Towards representation learning for general weighting problems in causal inference

Oscar Clivio, Avi Feller, Chris Holmes
University of Oxford, UC Berkeley
{oscar.clivio, chris.holmes}@stats.ox.ac.uk, afeller@berkeley.edu

Abstract

Weighting problems in treatment effect estimation can be solved by minimising an appropriate probability distance. Choosing which distance to minimise, however, can be challenging as it depends on the unknown data generating process. An alternative is to instead choose a distance that depends on a suitable representation of covariates. In this work, we give errors that quantify the bias added to a weighting estimator when using a representation, giving clear objectives to minimise when learning the representation and generalising a large body of previous work on deconfounding, prognostic, balancing and propensity scores. We further outline a method minimising such objectives, and show promising numerical results on two semi-synthetic datasets.

1 Introduction

Estimating the causal effect of a treatment variable on an outcome of study is a fundamental task in multiple fields such as epidemiology [1], medicine [2], public policy [3] or economics [4]. Some challenges include removing the influence of confounders [5] or generalising a treatment effect estimated on a randomised control trial (RCT) to a target observational population [6, 7]. Both problems can be solved with weighting [8, 7] : we reweight an original distribution to target a causal effect of interest. A set of methods relies on minimising a probability distance between the weighted distribution and a reference one, however we do not know which distance to minimise as it depends on a model for the outcome, which we do not have access to [9].

In this paper, we choose a distance that depends on an adequate *representation*, that is a (potentially multivariate) mapping of covariates to another manifold, which we learn from data. Our main contributions are : 1) we show that a "deconfounding error" quantifies the added bias on the weighting estimator when using a distance on the representation and should be minimised for this purpose, 2) we deduce a "balancing score error" that does not depend on the unknown outcome information and measures how much representations are not balancing scores [10] while giving guarantees on the bias, adding flexibility compared to assuming well-specified balancing (or propensity) scores as does a significant portion of the literature [10, 11] ; 3) we outline a method inspired from RieszNet [12] that learns such representations from data and apply it to a popular dataset in treatment effect estimation.

2 Background

2.1 Notations

Let X denote pre-treatment covariates, A denote the treatment variable and Y denote the outcome. We assume that the values of A belong to a finite (and potentially binary) space $\mathcal{A}$. We assume we have access to i.i.d. data $\{(X_i, A_i, Y_i)\}$. For $a \in \mathcal{A}$, we note $Y(a)$ the potential outcome wrt a, that is the outcome that a subject receives if they were to receive treatment a. Further, in transportability,

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

we have an other binary variable S such that $S = 1$ denotes membership in the RCT population, i.e. $(Y(1), Y(0)) \perp\!\!\!\perp A|S = 1$, and we do not have access to A, Y when $S = 0$.

2.2 Framework

Let P be a **source** distribution, Q a **target** distribution. We assume that P, Q have densities p, q, respectively, and that we have access to (not necessarily disjoint) samples $\mathcal{P}$ from P and $\mathcal{Q}$ from Q. We note $\mathbb{E}_R[f(Z)]$ be the expectation of $f(Z)$ under a function f when Z follows the distribution R. We call **weight function** or **weights** any non-negative function $w(x)$ of covariates such that $\mathbb{E}_P[w(X)] = 1$. Any weight function w induces a distribution P_w with density $w(x)p(x)$, where we say that is P **reweighted** by $w(X)$, with $\mathbb{E}_{P_w}[f(X)] = \mathbb{E}_P[w(X)f(X)]$ for any function f. Let $m(x)$ be a function of interest which we call **the outcome model**. We assume there exists an observed random variable $\tilde{Y}$, which we call the **pseudo-outcome**, with $m(x) = \mathbb{E}_P[\tilde{Y}|X = x]$. We are interested in the **target estimand** $\mathbb{E}_Q[m(X)]$. However, we do not have access to either the outcome model or the target estimand. On the other hand, for any weight function $w(x)$, $\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} w(X_i)\tilde{Y}_i$ is an unbiased estimator of $\mathbb{E}_{P_w}[m(X)]$. All of this motivates our problem statement.

Problem 1 *(General weighting problem) Find a weight function $w(X)$ such that*

$$\mathbb{E}_{P_w}[m(X)] = \mathbb{E}_Q[m(X)]$$

where $m(x) = \mathbb{E}_P[\tilde{Y}|X = x]$.

This generalises many weighting problems in treatment effect estimation (details in Supplement 6 ; some related work on importance weighting is in Supplement 7). In ATT (average treatment effect on the treated) estimation [8], we focus on estimating $\mathbb{E}[Y(0)|A = 1]$: the source distribution is $P(X|A = 0)$, the target one is $P(X|A = 1)$, the outcome model is $\mathbb{E}[Y(0)|X = x]$, the pseudo-outcome is Y. In ATE estimation, we estimate $\mathbb{E}[Y(a)]$ for each $a \in \mathcal{A}$ [13], the source distribution is $P(X|A = a)$, the target one is the marginal $P(X)$, the outcome model is $\mathbb{E}[Y(a)|X = x]$, the pseudo-outcome is Y. In transportability [6, 7], we are interested in the average treatment effect $\mathbb{E}[Y(1) - Y(0)|S = 0]$ on a target population $S = 0$ where A, Y are accessible only for the RCT population ($S = 1$) : the source distribution is the RCT population $P(X|S = 1)$, the target one is the target population $P(X|S = 0)$, the outcome model is CATE(x) $= \mathbb{E}[Y(1) - Y(0)|X = x]$, the pseudo-outcome is $\frac{AY}{P(A=1)} - \frac{(1-A)Y}{P(A=0)}$. We make the following important assumption.

Assumption 1 *(Absolute continuity) Q is absolutely continuous wrt P, i.e. for any x such that $q(x) > 0$, we have $p(x) > 0$.*

This assumption is equivalent to overlap in ATT/ATE estimation [14] and support inclusion [7] in transportability. It ensures that the inverse probability weight function (also called Riesz representer [15, 16]) $w^*(x) = {}^{q(x)}/_{p(x)}$ is well-defined. Here we more simply refer to it as the **true weights**. Such weights w^* are uniquely defined by $Q = P_{w^*}$ or, equivalently, for any function f [8],

$$\mathbb{E}_{P_{w^*}}[f(X)] = \mathbb{E}_Q[f(X)].$$

In particular, this holds for $f = m$, solving our general weighting problem. In practice, we unfortunately do not have access to the true weights w^* and need to either estimate them or more generally find another way to obtain a solution weight function.

2.3 Common methods in weighting

In ATT/ATE estimation and transportability, w^* is proportional to the inverse of one of the propensity scores $p(A = a|X = x)$ [17, 8] or $P(S = 1|X = x)$ [18, 19, 20, 21]. Thus, an inverse probability weighting estimator $\widehat{w}$ of w^* is obtained by fitting a model for the indicated propensity score and inverting it, leading to potentially outsize errors due to misspecification [22, 23]. An alternative used in the automatic debiased machine learning (AutoDML) literature is to minimise the mean squared error between w^* and $\widehat{w}$, which can actually be estimated without exactly knowing the true weights w^* [16, 12, 24]. Another family of methods [25, 26, 27, 28, 29, 30, 31, 32] relies on imposing that weights w verify **balance** in some moments r, i.e. $\mathbb{E}_{P_w}[r(X)] = \mathbb{E}_Q[r(X)]$. Then one minimises some dispersion measure of weights under these constraints. However, balancing r does not guarantee balancing the unknown m [13] and the solution might not be feasible if r has too many moments [9]. A related family of methods finds parameterised weights or propensity scores that approximately induce balance in moments r through a generalised method of moments [33, 27].

Finally, another family of methods [34, 23, 8, 14] aims at finding weights w minimising $|\text{Bias}(w; m)|$ where we refer to $\text{Bias}(w; m) = \mathbb{E}_{P_w}[m(X)] - \mathbb{E}_Q[m(X)]$ as the **"bias"** of weights w, measuring how short they fall of solving Problem 1. It is usually assumed that m belongs to a class of functions $\mathcal{M}$ which leads to the bound

$$|\text{Bias}(w; m)| \leq \text{IPM}_{\mathcal{M}}(P_w(X), Q(X)) := \sup_{\bar{m} \in \mathcal{M}} |\mathbb{E}_{P_w}[\bar{m}(X)] - \mathbb{E}_Q[\bar{m}(X)]|$$

where the RHS is an integral probability metric (IPM) [35] on the class $\mathcal{M}$ and generally corresponds to a known generic probability discrepancy, for example the Wasserstein distance when $\mathcal{M}$ is the set of Lipschitz functions or the maximal mean discrepancy (MMD) wrt kernel k when $\mathcal{M}$ is the RKHS of k. Thus, adding some regularisation to take variance from finite samples into account [13], we obtain a solution w by solving the problem

$$\min_w \text{IPM}_{\mathcal{M}}(P_w(X), Q(X))^2 + \sigma^2 \cdot ||w||^2_{L_2(P)} \tag{1}$$

for a chosen $\sigma^2 > 0$ [14]. A key challenge is that as we do not know the outcome model m, we do not know the model class $\mathcal{M}$, thus an adequate probability discrepancy to minimise. In practice, one resorts to trying a specific discrepancy, thus making an implicit assumption on the function space $\mathcal{M}$ which can then be inadequate wrt the outcome model m at stake. Thus, a direction in the literature is to find a data-driven tailored function class $\mathcal{M}$ [23, 9].

2.4 Choosing a distance via a representation

Many methods minimise a probability discrepancy measure or more generally find weights that only depend on covariates x via a **representation** (vector-valued function) $\phi(x)$ that is selected by the method [10, 36, 37, 23, 38, 39, 40, 41].There are several motivations to do so. First, some function classes implicitly define a representation, such as the kernel feature spaces $k(., x)$ for the RKHS of kernel k [23]. In turn, every representation defines a function class. Further, a low-dimensional representation can mitigate undesirable effects of high dimensions in causal inference [42, 43] or probability distances [44, 45, 46] and improve efficiency by selecting essential covariate information wrt the DGP, generalising motivations behind variable selection [47, 7].

The question then becomes how to obtain suitable representations $\phi(x)$. It is generally well-known that weighting on the true outcome model [9], the propensity score [10] or a representation predicting either [36, 10, 48, 40] is a sensible choice as these representations preserve unconfoundedness. However, we do not have access to these true models or representations predicting them. Methods based on sufficient dimension reduction attempt at finding a representation under the constraint that it predicts either model [49, 50, 51, 52, 53], while others extract representations from a learnt model for the outcome, the treatment or the RCT indicator [36, 54, 55, 40, 18, 19]. However, to the best of our knowledge, there are no guarantees on the bias when the posited model is misspecified, while they are critical as one does not have access to either model in general. In particular, classification-based learning of propensity scores does not optimise for covariate balance but for prediction of the treatment or the RCT indicator, while (near-)deterministic prediction of either will violate (strict [56]) overlap, leading to poor matching or weighting performance in practice [11].

3 Representation learning with approximate deconfounding scores

3.1 From a bias decomposition to the deconfounding error

Replacing X with $\phi(X)$ in a discrepancy measure might lead us to lose important information useful for weighting, such as unconfoundedness. Thus one might wonder which representations lead to "not too much" loss of information. On one extreme case, we know that true representations $\phi(X)$ like balancing scores and prognostic scores perfectly preserve unconfoundedness. In the other extreme, a probability distance wrt a *constant* $\phi(X)$ will always be zero, while the bias will take any value : all information has been lost. In the following, we outline our main contribution : we characterise representations that lead to "acceptable" loss of information, and do so thanks to a "deconfounding error". We start with the following decomposition.

Theorem 1 *For any function $f(x)$, representation $\phi(x)$ and distribution $R(X)$, let $f_\phi^R(\varphi) := \mathbb{E}_R[f(X)|\phi(X) = \varphi]$. In the absence of an explicit distribution in the superscript, we take the source distribution P, or in other words, $f_\phi(\varphi) := f_\phi^P(\varphi)$. The bias can be decomposed as*

$$\forall w, \phi, \;\; \textit{Bias}(w; m) = \textit{Bias}(w, \phi; m) + \textit{CE}(w, \phi; m) + \textit{DE}(\phi; m)$$

where

$$\begin{aligned} \textit{Bias}(w, \phi; m) &:= \mathbb{E}_{P_w}[m_\phi(\phi(X))] - \mathbb{E}_Q[m_\phi(\phi(X))], \\ \textit{CE}(w, \phi; m) &:= \mathbb{E}_P[(m(X) - m_\phi(\phi(X))) \cdot (w(X) - w_\phi(\phi(X)))], \\ \textit{DE}(\phi; m) &:= -\mathbb{E}_P[(m(X) - m_\phi(\phi(X))) \cdot (w^*(X) - w_\phi^*(\phi(X)))] = -\textit{CE}(w^*, \phi; m). \end{aligned}$$

where $w^*(x) = \frac{q(x)}{p(x)}$ are the true weights and, in addition, $w_\phi(\phi(x)) = \frac{p_w(\phi(x))}{p(\phi(x))}$ where $r(\phi(x))$ denotes the density of the marginal $R(\phi(X))$.

All proofs are in Supplement 9. We now explain the signification of each of these three terms. As $m_\phi(\phi(x)) = \mathbb{E}[\tilde{Y}|\phi(X) = \phi(x)]$, the term $\text{Bias}(w, \phi; m)$ represents the bias of w wrt the representation ϕ, in the sense that it would be the bias if we replaced X with $\phi(X)$ in the equality of Problem 1. This is also the term that is bounded by some integral probability metric $\text{IPM}_{\mathcal{G}}(P_w(\phi(X)), Q(\phi(X)))$, where m_ϕ belongs to some class $\mathcal{G}$, which is precisely the sort of probability distance we would like to minimise. For example, when $m \in \mathcal{M}$, then $m_\phi \in \mathcal{G} = \phi(\mathcal{M}) := \{\varphi \to \mathbb{E}[\bar{m}(X)|\phi(X) = \varphi], \;\; \bar{m} \in \mathcal{M}\}$. Then the weights resulting from optimisation of the IPM will depend only on the representation $\phi(x)$.

Theorem 2 *For any ϕ with values in a space Φ and class of functions $\mathcal{G}$ on Φ, the solution $\tilde{w}(x)$ to the problem $\min_w \textit{IPM}_{\mathcal{G}}(P_w(\phi(X)), Q(\phi(X)))^2 + \sigma^2 \cdot ||w||_{L_2(P)}^2$ is a function of $\phi(x)$.*

Now we refer to the second and third terms as "errors" to indicate that they are additional biases from this "bias" wrt ϕ. The second term $\text{CE}(w, \phi; m)$ is equal to zero when $m(x)$ or $w(x)$ depends on $\phi(x)$, which is why we call this term "conservation error", as $\phi(x)$ "conserves" the essential information of $m(x)$ and/or $\phi(x)$. However, this term is actually not of concern : as the solution weights $\tilde{w}$ to the problem in Theorem 2 will depend on $\phi(x)$, we will obtain $\text{CE}(\tilde{w}, \phi; m) = 0$.

Thus, the only added bias to weights depending on $\phi(x)$ compared to the bias wrt ϕ is the third term, $\text{DE}(\phi; m)$, which we call the "deconfounding error". Indeed, for ATE estimation, it is the difference in estimating $\mathbb{E}[Y(a)]$ between adjusting on $\phi(X)$ and adjusting on X [57], measuring how much $\phi(X)$ preserves unconfoundedness. We now use this term to generalise common notions of "scores".

3.2 Deconfounding scores

We note that $\text{DE}(\phi; m) = 0$ (a property we call *deconfounding*) in three important cases : 1) $m(x)$ is a function of $\phi(x)$, where we call $\phi(x)$ a *generalised prognostic score*. 2) $w^*(x)$ is a function of $\phi(x)$, where we call $\phi(x)$ a *generalised balancing score*, 3) $\text{DE}(\phi; m) = 0$ without ϕ necessarily being a generalised prognostic or balancing score, where we call $\phi(x)$ a *generalised deconfounding score*. The following result connects these notions to previous literature.

Theorem 3 *In ATT/ATE estimation, a) balancing scores are equivalent to generalised balancing scores [10]. In ATE estimation, b) deconfounding scores [57] are equivalent to generalised deconfounding scores, c) prognostic scores [36] are generalised prognostic scores, and the converse is true if $\forall a \in \mathcal{A}, Y(a) \perp\!\!\!\perp X \mid m_a(X)$. In transportability [58], d) heterogeneity sets are generalised prognostic scores while sampling and separating sets are generalised deconfounding scores.*

Thus, these "generalised" scores extend existing notions of prognostic, balancing and deconfounding scores from the literature to the more general framework from Problem 1 and connect them to the deconfounding error, refining our understanding of why these scores are well-suited for weighting.[1] Hence, for the remainder of the paper, we omit the "generalised" adjective from all notions of scores.

[1] Note that propensity scores are special cases of balancing scores [10]

3.3 Approximate deconfounding scores

Importantly, a representation ϕ need not be a perfect deconfounding score, as we can allow a certain amount of deconfounding error to obtain a small bias.

Definition 1 *Let $\epsilon \geq 0$. ϕ is called an ϵ-approximate deconfounding score if $|DE(\phi; m)| \leq \epsilon$.*

Corollary 3.1 *Let $\epsilon \geq 0$, ϕ be an ϵ-approximate deconfounding score, w be weights depending on ϕ. Then, $|Bias(w; m) - Bias(w, \phi; m)| \leq \epsilon$.*

As a consequence, if we find a representation such that its deconfounding error $\text{DE}(\phi; m)$ is small, and then we manage to find solution weights $\tilde{w}$ depending on ϕ (which is guaranteed if we minimise the objective in Theorem 2) with a small $\text{Bias}(\tilde{w}, \phi; m)$, then $\text{Bias}(\tilde{w}; m)$ will also be small, providing an approximate solution to Problem 1. Thus, the notion of ϵ-approximate scores allows us some flexibility in specifying ϕ, and $\text{DE}(\phi; m)$ measures the degree of misspecification of ϕ, in the sense of how much "relevant information for weighting" we lose. Thus we now aim at finding an approximate deconfounding score ϕ.

3.4 The balancing score error

In our setting, we do not have access to the outcome model m. It is further common to assume the absence of any information about pseudo-outcomes $\tilde{Y}$ during the weighting step [59], making any estimation of m impossible. Thus, we isolate the impact of ϕ not being a balancing score on $\text{DE}(\phi; m)$ from the outcome model m. To do so, we define the *balancing score error* (BSE) of ϕ, $\text{BSE}(\phi) := \text{RMSE}_P(w^*(X), w^*_\phi(\phi(X)))$ where $\text{RMSE}_P(A, B) := ||A - B||_{L^2(P)}$, and note that it can help bound the deconfounding error, thus the bias, as $\text{DE}(\phi; m) \leq ||m||_{L^2(P)} \cdot \text{BSE}(\phi)$. This gives a further bound on the bias.

Corollary 3.2 *Let m be an outcome model, M such that*[2] *$||m||_{L^2(P)} \leq M$. Then for any function class $\mathcal{M}$ such that $m \in \mathcal{M}$, any representation ϕ, any weights w depending on $\phi(x)$,*

$$|Bias(w; m)| \leq IPM_{\phi(\mathcal{M})}(P_w(\phi(X)), Q(\phi(X))) + M \cdot BSE(\phi).$$

Thus, the balancing score error is significant as it measures how much ϕ is not a balancing score *while giving guarantees on the deconfounding error, thus the bias*. Unlike the IPM, it provides a bound that *does not* require knowledge on the the outcome model m. This gives a clear objective to minimise when we aim at learning representations. Notably, models for propensity scores can be learnt by optimising for it instead of prediction of the treatment or the selection indicator [11].

3.5 Using the balancing score error to learn representations

A key bottleneck remains for the balancing score error : we do not have access to the true weights $w^*(X)$ or their projection $w^*_\phi(\phi(X))$. A workaround consists in first removing the projection by using the definition of a conditional expectation, as for any function g on the image space of ϕ,

$$\text{BSE}(\phi) \leq \text{RMSE}_P(w^*(X), g(\phi(X))).$$

In particular, for a given $\epsilon > 0$, if there exists *any* function g on the image space of ϕ such that $\text{RMSE}_P(w^*(X), g(\Phi(X))) < \epsilon/M$ with the M from Corollary 3.2, then ϕ is an ϵ-approximate deconfounding score. This gives us more flexibility than working the true projection of w^*, and motivates finding an g (and ϕ) minimising the RHS. However we still do not have access to w^* there either. The next result taken from the AutoDML literature helps us remove w^* from the minimisation.

Lemma 1 *(AutoDML loss [16, 12, 24]) For any function v, $RMSE_P(w^*(X), v(X))^2$ is equal to $\mathcal{L}_{P,Q}(v)$ up to an additive constant wrt v, where $\mathcal{L}_{P,Q}(v) := \mathbb{E}_P[v(X)^2] - 2 \cdot \mathbb{E}_Q[v(X)]$.*

In particular, $\mathcal{L}_{P,Q}(v)$ can be estimated in finite samples for any known v. This motivates an approach to **learn a representation** ϕ : one can posit a parameterised representation $\phi(x; \theta_\phi)$ belonging to some space Φ and a scalar parameterised function $g(.; \theta_g)$ on the Φ space, and minimise

[2] M can be 1 when $\tilde{Y}$ is binary, or $||\tilde{Y}||_{L^2(P)}$ from Jensen's inequality

$\mathcal{L}_{P,Q}(g(\phi(.;\theta_\phi);\theta_g))$ wrt θ_ϕ, θ_g. Notably, due to the compositionality of neural networks, we use a modified version of the Riesz representer component of RieszNet [12] where a pre-specified, potentially low-dimensional hidden layer is later used as the representation ϕ [60]. More generally, as a second **weighting step**, ϕ is then plugged into a generic integral probability metric $\text{IPM}_{\mathcal{G}}(P_w(\phi(X)), Q(\phi(X)))$ and solution weights $\tilde{w}$ are obtained by solving the problem in Theorem 2 with this IPM.

We note that a potential alternative method for learning ϕ consists in plugging-in a flexible density ratio estimator $v(X)$ of w^* [61] and again learning a parameterised model $g(\phi(.;\theta_\phi);\theta_g)$ for v, with a representation module ϕ, as $\text{BSE}(\phi) \leq \text{RMSE}_P(w^*(X), v(X)) + \text{RMSE}_P(v(X), v_\phi(\phi(X)))$. In practice we observed better performance with the AutoDML loss. We give an application of this loss to selecting between two representations in Supplement 8.

3.6 Extension to simultaneous weightings

In ATE estimation, one aims at estimating all $\mu(a) := \mathbb{E}[Y(a)]$ for all $a \in \mathcal{A}$ simultaneously. This can be done [62] by finding a function $f(a)$ minimising the error $\mathbb{E}_A[(\mu(A) - f(A))^2]$ over functions f defined by $f(a) = \mathbb{E}[w_a(X)\mathbb{E}[Y|X, A = a] \mid A = a]$ where $w_a(X)$ is a weight function. This is a special case of minimising a *joint squared bias*

$$\text{Bias}^2_{P^\Lambda,Q^\Lambda,p_\Lambda}(w^\Lambda; m^\Lambda) := \mathbb{E}_{p_\Lambda(\alpha)}[\text{Bias}^2_{P^\alpha,Q^\alpha}(w^\alpha; m^\alpha)]$$

where α belongs to a set Λ endowed with a probability distribution $p_\Lambda(\alpha)$, $h^\Lambda := (h^\alpha)_{\alpha\in\Lambda}$ for any h, and for each $\alpha \in \Lambda$, $\text{Bias}_{P^\alpha,Q^\alpha}(w^\alpha; m^\alpha)$ is the bias for Problem 1 with source distribution P^α, target distribution Q^α, weight function w^α, outcome model m^α. Notably, for ATE estimation, $\Lambda = \mathcal{A}$, each α is a treatment value $a \in \mathcal{A}$, and for each $a \in \mathcal{A}$, $P^a(X) = P(X|a)$, $Q^a(X) = P(X)$, $m^a(x) = \mathbb{E}[Y|A = a, X = x]$. Single-weighting problems like ATT estimation and transportability can still be encompassed by this simultaneous weightings framework by taking a unit set Λ.

One can wonder if we can learn a family of representations $\phi^\Lambda := (\phi^\alpha)_{\alpha\in\Lambda}$ where each representation ϕ^α is suitable for its corresponding weighting problem. Assuming that each the $L_2(P^\alpha)$-norm of each m^α is below M, it follows from Corollary 3.2 that for any ϕ^Λ,

$$\frac{1}{2} \cdot \text{Bias}^2_{P^\Lambda,Q^\Lambda}(w^\Lambda; m^\Lambda) \leq \mathbb{E}_{p_\Lambda(\alpha)}\Big[\text{IPM}^2_{\phi^\alpha(\mathcal{M}^\alpha;P^\alpha)}(P^\alpha_{w^\alpha}(\phi^\alpha(X)), Q^\alpha(\phi^\alpha(X)))\Big] \\ + M^2 \cdot \text{BSE}^2_{P^\Lambda,Q^\Lambda,p_\Lambda}(\phi^\Lambda).$$

where $\text{BSE}^2_{P^\Lambda,Q^\Lambda,p_\Lambda}(\phi^\Lambda) := \mathbb{E}_{p_\Lambda(\alpha)}[\text{BSE}^2_{P^\alpha,Q^\alpha}(\phi^\alpha)]$ is a *joint squared balancing score error*, that can be thought as a weighted average of the individual problems' balancing score errors $\text{BSE}_{P^\alpha,Q^\alpha}(\phi^\alpha)$. Consequently, the approach to learn representations described in Section 3.5 can be immediately extended by averaging individual AutoDML losses, and minimising the average. This will be equivalent to solving each problem separately if parameters on each problem are variationally independent, however some structure can be shared, e.g. by having a common representation $\phi^\alpha = \phi$.

4 Numerical results

We now evaluate our method and alternatives on the News dataset [63] for ATE estimation and a Traumatic Brain Injury (TBI) dataset [7] for transportability.

For the specific IPM, we focus on the *energy distance*, which is the MMD for the kernel $k(x, x') = -||x - x'||_2$; minimising Equation 1 with this distance is known as *energy balancing* [13]. We evaluate energy balancing with original covariates ("Energy") as in [13], a representation learned according to our approach ("Ours + Energy"), PCA ("PCA + Energy"), the propensity score model vector ($(\hat{p}(a|x))_{a\in\mathcal{A}}$ for ATE estimation, $(\hat{p}(s|x))_{s=0,1}$ for transportability) learnt with a gradient boosting classifier ("PS + Energy"), representations from a layer of a neural network model of such propensity score models as in [60] ("NSM + Energy").

We also check IPW with the same propensity scores ("IPW"), entropy balancing with first-order moments ("Entropy"), the weights at the head of the RieszNet ("RieszNet Head"), and uniform weights ("Unweighted"). Weights from "IPW" and "RieszNet Head" were normalised to prevent outsize errors, while those from other methods were already normalised by design.

On energy balancing methods, we take $\sigma = 0$ as in [13]. All representations are 10-dimensional, and in our method ("Ours + Energy"), we use a common representation ϕ for all treatment arms. The neural network first has a 200-unit layer, a 10-unit layer corresponding to the representation, a second 200-unit layer, and finally the scalar head. Adam [64] was used to optimise the loss with early stopping with a patience of 3 epochs. We average results over 50 random seeds for News and 100 for TBI. We show standard errors after "±." As a metric, we consider the joint bias (JB),

$$\sqrt{\sum_{\alpha \in \Lambda} p_\Lambda(\alpha) \left(\frac{\sum_{i \in \mathcal{P}^\alpha} w^\alpha(x_i)\tilde{y}_i}{|\mathcal{P}^\alpha|} - \frac{\sum_{i \in \mathcal{Q}^\alpha} m^\alpha(x_i)}{|\mathcal{Q}^\alpha|} \right)^2},$$

which is the square-root of the joint squared bias where we have replaced the target estimand with a finite-sample estimator of it where we average the known outcome models over $\mathcal{Q}^\alpha$. We also look at computational time in seconds to evaluate the speed of algorithms.

Table 1: Joint Bias and time in seconds on the News and TBI Datasets

	Joint Bias		Time (s)	
Dataset	News	TBI	News	TBI
Ours + Energy	0.078±0.009	4.97±0.35	29.7±1.2	0.844±0.028
NSM + Energy	0.076±0.007	5.77±0.52	24.0±0.3	0.988±0.027
PS + Energy	0.118±0.009	14.33±1.11	53.1±0.1	0.970±0.007
PCA + Energy	0.229±0.016	13.72±1.14	14.8±0.1	1.044±0.009
Energy	0.292±0.019	14.01±1.11	645.4±0.5	0.896±0.009
Entropy	0.221±0.020	7.63±0.60	3366.1±62.8	0.241±0.005
IPW	0.280±0.018	2.28±0.18	40.8±0.1	0.254±0.000
RieszNet Head	0.746±0.122	59.91±2.49	3.2±0.5	0.222±0.002
Unweighted	0.611±0.053	7.67±0.15	0.34±0.03	0.050±0.000

Results on the joint bias are shown in Table 1. Our neural network-based representation ranks in the top 2 of methods for both datasets and outperforms most other representations as well as the head of the neural network, which shows the benefit of both our representation and plugging it into a distance instead of taking the neural network head as solution weights. Further, looking at the computational time, energy balancing with the original covariates is particularly long on the News dataset, while using a lower-dimensional representation makes it faster, notably more so than gradient boosting-based IPW. Despite neural network training, weighting with our representation only takes twice as long as with PCA, and is faster than with gradient boosting propensity score models. On the TBI dataset, our method is also faster than other representation-based methods.

5 Conclusion

We have shown the importance of a specific quantity in learning representations for weighting, the deconfounding error. From it, we have redefined classical notions of balancing, prognostic and deconfounding scores, and introduced an approximate version of deconfounding scores. We have also defined the balancing score error (BSE) that measures how far a representation is from a balancing score while bounding the deconfounding error, and does not depend on outcome information. We have outlined a method to minimise it, and first experimental results suggest that representations obtained from the method might help improve performance and speed computations for common optimisation-based weighting approaches. We note two key challenges for future work : 1) the errors depend on quantities we do not have access to, 2) learning a representation will also induce a new function class that we still have to characterise, e.g. based on an original posited class.

Acknowledgements

We sincerely thank David Bruns-Smith for valuable feedback. O.C. was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). A.F. was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. C.H. was supported by the EPSRC Bayes4Health programme grant and The Alan Turing Institute, UK.

References

[1] Daniel Westreich, Jessie K Edwards, Catherine R Lesko, Elizabeth Stuart, and Stephen R Cole. Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology*, 186(8):1010–1014, 2017.

[2] Paul R Rosenbaum. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21(1):57–71, 2012.

[3] Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy learning with asymmetric utilities. *arXiv preprint arXiv:2206.10479*, 2022.

[4] Jasjeet Singh Sekhon and Richard D Grieve. A matching method for improving covariate balance in cost-effectiveness analyses. *Health economics*, 21(6):695–714, 2012.

[5] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[6] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

[7] Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Reweighting the rct for generalization: finite sample error and variable selection. 2022.

[8] Eli Ben-Michael, Avi Feller, David A Hirshberg, and José R Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021.

[9] Leonard Wainstein. Targeted function balancing. 2022.

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

[11] Shomoita Alam, Erica E. M. Moodie, and David A. Stephens. Should a propensity score model be super? the utility of ensemble procedures for causal adjustment. *Statistics in Medicine*, 38(9):1690–1702, 2019.

[12] Victor Chernozhukov, Whitney Newey, Vıctor M Quintas-Martınez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022.

[13] Jared D Huling and Simon Mak. Energy balancing of covariate distributions. *arXiv preprint arXiv:2004.13962*, 2020.

[14] David A Bruns-Smith and Avi Feller. Outcome assumptions and duality theory for balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 11037–11055. PMLR, 2022.

[15] David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

[16] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

[17] 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.

[18] Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

[19] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

[20] Elizabeth Tipton. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266, 2013.

[21] Colm O'Muircheartaigh and Larry V Hedges. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2):195–210, 2014.

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

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

[24] Michael Newey and Whitney K Newey. Automatic debiased machine learning for covariate shifts. *arXiv preprint arXiv:2307.04527*, 2023.

[25] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

[26] Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust, 2016.

[27] Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

[28] Stefan Tübbicke. Entropy balancing for continuous treatments. *Journal of Econometric Methods*, 11(1):71–89, 2021.

[29] Brian G Vegetabile, Beth Ann Griffin, Donna L Coffman, Matthew Cefalu, Michael W Robbins, and Daniel F McCaffrey. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology*, 21(1):69–110, 2021.

[30] Kevin P Josey, Seth A Berkowitz, Debashis Ghosh, and Sridharan Raghavan. Transporting experimental results with entropy balancing. *Statistics in medicine*, 40(19):4310–4326, 2021.

[31] Juan Chen and Yingchun Zhou. Causal effect estimation for multivariate continuous treatments. *arXiv preprint arXiv:2205.08730*, 2022.

[32] Ambarish Chattopadhyay, Eric R Cohn, and Jose R Zubizarreta. One-step weighting to generalize and transport treatment effect estimates to a target population. *arXiv preprint arXiv:2203.08701*, 2022.

[33] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 243–263, 2014.

[34] Nathan Kallus. Generalized optimal matching methods for causal inference. *The Journal of Machine Learning Research*, 21(1):2300–2353, 2020.

[35] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. 2012.

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

[37] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

[38] Xiao Shou, Tian Gao, Dharmashankar Subramanian, and Kristin P Bennett. Match2: hybrid self-organizing map and deep learning strategies for treatment effect estimation. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10, 2021.

[39] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.

[40] Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. Neural score matching for high-dimensional causal inference. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7076–7110. PMLR, 28–30 Mar 2022.

[41] Bing Xue, Ahmed Sameh Said, Ziqi Xu, Hanyang Liu, Neel Shah, Hanqing Yang, Philip Payne, and Chenyang Lu. Assisting clinical decisions for scarcely available treatment via disentangled latent representation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5360–5371, 2023.

[42] Yang Ning, Peng Sida, and Kosuke Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020.

[43] 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.

[44] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[45] R. M. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40(1):40 – 50, 1969.

[46] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.

[47] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[48] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

[49] David Nelson and Siamak Noorbaloochi. Information preserving sufficient summaries for dimension reduction. *Journal of multivariate analysis*, 115:347–358, 2013.

[50] Ming-Yueh Huang and Kwun Chuen Gary Chan. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, 104(3):583–596, 2017.

[51] Shujie Ma, Liping Zhu, Zhiwei Zhang, Chih-Ling Tsai, and Raymond J Carroll. A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *Annals of statistics*, 47(3):1505, 2019.

[52] Trinetri Ghosh, Yanyuan Ma, and Xavier De Luna. Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Statistica Sinica*, 31(2):821, 2021.

[53] Debo Cheng, Jiuyong Li, Lin Liu, Thuc Duy Le, Jixue Liu, and Kui Yu. Sufficient dimension reduction for average causal effect estimation. *Data Mining and Knowledge Discovery*, 36(3):1174–1196, 2022.

[54] P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B: Methodological*, 45:212–218, 1983.

[55] Pengzhou Wu and Kenji Fukumizu. β-intact-vae: Identifying and estimating causal effects under limited overlap. *arXiv preprint arXiv:2110.05225*, 2021.

[56] 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.

[57] Alexander D'Amour and Alexander Franks. Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*, 2021.

[58] Naoki Egami and Erin Hartman. Covariate selection for generalizing experimental results: application to a large-scale development program in uganda. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4):1524–1548, 2021.

[59] Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010.

[60] Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. Neural score matching for high-dimensional causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 7076–7110. PMLR, 2022.

[61] David Arbour, Drew Dimmery, and Arjun Sondhi. Permutation weighting. In *International Conference on Machine Learning*, pages 331–341. PMLR, 2021.

[62] Guillaume Martinet. A balancing weight framework for estimating the causal effect of general treatments. *arXiv preprint arXiv:2002.11276*, 2020.

[63] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[64] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[65] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR, 2019.

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

[67] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019.

[68] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[69] 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.

[70] Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, and Jianhua Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35:37704–37718, 2022.

[71] Xin-Qiang Cai, Yao-Xiang Ding, Zi-Xuan Chen, Yuan Jiang, Masashi Sugiyama, and Zhi-Hua Zhou. Seeing differently, acting similarly: Heterogeneously observable imitation learning. *arXiv preprint arXiv:2106.09256*, 2021.

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

[73] Thorsten Joachims, Adith Swaminathan, and Maarten De Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

[74] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[75] Chester Holtz, Tsui-Wei Weng, and Gal Mishne. Learning sample reweighting for accuracy and adversarial robustness. *arXiv preprint arXiv:2210.11513*, 2022.

Appendices

6 Details on problems in causal inference

Under the assumptions of *no interference* and *consistency*, $A = a$ implies $Y = Y(a)$, which can written as $Y = \sum_{a \in \mathcal{A}} 1_{\{A=a\}} Y(a)$ or, more compactly, $Y = Y(A)$. Further, under *unconfoundedness* and *overlap* we have that $\mathbb{E}[Y(a)|X] = \mathbb{E}[Y|A = a, X]$, helping identify causal effects of interest which we detail below.

In **ATT estimation** [8], we are interested in the effect of a binary treatment on the population receiving it, that is $\mathbb{E}[Y(1) - Y(0)|A = 1]$. Thanks to consistency and no interference, $\mathbb{E}[Y(1)|A = 1]$ is accessible as the average of outcomes on the treated distribution, so the challenging part is estimating $\mathbb{E}[Y(0)|A = 1]$. The weighting approach is then to reweight the control distribution, on which $Y(0) = Y$, that is to find a function $w(x)$ such that $\mathbb{E}[Y(0)|A = 1] = \mathbb{E}[w(X)Y|A = 0] \approx \frac{1}{\{i:A_i=0\}} \sum_{i:A_i=0} w(X_i)Y_i$.

In **average potential outcome estimation** [13], for a fixed $a \in \mathcal{A}$, we are interested in the marginal effect of the potential outcome wrt a, that is $\mathbb{E}[Y(a)]$. The weighting approach is then to reweight the distribution of the population for which $A = a$, implying $Y(a) = Y$, i.e. find a function $w_a(x)$ such that $\mathbb{E}[Y(a)] = \mathbb{E}[w_a(X)Y|A = a] \approx \frac{1}{\{i:A_i=a\}} \sum_{i:A_i=a} w_a(X_i)Y_i$. We note that the closely related goal of **ATE estimation**, that is when $\mathcal{A}$ is binary and we want $\mathbb{E}[Y(1) - Y(0)]$, can be solved by average potential outcome estimation for both $a = 1$ and $a = 0$ separately. With some abuse of notation, we use the two names of average potential outcome estimation and ATE estimation interchangeably.

In **transportability** [7], A is binary again and we have an other binary variable S such that $S = 1$ denotes membership in the RCT population, that is $(Y(1), Y(0)) \perp\!\!\!\perp A|S = 1$. We do not have access to A, Y for the target, non-RCT population $S = 0$, but we are still interested in the treatment effect on the target population $\mathbb{E}[Y(1) - Y(0)|S = 0]$. Under the *transportability* assumption, that is $Y(1) - Y(0) \perp\!\!\!\perp S|X$ [7, 6], the *conditional average treatment effect* is identical between RCT and non-RCT populations, i.e. for any x, CATE$(x) := \mathbb{E}[Y(1) - Y(0)|X = x]$ is equal to both $\mathbb{E}[Y(1) - Y(0)|X = x, S = 1]$ and $\mathbb{E}[Y(1) - Y(0)|X = x, S = 0]$. In addition, the CATE is identified on the RCT population as CATE$(x) = \mathbb{E}[\frac{AY}{P(A=1)} - \frac{(1-A)Y}{P(A=0)}|X = x, S = 1]$. Then, the weighting approach is to reweight the distribution of the RCT population, i.e. find a function $w(x)$ such that $\mathbb{E}[Y(1) - Y(0)|S = 0] = \mathbb{E}[w(X) \cdot \text{CATE}(X)|S = 1] \approx \frac{1}{|\{i:S_i=1\}|} \sum_{i:S_i=1} w(X_i)\big(\frac{A_iY_i}{P(A=1)} - \frac{(1-A_i)Y}{P(A=0)}\big)$.

7 Related work : importance weighting

Outside causal inference, the framework in Problem 1 is also close to *importance weighting* [65], where we want to estimate *and minimise* $\mathbb{E}_P[w^*(X)l(X;\theta)] = \mathbb{E}_Q[l(X;\theta)]$ wrt $\theta \in \Theta$, where this time $l(x;\theta)$ is a known parameterised loss function. Classically, solution weights $\tilde{w}(x)$ are first estimated with general techniques like kernel mean matching [66] (i.e. Equation 1 with an RKHS and $\sigma = 0$), or techniques that are more domain-specific i.e. to label shift [67]. Then the focus shifts to minimising $\mathbb{E}_P[\tilde{w}(x)l(x, \theta)]$ wrt θ to solve the task under scrutiny [68]. Examples of tasks include domain adaptation [69], subpopulation shift [70], imitation learning [71], off-policy reinforcement learning [72, 73], variational inference [74], or adversarial robustness [75]. We note that any technique from Section 2.3, including propensity score estimation that is generally a form of density ratio estimation [61], as well as our method, could be applied to the weight estimation part of importance weighting.

8 Representation selection

Further, to **select between two representations** ϕ_1 and ϕ_2, one can choose the representation with the lowest BSE. This is equivalent to compare $\min_{g_1} \mathcal{L}_{P,Q}(g_1(\phi_1(.))$ and $\min_{g_2} \mathcal{L}_{P,Q}(g_2(\phi_2(.))$, where each minimisation is taken over all functions. These are inaccessible, but we can instead perform each minimisation under a rich parameterised class of functions. Particularly, this would help select between two fitted propensity score models and we expect that the one with the best prediction performance might not necessarily be selected.

Further, we note that the AutoDML loss makes us lose the ability of evaluating how approximately deconfounding is *one* representation, instead of comparing different representations. Flexible density ratio estimators could be plugged into the balancing score error, especially as both the true weights and their expectation conditional on the representation are density ratios from Assumption 1 and Theorem 1.

9 Proof of results

9.1 Proof of Theorem 1

Let w be weights and ϕ be a representation. From the tower property, for any distribution R and function f,

$$\mathbb{E}_R[f_\phi^R(\phi(X))] = \mathbb{E}_R[f(X)].$$

Thus, we note that the bias can be decomposed as

$$\text{Bias}(w;m) = \text{Bias}(w,\phi;m) + \text{CE}(w,\phi;m) + \text{DE}(\phi;m)$$

where

$$\begin{aligned}\text{Bias}(w,\phi;m) &:= \mathbb{E}_{P_w}[m_\phi(\phi(X))] - \mathbb{E}_Q[m_\phi(\phi(X))],\\ \text{CE}(w,\phi;m) &:= \mathbb{E}_{P_w}[m_\phi^{P_w}(\phi(X)) - m_\phi(\phi(X))],\\ \text{DE}(\phi;m) &:= \mathbb{E}_Q[m_\phi(\phi(X)) - m_\phi^Q(\phi(X))] = -\text{CE}(w^*,\phi;m).\end{aligned}$$

Thus the proof relies in showing that for any w,

$$\mathbb{E}_{P_w}[m_\phi^{P_w}(\phi(X)) - m_\phi(\phi(X))] = \mathbb{E}_P[(m(X) - m_\phi(\phi(X)))\cdot(w(X) - w_\phi(\phi(X)))]$$

and $w_\phi(\phi(x)) = \frac{p_w(\phi(x))}{p(\phi(x))}$. For any x, x' and w,

$$p_w(x|\phi(x')) = \frac{p_w(x,\phi(x'))}{p_w(\phi(x'))} = 1_{\{\phi(x)=\phi(x')\}}\frac{p_w(x)}{p_w(\phi(x'))},$$

thus for any x and w,

$$\begin{aligned}w_\phi(\phi(x)) &= \mathbb{E}_P[w(X)|\phi(X)=\phi(x)]\\ &= \int w(x')p(x'|\phi(x))dx'\\ &= \int w(x')\cdot 1_{\{\phi(x)=\phi(x')\}}\frac{p(x')}{p_w(\phi(x))}dx'\\ &= \frac{\int 1_{\{\phi(x)=\phi(x')\}}w(x')p(x')dx'}{p(\phi(x))}\\ &= \frac{\int 1_{\{\phi(x)=\phi(x')\}}p_w(x')dx'}{p(\phi(x))}\\ &= \frac{p_w(\phi(x))}{p(\phi(x))}.\end{aligned}$$

Thus, for any x, x' such that $\phi(x) = \phi(x')$ and w,

$$\frac{p_w(x'|\phi(x))}{p(x'|\phi(x))} = \frac{p_w(x')/p_w(\phi(x))}{p(x')/p(\phi(x))} = \frac{p_w(x')/p(x')}{p_w(\phi(x))/p(\phi(x))} = \frac{w(x')}{w_\phi(\phi(x))}.$$

As a consequence,

$$\begin{aligned}m_\phi^{P_w}(\phi(X)) &= \mathbb{E}_{P_w}[m(X)|\phi(X)]\\ &= \mathbb{E}_P\Big[\frac{p_w(X|\phi(X))}{p(X|\phi(X))}m(X)|\phi(X)\Big]\\ &= \mathbb{E}_P\Big[\frac{w(X)}{w_\phi(\phi(X))}m(X)|\phi(X)\Big]\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{P_w}[m_\phi^{P_w}(\phi(X)) - m_\phi(\phi(X))] &= \mathbb{E}_{P_w}\Big[\mathbb{E}_P\Big[\Big(\frac{w(X)}{w_\phi(\phi(X))} - 1\Big)m(X)\Big|\phi(X)\Big]\Big] \\
&= \mathbb{E}_P\Big[w(X)\mathbb{E}_P\Big[\Big(\frac{w(X)}{w_\phi(\phi(X))} - 1\Big)m(X)\Big|\phi(X)\Big]\Big] \\
&= \mathbb{E}_P\Big[w_\phi(\phi(X)) \cdot \mathbb{E}_P\Big[\Big(\frac{w(X)}{w_\phi(\phi(X))} - 1\Big)m(X)\Big|\phi(X)\Big]\Big] \text{ from the tower property} \\
&= \mathbb{E}_P\Big[\mathbb{E}_P\Big[w_\phi(\phi(X)) \cdot \Big(\frac{w(X)}{w_\phi(\phi(X))} - 1\Big)m(X)\Big|\phi(X)\Big]\Big] \\
&= \mathbb{E}_P\Big[\mathbb{E}_P\Big[\Big(w(X) - w_\phi(\phi(X))\Big)m(X)\Big|\phi(X)\Big]\Big] \\
&= \mathbb{E}_P\Big[\Big(w(X) - w_\phi(\phi(X))\Big)m(X)\Big|\Big] \text{ from the tower property.}
\end{aligned}$$

Further, from the tower property, this RHS is actually zero for any m depending on $\phi(x)$, in particular $m_\phi(\phi(x))$. Thus,

$$\mathbb{E}_{P_w}[m_\phi^{P_w}(\phi(X)) - m_\phi(\phi(X))] = \mathbb{E}_P\Big[\Big(w(X) - w_\phi(\phi(X))\Big) \cdot \Big(m(X) - m_\phi(\phi(X))\Big)\Big]$$

which concludes the proof.

9.2 Proof of Theorem 2

First, for any function $g \in \mathcal{G}$, we have

$$\begin{aligned}
\Big|\mathbb{E}_{P_w}[(g \circ \phi)(X)] - \mathbb{E}_Q[(g \circ \phi)(X)]\Big| &= \Big|\mathbb{E}_{P_w}[g(\phi(X))] - \mathbb{E}_Q[g(\phi(X))]\Big| \\
&= \Big|\mathbb{E}_{\varphi \sim P_w(\phi(X))}[g(\varphi)] - \mathbb{E}_{\varphi \sim Q(\phi(X))}[g(\varphi)]\Big|.
\end{aligned}$$

Taking the supremum over $g \in \mathcal{G}$, we have

$$\text{IPM}_{\mathcal{M}}(P_w(X), Q(X)) = \text{IPM}_{\mathcal{G}}(P_w(\phi(X)), Q(\phi(X)))$$

where $\mathcal{M} = \{x \to (g \circ \phi)(x),\ g \in \mathcal{G}\}$. Note that this also justifies the claim that Bias$(w, \phi; m)$ is bounded by IPM$_{\mathcal{G}}(P_w(\phi(X)), Q(\phi(X)))$. Thus, we are solving

$$\min_w \text{IPM}_{\mathcal{M}}(P_w(X), Q(X))^2 + \sigma^2 \cdot ||w||^2_{L_2(P)}.$$

From Bruns-Smith et al. (2022) [14] (Section 3.2 then Theorem 4.3), the solution $\tilde{w}$ depends affinely on a function $\tilde{m}$ in $\mathcal{M}$: there exists $s_1, s_2 \in \mathbb{R}$ such that $\tilde{w}(x) = s_1 + s_2\tilde{m}(x)$. In particular, as $\tilde{m}(x)$ depends on $\phi(x)$ then so does $\tilde{w}(x)$.

9.3 Proof of Theorem 3

First, let's note two useful properties :

- For any distribution R and random variable Z,
$$\mathbb{E}_R[\mathbb{E}_R[Z|X] \mid \phi(X) = \phi(x)] = \mathbb{E}_R[Z|\phi(X) = \phi(x)]. \tag{S2}$$
- For any distribution R and function f,
$$\Big(\exists g,\ f(x) = g(\phi(x))\Big) \Leftrightarrow f(x) = \mathbb{E}_R[f(X) \mid \phi(X) = \phi(x)]. \tag{S3}$$

Proof of a), ATT case :

$$\begin{aligned}
&\phi \text{ is a balancing score} \\
&\Leftrightarrow \exists g,\ \ e(x) = g(\phi(x)) \text{ from Rosenbaum and Rubin (1983) [10]} \\
&\Leftrightarrow \exists g,\ \ w^*(x) = g(\phi(x)) \text{ as } w^*(x) \text{ is a bijective function of } e(x) \\
&\Leftrightarrow \phi(x) \text{ is a generalised balancing score.}
\end{aligned}$$

Proof of a), ATE case : we fix $a \in \mathcal{A}$ and work with the following definition of a balancing score for non-binary treatments : $1_{\{A=a\}} \perp\!\!\!\perp X|\phi(X)$. Indeed, as the problem is arm-specific, the definitions of generalised deconfounding, balancing and prognostic scores are arm-specific *a priori*. An extension to an alternative definition $A \perp\!\!\!\perp X|\phi(X)$ is straightforward by replacing a fixed $a \in \mathcal{A}$ with $\forall a \in \mathcal{A}$ at the start of each of the following statements involving a. Then,

$$\begin{aligned}
&\phi \text{ is a balancing score}\\
&\Leftrightarrow p(a|x) = p(a|\phi(x))\\
&\Leftrightarrow p(a|x) = \mathbb{E}[p(a|X)|\phi(X) = \phi(x)] \text{ using S2 with } Z = 1_{\{A=a\}}\\
&\Leftrightarrow \exists g_a,\ \ p(a|x) = g_a(\phi(x)) \text{ from S3}\\
&\Leftrightarrow \exists g_a,\ \ w_a^*(x) = g_a(\phi(x)) \text{ where } w_a^*(x) \text{ is the true weights and is a bijective function of } p(a|x)\\
&\Leftrightarrow \phi(x) \text{ is a generalised balancing score.}
\end{aligned}$$

Proof of b) : we slightly change the definition of deconfounding scores [57] to $\forall a \in \mathcal{A},\ \mathbb{E}[\mathbb{E}[Y|\phi(X), A = a]] = \mathbb{E}[Y(a)]$, where the representation ϕ is now shared across treatment arms, in the spirit of D'Amour and Franks (2021)[57]. To this aim, it is sufficient to show that, in Problem 1 applied to estimation of $\mathbb{E}[Y(a)]$, DE$(\phi; m) = \mathbb{E}[\mathbb{E}[Y|\phi(X), A = a]] - \mathbb{E}[Y(a)]$. From the original definition of DE$(\phi; m)$ in the proof of Theorem 1, this simplifies to $\mathbb{E}\Big[\mathbb{E}[m_a(X)|\phi(X), a]\Big] = \mathbb{E}[\mathbb{E}[Y|\phi(X), A = a]]$. This is true, as

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}[m_a(X)|\phi(X), a]\Big]\\
&= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{E}[Y|X, A = a]\Big|\phi(X), a\Big]\Big]\\
&= \mathbb{E}\Big[\mathbb{E}[Y|\phi(X), A = a]\Big] \text{ from S2 applied to } R = P(.|A = a) \text{ and } Z = Y,
\end{aligned}$$

which concludes the proof.

Proof of c) : again, $a \in \mathcal{A}$ is fixed. Assume $\phi(x)$ is a prognostic score for $Y(a)$, that is $Y(a) \perp\!\!\!\perp X|\phi(X)$. Then,

$$\begin{aligned}
m_a(x) &:= \mathbb{E}[Y(a)|x]\\
&= \mathbb{E}[Y(a)|x, \phi(x)]\\
&= \mathbb{E}[Y(a)|\phi(x)] \text{ by application of the definition of a prognostic score,}
\end{aligned}$$

so $m_a(x)$ is a function $\phi(x)$, making the latter a generalised prognostic score.

Now assume that $m_a(X)$ itself is a prognostic score, that is $Y(a) \perp\!\!\!\perp X|m_a(X)$. Then, $p(Y(a)|x) = p(Y(a)|m_a(x))$. Let $\phi(X)$ be a generalised prognostic score. Then, there exists a function g_a such that $m_a(x) = g_a(\phi(x))$. In particular, as $p(Y(a)|x)$ is already a function of $m_a(x)$, it is also a function of $\phi(x)$. So there exists a function h_a such that $p(Y(a)|x) = h_a(\phi(x))$. In particular, by application of S3, $p(Y(a)|x) = \mathbb{E}[p(Y(a)|X)|\phi(X) = \phi(x)]$ and by application of S2 to $Z = 1_{\{Y(a)\}}$, $p(Y(a)|x) = p(Y(a)|\phi(x))$. Thus, $\phi(x)$ is a prognostic score.

Proof of d) : let $X_{\mathcal{I}}$ be covariates selected according to indices $\mathcal{I}$ and $X_{-\mathcal{I}}$ be their complement.

If $x_{\mathcal{I}}$ is a heterogeneity set, i.e. $Y(1) - Y(0) \perp\!\!\!\perp (S, X_{-\mathcal{I}})|X_{\mathcal{I}}$ then

$$\begin{aligned}
m(x) &= \text{CATE}(x)\\
&= \mathbb{E}[Y(1) - Y(0)|x]\\
&= \mathbb{E}[Y(1) - Y(0)|x_{-\mathcal{I}}, x_{\mathcal{I}}]\\
&= \mathbb{E}[Y(1) - Y(0)|x_{\mathcal{I}}] \text{ by definition of a heterogeneity set}
\end{aligned}$$

so $m(x)$ is a function of $x_{\mathcal{I}}$, making the latter a generalised prognostic score.

If $x_{\mathcal{I}}$ is a separating set, that is $Y(1) - Y(0) \perp\!\!\!\perp S|X_{\mathcal{I}}$, then, noting $e(Z) := P(S = 1|Z)$, from the law of total covariance,

$$\begin{aligned}
\text{Cov}(m(X), e(X)|X_{\mathcal{I}}) &= \text{Cov}(Y(1) - Y(0), S|X_{\mathcal{I}}) - \mathbb{E}[\text{Cov}(Y(1) - Y(0), S|X)|X_{\mathcal{I}}]\\
&= 0 \qquad \text{(S4)}
\end{aligned}$$

as the second term of the sum is zero by the transportability assumption (see Section 6), and the first term is also zero as $X_{\mathcal{I}}$ is a separating set. Then, noting $\phi(x) = x_{\mathcal{I}}$

$$\begin{aligned}
\text{DE}(x_{\mathcal{I}}) &:= -\mathbb{E}\Big[\Big(m(X) - m_\phi(X_{\mathcal{I}})\Big) \cdot \Big(w^*(X) - w^*_\phi(X_{\mathcal{I}})\Big)\Big|S = 1\Big] \\
&:= -\mathbb{E}\Big[w^*(X) \cdot \Big(m(X) - m_\phi(X_{\mathcal{I}})\Big) \cdot \Big(1 - \frac{w^*_\phi(X_{\mathcal{I}})}{w^*(X)}\Big)\Big|S = 1\Big] \\
&\qquad \text{where } \frac{w^*_\phi(x_{\mathcal{I}})}{w^*(x)} = \frac{1 - e(x_{\mathcal{I}})}{1 - e(x)} \frac{e(x)}{e(x_{\mathcal{I}})} \\
&= -\mathbb{E}\Big[\Big(m(X) - m_\phi(X_{\mathcal{I}})\Big) \cdot \Big(1 - \frac{1 - e(X_{\mathcal{I}})}{1 - e(X)} \frac{e(X)}{e(X_{\mathcal{I}})}\Big)\Big|S = 0\Big] \\
&\qquad \text{as } w^*(X) \text{ is the likelihood ratio from } P(X|S = 1) \text{ to } P(X|S = 0) \\
&= \mathbb{E}\Big[\frac{\Big(m(X) - m_\phi(X_{\mathcal{I}})\Big) \cdot \Big(e(X) - e(X_{\mathcal{I}})\Big)}{(1 - e(X)) \cdot e(X_{\mathcal{I}})}\Big|S = 0\Big] \\
&= \mathbb{E}\Big[\frac{\Big(m(X) - m_\phi(X_{\mathcal{I}})\Big) \cdot \Big(e(X) - e(X_{\mathcal{I}})\Big)}{P(S = 0) \cdot e(X_{\mathcal{I}})}\Big] \\
&\qquad \text{as } \frac{P(S = 0)}{1 - e(x)} \text{ is the likelihood ratio from } P(X|S = 0) \text{ to } P(X) \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\frac{\Big(m(X) - m_\phi(X_{\mathcal{I}})\Big) \cdot \Big(e(X) - e(X_{\mathcal{I}})\Big)}{P(S = 0) \cdot e(X_{\mathcal{I}})}\Big]\Big|X_{\mathcal{I}}\Big] \text{ from the tower property} \\
&= \mathbb{E}\Big[\frac{\mathbb{E}[(m(X) - m_\phi(X_{\mathcal{I}})) \cdot (e(X) - e(X_{\mathcal{I}})) \mid X_{\mathcal{I}}]}{P(S = 0) \cdot e(X_{\mathcal{I}})}\Big] \\
&= \mathbb{E}\Big[\frac{\text{Cov}(m(X), e(X)|X_{\mathcal{I}})}{P(S = 0) \cdot e(X_{\mathcal{I}})}\Big] \\
&= 0 \text{ from } S4.
\end{aligned}$$

So $x_{\mathcal{I}}$ is a generalised deconfounding score. As sampling sets are also separating sets, the proof is concluded.

9.4 Proof of Corollary 3.1

As w depends on ϕ, $\text{CE}(w, \phi; m) = 0$ so $\text{Bias}(w; m) - \text{Bias}(w, \phi; m) = \text{DE}(\phi; m)$. As ϵ-approximate deconfounding scores verify $|\text{DE}(\phi; m)| \leq \epsilon$, the proof is concluded.

9.5 Proof of Corollary 3.2

As we have seen in the proof of Theorem 1, $\text{DE}(\phi; m) = \mathbb{E}\Big[m(X) \cdot \Big(w^*(X) - w^*_\phi(\phi(X))\Big)\Big]$ so the Cauchy-Schwarz inequality gives the bound $|\text{DE}(\phi; m)| \leq ||m||_{L_2(P)} \cdot \text{BSE}(\phi)$. Thus, from Theorem 1, for any w depending on ϕ,

$$\begin{aligned}
|\text{Bias}(w; m)| &\leq \Big|\mathbb{E}_{P_w}[m_\phi(\phi(X))] - \mathbb{E}_Q[m_\phi(\phi(X))]\Big| + |\text{DE}(\phi; m)| \text{ where } m \in \mathcal{M} \text{ so } m_\phi \in \phi(\mathcal{M}) \\
&\leq \text{IPM}_{\phi(\mathcal{M})}(P_w(\phi(X)), Q(\phi(X))) + |\text{DE}(\phi; m)| \\
&\leq \text{IPM}_{\phi(\mathcal{M})}(P_w(\phi(X)), Q(\phi(X))) + ||m||_{L_2(P)} \cdot \text{BSE}(\phi)| \text{ from the bound above} \\
&\leq \text{IPM}_{\phi(\mathcal{M})}(P_w(\phi(X)), Q(\phi(X))) + M \cdot \text{BSE}(\phi)| \text{ by assumption over } M.
\end{aligned}$$

9.6 Proof of Lemma 1

For any function v,

$$\begin{aligned}\text{RMSE}_P(w^*(X), v(X))^2 &= \mathbb{E}_P[(v(X) - w^*(X))^2] \\ &= \mathbb{E}_P[v(X)^2] - 2\mathbb{E}_P[w^*(X)v(X)] + \mathbb{E}_P[w^*(X)^2] \\ &= \mathbb{E}_P[v(X)^2] - 2\mathbb{E}_Q[v(X)] + \mathbb{E}_P[w^*(X)^2]\end{aligned}$$

as by definition of w^*, so for any f, $\mathbb{E}_P[w^*(X)f(X)] = \mathbb{E}_Q[f(X)]$, concluding the proof.