mid X])]$. Then, for a weighting function $w(x)$ such that $\mathbb{E}_X[w(X)] = 1$, with probability at least $1 - \delta$ over a sample $((x_1, y_1), ..., (x_n, y_n))$, with empirical distribution $\hat{p}$,*

$$R^w(h) \le \hat{R}^w(h) + V_{p,\hat{p}}[w(x)l_h(x)] \frac{\mathcal{C}_n^{\mathcal{H}}}{n^{3/8}} + \sigma_Y^2 \quad \text{where} \quad \mathcal{C}_n^{\mathcal{H}} = 2^{5/4} \left(d \log \frac{2ne}{d} + \log \frac{4}{\delta} \right)^{3/8}$$

and $V_{p,\hat{p}}[w(x)l_h(x)] = \max \left(\sqrt{\mathbb{E}_X[w^2(X)\ell_h^2(X)]}, \sqrt{\mathbb{E}_{X \sim \hat{p}}[w^2(X)\ell_h^2(X)]} \right)$.

Lemma 4 applies to any valid weighting function w, not only importance weights or weights based on balancing scores. Used in conjunction with Corollary 1, Lemma 4 allows us to separate the bias (the IPM-term) and variance (see above) introduced by w.

The efficiency with which a sample may be used to estimate $\mathrm{IPM}_{\mathcal{L}}$ depends on the chosen function family $\mathcal{L}$. In particular, the sample complexity of learning the Wasserstein distance between two densities on $\mathcal{X}$ scales as $\mathcal{O}(d)$ with the dimension d of $\mathcal{X}$, whereas the kernel-based MMD has $\mathcal{O}(1)$ dependence. Below, we state a result bounding the sample complexity for the MMD with universal kernels.

Lemma 5 (Sriperumbudur et al. (2009))**.** *Let $\mathcal{X}$ be a measurable space. Suppose k is a universal, measurable kernel such that $\sup_{x \in \mathcal{X}} k(x, x) \le C \le \infty$ and $\mathcal{L}$ the reproducing kernel Hilbert space induced by k, with $\nu := \sup_{x \in \mathcal{X}, f \in \mathcal{L}} f(x) < \infty$. Then, with $\hat{p}, \hat{q}$ the*

---

empirical distributions of p, q from m and n samples, and with probability at least $1 - \delta$,

$$|\text{IPM}_{\mathcal{L}}(p, q) - \text{IPM}_{\mathcal{L}}(\hat{p}, \hat{q})| \leq \sqrt{18\nu^2 \log \frac{4}{\delta} C} \left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) \ .$$

The Gaussian RBF kernel $k(x, x') = e^{-\|x - x'\|_2^2 / (2\sigma^2)}$, with bandwidth $\sigma > 0$, is an important class of universal kernels to which Lemma 5 applies.

With Lemmas 3–5 in place, we can now state our main result.

Theorem 1. *Assume that ignorability (Assumption 1) holds w.r.t. X. Given is a sample $(x_1, t_1, y_1), ..., (x_n, t_n, y_n) \overset{i.i.d.}{\sim} p(X, T, Y)$ with empirical measure $\hat{p}^n$ and $n_t := \sum_{i=1}^n \mathbb{1}[t_i = t]$ for $t \in \{0, 1\}$. Let $f_t(x) \in \mathcal{H}$ be a hypothesis of $\mathbb{E}_{Y(t)|X}[Y(t) \mid X = x]$ and $\ell_{f_t}(x) := \mathbb{E}_{Y|X}[L(f_t(x), Y(t)) \mid X = x]$ where $L(y, y') = (y - y')^2$. Assume that there exists a constant $B > 0$ such that, $\ell_{f_t}(x)/B \in \mathcal{L}$, where $\mathcal{L}$ is a reproducing kernel Hilbert space of a kernel, k such that $k(x, x) < \infty$. Finally, let $w_t : \mathcal{X} \to \mathbb{R}_+$ be a valid re-weighting of p_t, $\mathbb{E}[w(X) \mid T = t] = 1$, and let $\tilde{w}(x) = \pi_t + (1 - \pi_t)w(x)$, where $\pi_t = p(T = t)$. With probability at least $1 - 2\delta$,*

$$
\begin{aligned}
R(f_t) \quad \leq \quad & \hat{R}_t^{\tilde{w}}(f_t) + B(1 - \pi_t)\text{IPM}_{\mathcal{L}}(\hat{p}_t^{w_t}, \hat{p}_{1-t}) \\
+ \quad & V_{p_t}(\tilde{w}, \ell_{f_t}) \frac{\mathcal{C}_{n_t, \delta}^{\mathcal{H}}}{n_t^{3/8}} + \mathcal{D}_{n_0, n_1, \delta}^{\mathcal{L}} \left(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}} \right) + \sigma_{Y(t)}^2
\end{aligned}
$$

where $\mathcal{C}_{n_t, \delta}^{\mathcal{H}}$ is a function of the pseudo-dimension of $\mathcal{H}$, $\mathcal{D}_{n_0, n_1, \delta}^{\mathcal{L}}$ is a function of the kernel norm of $\mathcal{L}$ (see Lemma 5), both only with logarithmic dependence on n, $\sigma_{Y(t)}^2$ is the expected variance in $Y(t)$, and $V_{p_t}(w_t, \ell_{f_t}) = \max \left(\sqrt{\mathbb{E}_{p_t}[\tilde{w}_t^2 \ell_{f_t}^2]}, \sqrt{\mathbb{E}_{\hat{p}_t}[\tilde{w}_t^2 \ell_{f_t}^2]} \right)$. A similar bound exists where $\mathcal{L}$ is the family of functions with Lipschitz constant at most 1 and $\text{IPM}_{\mathcal{L}}$ the Wasserstein distance, but with worse sample complexity.

Proof. The result follows from application Lemmas 4–5 to Lemma 3 and is given in larger generality in Theorem 2. The $n^{3/8}$ rate may be improved at the cost of a more complicated expression, see discussion following Theorem 3 in (Cortes et al., 2010). $\qquad \square$

Remark 3. The spaces of hypotheses $\mathcal{H}$ and the space of loss functions $\mathcal{L}$ appear in Theorem 1 through two different measures of complexity, the pseudo-dimension of $\mathcal{H}$ and the kernel norm of losses in $\mathcal{L}$. It is worth pointing out that $\mathcal{L}$, which covers the loss functions of the hypotheses in $\mathcal{H}$, may be larger than $\mathcal{H}$, while the bound remains informative. For example, in experiments, we let $\mathcal{H}$ be a family of fixed-size deep neural networks and $\mathcal{L}$ the functions of an RKHS given by an RBF kernel of a given bandwidth. The pseudo-dimension of such an $\mathcal{H}$, and therefore $\mathcal{C}^{\mathcal{H}}$, is bounded. The second complexity term, $\mathcal{D}^{\mathcal{L}}$, is bounded too—irrespective of whether the pseudo dimension of $\mathcal{L}$ is bounded.

We can also immediately state the following corollary.

Corollary 2. *Assume that the conditions of Theorem 1 hold. Let $f(x,t) := f_t(x)$ and let $\hat{R}_p^w(f) := \sum_{i=1}^n w(x_i, t_i) L(f(x_i, t_i), y_i)/n$ represent the weighted empirical factual risk. Then, with $\tilde{w}(x,t) := w_t(x)/\pi_t$, $n_{\min} = \min(n_0, n_1)$ and $\sigma^2 = \max(\sigma_{Y(0)}^2, \sigma_{Y(1)}^2)$ there is a term $K_{\mathcal{L}, \mathcal{H}, w, \delta, n_0, n_1}$ with at most logarithmic dependence on n_0, n_1, such that*

$$\frac{R(\hat{\tau})}{2} \leq \hat{R}_p^{\tilde{w}_t}(f) + B\left[\pi_0 \text{IPM}_{\mathcal{L}}(\hat{p}_0, \hat{p}_1^{w_1}) + \pi_1 \text{IPM}_{\mathcal{L}}(\hat{p}_1, \hat{p}_0^{w_0})\right] + \frac{K_{\mathcal{L}, \mathcal{H}, w, \delta, n_0, n_1}}{n_{\min}^{3/8}} + 2\sigma^2 \ .$$

A tighter result may be obtained by decomposing the constant K.

Theorem 1 and Corollary 2 hint at several interesting dependencies between generalization error, treatment group imbalance, sample re-weighting schemes and the choice of hypothesis class. We comment on these below.

Bounds with overlap and known propensity scores. If overlap is satisfied and propensity scores known, applying importance weights $w_t(x) = p(T = t \mid X = x)/p(T = 1 - t \mid X = x)$ in Theorem 1 leads to a tight bound in the limit of infinite samples (IPM and variance terms approach zero, re-weighted risk approaches desired population risk). A special case of this is the randomized controlled trial (RCT), in which T has no dependence on X. In this setting, the IPM-terms depend only on the finite-sample differences between treatment groups—which may still be useful to characterize. It has been shown that under overlap, in the asymptotic limit, the best achievable sample complexity is unrelated to the imbalance of p_0 and p_1 (Alaa and Schaar, 2018). However, this setting is not our main concern as we are specification agnostic and focus on the finite-sample case.

Bounds without overlap. Theorem 1 *does not* rely on treatment group overlap. Instead, it relies on an assumption that the true loss (w.r.t. features X and potential outcome $Y(t)$) is a function in the given family $\mathcal{L}$. The bound requires that the weights used for re-weighting, $w_t(x)$, are defined everywhere on $p_t(x)$ for $t \in 0, 1$. Importantly, if for some $x \in \mathcal{X}$, $p_{1-t}(x) = 0$ and $p_t(x) > 0$, the density ratio $p_t(x)/p_{1-t}(x)$ is not defined and is not admissible as weighting function. Hence, the re-weighted densities cannot be equal in such regions and the IPM will be non-zero in this case. As a result, the bound cannot be tight in the general case without overlap, but may be informative if the lack of overlap is small. We return to the question of overlap in the next section, following Theorem 2.

Bias and variance. The term $V_{p_t}(\tilde{w}, \ell_{f_t})$ in Theorem 1 shows that a less uniform re-weighting w leads to larger variance (dependence on n). However, if p_0 and p_1 are very different, a non-uniform (balancing) w is required to ensure unbiasedness, e.g., by making $p_t^{w_t} = p_{1-t}$. This indicates that w introduces a bias-variance trade-off on top of the one typical for supervised learning. In particular, even if the true treatment propensity η is known, a biased weighting scheme may lead to a smaller bound on the population risk when p_0 and p_1 are far apart.

Imbalance in non-confounders. The size of the bounds in Theorem 1 and Corollary 2 clearly depends on the quality of the hypothesis f and the choice of re-weighting w. In addition, the IPM terms depend heavily on the input space $\mathcal{X}$. In particular, if variables included in $\mathcal{X}$ are predictive of T but not predictive of Y, e.g., if they are instrumental variables (Ding et al., 2017), they will contribute to the IPM term but not to the expected
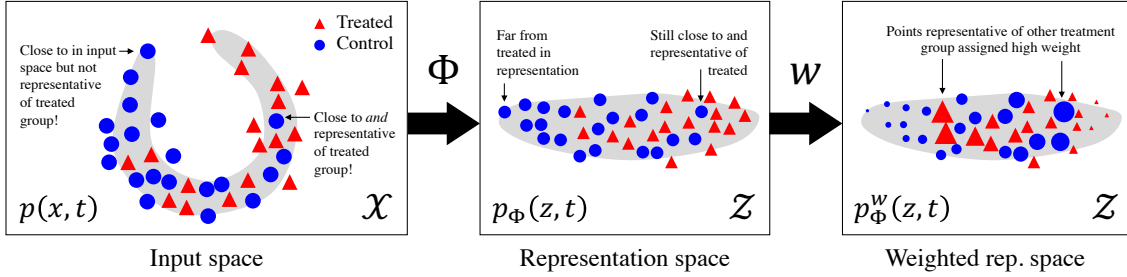
Figure 5: Treatment group imbalance in the observed context X (left). Treated and controls have different distributions. Mapping contexts $x \in \mathcal{X}$ to a representation space $\mathcal{Z}$ using an embedding function Φ (middle) can allow for more efficient comparison of treatment groups and re-weighting (right) of observations to achieve a more balanced sample. In Section 5, we give generalization bounds for learning such representations and in Section 7, we use neural networks to fit them.

risk, loosening the bound needlessly. If we can learn to ignore such information, we may obtain a tighter bound. To this end, in the next section, we derive bounds for representations $\Phi(X)$ of the original feature space.

Generalization under policy and domain shift. Predicting the conditional treatment effect for an individual may be viewed as predicting the effect of a change in treatment policy from one alternative to another. This notion may be generalized further by considering the estimation of treatment effects for change in policy on a population that differs from the one learned from. Specifically, this would involve a change not only in $p(T \mid X)$ but in $p(X)$ as well. We studied this extended problem in Johansson et al. (2018), and referred to changes in both policy $p(T \mid X)$ and domain $p(X)$ as a change in *design*. We do not cover this setting in detail here.

5. Generalization Bounds for Representation Learning

When the input space $\mathcal{X}$ increases in dimension, treatment groups $p_t(X)$ tend to grow increasingly different (D'Amour et al., 2021) and, in general, this will lead to a looser bound in Theorem 1. To some extent, this can be mitigated by appropriately chosen weights w, but the additional finite-sample variance introduced by highly non-uniform weights may prevent tightening of the bound. In this section, we introduce another tool for minimizing bounds on the marginal risk in hypotheses that act on learned (potentially lower-dimensional) representations of the covariates X, see Figure 5 for an illustration. This allows hypotheses to focus their modeling power on particular aspects of the covariate space, ignoring others.

In many applications, the input distribution $p(X)$ is believed to be a low-dimensional manifold embedded in a high-dimensional space $\mathcal{X}$, for example, the space of portraits embedded in the pixels of a photograph. In such settings, the best hypotheses are often simple functions of low-dimensional representations $\Phi(X)$ of the input (Bengio et al., 2013). The most famous examples of this are image and speech recognition for which representation learning using convolutional and recurrent neural networks advanced each field tremendously in only a few years (LeCun et al., 2015).

Let $\mathcal{E} \subset \{\mathcal{X} \to \mathcal{Z}\}$ denote a family of representation functions (embeddings) of the input space $\mathcal{X}$ into a space $\mathcal{Z} \subset \mathbb{R}^v$ and let $\Phi \in \mathcal{E}$ denote such an embedding function. For all theoretical results, we will assume that Φ is twice-differentiable and invertible. Further, let $\mathcal{G} \subseteq \{h : \mathcal{Z} \to \mathcal{Y}\}$ denote a set of hypotheses $h(\Phi)$ operating on the representation Φ and let $\mathcal{H}$ be the space of all such compositions, $\mathcal{H} = \{f = h \circ \Phi : h \in \mathcal{G}, \Phi \in \mathcal{E}\}$. Generalizing our discussion up to this point, we consider learning of $\Phi(X)$ from data with the goal of minimizing the marginal risk of hypotheses $h_t \circ \Phi$ for treatments $t \in \{0, 1\}$.

For CATE to be identifiable from observations of $p(\Phi(X), T, Y)$, we need precisely the same requirements on Φ as previously on X, *ignorability* and *overlap*,

$$\forall t \in \{0, 1\} : \underbrace{Y(t) \perp\!\!\!\perp T \mid \Phi(X)}_{\text{Ignorability}} \quad \text{and} \quad \forall z \in \mathcal{Z} : \underbrace{p(T = t \mid \Phi(X) = z) > 0}_{\text{Overlap}} \ . \tag{18}$$

Verifying the assumptions in (18) for a given Φ, based on observational data alone, is impossible, just as for X. To address this, we consider learning twice-differentiable, invertible representations $\Phi : \mathcal{X} \to \mathcal{Z}$ where $\Psi : \mathcal{Z} \to \mathcal{X}$ is the *inverse* representation, such that $\Psi(\Phi(x)) = x$ for all x. For treatment groups $t \in \{0, 1\}$, we let $p_{\Phi,t}(z)$ be the distribution induced by Φ over $\mathcal{Z}$, with $p_{\Phi,t}^w(z) := p_{\Phi,t}(z)w(\Psi(z))$ its re-weighted form and $\hat{p}_{\Phi,t}^w$ its re-weighted empirical form, following our previous notation. If Φ is invertible, ignorability and overlap in X is equivalent to ignorability and overlap in $\Phi(X)$, as $p(\Phi(X) = z) = p(X = \Psi(z))$.

Building on Section 4, we can now relate the expected marginal risk $R(h_t \circ \Phi)$ to the (expected) re-weighted factual risk $R^w(h_t \circ \Phi)$.

Lemma 6. *Suppose that Φ is a twice-differentiable, invertible representation, that $h_t(\Phi) \in \mathcal{G}$ is a hypothesis, and that $f_t(x) = h_t \in \mathcal{H}$ for $t \in \{0, 1\}$. Let $\ell_{\Phi,h_t}(z) := \mathbb{E}_Y[L(h_t(z), Y(t)) \mid X = \Psi(z)]$ be the expected pointwise loss given a representation z, where $L(y, y') = (y - y')^2$. Let C_Φ be a constant such that $\ell_{\Phi,h_t} \cdot |J_\Psi|/C_\Phi \in \mathcal{L}$ with $\mathcal{L} \subseteq \{\ell : \mathcal{Z} \to \mathbb{R}_+\}$ where $J_\Psi(z)$ is the Jacobian determinant of the representation inverse $\Psi = \Phi^{-1}$. Then, with $\pi_t = p(T = t)$ and w a valid re-weighting of $p_{\Phi,t}$, with $\tilde{w}$ defined as in Theorem 1, $\tilde{w}(x) = \pi_t + (1 - \pi_t)w(x)$, s*

$$R(f_t) \leq \pi_t R_t^{\tilde{w}}(f_t) + (1 - \pi_t) \, C_\Phi \cdot \text{IPM}_{\mathcal{L}}(p_{\Phi,1-t}, p_{\Phi,t}^w) \ . \tag{19}$$

Proof. By (14), with $\ell_{f_t}(x) := \mathbb{E}_Y[L(f_t(x), Y(t)) \mid X = x]$, we have that

$$R(f_t) \;=\; \pi_t R_t^{\tilde{w}}(f_t) + (1 - \pi_t) \int_{x \in \mathcal{X}} \ell_{f_t}(x) \, (p_{1-t}(x) - p_t^w(x)) \, dx \ .$$

Then, by the standard change of variables, assuming that Φ is invertible, we have

$$\begin{aligned}
\int_{x \in \mathcal{X}} \ell_{f_t}(x) \, (p_{1-t}(x) - p_t^w(x)) \, dx \;&=\; \int_{z \in \mathcal{Z}} \big(p_{\Phi,1-t}(z) - p_{\Phi,t}^w(z)\big) \, \ell_{f_t}(\Psi(z)) |J_\Psi(z)| dz \\
&=\; \int_{z \in \mathcal{Z}} \big(p_{\Phi,1-t}(z) - p_{\Phi,t}^w(z)\big) \, \ell_{\Phi,h_t}(z) |J_\Psi(z)| dz \\
&\leq\; C_\Phi \cdot \sup_{\ell \in \mathcal{L}} \int_{z \in \mathcal{Z}} \ell(z) \, \big(p_{\Phi,1-t}(z) - p_{\Phi,t}^w(z)\big) \, dz \ . \\
&=\; C_\Phi \cdot \text{IPM}_{\mathcal{L}}(p_{\Phi,1-t}, p_{\Phi,t}^w) \ .
\end{aligned}$$

22

Here, we have used the fact that, for invertible Φ, $p_\Phi(Z = \Phi(x)) = p(X = x)$. $\qquad\square$

The scale of Φ and the factor C_Φ. Comparing Lemma 6 (bound in representation) to Corollary 1 (original space), we notice two immediate differences: the additional factor C_Φ and the change from measuring distributional distance in X to doing so in Z, via Φ. The most illustrative example for why C_Φ is necessary is when Φ simply reduces the scale of X, i.e., when $\Phi(x) = x/a$ for $a > 1$. IPMs often vary with the scale of the space in which they are applied and we could reduce the right-hand side of the bound simply by scaling down X were it not for C_Φ, which counteracts this reduction.

The influence of Φ on the IPM Measuring distributional distance in Φ with a fixed IPM family $\mathcal{L}$ means that we may emphasize or de-emphasize part of the covariate space, even when Φ is invertible. For example, if Φ is a linear function that scales down a component X_d of X substantially, and $\mathcal{L}$ is a family of linear functions with bounded norm, the influence of distributional differences in $X(d)$ on the IPM is reduced. With Lemma 6 in place, we can

now state the result for the finite-sample case by following the same steps as in Section 4.

Theorem 2. *Given is a sample $(x_1, t_1, y_1), ..., (x_n, t_n, y_n) \overset{i.i.d.}{\sim} p(X, T, Y)$ with empirical measure $\hat{p}$. Assume that ignorability (Assumption 1) holds w.r.t. X. Suppose that Φ is a twice-differentiable, invertible representation, that $h_t(\Phi)$ is a hypothesis on $\mathcal{Z}$, and $f_t = h_t(\Phi(x)) \in \mathcal{H}$. Let $\ell_{\Phi, h_t}(z) := \mathbb{E}_Y[L(h_t(z), Y(t)) \mid X = \Psi(z)]$ where $L(y, y') = (y - y')^2$. Further, let A_Φ be a constant such that $\forall z \in \mathcal{Z} : A_\Phi \geq |J_\Psi(z)|$, where $J_\Psi(z)$ is the Jacobian of the representation inverse Ψ, and assume that there exists a constant $B_\Phi > 0$ such that, with $C_\Phi := A_\Phi B_\Phi$, $\ell_{\Phi, h_t}/C_\Phi \in \mathcal{L}$, where $\mathcal{L}$ is a reproducing kernel Hilbert space of a kernel, k such that $k(x, x) < \infty$. Finally, let w be a valid re-weighting of $p_{\Phi, t}$. Then, with probability at least $1 - 2\delta$,*

$$R_{1-t}(f_t) \leq \hat{R}_t^w(f_t) + C_\Phi \cdot \mathrm{IPM}_\mathcal{L}(\hat{p}_{\Phi, 1-t}, \hat{p}_{\Phi, t}^w)$$

$$+ V_{p_t}(w, \ell_{f_t}) \frac{\mathcal{C}_{n_t, \delta}^\mathcal{H}}{n_t^{3/8}} + \mathcal{D}_{n_0, n_1, \delta}^{\Phi, \mathcal{L}} \left(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}} \right) + \sigma_{Y(t)}^2 \qquad (20)$$

where $\mathcal{C}_{n, \delta}^\mathcal{H}$ is a function of the pseudo-dimension of $\mathcal{H}$, $\mathcal{D}_{n_0, n_1, \delta}^\mathcal{L}$ is a function of the kernel norm of $\mathcal{L}$ (see Lemma 5), both only with logarithmic dependence on n and m, $\sigma_{Y(t)}^2$ is the expected variance in $Y(t)$, and $V_p(w, \ell_f) = \max\left(\sqrt{\mathbb{E}_p[w^2 \ell_f^2]}, \sqrt{\mathbb{E}_{\hat{p}}[w^2 \ell_f^2]} \right)$. A similar bound exists where $\mathcal{L}$ is the family of functions Lipschitz constant at most 1 and $\mathrm{IPM}_\mathcal{L}$ the Wasserstein distance, but with worse sample complexity.

Proof. The result follows from application Lemmas 4–5 to Lemma 6. The $n^{3/8}$ rate may be improved at the cost of a more complicated expression, see discussion following Theorem 3 in (Cortes et al., 2010). $\qquad\square$

Overlap, ignorability and invertibility. Theorem 2 holds both with and without treatment group overlap in X. It is important to note, however, that when we change the

covariate space from X to Φ, the assumption that $\ell_{\Phi,h_t}/C_\Phi \in \mathcal{L}$ is not guaranteed, even for large C_Φ, since information on which ℓ depends may have been (approximately) removed. In the context of risk minimization, information is only excluded from Φ if it is not predictive of the outcome Y, in which case it is independent also of ℓ. Thus, under the additional assumption of overlap, the assumption that $\ell_{\Phi,h_t}/C_\Phi \in \mathcal{L}$ is verifiable in the limit of infinite data. In Johansson et al. (2019), we expand on the effects of non-invertibility on identifiability of the marginal risk in much greater detail. In particular, we show that for non-invertible Φ, without overlap, the marginal risk may be bounded under the assumption that information removed in Φ is as important to the risk of the factual outcome as to that of the counterfactual. This assumption, however, is also unverifiable in general.

5.1 Relation to Unsupervised Domain Adaptation

Connections between the problem of estimating causal effects and learning under distributional shift have been pointed out in several contexts (Tian and Pearl, 2001; Zhang et al., 2013). In particular, Johansson et al. (2016) showed that estimating counterfactual outcomes under ignorability is mathematically equivalent to unsupervised domain adaptation between domains $D \in \{0,1\}$ under covariate shift. We make this connection precise below.

Task	Data	Goal	Assumption
Causal estimation	Factual $(x,t,y) \sim p(X,T,Y)$	Counterfactual $p(Y(1-T) \mid X,T)$	Ignorability $Y(t) \perp\!\!\!\perp T \mid X$
Domain adaptation	Source domain $(x,y) \sim p(X,Y \mid D=0)$	Target label $p(Y \mid X, D=1)$	Covariate shift $Y \perp\!\!\!\perp D \mid X$

The bounds we present in this work are related to a series of work on generalization theory for unsupervised domain adaptation (Ben-David et al., 2007; Mansour et al., 2009; Long et al., 2015), but differ in significant ways. Superficially, the bounds given in these papers have a similar form using the sum of observed risk in the source domain and distributional distance w.r.t. a function class $\mathcal{H}$ to bound the risk in the target domain:

$$R_{D=1}(f) \le R_{D=0}(f) + d_\mathcal{H}(p(X \mid D=1), p(X \mid D=0)) + \lambda_\mathcal{H}.$$

Similarly, these bounds do not rely on overlap but cannot guarantee consistent estimation in the general case. In fact, because they do not allow for re-weighting of domains, even when source and target domains completely overlap, these bounds are often unnecessarily loose. Furthermore, while they are used to motivate representation learning algorithms, these bounds do not apply to learned representations without modification (Johansson et al., 2019). In this work, we overcome this issue by requiring that representations Φ are invertible.

6. Estimation

We turn now to deriving practical algorithms inspired by our theoretical results. First, we give learning objectives for estimating potential outcomes and CATE grounded in the theoretical results of Sections 4–5, and show that they lead to asymptotically consistent estimation. Then, we describe possible parameterizations of representation functions Φ, discuss possible drawbacks of estimating potential outcomes separately, and introduce a neural network architecture and objective for shared representation learning between treatment groups. While these parameterizations are not always guaranteed to minimize our bounds, they benefit from the insights gathered from them, as shown in empirical results.

6.1 A First Learning Objective and Asymptotic Consistency

Let $D = \{(x_1, t_1, y_1), ..., (x_n, t_n, y_n)\}$ be a set of samples drawn i.i.d. from $p(X, T, Y)$ and let each sample i be endowed with weights $w_i = w(x_i, t_i)$ for some function $w : \mathcal{X} \times \{0, 1\} \to \mathbb{R}_+$. Let $n_t = \sum_{i=1}^{n} \mathbb{1}[t_i = t]$ be the number of samples with treatment assignment t. Further, let $\lambda, \alpha > 0$ be hyperparameters controlling the strength of the regularization of functional complexity, as measured by $\mathcal{R}$, and distributional distance $\text{IPM}_{\mathcal{L}}$ respectively. Recall that $\hat{p}_{\Phi,t}^w$ is the re-weighted factual distribution of representations Φ under p_t. Now, we consider using compositions $f_t = (h_t \circ \Phi) \in \mathcal{H}$ of hypotheses $h_t \in \mathcal{G} \subseteq \{\mathcal{Z} \to \mathcal{Y}\}$ and representations $\Phi \in \mathcal{E} \subseteq \{\mathcal{X} \to \mathcal{Z}\}$ to estimate a single expected potential outcome $\mathbb{E}[Y(t) \mid X]$. Then, directly motivated by Theorems 1–2, we propose to minimize the following learning objective, with hyperparameters $\beta = (\lambda, \alpha)$,

$$\mathcal{O}_t(h_t, \Phi; \beta) = \underbrace{\sum_{i:t_i=t} \frac{\tilde{w}_i}{n_t} L(h_t(\Phi(x_i)), y_i)}_{\text{Weighted factual risk}} + \underbrace{\frac{\lambda}{\sqrt{n_t}} \mathcal{R}(h_t)}_{\text{Regularization}} + \underbrace{\alpha \pi_{1-t} \, \text{IPM}_{\mathcal{L}}(\hat{p}_{\Phi,t}^w, \hat{p}_{\Phi,1-t})}_{\text{Distributional regularization}} \quad (21)$$

where $\tilde{w}_i = \pi_t + \pi_{1-t} w_i$ and $\pi_t = p(T = t)$.

Under Assumptions 1 (ignorability) and 2 (overlap), for balancing weights $w_i = p(T = t_i)/p(T = t_i \mid X = x_i)$, objective (21) reduces to inverse propensity-weighted regression in the limit of infinite samples (Freedman and Berk, 2008). In the finite-sample regime, the IPM does not vanish even if $p_0 = p_1$, because of sample variance.

As pointed out previously, the results of Theorems 1–2 remain upper bounds on the CATE risk for weights other than the inverse propensity score. In fact, as shown by Hirano et al. (2003), weighting using the true propensity score does not always lead to the most efficient estimation. Thus, in addition to learning representations and hypotheses, we may consider learning the sample weights w jointly, controlling the variance introduced by non-uniform weights by regularizing the norm of w (Johansson et al., 2018). With $\beta = (\alpha, \lambda_h, \lambda_w)$ a set of hyperparameters, we consider the following objective,

$$\begin{aligned} \mathcal{O}_t(h_t, \Phi, w^t; \beta) &= \sum_{i:t_i=t} \frac{\tilde{w}_i^t}{n_t} L(h_t(\Phi(x_i)), y_i) + \frac{\lambda_h}{\sqrt{n_t}} \mathcal{R}(h_t) \\ &\quad + \alpha \pi_{1-t} \, \text{IPM}_{\mathcal{L}}(\hat{p}_{\Phi,t}^{w^t}, \hat{p}_{\Phi,1-t}) + \frac{\lambda_w}{n_t} \|w^t\|_2 \end{aligned} \quad (22)$$

with $\tilde{w}_i^t = \pi_t + \pi_{1-t} w_i^t$ and subject to the constraint $\sum_i w_i^t = n_t$.

A theoretical advantage of using objective (22) is that it allows for an explicit tradeoff between bias and variance induced by the sample weights. In principle, the weighting function can be implemented using one free parameter per sample point. However, our formulation allows the weights to be defined in terms of the learned representation Φ rather than free parameters or as a function of X. This imposes further restrictions on the variance in and induced by w. We proceed to give conditions under which minimization of (21) leads to consistent estimation of conditional expected potential outcomes.

Theorem 3. *Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) given by a bounded kernel k, such that for all $h_t \in \mathcal{G}, \Phi \in \mathcal{E}, h_t \circ \Phi \in \mathcal{H}$. Suppose weak overlap holds in that*

$$\forall t \in \{0,1\} : \mathbb{E}_X[(p_t(X)/p_{1-t}(X))^2] < \infty .$$

Let $\mathcal{O}_t$ be the objective defined in (22), with the IPM-space $\mathcal{L}$ also an RKHS with bounded kernel $k_{\mathcal{L}}$, and $n_t = \sum_{i=1}^n \mathbb{1}[t_i = t]$ for $t \in \{0,1\}$. Then,

$$\min_{h_t,\Phi,w^t} \mathcal{O}_t(h_t, \Phi, w^t; \beta) \leq \min_{f_t \in \mathcal{H}} R(f_t) + O_p(1/\sqrt{n_0} + 1/\sqrt{n_1}) ,$$

where O_p denotes stochastic boundedness. Thus, under the assumptions of Thm. 2, for sufficiently large α and λ_w, with $\hat{f}^n$ the minimizer of (22) for n samples,

$$R(\hat{f}_t^n) \leq \min_{f_t \in \mathcal{H}} R(f_t) + O_p(1/n_0^{3/8} + 1/n_1^{3/8}).$$

In words, the minimizers of (22) converge to the representation and hypothesis that minimize the counterfactual risk, in the limit of infinite samples.

Thorem 3 is proven in Appendix A. The result may be generalized further to the case of estimating the value of an abritrary treatment policy under a shift in marginal distribution $p(X)$. See the Appendix and Johansson et al. (2018) for results in this setting.

Relating bound and learning objective. Learning objectives (21)–(22) are direct results of Theorem 2. However, unlike the factor C_Φ in (20), the hyperparameter α, which takes its place in the objectives, is independent of the representation Φ. As remarked previously, when Φ shrinks the representation space, C_Φ must grow for Theorem 2 to be valid. Hence, any fixed α might be either too small during the learning of Φ, implying that the learning objective is not a bound on the counterfactual risk, or too large, over-penalizing distributional distance (see Figure 8a). We view the role of α similar to that of the regularization parameters of LASSO and ridge regression—not for directly evaluating a generalization bound, but for controlling the tradeoff between bias and variance (Friedman et al., 2001). In empirical results, we find that a fixed value of α can help learning an estimator with lower counterfactual risk.

6.2 Practical Representation Learning and Learning Shared Representations

A well-selected or learned representation Φ can lead to estimates of potential outcomes and CATE with improved counterfactual risk. The two most prominent approaches to learning
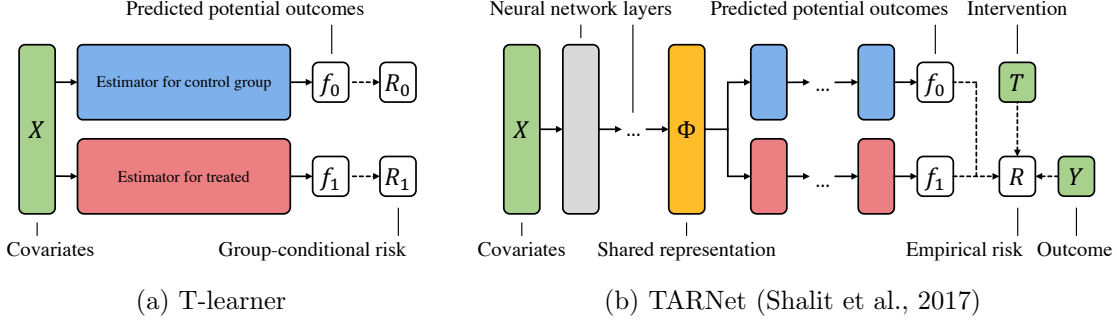
Figure 6: Estimator architectures for potential outcomes and conditional average treatment effects. Green boxes indicate inputs, white boxes outputs and loss terms, yellow boxes shared representations and blue/red boxes estimators of potential outcomes. Solid lines indicate transformation part of the prediction function and dashed lines indicate computations part of the learning procedure.

representations in the literature are i) deep neural networks (Bengio et al., 2013) and ii) variable selection (Schneeweiss et al., 2009). As non-trivial examples of the latter cannot satisfy the assumption of invertibility of Φ even approximately, we restrict our attention to parameterizations of Φ as neural networks. Neural networks have been made invertible by design in the context of normalizing flows (Kobyzev et al., 2020) and have been observed to be approximately invertible spontaneously (Gilbert et al., 2017). In experiments, we do not enforce or guarantee invertibility of our neural networks but nevertheless benefit from the insights in our theoretical results in regularizing treatment group distance in representations.

Parameterizing Φ with neural networks leaves a lot of freedom in the design of estimators. A straight-forward application minimizes the factual risk or objective (22) independently for each potential outcome, with separate neural network architectures. This approach is that of T-learners, which fit two models, one for each treatment group, as described Section 3. A potential drawback of this approach is that in many cases, much of the variance in *both* potential outcomes may be explained by the same patterns in the input (Künzel et al., 2019). Estimating each outcome independently fails to take advantage of this fact.

We proposed a natural extension in the Treatment-Agnostic Representation Network (TARNet) in Shalit et al. (2017). In TARNet, a T-learner architecture is appended to a representation Φ shared between treatment groups (see Figure 6 for a comparison with T-learners). TARNet has the advantage of sharing samples between treatment groups in learning the representation which may be useful when τ is a simpler function of X than $Y(0), Y(1)$. It improves on classical T-learning estimators by allowing estimators of different potential outcomes to share information through representation functions learned from both treatment groups. TARNet minimizes the following objective,

$$\mathcal{O}_{\text{TARNet}}(h, \Phi) = \frac{1}{n} \sum_{i=1}^{n} L(h(\Phi(x_i), t_i), y_i) + \frac{\lambda_h}{\sqrt{n}} \mathcal{R}(h) \tag{23}$$

By also regularizing treatment group distance in learned representations, we may enable better counterfactual generalization, in line with Theorem 2. The CounterFactual Regres-
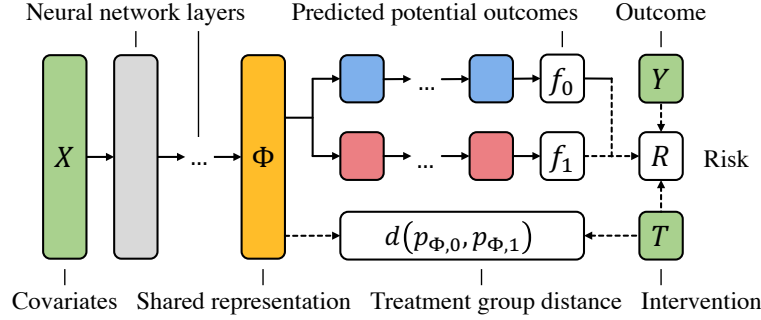
Figure 7: Illustration of the Counterfactual Regression (CFR) estimator. Here, d represents a distributional distance such as an IPM. The visual elements are described in Figure 6.

sion (CFR) estimator, introduced in Shalit et al. (2017) and illustrated in Figure 7, applies this idea to TARNet, minimizing the objective

$$\mathcal{O}_{\text{CFR}}(h, \Phi) = \frac{1}{n} \sum_{i=1}^{n} L(h(\Phi(x_i), t_i), y_i) + \alpha \cdot \text{IPM}_{\mathcal{L}}(\hat{p}_{\Phi,0}, \hat{p}_{\Phi,1}) + \frac{\lambda_h}{\sqrt{n}} \mathcal{R}(h) \ . \tag{24}$$

Finally, the re-weighted CFR (RCFR) minimizes a weighted objective analogous to objective (22) but with a shared representation (Johansson et al., 2018).

$$\mathcal{O}_{\text{RCFR}}(h, \Phi, w) = \sum_{i=1}^{n} \frac{w_i}{n} L(h(\Phi(x_i), t_i), y_i) + \alpha \cdot \text{IPM}_{\mathcal{L}}(\hat{p}_{\Phi,0}^w, \hat{p}_{\Phi,1}^w) + \frac{\mathcal{R}(h)}{\lambda_h^{-1}\sqrt{n}} + \frac{\|w\|_2}{\lambda_w^{-1} n} \tag{25}$$

subject to the constraint $\sum_{i:t_i=t} w_i = n_t$ for $t \in \{0, 1\}$. The weights are given by a learned function, parameterized by a neural network, to avoid over-penalizing distributional distance and to trade off the effects of selection bias and variance. See Appendix C for an illustration of the neural network architecture used in RCFR.

6.3 Implementing Regularization of Distributional Distance

The idea of regularizing models to be invariant to changes in a variable, in our case the treatment indicator, is prevalent through-out machine learning (Ganin et al., 2016; Goodfellow et al., 2014; Long et al., 2015). As a result, several families of distance metrics between distributions have been used to impose such constraints. The most common of these are f-divergences (e.g. the KL-divergence) (Nowozin et al., 2016), integral probability metrics (e.g., the maximum-mean discrepancy) and adversarial discriminators (Ganin et al., 2016). f-divergences are often ill-suited for comparing two empirical densities as they are based on the density ratio which is undefined in any point outside of the support of either density. In contrast, IPMs are based on the density difference which is defined everywhere. Adversarial methods are based on the metric implied by a learned discriminator function which is trained to distinguish samples from the two densities. The flexibility of this approach—that it tailors the metric to observed data—is also its weakness since optimization of adversarial discriminators is fraught with difficulty. In implementations of CFR and RCFR, we use the

empirical kernel MMD (Gretton et al., 2012) and the Wasserstein distance (Villani, 2008), both examples of integral probability metrics.

Minimizing the empirical maximum mean discrepancy. The maximum mean discrepancy (MMD) was popularized in machine learning through its kernel-based incarnation in which the associated function family is a reproducing kernel Hilberg space (RKHS) (Gretton et al., 2012). We restrict our attention to this family here. An unbiased estimator of the MMD distance between densities p, q on $\mathcal{X}$, with respect to a kernel k, may be obtained from samples $x_1, ..., x_m \sim p$, $x'_1, ..., x'_n \sim q$ as follows.

$$\widehat{\text{MMD}}_k^2(p, q) := \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, x'_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x'_i, x'_j)$$

By choosing a differentiable kernel k, such as the Gaussian RBF-kernel, we can ensure that the MMD is amenable to gradient-based learning. In applications where the quadratic time complexity w.r.t. sample size is prohibitively large, another unbiased estimator (but with larger variance) may be obtained by sampling pairs of points $(x_1, x'_1), ..., (x_{2n}, x'_{2n}) \sim p \times q$ and comparing only elements within pairs (Long et al., 2015),

$$\widehat{\text{MMD}}_k^2(p, q) := \frac{1}{n} \sum_{i=1}^{n} \left[k(x_{2i-1}, x_{2i}) + k(x'_{2i-1}, x'_{2i}) - k(x_{2i-1}, x'_{2i}) - k(x_{2i}, x'_{2i-1}) \right] .$$

Minimizing the Wasserstein distance. The Wasserstein distance is typically computed as the solution to a linear program (LP). The gradient of this solution with respect to the learned representation may be obtained through the KKT conditions of the problem and the solution for the current representation (Amos and Kolter, 2017). However, solving the LP at each gradient update is prohibitively expensive for many applications. Instead, we minimize an approximation of the distance known as Sinkhorn distances (Cuturi, 2013), computed using fixed-point iteration. In previous work (Shalit et al., 2017), we computed the distance and its gradient by forward and backpropagation through a recurrent neural network with transition matrix corresponding to the fixed-point update. For a full description, see Appendix D. Alternative methods for minimizing Wasserstein distances have been developed in the context of generative adversarial networks (GANs) (Arjovsky et al., 2017).

7. Experiments

Evaluating estimates of potential outcomes and causal effects from observational data is notoriously difficult as ground-truth labels are hard or impossible to come by. Cross-validation and other sample splitting schemes frequently used to evaluate supervised learning are not immediately applicable to our setting for this reason. Moreover, the task of producing the labels themselves is exactly the task we are attempting to solve. As a result, estimation methods are often evaluated on synthetic or semi-synthetic data, where consistent estimation or computation of the labels are guaranteed. Another alternative is using real-world data where the treatment-assignment randomization is known, e.g. data from an RCT. In this section, we give a suite of experimental results on synthetic, semi-synthetic and real-world data. Our experiments are developed to separately highlight the impact of architecture choice and the balancing regularization scheme.

7.1 Experimental Setup & Baselines

We evaluate learning to predict potential outcomes using the TARNet and CFR objectives, (23) and (24). Recall that TARNet is equivalent to CFR with the parameter $\alpha = 0$.

We specify the function family used in the IPM by a subscript, e.g., $\mathrm{CFR_{MMD}}$, and point out for which experiments the weighting function is learned and for which it is set to the uniform weighting. All variants of CFR were implemented as feed-forward neural networks with exponential-linear units and architectures as described in Section 6. These implementations do not guarantee the invertibility condition of our theoretical results but benefit from the regularization scheme derived from them. Further, in Johansson et al. (2019); Gilbert et al. (2017), it was found that similar networks satsify invertibility approximately without explicitly being made to. Ranges for hyperparameters, such as layer sizes, learning rates et cetera, are described in Appendix B and specific values were selected according to a procedure below. An implementation of CFR with uniform sample weights may be found at `https://github.com/clinicalml/cfrnet`.

As our primary baseline, we use two variants of Ordinary Least Squares (linear regression). The first (OLS-S) adopts the S-learner paradigm and includes the treatment variable T as a feature in the regression. The second (OLS-T) is a T-learner where the outcome in each treatment arm is modeled using a separate linear regression. Our other simple baseline is a k-nearest neigbor regression which imputes counterfactual outcomes of a unit by the average of its k-nearest neighbors with the opposite treatment assignment.

For a more challenging comparison, we use Targeted Maximum Likelihood, which is a doubly robust method (TMLE) (Gruber and van der Laan, 2012) which uses an ensemble of machine-learning methods. We also compare with a suite of tree-based estimators: First, we use a Random Forest (RAND. FOR.) (Breiman, 2001) in the S-learner paradigm by including T as a feature. Second, we include tree-based methods specifically designed or adapted for causal effect estimation: Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Chipman and McCulloch, 2016) and Causal Forests (CAUS. FOR.) (Wager and Athey, 2018; Athey, 2016). Finally, we also compare with our earlier work on Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) (Johansson et al., 2016).

7.1.1 EVALUATION CRITERIA & HYPERPARAMETER SELECTION

To assess the quality of CATE estimates, either knowledge of the propensity score or the outcome function is required. Where labels are available, our primary criterion for evaluation is the mean squared error in the imputed CATE as defined in (6). When only the propensity score is available, such as in a randomized controlled trial or other experiments, we instead estimate the *policy risk* as defined below.

A policy π is any (possibly stochastic) function that maps from covariates x to treatment decision $t \in \{0, 1\}$; we will only consider deterministic policies. The risk of a policy π for outcomes $Y \in [0, 1]$, where large Y is considered beneficial, is

$$R_{\mathrm{Pol}}(\pi) := 1 - \mathbb{E}_X[E_{(Y(0), Y(1))}[Y(\pi(x)) \mid X = x]] .$$

A good policy is one that for a given x will choose the potential outcome with the higher conditional expectation given x. If we know the true propensity scores $p^*(t_i = 1 | x_i)$ used in generating the data set, then the risk of a deterministic policy $R_{pol}(\pi)$ may be estimated

using rejection sampling based on a sample $(x_1, t_1, y_1), \ldots, (x_m, t_m, y_m)$ and propensity scores $p^*(t_1 = 1|x_1), \ldots, p^*(t_m = 1|x_m)$ by considering only the *propensity re-weighted effective sample* on which the proposed policy agrees with the observed one:

$$\hat{R}_{\text{Pol}}(\pi) := 1 - \frac{\sum_{i=1}^m y_i \mathbb{1}[\pi(x_i) = t_i] \frac{1}{p^*(t_i = \pi(x_i)|x_i)}}{\sum_{i=1}^m \mathbb{1}[\pi(x_i) = t_i]} \ . \tag{26}$$

A downside of this estimator is that it has very high variance for policies that are very different from the observed policy. Note that in the case where the data was generated by an RCT with equal probability of treatment and control, the propensity scores have a particularly simple form: $p(t = 1|x) = 0.5$ for all x.

In our experiments, we evaluate the policy $\pi_f : \mathcal{X} \to \mathcal{T}$ induced by an estimator $f(x, t)$ of potential outcomes and a threshold λ such that

$$\pi_f(x) := \begin{cases} 1, & \text{if } f(x, 1) - f(x, 0) > \lambda \\ 0, & \text{otherwise} \end{cases} \ . \tag{27}$$

By varying λ from low to high we obtain a curve that interpolates between liberal and conservative allocation of treatment.

In all experiments we fit a model on a training set and then evaluate on a held-out set. We always report results both within-sample and out-of-sample. We wish to emphasize that the within-sample results should *not* be thought of as training-loss in standard ML problems. Even within-sample results include the challenging task of inferring unobserved counterfactuals for the training samples.

Hyperparameter selection. We choose hyperparameters for all estimators in the same way. As the ground truth potential outcomes are unavailable to us, we use pseudo-labels for the true CATE imputed using a nearest-neighbor estimator. With $j(i)$ the nearest "counterfactual" neighbor of sample i in Euclidean distance, such that $t_{j(i)} \neq t_i$, we define

$$\widehat{\text{MSE}}_{\text{nn}}(f) := \frac{1}{n} \sum_{i=1}^n \left((1 - 2t_i)(y_{j(i)} - y_i) - (f(x_i, 1) - f(x_i, 0)) \right)^2$$

and use its value on a held-out validation set as a surrogate for the true MSE in $\hat{\tau}$ in hyperparameter section. This choice may bias selection of hyperparameters towards preferring models close to a nearest-neighbor estimator, but we anticipate this effect to be mild as $\widehat{\text{MSE}}_{\text{nn}}$ is not used as a training objective. For neural network estimators, we perform early stopping based on the training objective evaluated on a held-out validation set in the IHDP study, and based on held-out policy risk in the Jobs study (both described below). Ranges for hyperparameters for CFR are presented in Appendix B.

7.2 Synthesized Outcome: IHDP

The Infant Health and Development Program (IHDP) data set has been frequently used to evaluate machine learning approaches to causal effect estimation in recent years (Hill, 2011). The orginal data comes from a randomized study of the impact on educational and follow-up interventions on child cognitive development (Brooks-Gunn et al., 1992). Each

	Within sample		Out of sample	
	MSE CATE	MSE ATE	MSE CATE	MSE ATE
OLS-S	5.8 ± 0.3	0.73 ± 0.04	5.8 ± 0.3	0.94 ± 0.06
OLS-T	2.4 ± 0.1	0.14 ± 0.01	2.5 ± 0.1	0.31 ± 0.02
BLR	5.8 ± 0.3	0.72 ± 0.04	5.8 ± 0.3	0.93 ± 0.05
k-NN	2.1 ± 0.1	0.14 ± 0.01	4.1 ± 0.2	0.79 ± 0.05
TMLE	5.0 ± 0.2	0.30 ± 0.01	$\dagger$	$\dagger$
BART	2.1 ± 0.1	0.23 ± 0.01	2.3 ± 0.1	0.34 ± 0.02
R.For.	4.2 ± 0.2	0.73 ± 0.05	6.6 ± 0.3	0.96 ± 0.06
C.For.	3.8 ± 0.2	0.18 ± 0.01	3.8 ± 0.2	0.40 ± 0.03
BNN	2.2 ± 0.1	0.37 ± 0.03	2.1 ± 0.1	0.42 ± 0.03
TARNet	0.88 ± 0.02	0.26 ± 0.01	0.95 ± 0.02	0.28 ± 0.01
CFR_{MMD}	0.73 ± 0.01	0.30 ± 0.01	0.78 ± 0.02	0.31 ± 0.01
CFR_{Wass}	0.71 ± 0.02	0.25 ± 0.01	0.76 ± 0.02	0.27 ± 0.01

Table 1: Mean squared error, and standard error over 1000 random draws of the outcome model, in estimates of CATE and ATE on IHDP within-sample (left) and out-of-sample (right). Lower is better. $\dagger$Not applicable.

observation represents a single child in terms of 25 features of their birth and their mothers. To introduce confounding, Hill (2011) removed a biased subset of the treatment group—all treated children with nonwhite mothers—leaving 747 subjects in total. This induces not only confounding, but also lack of overlap in variables strongly correlated with race (race itself was removed from the feature set following the biased selection). To enable consistent evaluation, the outcome of the IHDP data set was synthesized according to several different stochastic models on the *observed* feature set. In this way, ignorability is guaranteed. Depending on the specific sample of the outcome model, i.e., whether variables correlated with race have strong influence or not, the lack of overlap varies in its impact on the results.

In our experiments, we use observations generated using setting "A" in the NPCI package (Dorie, 2016), corresponding to response surface (outcome function) "B" in Hill (2011). This model follows an exponential-linear form for the outcome under treatment and a linear form for the controls, ensuring that their difference, CATE, is a nonlinear function. Sparsity in the coefficients is introduced through random sampling with a probability 0.6 that a coefficient is exactly equal to 0. The full description of the model may be found in Hill (2011), Section 4.1. The specific realizations (draws) used in our evaluation can be accessed at `http://www.mit.edu/~fredrikj/`.

Results. The error in estimates of CATE on IHDP can be seen in Table 1. Here, we present only the variants of CFRwith uniform weighting, and refer to Figure 8b for a comparison between learned and uniform weights. First, we note that all of the proposed neural network estimators (TARNet and CFR variants) outperform the selected baselines. We attribute this, to a large extent, to multi-layer neural networks being a suitable function class for this data set. CFR improves marginally over TARNet, indicating that regular-
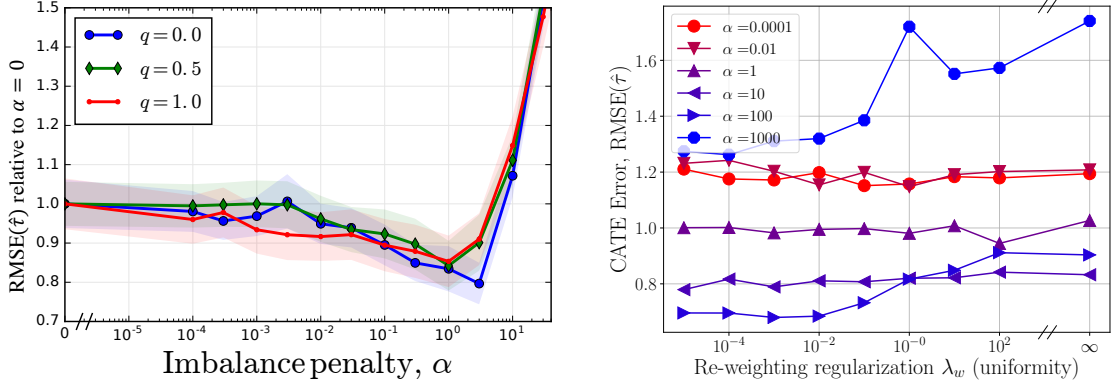
(a) Ratio of mean squared squared error in estimated CATE relative to $\alpha = 0$, as a function of the imbalance regularization α with uniform sample weights w, for different levels of introduced additional treatment group imbalance q. Uncertainty bands show standard errors over 500 realizations.

(b) RMSE for CFR estimates combining representation learning and re-weighting, minimizing (22), for varying weight regularization λ_w and imbalance penalty α. Higher λ_w leads to more uniform weights.

Figure 8: Results for estimating CATE on IHDP with different variants of the CFR model. In (a), we show results for the best performing architecture with uniform sample weights, varying the imbalance regularization α. In (b), we show the results for a smaller architecture and their dependence on the uniformity of learned weights.

izing distributional invariance is beneficial for prediction of CATE. We note also that, in general, the S-learner estimators (OLS-S and BNN) perform worse than separate or partially separate estimators (OLS-T, TARNet). The biggest differences between in-sample and out-of-sample performance are attained by the k-NN and random forest estimators.

Increasing imbalance. In Figure 8a, we study the effect of increasing the imbalance between treatment groups through biased subsampling. To do this, we fit a logistic regression propensity score model $\hat{p}(T = 1 \mid X = x)$ and for a parameter $q \geq 0$, we repeatedly remove the *control* sample with largest estimated propensity with probability, q and a random control observation with probability $1 - q$, until 400 samples remain. For three values of q, we estimate CATE using CFR with uniform sample weights for different values of the penalty α of treatment group distance in the learned representation Φ. We see that for small α, as expected, the relative error is comparable to TARNet ($\alpha = 0$), but that it decreases until $\alpha \approx 1$. For $\alpha > 2$, the performance deteriorates as the influence of the input on the representation is constrained too heavily. As we'll see below, this may be partially remedied by sample weighting.

Learning the sample weights. In Figure 8b, we study the quality of CFR estimates when sample weights are learned by minimizing objective (22). In this setting, the chosen model is intentionally restricted to have representations Φ of two layers with 32 and 16 hidden units each and hypotheses $h(\Phi)$ of a single layer with 16 units. This choice was made to emphasize the value of reweighting under model misspecification. The weighting function

33

was modeled using two layers of 32 units each. We see in Figure 8b that a model using non-uniform sample weights (λ_w small) is less sensitive to excessively large penalties α. This is because the IPM term may now be minimized also by learning the weights, rather than only by constraining the capacity of Φ. In the small-α regime, the non-uniformity of weights has almost no impact, as the incentive to reduce the IPM using the weights is too small. In this experiment, the best results are attained for combination of a considerably larger value of α and small penalty on the non-uniformity of weights. In general, we do not observe any adverse effects of having a small value of λ_w. This is likely due to the choice of architecture for the learned weighting function already constraining the weights.

7.3 Partially Randomized Study: National Supported Work Program

LaLonde (1986) carried out a widely known experimental study of the effect of job training on future income and employment status based on the National Supported Work (NSW) program. Later, Smith and Todd (2005) combined the LaLonde study with observational data to form a larger data set which has been used frequently as a benchmark in the causal inference community. The presence of the randomized subgroup allows for straightforward estimation of average treatment effects and policy value.

The original study by Smith and Todd (2005) includes 8 covariates such as age and education, as well as previous earnings. The treatment indicates participation in the NSW job training program. By construction, all treated subjects belong to the LaLonde experimental cohort; the observational cohort includes only controls. Additionally, the nature of the observational cohort is such that overlap is minimal at best—the experimental cohort may be separated from the observational using a linear classifier with 96% accuracy. This means that global estimators of the control outcome applied to the treated, such as linear models or difference-in-means estimators of causal effects are likely to suffer severe bias if not re-weighted.

Based on the original outcome measuring yearly earnings at the end of the study, we construct a binary classification task called Jobs, in which the goal is to predict unemployment. Following Dehejia and Wahba (2002), we use an expanded feature set that introduces interaction terms between some of the covariates. The task is based on the cohort used by Smith and Todd (2005) which combines the LaLonde experimental sample (297 treated, 425 control) and the "PSID" comparison group (2490 control). There were 482 (15%) subjects unemployed by the end of the study. In our experiments, we average results over 10 train/validation/test splits of the full cohort with ratios 56/24/20. We train CFR methods with uniform weighting, according to (21), selecting the imbalance parameter α according to held-out policy risk.

Results. In Table 2, we give the policy risk $\hat{R}_{\mathrm{Pol}}$ evaluated over the randomized component of Jobs, as defined in (26), and mean squared error in the estimated average treatment effect on the treated (MSE ATT). The policy we consider in the table assigns treatment to the top subject for which the CATE is estimated to be positive ($\lambda = 0$). We can see from the results that, despite the significant lack of overlap, the difference between linear and non-linear estimators is much less pronounced than for IHDP. This is likely partly due to the features used in the jobs data set which have been handcrafted to predict the outcome of interest well. In contrast, the IHDP outcome is non-linear by construction.

	Within sample		Out of sample	
	$\hat{R}_{\mathrm{Pol}}$	MSE ATT	$\hat{R}_{\mathrm{Pol}}$	MSE ATT
LR-S	0.22 ± 0.00	0.01 ± 0.00	0.23 ± 0.02	0.08 ± 0.04
LR-T	0.21 ± 0.00	0.01 ± 0.01	0.24 ± 0.01	0.08 ± 0.03
BLR	0.22 ± 0.01	0.01 ± 0.01	$0.25 \pm .02$	0.08 ± 0.03
k-NN	0.02 ± 0.00	0.21 ± 0.01	0.26 ± 0.02	0.13 ± 0.05
TMLE	0.22 ± 0.00	0.02 ± 0.01	†	†
BART	0.23 ± 0.00	0.02 ± 0.00	0.25 ± 0.02	0.08 ± 0.03
R.For.	0.23 ± 0.01	0.03 ± 0.01	0.28 ± 0.02	0.09 ± 0.04
C.For.	0.19 ± 0.00	0.03 ± 0.01	0.20 ± 0.02	0.07 ± 0.03
BNN	0.20 ± 0.01	0.04 ± 0.01	0.24 ± 0.02	0.09 ± 0.04
TARNet	0.17 ± 0.01	0.05 ± 0.02	0.21 ± 0.01	0.11 ± 0.04
CFR$_{\mathrm{MMD}}$	0.18 ± 0.00	0.04 ± 0.01	0.21 ± 0.01	0.08 ± 0.03
CFR$_{\mathrm{Wass}}$	0.17 ± 0.01	0.04 ± 0.01	0.21 ± 0.01	0.09 ± 0.03

Table 2: Policy risk and mean squared error in estimates of ATT on Jobs within-sample (left) and out-of-sample (right). Lower is better. †Not applicable.
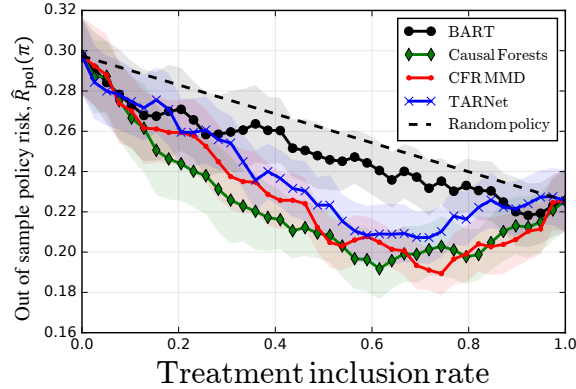

Figure 9: Policy risk as a function of treatment inclusion rate on Jobs. Lower is better. Subjects are included in treatment in order of their estimated treatment effect given by the various methods. CFR Wass is similar to CFR MMD and is omitted to avoid clutter

We also see that straightforward logistic regression does remarkably well in estimating the ATT. However, being a linear model, logistic regression can only ascribe a uniform policy – in this case, "treat everyone". The more nuanced policies offered by non-linear methods achieve lower policy risk, though this difference is less pronounced in the out of sample case, indicating that part of the difference may be due to overfitting. The nearest-neighbor estimator k-NN appears to perform incredibly well within-sample, but generalizes poorly to the hold-out. Additionally, its estimate of the ATT is the worst among the baselines.

In Figure 9, we plot policy risk as a function of treatment threshold λ, as defined in (27). This is described in the figure as varying the fraction of subjects treated in a policy that

treats only the subjects with the largest estimated CATE. Overall, the benefits of imbalance regularization of the CFR models offer less advantage than on IHDP. This may be due to the smaller covariate set of Jobs containing less redundant features than those in IHDP. Recall that the IHDP outcome coefficients have 60% sparsity in the feature set, by design. In contrast, the Jobs covariate set has been hand-picked to account for confounding. This means that one of the benefits of imbalance-regularizations of representations—to exclude variables only predictive of treatment—is likely to have a smaller effect in comparison.

8. Discussion

We have presented generalization bounds for estimation of potential outcomes and causal effects from observational data. These bounds were used to derive learning objectives for estimation algorithms that proved successful in empirical evaluation. The bounds are based on integral probability metric distances between treatment groups, rather than the density ratios commonly used in observational studies of causal effects, making them valid also when overlap is violated, albeit loose when the overlap is small. This assumption states that for any one observed subject, there is some probability that they were prescribed either treatment option. Removing this assumption means that we cannot identify the causal effect non-parametrically but, as we show in this work, we can still bound the expected error (risk) of any hypothesis in a given class.

When can we expect overlap to not hold yet identification to be possible? One example is when many of the covariates in the conditioning set $\mathcal{X}$ have a strong effect on treatment but only a weak or non-existent effect on the outcome. For example, if some of the covariates in $\mathcal{X}$ are actually instrumental variables, conditioning on them might substantially increase variance and reduce overlap, with no gain in estimating the CATE function (Brookhart et al., 2010; Shortreed and Ertefaie, 2017). We conjecture that this might often be the case in high-dimensional cases: in aggregate, there might not be nominal overlap with respect to the measured covariates, while at the same time many of them are actually only weak confounders, or even not confounders at all; see also D'Amour et al. (2021).

Our results offer several new perspectives on causal effect estimation. In particular, they bring together two hitherto separate approaches to dealing with treatment group shift—representation learning and sample re-weighting—and give insight as to when either approach is likely to be more successful, and when should they be used together. It is well known that under the overlap and ignorability assumptions, ordinary risk minimization leads to consistent estimation of causal effects (Pearl, 2009; Ben-David and Urner, 2012; Alaa and Schaar, 2018) in the limit of infinite samples, but the hardness of the problem is less well understood in the finite sample case, or when overlap is violated. Our results provide some insight in this setting.

The bounds in this work are limited to binary treatment choices. A natural question to ask is how to generalize our results to settings with multiple discrete treatments or continuous ones. In the former case, our bounds on the risk in predicting potential outcomes can be trivially extended using a one-vs-all approach, evaluating the factual risk for subjects on a particular treatment and the counterfactual risk of subjects on any other treatment. However, the bounds for causal effects do not extend in the same way to this approach since they are aimed at evaluating the difference between two potential outcomes. An

alternative for this purpose is to consider measuring the all-pairs distributional difference between treatment groups, but this is unlikely to be efficient in practice for large numbers of treatments. The right approach will depend on the causal estimand of interest. In real-world applications, comparing the value of treatment policies over a large number of treatments may be more important than comparing the effects of each pair of treatments. To this end, Liu et al. (2018) applied ideas put forth in this work to off-policy policy evaluation.

It is customary in machine learning to evaluate methodological progress based on performance on a small number of well-established benchmarks, such as MNIST (LeCun et al., 1998) or ImageNet (Deng et al., 2009). Similarly, IHDP has become a de facto benchmark for causal effect estimation (Hill, 2011; Shalit et al., 2017; Alaa and Schaar, 2018; Shi et al., 2019; Curth et al., 2021). However, IHDP is smaller than most machine learning benchmarks and susceptible to "test set overfitting". Even disregarding the size issue, it may be argued that benchmarks for causal effect estimation are more prone to going stale as the strong assumptions we make (or synthesize) need not hold in the tasks we wish to apply our models to (Curth et al., 2021). Moreover, the relatively simple form of the outcome model, the small dimensionality, and the structured fully observed nature of the data makes IHDP a much easier challenge than what we may face in for example analysis of electronic healthcare records. The Jobs data set, on the other hand, offers realistic data but suffers from small sample size and a small number of measured covariates. Towards understanding the behavior of different estimators, data sets like IHDP and Jobs are not necessarily representative samples of the problems we may encounter in applications. However, we believe that the least a proposed method should show is that under controlled, sometimes idealized conditions, it performs on par with similar methods (Dorie et al., 2019). Indeed, as Curth et al. (2021) advocate, we can see that among a class of similar neural network models our proposed methods perform well, and the addition of balancing terms improved performance; we might however set only limited stock in comparisons with methods which makes completely different assumptions about the data, such as Causal Forest (Wager and Athey, 2018). Finally, we believe it is of utmost importance for the field as a whole to produce a larger set of benchmarks that reflect the diversity of real-world observational studies, and that the recent ACIC challenge (Shimoni et al., 2018) is a good step in this direction.

Acknowledgments

Appendix A. Proof of Theorem 3

We prove Theorem 3, first repeating the original statement below. Recall that $\mathcal{O}_t(h_t, \Phi, w^t; \beta)$ is the weighted learning objective (22) and that $R(f_t)$ is the marginal risk for predicting a single potential outcome for the whole population, as in (4),

$$R(f_t) = \mathbb{E}_{X,Y}[L(Y(t), f_t(x))] .$$

Theorem 3 (Restated). Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) given by a bounded kernel k, such that for all $h_t \in \mathcal{G}, \Phi \in \mathcal{E}, h_t \circ \Phi \in \mathcal{H}$. Suppose weak overlap holds in that

$$\forall t \in \{0,1\} : \mathbb{E}_X[(p_t(X)/p_{1-t}(X))^2] < \infty .$$

Then, with $\mathcal{O}_t$ the objective defined in (22), with the IPM-space $\mathcal{L}$ also an RKHS with bounded kernel $k_{\mathcal{L}}$, and $n_t = \sum_{i=1}^n \mathbb{1}[t_i = t]$ for $t \in \{0,1\}$,

$$\min_{h_t, \Phi, w^t} \mathcal{O}_t(h_t, \Phi, w^t; \beta) \leq \min_{f_t \in \mathcal{H}} R(f_t) + O_p(1/\sqrt{n_0} + 1/\sqrt{n_1}) ,$$

where O_p denotes stochastic boundedness. Thus, under the assumptions of Thm. 2, for sufficiently large α and λ_w, with $\hat{f}^n$ the minimizer of (22) for n samples,

$$R(\hat{f}_t^n) \leq \min_{f_t \in \mathcal{H}} R(f_t) + O_p(1/n_0^{3/8} + 1/n_1^{3/8}).$$

In words, the minimizers of (22) converge to the representation and hypothesis that minimize the counterfactual risk, in the limit of infinite samples.

Proof. Let $f_t^* = \Phi^* \circ h_t^* \in \arg\min_{f_t \in \mathcal{H}} R(f_t)$ and let $w^{t,*}(x) := p_{\Phi,1-t}(\Phi^*(x))/p_{\Phi,t}(\Phi^*(x)) = p_{1-t}(x)/p_t(x)$ be the importance ratio between densities of treatment group $1-t$ and t. The second equality follows since Φ is assumed invertible. Since $\min_{h_t, \Phi, w^t} \mathcal{O}_t(h_t, \Phi, w^t; \beta) \leq \mathcal{O}_t(h_t^*, \Phi^*, w^{t,*}; \beta)$, it suffices to show that[4]

$$\mathcal{O}_t(h_t^*, \Phi^*, w^{t,*}; \beta) = R(f_t^*) + O_p(1/\sqrt{n_0} + 1/\sqrt{n_1}) ,$$

where O_p denotes stochastic boundedness, defined as follows. $X_n = O_p(a_n)$ holds, i.e., X_n/a_n is stochastically bounded, if there exist finite constants $M > 0, N > 0$, such that for every $\epsilon > 0$

$$p(|X_n/a_n| > M) < \epsilon, \forall n > N .$$

We will work term by term with the learning objective (22), with $\beta = (\alpha, \lambda_w, \lambda_h)$, $\pi_t = p(T = t)$, and $\tilde{w}_i^t = \pi_t + \pi_{1-t} w_i^t$,

$$\mathcal{O}_t(h_t, \Phi, w^t; \beta) = \underbrace{\sum_{i:t_i=t} \frac{\tilde{w}_i^t}{n_t} L(h_t(\Phi(x_i)), y_i)}_{\text{A}} + \underbrace{\frac{\mathcal{R}(h_t)}{\lambda_h^{-1}\sqrt{n_t}}}_{\text{B}} + \alpha \pi_{1-t} \underbrace{\text{IPM}_{\mathcal{L}}(\hat{p}_{\Phi,t}^{w^t}, \hat{p}_{\Phi,1-t})}_{\text{C}} + \underbrace{\frac{\|w^t\|_2}{\lambda_w^{-1} n_t}}_{\text{D}} .$$

---

4. Note the distinction between the objective $\mathcal{O}$ and the O of "Big O" notation.

for the case $h_t \circ \Phi = f^*, w^t = w^{t,*}$. For term $\boxed{D}$, letting $w_i^{t,*} = w^{t,*}(x_i)$, we have that by weak overlap

$$\boxed{D}^2 = \frac{1}{n_t} \times \frac{1}{n_t} \sum_{i:t_i=t} (w_i^{t,*})^2 = O_p(1/n_t),$$

so that $\boxed{D} = O_p(1/\sqrt{n})$.

For term $\boxed{A}$, under ignorability, each term in the sum in the first term has expectation equal to the population marginal risk $R(f_t^*)$ over $p(X)$, since $p_t(x_i)\tilde{w}_i^{t,*} = p_t(x_i)(\pi_t + \pi_{1-t}w_i^{t,*}) = p(X = x_i)$. As a result, by weak overlap and bounded second moments of loss, we have $\boxed{A} = R(f_t^*) + O_p(1/\sqrt{n})$. For term $\boxed{B}$, since h_t^* is fixed, we have deterministically that $\boxed{B} = O(1/\sqrt{n_t})$.

Finally, we address term $\boxed{C}$, which when expanded can be written as

$$\sup_{\|\ell\|_{\mathcal{L}} \leq 1} \left| \frac{1}{n_{1-t}} \sum_{i:t_i \neq t} \ell(\Phi^*(x_i)) - \frac{1}{n_t} \sum_{i:t_i=t} w_i^{t,*} \ell(\Phi^*(x_i)) \right|.$$

Let x_i' for $i = 1, \ldots, n_{1-t}$ and x_i'' for $i = 1, \ldots, n_t$ replicates of treatment group $1 - t$, i.e., new ghost samples, drawn from the treatment group opposite to t. Recall that $n = n_0 + n_1$. Without loss of generality, assume that $t = 1$ and that samples are ordered such that $t_i = 0$ for $i \in \{1, ..., n_0\}$ and $t_i = 1$ for $i \in \{n_0 + 1, ..., n\}$. Ghost samples are therefore drawn from $p_0(X)$. The case for $t = 0$ follows immediately. By Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}[\boxed{C}^2] &= \mathbb{E}\left[\sup_{\|\ell\|_{\mathcal{L}} \leq 1} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} \ell(\Phi^*(x_i)) - \frac{1}{n_1} \sum_{i=n_0+1}^{n} w_i^{1,*} \ell(\Phi^*(x_i)) \right)^2 \right] \\
&= \mathbb{E}\left[\sup_{\|\ell\|_{\mathcal{L}} \leq 1} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} (\ell(\Phi^*(x_i')) - \mathbb{E}[\ell(\Phi^*(x_i'))]) \right. \right. \\
&\qquad \left. \left. - \frac{1}{n_1} \sum_{i=n_0+1}^{n} (w_i^* \ell(\Phi^*(x_i)) - \mathbb{E}[\ell(\Phi^*(x_i''))]) \right)^2 \right] \\
&\leq \mathbb{E}\left[\sup_{\|\ell\|_{\mathcal{L}} \leq 1} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} (\ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i'))) \right. \right. \\
&\qquad \left. \left. - \frac{1}{n_1} \sum_{i=n_0+1}^{n} (w_i^{1,*} \ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i''))) \right)^2 \right] \\
&\leq 2\mathbb{E}\left[\sup_{\|\ell\|_{\mathcal{L}} \leq 1} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} (\ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i'))) \right)^2 \right] \\
&\qquad + 2\mathbb{E}\left[\sup_{\|\ell\|_{\mathcal{L}} \leq 1} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} (w_i^{1,*} \ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i''))) \right)^2 \right]
\end{aligned}
$$

Let $\xi_i(\ell) = \ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i'))$ and let $\zeta_i(h) = w_i^{1,*} \ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i''))$. Note that for every ℓ, $\mathbb{E}[\zeta_i(\ell)] = \mathbb{E}[\xi_i(\ell)] = 0$. Moreover, since the RKHS $\mathcal{L}$ has bounded kernel by

assumption,

$$\mathbb{E}[\|\zeta_i\|^2] \leq 4\mathbb{E}[k_\mathcal{L}(\Phi^*(x_i'), \Phi^*(x_i'))] \leq M$$

for some M. Similarly, $\mathbb{E}[\|\xi_i\|^2] \leq 2\mathbb{E}[(w_i^{1,*})^2]M + 2M \leq M' < \infty$ because of weak overlap. Let ζ_i' for $i = 1, \ldots, n$ be i.i.d. replicates of ζ_i (ghost sample) and let ϵ_i be i.i.d. Rademacher random variables. Because $\mathcal{L}$ is a Hilbert space, we have that $\sup_{\|\ell\|_\mathcal{L} \leq 1}(A(\ell))^2 = \|A\|^2 = \langle A, A \rangle$. Therefore, by Jensen's inequality,

$$\mathbb{E}\left[\sup_{\|\ell\|_\mathcal{L} \leq 1} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} (w_i^{1,*}\ell(\Phi^*(x_i)) - \ell(\Phi^*(x_i'')))\right)^2\right]$$

$$= \mathbb{E}\left[\sup_{\|\ell\|_\mathcal{L} \leq 1} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} \zeta_i(\ell)\right)^2\right]$$

$$= \mathbb{E}\left[\sup_{\|\ell\|_\mathcal{L} \leq 1} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} (\zeta_i(\ell) - \mathbb{E}[\zeta_i'(\ell)])\right)^2\right]$$

$$\leq \mathbb{E}\left[\sup_{\|\ell\|_\mathcal{L} \leq 1} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} (\zeta_i(\ell) - \zeta_i'(\ell))\right)^2\right]$$

$$= \mathbb{E}\left[\sup_{\|\ell\|_\mathcal{L} \leq 1} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} \epsilon_i(\zeta_i(\ell) - \zeta_i'(\ell))\right)^2\right]$$

$$\leq \frac{4}{n_1^2}\mathbb{E}\left[\sup_{\|\ell\|_\mathcal{L} \leq 1} \left(\sum_{i=n_0+1}^{n} \epsilon_i\zeta_i(\ell)\right)^2\right]$$

$$= \frac{4}{n_1^2}\mathbb{E}\left[\|\sum_{i=n_0}^{n} \epsilon_i\zeta_i\|^2\right]$$

$$= \frac{4}{n_1^2}\mathbb{E}\left[\sum_{i,j=n_0+1}^{n} \epsilon_i\epsilon_j \langle \zeta_i, \zeta_j \rangle\right]$$

$$= \frac{4}{n_1^2}\mathbb{E}[\sum_{i=n_0+1}^{n} \|\zeta_i\|^2]$$

$$= \frac{4}{n_1^2}\sum_{i=n_0+1}^{n} \mathbb{E}[\|\zeta_i\|^2]$$

$$\leq \frac{4M'}{n_1}$$

An analogous argument can be made of ξ_i's, showing that $\mathbb{E}[\text{Ⓒ}^2] = O(1/n)$ and hence $\text{Ⓒ} = O(1/\sqrt{n})$ by Markov's inequality. The final result follows from Theorem 2. $\qquad \square$

We may generalize Theorem 3 to bound the risk in predicting the outcome $Y = Y(T)$ in expectation over a treatment policy $p_\pi(T \mid X)$ based on observations from a policy $p_\mu(T \mid X)$. The risk in predicting a single fixed potential outcome t follows as a special (deterministic) case of $p_\pi(T = t \mid X) = 1$. With this in mind, let

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

Table 3: Hyperparameters and ranges.

$$R_\pi(f) = \mathbb{E}_X[\mathbb{E}_{T|X \sim p_\pi}[\ell_f(X, T)]] \ ,$$

where $\ell_f(x, t) = \mathbb{E}[L(f(x, t), Y(t)) \mid X = x, T = t]$. As previously, we consider hypotheses $f(x, t) = h(\Phi(x), t)$ for functions $h \in \mathcal{F}$ and embeddings $\Phi \in \mathcal{E} \subset \{\mathcal{X} \to \mathcal{Z}\}$ for some embedding space $\mathcal{Z} \subset \mathbb{R}^v$.

Corollary 3. *Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space given by a bounded kernel. Further suppose that weak overlap holds in that $\mathbb{E}[(p_\pi(x, t)/p_\mu(x, t))^2] < \infty$. Assume that n labeled samples $\{(x_i, t_i, y_i)\}_{i=1}^{n} \sim p_\mu$ and m unlabeled samples $\{(x_i, t_i)\}_{i=n+1}^{m+n} \sim p_\pi$ are available. Then,*

$$\min_{h, \Phi, w} \mathcal{O}_\pi(h, \Phi, w; \beta) \leq \min_{f \in \mathcal{F}} R_\pi(f) + O(1/\sqrt{n} + 1/\sqrt{m}) \ .$$

Corollary 3 may be proven using an anologous argument to that of Theorem 3.

Appendix B. Experiment Details

See Table 3 for a description of hyperparameters and search ranges for TARNet, CFR Wass and CFR MMD.

Appendix C. Architecture for Joint Learning of Sample Weights

For an illustration of the re-weighed CFR estimator, see Figure 10.

Appendix D. Minimization of Approximate Wasserstein Distances

Computing (and minimizing) the Wasserstein distance traditionally involves solving a linear program, which may be prohibitively expensive for many practical applications. Cuturi (2013) showed that introducing entropic regularization in the optimization problem results in an approximation computable through the Sinkhorn-Knopp matrix scaling algorithm, at orders of magnitude faster speed. The approximation, called Sinkhorn distances, is computed using a fixed-point iteration involving repeated multiplication with a kernel matrix K. We use the algorithm of Cuturi (2013) in our framework by differentiating through the iterations. See Algorithm 1 for an overview of how to compute the gradient g_1 in Algorithm 21. When computing g_1, disregarding the gradient $\nabla_{\mathbf{W}} T^*$ amounts to minimizing an
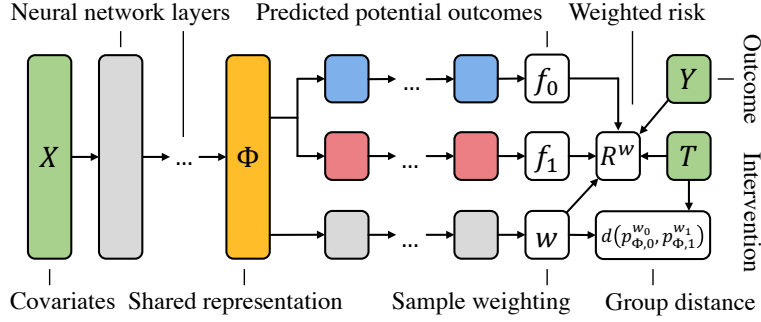
Figure 10: Illustration of the Re-weighted Counterfactual Regression (RCFR) estimator. Green boxes indicate inputs, white boxes outputs and loss terms, yellow boxes shared representations and blue/red boxes estimators of potential outcomes. Solid lines indicate transformation part of the pr ediction function and dashed lines indicate computations part of the learning procedure.

Algorithm 1 Computing the stochastic gradient of the Wasserstein distance

---

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

---

upper bound on the Sinkhorn transport. More advanced ideas for stochastic optimization of this distance have recently proposed by Genevay et al. (2016), and might be used in future work.

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

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

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

References

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 136–145. JMLR. org, 2017.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations.* Cambridge University Press, 2009.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Susan Athey. causalTree. `https://github.com/susanathey/causalTree`, 2016.

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

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

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

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Dimitris Bertsimas, Nathan Kallus, Alexander M Weinstein, and Ying Daisy Zhuo. Personalized diabetes management using electronic medical records. *Diabetes care*, 40(2): 210–217, 2017.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008.

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

M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0):S114, 2010.

Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

Peipei Chen, Wei Dong, Xudong Lu, Uzay Kaymak, Kunlun He, and Zhengxing Huang. Deep representation learning for individualized treatment effect estimation using electronic health records. *Journal of biomedical informatics*, 100:103303, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

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

Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pages 308–323. Springer, 2011.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

Alicia Curth and Mihaela Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Peng Ding, TJ VanderWeele, and JM Robins. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302, 2017.

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

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Rejoinder: Response to discussions and a look ahead. *Statistical Science*, 34(1):94–99, 2019.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Frederick Eberhardt. Causal discovery as a game. In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*, pages 87–96. JMLR. org, 2008.

David A Freedman and Richard A Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1 (10). Springer series in statistics New York, 2001.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning*, pages 1917–1925, 2015.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. Pac-bayes and domain adaptation. *Neurocomputing*, 379:379–397, 2020.

Anna C Gilbert, Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. Towards understanding the invertibility of convolutional neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1703–1710, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. In *The annual summer meeting of the society of political methodology*, 2010.

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

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Susan Gruber and Mark van der Laan. tmle: An r package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(1):1–35, 2012.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

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

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

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

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, volume 139, pages 4829–4838, 2021.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 527–536. PMLR, 16–18 Apr 2019.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10186–10193, 2020a.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in neural information processing systems*, 33, 2020b.

Nathan Kallus. A framework for optimal matching for causal inference. In *Artificial Intelligence and Statistics*, pages 372–381, 2017.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020a.

Nathan Kallus. Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.*, 21:62–1, 2020b.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in neural information processing systems*, 2018.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in neural information processing systems*, 2018.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Damien Lefortier, Adith Swaminathan, Xiaotao Gu, Thorsten Joachims, and Maarten de Rijke. Large-scale validation of counterfactual learning methods: A test-bed. *arXiv preprint arXiv:1612.00367*, 2016.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Y Liu, O Gottesman, A Raghu, M Komorowski, A Faisal, F Doshi-Velez, and E Brunskill. Representation balancing MDPs for off-policy policy evaluation. *Advances in neural information processing systems*, 2018.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada, 2009.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

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

Judea Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

D Pollard. *Convergence of Stochastic Processes*. David Pollard, 1984.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Paul R Rosenbaum. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer, 2002.

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

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4):512, 2009.

Uri Shalit, Fredrik Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Ricardo Silva, Richard Scheine, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

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

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521. Morgan Kaufmann Publishers Inc., 2001.

V Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

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

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020.