Learning Weighted Representations for Generalization Across Designs

Fredrik D. Johansson [1] **Nathan Kallus** [2] **Uri Shalit** [3] **David Sontag** [1]

Abstract

Predictive models that generalize well under distributional shift are often desirable and sometimes crucial to building robust and reliable machine learning applications. We focus on distributional shift that arises in causal inference from observational data and in unsupervised domain adaptation. We pose both of these problems as prediction under a shift in *design*. Popular methods for overcoming distributional shift make unrealistic assumptions such as having a well-specified model or knowing the policy that gave rise to the observed data. Other methods are hindered by their need for a pre-specified metric for comparing observations, or by poor asymptotic properties. We devise a bound on the generalization error under design shift, incorporating both representation learning and sample re-weighting. Based on the bound, we propose an algorithmic framework that does not require any of the above assumptions and which is asymptotically consistent. We empirically study the new framework using two synthetic datasets, and demonstrate its effectiveness compared to previous methods.

1. Introduction

A long-term goal in artificial intelligence is for agents to learn how to *act*. This endeavor relies on accurately predicting and optimizing for the outcomes of actions, and fundamentally involves estimating counterfactuals—what would have happened if the agent acted differently? In many applications, such as the treatment of patients in hospitals, experimentation is infeasible or impractical, and we are forced to learn from *biased*, observational data. Doing so requires adjusting for the *distributional shift* that exists between groups of patients that received different treatments. A related kind of distributional shift arises in unsupervised domain adaptation, the goal of which is to learn predictive models for a target domain, observing ground truth only in a source domain.

In this work, we pose both domain adaptation and treatment effect estimation as special cases of prediction across shifting *designs*, referring to changes in both action policy and feature domain. We separate policy from domain as we wish to make *causal* statements about the policy, but not about the domain. For example, to learn a treatment policy from observational data, personalizing the choice between medication A and B, one must adjust for the fact that treatment A was systematically given to patients of different characteristics from those who received treatment B. We call this predicting *under a shift in policy*. Furthermore, if all of our observational data comes from hospital P, but we wish to predict counterfactuals for patients in hospital Q, with a population that differs from P, an additional source of distributional shift is at play. We call this *a shift in domain*. Together, we refer to the combination of domain and policy as the *design*. The design for which we observe ground truth is called the *source*, and the design of interest the *target*.

The two most common approaches for addressing distributional shift are to learn shift-invariant representations of the data (Ajakan et al., 2014) or to perform sample re-weighting or matching (Shimodaira, 2000; Kallus, 2016). Representation learning approaches attempt to extract only information from the input that is invariant to a change in design *and* predictive of the variable of interest. Such representations are typically learned by fitting deep neural networks in which activations of deeper layers are regularized to be distributionally similar across designs (Ajakan et al., 2014; Long et al., 2015). Although representation learning can be shown to reduce the error associated to distributional shift (Long et al., 2015) in some cases, standard approaches are biased, even in the limit of infinite data, as they also penalize the use of predictive information. In contrast, re-weighting methods correct for distributional shift by assigning higher weight to samples from the source design that are representative of the target design, often using importance sampling. This idea has been well studied in, for example, causal inference (Rosenbaum & Rubin, 1983), domain adaptation (Shimodaira, 2000) and reinforcement learning (Precup et al., 2001). For example, in causal effect estimation, importance sampling is equivalent to re-weighting units by the inverse probability of observed treatments (treatment propensity). Re-weighting with knowledge of importance sampling weights often leads to asymptotically unbiased

[1] MIT [2] Cornell Tech [3] Technion. Correspondence to: Fredrik D. Johansson <fredrikj@mit.edu>.

estimators of the target outcome, but may suffer from high variance in finite samples (Swaminathan & Joachims, 2015).

A significant hurdle in applying re-weighting methods is that optimal weights are rarely known in practice. Weights can be estimated as the inverse of estimated feature or treatment densities (Rosenbaum & Rubin, 1983; Freedman & Berk, 2008) but this plug-in approach can lead to highly unstable estimates. More stable methods learn weights by minimizing distributional distance metrics (Gretton et al., 2009; Kallus, 2016; 2017; Zubizarreta, 2015). Closely related, matching (Stuart, 2010) produces weights by finding units in the source design that are similar in some metric to units in the target design. Specifying a distributional or unit-wise metric is challenging, especially if the input space is high-dimensional where no metric incorporating all features can can also be made small though weighting. This has inspired heuristics such as first performing variable selection and then balancing or finding matches only in the selected covariates.

In this work, we bring together shift-invariant representation learning and re-weighting methods. We show that existing representation learning approaches minimize an upper bound on the generalization under design-shift, implicitly using uniform sample weights, and that there exist weights that improve the tightness of these bounds. Our key algorithmic contribution is to jointly learn a representation Φ of the input space and a weighting function $w(\Phi)$ to minimize a) the re-weighted empirical risk and b) a re-weighted measure of distributional shift between designs. This is useful also for the identity representation $\Phi(x) = x$, as it allows for principled control of the variance of estimators through regularization of the re-weighting function $w(x)$, mitigating the issues of exact importance sampling methods. Further, this allows us to evaluate w on hold-out samples to select hyperparameters or do early stopping. Finally, letting w depend on Φ alleviates the problem of choosing a metric by which to optimize sample weights, as Φ is trained to extract information predictive of the outcome. We apply our theory and algorithmic framework for generalization error under a shift in design to the case of treatment effect estimation.

Main contributions We bring together two techniques used to overcome distributional shift between designs—re-weighting and representation learning, with complementary robustness properties, generalizing existing methods based on either technique. We give finite-sample generalization bounds for prediction under design shift, *without* assuming access to importance sampling weights *or* to a well-specified model, and develop an algorithmic framework to minimize these bounds. We propose a neural network architecture that jointly learns a representation of the input and a weighting function to improve balance across changing settings. Finally, we apply our proposed algorithm to the task of predicting causal effects from observational data, achieving state-of-the art results on a widely used benchmark.

2. Predicting outcomes under design shift

The goal of this work is to accurately predict outcomes of interventions $T \in \mathcal{T}$ in contexts $X \in \mathcal{X}$ drawn from a *target design* $p_\pi(X, T)$. The result of intervening with $t \in \mathcal{T}$ is the potential outcome $Y(t) \in \mathcal{Y}$ (Imbens & Rubin, 2015, Ch. 1–2), which has a stationary distribution $p(Y(t) \mid X)$ given context X. Assuming a stationary outcome is akin to the *covariate shift* assumption (Shimodaira, 2000), often used in domain adaptation.[1] For example, in the binary intervention setting, $Y(1)$ represents the outcome under treatment and $Y(0)$ the outcome under control. The target design consists of two components: the *target policy* $p_\pi(T \mid X)$, which describes how one intends to map observations of contexts (such as patient prognostics) to interventions (such as pharmacological treatments) and the *target domain* $p_\pi(X)$, which describes the population of contexts to which the policy will be applied. The target design is known to us only through m unlabeled samples $(x'_1, t'_1), \ldots, (x'_m, t'_m)$ from $p_\pi(X, T)$. Outcomes are only available to us in labeled samples from a source domain: $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$, where (x_i, t_i) are draws from a source design $p_\mu(X, T)$ and $y_i = y_i(t_i)$ is a draw from $p_T(Y \mid X)$, corresponding only to the *factual* outcome $Y(T)$ of the treatment administered. Like the target design, the source design consists of a domain of contexts for which we have data and a policy, which describes the (unknown) historical administration of treatment in the data. Only the factual outcomes of the treatments administered are observed, while the counterfactual outcomes $y_i(t)$ for $t \neq t_i$ are, naturally, unobserved.

Our focus is the *observational* or *off-policy* setting, in which interventions in the source design are performed dependent on attributes X, $p_\mu(T \mid X) \neq p_\mu(T)$, and that covariate marginals are shifted in general, $p_\mu(X) \neq p_\pi(X)$. This encapsulates both the covariate shift often observed between treated and control populations in observational studies and the covariate shift between the domain of the study and the domain of an eventual wider intervention. Examples of this problem are plentiful: in addition to the example given in the introduction, consider predicting the return of an advertising policy based on the historical results of a different policy, applied to a different population of customers. We stress that we are interested in the *causal effect* of an intervention T on Y, conditioned on X. As such, we cannot think of X and T as a single variable. Without additional assumptions, it is impossible to deduce the effect of an intervention based on observational data alone (Pearl, 2009), as it amounts disentangling correlation and causation. Crucially, for any

[1] Equivalently, we may write $p_\pi(Y(t) \mid X) = p_\mu(Y(t) \mid X)$.

unit i, we can observe the potential outcome $y_i(t)$ of at most one intervention t. In our analysis, we make the following standard assumptions.

Assumption 1 (Consistency, ignorability and overlap). *For any unit i, assigned to intervention t_i, we observe $Y_i = Y(t_i)$. Further, $\{Y(t)\}_{t\in\mathcal{T}}$ and the data-generating process $p_\mu(X, T, Y)$ satisfy strong ignorability: $\{Y(t)\}_{t\in\mathcal{T}} \perp\!\!\!\perp T \mid X$ and overlap: $\Pr_{p_\pi}(p_\mu(T \mid X) > 0) = 1$.*

Assumption 1 is a sufficient condition for *causal identifiability* (Rosenbaum & Rubin, 1983). Ignorability is also known as the *no hidden confounders* assumption, indicating that all variables that cause both T and Y are assumed to be measured. Under ignorability therefore, any domain shift in $p(X)$ cannot be due to variables that causally influence T and Y, other than through X. Under Assumption 1, potential outcomes equal conditional expectations: $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid X = x, T = t]$, and we may predict $Y(t)$ by regression. We further assume common domain support, $\forall x \in \mathcal{X} : p_\pi(X = x) > 0 \Rightarrow p_\mu(X = x) > 0$. Finally, we adopt the notation $p(x) := p(X = x)$.

2.1. Re-weighted risk minimization

We attempt to learn predictors $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ such that $f(x, t)$ approximates $\mathbb{E}[Y \mid X = x, T = t]$. Recall that under Assumption 1, this conditional expectation is equal to the (possibly counterfactual) potential outcome $Y(t)$, conditioned on X. Our goal is to ensure that hypotheses f are accurate under a design p_π that deviates from the data-generating process, p_μ. This is unlike standard supervised learning for which $p_\pi = p_\mu$. We measure the (in)ability of f to predict outcomes under π, using the expected risk,

$$R_\pi(f) := \mathbb{E}_{x,t,y\sim p_\pi}[\ell_f(x, t, y)] \tag{1}$$

based on a sample from μ, $D^n_\mu = \{(x_i, t_i, y_i) \sim p_\mu; i = 1, ..., n\}$. Here, $\ell_f(x, t, y) := L(f(x, t), y)$ is an appropriate loss function, such as the squared loss, $L(y, y') := (y - y')^2$ or the log-loss, depending on application. As outcomes under the target design p_π are not observed, even through a Monte Carlo sample, we cannot directly estimate (1) using the empirical risk under p_π. A common way to solve this is to use *importance sampling* (Shimodaira, 2000)—the observation that if p_μ and p_π have common support, with $w^*(x, t) = p_\pi(x, t)/p_\mu(x, t)$,

$$R^{w^*}_\mu(f) := \mathbb{E}_{x,t,y\sim p_\mu}[w^*(x, t)\ell_f(x, t, y)] = R_\pi(f) \ . \tag{2}$$

Hence, with access to w^*, an unbiased estimator of $R_\pi(f)$ may be obtained by re-weighting the (factual) empirical risk under μ,

$$\hat{R}^{w^*}_\mu(f) := \frac{1}{n}\sum_{i=1}^{n} w^*(x_i, t_i)\ell_f(x_i, t_i, y_i) \ . \tag{3}$$

Unfortunately, importance sampling weights can be very large when p_π is large and p_μ small, resulting in large variance in $\hat{R}^{w^*}_\mu(f)$ (Swaminathan & Joachims, 2015). More importantly, $p_\mu(x, t)$ is rarely known in practice, and neither is w^*. In principle, however, any re-weighting function w with the following property yields a valid risk under the re-weighted distribution p^w_μ.

Definition 1. *A function $w : \mathcal{X} \times \mathcal{T} \to \mathbb{R}_+$ is a valid re-weighting of p_μ if*

$$\mathbb{E}_{x,t\sim p_\mu}[w(x, t)] = 1 \ \text{ and } \ p_\mu(x, t) > 0 \Rightarrow w(x, t) > 0.$$

We denote the re-weighted density $p^w_\mu(x, t) := w(x, t)p_\mu(x, t)$.

A natural candidate in place of w^* is an estimate $\hat{w}^*$ formed by estimating densities $p_\pi(x, t)$ and $p_\mu(x, t)$. In this work, we adopt a different strategy, learning parameteric re-weighting functions w from observational data, that minimize an upper bound on the risk under p_π.

2.2. Conditional treatment effect estimation

An important special case of our setting is when treatments are binary, $T \in \{0, 1\}$, often interpreted as treating ($T = 1$) or not treating ($T = 0$) a unit, and the domain is fixed across designs, $p_\mu(X) = p_\pi(X)$. This is the classical setting for estimating treatment effects—the effect of choosing one intervention over another (Morgan & Winship, 2014).[2] The effect of an intervention $T = 1$ in context X, is measured by the *conditional average treatment effect* (CATE), $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$. Predicting τ for unobserved units typically involves prediction of both potential outcomes[3]. In a clinical setting, knowledge of τ is necessary to assess which medication should be administered to a certain individual. Historically, the (population) *average treatment effect*, ATE $= \mathbb{E}_{x\sim p}[\tau(x)]$, has received comparatively much more attention (Rosenbaum & Rubin, 1983), but is inadequate for personalized decision making. Using predictors $f(x, t)$ of potential outcomes $Y(t)$ in contexts $X = x$, we can estimate the CATE by $\hat{\tau}(x) = f(x, 1) - f(x, 0)$ and measure the quality using the mean squared error (MSE),

$$\text{MSE}(\hat{\tau}) = \mathbb{E}_p\left[(\hat{\tau}(x) - \tau(x))^2\right] \tag{4}$$

In Section 4, we argue that estimating CATE from observational data requires overcoming distributional shift with respect to the treat-all and treat-none policies, in predicting each respective potential outcome, and show how this can be used to derive generalization bounds for CATE.

[2] Effects for non-binary interventions are not considered here.
[3] This is sufficient but not necessary.

3. Related work

A large body of work has shown that under assumptions of ignorability and having a well-specified model, various regression methods for counterfactual estimation are asymptotically consistent (Chernozhukov et al., 2017; Athey & Imbens, 2016; Belloni et al., 2014). However, consistency results like these provide little insight into the case of model misspecification. Under model misspecification, regression methods may suffer from additional bias when generalizing across designs due to distributional shift. A common way to alleviate this is *importance sampling*, see Section 2. This idea is used in propensity-score methods (Austin, 2011), that use the observed treatment policy to re-weight samples for causal effect estimation, and more generally in re-weighted regression, see e.g. (Swaminathan & Joachims, 2015). A major drawback of these methods is the assumption that the design density is known. To address this, others (Gretton et al., 2009; Kallus, 2016), have proposed *learning* sample weights w to minimize a distributional distance between samples under p_π and p_μ^w, but rely on specifying the data representation a priori, without regard for which aspects of the data matter for outcome prediction.

On the other hand, Johansson et al. (2016); Shalit et al. (2017) proposed *learning representations* for counterfactual inference, inspired by work in unsupervised domain adaptation (Mansour et al., 2009). The drawback of this line of work is that the generalization bounds of Shalit et al. (2017) and Long et al. (2015) are loose—even if infinite samples are available, they are not guaranteed to converge to the lowest possible error. Moreover, these approaches do not make use of important information that can be estimated from data: the treatment/domain assignment probabilities.

4. Generalization under design shift

We give a bound on the risk in predicting outcomes under a target design $p_\pi(T, X)$ based on unlabeled samples from p_π and labeled samples from a source design $p_\mu(T, X)$. Our result combines representation learning, distribution matching and re-weighting, resulting in a tighter bound than the closest related work, Shalit et al. (2017). The predictors we consider are compositions $f(x, t) = h(\Phi(x), t)$ where Φ is a representation of x and h an hypothesis. We first give an upper bound on the risk in the general design shift setting, then show how this result can be used to bound the error in prediction of treatment effects. In Section 5 we give a result about the asymptotic properties of the minimizers of this upper bound.

Risk under distributional shift Our bounds on the risk under a target design capture the intuition that if either a) the target design π and source design μ are close, or b) the true outcome is a simple function of x and t, the gap between the target risk and the re-weighted source risk is small. These notions can be formalized using integral probability metrics (IPM) (Sriperumbudur et al., 2009) that measure distance between distributions w.r.t. a normed vector space of functions $\mathcal{H}$.

Definition 2. *The integral probability metric (IPM) distance, associated with a normed vector space of functions $\mathcal{H}$, between distributions p and q is,* $\mathrm{IPM}_\mathcal{H}(p, q) := \sup_{h\in\mathcal{H}} |\mathbb{E}_p[h] - \mathbb{E}_q[h]|$.

Important examples of IPMs include the Wasserstein distance, for which $\mathcal{H}$ is the family of functions with Lipschitz constant at most 1, and the Maximum Mean Discrepancy for which $\mathcal{H}$ are functions in the norm-1 ball in a reproducing kernel Hilbert space. Using definitions 1–2, and the definition of re-weighted risk, see (2), we can state the following result (see the Appendix for a proof).

Lemma 1. *For hypotheses f with loss ℓ_f such that $\ell_f/\|\ell_f\|_\mathcal{H} \in \mathcal{H}$, and p_μ, p_π with common support, there exists a valid re-weighting w, see Definition 1, such that,*

$$\begin{aligned} R_\pi(f) &\le R_\mu^w(f) + \|\ell_f\|_\mathcal{H}\mathrm{IPM}_\mathcal{H}(p_\pi, p_\mu^w) \\ &\le R_\mu(f) + \|\ell_f\|_\mathcal{H}\mathrm{IPM}_\mathcal{H}(p_\pi, p_\mu)\ . \end{aligned} \tag{5}$$

The first inequality is tight for importance sampling weights, $w(x, t) = p_\pi(x, t)/p_\mu(x, t)$. The second inequality is not tight for general f, even if $\ell_f/\|\ell_f\|_\mathcal{H} \in \mathcal{H}$, unless $p_\pi = p_\mu$.

The bound of Lemma 1 is tighter if p_μ and p_π are close (the IPM is smaller), and if the loss lives in a small family of functions $\mathcal{H}$ (the supremum is taken over a smaller set). Lemma 1 also implies that there exist weighting functions $w(x, t)$ that achieve a tighter bound than the uniform weighting $w(x, t) = 1$, implicitly used by Shalit et al. (2017). While importance sampling weights result in a tight bound in expectation, neither the design densities nor their ratio are known in general. Moreover, exact importance weights often result in large variance in finite samples (Cortes et al., 2010). Here, we will search for a weighting function w, that minimizes a finite-sample version of (5), trading off bias and variance. We examine the empirical value of this idea alone in Section 6.1.

Representation learning The idea of learning representations that reduce distributional shift in the induced space, and thus the source-target generalization gap, has been applied in domain adaptation (Ajakan et al., 2014), algorithmic fairness (Zemel et al., 2013) and counterfactual prediction (Shalit et al., 2017). The hope of these approaches is to learn predictors that predominantly exploit information that is common to both source and target distributions. For example, a face detector should be able to recognize the structure of human features even under highly variable environment conditions, by ignoring background, lighting etc. We argue that re-weighting (e.g. importance sampling) should also be done only with respect to features that are predictive of the outcome. Hence, in Section 5, we propose using re-weightings that are functions of learned representations.

We follow the setup of Shalit et al. (2017), and consider learning twice-differentiable, invertible representations $\Phi : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{Z}$ is the representation space, and $\Psi : \mathcal{Z} \rightarrow \mathcal{X}$ is the *inverse* representation, such that $\Psi(\Phi(x)) = x$ for all x. Let $\mathcal{E}$ denote space of such representation functions. For a design π, we let $p_{\pi,\Phi}(z,t)$ be the distribution induced by Φ over $\mathcal{Z} \times \mathcal{T}$, with $p^w_{\pi,\Phi}(z,t) := p_{\pi,\Phi}(z,t)w(\Psi(z),t)$ its re-weighted form and $\hat{p}^w_{\pi,\Phi}$ its re-weighted empirical form, following our previous notation. Finally, we let $\mathcal{G} \subseteq \{h : \mathcal{Z} \times \mathcal{T} \rightarrow \mathcal{Y}\}$ denote a set of hypotheses $h(\Phi, t)$ operating on the representation Φ and let $\mathcal{F}$ the space of all compositions, $\mathcal{F} = \{f = h(\Phi(x), t) : h \in \mathcal{G}, \Phi \in \mathcal{E}\}$. We can now relate the expected target risk $R_\pi(f)$ to the re-weighted empirical source risk $\hat{R}^w_\mu(f)$.

Theorem 1. *Given is a labeled sample $(x_1, t_1, y_1), ..., (x_n, t_n, y_n)$ from p_μ, and an unlabeled sample $(x'_1, t'_1), ..., (x'_m, t'_m)$ from p_π, with empirical measures $\hat{p}_\mu$ and $\hat{p}_\pi$. Suppose that Φ is a twice-differentiable, invertible representation, that $h(\Phi, t)$ is an hypothesis, and $f = h(\Phi(x), t) \in \mathcal{F}$. Define $m_t(x) = \mathbb{E}_Y[Y \mid X = x, T = t]$, let $\ell_{h,\Phi}(\Psi(z), t) := L(h(z,t), m_t(\Psi(z)))$ where L is the squared loss, $L(y, y') = (y - y')^2$, and assume that there exists a constant $B_\Phi > 0$ such that $\ell_{h,\Phi}/B_\Phi \in \mathcal{H} \subseteq \{h : \mathcal{Z} \times \mathcal{T} \rightarrow \mathcal{Y}\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space of a kernel, k such that $k((z,t),(z,t)) < \infty$. Finally, let w be a valid re-weighting of $p_{\mu,\Phi}$. Then with probability at least $1 - 2\delta$,*

$$\begin{aligned} R_\pi(f) \leq{}& \hat{R}^w_\mu(f) + B_\Phi \mathrm{IPM}_{\mathcal{H}}(\hat{p}_{\pi,\Phi}, \hat{p}^w_{\mu,\Phi}) \\ &+ V_\mu(w, \ell_f)\frac{\mathcal{C}^{\mathcal{F}}_{n,\delta}}{n^{3/8}} + \mathcal{D}^{\Phi,\mathcal{H}}_\delta\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) + \sigma^2_Y \end{aligned} \tag{6}$$

where $\mathcal{C}^{\mathcal{F}}_{n,\delta}$ is a function of the pseudo-dimension of $\mathcal{F}$, $\mathcal{D}^{\mathcal{H}}_{m,n,\delta}$ a function of the kernel norm of $\mathcal{H}$, both only with logarithmic dependence on n and m, σ^2_Y is the expected variance in Y, and

$$V_\mu(w, \ell_f) = \max\left(\sqrt{\mathbb{E}_{p_\mu}[w^2\ell_f^2]}, \sqrt{\mathbb{E}_{\hat{p}_\mu}[w^2\ell_f^2]}\right) .$$

A similar bound exists where $\mathcal{H}$ is the family of functions Lipschitz constant at most 1, and $\mathrm{IPM}_{\mathcal{H}}$ the Wasserstein distance, but with worse sample complexity.

See the Appendix for a proof of Theorem 1 that involves applying finite-sample generalization bounds to Lemma 1, as well and a change of variables to the space induced by the representation Φ.

Theorem 1 has several implications: non-identity feature representations, non-uniform sample weights, and variance control of these weights can all contribute to a lower bound. Using uniform weights $w(x,t) = 1$ in (6), results in a bound similar to that of Shalit et al. (2017) and Long et al. (2015). When $\pi \neq \mu$, minimizing uniform-weight bounds results in biased hypotheses, even in the asymptotical limit, as the IPM term does not vanish with increased sample size. This is an undesirable property, as even k-nearest-neighbor classifiers are consistent in the limit of infinite samples. We consider minimizing (6) with respect to w, improving the tightness of the bound. Further, Theorem 1 indicates that even though importance sampling weights w^* yield estimators with small bias, they can suffer from high variance, as captured by the factor $V_\mu(w, \ell_f)$.

The factor B_Φ in (6) is not known in general as it depends on the true outcome, and is determined by $\|\ell_f\|_{\mathcal{H}}$ as well as the determinant of the Jacobian of Ψ, see the Appendix for proofs. Qualitatively, B_Φ measures the joint complexity of Φ and ℓ_f and is sensitive to the scale of Φ—as the scale of Φ vanishes, B_Φ blows up. To prevent this in practice, we normalize Φ. As B_Φ is unknown, Shalit et al. (2017) substituted a hyperparameter α for B_Φ, but discussed the difficulties of selecting its value without access to counterfactual labels. In our experiments, we explore a heuristic for adaptively choosing α, based on measures of complexity of the observed held-out loss as a function of the input. Finally, the term $\mathcal{C}^{\mathcal{F}}_{n,\delta}$ follows from standard learning theory results (Cortes et al., 2010) and $\mathcal{F}$, and $\mathcal{D}^{\Phi,\mathcal{H}}_\delta$ from concentration results for estimating IPMs (Sriperumbudur et al., 2012), see the Appendix.

Theorem 1 is immediately applicable to the case of unsupervised domain adaptation in which there is only a single potential outcome of interest, $\mathcal{T} = \{0\}$. In this case, $p_\mu(T \mid X) = p_\pi(T \mid X) = 1$. Another important special case is where $p_\mu(X) = p_\pi(X)$, such as in the classical setting of causal effect estimation.

Conditional average treatment effects A simple argument shows that the error in predicting the conditional average treatment effect, $\mathrm{MSE}(\hat{\tau})$ can be bounded by the sum of risks under the constant treat-all and treat-none policies. As in Section 2.2, we consider the case of a fixed domain $p_\pi(X) = p_\mu(X)$ and binary treatment $\mathcal{T} = \{0, 1\}$. Let $R_{\pi_t}(f)$ denote the risk under the constant policy π_t such that $\forall x \in \mathcal{X} : p_{\pi_t}(T = t \mid X = x) = 1$.

Proposition 1. *We have with $\mathrm{MSE}(\hat{\tau})$ as in (4) and $R_{\pi_t}(f)$ the risk under the constant policy π_t,*

$$\mathrm{MSE}(\hat{\tau}) \leq 2(R_{\pi_1}(f) + R_{\pi_0}(f)) - 4\sigma^2 \tag{7}$$

where σ is such that $\forall t \in \mathcal{T}, x \in \mathcal{X}, \sigma_Y(x,t) \geq \sigma$ and $\sigma^2_Y(x,t)$ is variance of $Y(t)$ conditioned on $X = x$.

The proof involves the relaxed triangle inequality and the law of total probability. By Proposition 1, we can apply Theorem 1 to R_{π_1} and R_{π_0} separately, to obtain a bound on $\mathrm{MSE}(\tau)$. For brevity, we refrain from stating the full result, but emphasize that it follows from Theorem 1. In Section 6.2, we evaluate our framework in treatment effect estimation, minimizing this bound.

5. Joint learning of representations and sample weights

Motivated by the theoretical insights of Section 4, we propose to jointly learn a representation $\Phi(x)$, a re-weighting $w(x, t)$ and an hypothesis $h(\Phi, t)$ by minimizing a bound on the risk under the target design, see (6). This approach improves on previous work in two ways: it alleviates the bias of Shalit et al. (2017) when sample sizes are large, see Section 4, and it increases the flexibility of the balancing method of (Gretton et al., 2009) by learning the representation to balance.

For notational brevity, we let $w_i = w(x_i, t_i)$. Recall that $\hat{p}^w_{\pi,\Phi}$ is the re-weighted empirical distribution of representations Φ under p_π. The training objective of our algorithm is the RHS of (6), with hyperparameters $\beta = (\alpha, \lambda_h, \lambda_w)$ substituted for model (and representation) complexity terms,

$$\begin{aligned}\mathcal{L}_\pi(h, \Phi, w; \beta) = & \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_i \ell_h(\Phi(x_i), t_i) + \frac{\lambda_h}{\sqrt{n}}\mathcal{R}(h)}_{\mathcal{L}^h_\pi(h,\Phi,w;D,\alpha,\lambda_h)} \\ & + \underbrace{\alpha \, \mathrm{IPM}_G(\hat{p}_{\pi,\Phi}, \hat{p}^w_{\mu,\Phi}) + \lambda_w \frac{\|w\|_2}{n}}_{\mathcal{L}^w_\pi(\Phi,w;D,\alpha,\lambda_w)} \end{aligned} \tag{8}$$

where $\mathcal{R}(h)$ is a regularizer of h, such as ℓ_2-regularization. We can show the following result.

Theorem 2. *Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space given by a bounded kernel. Suppose weak overlap holds in that $\mathbb{E}[(p_\pi(x,t)/p_\mu(x,t))^2] < \infty$. Then,*

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

Consequently, under the assumptions of Thm. 1, for sufficiently large α and λ_w,

$$R_\pi(\hat{f}_n) \leq \min_{f \in \mathcal{F}} R_\pi(f) + O_p(1/n^{3/8} + 1/\sqrt{m}).$$

In words, the minimizers of (8) converge to the representation and hypothesis that minimize the counterfactual risk, in the limit of infinite samples.

Implementation Minimization of $\mathcal{L}_\pi(h, \Phi, w; \beta)$ over h, Φ and w is, while motivated by Theorem 2, a difficult optimization problem to solve in practice. For example, adjusting w to minimize the empirical risk term may result in overemphasizing "easy" training examples, resulting in a poor local minimum. Perhaps more importantly, ensuring invertibility of Φ while maintaining good accuracy is nontrivial for many representation learning frameworks, such as deep neural networks. In our implementation, we deviate from theory on these points, by fitting the re-weighting w based only on imbalance and variance terms, and don't explicitly enforce invertibility. As a heuristic, we split the objective, see (8), in two and use only the IPM term and regularizer to learn w. In short, we adopt the following alternating procedure.

$$h^k, \Phi^k = \operatorname*{arg\,min}_{h,\Phi} \;\; \mathcal{L}^h_\pi(h, \Phi, w; D, \alpha, \lambda_h), \tag{9}$$

$$w^{k+1} = \operatorname*{arg\,min}_{w} \;\; \mathcal{L}^w_\pi(\Phi^k, w; D, \alpha, \lambda_w) \tag{10}$$

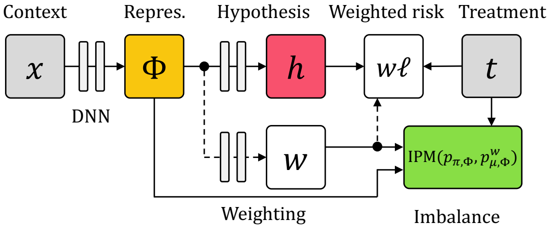

Figure 1: Architecture for predicting outcomes under design shift. A re-weighting function w is fit jointly with a representation Φ and hypothesis h of the potential outcomes, minimizing a bound on the target risk. Dashed lines are not back-propagated through. Regularization not shown.

The re-weighting function $w(x, t)$ could be represented by one free parameter per training point, as it is only used to learn the model, not for prediction. However, we propose to let w be a parametric function of $\Phi(x)$. Doing so ensures that information predictive of the outcome is used for balancing, and lets us compute weights and the objective on a hold-out set, to perform early stopping or select hyperparameters. This is not possible with existing re-weighting methods such as Gretton et al. (2009); Kallus (2016). An example architecture for the treatment effect estimation setting is presented in Figure 1. By Proposition 1, estimating treatment effects involves predicting under the two constant policies—treat-everyone and treat-no-one. In Section 6, we evaluate our method in this task.

As noted by Shalit et al. (2017), choosing hyperparameters for counterfactual prediction is fundamentally difficult, as we cannot observe ground truth for counterfactuals. In this work, we explore setting the balance parameter α adaptively. α is used in (8) in place of B_Φ, a factor measuring the complexity of the loss and representation function as functions of the input, a quantity that changes during training. As a heuristic, we use an approximation of the Lipschitz constant of ℓ_f, with $f = h(\Phi(x), t)$, based on observed examples: $\alpha_{h,\Phi} = \max_{i,j\in[n]} |\ell_f(x_i, t_i, y_i) - \ell_f(x_j, t_j, y_j)|/\|x_i - x_j\|_2$. We use a moving average to improve stability.

6. Experiments

6.1. Synthetic experiments for domain adaptation

We create a synthetic domain adaptation experiment to highlight the benefit of using a learned re-weighting function to minimize weighted risk over using importance sampling weights $w^*(x) = p_\pi(x)/p_\mu(x)$ for small sample sizes. We observe n labeled source samples, distributed according to $p_\mu(x) = \mathcal{N}(x; m_\mu, I_d)$ and predict for n unlabeled target samples drawn according to $p_\pi(x) = \mathcal{N}(x; m_\pi, I_d)$ where I_d is the d-dimensional identity matrix, $m_\mu = \mathbf{1}_d/2$, $m_\pi = -\mathbf{1}_d/2$ and $\mathbf{1}_d$ is the d-dimensional vector of all 1:s, here with $d = 10$. We let $\beta \sim \mathcal{N}(\mathbf{0}_d, 1.5I_d)$ and $c \sim \mathcal{N}(0, 1)$ and let $y = \sigma(\beta^\top x + c)$ where $\sigma(z) = 1/(1 + e^{-z})$. Importance sampling weights, $w^*(x) = p_\pi(x)/p_\mu(x)$, are known. In experiments, we vary n from 10 to 600. We fit (misspecified) linear models—the identity representation $\Phi(x) = x$ is used for both approaches—$f(x) = \beta^\top x + \gamma$ to the logistic outcome, and compare minimizing a weighted source risk by a) parameterizing sample weights as a small feed-forward neural network to minimize (8) (ours) b) using importance sampling weights (baseline), both using gradient descent. For our method, we add a small variance penalty, $\lambda_w = 10^{-3}$, to the learned weights, use MMD with an RBF-kernel of $\sigma = 1.0$ as IPM, and let $\alpha = 10$. We compare to exact importance sampling weights (IS) as well as clipped IS weights (ISC), $w_M(x) = \min(w(x), M)$ for $M \in \{5, 10\}$, a common way of reducing variance of re-weighting methods (Swaminathan & Joachims, 2015).

In Figure 2, we see that our proposed method behaves well at small sample sizes compared to importance sampling methods. The poor performance of exact IS weights is expected at smaller samples, as single samples are given very large weight, resulting in hypotheses that are highly sensitive to the training set. While clipped weights alleviates this issue, they do not preserve relevance ordering of high-weight samples, as many are given the truncation value M, in contrast to the re-weighting learned by our method. True domain densities are known only to IS methods.

6.2. Conditional average treatment effects — IHDP

We evaluate our framework in the CATE estimation setting, see Section 2.2. Our task is to predict the expected difference between potential outcomes conditioned on pre-treatment variables, *for a held-out sample* of the population. We compare our results to ordinary least squares (OLS) (with one regressor per outcome), OLS-IPW (re-weighted OLS according to a logistic regression estimate of propensities), Random Forests, Causal Forests (Wager & Athey, 2017), BART (Chipman et al., 2010), and CFRW (Shalit et al., 2017) (with Wasserstein penalty). Finally, we use as baseline (IPM-WNN): first weights are found by IPM minimization in the input space (Gretton et al., 2009; Kallus, 2016), then used in a re-weighted neural net regression, with the same architecture as our method.

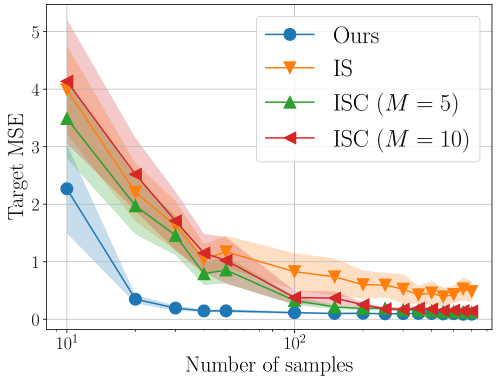

Figure 2: Target prediction error on synthetic domain adaptation experiment, comparing learned re-weighting (RCFR) and exact/clipped importance sampling weights (IS/ISC). Variance of IS hurts performance for small sample sizes.

Our implementation, dubbed RCFR for *Re-weighted CounterFactual Regression*, parameterizes representations $\Phi(x)$, weighting functions $w(\Phi, t)$ and hypotheses $h(\Phi, t)$ using neural networks, trained by minimizing (8). We use the RBF-kernel maximum mean discrepancy as the IPM (Gretton et al., 2012). For a description of the architecture, training procedure and hyperparameters, see the Appendix. We compare results using uniform $w = 1$ and learned weights, setting the balance parameter α either fixed, by an oracle (test-set error), or adaptively using the heuristic described in Section 5. To pick other hyperparameters, we split training sets into one part used for function fitting and one used for early stopping and hyperparameter selection. Hyperparameters for regularization are chosen based on the empirical loss on a held-out source (factual) sample.

The Infant Health and Development Program (IHDP) dataset is a semi-synthetic binary-treatment benchmark (Hill, 2011), split into training and test sets by Shalit et al. (2017). IHDP has a set of $d = 25$ real-world continuous and binary features describing $n = 747$ children and their mothers, a real-world binary treatment made non-randomized through biased subsampling by Hill (2011), and a synthesized continuous outcome that can be used to compute the ground-truth CATE error. Average results over 100 different realizations/settings of the outcome are presented in Table 1.

	RMSE($\hat{\tau}$)	$\hat{R}_\pi(f)$
OLS	$2.3 \pm .11$	$1.1 \pm .05$
OLS-IPW	$2.4 \pm .11$	$1.2 \pm .05$
Random For.	$6.6 \pm .30$	$4.1 \pm .18$
Causal For.	$3.8 \pm .18$	$1.8 \pm .08$
BART	$2.3 \pm .10$	$1.7 \pm .07$
IPM-WNN	$1.2 \pm .12$	$.65 \pm .02$
CFRW	$.76 \pm .02$	$.46 \pm .01$
RCFR Oracle α, $w = 1$	$.81 \pm .07$	$.47 \pm .03$
RCFR Oracle α	$.65 \pm .04$	$.38 \pm .01$
RCFR Adapt. α	$.67 \pm .05$	$.37 \pm .01$

Table 1: Causal effect estimation on IHDP. CATE error RMSE($\hat{\tau}$), target prediction error $\hat{R}_\pi(f)$ and std errors. Lower is better.

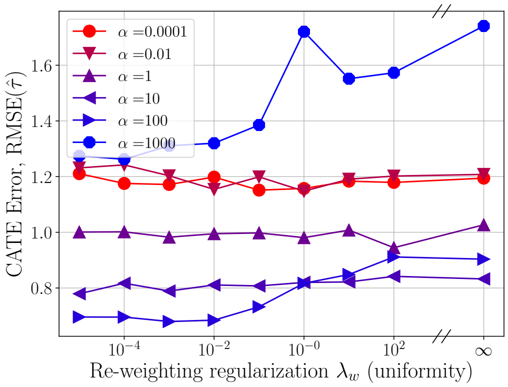

Figure 3: For small imbalance penalties α, re-weighting (low λ_w) has no effect. For moderate α, non-uniform re-weighting (smaller λ_w) lowers error, c) for large α, weighting helps, but overall error increases. Best viewed in color.

We see that our proposed method achieves state-of-the-art results, and that adaptively choosing α does not hurt performance much. Furthermore, we see a substantial improvement from using non-uniform sample weights. In Figure 3 we take a closer look at the behavior of our model as we vary its hyperparameters on the IHDP dataset. Between the two plots we can draw the following conclusions: a) For moderate to large $\alpha \in [10, 100]$, we observe a marginal gain from using the IPM penalty. This is consistent with the observations of Shalit et al. (2017). b) For large $\alpha \in [10, 1000]$, we see a large gain from using a non-uniform re-weighting (small λ_w). c) While large α makes the factual error more representative of the counterfactual error, using it without re-weighting results in higher absolute error. We believe that the moderate sample size of this dataset is one of the reasons for the usefulness of our method. See the Appendix for a complementary view of these results.

7. Discussion

We have proposed a theory and an algorithmic framework for learning to predict outcomes of interventions under shifts in design—changes in both intervention policy and feature domain. The framework combines representation learning and sample re-weighting to balance source and target designs, emphasizing information from the source sample relevant for the target. Existing re-weighting methods either use pre-defined weights or learn weights based on a measure of distributional distance in the input space. These approaches are highly sensitive to the choice of metric used to measure balance, as the input may be high-dimensional and contain information that is not predictive of the outcome. In contrast, by learning weights to achieve balance in representation space, we base our re-weighting only on information that is predictive of the outcome. In this work, we apply this framework to causal effect estimation, but emphasize that joint representation learning and re-weighting is a general idea that could be applied in many applications with design shift.

Our work suggests that distributional shift should be measured and adjusted for in a representation space relevant to the task at hand. Joint learning of this space and the associated re-weighting is attractive, but several challenges remain, including improving optimization of the proposed bound and relaxing the invertibility constraint on representations. For example, variable selection methods are not covered by our current theory, as they induce a non-ivertible representation, but a similar intuition holds there—only predictive attributes should be used when measuring imbalance. We believe that addressing these limitations is a fruitful path forward for future work.

Acknowledgements

This work was supported by Office of Naval Research Award No. N00014-17-1-2791 (DS & FJ) and by the National Science Foundation under Grant No. 1656996 (NK).

References

Ajakan, Hana, Germain, Pascal, Larochelle, Hugo, Laviolette, François, and Marchand, Mario. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

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

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

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

Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, Newey, Whitney, Robins, James, et al. Double/debiased machine learning for treatment and causal parameters. Technical report, 2017.

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

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

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

Gretton, Arthur, Smola, Alexander J, Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten M, and Schölkopf, Bernhard. Covariate shift by kernel mean matching. 2009.

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

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

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

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

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

Kallus, Nathan. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017. doi: 10.1111/rssb.12240.

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

Mansour, Yishay, Mohri, Mehryar, and Rostamizadeh, Afshin. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Morgan, Stephen L and Winship, Christopher. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

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

Precup, Doina, Sutton, Richard S, and Dasgupta, Sanjoy. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

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

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

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

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

Sriperumbudur, Bharath K, Fukumizu, Kenji, Gretton, Arthur, Schölkopf, Bernhard, Lanckriet, Gert RG, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Stuart, Elizabeth A. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1, 2010.

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

Wager, Stefan and Athey, Susan. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 325–333, 2013.

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

Appendix

A. Proofs

A.1. Definitions

Distribution re-weighting

Definition 1 (Restated). *A function $w : \mathcal{X} \times \mathcal{T} \to \mathbb{R}_+$ is a valid re-weighting of p_μ if*

$$\mathbb{E}_{x,t\sim p_\mu}[w(x,t)] = 1 \;\; \textit{and} \;\; p_\mu(x,t) > 0 \Rightarrow w(x,t) > 0.$$

We denote the re-weighted density $p_\mu^w(x,t) := w(x,t)p_\mu(x,t)$.

Expected & empirical risk We let the (expected) risk of f measured by ℓ_h under p_μ be denoted

$$R_\mu(h) = \mathbb{E}_{p_\mu}[l_h(x,t)]$$

where l_h is an appropriate loss function, and the empirical risk over a sample $D_\mu = \{(x_1,t_1,y_1)...,(x_n,t_n,y_n)$ from p_μ

$$\hat{R}_\mu(f) = \frac{1}{n}\sum_{i=1}^{n} l_f(x_i,t_i,y_i) \; .$$

We use the superscript w to denote the re-weighted risks

$$R_\mu^w(f) = \mathbb{E}[w(x,t)l_f(x,t)]$$

$$\hat{R}_\mu^w(f) = \frac{1}{n}\sum_{i=1}^{n} w(x_i,t_i)l_h(x_i,t_i,y_i)$$

Definition A1 (Importance sampling)**.** *For two distributions p, q on $\mathcal{Z}$, of common support, $\forall z \in \mathcal{Z} : p(z) > 0 \iff q(z) > 0$, we call*

$$w_{IS}(z) := \frac{q(z)}{p(z)}$$

the importance sampling *weights of p and q.*

Definition 2 (Restated). *The integral probability metric (IPM) distance, associated with the function family $\mathcal{H}$, between distributions p and q is defined by*

$$\mathrm{IPM}_\mathcal{H}(p,q) := \sup_{h:\|h\|_\mathcal{H}=1} |\mathbb{E}_p[h] - \mathbb{E}_q[h]|$$

A.2. Learning bounds

We begin by bounding the expected risk under a distribution p_π in terms of the expected risk under p_μ and a measure of the discrepancy between p_π and p_μ. Using definition 2 we can show the following result.

Lemma 1 (Restated). *For hypotheses f with loss ℓ_f such that $\ell_f/\|\ell_f\|_\mathcal{H} \in \mathcal{H}$, and p_μ, p_π with common support, there exists a valid re-weighting w of p_μ, see Definition 1, such that,*

$$\begin{aligned} R_\pi(f) \;\; & \leq R_\mu^w(f) + \|\ell_f\|_\mathcal{H}\mathrm{IPM}_\mathcal{H}(p_\pi, p_\mu^w) \\ & \leq R_\mu(f) + \|\ell_f\|_\mathcal{H}\mathrm{IPM}_\mathcal{H}(p_\pi, p_\mu) \; . \end{aligned} \tag{11}$$

The first inequality is tight for importance sampling weights, $w(x,t) = p_\pi(x,t)/p_\mu(x,t)$. The second inequality is not tight for general f, even if $\ell_f \in \mathcal{H}$, unless $p_\pi = p_\mu$.

Proof. The results follows immediately from the definition of IPM.

$$\begin{aligned} R_\pi(f) - R_\mu^w(f) &= \mathbb{E}_\pi[\ell_f(x,t)] - \mathbb{E}_\mu[w(x,t)\ell_f(x,t)] \\ &\leq \sup_{h\in\mathcal{H}_\ell} |\mathbb{E}_\pi[h(x,t)] - \mathbb{E}_\mu[w(x,t)h(x,t)]| \\ &= \mathrm{IPM}_{\mathcal{H}_\ell}(p_\pi, p_\mu^w) \end{aligned}$$

Further, for importance sampling weights $w_{IS}(x,t) = \frac{\pi(t;x)}{\mu(t;x)}$, for any $h \in \mathcal{H}$,

$$\begin{aligned} & \mathbb{E}_\pi[h(x,t)] - \mathbb{E}_\mu[w_{IS}(x,t)h(x,t)] \\ = \; & \mathbb{E}_\pi[h(x,t)] - \mathbb{E}_\mu[\frac{\pi(t;x)}{\mu(t;x)}h(x,t)] = 0 \end{aligned}$$

and the LHS is tight. □

We could apply Lemma 1 to bound the loss under a distribution q based on the weighted loss under p. Unfortunately, bounding the expected risk in terms of another expectation is not enough to reason about generalization from an empirical sample. To do that we use Corollary 2 of Cortes et al. (2010), restated as a Theorem below.

Theorem A1 (Generalization error of re-weighted loss (Cortes et al., 2010))**.** *For a loss function ℓ_h of any hypothesis $h \in \mathcal{H} \subseteq \{h' : \mathcal{X} \to \mathbb{R}\}$, such that $d =$ Pdim$(\{\ell_h : h \in \mathcal{H}\})$ where* Pdim *is the pseudo-dimension, and a weighting function $w(x)$ such that $\mathbb{E}_p[w] = 1$, with probability $1-\delta$ over a sample $(x_1, ..., x_n)$, with empirical distribution $\hat{p}$,*

$$\begin{aligned} R_p^w(h) &\leq \hat{R}_p^w(h) \\ &\quad + 2^{5/4} V_{p,\hat{p}}[w(x)l(x)] \left(\frac{d\log\frac{2ne}{d} + \log\frac{4}{\delta}}{n}\right)^{3/8} \end{aligned}$$

with

$$\begin{aligned} & V_{p,\hat{p}}[w(x)l(x)] \\ & = \max(\sqrt{\mathbb{E}_p[w^2(x)\ell_h^2(x)]}, \sqrt{\mathbb{E}_{\hat{p}}[w^2(x)\ell_h^2(x)]}) \; . \end{aligned}$$

With

$$\mathcal{C}_n^\mathcal{H} = 2^{5/4}\left(d\log\frac{2ne}{d} + \log\frac{4}{\delta}\right)^{3/8}$$

we get the simpler form

$$R_p^w(h) \le \hat{R}_p^w(h) + V_{p,\hat{p}}[w(x)l(x)]\frac{\mathcal{C}_n^{\mathcal{H}}}{n^{3/8}} \ .$$

We will also need the following result about estimating IPMs from finite samples from Sriperumbudur et al. (2009).

Theorem A2 (Estimation of IPMs from empirical samples (Sriperumbudur et al., 2009))**.** *Let M be a measurable space. Suppose k is measurable kernel such that $\sup_{x\in M} k(x,x) \le C \le \infty$ and $\mathcal{H}$ the reproducing kernel Hilbert space induced by k, with $\nu := \sup_{x\in M, f\in\mathcal{H}} f(x) < \infty$. Then, with $\hat{p}, \hat{q}$ the empirical distributions of p, q from m and n samples respectively, and with probability at least $1-\delta$,*

$$\begin{aligned} &|\text{IPM}_{\mathcal{H}}(p,q) - \text{IPM}_{\mathcal{H}}(\hat{p},\hat{q})| \\ &\le \sqrt{18\nu^2 \log\frac{4}{\delta}C}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) \end{aligned}$$

We consider learning twice-differentiable, invertible representations $\Phi : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is the representation space, and $\Psi : \mathcal{Z} \to \mathcal{X}$ is the inverse representation, such that $\Psi(\Phi(x)) = x$ for all x. Let $\mathcal{E}$ denote space of such representation functions. For a design π, we let $p_{\pi,\Phi}(z,t)$ be the distribution induced by Φ over $\mathcal{Z}\times\mathcal{T}$, with $p^w_{\pi,\Phi}(z,t) := p_{\pi,\Phi}(z,t)w(\Psi(z),t)$ its re-weighted form and $\hat{p}^w_{\pi,\Phi}$ its re-weighted empirical form, following our previous notation. Note that we do not include t in the representation itself, although this could be done in principle. Let $\mathcal{G} \subseteq \{h : \mathcal{Z}\times\mathcal{T} \to \mathcal{Y}\}$ denote a set of hypotheses $h(\Phi, t)$ operating on the representation Φ and let $\mathcal{F}$ denote the space of all compositions, $\mathcal{F} = \{f = h(\Phi(x), t) : h \in \mathcal{G}, \Phi \in \mathcal{E}\}$. We now restate and prove Theorem 1.

Theorem 1 (Restated). *Given is a labeled sample $D_\mu = \{(x_1,t_1,y_1),...,(x_n,t_n,y_n)\}$ from p_μ, and an unlabeled sample $D_\pi = \{(x'_1,t'_1),...,(x'_m,t'_m)\}$ from p_π, with corresponding empirical measures $\hat{p}_\mu$ and $\hat{p}_\pi$. Suppose that Φ is a twice-differentiable, invertible representation, that $h(\Phi,t)$ is an hypothesis, and $f = h(\Phi(x),t) \in \mathcal{F}$. Define $m_t(x) = \mathbb{E}_Y[Y \mid X = x, T = t]$, let $\ell_{h,\Phi}(\Psi(z),t) := L(h(z,t), m_t(\Psi(z)))$ where L is the squared loss, $L(y,y') = (y-y')^2$, and assume that there exists a constant $B_\Phi > 0$ such that $\ell_{h,\Phi}/B_\Phi \in \mathcal{H} \subseteq \{h : \mathcal{Z}\times\mathcal{T} \to \mathcal{Y}\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space of a kernel, k such that $k((z,t),(z,t)) < \infty$. Finally, let w be a valid re-weighting of $p_{\mu,\Phi}$. Then with probability at least $1-2\delta$,*

$$\begin{aligned} R_\pi(f) \ \ &\le \hat{R}^w_\mu(f) + B_\Phi \text{IPM}_{\mathcal{H}}(\hat{p}_{\pi,\Phi}, \hat{p}^w_{\mu,\Phi}) \\ &+V_\mu(w,\ell_f)\tfrac{\mathcal{C}^{\mathcal{F}}_{n,\delta}}{n^{3/8}} \\ &+\mathcal{D}^{\Phi,\mathcal{H}}_\delta\left(\tfrac{1}{\sqrt{m}} + \tfrac{1}{\sqrt{n}}\right) + \sigma^2_Y \end{aligned} \tag{12}$$

where $\mathcal{C}^{\mathcal{F}}_{n,\delta}$ measures the capacity of $\mathcal{F}$ and has only logarithmic dependence on n, $\mathcal{D}^{\mathcal{H}}_{m,n,\delta}$ measures the capacity of $\mathcal{H}$, σ^2_Y is the expected variance in potential outcomes, and

$$\begin{aligned} &V_\mu(w,\ell_f) \\ &= \max\left(\sqrt{\mathbb{E}_{p_\mu}[w^2(x,t)\ell_f^2(x,t)]}, \sqrt{\mathbb{E}_{\hat{p}_\mu}[w^2(x,t)\ell_f^2(x,t)]}\right) . \end{aligned}$$

A similar bound exists where $\mathcal{H}$ is the family of functions Lipschitz constant at most 1, but with worse sample complexity.

Proof. We have by definition

$$\begin{aligned} &R_\pi(f) - R^w_\mu(f) = \mathbb{E}_\pi[\ell_f(x,t,y)] - \mathbb{E}_\mu[w(x,t)\ell_f(x,t,y)] \\ &= \int_{x,t,y} \ell_f(x,t,y)p(y \mid t,x)(p_\pi(x,t) - p^w_\mu(x,t))dxdtdy \end{aligned}$$

Define $\ell_{h,\Phi}(x,t) = L(h(\Phi(x),t), m_t(x))$ where $m_t(x) := E[Y \mid T = t, X = x]$). Then, with L, the squared loss, $L(y,y') = (y-y')^2$, we have,

$$\mathbb{E}_\pi[\ell_{h,\Phi}(x,t,y)] = \mathbb{E}_\pi[\ell_{h,\Phi}(x,t)] + \sigma^2_\pi$$

where $\sigma^2_\pi = \mathbb{E}_{p_\pi}[(Y - m_t(x))^2]$, and analogously for μ. We get that

$$\begin{aligned} &R_\pi(f) - R^w_\mu(f) = \\ &\int_{\substack{z\in\mathcal{Z}\\ t\in\mathcal{T}}} \ell_{h,\Phi}(x,t)(p_\pi(x,t) - p^w_\mu(x,t))dxdt + \sigma^2_\pi + \sigma^2_\mu \\ &= \int_{\substack{z\in\mathcal{Z}\\ t\in\mathcal{T}}} \ell_{h,\Phi}(\Psi(z),t)(p_{\pi,\Phi}(z,t) - p^w_{\mu,\Phi}(z,t))|J_\Psi(z)|dzdt \\ &\quad + \sigma^2_\pi + \sigma^2_\mu \\ &\le A_\Phi \int_{\substack{z\in\mathcal{Z}\\ t\in\mathcal{T}}} \ell_{h,\Phi}(\Psi(z),t)(p_\pi(z,t) - p^w_\mu(z,t))dzdt \\ &\quad + \sigma^2_\pi + \sigma^2_\mu \\ &\le A_\Phi\|\ell_{h,\Phi}\|_{\mathcal{H}} \sup_{h\in\mathcal{H}} \left| \int_{\substack{z\in\mathcal{Z}\\ t\in\mathcal{T}}} h(\Psi(z),t)\left(p_{\pi,\Phi}(z,t) - p^w_{\mu,\Phi}(z,t)\right)dzdt \right| \\ &\quad + \sigma^2_\pi + \sigma^2_\mu \\ &= B_\Phi \cdot \text{IPM}_{\mathcal{H}}(p_{\pi,\Phi}, p^w_{\mu,\Phi}) + \sigma^2_\pi + \sigma^2_\mu \end{aligned}$$

where $J_\Psi(z)$ is the Jacobian matrix of Ψ evaluated at z and $A_\Phi \ge |J_\Psi(z)|$ for all $z \in \mathcal{Z}$, where $|J|$ is the absolute determinant of J. By application of Theorem A1 we have with probability at least $1-\delta$,

$$R^w_\mu(f) \le \hat{R}^w_\mu(f) + V_\mu(w,\ell)\frac{\mathcal{C}^{\mathcal{H}}_{n,\delta}}{n^{3/8}} \ .$$

and by applying Theorem A2, we have with probability at least $1-\delta$,

$$\begin{aligned} &\left|\text{IPM}_{\mathcal{H}}(p_{\pi,\Phi}, p^w_{\mu,\Phi}) - \text{IPM}_{\mathcal{H}}(\hat{p}_{\pi,\Phi}, \hat{p}^w_{\mu,\Phi})\right| \\ &\le \sqrt{18\nu^2 \log\frac{4}{\delta}C}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) \end{aligned}$$

We let $\sigma_Y^2 = \sigma_\pi^2 + \sigma_\mu^2$ and

$$\mathcal{D}_\delta^{\Phi,\mathcal{H}} := B_\Phi \sqrt{18\nu^2 \log \frac{4}{\delta} C}$$

Combining these results, observing that $(1-\delta)^2 \geq 1 - 2\delta$, we obtain the desired result. □

A.3. Asymptotics

Theorem 2 (Restated). *Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space given by a bounded kernel. Suppose weak overlap holds in that $\mathbb{E}[(p_\pi(x,t)/p_\mu(x,t))^2] < \infty$. Then,*

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

Proof. Let $f^* = \Phi^* \circ h^* \in \arg\min_{f\in\mathcal{F}} R_\pi(f)$ and let $w^*(x,t) = p_{\pi,\Phi}(\Phi^*(x),t)/p_{\mu,\Phi}(\Phi^*(x),t)$. Since $\min_{h,\Phi,w} \mathcal{L}_\pi(h,\Phi,w;\beta) \leq \mathcal{L}_\pi(h^*,\Phi^*,w^*;\beta)$, it suffices to show that $\mathcal{L}_\pi(h^*,\Phi^*,w^*;\beta) = R_\pi(f^*) + O(1/\sqrt{n} + 1/\sqrt{m})$. We will work term by term:

$$\mathcal{L}_\pi(h^*,\Phi^*,w^*;\beta) = \underbrace{\frac{1}{n}\sum_{i=1}^n w_i \ell_h(\Phi(x_i),t_i)}_{\textcircled{A}}$$

$$+ \lambda_h \underbrace{\frac{\mathcal{R}(h)}{\sqrt{n}}}_{\textcircled{B}} + \alpha \underbrace{\text{IPM}_G(\hat{q}_\Phi, \hat{p}_\Phi^{w^k})}_{\textcircled{C}} + \lambda_w \underbrace{\frac{\|w\|_2}{n}}_{\textcircled{D}} .$$

For term $\textcircled{D}$, letting $w_i^* = w^*(x_i,t_i)$, we have that by weak overlap

$$\textcircled{D}^2 = \frac{1}{n} \times \frac{1}{n}\sum_{i=1}^n (w_i^*)^2 = O_p(1/n),$$

so that $\textcircled{D} = O_p(1/\sqrt{n})$. For term $\textcircled{A}$, under ignorability, each term in the sum in the first term has expectation equal to $R_\pi(f^*)$ and so, so by weak overlap and bounded second moments of loss, we have $\textcircled{A} = R_\pi(f^*) + O_p(1/\sqrt{n})$. For term $\textcircled{B}$, since h^* is fixed we have deterministically that $\textcircled{B} = O(1/\sqrt{n})$.

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

$$\sup_{\|h\|\leq 1} (\frac{1}{m}\sum_{i=1}^m h(\Phi^*(x_i'),t_i') - \frac{1}{n}\sum_{i=1}^n w_i^* h(\Phi^*(x_i),t_i)).$$

Let x_i'', t_i'' for $i = 1,\ldots,m$ and x_i''', t_i''' for $i = 1,\ldots,n$ be new iid replicates of x_1', t_1', i.e., new ghost samples drawn from the target design. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[\textcircled{C}^2] &= \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{m}\sum_{i=1}^m h(\Phi^*(x_i'),t_i') - \frac{1}{n}\sum_{i=1}^n w_i^* h(\Phi^*(x_i),t_i))^2] \\
&= \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{m}\sum_{i=1}^m (h(\Phi^*(x_i'),t_i') - \mathbb{E}[h(\Phi^*(x_i''),t_i'')]) \\
&\quad - \frac{1}{n}\sum_{i=1}^n (w_i^* h(\Phi^*(x_i),t_i) - \mathbb{E}[h(\Phi^*(x_i'''),t_i''')]))^2] \\
&\leq \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{m}\sum_{i=1}^m (h(\Phi^*(x_i'),t_i') - h(\Phi^*(x_i''),t_i'')) \\
&\quad - \frac{1}{n}\sum_{i=1}^n (w_i^* h(\Phi^*(x_i),t_i) - h(\Phi^*(x_i'''),t_i''')))^2] \\
&\leq 2\mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{m}\sum_{i=1}^m (h(\Phi^*(x_i'),t_i') - h(\Phi^*(x_i''),t_i'')))^2] \\
&\quad + 2\mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{n}\sum_{i=1}^n (w_i^* h(\Phi^*(x_i),t_i) - h(\Phi^*(x_i'''),t_i''')))^2]
\end{aligned}$$

Let $\xi_i(h) = h(\Phi^*(x_i'),t_i') - h(\Phi^*({X'}_i^q)$ and let $\zeta_i(h) = w_i^* h(\Phi^*(x_i),t_i) - h(\Phi^*(x_i'''),t_i''')$. Note that for every h, $\mathbb{E}[\zeta_i(h)] = \mathbb{E}[\xi_i(h)] = 0$. Moreover, $\mathbb{E}[\|\zeta_i\|^2] \leq 4E[K(\Phi^*(x_i'),t_i',\Phi^*(x_i'),t_i')] \leq M$. Similarly, $\mathbb{E}[\|\xi_i\|^2] \leq 2E[(w_i^*)^2]M + 2M \leq M' < \infty$ because of weak overlap. Let ζ_i' for $i = 1,\ldots,n$ be iid replicates of ζ_i (ghost sample) and let ϵ_i be iid Rademacher random variables. Because $\mathcal{H}$ is a Hilbert space, we have that $\sup_{\|h\|\leq 1}(A(h))^2 =$

$\|A\|^2 = \langle A, A \rangle$. Therefore, by Jensen's inequality,

$$
\begin{aligned}
&\mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{n}\sum_{i=1}^{n}(w_i^* h(\Phi^*(x_i), t_i) - h(\Phi^*(x_i'''), t_i''')))^2] \\
&= \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{n}\sum_{i=1}^{n}\zeta_i(h))^2] \\
&= \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{n}\sum_{i=1}^{n}(\zeta_i(h) - \mathbb{E}[\zeta_i'(h)]))^2] \\
&\leq \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{n}\sum_{i=1}^{n}(\zeta_i(h) - \zeta_i'(h)))^2] \\
&= \mathbb{E}[\sup_{\|h\|\leq 1} (\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\zeta_i(h) - \zeta_i'(h)))^2] \\
&\leq \frac{4}{n^2}\mathbb{E}[\sup_{\|h\|\leq 1} (\sum_{i=1}^{n}\epsilon_i\zeta_i(h))^2] \\
&= \frac{4}{n^2}\mathbb{E}[\|\sum_{i=1}^{n}\epsilon_i\zeta_i\|^2] \\
&= \frac{4}{n^2}\mathbb{E}[\sum_{i,j=1}^{n}\epsilon_i\epsilon_j\,\langle\zeta_i, \zeta_j\rangle] \\
&= \frac{4}{n^2}\mathbb{E}[\sum_{i=1}^{n}\|\zeta_i\|^2] \\
&= \frac{4}{n^2}\sum_{i=1}^{n}\mathbb{E}[\|\zeta_i\|^2] \\
&\leq \frac{4M'}{n}
\end{aligned}
$$

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

B. Implementation

We implemented all neural network models (IPM-WNN, RCFR) in TensorFlow as feed-forward fully-connected networks with ELU activations. All architectures have a representation with two hidden layers of 32 and 16 hidden units, and hypotheses (one for each outcome) of 1 layer of 16 hidden units. The networks were trained using stochastic gradient descent with the ADAM optimizer with a learning rate of 10^{-3}. The batch size was 128. Representations were normalized by dividing by the norm. Weight functions were implemented as 2 hidden layers of 32 units each, as functions of the representation Φ. σ in the RBF kernel was set to 1.0. λ_w was set to 0.1 and λ_h to 10^{-4}.

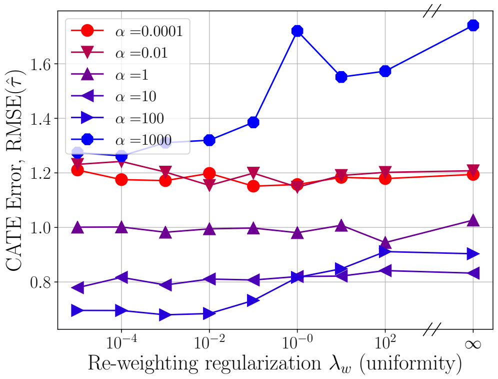

Figure 4: Error in CATE estimation on IHDP as a function of re-weighting regularization strength λ_w. We see that a) for small imbalance penalties α, re-weighting (low λ_w) has no effect, b) for moderate α, less uniform re-weighting (smaller λ_w) improves the error, c) for large α, weighting helps, but overall error increases.

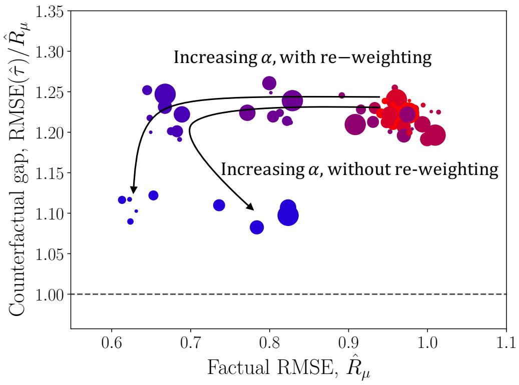

Figure 5: Source prediction error on IHDP. we compare the ratio of CATE error to source error. Color represents α (see left) and size λ_w. We see that for large α, the source error is more representative of CATE error, but does not improve in absolute value without weighting. Here, α was fixed. Best viewed in color.

C. Experiments

C.1. Synthetic

We use a two-layer MLP with ELU units and layer sizes 10, 10 as parameterization of the sample weights. Weights are normalized by dividing by the mean.

C.2. IHDP

In Figures 4–5, we see two different views of the IHDP results.