Towards Representation Learning for Treatment Effect Estimation with High-Dimensional Covariates

Oscar Clivio

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the intellectual contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification. My personal contributions are detailed in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text and authorship forms.

Oscar Clivio
Trinity 2025

This thesis is dedicated to
my parents and my aunt.

Acknowledgements

First, I would like to extend my deepest gratitude to my supervisor, Chris Holmes, for having introduced me to the field of causal inference, for his guidance and his patience throughout my doctoral journey, as well as his unwavering support for my various endeavours: internships, academic visits, and trips. Additionally, I am indebted to his leadership in setting up and pursuing the collaboration with Novo Nordisk, without which my PhD would not have been possible. I am also deeply grateful to Novo Nordisk for generously agreeing to fund my doctoral studies.

Further, I am deeply indebted to Avi Feller for generously receiving me as a visitor at UC Berkeley, for offering his expertise in critical portions of causal inference and for his deep involvement in two research papers of this thesis. I am also indebted to his extraordinary dedication to connecting me with other causal inference people at Berkeley, during conferences, and for a research paper.

I also thank my co-authors for their essential contributions to the papers I authored during my doctorate.

I am deeply grateful to Robin Evans for his leadership in the causal inference reading group, for offering his causal inference expertise, and for his rigorous examination of my work in all my exams: transfer, confirmation and viva. I also extend my gratitude to Uri Shalit for serving as the external examiner for my viva, as well as Judith Rousseau for assessing my confirmation and Arnaud Doucet for assessing my transfer. I thank the professors who monitored my progress, Geoff Nichols, Judith Rousseau and Garrett Morris.

I am grateful to the administrative staff of the Department of Statistics and of the StatML CDT, especially Joanna Stoneham, Frédérique Godin, Emma Bodger, Shabana Akthar, Olivier Cristini and the IT team, for their responsiveness, patience and assistance.

I am further indebted to mentors in my internships, Gorkem Ozkaya and Ben Marchi at Uber, Alexandre Drouin and Valentina Zantedeschi at ServiceNow, for having offered their mentorship, their expertise, and unforgettable experiences more generally. I am also grateful to the entire faculty of the Center for Targeted Machine Learning as well as Michael I. Jordan, Ahmed Alaa and Ioannis Mitliagkas for receiving me in their labs and group meetings.

On a more social note, I am thankful to my friends, notably at the Department of Statistics and my housemates, for great moments.

Finally, I extend my deepest gratitude to my family for their love and for helping me fully recharge my batteries whenever I was visiting back home. I am particularly indebted

to my parents for their incredible support, which was instrumental in the completion of this thesis. I dedicate this thesis to them, as well as my aunt: I hope that you are doing great wherever you are now and you watch over me.

Abstract

Treatment effect estimation is typically difficult in the presence of high-dimensional covariates. In this thesis, we propose to use a representation, that is, the mapping of covariates to a lower-dimensional manifold, as an adjustment set in estimators of treatment effects.

In an introductory chapter, we review treatment effect estimation as well as challenges with high-dimensional features, both in statistics or machine learning generally and in treatment effect estimation in particular. The thesis's contributions are then outlined.

As a first contribution, we extend the popular propensity score matching method to the use of multivariate representations in matching. This can be done by noticing that neural networks naturally satisfy the compositional nature of balancing scores from Rosenbaum and Rubin [1983], which are valid adjustment sets. As a result, intermediary hidden layers of a neural network modelling the propensity score can be used as adjustment sets. We also bound the original covariate imbalance using the representation imbalance in different settings. This method requires that the neural network model is correctly specified.

We address this in the second contribution, where we upper-bound the bias induced by adjusting for a representation instead of original covariates using a loss which builds on former work in density ratio regression and is differentiable. As a result, a measure of misspecification of the representation, which is added to the representation imbalance in the former bound, can directly be targeted. We also extend the analysis from matching to a more general weighting framework.

In a third contribution, we address the problem of poor overlap which is typical for high-dimensional covariates. We quantify the degree of overlap of a representation using a specific "overlap divergence" measure, and justify doing so by connecting it to the variance of estimators adjusting for the representation. We establish that outcome information is required to find a representation improving overlap, which stands in contrast to the two previous contributions. In a simplified setting with Gaussian covariates and generalised linear models for the outcome model and the propensity score model, we further strengthen the result: we describe a class of representations without confounding bias where the more predictive of the outcome the representation, the better its overlap.

We conclude with a summary of the contributions, key findings and limitations in all methods.

Contents

1 Introduction **1**

1.1 Fundamentals of treatment effect estimation 1

1.1.1 Potential outcomes 2

1.1.2 Structural causal models 5

1.1.3 Estimating treatment effects 6

1.1.4 Alternative estimands 9

1.2 Challenges in high dimensions 13

1.2.1 High dimensions in statistics and machine learning more generally 13

1.2.1.1 Difficulties with high dimensions in statistics and machine learning 13

1.2.1.2 Using representations to deal with high dimensions in statistics and machine learning 14

1.2.2 High-dimensional covariates in treatment effect estimation . . . 16

1.2.2.1 Difficulties with high-dimensional covariates in treatment effect estimation 16

1.2.2.2 Learning representations to deal with high-dimensional covariates in treatment effect estimation 18

1.3 Thesis outline 21

1.3.1 Work included in this thesis 21

1.3.2 Work not included in this thesis 22

2 Neural Score Matching for High-Dimensional Causal Inference **25**

3 Towards Representation Learning for Weighting Problems in Design-Based Causal Inference **57**

4 Deconfounding Scores and Representation Learning for Causal Effect Estimation with Weak Overlap **84**

5 Discussion and Conclusion **119**

5.1 Summary of contributions 120

5.1.1 Chapter 2 120

5.1.2 Chapter 3 120

5.1.3 Chapter 4 121
5.2 Key findings 121
5.2.1 More advanced representation learning for design-based causal inference 122
5.2.2 First steps towards controlling the confounding bias 122
5.2.3 An impossibility result on design-based causal inference and overlap 123
5.3 Limitations 124
5.3.1 Reliance on well-specified models 124
5.3.2 Unknown behaviour of representation-wise estimators 125
5.4 General conclusion 125

Bibliography **127**

1 Introduction

Contents

1.1 Fundamentals of treatment effect estimation 1
1.1.1 Potential outcomes 2
1.1.2 Structural causal models 5
1.1.3 Estimating treatment effects 6
1.1.4 Alternative estimands 9
1.2 Challenges in high dimensions 13
1.2.1 High dimensions in statistics and machine learning more generally 13
1.2.2 High-dimensional covariates in treatment effect estimation . . 16
1.3 Thesis outline 21
1.3.1 Work included in this thesis 21
1.3.2 Work not included in this thesis 22

In this introductory chapter, we first introduce the fundamentals of treatment effect estimation (Section 1.1), then review challenges that high-dimensional features pose for statistics, machine learning, and treatment effect estimation (Section 1.2), before finally presenting the contributions of the thesis (Section 1.3).

1.1 Fundamentals of treatment effect estimation

Within causal inference, we focus on the problem of treatment effect estimation. We want to evaluate the causal effect of a given *treatment* on some outcome of interest [Rubin, 1990,

Splawa-Neyman et al., 1990]. The treatment can refer to medical treatments but also any decision, policy or action [Hernan and Robins, 2024]. It is frequently assumed to be binary (in the form of "treat" versus "not treat"), a setting to which we stick throughout the thesis.

A key challenge in estimating a causal effect from observational data is *confounding*: the presence of variables (*confounders*) that influence both the treatment and the outcome. Such confounders can create a spurious association between the treatment and the outcome, and lead to incorrect conclusions. For example, it can be observed that increases in ice cream sales coincide with increases in violent crime, suggesting a causal relationship between the two phenomena. However, both increases typically occur during hot weather [McMahan, 2021, Pearl et al., 2016]: while heat intuitively leads to higher demand for refreshments such as ice cream, it might also cause more aggressive behaviour [Anderson, 1989]. Here, weather acts as a confounder between the decision to eat ice cream and violent crime, whereas the latter two are not connected by any causal relationship. In a more general setting, a treatment can cause an outcome, but the causal effect has to be disentangled from the effects of any confounders. For example, evaluating the effect of a drug on people with a given disease requires making sure that the effects of age, sex, weight, height, and other potential confounders are taken into account [Hernan and Robins, 2024]. Otherwise, one could incorrectly find that an effective drug reduces the chances of survival because it was actually mostly given to people in poorer health [Blyth, 1972]. The process of accounting for confounders when estimating treatment effects is called *adjustment* [Pearl et al., 2016].

Adjustment can be written in two different languages: potential outcomes [Rubin, 2005, Splawa-Neyman et al., 1990], and structural causal models [Pearl, 2009].

1.1.1 Potential outcomes

In the framework of potential outcomes (or Neyman-Rubin causal model) [Rubin, 2005, Splawa-Neyman et al., 1990], we assume that we have different *units* or *subjects* i, each with the following:

- A *treatment assignment* variable T_i indicating whether the unit i has been treated (when $T_i = 1$) or not (when $T_i = 0$);

- *Potential outcomes* $Y_i(1)$ and $Y_i(0)$ indicating the outcome we would have observed if the unit i was treated or not treated, respectively;

- *Covariates* X_i that encompass (pre-treatment) measurements about the unit i.

We assume that the data $(X_i, T_i, Y_i(1), Y_i(0))_i$ are i.i.d.. Thus, when relevant, we drop the i subscript for convenience. Throughout the thesis, we refer to any of the two values possibly taken by T_i as a treatment assignment or a treatment value, and we also refer to the distribution of T_i given X_i as the treatment assignment. One can measure the causal effect of the treatment by estimating the *average treatment effect* defined as

$$\text{ATE} := \mathbb{E}[Y(1) - Y(0)]. \tag{1.1}$$

The difficulty is that, in practice, *one does not observe the potential outcomes* $Y_i(1), Y_i(0)$ *but only an observed outcome* Y_i. As a result, the observed data is $(X_i, T_i, Y_i)_i$.

By assuming that a subject's treatment assignment does not influence the observed outcome of other subjects (no interference assumption), and that the observed outcome matches the potential outcome corresponding to the assigned treatment value (consistency assumption), one can write the observed outcome as $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$, or in other words, $Y_i = Y_i(1)$ when $T_i = 1$ and $Y_i = Y_i(0)$ when $T_i = 0$ [Hernan and Robins, 2024]. Critically, note that only one of the two potential outcomes can be observed for any subject i. As a result, the *individual treatment effect* $Y_i(1) - Y_i(0)$ cannot be observed for any i. This difficulty has been referred to as "the fundamental problem of causal inference" [Holland, 1986]. In contrast, as we will see shortly, the ATE can be estimated under additional assumptions even when only one potential outcome is observed for each subject.

For a treatment assignment $t = 0, 1$, one might consider estimating the average potential outcome $E[Y(t)]$ by simply averaging the observed outcomes for units that received that treatment assignment, i.e., $E[Y|T = t]$. Why is this approach invalid? This is due to the issue of confounding described previously. For a more concrete example, assume that T indicates taking a drug, X is a binary indicator of having a bad ($X = 1$) or good ($X = 0$) health, $Y(t)$ is Bernoulli and denotes remission or not, with a probability

of $\frac{1-X}{2} + \frac{t}{4}$. Then one can verify that the ATE is $\frac{1}{4}$. However, assume that the drug is given to all patients with bad health, and only to them. Then $T = X$, and

$$\begin{aligned}\mathbb{E}[Y|T=t] &= \mathbb{E}[Y(t)|T=t] = \mathbb{E}[Y(t)|T=t, X=t] \\ &= \frac{1-t}{2} + \frac{t}{4} = \frac{1}{2} - \frac{t}{4},\end{aligned}$$

where the first equality comes from consistency, and the second from $X = T$. As a result, taking the difference between $\mathbb{E}[Y|T=1]$ and $\mathbb{E}[Y|T=0]$ would give an ATE estimate of $-\frac{1}{4}$, the opposite of the true ATE.

As we note in this example, X confounds the relationship between the treatment T and the potential outcomes $(Y(1), Y(0))$. Could there be any other confounders? This is typically not testable [Hernan and Robins, 2024]. Instead we directly impose it by making the *unconfoundedness* assumption which states that

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X, \tag{1.2}$$

or, in other words, that X includes all confounders of the relationship between treatment T and potential outcomes $(Y(1), Y(0))$. Another critical assumption is called *overlap*, and states that every treatment assignment can be observed for a given value of confounders, or

$$0 < P(T=t|X=x) < 1, \quad \forall t = 0, 1, \ \forall x \text{ s.t. } P(X=x) > 0. \tag{1.3}$$

Letting $e(x) := P(T=1|X=x)$, a function called the *propensity score* [Rosenbaum and Rubin, 1983], overlap can be rewritten as

$$0 < e(X) < 1 \text{ a.s..}$$

Intuitively, overlap states that any unit could receive both treatment assignments. With all of these assumptions, we can finally rewrite the ATE as [Hernan and Robins, 2024]

$$\text{ATE} = \mathbb{E}[\mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X]] \tag{1.4}$$

where the quantity in the expectation with respect to X is equal to the *conditional average treatment effect* (CATE) [Imbens and Wooldridge, 2009], which is defined as

$$\text{CATE}(x) := \mathbb{E}[Y(1) - Y(0)|X=x].$$

In short, this is the potential outcomes vision of "adjusting for confounders". In Equation 1.4, we note that the ATE has been *identified*, that is, written using only observed random variables such as X, T, Y. More precisely, we have adjusted for X. We now look at the complementary version of adjustment for confounders, Pearl's structural causal model (SCM).

1.1.2 Structural causal models

In the structural causal model (SCM) framework [Pearl, 2009, Pearl et al., 2016], we assume that we can describe causal relationships thanks to a graph, which we refer to as the *causal graph*. Every variable W_i under scrutiny, where i now denotes the index of variables and not the subject, is caused by its parents in the graph and causes its children. Importantly, the graph is taken to be a directed acyclic graph (DAG) to avoid loops in the causation. Formally, noting $\mathcal{G}$ the causal graph, we assume that every variable W_i is assigned a value as

$$W_i \leftarrow f_i(W_{\mathrm{Pa}_i^{\mathcal{G}}}, U_i), \tag{1.5}$$

where f_i is the (deterministic) assignment function, $\mathrm{Pa}_i^{\mathcal{G}}$ the (potentially empty) set of parents of i in graph $\mathcal{G}$, $W_{\mathrm{Pa}_i^{\mathcal{G}}}$ the vector of variables indexed by this set, and U_i a noise variable. The collection of these equations for all variables makes up the structural causal model. Importantly, the U_i's are assumed mutually independent, so the random mechanisms giving every variable from its parents are independent. This is referred to as "independence of mechanisms" [Peters et al., 2017]. More succinctly, the joint probability density of W_i's can be factorised as

$$p(w) = \prod_i p(w_i | w_{\mathrm{Pa}_i^{\mathcal{G}}}). \tag{1.6}$$

However, different graphs can factorise the same distribution. What distinguishes the causal graph $\mathcal{G}$ from other graphs? It allows to perform the *intervention* [Pearl, 1995, 2009] do$(W_i = w_i^*)$ defined as replacing the $W_i \leftarrow f_i(W_{\mathrm{Pa}_i^{\mathcal{G}}}, U_i)$ operation in the SCM of Equation 1.5 by $W_i \leftarrow w_i^*$, or equivalently by replacing the $p(w_i | w_{\mathrm{Pa}_i^{\mathcal{G}}})$ term in the factorisation of Equation 1.6 by $1_{\{w_i = w_i^*\}}$. Other terms in the SCM and factorisation

are unaffected. The resulting distribution is denoted $p(.|\text{do}(W_i = w_i^*))$. Intuitively, this intervention corresponds to forcing a system to act beyond its observed or "natural" course. For example, instead of having the temperature of a room being decided by the weather, we force it at some fixed value. After an intervention on a variable, all dependencies of this variable on parents are cut off. However, children and descendants of the variable still depend on it, so the intervention affects them too. This is where the directions of the cause-effect relationships, as encoded by the causal graph $\mathcal{G}$, matter.

Then, we are interested in the distribution $P(W_j|\text{do}(W_i = w_i^*))$ of a variable W_j after the intervention $\text{do}(W_i = w_i^*)$. Notably, its expectation $\mathbb{E}[W_j|\text{do}(W_i = w_i^*)]$ gives the causal effect of the intervention $\text{do}(W_i = w_i^*)$ on W_j. In contrast, the conditional distribution $p(W_j|W_i = w_i^*)$ is typically subject to confounding, which the intervention removes. While the distribution $P(W_j|\text{do}(W_i = w_i^*))$ is typically not directly observable, it can be identified by finding an *adjustment set* $Z \subseteq \mathcal{W}\backslash\{W_i, W_j\}$, where $\mathcal{W}$ is the set of all variables, such that $P(W_j|\text{do}(W_i = w_i^*)) = \mathbb{E}[P(W_j|Z, W_i = w_i^*)]$. Classical criteria to find such an adjustment set are the back-door criterion and the front-door criterion [Pearl, 1995, 2009].

The potential outcomes framework can be rewritten as a structural causal model by assuming the graph in Figure 1.1 [Hernan and Robins, 2024, Imbens, 2020, Richardson and Robins, 2013, Wasserman, 2004]. Each $\mathbb{E}[Y(t)]$, $t = 0, 1$, in the ATE of Equation 1.1 is equal to $\mathbb{E}[Y|\text{do}(T = t)]$, and its identification in Equation 1.4 is a simple application of the back-door criterion. More generally, it can be shown that potential outcomes and structural causal models are equivalent frameworks [Pearl, 2009, Peters et al., 2017], and recent formulations have attempted to further bridge them [Dawid, 2021, Richardson and Robins, 2013]. We focus on the potential outcomes framework. However, the structural causal model is important to keep in mind, especially given the additional flexibility it gives.

1.1.3 Estimating treatment effects

Throughout this section, we assume unconfoundedness, overlap, consistency and no interference as defined in Section 1.1.1. First, if the covariates are *balanced* (we also

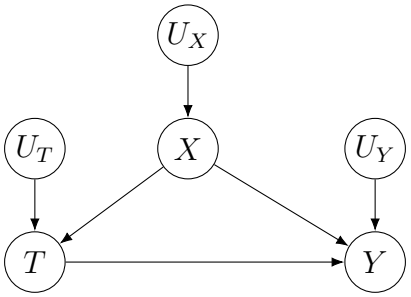


Figure 1.1: Causal graph for the potential outcomes framework with noise variables

say that *covariate balance* holds), that is, $X \perp\!\!\!\perp T$ which for a binary treatment means $P(X|T = 1) = P(X|T = 0)$, then covariates X no longer confound the treatment and potential outcomes, so we do have $\mathbb{E}[Y(t)] = \mathbb{E}[Y|T = t]$ [Hernan and Robins, 2024] and, denoting N_1 and N_0 as the number of treated and control samples in the data, respectively, the ATE can be estimated as

$$\widehat{\text{ATE}}_{\text{Diff}} := \frac{1}{N_1} \sum_{i:T_i=1} Y_i - \frac{1}{N_0} \sum_{i:T_i=0} Y_i.$$

Covariates are typically balanced (or assumed to be so) in randomised controlled trials (RCTs) [Moher et al., 2010]: indeed, the random assignment of treatments to subjects amounts to removing any dependence between patients' features and the treatment, making the former unable to act as confounders [Pearl et al., 2016]. In a general case, this does not hold and we have to resort to different techniques for estimating average treatment effects. One family sometimes referred to as "outcome regression" [Funk et al., 2011] takes advantage of Equation 1.4, and estimates the function $\mu(t, x) = \mathbb{E}[Y|T = t, X = x]$, which is often referred to as the "outcome model" [Brookhart et al., 2006, Hernan and Robins, 2024, Ning et al., 2020]. The resulting estimator $\hat{\mu}(t, x)$ is then used to estimate the ATE as [Hahn, 1998]

$$\widehat{\text{ATE}}_{\text{OR}} := \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\Big).$$

Another set of techniques consists in noting that

$$\forall t, \;\; \mathbb{E}[Y(t)] = \mathbb{E}\Big[\frac{\mathbb{1}_{\{T=t\}} Y}{P(T = t|X)}\Big],$$

so

$$\text{ATE} = \mathbb{E}\Big[\frac{TY}{e(X)}\Big] - \mathbb{E}\Big[\frac{(1-T)Y}{1-e(X)}\Big].$$

Thus, using an estimator $\hat{e}(x)$ of the propensity score, we can estimate the ATE from *inverse probability weighting* [Hernan and Robins, 2024, Horvitz and Thompson, 1952] as

$$\widehat{\text{ATE}}_{\text{IPW}} := \frac{1}{N}\left(\sum_{i=1}^{N}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}\right).$$

These two estimators can be combined to form the *augmented inverse probability weighted* estimator [Glynn and Quinn, 2010, Hernan and Robins, 2024, Robins et al., 1994], defined as

$$\widehat{\text{ATE}}_{\text{AIPW}} := \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\mu}(1,X_i) - \hat{\mu}(0,X_i) + \frac{T_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}(1,X_i)) - \frac{1-T_i}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}(0,X_i))\Big).$$

This estimator has the desirable property of *double robustness*, as it is consistent as soon as μ or e is correctly specified [Glynn and Quinn, 2010].

In general, inverse probability weighting can be related to a more general family of *weighting* techniques where the ATE is estimated as [Huling and Mak, 2024]

$$\widehat{\text{ATE}}_w := \frac{1}{N_1}\sum_{i:T_i=1} w_i Y_i - \frac{1}{N_0}\sum_{i:T_i=0} w_i Y_i$$

where w_i denotes weights, which are typically chosen so that covariate balance constraints hold in the reweighted distribution [Hainmueller, 2012, Zubizarreta, 2015], or more generally to minimise an "imbalance", i.e. a metric measuring how much covariate balance is not verified, again in this reweighted distribution [Ben-Michael et al., 2021, Kallus, 2020].

Finally, another set of techniques is *matching* [Abadie and Imbens, 2006, Stuart, 2010] which consists in finding for every unit i with treatment assignment T_i a match $j(i)$ in the other group, i.e. with treatment assignment $T_{j(i)} = 1 - T_i$. In *nearest-neighbour matching*, $j(i)$ is chosen to minimise a measure of dissimilarity between X_i and $X_{j(i)}$, and the counterfactual potential outcome is then taken to be $Y_{j(i)}$. The resulting ATE estimator can be written as

$$\widehat{\text{ATE}}_{\text{matching}} := \frac{1}{N}\sum_{i=1}^{N}\Big(T_i(Y_i - Y_{j(i)}) + (1-T_i)(Y_{j(i)} - Y_i)\Big).$$

In particular, the resulting assignments can be collapsed to unit-level weights, making matching a specific case of weighting. Notably, Kallus [2020] showed that weighting by minimisation of a Wasserstein imbalance returned the same weights as nearest-neighbour matching.

1.1.4 Alternative estimands

We briefly review alternative estimands to the ATE; notably, we underline connections with covariate shift.

Average treatment effect on the treated (ATT): as the name suggests, it is an analogue of the ATE of Equation 1.1 restricted to the treated population. More precisely, it is defined as

$$\text{ATT} := \mathbb{E}[Y(1) - Y(0)|T = 1].$$

The same unconfoundedness, overlap, and consistency assumptions are then used to identify the ATT as

$$\text{ATT} = \mathbb{E}[Y|T = 1] - \mathbb{E}\left[\mathbb{E}[Y|T = 0, X] \mid T = 1\right]$$

or as

$$\text{ATT} = \mathbb{E}[Y|T = 1] - \frac{\mathbb{E}\left[\frac{(1-T)e(X)}{1-e(X)}Y\right]}{\mathbb{E}[T]}.$$

Note that the challenging part resides in estimating the $\mathbb{E}[Y(0)|T = 1]$ part of the ATT, as the $\mathbb{E}[Y(1)|T = 1]$ part is equal to $\mathbb{E}[Y|T = 1]$, thus can be obtained by averaging $Y_i's$ across treated subjects. The previously seen outcome regression estimator can be adapted [Hahn, 1998] as

$$\widehat{\text{ATT}}_{\text{OR}} := \frac{1}{N_1}\sum_{i=1}^{N} T_i\Big(Y_i - \hat{\mu}(0, X_i)\Big),$$

the inverse probability weighting estimator as [Moodie et al., 2018]

$$\widehat{\text{ATT}}_{\text{IPW}} := \frac{1}{N_1}\sum_{i=1}^{N}\left(T_iY_i - (1 - T_i)\frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}Y_i\right),$$

its augmented counterpart as [Moodie et al., 2018]

$$\widehat{\text{ATT}}_{\text{AIPW}} := \frac{1}{N_1}\sum_{i=1}^{N}\left(T_i(Y_i - \hat{\mu}(0, X_i)) - (1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}(0, X_i))\right),$$

general weighting estimators as [Kallus, 2020]

$$\widehat{\text{ATT}}_w := \frac{1}{N_1}\sum_{i:T_i=1} Y_i - \frac{1}{N_0}\sum_{i:T_i=0} w_i Y_i,$$

and the matching estimator as [Abadie and Imbens, 2006]

$$\widehat{\text{ATT}}_{\text{matching}} := \frac{1}{N_1}\sum_{i:T_i=1}^{N}\left(Y_i - Y_{j(i)}\right)$$

where $j(i)$ is the control group.

Transportability: it consists in estimating the average treatment effect on a target population ($S = 0$) using the data of an RCT ($S = 1$) [Colnet et al., 2024, Degtiar and Rose, 2023]. More precisely, we have access to covariates X_i when $S_i = 0$ and to full data X_i, T_i, Y_i when $S_i = 1$, and our target estimand is

$$\text{ATE}_T := \mathbb{E}[Y(1) - Y(0)|S = 0].$$

As $S = 1$ is an RCT, we have $X \perp\!\!\!\perp T|S = 1$, and $(Y(1), Y(0)) \perp\!\!\!\perp T|S = 1$. Defining $e_S(x) := P(S = 1|X = x)$ the conditional probability of trial participation, as well as $\pi_S := P(S = 1)$ and $\pi_T := P(T = 1|S = 1)$, assume *support inclusion*, that is, $e_S(x) > 0$, as well as $0 < \pi_T < 1$ and $\pi_S < 1$. An important assumption is *transportability* which states that the CATE is identical across the target and RCT groups:

$$\mathbb{E}\left[Y(1) - Y(0)|S = 1, X\right] = \mathbb{E}\left[Y(1) - Y(0)|S = 0, X\right].$$

We denote both CATEs as $\text{CATE}(X)$. Then, the target ATE can be identified as both

$$\text{ATE}_T = \mathbb{E}\left[\text{CATE}(X)|S = 0\right]$$

and

$$\text{ATE}_T = \mathbb{E}\left[\frac{\pi_S}{1-\pi_S}\frac{1-e_S(X)}{e_S(X)}\left(\frac{TY}{\pi_T} - \frac{(1-T)Y}{1-\pi_T}\right)\,\middle|\, S = 1\right]$$

where we used the density ratio of $P(X|S=0)$ with respect to $P(X|S=1)$ and we identified the CATE as

$$\text{CATE}(X) = \mathbb{E}\left[\frac{TY}{\pi_T} - \frac{(1-T)Y}{1-\pi_T} \;\middle|\; X, S=1\right].$$

Thus, the previous estimators can be generalised to transportability, as we detail next.

Importance weighting: We note that ATE/ATT estimation and transportability can be framed as an importance weighting problem [Bruns-Smith and Feller, 2022] (see Chapter 3 for details on the connection). Indeed, let L be a binary random variable, and P^s, P^t two distributions for X, Y where $P^s = P(.|L=0)$ is the source distribution, and $P^t = P(.|L=1)$ is the target distribution. Denote $\mathbb{E}_R$ the expectation with respect to distribution R, and R_Z the law of random variable Z in R. Overlap is generalised by assuming that P^t_X is absolutely continuous with respect to P^s_X. We have access to samples $(X_i, Y_i) \sim P^s$ and $X_i \sim P^t$. Then we are interested in the estimand

$$\tau_{s,t} := \mathbb{E}_{P^t}\left[\mathbb{E}_{P^s}[Y|X]\right]$$

which is further equal to

$$\tau_{s,t} = \mathbb{E}_{P^t}\left[Y\right]$$

if one assumes $P^s_{Y|X} = P^t_{Y|X}$, which is a generalised form of unconfoundedness [Bruns-Smith and Feller, 2022]. While one can regress Y on X in P^s to estimate $\mathbb{E}_{P^s}[Y|X]$ and use it to estimate $\tau_{s,t}$ as in the previously seen outcome regression, weighting is also possible as we have

$$\tau_{s,t} = \mathbb{E}_{P^s}\left[\frac{\mathrm{d}P^t_X}{\mathrm{d}P^s_X}(X)Y\right]$$

where $\frac{\mathrm{d}P^t_X}{\mathrm{d}P^s_X}(X)$ is the density ratio of P^t_X with respect to P^s_X. This density ratio can be estimated through density ratio regression methods [Kanamori et al., 2009], or by estimating $e_{s,t}(X) := p(L=1|X)$ through classification [Sugiyama et al., 2008]. More general weights $w(X)$ such that

$$\tau_{s,t} \approx \mathbb{E}_{P^s}\left[w(X)Y\right]$$

can be computed through other techniques such as kernel mean matching [Gretton et al., 2009]. While these methods can be readily applied to ATE/ATT estimation and transportability, and are often equivalent to existing canonical techniques for these estimands (e.g. fitting $e_{s,t}(X)$ through classification and plugging it in an estimator of $\frac{\mathrm{d}P^t_X}{\mathrm{d}P^s_X}(X)$ is equivalent to IPW for the ATT), these are more commonly used to estimate the average loss on a testing set using the training set [Bickel et al., 2009, Sugiyama et al., 2007]. In this setup, $Y = \ell(f(X), Y')$ where f is a predictor, Y' a label available on training samples but not testing samples and ℓ a loss function, P^s is the training distribution and P^t the testing distribution. Unconfoundedness is typically written as *covariate shift*, i.e. $P^s_{Y'|X} = P^t_{Y'|X}$ [Bickel et al., 2009, Sugiyama et al., 2007], directly implying $P^s_{Y|X} = P^t_{Y|X}$, while in contrast $P^s_X \neq P^t_X$. While outcome regression and doubly robust methods are typically not used when Y is a loss function as aforementioned, they have been explored when $Y = Y'$ [Li et al., 2020, Yu and Szepesvári, 2012].

General linear functionals of outcome functions: Finally, we present the most general estimand that encompasses all estimands seen before [Chernozhukov et al., 2022b]. Noting the data as (W_i, Y_i) where Y refers to the outcome and W to features (encompassing both covariates and treatment for ATE/ATT estimation), we are interested in

$$\theta_0 := \theta(\mu_0)$$

where $\theta(\mu) := \mathbb{E}\left[m(W, \mu)\right]$ for some known function m that is linear in μ and where $\mu_0(w) := \mathbb{E}[Y|W = w]$. From there, it is clear that outcome regression can be generalised to such estimands by regressing Y on W to obtain an estimator $\hat{\mu}$ of μ_0 and plugging it into θ. However, it turns out that weighting can be generalised too, if one assumes that θ is a mean-square continuous functional. Then, from the Riesz representation theorem, there exists α_0 in $L_2(P_W)$ such that $\theta(\mu) = \mathbb{E}[\alpha_0(W)\mu(W)]$ for any $\mu \in L_2(P_W)$, where $L_2(P_W)$ is the set of functions f of W such that $f(W)$ is square-integrable. Thus if $\mu_0 \in L_2(P_W)$ then the target estimand can be expressed as

$$\theta(\mu_0) = \mathbb{E}[\alpha_0(W)Y].$$

Consequently, an analogue of previously seen IPW and AIPW estimators can be derived by estimating α_0. A general method, regardless of m, is to notice that

$$\begin{aligned}\alpha_0 &= \text{argmin}_{\alpha \in L_2(P_W)} \mathbb{E}\left[(\alpha(W) - \alpha_0(W))^2\right] \\ &= \text{argmin}_{\alpha \in L_2(P_W)} \left(\mathbb{E}\left[\alpha(W)^2\right] - 2\mathbb{E}\left[m(W, \alpha)\right]\right)\end{aligned}$$

so minimising an empirical analogue of $\mathbb{E}\left[\alpha(W)^2\right] - 2\mathbb{E}\left[m(W, \alpha)\right]$ over a hypothesis class gives an estimator $\hat{\alpha}$ of α_0. This method is referred to as Riesz regression [Chernozhukov et al., 2021] and is a generalisation of density ratio least-squares regression [Kanamori et al., 2009]. Riesz regression and outcome regression are typically combined to form *automatic debiased machine learning* estimators [Chernozhukov et al., 2022b]; those are generalisations of AIPW estimators used in ATE/ATT estimation.

1.2 Challenges in high dimensions

Many modern datasets, such as images or electronic health records (EHRs) in medicine, are high-dimensional [Bengio et al., 2013, Hripcsak and Albers, 2013]. As we will see, this creates challenges for prediction and inference in general, and treatment effect estimation in particular.

1.2.1 High dimensions in statistics and machine learning more generally

1.2.1.1 Difficulties with high dimensions in statistics and machine learning

The presence of many covariates, i.e. when the features X have many dimensions, e.g., $D \approx N$ or $D \geq N$ where D is the dimension of X and N the sample size, is a challenging setting in statistics and machine learning. In the absence of regularisation, classical methods will not only capture the relationship between predictors X and the response variable Y, but also the noise from the sample: this is overfitting [Belloni et al., 2014a, Chernozhukov et al., 2018]. Statistical inference itself becomes difficult as classical asymptotic theory no longer provides useful statements, such as consistency results, confidence intervals or p-values [Javanmard and Montanari, 2018, Negahban et al., 2012, Wainwright, 2019]. Nearest-neighbour estimators also fail in high dimensions due

to Euclidean distances of any two points becoming hardly distinguishable [Aggarwal et al., 2001, Beyer et al., 1999]. Other methods breaking down or struggling in high dimensions include second-order optimisation methods [Goodfellow et al., 2016], kernel methods [Bach, 2024], series regression [Athey and Imbens, 2017], kernel- and distance-based hypothesis tests [Ramdas et al., 2015], Markov Chain Monte-Carlo [Goodfellow et al., 2016], covariance matrix learning [Raymaekers and Rousseeuw, 2021, Wainwright, 2019]. Intuitively, the number of observations needed for a dense coverage of the entire feature space grows exponentially with the dimension [Goodfellow et al., 2016, Wainwright, 2019], making generalisation difficult. Additionally, high dimensions can lead to intractable computations [Fan and Lv, 2008, Goodfellow et al., 2016].

1.2.1.2 Using representations to deal with high dimensions in statistics and machine learning

To solve all of this, a canonical solution is to assume that data or models need not be as complex as the dimension suggests, and actually follow a form of low-dimensional structure [Bengio et al., 2013, Wainwright, 2019]. In linear models, where there are D parameters, as many as dimensions, the rationale is to assume that only $s \ll D$ entries of the parameter vector are actually non-zero (*sparsity*), or that the model can be well-approximated by such a lower-dimensional parameter with s non-zero entries (*approximate sparsity*) [Belloni and Chernozhukov, 2013, Belloni et al., 2014a, 2017, Bühlmann and Van De Geer, 2011, Wainwright, 2019]. A related but distinct assumption is the *manifold hypothesis* which states that high-dimensional features lie near a low-dimensional manifold [Bengio et al., 2013, Goodfellow et al., 2016]. The manifold hypothesis is well-motivated by existing high-dimensional data: images typically have outsize pixel dimensions but can be compressed to low-dimensional formats [Rigollet and Hütter, 2023], and like texts they generally follow a precise, lower-dimensional structure, as uniformly sampling them will lead to nonsensical samples [Bengio et al., 2013, Goodfellow et al., 2016]. All of this suggest learning *representations* of features, that is, low-dimensional mappings $\phi(X)$ of features X that capture their essential information while discarding redundancies [Bengio et al., 2013, Goodfellow et al., 2016].

How to learn such representations, and generally obtain accurate estimators and inference from high-dimensional data? A first set of approaches is unsupervised, in the sense that such methods learn a representation $\phi(X)$ using only the feature space X, before feeding them to downstream tasks; they include principal component analysis (PCA) [Hotelling, 1933, Pearson, 1901], random projections [Bach, 2024, Bingham and Mannila, 2001, Johnson et al., 1984] and deterministic autoencoders [Hinton and Salakhutdinov, 2006]. However, for the sake of treatment effect estimation, we are mostly interested in supervised approaches, where the representation $\phi(X)$ is chosen in accordance with a fixed goal, such as predicting an outcome Y or inferring parameters. The earliest approaches include stepwise selection, where features are added (forward selection) or removed (backward selection) iteratively, one-by-one [Hocking, 1976]. However these were found to be computationally expensive and unstable to data changes [Fan and Li, 2001, Kreif and DiazOrdaz, 2019, Tibshirani, 1996]. Since then, the canonical approach for linear models has been the LASSO [Tibshirani, 1996], which consists in adding a l_1 penalty on weights. This will typically lead to many weights being set to zero, effectively performing variable selection. The LASSO is typically near-optimal for estimation of the regression function [Belloni and Chernozhukov, 2013], however it can require more stringent assumptions for other tasks, such as recovering the true set of relevant covariates [Meinshausen and Bühlmann, 2010, Zhao and Yu, 2006], especially when there are more of them than samples [Zou and Hastie, 2005], or selecting similar covariates together [Yuan and Lin, 2006, Zou and Hastie, 2005]. Alternatives to alleviate such issues include Elastic Net [Zou and Hastie, 2005], Adaptive LASSO [Zou, 2006] and Group LASSO [Yuan and Lin, 2006]. In addition, it is difficult to perform statistical inference for model parameters estimated through the LASSO [Javanmard and Montanari, 2014, 2018], notably as non-zero entries of the parameter vector estimated by LASSO are typically biased towards zero [Belloni et al., 2014a]. More generally, estimators of parameters defined as functionals of nuisance functions typically converge less fast than root-N when nuisance functions are estimated using regularised models [Chernozhukov et al., 2018, 2022b]. Thus, approaches have been developed to "de-bias" the LASSO and other regularised models using an additional

step. These include (i) post-selection, i.e., where models are re-trained while keeping only the selected parameters [Belloni and Chernozhukov, 2013]; or (ii) adding carefully chosen additive bias-correction terms to estimators of these parameters [Javanmard and Montanari, 2014, 2018]. The latter approach has been generalised to correcting moments in estimating equations, replacing them with "orthogonal" moments that are more robust to changes in nuisance functions [Chernozhukov et al., 2018].

Other representation learning methods outside linear regression include Fisher's linear discriminant [Fisher, 1936, Rao, 1948], partial least squares [Wold, 1966, 1975], sufficient dimension reduction [Cook, 2009, Li, 2018, 1991], sparse additive methods [Ravikumar et al., 2009] and neural networks [LeCun et al., 2015, Rumelhart et al., 1986]. The latter inherently learn representations in each hidden layer, with increasing levels of abstraction [LeCun et al., 2015]; examples include transformers [Vaswani et al., 2017] for texts and other sequential inputs, and convolutional neural networks [LeCun et al., 1989] for images. Finally, methods learning non-deterministic representations of features include probabilistic PCA [Tipping and Bishop, 1999] and variational auto-encoders [Kingma and Welling, 2013], while manifold learning methods [McInnes et al., 2018, Roweis and Saul, 2000] typically do not learn a deterministic or random mapping ϕ of individual feature values, but rather of the entire dataset.

1.2.2 High-dimensional covariates in treatment effect estimation

1.2.2.1 Difficulties with high-dimensional covariates in treatment effect estimation

We now turn to treatment effect estimation with high-dimensional covariates X. It involves aforementioned challenges with high-dimensional features from statistics and machine learning, but also has unique issues. Because the ATE is obtained by adjusting for confounders, estimating it will critically differ from classical prediction in the sense that it typically is obtained as a functional of nuisance functions [Athey et al., 2018, Chernozhukov et al., 2018, 2022a,b]. These nuisance functions usually are the propensity score and the outcome model. They can be fitted using automatic variable selection methods such as the LASSO, but will then have to be "combined" appropriately as in Sections 1.1.3 and 1.1.4 to obtain an estimate of the average treatment effect, as well as

confidence intervals. This is where de-biasing methods described in Section 1.2.1.2 will become relevant. Indeed, doubly robust estimators such as AIPW are specialisations of de-biasing methods [Kennedy, 2024]. Those are particularly important as a single variable selection method on the outcome model can exclude important confounders [Belloni et al., 2014a, Ertefaie et al., 2018]. However, such methods typically involve fitting a propensity score model and inverting it, thus can exhibit outsize errors when the estimated propensity score reaches near-zero or near-one values [Athey et al., 2018, Kang and Schafer, 2007].

Unfortunately, this is common in high dimensions due to a higher level issue in treatment effect estimation with many covariates: the overlap assumption might then be harder to verify [Hill and Su, 2013]. Intuitively, this is related to the idea developed in Section 1.2.1.1 that it becomes exponentially harder to densely cover the entire feature space with samples. Notably, it becomes harder to find two units in the same neighbourhood but with different treatment assignments, effectively precluding overlap; this is referred to as a "practical violation" of overlap [Petersen et al., 2012]. More precisely, D'Amour et al. [2021] have shown that assuming strict overlap, a slightly stronger version of overlap where the propensity score $e(X)$ is bounded away from 0 and 1, becomes unrealistic as the dimension grows: it implies that each covariate nearly verifies balance and brings negligible information compared to others. While the standard overlap assumption is critical for identification of the ATE, plain identification is typically not enough and strict overlap or other assumptions ensuring that $e(X)$ is not "too close" to 0 or 1 are typically needed for accurate statistical inference as it becomes inherently difficult to construct estimators involving inverse propensity scores without outsize variance [D'Amour et al., 2021]. Notably, the efficiency bound which lower-bounds the asymptotic variance of a large class of estimators is controlled by the magnitude of the inverse propensity score [Hahn, 1998, Khan and Tamer, 2010]. Thus, poor overlap will usually translate to high variance of many treatment effect estimators.

As a result, the aforementioned de-biasing methods involving inverse propensity scores will exhibit high errors in practice. Possible mitigations include fitting propensity scores using losses based on covariate balance constraints [Bradic et al., 2019, Tan, 2020, Zhao, 2019], or bypassing propensity score estimation to directly estimate weights, again

using covariate balance constraints [Athey et al., 2018, Hirshberg and Wager, 2021]. This is further motivated by the fact that the propensity score estimate should not accurately discriminate the treatment and control groups [Westreich et al., 2011]. However, these approaches based on balance constraints might not scale well to many covariates as they will involve as many constraints to satisfy [Ning et al., 2020]. Another approach is to directly fit the inverse propensity score model, instead of inverting a fitted propensity score model, as illustrated by Riesz representers [Chernozhukov et al., 2021, 2022a,c]. However, the variance of such estimators will remain vulnerable to poor overlap which again translates to high-amplitude inverse propensity scores or more generally high-amplitude Riesz representers [Chernozhukov et al., 2021, 2022c]. Another illustration of the combined impact of poor overlap and failure of classical machine learning methods is that matching estimators might struggle to find matches when covariates are high-dimensional, due to both poor overlap and the failure of nearest-neighbour estimators [Dehejia and Wahba, 2002, Wang et al., 2021]. As a result, the bias of matching estimators increases with the dimension [Abadie and Imbens, 2006].

1.2.2.2 Learning representations to deal with high-dimensional covariates in treatment effect estimation

Thus, we turn to selecting variables, or learning representations. The key distinction with general statistics and machine learning is that the representation $\phi(X)$ should still verify unconfoundedness, or at least be a valid adjustment set [Hernan and Robins, 2024]. In a sense, this is analogous to the sparsity assumption and the manifold hypothesis outlined earlier for general statistics and machine learning: while the covariates X might be high-dimensional, only a small portion of it might actually describe the confounding information we are interested in adjusting for. More practically, it is further discouraged to include variables in X or more generally information in X that is related to the treatment assignment and not the outcome, as this can increase variance of estimators [Hernan and Robins, 2024, Schneeweiss et al., 2009] and even inflate bias due to hidden confounders that would arise if unconfoundedness was violated [Pearl, 2011, Wooldridge, 2016]. Conversely, including information related to the outcome but not the treatment assignment is recommended as (i) it can decrease the variance of estimators without

adding bias [Brookhart et al., 2006]; and (ii) overlap on the resulting representation might be less restrictive than overlap on the full covariate set [D'Amour et al., 2021, Luo et al., 2017]. Such guidance does not only apply to classical ATE/ATT estimations but also transportability [Colnet et al., 2024], which suggests a general property of representation learning for covariate shift problems. Thus, using representations to reduce overlap-related variance is an attractive alternative to other canonical approaches that might add bias, such as (i) changing the estimand to the treatment effect on a population with better overlap [Crump et al., 2009, Kallus, 2021, Li et al., 2018, Matsouaka and Zhou, 2020], and (ii) trimming extreme (inverse) propensity scores [Mehrabi and Wager, 2024, Stürmer et al., 2010], which typically requires additional de-biasing steps [Chaudhuri and Hill, 2025, Ma and Wang, 2020, Sasaki and Ura, 2022]. However, outcome information might not be available at the time of representation learning, for example in design-based settings [Rubin, 2008, Stuart, 2010] where outcomes are only available after a matching or weighting stage. Finally, we mention that some variables should not even be included in the adjustment set, especially post-treatment variables, as they might lead to violations of unconfoundedness [Hernan and Robins, 2024]. Variable selection methods might not exclude these variables, leading to poor performance of de-biasing estimators [Hünermund et al., 2023].

Two canonical, extreme representations preserving unconfoundedness exist: (i) *balancing scores* [Rosenbaum and Rubin, 1983], that contain all the covariate information needed to predict treatments, i.e. $X \perp\!\!\!\perp T|\phi(X)$, which notably includes the propensity score $e(X)$, (ii) *prognostic scores* [Hansen, 2008] that contain all the covariate information needed to predict outcomes, i.e. $X \perp\!\!\!\perp (Y(0), Y(1))|\phi(X)$. In practice, such representations are not readily available and must be estimated, typically using standard outcome regression and propensity score fitting, which themselves can be done using classical penalised regression [Antonelli et al., 2018, Wyss et al., 2024]. Sparsity conditions might need to apply to both the outcome model and the propensity score model [Belloni et al., 2014b, Bradic et al., 2019]. Besides classical variable selection to fit each of the outcome and propensity score models, variable selection methods attempting to combine outcome information and treatment assignment information include the

following: (i) selecting each variable or not using an empirical measure of confounding [Schneeweiss et al., 2009]; (ii) combining both outcome and propensity models in the penalised regression loss [Ertefaie et al., 2018]; (iii) performing separate variable selection on the two models then post-selection treatment effect estimation on the concatenated set of selected variables [Belloni et al., 2014b]; (iv) performing outcome regression first and then using information from this outcome regression to guide the propensity score fitting [Ning et al., 2020, Shortreed and Ertefaie, 2017]. Approaches to learn representations beyond simple variable selection include sufficient dimension reduction to find representations that are predictive of the outcome [Luo et al., 2017] or of the treatment [Ghosh, 2011] or both [Ma et al., 2019], partial least squares [Ghosh, 2011], linear transformations of fixed kernel feature maps [Li and Fu, 2017], or using neural networks where an intermediate hidden layer can be used as the representation. Such neural networks can predict only the outcome [Shalit et al., 2017] or both the outcome and the treatment [Shi et al., 2019], and the classical losses might be combined with a term penalising poor overlap of the representation [Johansson et al., 2016, Shalit et al., 2017, Zhang et al., 2020] or with sample weighting [Assaad et al., 2021, Johansson et al., 2022].

Also, while these methods attempt to learn a representation that preserves unconfoundedness or generally is a valid adjustment set, it is difficult to ensure that the learnt representation actually does so. Common practices are restrictive assumptions such as linear models [Antonelli et al., 2018, Belloni et al., 2014b, Ertefaie et al., 2018, Ning et al., 2020, Shortreed and Ertefaie, 2017] or conditional independence statements [Ghosh, 2011, Luo et al., 2017], or simply assuming that the representation is one-to-one [Johansson et al., 2022, Shalit et al., 2017]. However, similarly as sparsity might not be exact but approximate, an emerging direction in the literature is to quantify the loss of confounding information induced by the representation and how it affects the estimation of treatment effects [Curth et al., 2021, Johansson et al., 2019]. As this is typically unobservable, an alternative approach [Melnychuk et al., 2024] involves adapting tools from the sensitivity literature where one generally quantifies the impact of an unobserved confounder using assumed correlations between unobserved and observed confounders [Cinelli et al., 2019, Huang and Pimentel, 2025, Tan, 2006].

1.3 Thesis outline

This thesis focuses on learning suitable representations of covariates for treatment effect estimation; notably, they can be applied to finding low-dimensional representations of high-dimensional covariates. It follows an integrated format, where each non-introductory and non-concluding chapter corresponds to an individual paper. We now detail each such paper included in Chapters 2, 3 and 4, as well as other works. Chapter 5 provides a discussion and a conclusion.

1.3.1 Work included in this thesis

In Chapter 2, we learn low-dimensional but multivariate balancing scores by observing that hidden layers of neural networks predicting the treatment from covariates naturally satisfy a key characterisation of such scores. We then bound the imbalance of the original covariates with the imbalance of the learnt representations, assuming that those are balancing scores, for a variety of imbalance measures. The learnt representations generally exhibit competitive performance on matching for datasets with high-dimensional covariates. Chapter 2 was published as follows.

- Clivio, O., Falck, F., Lehmann, B., Deligiannidis, G. & Holmes, C. (2022). **Neural Score Matching for High-Dimensional Causal Inference**. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 151:7076-7110. [Clivio et al., 2022]

However, Chapter 2 relies on the assumption that the neural network from which the representation is extracted is well-specified. This is what we focus on in Chapter 3: besides working on a more general weighting framework that accommodates multiple problems in causal inference, we give an explicit characterisation of the information lost when using any representation for weighting. This allows to derive a "balancing score error" which, as the name suggests, quantifies how much the representation is not a balancing score. This term is further upper-bounded by a quantity that can be estimated from finite data up to a constant: we optimise it to derive approximate analogues of balancing scores. Chapter 3 was published as follows.

- Clivio, O., Feller, A. & Holmes, C. (2024). **Towards Representation Learning for Weighting Problems in Design-Based Causal Inference.** Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence, in Proceedings of Machine Learning Research 244:856-880. [Clivio et al., 2024]

However, while we previously highlighted the critical role of high dimensions in poor overlap, Chapters 2 and 3 do not explicitly aim to optimise overlap when learning representations. This is precisely what we do in Chapter 4: we define an "overlap divergence" metric that quantifies representations' degree of overlap and its impact on estimators' performance. We justify it, underline some of its properties, and minimise it under the constraint that the representation is a valid adjustment set. Notably, we depart from previous Chapters where only information on the treatment assignment was used for representation learning and now also use outcome information to this aim. Specifically, in a simple setting, we compute a closed form for representations that are valid adjustment sets and show that among them, prognostic scores are optimal. We assess this behaviour in practice. A reference can be given as:

- Clivio, O.*, D'Amour, A.*, Franks, A.*, Bruns-Smith, D., Holmes, C. & Feller, A. (2026). **Deconfounding Scores and Representation Learning for Causal Effect Estimation with Weak Overlap**. Proceedings of The 29th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 300. (* for equal contribution) [Clivio et al., 2026]

1.3.2 Work not included in this thesis

The following works were performed during doctoral studies; however, they were not included to maintain a cohesive thesis.

In one line of work, we focused on links between Shapley values [Lundberg and Lee, 2017, Shapley, 1953] and treatment effect estimation. When evaluating the impact of a binary feature on an outcome, it turns out that its Shapley value is a weighted decomposition of coalition-wise conditional average treatment effects. We leveraged this to introduce PWSHAP, a method to provide path-specific Shapley values that can have

causal interpretations depending on an assumed causal graph. We illustrate on specific examples that such Shapley values can identify phenomena such as confounding, effect modification and mediation. This method was published as follows.

- Ter-Minassian, L.*, Clivio, O.*, Diazordaz, K., Evans, R.J. & Holmes, C. (2023). **PWSHAP: A Path-Wise Explanation Model for Targeted Variables**. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research 202:34054-34089. (* for equal contribution) [Ter-Minassian et al., 2023]

In another line of work, we reviewed causal reasoning benchmarks for large language models [Achiam et al., 2023, Radford et al., 2018, Touvron et al., 2023]. Typically, many of them amount to pure knowledge retrieval, making their ability to measure causal reasoning questionable. While recent benchmarks incorporate more interventional and counterfactual reasoning, they still suffer from key design issues, such as inappropriate metrics, poor data quality or memorisation. We propose four criteria for benchmarks to appropriately assess causal reasoning abilities of large language models: causal rather than correlative, open-ended, scalable, non-retrievable. It was published as follows.

- Yang, L., Shirvaikar, V., Clivio, O. & Falck, F. (2024). **A Critical Review of Causal Reasoning Benchmarks for Large Language Models.** In AAAI Workshop on "Are Large Language Models Simply Causal Parrots?". [Yang et al., 2024]

Finally, we studied how to incorporate knowledge returned by large language models into causal discovery algorithms [Pearl, 2009, Spirtes et al., 2000]. Traditional algorithms integrating background knowledge into causal discovery algorithms typically do not consider the fact that large language models are imperfect experts, that is, they might return incorrect knowledge. We build on literature on learning to defer [Madras et al., 2018, Mao et al., 2023] to propose L2D-CD, a method for bivariate causal discovery that learns whether each of a metadata-based imperfect expert and a numerical causal discovery algorithm returns the correct causal direction depending on the pair of variables, helping decide which method to use on new pairs. We show that the resulting combined

method outperforms each separate method in isolation, and can identify domains where the imperfect expert is strong and where it is weak. We also outline how to generalise this approach to graphs with more than 2 variables. It can be referenced as follows.

- Clivio, O., Mahajan, D., Taslakian, P.*, Magliacane, S.*, Mitliagkas, I., Zantedeschi, V.† & Drouin, A.† (2025), **Learning to Defer for Causal Discovery with Imperfect Experts.** In ICLR Workshop on "Reasoning and Planning for Large Language Models". (* for equal contribution, † for equal supervision) [Clivio et al., 2025]

2

Neural Score Matching for High-Dimensional Causal Inference

Neural Score Matching for High-Dimensional Causal Inference

Oscar Clivio[1] **Fabian Falck**[1]
Brieuc Lehmann[2] **George Deligiannidis**[1] **Chris Holmes**[1,3]
[1]University of Oxford [2]University College London [3]Alan Turing Institute

Abstract

Traditional methods for matching in causal inference are impractical for high-dimensional datasets. They suffer from the curse of dimensionality: exact matching and coarsened exact matching find exponentially fewer matches as the input dimension grows, and propensity score matching may match highly unrelated units together. To overcome this problem, we develop theoretical results which motivate the use of neural networks to obtain non-trivial, multivariate balancing scores of a chosen level of coarseness, in contrast to the classical, scalar propensity score. We leverage these balancing scores to perform matching for high-dimensional causal inference and call this procedure *neural score matching*. We show that our method is competitive against other matching approaches on semi-synthetic high-dimensional datasets, both in terms of treatment effect estimation and reducing imbalance.

1 INTRODUCTION

Estimating the causal effect of a treatment or a policy is the fundamental task of causal inference. For binary treatments, the quantity of interest is the difference between the outcome of a subject receiving a treatment (a *treated* subject) and the outcome of that subject in the absence of treatment (a *control* subject). The main difficulty in estimating a causal effect from observational data is that one cannot observe the outcome of both the true and the alternative scenario for the same subject – also called the factual and counterfactual outcomes. For instance, to evaluate the effect of a lockdown on reducing infection case numbers in a given country, one cannot create an exact copy of that country to study the consequences of its absence.

One possible solution would be to find a country that is very similar to the country under study, yet which did not experience a lockdown. This is the general idea behind *matching* whereby each treated subject in the sample data is assigned to one or more subjects from the control group (Stuart, 2010). Matching is among the dominant techniques used in medicine and other domains to estimate the effect of a treatment from observational data (Su et al., 2019; Farzadfar et al., 2012; Razonable et al., 2021; Webb et al., 2020). Besides estimating the treatment effect, matching can serve additional objectives. For example, matching can reduce imbalance, i.e. distributional differences between the treated and control groups that indicate confounding and consequently make treatment effect estimation more difficult. Matching can also help to decrease costs by reducing the number of control samples required when the collection of data (e.g. subjects' outcome) is expensive (Stuart, 2010). Matching methods, however, generally suffer from the *curse of dimensionality* (Abadie and Imbens, 2006a; Roberts et al., 2020), rendering them impractical for many modern high-dimensional datasets, such as electronic health records or clinical images.

In this work, we address the curse of dimensionality by first compressing the input covariates into a lower-dimensional matching space with a neural network and then matching in this space. Our contributions are as follows: (a) We develop novel theoretical results that bound the imbalance in the original covariate space via imbalance in a lower-dimensional balancing score space. We also extend these results to functions of covariates that violate the balancing score condition and which we refer to as "non-balancing scores". (b) These theoretical results motivate *neural score matching*, a procedure to match on low-dimensional balancing scores obtained from the intermediate layers of a neural network modelling the propensity score. This yields a simple method for estimating average or group-based treatment effects in the presence of high-dimensional

Preprint.

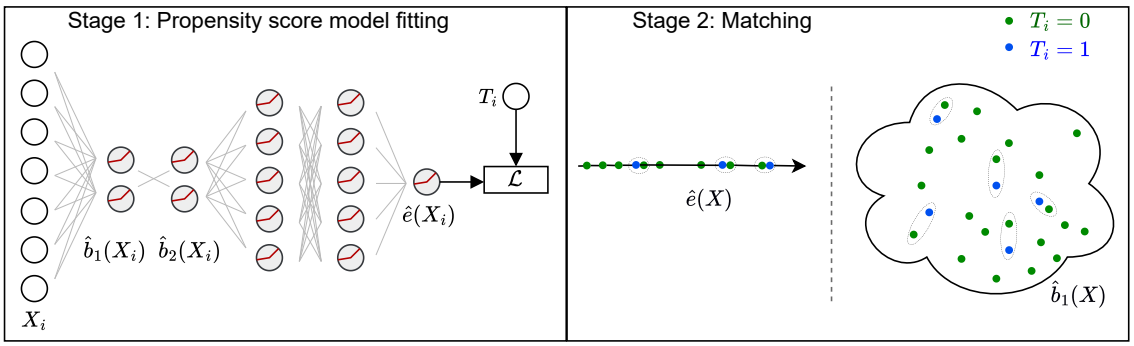


Figure 1: An illustration of neural score matching. In the first stage, a propensity score model is fitted to obtain low-dimensional balancing scores. In the second stage, samples are matched (to one neighbour) in the balancing score space based on a given distance metric. Matched samples are subsequently used to estimate the ATT.

covariates without regressing on outcomes. The intuition of neural score matching is illustrated in Fig. 1. (c) We show that neural score matching is competitive against other matching methods on two causal inference benchmarks in terms of calibration error, treatment effect estimation and balance.

2 MATCHING IN CAUSAL INFERENCE

2.1 Problem Setup

Let $(X_i, T_i, Y_i) \sim P$ be a dataset where X_i denotes (pre-treatment) covariates, T_i is the binary variable indicating whether the treatment under scrutiny has been applied to the subject or not, and Y_i is the observed outcome after the treatment or absence of treatment, all corresponding to subject i. In the potential outcomes framework (Rubin, 2005), $Y_i(1)$ is the outcome which would have happened (is "potential") if $T_i = 1$, and $Y_i(0)$ is the analogous outcome for when $T_i = 0$. Then, $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. We denote N_t as the number of treated units in the dataset, and N_c the number of control units. Our task is to estimate the *average treatment effect on the treated (ATT)*, defined as

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid T = 1].$$

This quantity measures the treatment effect for patients under treatment, and is typically the primary interest of medical applications (Ho et al., 2007). Here, covariates, such as age or BMI that are related to a treatment are of particular interest (Greifer and Stuart, 2021). The ATT can be approximated by the *sample average treatment effect on the treated (SATT)*, defined as

$$\text{SATT} = \frac{1}{N_t} \sum_{i:T_i=1} Y_i(1) - Y_i(0).$$

We make the following standard assumptions :

- Consistency: $\forall t,\ T_i = t \implies Y_i(t) = Y_i$.
- Ignorability: $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i | X_i$.
- Overlap: $\forall \boldsymbol{x},\ 0 < P(T_i = 1 | X_i = \boldsymbol{x}) < 1$.

Consistency ensures that $Y_i(1)$ is the observed outcome Y_i when $T_i = 1$. However, $Y_i(0)$ is not observed and must be estimated, for instance through matching. In addition, the ATT can also be expressed using *conditional average treatment effects* as

$$\text{ATT} = \mathbb{E}_X\Big[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\Big|T=1\Big], \quad (1)$$

which can be approximated by taking the sample mean over units as

$$\text{ATT} \approx \frac{1}{N_t} \sum_{i:T_i=1} \mathbb{E}[Y_i|T_i=1,X_i] - \mathbb{E}[Y_i|T_i=0,X_i]. \quad (2)$$

While we focus on the potential outcomes framework in this work, we note that an alternative is Pearl's framework of directed acyclic graphs (DAGs) and structural causal models (SCMs) (Pearl, 2009).

2.2 Key Concepts

In general, a matching procedure generates weights w_{ij} denoting the assignment of one or many control units j to a treated unit i (Morgan and Winship, 2014, Chapter 5). Typically, matching only assigns few control units, i.e. for a treated unit i, there is a small number of control units j such that $w_{ij} > 0$, and $w_{ij} = 0$, otherwise. This yields a new, weighted dataset $(w_i, X_i, T_i, Y_i) \sim P'$, where $w_i = 1$ for all treated units i and $w_j = \sum_i (w_{ij} / \sum_j w_{ij})$ for control units j. The matching procedure serves two main goals. One is to

estimate the ATT through the following estimator of the potential outcome $Y_i(0)$:

$$\hat{Y}_i(0) = \frac{1}{\sum_{j:T_j=0} w_{ij}} \sum_{j:T_j=0} w_{ij} Y_j.$$

Another is to obtain *balance* or, when it is not possible, reduce *imbalance* in P' compared to the original distribution P. Balance occurs when the distributions of covariates X given $T = 0$ on the one hand and $T = 1$ on the other hand are equal. Perfect balance thus corresponds to zero imbalance, and is desirable because it eliminates confounding. In this ideal setting, the treatment effect can then be estimated as the difference between averaged outcomes in both distributions. In this sense, the two goals of treatment effect estimation and balance are related. However, there is also a bias-variance trade-off at stake, as selecting fewer matching units will reduce imbalance and thus the expected treatment estimation error or "bias", at the cost of increased variance.

Formally, for a random variable A, we refer to the statement

$$P(A|T = 1) = P(A|T = 0)$$

as "balance in A", and, for a function D of two probability distributions, we refer to the quantity

$$D(P(A|T = 1), P(A|T = 0))$$

as "D-imbalance in A". When D is a probability distance, e.g. total variation or Wasserstein distance, then a zero D-imbalance in A implies balance in A. This is not true when D is *not* a probability distance, e.g. linear MMD. Note to distinguish D from the notation for a distance metric d in Section 3. We omit the mention of D or A when obvious from the context.

There are different ways to measure imbalance, such as a (standardised) difference in means (Austin, 2011), integral probability metrics such as the Wasserstein distance, the maximum mean discrepancy (MMD) and the total variation (TV) (Sriperumbudur et al., 2012; Kallus, 2020a), or histogram-based L_1 distances (Iacus et al., 2012). Balance and imbalance can also apply to other variables than covariates, such as transformations of covariates (Johansson et al., 2016; Shalit et al., 2017; Iacus et al., 2011).

3 RELATED WORK

We now discuss existing work on matching and alternative approaches in causal inference that aim to reduce imbalance or estimate the ATT. Most commonly, choosing matched control units j is done through a nearest neighbours search among all control units j according to some distance metric $d(.,.)$ (Stuart, 2010). Nearest neighbour search can be performed with or without replacement, and additionally, one may enforce a caliper, i.e. a maximal distance between matches. Alternatively, one might consider all matches simultaneously through an optimisation programme (optimal matching) (Rosenbaum, 1989). The choice of the distance metric d differs between common matching techniques:

- Exact matching (Rosenbaum and Rubin, 1985): $d(X_i, X_j) = \infty$, if $X_i \neq X_j$, and $d(X_i, X_j) = 0$, otherwise.
- Coarsened exact matching (Iacus et al., 2012): for a function f, $d(X_i, X_j) = \infty$, if $f(X_i) \neq f(X_j)$, and $d(X_i, X_j) = 0$, otherwise. f is typically an element-wise function, mapping to some (aggregated) value.
- Mahalanobis distance matching (Stuart, 2010): $d(X_i, X_j) = (X_i - X_j)^T \Sigma^{-1} (X_i - X_j)$, where Σ is the estimated covariance matrix of the control dataset in the case of ATT estimation.
- Propensity score matching (Austin, 2011): $d(X_i, X_j) = |\hat{e}(X_j) - \hat{e}(X_i)|$ where $\hat{e}(\boldsymbol{x})$ is an estimate of the propensity score $e(\boldsymbol{x}) := P(T = 1|X = \boldsymbol{x})$. This method is based on the property that $X \perp\!\!\!\perp T \mid e(X)$. We provide more details on implications of this property in Section 4.1.

Other than coarsened exact matching for which the weights have a different formulation, these methods set $w_{ij} = 1$ for matched units i and j, and $w_{ij} = 0$, otherwise.

All the above matching methods suffer from the *curse of dimensionality*, rendering them impractical in high-dimensional datasets. In general, theoretical results on nearest neighbour matching, to which the above techniques belong, show that the bias of the resulting ATT estimator grows with the data dimension D at a rate $\mathcal{O}(N^{-r/D})$, where N is the sample size and $r \geq 1$ is a constant (Abadie and Imbens, 2006b). More precisely, exact matching and coarsened exact matching remove more and more control items as the number of covariates increases. Further, matching based on the Mahalanobis distance performs poorly in high dimensions, likely because all covariate interactions are assumed to be equally important (Stuart, 2010).

In the literature, the preferred method for high dimensions is propensity score matching. However, compression into a single dimension can lead to matches with very different characteristics in the original covariate space, as for a fixed compression, there is no other information used to choose matches: matching is then done at random. This applies to all compressions of covariates, however as the propensity score $p(T = 1|X)$ is the coarsest compression which can be used for matching (see Section 4.1), with the least information from X, it

is most prone to actually matching at random. This can increase imbalance and consequently bias (King and Nielsen, 2019). Other than propensity score methods, approaches for matching in high dimensions include penalised regression techniques such as LASSO to perform variable selection before matching (Schneeweiss et al., 2009; Belloni et al., 2013; Farrell, 2015), sufficient dimension reduction (Luo and Zhu, 2020; Cheng et al., 2020), and distance metric learning (Li et al., 2016; Wang et al., 2021).

An alternative to matching is *weighting*, where weights w_j in the weighted dataset are directly estimated, generalising the problem formulation of matching (Kallus, 2020b). Examples include leveraging the estimated propensity score for inverse probability weighting (Horvitz and Thompson, 1952) or learning weights directly (Kallus, 2020a). A second alternative to matching is *outcome regression*. These methods estimate the quantity $\mathbb{E}[Y|T = t, X = \boldsymbol{x}]$ through a regressor $Q(t, \boldsymbol{x})$ that can be fitted through various methods related to linear regression (Imbens and Rubin, 2015), tree models (Athey et al., 2019), or neural networks (Shi et al., 2019; Shalit et al., 2017). Combining weighting through the propensity score estimate and outcome regression leads to the popular doubly robust methods, such as the augmented inverse probability weighted (AIPW) method (Robins et al., 1994). Recent efforts have been made to recategorise and benchmark outcome regression and doubly robust methods (Curth and Schaar, 2021).

4 NEURAL SCORE MATCHING

In this section, we present theoretical results that motivate the use of neural networks to obtain non-trivial, multivariate balancing scores. This approach aims to address the curse of dimensionality problem, as outlined in the previous section. In addition, some of these results have wider applicability to other models than neural networks.

4.1 Balancing Scores

We start by defining and analysing the use of *balancing scores*. This notion also motivated propensity score matching (Rosenbaum and Rubin, 1983).

Definition 1. *A* balancing score *is a function b of X such that $X \perp\!\!\!\perp T \mid b(X)$.*

As a consequence, for a fixed value $\boldsymbol{\beta}$ of $b(X)$, it holds that

$$P(X \mid b(X) = \boldsymbol{\beta}, T = 1) = P(X \mid b(X) = \boldsymbol{\beta}, T = 0),$$

i.e. the treatment and control distributions in the covariate space are equal for any fixed realisation of $b(X)$. Notably, it is possible to show that average treatment effects can be estimated by conditioning on $b(X)$ instead of X in Equation (1) (Rosenbaum and Rubin, 1983).

We can further connect (im)balance in $b(X)$ to (im)balance in X. The following Proposition shows that TV-imbalance in X is equal to TV-imbalance in $b(X)$, where TV is the total variation distance.

Proposition 1. *Let b be a function such that $b(X)$ is a balancing score. Then,*

$$\begin{aligned} &TV\left(P(X \mid T = 1), P(X \mid T = 0)\right) \\ &= TV\left(P(b(X) \mid T = 1), P(b(X) \mid T = 0)\right). \end{aligned}$$

Proof: *See Appendix A.1.* □

This allows us to potentially use lower-dimensional balancing scores $b(X)$ instead of high-dimensional covariates to achieve balance in X, as the following corollary shows that balance in $b(X)$ ensures balance in X :

Corollary 1.1. *Under the same conditions as Proposition 1,*

$$\begin{aligned} &P(b(X) \mid T = 1) = P(b(X) \mid T = 0) \\ &\implies P(X \mid T = 1) = P(X \mid T = 0). \end{aligned}$$

Proof: *See Appendix A.1.* □

Matching on a given balancing score $b(X)$ is commonly used to reduce imbalance in $b(X)$, with the aim of consequently reducing imbalance in X. Proposition 1 shows that a lower TV-imbalance in $b(X)$ will also mean a lower TV-imbalance in X, but only if $b(X)$ remains a balancing score in the post-matching distribution P'. Thankfully, the following Proposition shows that $b(X)$ remains a balancing score after matching.

Proposition 2. *Let b be a function such that $b(X)$ is a balancing score, P' be a distribution obtained from matching every treated unit with control units using $b(X)$ only. Then $b(X)$ is also a balancing score in P'.*

Proof: *See Appendix A.1.* □

Thus, all further theoretical results involving balancing scores in the original distribution will also be valid in the matched distribution. An important question left open at this point is how to find such a function b such that $b(X)$ is a balancing score.

Leveraging theoretical results in (Rosenbaum and Rubin, 1983), balancing scores can be linked to the propensity score $e(X) = P(T = 1 \mid X)$.

Proposition 3. *A function $b(X)$ is a balancing score, if and only if $b(X)$ can be mapped deterministically to the propensity score $e(X)$ through a function f, i.e.*

$$e(X) = f(b(X)).$$

Proof: See (Rosenbaum and Rubin, 1983, Thm. 2). □

It follows from Proposition 3 that $e(X)$ is itself a balancing score for the identity map. When this identity does not hold, $b(X)$ is said to be "finer" than $e(X)$, and conversely, $e(X)$ is "coarser" than $b(X)$. As noted in (Rosenbaum and Rubin, 1983), X is the finest balancing score, containing the most information; $e(X)$ is the coarsest balancing score, containing the least information; and any other $b(X)$ such that $e(X) = f(b(X))$ lies between the two. Choosing the degree of coarseness via multi-dimensional balancing scores to achieve optimal matching results rather than assuming a one-dimensional balancing score (i.e. the propensity score) is what we exploit in our method which we introduce in the following.

4.2 Introducing Neural Score Matching

Previous work has largely focused on the use of the propensity score $e(X)$ as a balancing score, and relatively little attention has been paid to non-trivial balancing scores that are neither X nor $e(X)$. Neural networks provide a natural mechanism by which to construct such balancing scores: fundamentally, a multi-layer neural network is a composition of functions $f_1, f_2, \dots, f_L$. Let us for a moment assume this network (perfectly) estimates the propensity score, i.e. $\hat{e}(X) = f_L \circ f_{L-1} \circ \dots \circ f_1(X) = e(X)$. Then, by Proposition 3, this provides us with $L+1$ balancing scores (X, the $L-1$ intermediate hidden representations and the estimated propensity score) that are coarser and coarser with increasing "depth" of the composition. We note that instead of neural networks parameterising $f_1, f_2, \dots, f_L$, one may consider other hierarchical models. We formalise this general principle, which we call *neural score matching*, in the following Proposition:

Proposition 4. *Assume that* $e(X) = f_L \circ f_{L-1} \circ \cdots \circ f_1(X)$ *for some functions* $f_1, \dots, f_L$. *Define* $b_0(X) := X$ *and* $b_l(X) = f_l \circ f_{l-1} \circ \cdots \circ f_1(X)$ *for* $l = 1, \dots, L$. *Then, every* $b_l(X)$ *is a balancing score, and for any* $l < L$, $b_{l+1}(X)$ *is coarser than* $b_l(X)$.

Proof: See Appendix A.2. □

Using this Proposition, we can now connect these balancing scores to our goal of achieving balance in X:

Corollary 4.1. *Under the same conditions and notation as Proposition 4, for any* $l, l' = 0, \dots, L$,

$$\begin{aligned} &TV\left(P(b_l(X) \mid T=1), P(b_l(X) \mid T=0)\right) \\ &= TV\left(P(b_{l'}(X) \mid T=1), P(b_{l'}(X) \mid T=0)\right), \end{aligned}$$

and balance in $b_{l'}(X)$ *is equivalent to balance in* $b_l(X)$.

Proof: See Appendix A.2. □

This Proposition gives us a choice of balancing scores with varying degree of coarseness which we can use for matching. Note that achieving balance in *any* of the scores will yield balance in *all* of them, and particularly in $X = b_0(X)$. On the other hand, perfect balance is difficult to attain, but we can still aim to achieve the lowest imbalance possible. Importantly, although imbalance is identical for two given balancing scores in the same hierarchical propensity score model *when the distribution is fixed*, matching on these two balancing scores will in general result in different distributions and consequently different imbalances.

Thus, if we can compute TV-imbalances, the Proposition ensures that selecting the balancing score and matching procedure with the lowest resulting TV-imbalance will also reach the lowest TV-imbalance in covariate distributions X among the candidate balancing scores and matching procedures.

It is important to note that Proposition 4, Corollary 4.1 and the following theoretical results all assume that $\hat{e}(X) = e(X)$, i.e. a well-calibrated propensity score model, or at least that the obtained scores are indeed balancing scores. In Section 4.4, we will relax this assumption and provide theoretical bounds when scores violate the balancing score assumption from Definition 1.

In practice, however, the total variation distance is not suitable for this purpose of balancing score comparison due to the difficulties with estimating it in finite samples (Kallus, 2020a). We provide results with alternative metrics which overcome this issue in Section 4.3

4.3 Bounds With Estimable Integral Probability Metrics

We start with a general inequality that shows that any imbalance in X measured using an integral probability metric (IPM) is also upper-bounded by such an imbalance in $b(X)$.

Proposition 5. *Let* $\mathcal{F}$ *be a set of functions of* X. *For any function* b *of* X, *define*

$$\mathcal{F}_b := \{\ \boldsymbol{\beta} \mapsto \mathbb{E}[f(X) \mid b(X) = \boldsymbol{\beta}], \quad f \in \mathcal{F}\}\,.$$

Then, for any balancing score $b(X)$ *and any set* $\mathcal{G}$ *of functions on the image set of* b *such that* $\mathcal{F}_b \subseteq \mathcal{G}$,

$$\begin{aligned} &IPM_{\mathcal{F}}\left(P(X \mid T=1), P(X \mid T=0)\right) \\ &\leq IPM_{\mathcal{G}}\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right) \end{aligned}$$

with equality when $\mathcal{G} = \mathcal{F}_b$.

As a result, any measure of imbalance of original covariates based on an IPM, including using popular ones

such as the linear MMD or the Wasserstein distance, can be controlled using another measure of imbalance depending on an IPM. Thus, as in Corollary 4.1, we expect to reduce any IPM-imbalance in X when reducing another IPM-imbalance in $b(X)$, further justifying matching on $b(X)$ as an alternative to matching on X when the measure of interest for imbalance in X is an IPM. Further, if we had access to the $\text{IPM}_{\mathcal{F}_b}$-imbalance in $b(X)$, we could again use it to select the appropriate balancing score, as for the total variation distance. One caveat is that it is precisely unclear *which* IPM-imbalance in $b(X)$ is suitable in Proposition 5 as the class $\mathcal{F}_b$ is non-trivial due to the conditional expectation in $b(X)$, even for common base classes $\mathcal{F}$ such as linear or Lipschitz functions. Thus, the question remains whether we can bound the IPM-imbalance of X using a *computable* IPM-imbalance.

To solve this, we consider a *linear* balancing score $b(X)$, meaning that b is a linear function. For example, this can be realised by considering the first layer of a neural network before applying an activation function. In this simple case, and under strong assumptions on the distribution of X, we can leverage popular integral probability metrics which *can* be estimated with finite samples, in contrast to the total variation distance.

Proposition 6. *Let b be a function such that $\forall \boldsymbol{x},\ b(\boldsymbol{x}) = W\boldsymbol{x}$ for some matrix W and $b(X)$ is a balancing score. Let $\|.\|$ be the Euclidean norm on any vector space, and $|||.|||$ be a norm[1] on any matrix space such that $\forall \boldsymbol{x}, A, \|A\boldsymbol{x}\| \leq |||A||| \cdot \|\boldsymbol{x}\|$. Further, let A^+ be the Moore-Penrose pseudo-inverse of A, Wass be the Wasserstein distance, MMD be the linear MMD[2]. Let $W_\Sigma^+ := \Sigma W^T (W\Sigma W^T)^+$. If X is elliptical with covariance matrix Σ then*

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot MMD\big(P(b(X) \mid T=1), P(b(X) \mid T=0)\big) \\
&\leq MMD\big(P(X \mid T=1), P(X \mid T=0)\big) \\
&\leq |||W_\Sigma^+||| \cdot MMD\big(P(b(X) \mid T=1), P(b(X) \mid T=0)\big)
\end{aligned}$$

If X is Gaussian with positive-definite covariance matrix Σ and W has full row rank then

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot Wass\big(P(b(X) \mid T=1), P(b(X) \mid T=0)\big) \\
&\leq Wass\big(P(X \mid T=1), P(X \mid T=0)\big) \\
&\leq |||W_\Sigma^+||| \cdot Wass\big(P(b(X) \mid T=1), P(b(X) \mid T=0)\big)\,.
\end{aligned}$$

This Proposition provides lower- and upper-bounds (in contrast to Proposition 1) for the Wasserstein- or linear MMD-imbalance in X which depend linearly on the corresponding imbalance in $b(X)$.

One could exploit these bounds by computing them for different balancing scores and choose the one with the lowest (lower or upper) bounds of the interval, or the narrowest bounds. One might also perform a type of optimal matching minimising the Wasserstein or linear MMD imbalance in $b(X)$. However, it is important to point out that these bounds may be wide depending on the singular values of W. For example, assume $\Sigma = I$, then $W_\Sigma^+ = W^+$. When using the operator norm and denoting $\sigma_{\min}(W)$ and $\sigma_{\max}(W)$ as the minimal and maximal non-zero singular values of W [3], respectively, we have $\frac{1}{|||W|||} = \frac{1}{\sigma_{\max}(W)}$ and $|||W^+||| = \frac{1}{\sigma_{\min}(W)}$. As a consequence, values within the bounds can vary by a factor of $\frac{\sigma_{\max}(W)}{\sigma_{\min}(W)}$. Further, the strong assumptions on the distribution of X might not hold in practice, especially in the post-matching distribution.

In addition, the imbalance in $b(X)$ might also help speed up computations. In Appendix B, we show how the computational complexity of the estimators of the Wasserstein distance can be reduced on a lower-dimensional space.

From the insights of Proposition 6, we only use the first layer of a neural network for the purpose of matching; the other layers serve to achieve a better fit of the propensity score model.

4.4 Bounds For Non-Balancing Scores

As mentioned above, a requirement for applying the above Propositions within the context of hidden representations of a neural network is that either the estimated propensity score of said network equals the true propensity score, or more generally, every learned function b is indeed a balancing score. When this is not the case, as the next Proposition shows, we can still bound the imbalance in X in terms of the imbalance in $b(X)$ and some quantification of "how much" the assumption $X \perp\!\!\!\perp T | b(X)$ is violated.

Proposition 7. *Let*

$$\mathcal{E}_{t,b}^D(\boldsymbol{\beta}) := D\Big(P\big(X|b(X)=\boldsymbol{\beta}, T=t\big), P\big(X|b(X)=\boldsymbol{\beta}\big)\Big)$$

where D is a probability discrepancy measure, b is a function of X, $t \in \{0,1\}$ is a realisation of T, $\boldsymbol{\beta}$ is a realisation of $b(X)$. For any function b,

$$\begin{aligned}
&TV\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&\leq TV\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \\
&\leq TV\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big)
\end{aligned}$$

[1] Examples include the operator norm or the Euclidean norm.

[2] Note that these theoretical results also hold when $b(X)$ has a bias term.

[3] This assumes $W \neq 0$, i.e. we do not have balance in X.

$$+ \mathbb{E}\big[\mathcal{E}^{TV}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{TV}_{0,b}\big(b(X)\big)|T=0\big]$$

and, using the notations of Proposition 5,

$$\begin{aligned}
&IPM_{\mathcal{F}}\left(P(X \mid T=1), P(X \mid T=0)\right) \\
&\leq IPM_{\mathcal{F}_b}\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right) \\
&+ \mathbb{E}\big[\mathcal{E}^{IPM_{\mathcal{F}}}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{IPM_{\mathcal{F}}}_{0,b}\big(b(X)\big)|T=0\big].
\end{aligned}$$

For a linear function $b(x) = Wx$, if X is elliptical with covariance matrix Σ, then

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot MMD\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&\leq MMD\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \\
&\leq |||W^{+}_{\Sigma}||| \cdot MMD\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&+ \mathbb{E}\big[\mathcal{E}^{MMD}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{MMD}_{0,b}\big(b(X)\big)|T=0\big]
\end{aligned}$$

and if X is Gaussian with positive-definite covariance matrix Σ while W has full row rank, then

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot Wass\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&\leq Wass\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \\
&\leq |||W^{+}_{\Sigma}||| \cdot Wass\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&+ \mathbb{E}\big[\mathcal{E}^{Wass}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{Wass}_{0,b}\big(b(X)\big)|T=0\big].
\end{aligned}$$

Unlike the calibration error, i.e. the mean difference between true and predicted propensity scores, the extra balancing error term in the Proposition does not rely on access to the true propensity score. Therefore, it could be computed and used to obtain an upper bound of covariate imbalance in any dataset. In practice, however, it might be challenging to estimate as it relies on conditional expectations for which few samples may be available.

5 EXPERIMENTS

We now evaluate neural score matching on two semi-synthetic datasets and benchmark it against other matching methods. We provide code to implement neural score matching and reproduce the main results at https://github.com/oscarclivio/neuralscorematching.

5.1 Experimental Setup

Our general procedure for matching and in particular neural score matching follows two stages: in the first stage, we learn a model to obtain some representation or score s from datapoints. In the second stage, we perform matching on these scores using the Euclidean distance[4] $d(s_i, s_j) = \|s_i - s_j\|_2$. We use nearest neighbour matching with replacement using one neighbour.

To perform neural score matching, we train a neural network predicting treatment assignment from covariates, with the final one-dimensional layer being an estimator of the propensity score. Training is performed using a standard binary cross-entropy loss. The neural network has the following architecture: one low-dimensional layer with 5 hidden units, two layers with 100 units and one final 1-dimensional layer. We use the leaky ReLU activation function in all layers except the last one where we use the sigmoid function. When using the hidden representation in the first layer before applying the activation function as a score, we refer to the resulting method as `NN Layer 1`. Notably, if it is indeed a balancing score, it meets the assumptions of Proposition 6. From the insights of this Proposition, we choose to focus on one single multivariate layer for matching, and dedicate other layers to model fitting (with corresponding high dimensions as given above). The final activation of the network estimates the propensity score and is also used for matching as a balancing score. We refer to it as `NN PS`.

We benchmark these scores obtained by the neural network against other scores, namely X (`X`) and a five-dimensional PCA reduction of X (`PCA`). We also benchmark against a logistic regression estimating the propensity score given X or PCA features, which we refer to as `LogReg PS` and `PCA + LogReg PS`, respectively. In addition, we consider matching uniformly at random (`Random matching`) and leaving the treatment and control datasets unchanged w.r.t. balance by not matching at all (`No Matching`). All methods were evaluated using 10 different training random seeds.

We use variants of two standard datasets for treatment effect estimation: *ACIC 2016* and *News*. Both datasets have a large number of covariates (82 and 3477, respectively), rendering them challenging for standard matching techniques. They are both semi-synthetic: the covariates come from real-world studies, while the treatments and outcomes were generated through a data generating process. For every dataset, we will average results over different draws of the data generating process (100 for ACIC 2016 and 50 for News). Early stopping was used on News. Results on a third dataset, IHDP, are presented in Appendix D.

To evaluate the methods, we report three metrics: *calibration error*, defined as the mean absolute difference between the estimated and true propensity score, *ATT error*, defined as the absolute difference between the

[4]We use the Euclidean distance as the Mahalanobis distance was prohibitively slow to compute in high-dimensional and large sample settings.

ATT estimated by the method and a ground-truth ATT, and *sample imbalance* $\hat{I}$, defined as the squared Euclidean distance between sample means of covariates of treated and control groups from the dataset $\mathcal{D}'$ obtained from the original dataset $\mathcal{D}$ after matching. To reliably assess the performance of the methods under investigation, we average and present standard deviations over the repeated draws of the data generating processes and additionally over the different random seeds for model fitting/training.

We refer to Appendix C for further details about implementation and experimental setup.

5.2 Experimental Results

In this section, we present our experimental results as Tables (and refer to Appendix E for their visualisation as boxplots).

5.2.1 ACIC 2016

Results for the different matching methods under consideration are presented in Table 1. Propensity score models for the two dimensionality reduction methods (`NN Layer 1` and `PCA`) have better calibration than the standard logistic-regression propensity score (`LogReg PS`), with a slight advantage for `NN PS`. The relevance of using a multivariate score is demonstrated: on ATT errors and imbalances, `NN Layer 1` most often outperforms `NN PS`, and all other methods except:

- Logistic regression propensity score (`LogReg PS`) on in-sample metrics. It is possible that the dimensionality remains sufficiently low for this method to handle (unlike News, see next section). However, the method might also overfit, as shown by the hold-out performance.
- `No Matching` and `PCA` on hold-out imbalances. Neural scores might need better generalisation as they increase imbalance compared to the original dataset, unlike `PCA`. Other methods also increase imbalance, as expected.

5.2.2 News

Results for the News dataset are presented in Table 2. Multivariate dimensionality-reduced scores (`NN Layer 1` and `PCA`) generally outperform their respective propensity scores (except `NN Layer 1` and `NN PS` having similar performance on ATT errors), as well as `Random matching`, `X` and `LogReg PS`. The two latter have particularly high ATT errors and imbalances, even compared to `Random matching`. This shows that multivariate, but lower-dimensional scores can improve matching on high-dimensional datasets. The performance is more balanced between `PCA` and `NN Layer 1`: `PCA` is better on imbalances, `NN Layer 1` on in-sample ATT errors, and their hold-out ATT errors are not significantly different according to standard errors.

Table 1: Results on the ACIC2016 dataset.

Calibration errors	In-Sample	Hold-Out
`NN PS` (ours)	0.055±0.000	0.055±0.000
`LogReg PS`	0.067±0.000	0.069±0.000
`PCA + LogReg PS`	0.058±0.001	0.058±0.001
ATT errors	**In-Sample**	**Hold-Out**
`NN Layer 1` (ours)	0.707±0.012	0.918±0.018
`NN PS` (ours)	0.735±0.012	1.008±0.019
`X`	0.848±0.018	0.990±0.019
`Random matching`	1.209±0.019	1.301±0.023
`LogReg PS`	0.678±0.012	1.036±0.018
`PCA`	0.927±0.016	1.007±0.020
`PCA + LogReg PS`	0.962±0.016	1.097±0.021
Sample imbalance	**In-Sample**	**Hold-Out**
`NN Layer 1` (ours)	0.107±0.001	0.422±0.003
`NN PS` (ours)	0.105±0.001	0.498±0.004
`X`	0.438±0.002	0.739±0.004
`Random matching`	0.232±0.003	0.558±0.006
`LogReg PS`	0.056±0.001	0.511±0.004
`PCA`	0.117±0.001	0.342±0.003
`PCA + LogReg PS`	0.134±0.001	0.488±0.004
`No Matching`	0.192±0.003	0.396±0.006

6 DISCUSSION AND CONCLUSION

In this work, we have provided novel theoretical results motivating neural score matching: using neural networks to obtain balancing scores which can be readily used for matching. In contrast to lower-dimensional representations obtained from classical dimensionality reduction techniques (e.g. PCA), our method estimates lower-dimensional balancing scores as defined in Proposition 3, which can be mapped back to the propensity score "for free" due to the inherent compositionality of neural networks, allowing more flexibility in choosing the degree of coarseness. This applies only if the model is correctly specified, however. Proposition 7 paves the way to rigorous analysis of situations when the constraint is violated. We found that in popular semi-synthetic datasets, neural score matching is competitive against other matching methods. In addition, our results indicate the general utility of dimensionality reduction techniques for matching in causal inference. This leads the way towards learning suitable represen-

Table 2: Results on the News dataset.

ATT errors	In-Sample	Hold-Out
`NN Layer 1` (ours)	0.071±0.002	0.106±0.004
`NN PS` (ours)	0.073±0.002	0.105±0.004
`X`	0.510±0.015	0.765±0.024
`Random matching`	0.100±0.003	0.114±0.004
`LogReg PS`	1.460±0.052	0.505±0.020
`PCA`	0.080±0.002	0.103±0.003
`PCA + LogReg PS`	0.095±0.003	0.100±0.003
Sample imbalance	In-Sample	Hold-Out
`NN Layer 1` (ours)	1.518±0.022	3.886±0.045
`NN PS` (ours)	2.104±0.035	5.105±0.079
`X`	12.531±0.032	18.178±0.052
`Random matching`	2.121±0.041	4.581±0.043
`LogReg PS`	371.070±36.672	131.192±4.682
`PCA`	1.097±0.013	3.608±0.030
`PCA + LogReg PS`	1.444±0.017	4.600±0.046
`No Matching`	1.844±0.040	3.432±0.038

tations for matching which might be useful for downstream tasks to gain scientific insight, notably in areas where the use of neural networks is ubiquitous, such as medical imaging (Zhou et al., 2021), text classification (Minaee et al., 2021) and audio processing (Purwins et al., 2019).

Our work has the following two limitations: 1) It is difficult to properly specify and train neural networks for the task of matching. In particular, there is a trade-off between finding low-dimensional balancing scores, which implies low-dimensional hidden layers, and fitting the propensity score model, which implies wide hidden layers not suitable for matching. We also did not find hyperparameters that performed consistently better than others across all datasets, nor a correlation between matching performance and hold-out loss. More complex architectures than our experimental setup and a deeper understanding of the hyperparameter space should be explored. 2) Most of our theoretical results assume the propensity score model is correct, or, more generally speaking, that the hidden layers are indeed balancing scores. Most often, neither is true. Proposition 7 is a first step towards theoretical guarantees for scores that are not perfectly balancing.

Future work will investigate the following ideas: 1) As outlined earlier, our propensity score model might be miscalibrated and the balancing scores might not perfectly balance covariates. Empirically measuring calibration error and the violation of the balancing score property via Proposition 7, we aim at using this to inform model training and hence improve performance. 2) We plan to extend the relatively simple setup of neural score matching as presented here to, for instance, using multiple intermediate balancing scores. This entails further questions, such as how to choose the degree of coarseness of the balancing scores, which might be assessed via empirical out-of-sample evaluation, and where to best place layers with few hidden units that are suited for matching. 3) We aim to develop a form of optimal matching which uses more general bounds of Wass- or MMD-imbalance in $b(X)$ than those of Proposition 6, and use them directly in a loss function, which in turn should reduce imbalance in X. 4) Our obtained balancing scores might enable the use of coarsened exact matching (CEM), offering the possibility to pre-specify the desired level of imbalance before matching (Iacus et al., 2012). 5) We aim to explore more in depth how intermediate balancing scores compare to propensity scores, e.g. by visualising how their spaces capture features of the covariate space. We also expect these intermediate balancing scores to be preferable to propensity scores for CATE estimation as they provide less coarse representations of covariates.

Acknowledgements

O.C. is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and Novo Nordisk. F.F. acknowledges the receipt of a studentship award from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z). B.L. was supported by the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme (grant number EP/R018561/1) and gratefully acknowledges funding from Jesus College, Oxford. C.H. acknowledges support from the Medical Research Council Programme Leaders award MC_UP_A390_1107, The Alan Turing Institute, Health Data Research, U.K., and the U.K. Engineering and Physical Sciences Research Council through the Bayes4Health programme grant.

We would like to thank the anonymous reviewers for helpful feedback.

References

Abadie, A. and Imbens, G. W. (2006a). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2006b). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424. PMID: 21818162.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies*, 81(2):608–650.

Bertsekas, D. P. (1998). Network optimization: Continuous and discrete models.

Cheng, D., Li, J., Liu, L., and Liu, J. (2020). Sufficient dimension reduction for average causal effect estimation. *arXiv preprint arXiv:2009.06444*.

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

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Farzadfar, F., Murray, C. J., Gakidou, E., Bossert, T., Namdaritabar, H., Alikhani, S., Moradi, G., Delavari, A., Jamshidi, H., and Ezzati, M. (2012). Effectiveness of diabetes and hypertension management by rural primary health-care workers (behvarz workers) in iran: a nationally representative observational study. *The Lancet*, 379(9810):47–54.

Greifer, N. and Stuart, E. A. (2021). Choosing the estimand when matching or weighting in observational studies. *arXiv preprint arXiv:2106.10577*.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

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

Iacus, S. M., King, G., and Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493):345–361.

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

Imbens, G. W. and Rubin, D. B. (2015). *Regression Methods for Completely Randomized Experiments*, page 113–140. Cambridge University Press.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR.

Kallus, N. (2020a). DeepMatch: Balancing deep covariate representations for causal inference using adversarial training. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5067–5077. PMLR.

Kallus, N. (2020b). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.

Li, S., Vlassis, N., Kawale, J., and Fu, Y. (2016). Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*.

Luo, W. and Zhu, Y. (2020). Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, 38(4):888–900.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2 edition.

Pearl, J. (2009). *Causality*. Cambridge University Press, 2 edition.

Peyré, G. and Cuturi, M. (2020). Computational optimal transport.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Razonable, R. R., Pawlowski, C., O'Horo, J. C., Arndt, L. L., Arndt, R., Bierle, D. M., Borgen, M. D., Hanson, S. N., Hedin, M. C., Lenehan, P., et al. (2021). Casirivimab–imdevimab treatment is associated with reduced rates of hospitalization among high-risk patients with mild to moderate coronavirus disease-19. *EClinicalMedicine*, page 101102.

Roberts, M. E., Stewart, B. M., and Nielsen, R. A. (2020). Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

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

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

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

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

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

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

Shi, C., Blei, D. M., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *NeurIPS*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599.

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

Su, M., Zhou, Z., Si, Y., and Wei, X. (2019). Effect of health alliances on the quality of primary care in urban china: A coarsened exact matching difference-in-differences analysis. *The Lancet*, 394:S86.

Wang, T., Morucci, M., Awan, M. U., Liu, Y., Roy, S., Rudin, C., and Volfovsky, A. (2021). Flame: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, 22(31):1–41.

Webb, G. J., Marjot, T., Cook, J. A., Aloman, C., Armstrong, M. J., Brenner, E. J., Catana, M.-A., Cargill, T., Dhanasekaran, R., García-Juárez, I., et al. (2020). Outcomes following sars-cov-2 infection in liver transplant recipients: an international registry study. *The lancet Gastroenterology & hepatology*, 5(11):1008–1016.

Zhou, S., Greenspan, H., Davatzikos, C., Duncan, J., Van Ginneken, B., Madabhushi, A., Prince, J., Rueckert, D., and Summers, R. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the Institute of Radio Engineers*, 109(5):820–838.

Neural Score Matching for High-Dimensional Causal Inference: Appendices

A PROOFS OF THEORETICAL RESULTS

A.1 Balance on $b(X)$ and X

Proposition 1. *Let b be a function such that $b(X)$ is a balancing score. Then,*

$$\begin{aligned} &TV\left(P(X \mid T=1), P(X \mid T=0)\right) \\ &= TV\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right). \end{aligned}$$

Proof:

- First, let us note that for any random variable V, and by definition of the total variation distance:

$$\begin{aligned} TV(P(V|T=1), P(V|T=0)) &= \sup_{\|f\|_{L^\infty} \leq 1} \left|\mathbb{E}[f(V)|T=1] - \mathbb{E}[f(V)|T=0]\right| \\ &= \mathrm{IPM}_{\{f:\ \|f\|_{L^\infty} \leq 1\}}(P(V|T=1), P(V|T=0)), \end{aligned}$$

  where IPM is defined in Equation S6, $\|\cdot\|_{L^\infty}$ is the uniform norm, and f is a function.

- For any function f on the $\mathcal{B}$ space (i.e. the image space of $b(X)$) such that $\|f\|_{L^\infty} < 1$:

$$\begin{aligned} \left|\mathbb{E}[f(b(X))|T=1] - \mathbb{E}[f(b(X))|T=0]\right| &= \left|\mathbb{E}[(f \circ b)(X)|T=1] - \mathbb{E}[(f \circ b)(X)|T=0]\right| \\ &\leq TV(P(X|T=1), P(X|T=0)) \end{aligned}$$

  as $(f \circ b)$ is a function on the $\mathcal{X}$ space (i.e. the image space of X) with $\|f\|_{L^\infty} < 1$. Thus, taking the supremum over all such functions f,

$$TV(P(b(X)|T=1), P(b(X)|T=0)) \leq TV(P(X|T=1), P(X|T=0)).$$

- Let $g(\boldsymbol{\beta}) := \mathbb{E}[f(X)|b(X)=\boldsymbol{\beta}]$. We show that g is a function on the $\mathcal{B}$ space with $\|g\|_{L^\infty} < 1$:

$$\begin{aligned} \forall \boldsymbol{\beta}, |g(\boldsymbol{\beta})| &= \left|\mathbb{E}[f(X) \mid b(X)=\boldsymbol{\beta}]\right| \\ &\leq \mathbb{E}[|f(X)| \ \mid \ b(X)=\boldsymbol{\beta}] \text{ from Jensen's inequality} \\ &\leq \mathbb{E}[1 \mid b(X)=\boldsymbol{\beta}] \text{ as } \|f\|_{L^\infty} < 1 \\ &= 1. \end{aligned}$$

  Therefore, as a consequence of Proposition 5,

$$TV(P(X|T=1), P(X|T=0)) \leq TV(P(b(X)|T=1), P(b(X)|T=0)).$$

- Consequently, it follows that

$$TV(P(X|T=1), P(X|T=0)) = TV(P(b(X)|T=1), P(b(X)|T=0)).$$

□

Corollary 1.1. *Under the same conditions as Proposition 1,*

$$P(b(X) \mid T = 1) = P(b(X) \mid T = 0) \implies P(X \mid T = 1) = P(X \mid T = 0).$$

Proof: One should note that $P(b(X) \mid T = 1) = P(b(X) \mid T = 0)$ implies $TV(P(b(X)|T = 1), P(b(X)|T = 0)) = 0$. As a consequence, $TV(P(X|T = 1), P(X|T = 0)) = 0$ from Proposition 1. As the total variation is a distance, we obtain that $P(X|T = 1) = P(X|T = 0)$. □

Proposition 2. *Let b be a function such that $b(X)$ is a balancing score, P' be a distribution obtained from matching every treated unit with control units using $b(X)$ only. Then $b(X)$ is also a balancing score in P'.*

Proof: Let $\boldsymbol{\beta}$ be a value of $b(X)$. Any matching method using $b(X)$ only to match units does not change the conditional distribution of X given $b(X) = \boldsymbol{\beta}$ in the control group (Rosenbaum and Rubin, 1983). Then,

$$P'(X|b(X) = \boldsymbol{\beta}, T = 0) = P(X|b(X) = \boldsymbol{\beta}, T = 0). \tag{S3}$$

The conditional distribution of X given $b(X) = \boldsymbol{\beta}$ in the treated group is likewise left unchanged as the matching method does not change the treated distribution in any way. Thus,

$$P'(X|b(X) = \boldsymbol{\beta}, T = 1) = P(X|b(X) = \boldsymbol{\beta}, T = 1). \tag{S4}$$

Also, as $b(X)$ is a balancing score in P, $P(X|b(X) = \beta, T = 0) = P(X|b(X) = \beta, T = 1)$. Tying it all together, we have

$$\begin{aligned} P'(X|b(X) = \beta, T = 0) &= P(X|b(X) = \beta, T = 0) && \text{from Equation S3} \\ &= P(X|b(X) = \beta, T = 1) && \text{as } b(X) \text{ is a balancing score.} \\ &= P'(X|b(X) = \beta, T = 1) && \text{from Equation S4} \end{aligned}$$

so $b(X)$ is a balancing score in P'. □

A.2 Further Balancing Scores

Proposition 4. *Assume that $e(X) = f_L \circ f_{L-1} \circ \cdots \circ f_1(X)$ for some functions $f_1, \ldots, f_L$. Define $b_0(X) \coloneqq X$ and $b_l(X) = f_l \circ f_{l-1} \circ \cdots \circ f_1(X)$ for $l = 1, \ldots, L$. Then, every $b_l(X)$ is a balancing score, and for any $l < L$, $b_{l+1}(X)$ is coarser than $b_l(X)$.*

Proof: According to Proposition 3, $b_l(X)$ where $l < L$ is a balancing score as

$$e(X) = f_L \circ f_{L-1} \circ \ldots \circ f_{l+1}(b_l(X)),$$

and $e(X)$ is the propensity score with the property $X \perp\!\!\!\perp T|e(X)$. Also for any $l < L$, $b_{l+1}(X)$ is coarser than $b_l(X)$ as $b_{l+1}(X) = f_{l+1}(b_l(X))$.

Corollary 4.1. *Under the same conditions and notation as Proposition 4, for any $l, l' = 0, \ldots, L$,*

$$\begin{aligned} &TV\left(P(b_l(X) \mid T = 1), P(b_l(X) \mid T = 0)\right) \\ &= TV\left(P(b_{l'}(X) \mid T = 1), P(b_{l'}(X) \mid T = 0)\right), \end{aligned}$$

and balance in $b_{l'}(X)$ is equivalent to balance in $b_l(X)$.

Proof: First, for any $l < L$, as $b_{l+1}(X)$ is a balancing score w.r.t. $b_l(X)$ from Proposition 4, we note that Proposition 1 can also be applied to $b_l(X)$ and $b_{l+1}(X)$ instead of X and $b(X)$, respectively. Thus, it follows from Proposition 1 that

$$TV(P(b_{l+1}(X)|T = 1), P(b_{l+1}(X)|T = 0)) = TV(P(b_l(X)|T = 1), P(b_l(X)|T = 0)).$$

Consequently, it follows by induction that for any $l, l' = 0, \ldots, L$,

$$TV(P(b_l(X)|T = 1), P(b_l(X)|T = 0)) = TV(P(b_{l'}(X)|T = 1), P(b_{l'}(X)|T = 0)).$$

Then, the proof that balance in $b_{l'}(X)$ is equivalent to balance in $b_l(X)$ is analogous to Corollary 1.1. □

A.3 Other Integral Probability Metrics

Proposition 5. *Let $\mathcal{F}$ be a set of functions of X. For any function b of X, define*

$$\mathcal{F}_b := \{\ \boldsymbol{\beta} \mapsto \mathbb{E}[f(X) \mid b(X) = \boldsymbol{\beta}],\quad f \in \mathcal{F}\}\,.$$

Then, for any balancing score $b(X)$ and any set $\mathcal{G}$ of functions on the image set of b such that $\mathcal{F}_b \subseteq \mathcal{G}$,

$$\begin{aligned}&IPM_{\mathcal{F}}\left(P(X \mid T=1), P(X \mid T=0)\right)\\ &\leq IPM_{\mathcal{G}}\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right)\end{aligned}$$

with equality when $\mathcal{G} = \mathcal{F}_b$.

Proof. As $b(X)$ is a balancing score, we have $T \perp\!\!\!\perp X | b(X)$ and for any measurable function f:

$$\mathbb{E}[f(X) \mid b(X), T] = \mathbb{E}[f(X) \mid b(X)] \tag{S5}$$

Also, by definition, for any random variable V,

$$\mathrm{IPM}_{\mathcal{F}}(P(V|T=1), P(V|T=0)) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}[f(V) \mid T=1] - \mathbb{E}[f(V) \mid T=0]\right|. \tag{S6}$$

Let f be a measurable function, then

$$\begin{aligned}\mathbb{E}[f(X) \mid T=t] &= \mathbb{E}\Big[\mathbb{E}[f(X) \mid b(X), T=t] \ \mid\ T=t\Big] \text{ due to the law of total expectation}\\ &= \mathbb{E}\Big[\mathbb{E}[f(X) \mid b(X)] \ \mid\ T=t\Big] \text{ due to Equation (S5)}\\ &= \mathbb{E}\Big[g(b(X)) \mid T=t\Big],\end{aligned} \tag{S7}$$

where $g(\boldsymbol{\beta}) = \mathbb{E}[f(X) \mid b(X) = \boldsymbol{\beta}]$. By definition, if $f \in \mathcal{F}$ then $g \in \mathcal{F}_b$ and, by definition of $\mathcal{G}$, $g \in \mathcal{G}$. Thus, for any $f \in \mathcal{F}$,

$$\begin{aligned}\left|\mathbb{E}[f(X) \mid T=1] - \mathbb{E}[f(X) \mid T=0]\right| &= \left|\mathbb{E}[g(b(X)) \mid T=1] - \mathbb{E}[g(b(X)) \mid T=0]\right| \text{ for some } g \in \mathcal{G}\\ &\leq \mathrm{IPM}_{\mathcal{G}}(P(b(X) \mid T=1), P(b(X) \mid T=0)).\end{aligned}$$

Taking the supremum wrt $\mathcal{F}$ on the LHS gives that

$$\mathrm{IPM}_{\mathcal{F}}(P(X|T=1), P(X|T=0)) \leq \mathrm{IPM}_{\mathcal{G}}(P(b(X) \mid T=1), P(b(X) \mid T=0)).$$

Further, let $g \in \mathcal{F}_b$. By definition, there exists $f \in \mathcal{F}$, such that $g(\boldsymbol{\beta}) = \mathbb{E}[f(X) \mid b(X) = \boldsymbol{\beta}]$. By Equation S7,

$$\forall t = 0, 1,\quad \mathbb{E}[g(b(X)) \mid T=t] = \mathbb{E}[f(X) \mid T=t].$$

Thus,

$$\begin{aligned}|\mathbb{E}[g(b(X)) \mid T=1] - \mathbb{E}[g(b(X)) \mid T=0]| &= |\mathbb{E}[f(X) \mid T=1] - \mathbb{E}[f(X) \mid T=0]| \text{ for some } f \in \mathcal{F}\\ &\leq \mathrm{IPM}_{\mathcal{F}}(P(X|T=1), P(X|T=0)).\end{aligned}$$

Taking the supremum wrt $\mathcal{F}_b$ on the LHS gives

$$\mathrm{IPM}_{\mathcal{F}_b}(P(b(X) \mid T=1), P(b(X) \mid T=0)) \leq \mathrm{IPM}_{\mathcal{F}}(P(X|T=1), P(X|T=0)),$$

concluding the proof. □

Proposition 6. *Let b be a function such that $\forall \boldsymbol{x},\ b(\boldsymbol{x}) = W\boldsymbol{x}$ for some matrix W and $b(X)$ is a balancing score. Let $\|.\|$ be the Euclidean norm on any vector space, and $|||.|||$ be a norm[5] on any matrix space such that*

[5]Examples include the operator norm or the Euclidean norm.

$\forall \boldsymbol{x}, A, \|A\boldsymbol{x}\| \leq |||A||| \cdot \|\boldsymbol{x}\|$. *Further, let* A^+ *be the Moore-Penrose pseudo-inverse of* A*, Wass be the Wasserstein distance, MMD be the linear MMD*[6]*. Let* $W_\Sigma^+ := \Sigma W^T (W\Sigma W^T)^+$*. If* X *is elliptical with covariance matrix* Σ *then*

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot MMD\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right) \\
&\leq MMD\left(P(X \mid T=1), P(X \mid T=0)\right) \\
&\leq |||W_\Sigma^+||| \cdot MMD\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right)
\end{aligned}$$

If X *is Gaussian with positive-definite covariance matrix* Σ *and* W *has full row rank then*

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot Wass\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right) \\
&\leq Wass\left(P(X \mid T=1), P(X \mid T=0)\right) \\
&\leq |||W_\Sigma^+||| \cdot Wass\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right).
\end{aligned}$$

Proof. We prove separately the bounds on the Wass distance and on the MMD.

- First, note that for any random variable V,

$$\begin{aligned}
\text{Wass}(P(V|T=1), P(V|T=0)) &= \sup_{f \text{ 1-Lipschitz}} \left|\mathbb{E}[f(V) \mid T=1] - \mathbb{E}[f(V) \mid T=0]\right| \\
&= \text{IPM}_{\{f:\ f \text{ 1-Lipschitz}\}}(P(V|T=1), P(V|T=0)),
\end{aligned}$$

  where IPM is defined in Equation S6.

- Let f be a 1-Lipschitz function f on the $\mathcal{B}$ space of $b(X)$, and define $g(x) = \frac{1}{|||W|||} f(Wx)$. The function g is also 1-Lipschitz, since for any x, x',

$$\begin{aligned}
|g(x) - g(x')| &= \frac{1}{|||W|||} |f(Wx) - f(Wx')| \\
&\leq \frac{1}{|||W|||} \|Wx - Wx'\| \quad \text{by 1-Lipschitzness of } f \\
&= \frac{1}{|||W|||} \|W(x - x')\| \\
&\leq \frac{1}{|||W|||} \cdot |||W||| \cdot \|x - x'\| \\
&= \|x - x'\|.
\end{aligned}$$

  Thus,

$$\begin{aligned}
\left|\mathbb{E}[f(b(X))|T=1] - \mathbb{E}[f(b(X))|T=0]\right| &= |||W||| \cdot \left|\mathbb{E}[g(X)|T=1] - \mathbb{E}[g(X)|T=0]\right| \\
&\leq |||W||| \cdot \text{Wass}(P(X|T=1), P(X|T=0)).
\end{aligned}$$

  It follows that

$$\text{Wass}(P(b(X)|T=1), P(b(X)|T=0)) \leq |||W||| \cdot \text{Wass}(P(X|T=1), P(X|T=0)).$$

- Now, let f is a 1-Lipschitz real-valued function on $\mathcal{X}$. We show that $g(\boldsymbol{\beta}) = \mathbb{E}[f(X)|b(X) = \boldsymbol{\beta}]$ is Lipschitz. First, as $\text{Cov}(X, WX) = \Sigma W^T$, $\text{Cov}(WX, X) = W\Sigma$, and $\text{Cov}(WX, WX) = W\Sigma W^T$ which is invertible as W has full row rank, we have that for any $\boldsymbol{\beta}$,

$$X|WX = \boldsymbol{\beta} \sim \mathcal{N}\left(\mu + \Sigma W^T (W\Sigma W^T)^{-1}(\boldsymbol{\beta} - W\mu), \Sigma - \Sigma W^T (W\Sigma W^T)^{-1} W\Sigma\right).$$

[6] Note that these theoretical results also hold when $b(X)$ has a bias term.

As a result, for any $\boldsymbol{\beta}$,

$$g(\boldsymbol{\beta}) = \mathbb{E}_{Z \in \mathcal{N}((I-\Sigma W^T(W\Sigma W^T)^{-1}W)\mu, \Sigma-\Sigma W^T(W\Sigma W^T)^{-1}W\Sigma)}[f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} + Z)]$$

We simplify the notation of this expectation into $\mathbb{E}_Z[...]$. Note that, critically, the distribution of Z does not depend on $\boldsymbol{\beta}$.

Then, for any $\boldsymbol{\beta}, \boldsymbol{\beta}'$,

$$\begin{aligned}
|g(\boldsymbol{\beta}) - g(\boldsymbol{\beta}')| &= \left|\mathbb{E}_Z[f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} + Z)] - \mathbb{E}_Z[f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta}' + Z)]\right| \\
&= \left|\mathbb{E}_Z[f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} + Z) - f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta}' + Z)]\right| \\
&\leq \mathbb{E}_Z[\left|f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} + Z) - f(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta}' + Z)\right|] \text{ from Jensen's inequality} \\
&\leq \mathbb{E}_Z[||(\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} + Z) - (\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta}' + Z)||] \quad \text{from the 1-Lipschitzness of } f \\
&= \mathbb{E}_Z[||\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} - \Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta}'||] \\
&= ||\Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta} - \Sigma W^T(W\Sigma W^T)^{-1}\boldsymbol{\beta}'|| \\
&= ||\Sigma W^T(W\Sigma W^T)^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}')|| \\
&\leq |||\Sigma W^T(W\Sigma W^T)^{-1}||| \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}'||
\end{aligned}$$

so g is $|||\Sigma W^T(W\Sigma W^T)^{-1}|||$-Lipschitz. Therefore, as a consequence of Proposition 5,

$$\text{Wass}(P(X|T=1), P(X|T=0)) \leq |||\Sigma W^T(W\Sigma W^T)^{-1}||| \cdot \text{Wass}(P(b(X)|T=1), P(b(X)|T=0)).$$

- For MMD, note that for any random variable V :

$$\begin{aligned}
\text{MMD}(P(V|T=1), P(V|T=0)) &= \sup_{a \in \mathbb{R}^{\dim(V)},\ ||a|| \leq 1} \left|\mathbb{E}[a^T V|T=1] - \mathbb{E}[a^T V|T=0]\right| \\
&= ||\mathbb{E}[V|T=1] - \mathbb{E}[V|T=0]||
\end{aligned}$$

- We note that

$$\begin{aligned}
&\text{MMD}\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right) \\
&= ||\mathbb{E}[b(X)|T=1] - \mathbb{E}[b(X)|T=0]|| \\
&= ||\mathbb{E}[WX|T=1] - \mathbb{E}[WX|T=0]|| \\
&= ||W(\mathbb{E}[X|T=1] - \mathbb{E}[X|T=0])|| \\
&\leq |||W||| \cdot ||\mathbb{E}[X|T=1] - \mathbb{E}[X|T=0]|| \\
&= |||W||| \cdot \text{MMD}\left(P(X \mid T=1), P(X \mid T=0)\right)
\end{aligned}$$

- From Equation S7, $\mathbb{E}[X|T=t] = \mathbb{E}\big[g(b(X)) \mid T=t\big]$, where $g(\boldsymbol{\beta}) \coloneqq \mathbb{E}[X|b(X)=\boldsymbol{\beta}]$. From Section 2 of (Cambanis et al., 1981), if if X is elliptical with location μ and covariance matrix Σ then $\begin{pmatrix} I \\ W \end{pmatrix} X$ is elliptical with location $\begin{pmatrix} I \\ W \end{pmatrix} \mu$ and covariance matrix $\begin{pmatrix} \Sigma & \Sigma W^T \\ W\Sigma & W\Sigma W^T \end{pmatrix}$. Then, from Corollary 5 of (Cambanis et al., 1981),

$$\begin{aligned}
g(\boldsymbol{\beta}) &= \mathbb{E}[X|WX = \boldsymbol{\beta}] \\
&= \mu + \Sigma W^T(W\Sigma W^T)^{+}(\boldsymbol{\beta} - W\mu).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[X|T=t] &= \mathbb{E}\big[g(b(X)) \mid T=t\big] \\
&= \mathbb{E}\big[\mu + \Sigma W^T(W\Sigma W^T)^{+}(b(X) - W\mu) \mid T=t\big]
\end{aligned}$$

$$= \Sigma W^T (W\Sigma W^T)^+ \mathbb{E}[b(X) \mid T = t] + \left(I - \Sigma W^T (W\Sigma W^T)^+ W\right)\mu$$

Then,

$$\begin{aligned}
&\mathrm{MMD}\left(P(X \mid T=1), P(X \mid T=0)\right) \\
&= \|\mathbb{E}[X|T=1] - \mathbb{E}[X|T=0]\| \\
&= \Big\| \Sigma W^T (W\Sigma W^T)^+ \mathbb{E}[b(X) \mid T=1] + \left(I - \Sigma W^T (W\Sigma W^T)^+ W\right)\mu \\
&\quad - \left(\Sigma W^T (W\Sigma W^T)^+ \mathbb{E}[b(X) \mid T=0] + \left(I - \Sigma W^T (W\Sigma W^T)^+ W\right)\mu\right) \Big\| \\
&= \|\Sigma W^T (W\Sigma W^T)^+ (\mathbb{E}[b(X) \mid T=1] - \mathbb{E}[b(X) \mid T=0])\| \\
&\leq |||\Sigma W^T (W\Sigma W^T)^+||| \cdot \|\mathbb{E}[b(X)|T=1] - \mathbb{E}[b(X)|T=0]\| \\
&= |||\Sigma W^T (W\Sigma W^T)^+||| \cdot \mathrm{MMD}\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right).
\end{aligned}$$

□

A.4 Bounds For Non-Balancing Scores

Proposition 7. *Let*

$$\mathcal{E}^D_{t,b}(\boldsymbol{\beta}) := D\Big(P\big(X|b(X)=\boldsymbol{\beta}, T=t\big), P\big(X|b(X)=\boldsymbol{\beta}\big)\Big)$$

where D is a probability discrepancy measure, b is a function of X, $t \in \{0,1\}$ is a realisation of T, $\boldsymbol{\beta}$ is a realisation of $b(X)$. For any function b,

$$\begin{aligned}
&TV\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&\leq TV\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \\
&\leq TV\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&+ \mathbb{E}\big[\mathcal{E}^{TV}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{TV}_{0,b}\big(b(X)\big)|T=0\big]
\end{aligned}$$

and, using the notations of Proposition 5,

$$\begin{aligned}
&IPM_{\mathcal{F}}\left(P(X \mid T=1), P(X \mid T=0)\right) \\
&\leq IPM_{\mathcal{F}_b}\left(P(b(X) \mid T=1), P(b(X) \mid T=0)\right) \\
&+ \mathbb{E}\big[\mathcal{E}^{IPM_{\mathcal{F}}}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{IPM_{\mathcal{F}}}_{0,b}\big(b(X)\big)|T=0\big].
\end{aligned}$$

For a linear function $b(x) = Wx$, if X is elliptical with covariance matrix Σ, then

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot MMD\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&\leq MMD\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \\
&\leq |||W^+_\Sigma||| \cdot MMD\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&+ \mathbb{E}\big[\mathcal{E}^{MMD}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{MMD}_{0,b}\big(b(X)\big)|T=0\big]
\end{aligned}$$

and if X is Gaussian with positive-definite covariance matrix Σ while W has full row rank, then

$$\begin{aligned}
&\frac{1}{|||W|||} \cdot Wass\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&\leq Wass\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \\
&\leq |||W^+_\Sigma||| \cdot Wass\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\
&+ \mathbb{E}\big[\mathcal{E}^{Wass}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{Wass}_{0,b}\big(b(X)\big)|T=0\big].
\end{aligned}$$

Proof. The lower bounds were established in the previous Propositions, while the upper bounds follow as a corollary of the following Proposition. Indeed, the proofs Propositions 1, 5 and 6 directly show that Equation (S8) follows for their respective assumptions on classes of functions, distributions of X and balancing scores. □

Proposition 8. *Let b a function of X, $\mathcal{F}$ a class of functions of X. Assume that for some constant C_b and some class of function $\mathcal{F}'_b$ of functions on the image space on b, both depending on b:*

$$\forall f \in \mathcal{F}, \Big|\mathbb{E}\big[\ \mathbb{E}[f(X)|b(X)]\ \big|\ T=1\big] - \mathbb{E}\big[\ \mathbb{E}[f(X)|b(X)]\ \big|\ T=0\big]\Big| \leq C_b \cdot IPM_{\mathcal{F}'_b}\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big). \tag{S8}$$

Then, letting $\mathcal{E}^D_{t,b}(\boldsymbol{\beta}) = D\Big(P\big(X|b(X)=\boldsymbol{\beta}, T=t\big), P\big(X|b(X)=\boldsymbol{\beta}\big)\Big)$ where D is a probability distance, we have

$$\begin{aligned} IPM_{\mathcal{F}}\Big(P\big(X|T=1\big), P\big(X|T=0\big)\Big) \leq\ & C_b \cdot IPM_{\mathcal{F}'_b}\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\ & + \mathbb{E}\big[\mathcal{E}^{IPM_{\mathcal{F}}}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{IPM_{\mathcal{F}}}_{0,b}\big(b(X)\big)|T=0\big] \end{aligned}$$

Proof. Denote $\Delta_t(\boldsymbol{\beta}; f) \coloneqq \mathbb{E}\Big[f(X)\Big|b(X)=\boldsymbol{\beta}, T=t\Big] - \mathbb{E}\Big[f(X)\Big|b(X)=\boldsymbol{\beta}\Big]$, so that $\mathcal{E}^{\mathrm{IPM}_{\mathcal{F}}}_{t,b}\big(\boldsymbol{\beta}\big) = \sup_{f\in\mathcal{F}}\Big|\Delta_t(\boldsymbol{\beta}; f)\Big|$.

We fix $f \in \mathcal{F}$, noting that

$$\begin{aligned} \mathbb{E}[f(X)|T=t] &= \mathbb{E}\big[\mathbb{E}[f(X)|b(X), T=t]\ \big|\ T=t\big] \text{ due to the law of total expectation} \\ &= \mathbb{E}\big[\Delta_t\big(b(X); f\big) + \mathbb{E}[f(X)|b(X)]\ \big|\ T=t\big] \\ &= \mathbb{E}\big[\Delta_t\big(b(X); f\big)\ \big|\ T=t\big]\ +\ \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=t\big]. \end{aligned}$$

As a consequence,

$$\begin{aligned} &\Big|\mathbb{E}[f(X)|T=1] - \mathbb{E}[f(X)|T=0]\Big| \\ &= \Big|\ \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=1\big] - \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=0\big] + \mathbb{E}\big[\Delta_1\big(b(X); f\big)\ \big|\ T=1\big] \\ &\quad - \mathbb{E}\big[\Delta_0\big(b(X); f\big)\ \big|\ T=0\big]\ \Big| \\ &\leq \Big|\ \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=1\big] - \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=0\big]\ \Big|\ +\ \Big|\ \mathbb{E}\big[\Delta_1\big(b(X); f\big)\ \big|\ T=1\big]\ \Big| \\ &\quad +\ \Big|\ \mathbb{E}\big[\Delta_0\big(b(X); f\big)\ \big|\ T=0\big]\ \Big|, \end{aligned}$$

where

$$\Big|\ \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=1\big] - \mathbb{E}\big[\mathbb{E}[f(X)|b(X)]\ \big|\ T=0\big]\ \Big| \leq C_b \cdot \mathrm{IPM}_{\mathcal{F}'_b}\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \quad \text{by assumption}$$

and, for $t \in \{0,1\}$,

$$\begin{aligned} \Big|\ \mathbb{E}\big[\Delta_t\big(b(X); f\big)\ \big|\ T=t\big]\ \Big| &\leq \mathbb{E}\big[\ |\Delta_t\big(b(X); f\big)|\ \big|\ T=t\big] \\ &\leq \mathbb{E}\big[\sup_{f\in\mathcal{F}}|\Delta_t\big(b(X); f\big)|\ \big|\ T=t\big] \\ &= \mathbb{E}\big[\mathcal{E}^{\mathrm{IPM}_{\mathcal{F}}}_{t,b}\big(b(X)\big)|T=t\big]. \end{aligned}$$

Thereby, for any $f \in \mathcal{F}$,

$$\begin{aligned} \Big|\mathbb{E}[f(X)|T=1] - \mathbb{E}[f(X)|T=0]\Big| \leq\ & C_b \cdot \mathrm{IPM}_{\mathcal{F}'_b}\Big(P\big(b(X)|T=1\big), P\big(b(X)|T=0\big)\Big) \\ & + \mathbb{E}\big[\mathcal{E}^{\mathrm{IPM}_{\mathcal{F}}}_{1,b}\big(b(X)\big)|T=1\big] + \mathbb{E}\big[\mathcal{E}^{\mathrm{IPM}_{\mathcal{F}}}_{0,b}\big(b(X)\big)|T=0\big]. \end{aligned}$$

Taking the supremum over $f \in \mathcal{F}$ yields the desired result.

□

B A FEW NOTES ABOUT COMPUTATIONAL COMPLEXITIES OF BOUNDS

Computational complexity of bounds in Proposition 6 Denoting $N \coloneqq N_t + N_c$, the computational complexity of the linear MMD estimator is in $\mathcal{O}(ND)$, and the computational complexity of the the Wasserstein distance estimator is in $\mathcal{O}(N^2D + N^3 \log N + N^3 \log D)$ when using the auction algorithm (Peyré and Cuturi, 2020; Bertsekas, 1998), assuming that covariates and balancing scores have bounded second-order moments; we refer to the paragraph below on "Computational complexity of Wasserstein distance". As a result, assuming that the balancing score is of dimension $d << D$ and that we have already computed the ground-truth balancing scores, these complexities can be decreased to $\mathcal{O}(Nd)$ and $\mathcal{O}(N^2d + N^3 \log N + N^3 \log d)$, respectively. Thus, if we assume $d << D \sim N$, there is a clear decrease of the complexity for the linear MMD. The decrease is less stark for the Wasserstein distance, as the dominant $N^3 \log N$ term is untouched; however other terms are clearly decreased.

The decrease of complexity should be nuanced if we compute the entire bounds of Proposition 6, and not just the probability distances, as we have to (1) compute the constants $|||W|||$ and $|||W_\Sigma^+|||$, and (2) compute the balancing scores $(WX_i)_i$. We assume that $|||.|||$ is the operator norm. For (1), when $\Sigma = I$, then $W_\Sigma^+ = W^+$, both constants can be handled simultaneously by computing the singular value decomposition of W, which has a complexity $\mathcal{O}(Dd^2)$ (Golub and Van Loan, 2013; Vasudevan and Ramakrishna, 2017). For a general Σ, then $W_\Sigma^+ = \Sigma W^T (W \Sigma W^T)^+$ and computing $|||W_\Sigma^+|||$ has a complexity $\mathcal{O}(D^2d + Dd^2 + d^3)$, as computing ΣW^T (present twice in the formula of W_Σ^+) is in $\mathcal{O}(D^2d)$, further computing $W \Sigma W^T$ is in $\mathcal{O}(d^2D)$, computing the pseudo-inverse through computing the singular value decomposition is in $\mathcal{O}(d^3)$, deducing W_Σ^+ from both the inverse matrix and the already computed ΣW^T is in $\mathcal{O}(Dd^2)$, and another singular value decomposition for the norm is in $\mathcal{O}(Dd^2)$. For (2), we further increase computational complexity by a term $\mathcal{O}(NDd)$ due to the additional matrix multiplication operations. As a result, when $d << D \sim N$, the bounds for the linear MMD imbalance actually have higher computational complexity than the original imbalance itself, while those for the Wasserstein distance imbalance have slightly lower computational complexity than the original imbalance itself.

Computational complexity of Wasserstein distance More precisely, the computational complexity of the Wasserstein distance is $\mathcal{O}(N^2D + \min\{N^3 \log C_{\infty,X}, N^2 C_{\infty,X}^2 \log N\})$, where the first term corresponds to computing the L_2 distance matrix wrt X, and the second term corresponds to the minimum of the computational complexities of the auction algorithm (Peyré and Cuturi, 2020; Bertsekas, 1998) and Sinkhorn's algorithm (Dvurechensky et al., 2018), assuming we choose the algorithm with the lowest complexity. $C_{\infty,X}$ is an upper bound of the maximal value of the distance matrix wrt X and can further depend on N and D. We assume covariates have a bounded second-order moment: noting X_i covariates of treated units, X'_j those of control units, k the dimension index, we assume that $\forall i,k, \;\; \mathbb{E}[|X_i^k|^2] < M$ and $\forall j,k, \;\; \mathbb{E}[|X_j'^k|^2] < M$. Then

$$
\begin{aligned}
\mathbb{E}[C_{\infty,X}] &= \mathbb{E}[\max_{i,j} \|X_i - X'_j\|] \\
&= \mathbb{E}\left[\max_{i,j} \sqrt{\sum_{k=1}^{D} |X_i^k - X_j'^k|^2}\right] \\
&= \mathbb{E}\left[\sqrt{\max_{i,j} \sum_{k=1}^{D} |X_i^k - X_j'^k|^2}\right] \\
&\leq \sqrt{\mathbb{E}\left[\max_{i,j} \sum_{k=1}^{D} |X_i^k - X_j'^k|^2\right]} \text{ from Jensen's inequality as } \sqrt{.} \text{ is concave} \\
&\leq \sqrt{\mathbb{E}\left[\sum_{k=1}^{D} \max_{i,j} |X_i^k - X_j'^k|^2\right]} \\
&\leq \sqrt{\mathbb{E}\left[\sum_{k=1}^{D} \max_{i,j} 2(|X_i^k|^2 + |X_j'^k|^2)\right]} \quad \text{from } (a-b)^2 \leq 2(a^2+b^2) \;\; \forall a,b \\
&= \sqrt{\mathbb{E}\left[\sum_{k=1}^{D} 2(\max_i |X_i^k|^2 + \max_j |X_j'^k|^2)\right]}
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{\sum_{k=1}^{D} 2\mathbb{E}\left[\max_i |X_i^k|^2 + \max_j |X_j'^k|^2\right]} \\
&\leq \sqrt{2\sum_{k=1}^{D} \mathbb{E}\left[\sum_i |X_i^k|^2 + \sum_j |X_j'^k|^2\right]} \\
&= \sqrt{2\sum_{k=1}^{D} \left(\sum_i \mathbb{E}\left[|X_i^k|^2\right] + \sum_j \mathbb{E}\left[|X_j'^k|^2\right]\right)} \\
&\leq \sqrt{2 \cdot D \cdot (N_t + N_c) \cdot M} \\
&= \sqrt{2 \cdot D \cdot N \cdot M}
\end{aligned}
$$

so, from Jensen's inequality applied to the log function,

$$
\mathbb{E}[\log C_{\infty,X}] \leq \log \mathbb{E}[C_{\infty,X}] = \frac{1}{2} \cdot (\log 2 + \log N + \log D + \log M)
$$

and

$$
\begin{aligned}
\mathbb{E}[C_{\infty,X}^2] &= \mathbb{E}\left[\left(\max_{i,j} \|X_i - X_j'\|\right)^2\right] \\
&= \mathbb{E}\left[\left(\max_{i,j} \sqrt{\sum_{k=1}^{D} |X_i^k - X_j'^k|^2}\right)^2\right] \\
&= \mathbb{E}\left[\max_{i,j}\left(\sqrt{\sum_{k=1}^{D} |X_i^k - X_j'^k|^2}\right)^2\right] \\
&= \mathbb{E}\left[\max_{i,j} \sum_{k=1}^{D} |X_i^k - X_j'^k|^2\right] \\
&\leq 2DNM.
\end{aligned}
$$

where we repeated the above expectations from after Jensen's inequality without the square root. Thus, assuming $D \sim N$ or $D \leq N$ and substituting those complexities in expectation into the computational complexities above, the auction algorithm is in $\mathcal{O}(N^3 \log N + N^3 \log D)$ in expectation, and Sinkhorn's algorithm is in $\mathcal{O}(N^3 D \log N)$ in expectation, so the auction algorithm might be preferable.

C IMPLEMENTATION DETAILS

ACIC 2016 Dataset. This dataset is taken from the ACIC competition of 2016 (Dorie et al., 2017). Covariates were obtained from a study about developmental disorders, measuring data from pregnant women and their children. Treatment assignments and outcomes were synthetically generated from transformed versions of covariates using different data generating processes. Importantly, as treatments are synthetically generated, ground-truth propensity scores are made readily available, allowing us to compute calibration errors. We chose the provided data generating process setting number 4, which has polynomial treatment assignment, an exponential outcome model, 35% of treated units, full overlap, and high treatment heterogeneity. To preprocess the data, categorical covariates with F factors were converted to $F-1$ binary covariates, where the f-th binary covariate encodes factor $f+1$. Due to high heterogeneity between subjects, we also centered and scaled continuous covariates to improve performance of all models. Binary covariates were left unprocessed. 4802 subjects were present in the dataset. The subjects have 82 covariates after preprocessing (23 continuous and 59 binary). In our experiments, we considered 100 versions of this dataset, each corresponding to a different random seed for the data generating process.

News Dataset. This dataset contains 5000 documents extracted from the NYT Corpus, where each of the 3477 covariates represents counts of a word in news articles. The treatment indicator T represents the use of a desktop ($T = 0$) or a mobile device ($T = 1$). The real-valued outcome Y measures the opinion of the reader about the news article. Both treatments and outcomes are generated using a data generating process. Here, 50 random seeds from the data generating process are considered. In contrast to ACIC 2016, we did not choose these random seeds ourselves as they were already provided by the original authors[7] (Johansson et al., 2016).

IHDP Dataset. For this dataset, covariates and treatment assignments are used from 747 subjects in real-world data of the Infant Health Development Program. Outcomes, however, are synthetically generated. We further apply the same scaling of outcomes as in Curth and Schaar (2021), as the absence of scaling led to a few outliers causing very high ATT errors in all methods, making comparisons very challenging. Here, 50 seeds from the data generating process are considered, directly used from the implementation of Dragonnet (Shi et al., 2019). 25 covariates are present (9 are continuous, 16 are binary). Experimental results on this dataset are presented in Appendix D.

Evaluation Metrics. To evaluate and compare experimental results, we use the following metrics:

- The *calibration error*, defined as the mean absolute difference between the estimated and true propensity score. This metric can only be computed when the true propensity score is assumed to be known in the dataset. The smaller the calibration error, the more suitable the estimated propensity score and estimated balancing scores obtained from a model are for matching, as we will be closer to the assumption that the propensity score is correctly estimated. Connecting the calibration error to the balancing error term in Proposition 7 is left for future work.
- The *ATT error*, defined as the absolute difference between the ATT estimated by the method and a ground-truth ATT. For every dataset, we compute the ground-truth ATT as the approximation from Equation (2), as we have access to the conditional expectations of Y.
- We empirically quantify *sample imbalance* $\hat{I}$, defined as the squared Euclidean distance between sample means of covariates of treated and control groups from the dataset $\mathcal{D}'$, which is obtained from the original dataset $\mathcal{D}$ via matching, or formally,

$$\hat{I} = \left\| \frac{1}{N_t} \sum_{i \in \mathcal{D}: T_i = 1} X_i - \frac{\sum_{j \in \mathcal{D}: T_j = 0} w_j X_j}{\sum_{j \in \mathcal{D}: T_j = 0} w_j} \right\|_2^2,$$

  where N_t is the number of treated samples, and w_j is the total weight of control sample j after matching. As we can see from this equation, only the sample means of covariates from the control group may change due to matching; the sample means of covariates from the treated group remain unchanged. We note that this measure of imbalance is proportional to the squared linear MMD (Sriperumbudur et al., 2012).

Data Splits. The neural networks were trained using a 60/20/20 training/validation/testing split. The benchmarks logistic regression-based propensity score estimate and PCA were trained using the combined training and validation sets. In-sample metrics were also computed on the combined training and validation datasets, and hold-out metrics were evaluated using the testing set. Alternatively, one might also use controls from the in-sample set when computing hold-out metrics. However, for simplicity of the definition of the hold-out imbalance, we preferred to just use controls from the testing set.

Neural Architecture. The architecture of the neural networks used for matching is as follows : a low-dimensional layer corresponding to the multivariate balancing score (which we also call the "balancing score layer"), then wide hidden layers which are not used as balancing scores, and finally the propensity score head. This architecture is designed to focus on a linear balancing score as in Proposition 6 while keeping flexibility in the rest of the architecture to fit the propensity score model.

Hyperparameters. To choose hyperparameters, we ran a grid search over the following hyperparameter values, minimising validation error on the first dataset version of ACIC 2016 (setting 4, as discussed above).

- Number of hidden layers in addition to the balancing score (hidden) layer: 1, 2.

[7] See "News" link in the "Software and Data" section here: https://www.fredjo.com/

- Number of hidden units per hidden layer (besides the balancing score layer): 100, 200, 300.
- Learning rate: 10^{-2}, 10^{-3}, 10^{-4}.
- Weight decay: 0, 0.001, 0.01.

Other hyperparameters which we did not tune include a batch size of 100, and stochastic gradient descent with fixed learning rate as the optimiser. The chosen values by the hyperparameter search were 2 hidden layers besides the balancing score layer, 100 hidden units per hidden layer other than the balancing score layer, a learning rate of 10^{-2}, weight decay with 0.01, and leaky ReLU as an activation. Additionally, on News datasets, the chosen hyperparameters caused the validation loss to diverge after a period of decrease, causing the training to fail. Thus, for this dataset, we used early stopping as a remedy.

Code. We provide our code to implement neural score matching and reproduce our main results at `https://github.com/oscarclivio/neuralscorematching`.

Resources and Assets. Experiments were run on a laptop with a GeForce GTX 1070 GPU with Max-Q Design for training models with neural networks, and on 12 CPU cores for other tasks. For all datasets, we used our own implementation of them in NumPy and PyTorch (after downloading the data in the case of ACIC 2016 and IHDP, as discussed above), and used our own PyTorch implementation for neural network training.

D IHDP

In addition to the experimental results in the main paper, we also provide results for the IHDP dataset (Hill, 2011) in Table 3. Boxplots are presented in Section E.

On IHDP, our method is not outperforming other methods. Plain covariates `X` consistently rank as the best or second best method for each metric and setting (in-sample or hold-out). This might indicate that IHDP, which is a rather low-dimensional dataset with only 25 covariates, is not suited for dimensionality reduction methods, but further work should investigate these results. We also note that matching in the raw covariate space is probably facilitated by the fact that 16 of covariates are binary.

Table 3: Results on the IHDP dataset.

ATT errors	In-Sample	Hold-Out
`NN Layer 1` (ours)	0.156±0.005	0.311±0.011
`NN PS` (ours)	0.190±0.006	0.330±0.011
`X`	0.144±0.005	0.295±0.011
`Random matching`	0.216±0.007	0.342±0.012
`LogReg PS`	0.164±0.005	0.294±0.009
`PCA`	0.159±0.005	0.307±0.011
`PCA + LogReg PS`	0.146±0.005	0.372±0.011
Imbalances	**In-Sample**	**Hold-Out**
`NN Layer 1` (ours)	0.159±0.005	0.442±0.009
`NN PS` (ours)	0.335±0.006	0.511±0.008
`X`	0.07±0.000	0.223±0.000
`Random matching`	0.592±0.006	0.658±0.012
`LogReg PS`	0.033±0.000	0.318±0.000
`PCA`	0.129±0.000	0.407±0.000
`PCA + LogReg PS`	0.137±0.001	0.909±0.003
`No Matching`	0.492±0.000	0.421±0.000

E BOXPLOTS OF ATT ERRORS AND IMBALANCES

We show boxplots corresponding to Tables 1 to 3 in Figures S2 to S8. We provide boxplots with and without outliers. Outliers are defined as values above $Q3 + 1.5 \cdot IQ$ and below $Q1 - 1.5 \cdot IQ$ where $Q1, Q3, IQ$ are the lower quartile, the upper quartile and the interquartile range of the underlying data, respectively.

F SOCIETAL IMPACT

Possible positive societal impacts of our method include improving decision-making for various real-world applications in politics, economics or medicine. Possible negative societal impacts include the misuse of individualised treatment effect estimation to discriminate against individuals or groups, and of matching to identify protected characteristics of individuals or groups. To mitigate such impacts, we emphasise the importance of continued oversight and evaluation in the deployment of AI tools in society as well as the protection of data confidentiality via rigorous anonymisation, particularly with regards to protected characteristics.

References (Appendices)

Bertsekas, D. P. (1998). Network optimization: Continuous and discrete models.

Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385.

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

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2017). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR.

Golub, G. and Van Loan, C. (2013). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR.

Peyré, G. and Cuturi, M. (2020). Computational optimal transport.

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

Shi, C., Blei, D. M., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *NeurIPS*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599.

Vasudevan, V. and Ramakrishna, M. (2017). A hierarchical singular value decomposition algorithm for low rank matrices. *ArXiv*, abs/1710.02812.

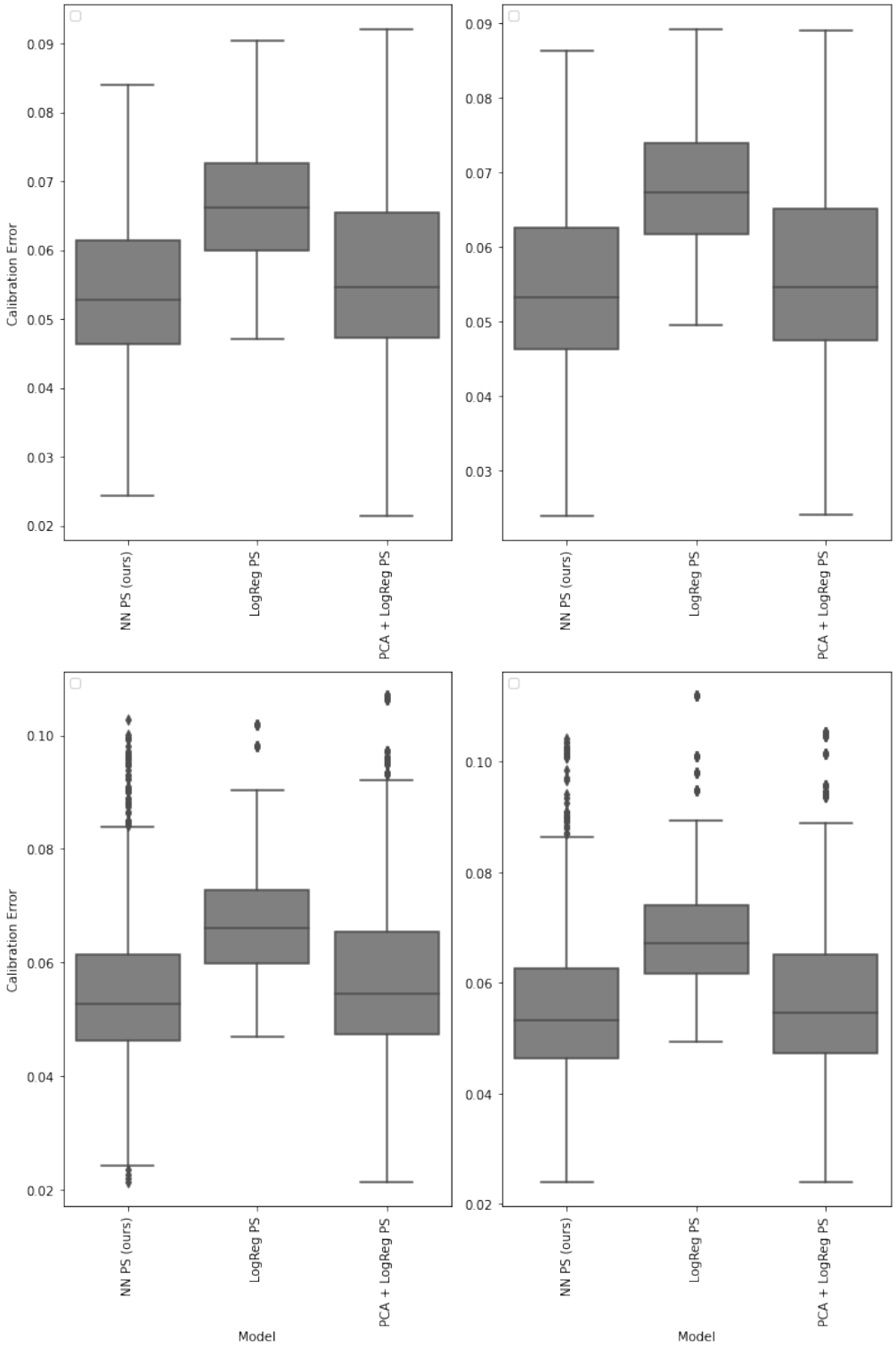


Figure S2: Calibration error boxplots on the ACIC2016 dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to the average calibration error across a dataset version, corresponding to a single draw of the random seed, and a training seed.

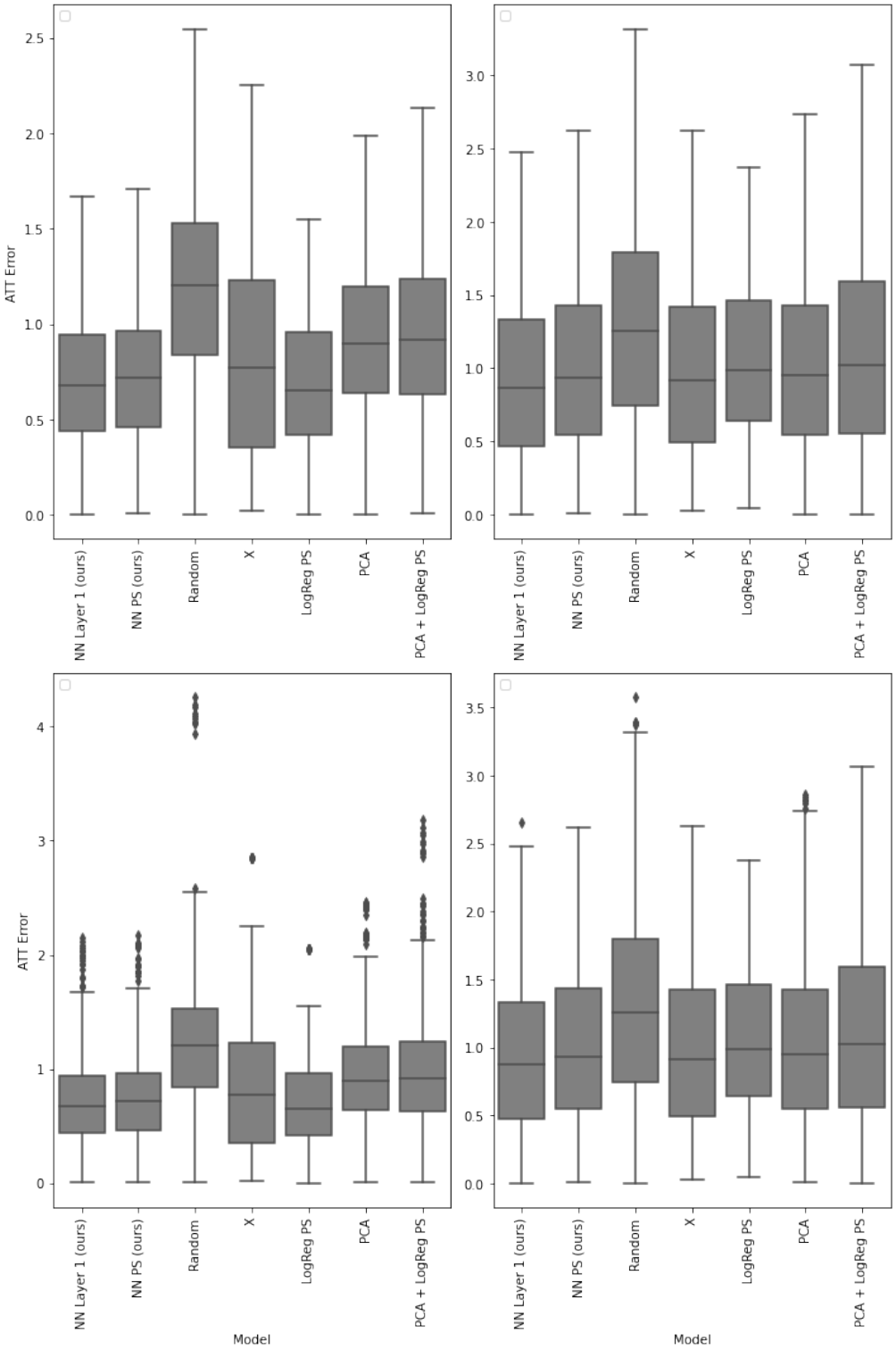


Figure S3: ATT error boxplots on the ACIC2016 dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to the ATT computed on a dataset version, corresponding to a single draw of the random seed, and a training seed.

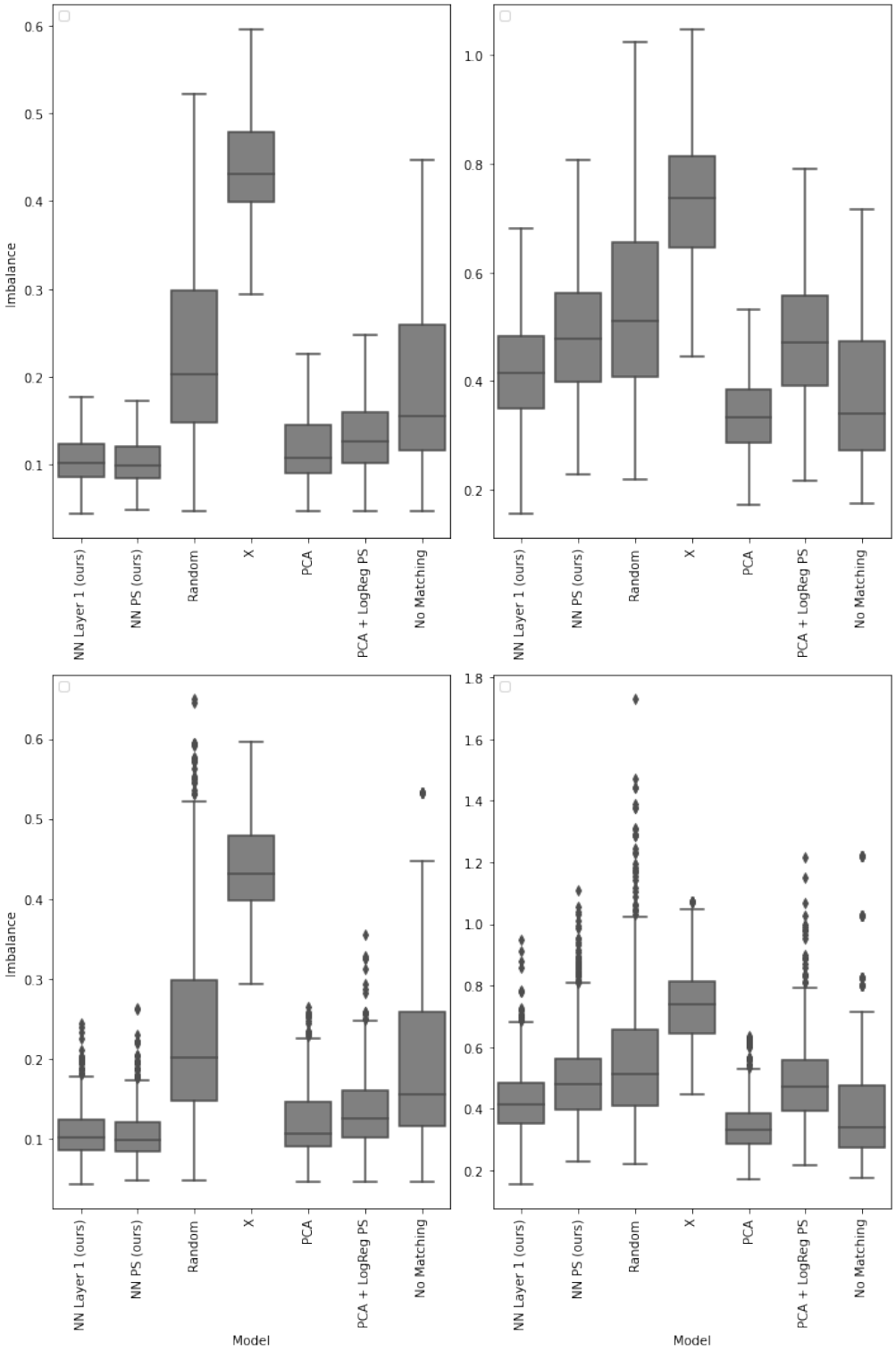


Figure S4: Sample imbalance boxplots on the ACIC2016 dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to sample imbalance computed on a dataset version, corresponding to a single draw of the random seed, and a training seed.

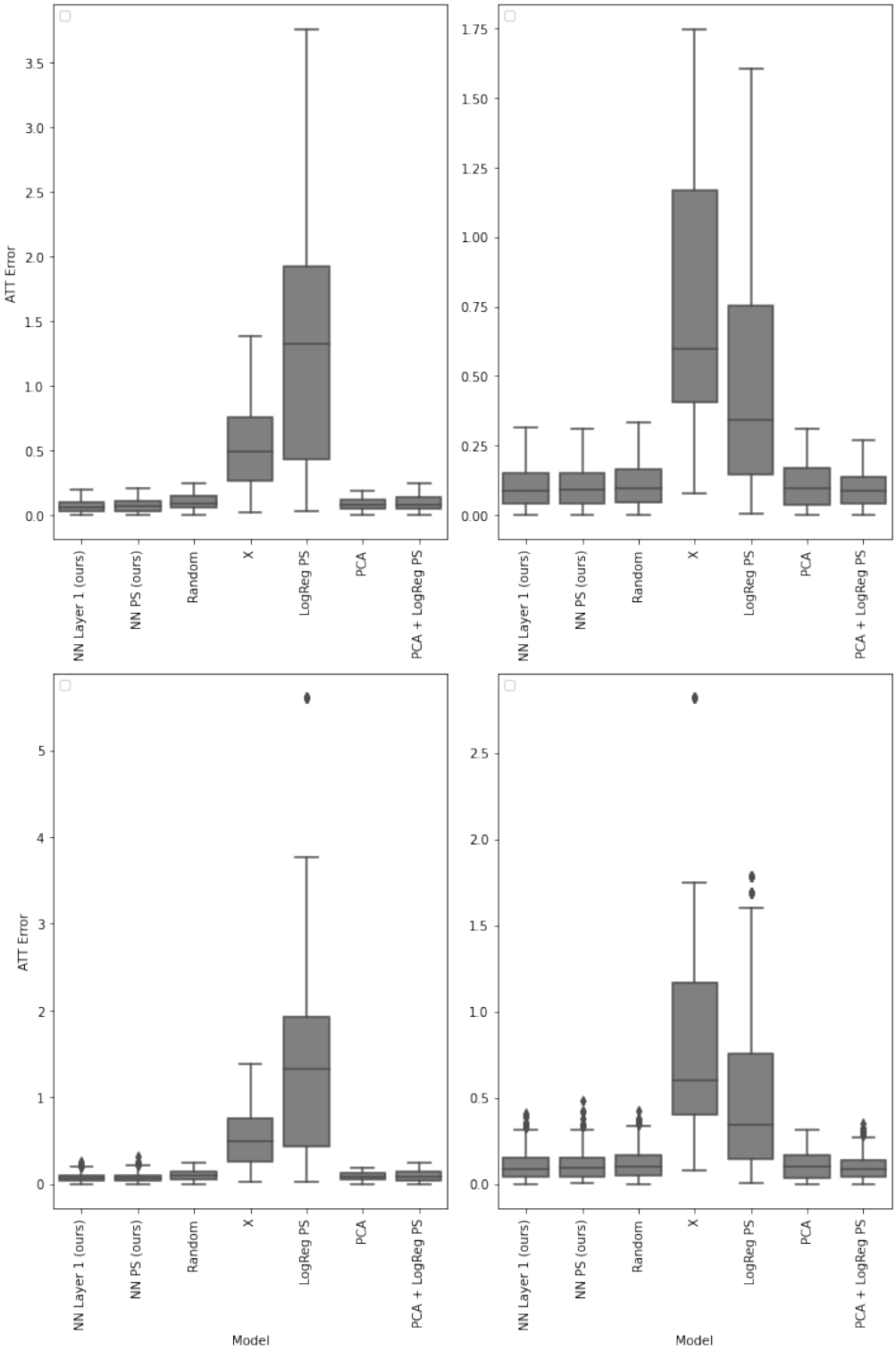


Figure S5: ATT error boxplots on the News dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to the ATT computed on a dataset version, corresponding to a single draw of the random seed, and a training seed.

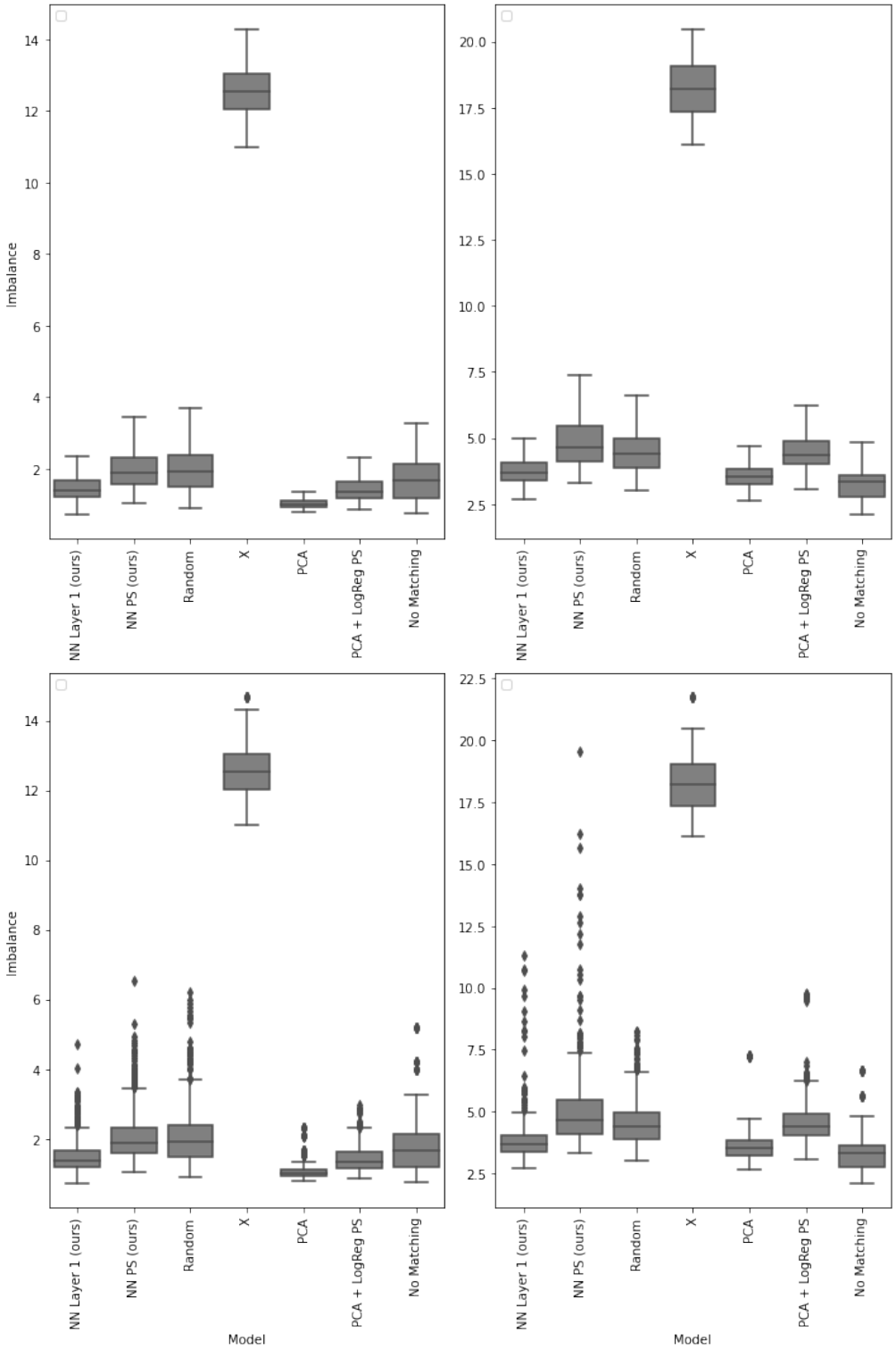


Figure S6: Sample imbalance boxplots on the News dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to sample imbalance computed on a dataset version, corresponding to a single draw of the random seed, and a training seed. Note that we do not show the boxplot for LogReg PS, whose exceptionally high values were hindering the readability of the Figure.

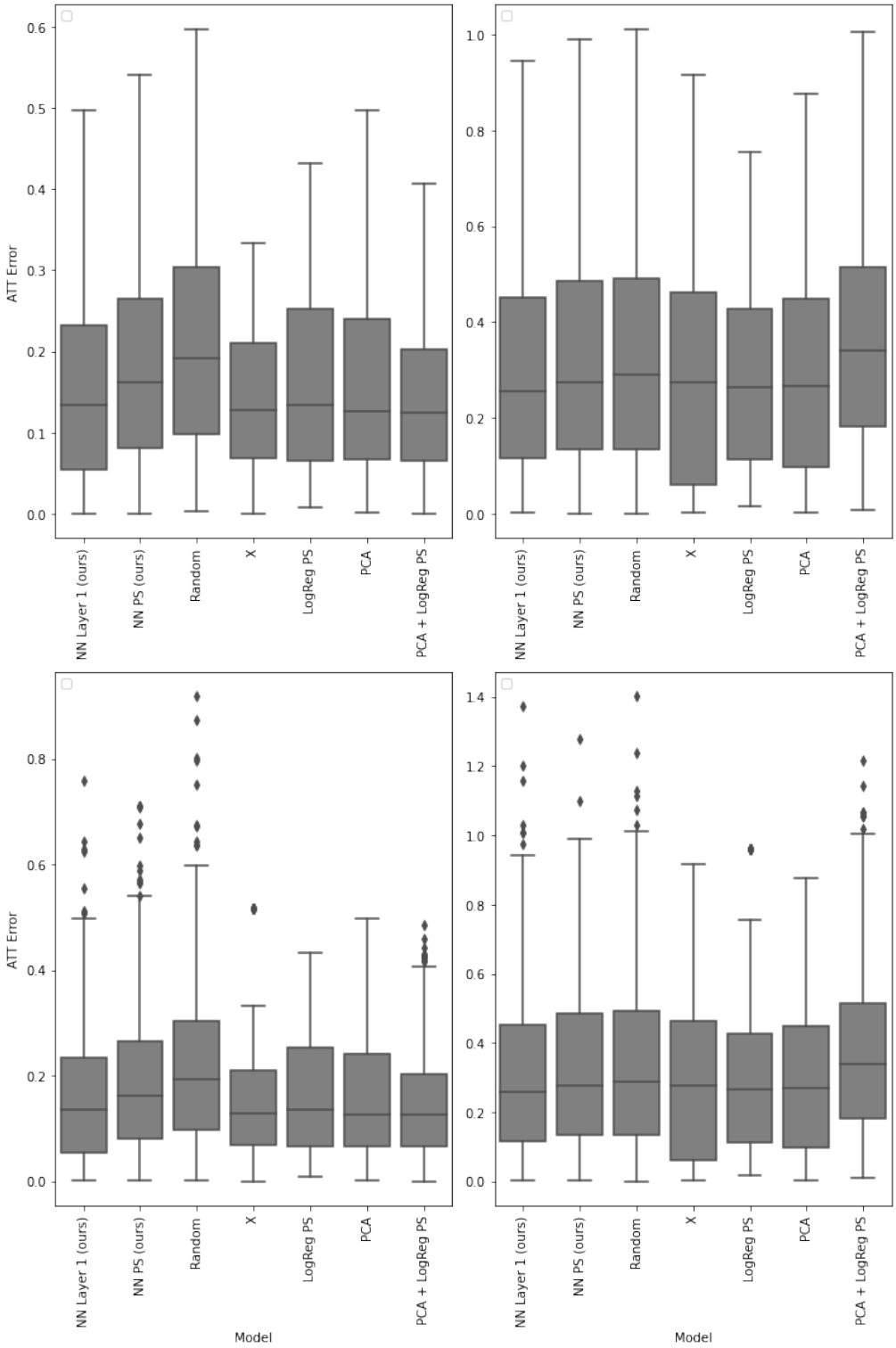


Figure S7: ATT error boxplots on the IHDP dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to the ATT computed on a dataset version, corresponding to a single draw of the random seed, and a training seed.

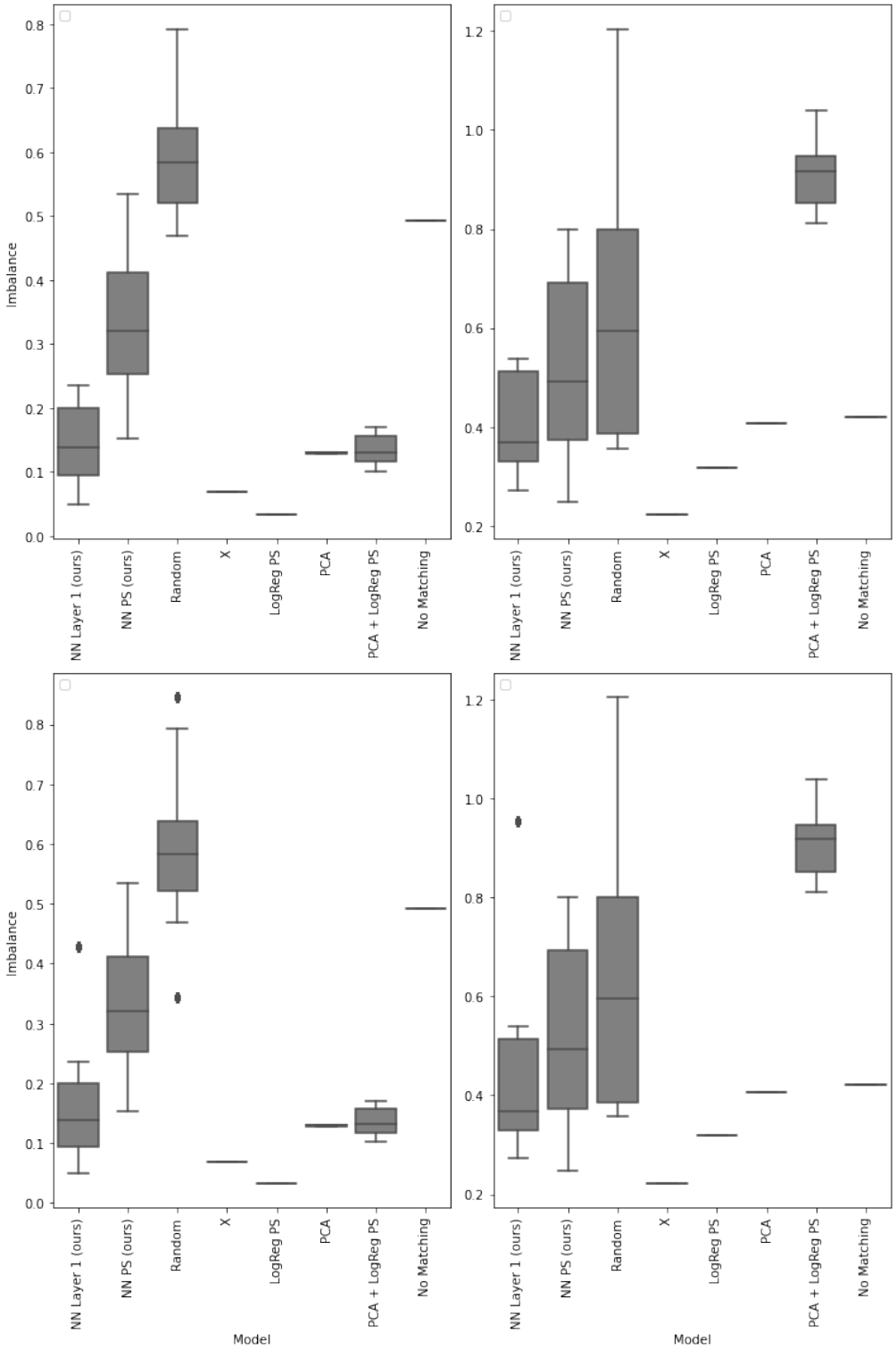


Figure S8: Sample imbalance boxplots on the IHDP dataset: in-sample (left) and hold-out (right), without (up) and with (bottom) outliers. The data points underlying this figure refer to sample imbalance computed on a dataset version, corresponding to a single draw of the random seed, and a training seed.

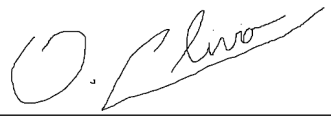
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Neural Score Matching for High-Dimensional Causal Inference
Publication Status	X Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Clivio, O., Falck, F., Lehmann, B., Deligiannidis, G. & Holmes, C.. (2022). Neural score matching for high-dimensional causal inference. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 151:7076-7110

Student Confirmation

Student Name:	Oscar Clivio		
Contribution to the Paper	As the first author, I found the idea, derived all theoretical results, expanded the codebase, conducted all experiments, wrote the first iteration of the paper and took the lead in the writing. Co-authors wrote a first version of the codebase, designed Figure 1, gave a very helpful pointer that was decisive in deriving Proposition 2, and assisted in checking theoretical and numerical results as well as writing.		
Signature		Date	2025/02/20

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Chris Holmes		
Supervisor comments I agree with the student's comments		
Signature 	Date	2025/06/26

This completed form should be included in the thesis, at the end of the relevant chapter.

3

Towards Representation Learning for Weighting Problems in Design-Based Causal Inference

Towards Representation Learning for Weighting Problems in Design-Based Causal Inference

Oscar Clivio[1] **Avi Feller**[2] **Chris Holmes**[1]

[1]Department of Statistics, University of Oxford
[2]Goldman School of Public Policy and Department of Statistics, University of California, Berkeley

Abstract

Reweighting a distribution to minimize a distance to a target distribution is a powerful and flexible strategy for estimating a wide range of causal effects, but can be challenging in practice because optimal weights typically depend on knowledge of the underlying data generating process. In this paper, we focus on design-based weights, which do not incorporate outcome information; prominent examples include prospective cohort studies, survey weighting, and the weighting portion of augmented weighting estimators. In such applications, we explore the central role of representation learning in finding desirable weights in practice. Unlike the common approach of assuming a well-specified representation, we highlight the error due to the choice of a representation and outline a general framework for finding suitable representations that minimize this error. Building on recent work that combines balancing weights and neural networks, we propose an end-to-end estimation procedure that learns a flexible representation, while retaining promising theoretical properties. We show that this approach is competitive in a range of common causal inference tasks.

1 INTRODUCTION

Estimating causal effects is a fundamental task in multiple fields such as epidemiology [Westreich et al., 2017], medicine [Rosenbaum, 2012], public policy [Eli Ben-Michael and Jiang, 2024] or economics [Sekhon and Grieve, 2012]. Some challenges include removing the influence of confounders [Pearl et al., 2016] or generalizing a treatment effect estimated on a randomized control trial (RCT) to a target observational population [Degtiar and Rose, 2023, Colnet et al., 2024]. Weighting approaches, which target a causal effect as an expectation under a reweighting of the original distribution, can address many of these problems [Ben-Michael et al., 2021, Colnet et al., 2024, Johansson et al., 2022].

In this paper, we focus on finding so-called *design-based weights*, which do not incorporate any outcome information, either out of principle or out of necessity; as such, we cannot apply existing approaches involving outcomes off the shelf. Most prominently, design-based weights arise in the classical literature on the design of observational studies, which stresses the importance of separating the "design" and "analysis" phases of a non-randomized study [Rubin, 2008], and therefore stresses the importance of estimating weights without using the outcome. Such weights also arise in *prospective cohort studies* [Song and Chung, 2010] and in *survey design* [Lohr, 2021], in which researchers have not yet collected outcomes, as well as in applications in which it is useful to develop a single set of outcome-agnostic weights, such as in analyses with multiple outcomes of interest [Ben-Michael et al., 2024]. Finally, in doubly robust methods that combine outcome and weighting models, such as in Automatic Debiased Machine Learning (AutoDML) [Chernozhukov et al., 2022b] or augmented balancing weights [Ben-Michael et al., 2021], the weights are typically estimated without using outcomes.

Such methods for finding design-based weights generally rely on minimizing a probability distance between the weighted distribution and a reference distribution. The optimal distance, however, typically depends on the unknown data generating process (DGP). This has led to a large literature on learning an adequate *representation*, a mapping of the covariate space to another manifold, that retains important properties of the DGP. Standard representations include balancing scores [Rosenbaum and Rubin, 1983b], sufficient dimension reduction [Luo and Zhu, 2020], and variable selection [Brookhart et al., 2006]. The correctness of these representations typically relies on unverifiable assumptions and the analyst is left without guarantees on the bias of the weighting estimator if they are not met, leading to poor per-

formance in practice [Kang and Schafer, 2007]. More recent approaches learn a representation implicitly by, for example, modelling weights directly as neural networks, however they only provide guarantees on the bias for specific DGPs, e.g. when the outcome model is piece-wise constant [Ozery-Flato et al., 2018] or follows a neural network architecture [Kallus, 2020a]. Despite these advances, there are not currently principled procedures to directly assess and control the quality of a representation and its impact of the bias on the weighted estimator for any possible data generating process.

This constitutes our two main contributions. (1) We quantify the information lost by using a weighted estimator based on a representation, rather on the original covariates, through a "confounding bias" and a "balancing score error", and give guarantees on the resulting bias of the estimator for any (posited class of the) outcome model. (2) We develop a method inspired by DeepMatch [Kallus, 2020a] and RieszNet [Chernozhukov et al., 2022a] that learns such representations from data. Unlike the original RieszNet application, however, we do not incorporate outcome information. We show promising performance of this approach on benchmark datasets in treatment effect estimation. This learnt representation can serve as the input of any weighting method, making it a generic pre-processing method.

2 BACKGROUND

2.1 SETUP AND NOTATION

Let $P(X, \tilde{Y})$ be a **source** distribution on *covariates* X and some *pseudo-outcomes* $\tilde{Y}$, and Q be a **target** distribution on covariates. For any distribution R and random variable Z, denote R_Z the law $R(Z)$. We assume that we have access to (not necessarily disjoint) i.i.d. samples $\mathcal{P}$ from P and $\mathcal{Q}$ from Q. Let $\mathbb{E}_R[Z]$ be the expectation of a random variable Z under the distribution R. We call a **weight function wrt** P or **weights wrt** P any measurable P_X-a.s. non-negative function $w(x)$ of covariates such that $\mathbb{E}_P[w(X)] = 1$. Any weight function w wrt P induces a distribution P^w such that $\frac{\mathrm{d}P^w_X}{\mathrm{d}P_X}(x) = w(x)$ and $P^w(\tilde{Y}|X) = P(\tilde{Y}|X)$, where we say that P is **reweighted** by $w(X)$, with $\mathbb{E}_{P^w}[f(X)] = \mathbb{E}_P[w(X)f(X)]$ for any function f. Let $\mathbb{E}_P[\tilde{Y}|X = x]$ be a function of interest, which we call **the outcome model**. We are interested in the **target estimand** $\mathbb{E}_Q[\mathbb{E}_P[\tilde{Y}|X]]$. In general, we do not have access to either the outcome model or the target estimand. That said, for any weight function $w(x)$ wrt P, $\hat{\tau}_w := \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} w(X_i)\tilde{Y}_i$ is an unbiased estimator of $\mathbb{E}_{P^w}[\mathbb{E}_P[\tilde{Y}|X]]$ as soon as $\mathbb{E}_P[w(X)\tilde{Y}]$ is well-defined. All of this motivates our problem statement.

Problem 2.1. Find a weight function $w(X)$ wrt P such that

$$\mathbb{E}_{P^w}\left[\mathbb{E}_P[\tilde{Y}|X]\right] = \mathbb{E}_Q\left[\mathbb{E}_P[\tilde{Y}|X]\right]$$

This generalizes many weighting problems in causal inference. Generally, let A denote the treatment variable, and Y denote the outcome. We assume that the values of A belong to a finite space $\mathcal{A}$. For $a \in \mathcal{A}$, we denote $Y(a)$ the potential outcome wrt a, which is the realized outcome if the subject were to receive treatment a. In the context of transportability, we also introduce a binary indicator S for membership in a RCT population, thus $A \perp\!\!\!\perp X|S = 1$ and $(Y(1), Y(0)) \perp\!\!\!\perp A|S = 1$. Let $P^{\text{data}}(X, Y, S, A, (Y(a))_{a\in\mathcal{A}})$ be the true data distribution. In the absence of subscript, we assume that the expectation operator is that wrt P^{data}, that is $\mathbb{E} := \mathbb{E}_{P^{\text{data}}}$. Then, Problem 2.1 can be applied to the following weighting problems (details in Appendix A):

- *Average Treatment Effect on the Treated (ATT).* The pseudo-outcome is Y; the source and target distributions are $P^{\text{data}}(X, Y|A = 0)$ and $P^{\text{data}}(X|A = 1)$, respectively; the outcome model is $\mathbb{E}[Y(0)|X = x]$; the estimand is $\mathbb{E}[Y(0)|A = 1]$.
- *Average Treatment Effect (ATE).* Let $a \in \mathcal{A}$ be fixed. The pseudo-outcome is Y; the source and target distributions are $P^{\text{data}}(X, Y|A = a)$ and $P^{\text{data}}(X)$, respectively; the outcome model is $\mathbb{E}[Y|A = a, X = x]$; the estimand is $\mathbb{E}[Y(a)]$.
- *Transportability.* The pseudo-outcome is

  $$\tilde{Y} := \frac{AY}{P^{\text{data}}(A = 1|S = 1)} - \frac{(1 - A)Y}{P^{\text{data}}(A = 0|S = 1)};$$

  the source distribution is the joint covariate and pseudo-outcome distribution in the RCT $P^{\text{data}}(X, \tilde{Y}|S = 1)$; the target distribution is the covariate distribution in the target population $P^{\text{data}}(X|S = 0)$; the outcome model is the conditional average treatment effect (CATE) $\mathbb{E}[Y(1) - Y(0)|X = x]$; the estimand is the ATE on the target population, $\mathbb{E}[Y(1) - Y(0)|S = 0]$.

One solution to these problems has the following form:

Definition 2.2. We call **true weights between** P **and** Q the Radon-Nikodym derivative $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$, which is a weight function wrt P.

These weights are also known as *inverse probability weights* or the *Riesz representer* [Hirshberg and Wager, 2021, Chernozhukov et al., 2022b]. They are uniquely defined [Ben-Michael et al., 2021] by, for any measurable function f,

$$\mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)f(X)\right] = \mathbb{E}_Q[f(X)].$$

In particular, this holds for $f(x) = \mathbb{E}_P[\tilde{Y}|x]$, which solves Problem 2.1. In practice, the true weights $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ are unknown; we turn to estimating them and more generally obtaining solution weights in the next section.

Finally, to ensure that true weights are well-defined, we make the following assumption, which is equivalent to *overlap* in ATE estimation [Bruns-Smith et al., 2023] and *support inclusion* [Colnet et al., 2024] in transportability.

Assumption 2.3. Q_X is absolutely continuous wrt P_X.

As we discuss in the introduction, we are in the setting where outcomes Y_i and pseudo-outcomes $\tilde{Y}_i$ for $i \in \mathcal{P}$ are not observed and cannot be used when trying to find weights solving Problem 2.1, and are only available for estimating the final estimate $\hat{\tau}_w$ *after* weights have been found.

2.2 COMMON METHODS IN WEIGHTING

In ATT/ATE estimation and transportability, true weights are proportional to the inverse of one of the propensity scores $p(A = a|X = x)$ [Ben-Michael et al., 2021] or $P(S = 1|X = x)$ [Cole and Stuart, 2010]. Thus, an inverse probability weighting estimator $\widehat{w}$ of $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ is obtained by fitting a model for the indicated propensity score and inverting it, leading to potentially outsize errors due to misspecification [Zubizarreta, 2015]. An alternative used in the automatic debiased machine learning (AutoDML) literature is to minimize the mean squared error between $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ and $\widehat{w}$, which can actually be estimated without exactly knowing the true weights $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ [Chernozhukov et al., 2022b,a, Newey and Newey, 2023]. Another family of methods [Hainmueller, 2012, Fong et al., 2018] 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 minimizes some dispersion measure of weights under these constraints. However, balancing $r(X)$ does not guarantee balancing the unknown $\mathbb{E}_P[Y|X]$ and the solution might not be feasible if r has too many moments [Wainstein, 2022]. Similar methods enforce such balance approximately through a generalized method of moments [Imai and Ratkovic, 2014, Fong et al., 2018].

Finally, another family of methods [Ben-Michael et al., 2021] aims at finding weights w minimizing $|\mathrm{Bias}_{P,Q}(w)|$ where we refer to

$$\mathrm{Bias}_{P,Q}(w) = \mathbb{E}_{P^w}[\mathbb{E}_P[Y|X]] - \mathbb{E}_Q[\mathbb{E}_P[Y|X]]$$

as the **"bias"** of weights w, measuring how short they fall of solving Problem 2.1 and which is also equal to the bias of the estimator $\hat{\tau}_w$ wrt the target estimand. It is usually assumed that $\mathbb{E}_P[\tilde{Y}|x]$ belongs to a class of functions $\mathcal{M}$ which leads to the bound

$$\begin{aligned}|\mathrm{Bias}_{P,Q}(w)| &\leq \mathrm{IPM}_{\mathcal{M}}(P_X^w, Q_X) \\ &:= \sup_{\bar{m}\in\mathcal{M}} |\mathbb{E}_{P^w}[\bar{m}(X)] - \mathbb{E}_Q[\bar{m}(X)]|\end{aligned}$$

where the RHS is an integral probability metric (IPM) [Sriperumbudur et al., 2012] on the class $\mathcal{M}$ and generally corresponds to a known 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 a term to control the variance of the weighting estimator [Kallus, 2020b, Ben-Michael et al., 2021], we obtain a solution w by solving

$$\min_w \ \mathrm{IPM}_{\mathcal{M}}(P_X^w, Q_X)^2 + \sigma^2 \cdot ||w(X)||^2_{L_2(P)} \qquad (1)$$

for a chosen $\sigma > 0$ that controls a bias-variance trade-off [Bruns-Smith and Feller, 2022]. A key challenge is that as we do not know the outcome model $\mathbb{E}_P[\tilde{Y}|x]$, we do not know the model class $\mathcal{M}$, thus an adequate probability discrepancy to minimize. 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 $\mathbb{E}_P[\tilde{Y}|x]$ at stake. Recognizing this, directions in the literature include finding a data-driven tailored function class $\mathcal{M}$ [Kallus, 2020a, Wainstein, 2022] or finding guarantees when the function class is misspecified [Bruns-Smith and Feller, 2022].

2.3 CHOOSING A DISTANCE VIA A REPRESENTATION

Many methods minimize a probability discrepancy measure or more generally find weights that only depend on covariates x via a vector-valued function $\phi(x)$ known as a **representation** [Kallus, 2020a, Xue et al., 2023]. Indeed, assuming any function class $\mathcal{M}$ implicitly assumes that any function linearly depends on a representation $\phi(x)$, e.g. the first-order moment x for linear functions, the kernel feature spaces $k(., x)$ for the RKHS of kernel k [Hazlett, 2020, Kallus, 2020a], and more generally $(m(x))_{m\in\mathcal{M}}$ for any class $\mathcal{M}$ (note that such a representation is not unique). In turn, every representation defines a function class. Thus, choosing a function class $\mathcal{M}$ means *implicitly* choosing a representation $\phi(x)$ and assuming that the true outcome model $\mathbb{E}_P[\tilde{Y}|x]$ linearly depends on it.

Further, it is also common practice to *explicitly* define a representation $\phi(x)$ (on which the outcome model need *not* depend linearly) and apply a weighting method using it. Notable examples include propensity scores and balancing scores [Rosenbaum and Rubin, 1983b], prognostic scores [Hansen, 2008] or variable selection [Brookhart et al., 2006, Colnet et al., 2024]. One motivation to do so is that a low-dimensional representation can mitigate undesirable effects of high dimensions in causal inference [Ning et al., 2020, D'Amour et al., 2021] or probability distances [Dudley, 1969, Ramdas et al., 2015] and improve efficiency by selecting essential covariate information wrt the DGP.

The question then becomes how to obtain suitable repre-

sentations $\phi(x)$. It is well-known that weighting on the true outcome model, the propensity score or a representation predicting either [Rosenbaum and Rubin, 1983b, Hansen, 2008] 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 to find a linear representation under the constraint that it predicts either model [Cook, 2009, Luo and Zhu, 2020], while others extract representations from a learnt model for the outcome, the treatment or the RCT indicator [Rosenbaum and Rubin, 1983a, Hansen, 2008, Cole and Stuart, 2010]. However, to the best of our knowledge, there are no guarantees on the bias when any posited model is misspecified or more generally when any underlying assumption is violated, while they are critical as one cannot verify such assumptions. In particular, classification-based learning of propensity scores does not optimize for covariate balance but for prediction of the treatment or the RCT indicator, while (near-)deterministic prediction of either will violate (strict [D'Amour et al., 2021]) overlap, leading to poor matching or weighting performance in practice [Alam et al., 2019, King and Nielsen, 2019]. In addition, many such methods learn the representation using outcomes, which is done before weighting, thus is not permitted in an actual design-based setting. More recent works learn implicit representations by positing a rich parametric class $\mathcal{M}$ [Ozery-Flato et al., 2018, Kallus, 2020a], as a result bias can be controlled but only for outcome models belonging to this class.

Thus, one might wonder whether guarantees on the bias can be provided when using *any* representation ϕ and *any* class $\mathcal{M}$, without using outcome information and without relying on rigid well-specification assumptions. This is the main contribution of our paper, which we develop next.

3 THEORY AND METHOD

3.1 QUANTIFYING THE INFORMATION LOSS

Choosing a representation $\phi(X)$ introduces many trade-offs. At one extreme, oracle representations, such as balancing scores or prognostic scores, perfectly preserve unconfoundedness; that is, unconfoundedness given $\phi(X)$ implies unconfoundedness given X. These are largely unknown, however. At the other extreme, degenerate representations, such as a *constant* $\phi(X)$, will destroy all the information in the original X. We now characterize representations that minimize the information lost relative to X.

Indeed, we first make technical assumptions ensuring that all expectations are well-defined. For any distribution R, random variable Z and integer $p \geq 1$, let

$$||Z||_{L_p(R)} := (E_R[|Z|^p])^{\frac{1}{p}},$$

and note $Z \in L_P(R)$ iff $||Z||_{L_P(R)} < \infty$. Notably, for a measurable function f of values of Z,

$$||f||_{L_p(R_Z)} = (E_R[|f(Z)|^p])^{\frac{1}{p}} = ||f(Z)||_{L_p(R)}$$

We then make the following assumptions.

Assumption 3.1. $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) \in L_2(P)$

Assumption 3.2. $\tilde{Y} \in L_2(P)$

Then, under Assumptions 2.3, 3.1, 3.2 by noting that for any weights w wrt P that are in $L_2(P_X)$, and for any measurable mapping $\phi(x)$ of covariates, the bias can be decomposed as

$$\begin{aligned}
&\text{Bias}_{P,Q}(w) = \mathbb{E}_{P^w}\left[\mathbb{E}_P[\tilde{Y}|X]\right] - \mathbb{E}_Q\left[\mathbb{E}_P[\tilde{Y}|X]\right] \\
&= \underbrace{\mathbb{E}_{P^w}\left[\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] - \mathbb{E}_Q\left[\mathbb{E}_P[\tilde{Y}|\phi(X)]\right]}_{\text{Bias wrt the representation}} \\
&+ \underbrace{\mathbb{E}_{P^w}\left[\mathbb{E}_P[\tilde{Y}|X] - \mathbb{E}_P[\tilde{Y}|\phi(X)]\right]}_{\text{Chosen weights bias}} \\
&+ \underbrace{\mathbb{E}_Q\left[\mathbb{E}_P[\tilde{Y}|\phi(X)] - \mathbb{E}_P[\tilde{Y}|X]\right]}_{\text{Confounding bias}}. \qquad (2)
\end{aligned}$$

We now explain each term in the RHS. First, if the weights $w(X)$ are a function of the representation $\phi(X)$, the *bias wrt the representation* would be the bias if we replaced X with $\phi(X)$ in the equality of Problem 2.1. This interpretation still holds for general weights $w(X)$ as from the tower property applied to $\mathbb{E}_{P^w}[\mathbb{E}[\tilde{Y}|\phi(X)]]$, they can be replaced with $\mathbb{E}_P[w(X)|\phi(X)]$, which is a $L_2(P_X)$ weight function wrt P and is a function of $\phi(X)$, in the term. As in Section 2.2, we can directly bound the bias wrt the representation via an IPM of the form

$$\text{IPM}_{\mathcal{G}}(P^w_{\phi(X)}, Q_{\phi(X)}),$$

where for example, for a class $\mathcal{M}$ such that $\mathbb{E}_P[\tilde{Y}|x] \in \mathcal{M}$, the class $\mathcal{G}$ can contain

$$\phi(\mathcal{M}, P) := \{z \mapsto \mathbb{E}_P[m(X)|\phi(X) = z],\ \ m \in \mathcal{M}\}. \qquad (3)$$

Second, the *chosen weights bias* measures how much "chosen" weights $w(x)$ do not depend on $\phi(x)$. It turns out that this quantity is zero for weights $\hat{w}(x)$ that solve the canonical minimization in Equation 1 with the aforementioned IPM; as we show next, these weights only depend on $\phi(x)$.

Proposition 3.3. *Let $\phi(x)$ be a measurable mapping with values in a space Φ.*

1. *Under Assumptions 2.3, 3.1, if $\mathcal{G}$ is a class of $L_2(P_{\phi(X)})$ functions on Φ, $\sigma > 0$, there is a unique solution $\hat{w}(x)$ to the problem*

$$\min_{\substack{w \text{ weight} \\ \text{function} \\ \text{wrt } P}} \textit{IPM}_{\mathcal{G}}(P^w_{\phi(X)}, Q_{\phi(X)})^2 + \sigma^2 \cdot ||w(X)||^2_{L_2(P)}$$

and it is a function of $\phi(x)$ P_X-almost surely, i.e. there exists $\bar{w} : \Phi \to \mathbb{R}$ such that $\hat{w}(x) = \bar{w}(\phi(x))$ $\forall x$ P_X-a.s. ; and $\hat{w}(X) \in L_2(P)$.

2. *Under Assumption 3.2, for any $L_2(P_X)$ weight function $w(x)$ wrt P that is a function of $\phi(x)$ P_X-a.s., the chosen weights bias is zero.*

Finally, the **confounding bias** is the most important term of this decomposition, as it characterizes the information lost in $\phi(X)$ relative to X — and thus can be seen as the bias *of* ϕ, rather than the bias *wrt* ϕ that is applied to weights.

When the target is $\mathbb{E}[Y(a)]$, this quantity is the difference between $\mathbb{E}\left[\mathbb{E}[Y \mid A = a, \phi(X)]\right]$ and $\mathbb{E}\left[\mathbb{E}[Y \mid A = a, X]\right]$, measuring how much $\phi(X)$ preserves unconfoundedness [D'Amour and Franks, 2021, Melnychuk et al., 2024]. More generally, for solution weights $\hat{w}$ of Equation 1 with an IPM depending on $\phi(X)$, it is exactly the difference between the biases of $\hat{w}$ wrt original covariates X and their representation $\phi(X)$, as shown by Equation 2. Thus, assuming zero confounding bias, if $\hat{w}$ has a small (resp. zero) bias wrt ϕ then it will also have a small (resp. zero) bias overall.

To the best of our knowledge, this is the first extension of the confounding bias for the $\mathbb{E}[Y(a)]$ target estimand to more general weighting problems in causal inference. It has a similar formulae as the *excess target information loss* in Johansson et al. [2019] measuring the loss of information induced by a representation in domain adaptation. We further provide a characterisation for it that will prove useful.

Proposition 3.4. *Under Assumption 2.3, for any measurable $\phi(x)$, $Q_{\phi(X)}$ is absolutely continuous wrt $P_{\phi(X)}$, with*

$$\frac{dQ_{\phi(X)}}{dP_{\phi(X)}}(\phi(X)) = \mathbb{E}_P\left[\frac{dQ_X}{dP_X}(X)\middle|\phi(X)\right] \quad P\text{-a.s.}$$

and under the additional Assumptions 3.1 and 3.2, the confounding bias is equal to both

$$\mathbb{E}_P\Bigg[\left(\mathbb{E}_P[\tilde{Y}|\phi(X)] - \mathbb{E}_P[\tilde{Y}|X]\right) \times \left(\frac{dQ_X}{dP_X}(X) - \frac{dQ_{\phi(X)}}{dP_{\phi(X)}}(\phi(X))\right)\Bigg] \tag{4}$$

and

$$-\mathbb{E}_P\left[\mathbb{E}_P[\tilde{Y}|X]\left(\frac{dQ_X}{dP_X}(X) - \frac{dQ_{\phi(X)}}{dP_{\phi(X)}}(\phi(X))\right)\right] \tag{5}$$

When the confounding bias is zero, ϕ is known as a *deconfounding score* [D'Amour and Franks, 2021], and the overall bias is simply equal to the bias wrt ϕ. In particular, from Equation 4, the confounding bias will be zero in two special cases :

- When $\mathbb{E}_P[\tilde{Y}|X] = \mathbb{E}_P[\tilde{Y}|\phi(X)]$ $P-$a.s., that is

$$\mathbb{E}_P[\tilde{Y}|X] = \mathbb{E}_P\left[\mathbb{E}_P[\tilde{Y}|X]\middle|\phi(X)\right] \ P\text{-a.s.}$$

  from the tower property. This is equivalent to $\mathbb{E}_P[\tilde{Y}|x]$ being a function of $\phi(x)$ P_X-a.s., i.e. $\phi(X)$ being a prognostic score [Hansen, 2008].

- When $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) = \frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))$ P-a.s., that is

$$\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) = \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X)\right] \ P\text{-a.s.}$$

  from Proposition 3.4. This is equivalent to $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x)$ being a function of $\phi(x)$ P_X-a.s., i.e. $\phi(X)$ being a balancing score [Rosenbaum and Rubin, 1983b].

We make a more rigorous connection between the confounding bias and canonical scores from the literature as well as notions from transportability in Appendix B

Further, the confounding bias and its role in the decomposition of Equation 2 allow us to extend the idea of a deconfounding score to hold approximately, rather than exactly. Indeed, if the confounding bias of ϕ is not zero but remains small, then we can expect that a small bias wrt ϕ obtained by solving the problem of Proposition 3.3 will still yield a small overall bias. This gives us more flexibility than relying on well-specified models, where any guarantee on the bias is lost in case of misspecification. In contrast, the confounding bias directly quantifies the misspecification itself.

Thus, one might wonder whether we can minimize directly said misspecification to find an *approximate* deconfounding score ϕ. However Equation 4 involves ground-truth models we do not have access to like $\mathbb{E}_P[\tilde{Y}|x]$, $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x)$ as well as their projections on $\phi(x)$. Further, we do not observe any outcomes at this stage, precluding any estimation of $\mathbb{E}_P[\tilde{Y}|x]$. To address all of this, note that a direct application of the Cauchy-Schwarz inequality to Equation 5 yields

$$|\text{Confounding bias}| \leq ||\mathbb{E}_P[\tilde{Y}|X]||_{L^2(P)} \cdot \text{BSE}_{P,Q}(\phi) \tag{6}$$

where we further have $||\mathbb{E}_P[\tilde{Y}|X]||_{L^2(P)} \leq ||\tilde{Y}||_{L_2(P)}$ from Jensen's inequality, and we call

$$\text{BSE}_{P,Q}(\phi) := \left|\left|\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - \frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))\right|\right|_{L_2(P)} \tag{7}$$

the **balancing score error** (BSE). This name is justified as from Proposition 3.4, this quantity is equal to

$$\left|\left|\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X)\right]\right|\right|_{L_2(P)},$$

that is the root mean-squared error between $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)$ and its projection on $\phi(X)$, i.e. its best predictor from $\phi(X)$ in $L_2(P)$. In other words, it measures the extent to which $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x)$ is not a function of $\phi(x)$ P_X-a.s., and therefore the

extent to which $\phi(x)$ is not a balancing score. Importantly, it does not depend on the pseudo-outcome $\tilde{Y}$, only on the marginals P_X and Q_X. Note that the confounding bias can be zero and the balancing score error positive, even potentially arbitrary, for many representations $\phi(x)$ that contain information on the outcome model $\mathbb{E}_P[\tilde{Y}|x]$. Concrete examples include prognostic scores from Hansen [2008], or the deconfounding scores in the example of Section 5 in D'Amour and Franks [2021]. Our setup excludes such representations as it assumes that we do not observe outcomes at this stage. Alternatively, if one had access to outcomes, then similarly as for the balancing score error, we can bound the confounding bias with a "prognostic score error".

On the other hand, note that the balancing score error allows us to control the resulting bias with only mild assumptions on the outcome model. We formalize this next.

Corollary 3.5. *Under Assumptions 2.3, 3.1, 3.2, for any set $\mathcal{M}$ of $L_2(P_X)$ functions such that $E_P[\tilde{Y}|x] \in \mathcal{M}$, for any measurable representation ϕ, and for any $L_2(P_X)$ weights w wrt P depending on $\phi(x)$ P_X-a.s., defining $\phi(\mathcal{M}, P)$ as in Equation 3,*

$$|\mathit{Bias}_{P,Q}(w)| \leq \mathit{IPM}_{\phi(\mathcal{M},P)}(P^w_{\phi(X)}, Q_{\phi(X)}) + ||\tilde{Y}||_{L_2(P)} \cdot \mathit{BSE}_{P,Q}(\phi).$$

We note that the bound of Corollary 3.5 is "sharp" in the sense that when we replace the IPM and the BSE by the (unknown) terms they bound, namely the bias wrt the representation and the confounding bias, the inequality becomes an equality. It further suggests a two-step approach to minimize the overall bias on the LHS. First, learn a representation ϕ that minimizes the BSE, i.e. the second term of the RHS, plug this learnt representation ϕ into an IPM and find weights minimizing it, or in other words minimizing the first term of the RHS. To learn the representation, the BSE could be used in addition or in replacement of the traditional likelihood to learn propensity score models. Importantly, we can bypass propensity score estimation altogether and posit more general representations, including multivariate functions. We turn to this in the next sections.

3.2 OPERATIONALIZING AND MINIMIZING INFORMATION LOSS

While we have avoided the need to specify an outcome model $\mathbb{E}_P[\tilde{Y}|x]$, a key bottleneck remains for the balancing score error: we do not have access to the true weights $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)$ or their projection $\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))$. One possible workaround is to first remove the projection by using the definition of a conditional expectation: for any function g on the image space of ϕ,

$$\mathrm{BSE}_{P,Q}(\phi) \leq \left|\left|\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - g(\phi(X))\right|\right|_{L^2(P)}. \quad (8)$$

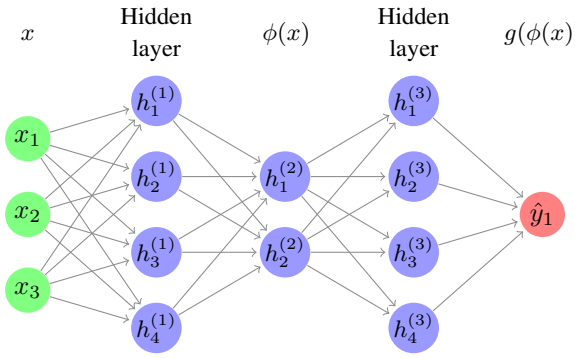


Figure 1: Neural Network to Learn a Representation ϕ.

In particular, for $\epsilon > 0$, if there exists *any* function g on the image space of ϕ such that the RHS of Equation 8 is below $\epsilon/||\tilde{Y}||_{L_2(P)}$, then ϕ has an absolute confounding bias at most ϵ. This gives us more flexibility than working with the true projection of $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$, and motivates finding an g and ϕ minimizing the RHS.

This approach, however, is insufficient since we still do not have access to $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$. A key result from the covariate shift literature [Kanamori et al., 2009], notably exploited in the AutoDML literature [Chernozhukov et al., 2022a,b], helps us remove $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ from the minimization entirely : for any distributions P, Q verifying Assumption 2.3, and for any function v, $\left|\left|\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - v(X)\right|\right|^2_{L^2(P)}$ is equal to $\mathcal{L}_{P,Q}(v)$ up to an additive constant wrt v, where we refer to

$$\mathcal{L}_{P,Q}(v) := \mathbb{E}_P[v(X)^2] - 2 \cdot \mathbb{E}_Q[v(X)]$$

as the *AutoDML loss*. In particular, $\mathcal{L}_{P,Q}(v)$ can be estimated in finite samples for any known v, as

$$\mathcal{L}_{\mathcal{P},\mathcal{Q}}(v) = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} v(X_i)^2 - \frac{2}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} v(X_i)$$

This motivates an approach to *learn a representation* ϕ. We posit a parameterized representation $\phi(x;\theta_\phi)$ with values in a space Φ, and a scalar parameterized function $g(.;\theta_g)$ on Φ. Then we minimize $\mathcal{L}_{\mathcal{P},\mathcal{Q}}(g(\phi(.;\theta_\phi);\theta_g))$ wrt θ_ϕ, θ_g. Due to the compositionality of neural networks, we parameterize g and ϕ jointly in a neural network which is plugged into the AutoDML loss, similarly to the Riesz representer component of RieszNet [Chernozhukov et al., 2022a], and where a pre-specified, potentially low-dimensional hidden layer is later used as the representation ϕ [Clivio et al., 2022]. This is illustrated in Figure 1. Unlike RieszNet, we do not use any outcome information and we do not use the final Riesz representer head as the solution weight function, but instead plug the representation into a probability distance to obtain such a solution, as we will see shortly. We later show that this yields lower biases in our experiments.

3.3 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 [Martinet, 2020, Huling and Mak, 2024] by finding a function $f(a)$ minimizing

$$\mathbb{E}[(\mu(A) - f(A))^2]$$

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 wrt $P^{\text{data}}_{X|A=a}$. This is equivalent to minimizing

$$\mathbb{E}[\text{Bias}^2_{P^{\text{data}}(.|A),P^{\text{data}}(.)}(w_A)],$$

which is a special case of minimizing the **joint squared bias**

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

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 $P^\alpha, Q^\alpha, w^\alpha$ are a source distribution, a target distribution, a weight function indexed by $\alpha \in \Lambda$, respectively. The following corollary extends previous results on the balancing score error to the setting of simultaneous weighting problems.

Corollary 3.6. *Let Λ be a set endowed with a distribution $p_\Lambda(\alpha)$, P^Λ, Q^Λ be mappings from Λ to a distribution such that for any $\alpha \in \Lambda$, P^α, Q^α satisfy Assumptions 2.3, 3.1, 3.2. Then for any $\mathcal{M}^\Lambda$ such that $\forall\alpha \in \Lambda,\ \mathbb{E}_{P^\alpha}[\tilde{Y}|x] \in \mathcal{M}^\alpha$ where $\mathcal{M}^\alpha$ is a set of $L_2(P_X)$ functions, for any mapping ϕ^Λ from Λ to measurable representations, for any w^Λ such that each $w^\alpha(x)$ is an $L_2(P^\alpha_X)$ weight function wrt P^α depending on $\phi^\alpha(x)$,*

$$\begin{aligned}
&\frac{1}{2}\cdot \textit{Bias}^2_{P^\Lambda,Q^\Lambda}(w^\Lambda)\\
&\leq \mathbb{E}_{p_\Lambda(\alpha)}\left[\textit{IPM}^2_{\phi^\alpha(\mathcal{M}^\alpha,P^\alpha)}(P^{\alpha,w^\alpha}_{\phi^\alpha(X)}, Q^\alpha_{\phi^\alpha(X)})\right]\\
&\quad + \left(\sup_{\alpha\in\Lambda} ||\tilde{Y}||^2_{L_2(P^\alpha)}\right)\cdot \textit{BSE}^2_{P^\Lambda,Q^\Lambda,p_\Lambda}(\phi^\Lambda).
\end{aligned}$$

where we call

$$\textit{BSE}^2_{P^\Lambda,Q^\Lambda,p_\Lambda}(\phi^\Lambda) := \mathbb{E}_{p_\Lambda(\alpha)}[\textit{BSE}^2_{P^\alpha,Q^\alpha}(\phi^\alpha)]$$

*the **joint squared balancing score error**.*

We also note that this framework is identical to Problem 2.1 when Λ is of cardinality 1. Finally, we can extend the previous section to simultaneous weights, where we now find an indexed representation ϕ^Λ that minimizes the joint squared balancing score error. We do so by first positing a parameterized representation $\phi(x, \alpha; \theta_\phi)$ belonging to some space Φ and a scalar parameterized function $g(\varphi, \alpha; \theta_g)$ on the $\Phi \times \Lambda$ space, and then minimizing

$$\mathcal{L}^{g,\phi,p_\Lambda}_{\mathcal{P}^\Lambda,\mathcal{Q}^\Lambda}(\theta) = \mathbb{E}_{p_\Lambda(\alpha)}\left[\mathcal{L}_{\mathcal{P}^\alpha,\mathcal{Q}^\alpha}(g(\phi(.,\alpha;\theta_\phi),\alpha;\theta_g))\right]$$

wrt θ_ϕ, θ_g, where $\mathcal{P}^\alpha$, $\mathcal{Q}^\alpha$ are samples from P^α, Q^α.

If desired, we can separate the problem of minimizing the joint squared bias into independent weighting problems, minimizing each individual bias separately, especially when Λ is finite and with few elements. However, we can also share parameters or dependencies between individual problems, e.g. use the same representation for all problems, i.e. $\phi^\alpha := \phi$ for some ϕ for all $\alpha \in \Lambda$, or share parameters θ_ϕ, θ_g between problems $\alpha \in \Lambda$, notably when there are few samples for every $\mathcal{P}_\alpha$ or $\mathcal{Q}_\alpha$ as in ATE estimation with high-cardinal $\mathcal{A}$.

For completeness, we now give examples of $\Lambda, \mathcal{P}^\Lambda, \mathcal{Q}^\Lambda$. In ATE estimation, we have access to samples $\{(x_i, a_i, y_i)\}_{i=1,\cdots,n}$ of $P^{\text{data}}(X, A, Y)$. Then, $\Lambda = \mathcal{A}$ and for each $\alpha = a \in \mathcal{A}$, $\mathcal{P}^a = \{(x_i, y_i)\}_{i:a_i=a}$, $\mathcal{Q}^a = \mathcal{Q}^0 := \{x_i\}_{i=1,\cdots,n}$. In ATT estimation, where $\mathcal{A}$ is binary, then $\Lambda = \{0\}$, $\mathcal{P}^0 = \{(x_i, y_i)\}_{i:a_i=0}$, $\mathcal{Q}^0 = \{x_i\}_{i:a_i=1}$. In transportability, $\Lambda = \{0\}$, we have access to samples $\{(x_i, a_i, y_i)\}_{i=1,\cdots,n}$ of the RCT distribution $P^{\text{data}}(X, A, Y|S = 1)$, samples $\{(x_i)\}_{i=n+1,\cdots,n+m}$ of some observational data $P(X|S = 0)$, and $\pi = P^{\text{data}}(A = 1|S = 1)$, so $\Lambda = \{0\}$, $\mathcal{P}^0 = \{(x_i, \tilde{y}_i = \frac{a_i y_i}{\pi} - \frac{(1-a_i)y_i}{1-\pi})\}_{i=1,\cdots,n}$, $\mathcal{Q}^0 = \{x_i\}_{i=n+1,\cdots,n+m}$.

3.4 WEIGHTING AND ALGORITHM

Learning a representation by minimizing a bound of the BSE helped us minimize the second term of the RHS of Corollary 3.5. We now turn to minimizing the first term, that is **finding weights**. In finite samples, we aim to find discrete weights $w_i = w(X_i)$ for $i \in \mathcal{P}$, with constraints $\forall i \in \mathcal{P},\ w_i \geq 0$ and $\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} w_i = 1$.

In line with Proposition 3.3, we would ideally obtain $\hat{w}$ by solving Equation 1 with $\text{IPM}_{\phi(\mathcal{M},\mathcal{P})}(\mathcal{P}^w_{\phi(X)}, \mathcal{Q}_{\phi(X)})$ where $\mathcal{P}^w$ is the empirical distribution over $\mathcal{P}$ with probabilities $w_i/|\mathcal{P}|$. However, as $\mathcal{M}$ is unknown, $\text{IPM}_{\phi(\mathcal{M},P)}$ will remain unknown. Proposition 6 of Clivio et al. [2022] suggests that if $\mathcal{M}$ is the set of L-Lipschitz functions, ϕ is linear and X follows a Gaussian distribution in P, then $\phi(\mathcal{M}, P)$ is contained in the class of L'-Lipschitz functions for some L' that might be significantly larger than L.

For computational simplicity, we work with a canonical IPM and choose the maximal mean discrepancy wrt some kernel k [Gretton et al., 2012], following common practice in the literature [Kallus, 2020b, Huling and Mak, 2024]. More generally, minimizing such an MMD under the above weight constraints is referred to as *kernel optimal matching*

Input : Distribution $p_\Lambda(\alpha)$ over $\alpha \in \Lambda$, model $g(\phi(.,\alpha;\theta_\phi),\alpha;\theta_g)$, for each α: samples $\mathcal{P}^\alpha$, $\mathcal{Q}^\alpha$, kernel k^α, hyperparameter $\sigma^\alpha \geq 0$.

Initialize $\theta := (\theta_\phi, \theta_g)$;
while θ *not converged* **do**
  Move θ in direction $-\nabla_\theta \mathcal{L}^{g,\phi,p}_{\mathcal{P}^\Lambda,\mathcal{Q}^\Lambda}(\theta)$;
end
for $\alpha \in \Lambda$ **do**
  $\phi^\alpha(x) \leftarrow \phi(x,\alpha,\theta_\phi)$;
  $\tilde{k}^\alpha(x,x') \leftarrow k^\alpha(\phi^\alpha(x),\phi^\alpha(x'))$;
  $\hat{w}^\alpha \leftarrow$ kernel optimal matching with simplex weights, kernel $\tilde{k}^\alpha$ and regularization hyperparameter $(\sigma^\alpha)^2$;
end
Result: $\hat{w}^\Lambda$

Algorithm 1: Representation Learning and Weighting.

(KOM) with simplex weights [Kallus, 2020b] in causal inference, where we change the setting from treated and control distributions to target and source distributions, or empirical *kernel mean matching* (KMM) [Huang et al., 2006] in covariate shift, where we add L_2 regularization. This minimization amounts to solving a quadratic program (QP) with linear constraints, which can be done using any off-the-shelf QP solver. The σ^2 hyperparameter for regularization can be selected either with a fixed value (e.g. 0 as in Huling and Mak [2024]) or from a principled procedure [Kallus, 2020b]. In the case of simultaneous weightings, this procedure can be repeated for each problem indexed $\alpha = 1, \cdots, \ell$. Our exact implementation of kernel optimal/mean matching for this purpose is given in Appendix E

We summarize all the previous steps in Algorithm 1. Each component $\hat{w}^\alpha$ of its result $\hat{w}^\Lambda$ is then plugged in an estimator $\hat{\tau}^\alpha_{\hat{w}^\alpha}$ of $\mathbb{E}_{Q^\alpha}[\mathbb{E}_{P^\alpha}[\tilde{Y}|X]]$ as

$$\hat{\tau}^\alpha_{\hat{w}^\alpha} = \frac{1}{|\mathcal{P}^\alpha|} \sum_{i \in \mathcal{P}^\alpha} \hat{w}^\alpha_i \tilde{Y}_i.$$

This estimator could be analyzed theoretically (e.g. for consistency, error rates, ...) by inspecting, for each $\alpha \in \Lambda$, two separate terms : (i) the confounding bias of the learnt representation $\hat{\phi}^\alpha$, and (ii) the difference between the estimator and the representation-wise estimand $\mathbb{E}_{Q^\alpha}[\mathbb{E}_{P^\alpha}[\tilde{Y}|\hat{\phi}^\alpha(X)]]$. As the representation is learnt using the same loss as Equation 2.6 of Chernozhukov et al. [2024] and the confounding bias of the learnt representation is bounded by its balancing score error, itself bounded by the Riesz representer error in Theorem 2.1 of Chernozhukov et al. [2024], we can resort to their results. Then, the difference between estimator and representation-wise estimand can be analyzed using previous work on analysis of KOM or KMM, such as Kallus [2020b] or Yu and Szepesvári [2012].

4 RELATED WORK

Generalization bounds. An adjacent line of work to ours is generalization bounds in domain adaptation, where one aims to bound the risk of a model on a target domain using the risk on a source domain. Typically, this involves a representation and the bound includes an IPM analogous to ours [Ben-David et al., 2006, Zhao et al., 2018a,b, Li et al., 2023]. In extensions of such bounds to causal inference, where a counterfactual risk is bounded using a factual risk and an IPM as before, the representation is usually assumed to be invertible [Shalit et al., 2017, Bellot et al., 2022, Johansson et al., 2022, Kazemi and Ester, 2024], precluding the study of misspecified or confounded representations. Thus, usually no term quantifying the "misspecification" of the representation is added. Notable exceptions are Johansson et al. [2019] in domain generalization and Curth et al. [2021] in causal inference which include an *information loss* without actively trying to minimize it. The information loss from Johansson et al. [2019] can be shown to be identical to our confounding bias with the outcome replaced by the loss function. A balancing score error analogous to ours will bound this information loss if the loss function is bounded above by a constant ; then our AutoDML loss-based approach can be used too.

Confounding bias, balancing score error D'Amour and Franks [2021] also define a confounding bias and their Proposition 2 can be shown to be a special case of our Proposition 3.4 for ATE estimation, which they only compute in a restricted case with Gaussian covariates and generalized linear models for outcome and propensity models. Melnychuk et al. [2024] define a conditional confounding bias and estimate bounds of it for a *fixed* representation, instead of learning the representation using their bounds, which does not seem trivial as their estimation relies on two different neural network fitting steps *after* fitting the representation. Clivio et al. [2022] provide an alternative error on how much the representation is not a balancing score but they mention that it is difficult to compute and do not use it to learn the representation, which relies on assuming a propensity score model. Further, note that approaches to sensitivity analysis generally derive or bound the confounding bias induced by not including unobserved confounders in the adjustment set [Imbens, 2003, Tan, 2006, Hartman and Huang, 2024], although this is done by making assumptions on the relationship between unobserved confounders and other variables in the data generating process ; in contrast, aforementioned methods and our work pertain to a setting without such unobserved confounders.

Learning representations for treatment effects. For weighting, besides points developed in Section 2.3, DeepMatch [Kallus, 2020a] requires a grid search involving multiple neural network trainings (50 in the experiments) and other methods [Averitt et al., 2020, Kitazawa, 2022] take an

f-divergence as the discrepancy measure but do not provide bounds on the bias, which is likely inherent to the non-intersection of IPMs and f-divergences [Sriperumbudur et al., 2012]. Other methods learn representations using outcome regression, alone or with weights [Shalit et al., 2017, Johansson et al., 2022, Chernozhukov et al., 2022a].

Outcome-based weights and representations Some methods use outcomes to derive (i) the outcome function class $\mathcal{M}$ e.g. as a confidence interval around a regressed outcome model as in Wainstein [2022]; (ii) the representation ϕ as in the canonical prognostic scores [Hansen, 2008] or more recent and more general deconfounding scores [D'Amour and Franks, 2021]; or (iii) the weights more generally, e.g. by directly estimating the density ratio between the source and target distributions of the outcome [Taufiq et al., 2023]. Finally, many standard outcome modeling approaches, such as (kernel ridge) regression are implicitly weighting estimators so one could use such approaches to derive weights; see, for example, Bruns-Smith et al. [2023].

5 NUMERICAL RESULTS

We now evaluate our method and alternatives on the IHDP [Hill, 2011] and News datasets [Johansson et al., 2016] for ATE estimation and a Traumatic Brain Injury (TBI) dataset [Colnet et al., 2024] for transportability.

For weighting, we focus on KOM for two kernels, 1) the *energy distance* kernel $k(x, x') = \frac{1}{2}\left(||x||_2 + ||x'||_2 - ||x - x'||_2\right)$; KOM with this kernel is known as *energy balancing* [Huling and Mak, 2024] ; 2) the linear kernel $k(x, x') = x^T x'$. We evaluate these two methods with original covariates ("Energy" and "Linear"), a representation learned according to our approach ("Ours + Energy" and "Ours + Linear"), one through the canonical Principal Component Analysis (PCA, Hotelling [1933]) approach ("PCA + Energy" and "PCA + Linear"), 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" and "PS + Linear"), representations from a layer of a neural network model of such propensity score models as in neural score matching (NSM, Clivio et al. [2022]) ("NSM + Energy" and "NSM + Linear"). We also check IPW with the same propensity scores ("IPW"), entropy balancing with first-order moments ("Entropy"), the weights head of the neural network used to train our representation ("NN Head"), and uniform weights ("Unweighted"). Weights from "IPW" and "NN Head" were normalized to prevent outsize errors, while those from other methods were already normalized by design.

On energy balancing or linear kernel methods, we take $\sigma^\alpha = 0.01$. KOM was performed using the `osqp` library in Python [Stellato et al., 2020], in line with the implementation of energy balancing in the `weightit` package [Greifer, 2024]. All representations are 10-dimensional, and we always use a common representation 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. Neural network implementation was performed in PyTorch [Paszke et al., 2019]. Adam [Kingma and Ba, 2015] was used to optimize the loss with a learning rate of 0.01 and early stopping with a patience of 3 epochs, and all other hyperparameters at their default PyTorch values. We average results over 50 random seeds for IHDP and News, 100 for TBI. We show standard errors in parentheses.

Table 1: Joint Bias on the IHDP, News and TBI Datasets

Method	IHDP	News	TBI
Ours + Energy	0.079 (0.011)	0.128 (0.014)	5.00 (0.37)
NSM + Energy	0.167 (0.040)	0.070 (0.013)	5.40 (0.53)
PS + Energy	0.096 (0.012)	0.381 (0.026)	7.79 (0.53)
PCA + Energy	0.080 (0.014)	0.314 (0.020)	10.65 (0.82)
Energy	0.078 (0.014)	0.397 (0.027)	10.69 (0.83)
Ours + Linear	0.087 (0.009)	0.122 (0.013)	18.50 (1.61)
NSM + Linear	0.183 (0.043)	0.113 (0.018)	19.20 (1.71)
PS + Linear	0.105 (0.017)	0.499 (0.036)	13.22 (1.15)
PCA + Linear	0.077 (0.013)	0.321 (0.023)	63.86 (2.29)
Linear	0.076 (0.011)	0.168 (0.011)	22.71 (1.75)
Entropy	0.087 (0.013)	0.221 (0.020)	7.63 (0.60)
IPW	0.114 (0.024)	0.280 (0.018)	2.28 (0.18)
NN Head	0.181 (0.031)	0.746 (0.121)	59.71 (2.52)
Unweighted	0.195 (0.050	0.611 (0.053)	7.67 (0.15)

As a metric, we consider the joint bias (JB), which is the square-root of the joint squared bias where the target estimand is replaced by the average (known) outcome model

over the empirical target distribution,

$$\sqrt{\sum_{\alpha\in\Lambda} p_\Lambda(\alpha)\left(\frac{\sum_{i\in\mathcal{P}^\alpha} w_i^\alpha \tilde{Y}_i}{|\mathcal{P}^\alpha|} - \frac{\sum_{i\in\mathcal{Q}^\alpha}\mathbb{E}_{\mathcal{P}^\alpha}[\tilde{Y}|x_i]}{|\mathcal{Q}^\alpha|}\right)^2}.$$

Results are shown in Table 1. For either energy or linear KOM, our representation typically outperforms all other representations; exceptions are NSM on News for both kernels, the propensity score on TBI with the linear kernel, original covariates for both kernels and PCA for the linear kernel on IHDP. It further outperforms baselines not relying on KOM on IHDP and News for both kernels and on TBI for the energy balancing kernel, except entropy balancing for the linear kernel on IHDP and IPW on News. We note that the linear kernel yields generally degraded performance on TBI compared to the energy balancing kernel, but not other datasets. On IHDP, each KOM method performs better using original covariates than using a representation, which suggests that dimensionality reduction in any form is not necessarily beneficial on such a dataset where 16 out of 25 covariates are binary. Notably, on all datasets, using our representation with any KOM outperforms the Riesz representer head of the same neural network used to train the representation. Further, on the 3477-dimensional News dataset, energy balancing was significantly sped up when using a lower-dimensional representation instead of original covariates.

On TBI, high biases are due to a wide range of pseudo-outcomes (e.g. from -8.37 to 174.18, with a target estimand at 56.89 on seed 5), and the highest biases to weights with most of their mass on a single point with an pseudo-outcome far away from the target estimand (e.g. 97% of the mass on an pseudo-outcome of 145.40 for NN Head, compared to at most 5% on an pseudo-outcome of 117.97 for entropy balancing, both on seed 5).

6 LIMITATIONS AND CONCLUSION

We have shown the importance of the confounding bias and the balancing score error (BSE) in learning representations for weighting, and have outlined a method to minimize the BSE. Experimental results suggest that representations obtained from the method might help improve performance for common optimization-based weighting approaches. The method could notably be applied to multimodal data involving tabular, text and image covariates [Klaassen et al., 2024].

One concern could be that the functions g, ϕ are generally not uniquely identifiable by minimizing the AutoDML loss. Without restrictions, many different (g, ϕ) tuples will indeed share the same value of the AutoDML loss, e.g. any $(g_h, \phi_h) = (g \circ h, h^{-1} \circ \phi))$ where h is invertible. However, restricting g and ϕ to be components of a neural network with a given architecture will exclude many possible invertible h's. Some h's will remain though, such as $h(z) = \lambda \odot z$ where $\odot$ is the Hadamard product and $\lambda_i \neq 0\ \forall i$, which means that the returned ϕ might have arbitrary amplitude or smoothness. A workaround could be in adding some regularization of ϕ in the AutoDML loss, eg through weight decay. We do not perform weight decay and still obtain competitive performance in later experiments, which suggests that Adam optimization might choose an appropriate ϕ in practice.

Directions for future work to address limitations of our current approach include: (1) check whether such quantification of the representation's quality can also be done for augmented estimators, (2) evaluate the currently unknown gaps between the confounding bias and the BSE, and between the BSE and the AutoDML objective ; the latter provides a worst-case error but can be overly conservative ; (3) characterize the function class of the projection of the outcome model on the representation, depending on the class of the original outcome model or that of the representation, instead of assuming a canonical RKHS as we do now ; (4) develop a more thorough theoretical analysis of the estimator than the strategy presented in this paper.

Acknowledgements

We sincerely thank David Bruns-Smith, Sam Pimentel, Erin Hartman and anonymous reviewers 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

Shomoita Alam, Erica EM 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.

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

Amelia J Averitt, Natnicha Vanitchanant, Rajesh Ranganath, and Adler J Perotte. The counterfactual χ-gan: Finding comparable cohorts in observational health data. *Journal of Biomedical Informatics*, 109:103515, 2020.

Alexis Bellot, Anish Dhir, and Giulia Prando. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage. *arXiv preprint arXiv:2205.14692*, 2022.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

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

Eli Ben-Michael, Avi Feller, and Erin Hartman. Multilevel calibration weighting for survey data. *Political Analysis*, 32(1):65–83, 2024.

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.

David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. Augmented balancing weights as linear regression. *arXiv preprint arXiv:2304.14545*, 2023.

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.

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, 2022a.

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

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression, 2024. URL `http://arxiv.org/pdf/2104.14737v3`.

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.

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.

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Re-weighting the randomized controlled trial for generalization: finite-sample error and variable selection. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnae043, 05 2024. ISSN 0964-1998. doi: 10.1093/jrsssa/qnae043.

R Dennis Cook. *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons, 2009.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 34:26740–26753, 2021.

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

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

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

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.

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.

Kosuke Imai Eli Ben-Michael and Zhichao Jiang. Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, 0(0):1–14, 2024. doi: 10.1080/01621459.2023.2300507.

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.

Noah Greifer. *WeightIt: Weighting for Covariate Balance in Observational Studies*, 2024. R package version 1.1.0, https://github.com/ngreifer/WeightIt.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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.

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

Erin Hartman and Melody Huang. Sensitivity analysis for survey weights. *Political Analysis*, 32(1):1–16, 2024.

Chad Hazlett. Kernel balancing. *Statistica Sinica*, 30(3): 1155–1189, 2020.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

David A. Hirshberg and Stefan Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206 – 3227, 2021. doi: 10.1214/21-AOS2080.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Jared D Huling and Simon Mak. Energy balancing of covariate distributions. *Journal of Causal Inference*, 12(1): 20220029, 2024.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132, 2003.

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

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166): 1–50, 2022.

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

Nathan Kallus. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020b.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

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

Amirreza Kazemi and Martin Ester. Adversarially balanced representation for continuous treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13085–13093, 2024.

Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political analysis*, 27 (4):435–454, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. URL `http://arxiv.org/abs/1412.6980`.

Yoshiaki Kitazawa. Generalized balancing weights via deep neural networks. *arXiv preprint arXiv:2211.07533*, 2022.

Sven Klaassen, Jan Teichert-Kluge, Philipp Bach, Victor Chernozhukov, Martin Spindler, and Suhas Vijaykumar. Doublemldeep: Estimation of causal effects with multimodal data. *arXiv preprint arXiv:2402.01785*, 2024.

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

Xiaotong Li, Zixuan Hu, Jun Liu, Yixiao Ge, Yongxing Dai, and Ling-Yu Duan. Modeling uncertain feature representation for domain generalization. *arXiv preprint arXiv:2301.06442*, 2023.

Sharon L Lohr. *Sampling: design and analysis*. Chapman and Hall/CRC, 2021.

Wei Luo and Yeying Zhu. Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, 38(4):888–900, 2020.

Simon Mak and V. Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562 – 2592, 2018. doi: 10.1214/17-AOS1629.

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

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. 2024.

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

David Newman. Bag of Words. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C5ZG6P.

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.

Michal Ozery-Flato, Pierre Thodoroff, Matan Ninio, Michal Rosen-Zvi, and Tal El-Hay. Adversarial balancing for causal inference. *arXiv preprint arXiv:1810.07406*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

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.

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

Paul R Rosenbaum and Donald 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(2):212–218, 1983a.

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

Donald B Rubin. For objective causal inference, design trumps analysis. 2008.

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.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

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

Jae W Song and Kevin C Chung. Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, 126(6):2234–2242, 2010.

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

Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. Osqp: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal density ratio for off-policy evaluation in contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2023.

Leonard Wainstein. Targeted function balancing. *arXiv preprint arXiv:2203.12179*, 2022.

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.

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.

Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL `http://icml.cc/2012/papers/330.pdf`.

Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. 31, 2018a.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018b.

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.

Towards Representation Learning for Weighting Problems in Design-Based Causal Inference (Supplementary Material)

Oscar Clivio[1] **Avi Feller**[2] **Chris Holmes**[1]

[1]Department of Statistics, University of Oxford
[2]Goldman School of Public Policy and Department of Statistics, University of California, Berkeley

A 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** [Ben-Michael et al., 2021], 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** [Huling and Mak, 2024], 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 **generalizability** and **transportability** [Colnet et al., 2024, Degtiar and Rose, 2023], A is binary again and we have an other binary variable S such that $S = 1$ denotes membership in the RCT population, that is $A \perp\!\!\!\perp X|S = 1$ and $(Y(1), Y(0)) \perp\!\!\!\perp A|S = 1$. We are interested in $\mathbb{E}[Y(1) - Y(0)]$ for generalizability and $\mathbb{E}[Y(1) - Y(0)|S = 0]$ for transportability. What motivates weighting here is that we do not have access to A, Y when $S = 0$. Under the *transportability assumption*, the *conditional average treatment effect* is identical between RCT and non-RCT populations, i.e. for any x, $\text{CATE}(x) := \mathbb{E}[Y(1) - Y(0)|x]$ is equal to both $\mathbb{E}[Y(1) - Y(0)|x, S = 1]$ and $\mathbb{E}[Y(1) - Y(0)|x, S = 0]$. In addition, the CATE is identified on the RCT population as $\text{CATE}(x) = \mathbb{E}\left[\frac{AY}{P(A=1|S=1)} - \frac{(1-A)Y}{P(A=0|S=1)}\middle| X = x, S = 1\right]$. Then, defining $\pi = P(A = 1|S = 1)$, the weighting approach is to reweight the distribution of the RCT population, i.e. find weights w such that

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[w(X) \cdot \text{CATE}(X)|S = 1] \approx \frac{1}{|\{i : S_i = 1\}|} \sum_{i:S_i=1} w(X_i) \left(\frac{A_i Y_i}{\pi} - \frac{(1 - A_i)Y}{1 - \pi}\right)$$

in generalizability or such that

$$\mathbb{E}[Y(1) - Y(0)|S = 0] = \mathbb{E}[w(X) \cdot \text{CATE}(X)|S = 1] \approx \frac{1}{|\{i : S_i = 1\}|} \sum_{i:S_i=1} w(X_i) \left(\frac{A_i Y_i}{\pi} - \frac{(1 - A_i)Y}{1 - \pi}\right)$$

in transportability. Due to the similarity of both frameworks, without loss of generality, we focus on transportability as in Colnet et al. [2024] and Egami and Hartman [2021] which study variable selection in this setting.

Note that the framework of Problem 2.1 generally does not allow for **CATE estimation**, as the CATE is a function and the target $\mathbb{E}_Q[\mathbb{E}_P[\tilde{Y}|X]]$ is a scalar. Alternatively, one can perform simultaneous weightings as in Section 3.3, where for every problem we fix a covariate value x_0 and a treatment value a and take the target estimand to be the $\mathbb{E}[Y|A=a, X=x_0]$. We can take the pseudo-outcome to be Y, the source distribution to be $P^{\text{data}}(X, Y|A=a)$ and the target distribution to be $P^{\text{data}}(X|X=x_0)$. However, this choice of target distribution would be a spike at $X=x_0$, potentially violating Assumption 2.3 in many widely-applicable situations, e.g. if the source distribution of covariates has a density wrt the Borel measure. As in Ben-Michael et al. [2021], such a problem with spikes could be mitigated with smoothing. If Assumption 2.3 does hold, e.g. if the source distribution of covariates is discrete and has a non-zero mass at x_0, then it could actually be possible to perform weighting, although we are not aware of such an approach in the previous literature.

B GENERALIZATION OF FORMER "SCORE" NOTIONS

More rigorously, the confounding bias is zero in three important cases:

1. $\mathbb{E}_P[\tilde{Y}|x]$ is a function of $\phi(x)$ P_X-a.s., where we call $\phi(x)$ a "generalized prognostic score" ;
2. $\frac{\mathrm{d}Q}{\mathrm{d}P}(x)$ is a function of $\phi(x)$ P_X-a.s., where we call $\phi(x)$ a "generalized balancing score" ;
3. The confounding bias is zero without ϕ necessarily being a generalized prognostic or balancing score, where we call $\phi(x)$ a "generalized deconfounding score".

The following result connects these notions to previous literature.

Proposition B.1. *In ATT/ATE estimation, a) balancing scores [Rosenbaum and Rubin, 1983b] are equivalent to generalized balancing scores . In ATE estimation, b) deconfounding scores [D'Amour and Franks, 2021] are equivalent to generalized deconfounding scores, c) prognostic scores [Hansen, 2008] are generalized prognostic scores, and the converse is true if*

$$\forall a \in \mathcal{A}, Y(a) \perp\!\!\!\perp X \mid \mathbb{E}[Y|X, A=a].$$

In transportability [Egami and Hartman, 2021], assuming that the transportability assumption holds for X, d) heterogeneity sets are generalized prognostic scores, e) sampling sets are generalized balancing scores, f) separating sets are generalized deconfounding scores.

Thus, these "generalized" scores extend existing notions of prognostic, balancing and deconfounding scores from the literature to the more general framework from Problem 2.1 and connect them to the confounding bias, refining our understanding of why these scores are well-suited for weighting. They also connect notions used for variable selection in transportability to the score notions from weighting for the ATT or the ATE.

We might say that the generalized notions clearly outline the "proper" definitions of their original counterparts, in a sense that they are either equivalent to them, or weaker than them while preserving properties required for deconfounding, as illustrated by generalized prognostic scores. Hence, for the remainder of the paper, we omit the "generalized" adjective from all notions of scores.

C REPRESENTATION SELECTION

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 minimization is taken over all functions. These are inaccessible, but we can instead perform each minimization under a rich parameterized 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 [Arbour et al., 2021] 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 2.3 and Proposition 3.4.

D DETAILS ON EXPERIMENTS

Code Our code is available at `https://github.com/oscarclivio/representations_weighting`.

Origin of datasets We extracted IHDP [Hill, 2011] from the GitHub repository for Dragonnet [Shi et al., 2019] at `https://github.com/claudiashi57/dragonnet/tree/master/dat/ihdp/csv`, News [Johansson et al., 2016] from `https://www.fredjo.com/files/NEWS_csv.zip` and TBI [Colnet et al., 2024] from `https://github.com/BenedicteColnet/IPSW-categorical`. In addition, TBI is covered by a MIT license, and the original data source for News [Newman, 2008] by a CC BY 4.0 license.

Infrastructure We ran experiments on a laptop with a GeForce GTX 1070 GPU with Max-Q Design and 12 CPU core. We used our own Python implementation for datasets (after downloading the data), weighting methods and representation learning techniques, including propensity score modelling and neural network fitting.

Choice of hyperparameters We tried different sets of hyperparameters for neural networks, and first chose a set such that our representation had good performance (outperformed by at most one other method) on two different datasets each in a separate task (ATT estimation, ATE estimation, transportability) under energy balancing as the kernel optimal matching. Several sets verified this property, however performance of individual methods and individual hyperparameters was generally unequal among datasets. For ATE estimation, our method had the same ranking as in the paper for many hyperparameters on News, but was outperformed by standard kernel balancing on IHDP [Hill, 2011, Shalit et al., 2017] and ACIC 2016 [Dorie et al., 2019]. For transportability, rankings were also identical to those in the paper across multiple hyperparameters on TBI. For ATT estimation, at the time of writing, we did not find such "good" hyperparameters on News and ACIC2016, but did so on IHDP and Jobs [LaLonde, 1986, Johansson et al., 2016]. This generally shows that different hyperparameters should be tested, especially for neural network-based methods. Defining a principled way (that does not use ground-truth target estimands) to select them for weighting has to be addressed in future work.

Misspecified outcome classes We note that in our experiments, the outcome class $\mathcal{M}$ is often not correctly specified. Indeed, note that in our datasets

- The control outcome model in IHDP in its setting B [Hill, 2011], as used to generate the data [Shi et al., 2019], linearly depends on an exponential $e^{\beta.x}$ function for some β.
- The outcome models for News linearly depend on the vector of topic probabilities $z(x)$ in Johansson et al. [2016], where we note that weights in this linear relationship are further positive. Noting k the number of topics, we then have for any x, $z_i(x) \geq \frac{1}{k}$ for at least one i, and noting w_0 the minimal weight we obtain that $\forall a = 0, 1,\ \mathbb{E}[Y|x, A = a] \geq \frac{w_0}{k}$, thus either treated or control outcome model is bounded away from 0.
- The outcome model for TBI is quadratic in x [Colnet et al., 2024].

Thus, outcome functions on IHDP, News and TBI are clearly not linear functions, thus misspecified for linear kernel optimal matching. For energy balancing, none of the outcome functions above is square-integrable (here *without* a probability measure; IHDP and TBI due to their functional forms, News due to being bounded away from 0), thus none of them is Sobolev of any order. As the covariate space for IHDP and News has an odd dimensionality, these functions are misspecified outcome functions for the class corresponding to the energy distance according to page 12 of Mak and Joseph [2018]. This is less clear for the outcome model of the TBI dataset which has even dimensionality ; we conjecture that this outcome model is misspecified too, as its outcome model is not Sobolev of any order and Sobolev spaces up to a certain order are invoked as canonical members of the outcome class corresponding in Huling and Mak [2024].

E DETAILS ON OUR IMPLEMENTATION OF KERNEL OPTIMAL OR MEAN MATCHING

When applied to the representation ϕ composed through a kernel k, the square of the MMD is:

$$\begin{aligned}\text{MMD}_k^2(\mathcal{P}_w, \mathcal{Q}) = &\frac{1}{|\mathcal{P}|^2} \sum_{i,j \in \mathcal{P}} w_i w_j k(\phi(x_i), \phi(x_j)) \\ &- \frac{2}{|\mathcal{P}||\mathcal{Q}|} \sum_{i \in \mathcal{P}, j \in \mathcal{Q}} w_i k(\phi(x_i), \phi(x_j))\end{aligned}$$

$$+ \frac{1}{|\mathcal{Q}|^2} \sum_{i,j \in \mathcal{Q}} k(\phi(x_i), \phi(x_j)).$$

Thus, its minimization with regularization is a quadratic problem (QP)

$$\min_w \frac{1}{2} w^T S w + v^T w \text{ subject to } l \leq Aw \leq u \tag{9}$$

that can be solved with any off-the-shelf solver $\texttt{solver}(S, v, l, A, u)$ (e.g. Stellato et al. [2020]). Noting $I_{\mathcal{P}}$ the identity matrix over $\mathcal{P}$, we have

$$\begin{aligned}
&S = S^{k,\phi,\sigma}_{\mathcal{P},\mathcal{Q}} := (2/|\mathcal{P}|^2 \cdot k(\phi(x_i), \phi(x_j)) + 2\sigma^2 \cdot I_{\mathcal{P}})_{i,j \in \mathcal{P}}, \\
&v := v^{k,\phi}_{\mathcal{P},\mathcal{Q}} = (-2/|\mathcal{P}||\mathcal{Q}| \cdot \sum_{j \in \mathcal{Q}} k(\phi(x_i), \phi(x_j)))_{i \in \mathcal{P}}, \\
&A := A_{\mathcal{P}} = \begin{pmatrix} I_{\mathcal{P}} \\ \hline 1 \cdots 1 \end{pmatrix}, \; l := l_{\mathcal{P}} = (\underbrace{0, \cdots, 0}_{|\mathcal{P}| \text{ times}}, |\mathcal{P}|)^T, \\
&u = u_{\mathcal{P}} = (\underbrace{+\infty, \cdots, +\infty}_{|\mathcal{P}| \text{ times}}, |\mathcal{P}|)^T
\end{aligned}$$

For the joint squared bias with a finite $\Lambda = \{1, \cdots, \ell\}$, we sum all objectives from Proposition 3.3 over each $\alpha = 1, \cdots, \ell$ with $\mathcal{P}^\alpha$, $\mathcal{Q}^\alpha$, ϕ^α, σ^α and with kernel k^α, giving

$$\begin{aligned}
S = S^{k^\Lambda,\phi^\Lambda,\sigma^\Lambda}_{\mathcal{P}^\Lambda,\mathcal{Q}^\Lambda} &:= \text{diag}\left((S^{k^i,\phi^i,\sigma^i}_{\mathcal{P}^i,\mathcal{Q}^i})_{i=1,\cdots,\ell}\right) \\
v = v^{k^\Lambda,\phi^\Lambda}_{\mathcal{P}^\Lambda,\mathcal{Q}^\Lambda} &:= \begin{pmatrix} v^{k^1,\phi^1}_{\mathcal{P}^1,\mathcal{Q}^1} \\ \vdots \\ v^{k^\ell,\phi^\ell}_{\mathcal{P}^\ell,\mathcal{Q}^\ell} \end{pmatrix}, \; \psi = \psi_{\mathcal{P}^\Lambda} := \begin{pmatrix} \psi_{\mathcal{P}^1} \\ \vdots \\ \psi_{\mathcal{P}^\ell} \end{pmatrix},
\end{aligned} \tag{10}$$

for $\psi \in A, l, u$. This step is agnostic to how σ^α is selected, either with a fixed value (e.g. 0 as in Huling and Mak [2024]) or from a principled procedure [Kallus, 2020b].

F PROOF OF RESULTS

F.1 PROOF OF PROPOSITION 3.3

F.1.1 Item 1

First, note that we can restrict our attention to weight functions w in $L_2(P_X)$, that is such that $w(X) \in L_2(P)$, as the objective will be ∞ for weights functions not in $L_2(P_X)$. For any $L_2(P_X)$ weight function and function $g \in \mathcal{G}$, we have

$$\begin{aligned}
\left|\mathbb{E}_{P^w_X}[g \circ \phi] - \mathbb{E}_{Q_X}[g \circ \phi]\right| &= \left|\mathbb{E}_{P^w}[(g \circ \phi)(X)] - \mathbb{E}_Q[(g \circ \phi)(X)]\right| \\
&= \left|\mathbb{E}_{P^w}[g(\phi(X))] - \mathbb{E}_Q[g(\phi(X))]\right| \\
&= \left|\mathbb{E}_{P^w_{\phi(X)}}[g] - \mathbb{E}_{Q_{\phi(X)}}[g]\right|.
\end{aligned}$$

where all integrals are well-defined, as $g(\phi(X)) \in L_2(P)$ by assumption in the Proposition and $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) \in L_2(P)$ from Assumptions 2.3 and 3.1. 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 $\mathcal{M} \subseteq L_2(P_X)$, and that all of this also justifies the claim that the bias wrt ϕ is bounded by $\text{IPM}_{\mathcal{G}}(P^w_{\phi(X)}, Q_{\phi(X)})$. Thus, we are solving

$$\min_{w \in \mathcal{A}} J(w)$$

where

$$\begin{aligned} A &:= \{w \in L_2(P_X) \mid w \geq 0 \; P_X\text{-a.s.}, \mathbb{E}_P[w(X)] = 1\} \\ J(w) &:= I_{\mathcal{M}}(w)^2 + \sigma^2 \cdot S(w) \\ I_{\mathcal{M}}(w) &:= \text{IPM}_{\mathcal{M}}(P_X^w, Q_X) \\ S(w) &:= \mathbb{E}_P[w(X)^2] \end{aligned}$$

where functions in $L_2(P_X)$ are identified P_X-a.s.. We note that $\inf_{w \in \mathcal{A}} J(w)$ is finite, as $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X} \in A$ from Assumption 3.1 and $J(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}) = \sigma^2 \cdot E_P\left[\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\right)^2\right] < \infty$.

We prove the first item of the Proposition in three parts :

1. There is at most one solution.
2. There is at least one solution.
3. Any solution is a function of $\phi(x)$ P_X-a.s.

Note that (i) only the third part uses the fact that functions in $\mathcal{M}$ are functions of $\phi(x)$ P_X-a.s., (ii) under stronger assumptions on the class $\mathcal{G}$, the result also follows directly from Theorem 4.1 of Bruns-Smith and Feller [2022], while the following analysis presents more relaxed assumptions over $\mathcal{G}$.

Part 1 : There is at most one solution A is clearly a convex subset of $L_2(P_X)$, and J is strictly convex. Indeed, for any $t \in [0,1], w_1, w_2 \in A$, $m \in \mathcal{M}$, letting $w_t = tw_1 + (1-t)w_2$

$$\begin{aligned} &|\mathbb{E}_{P^{w_t}}[m(X)] - \mathbb{E}_Q[m(X)]| \\ &= |t\left(\mathbb{E}_{P^{w_1}}[m(X)] - \mathbb{E}_Q[m(X)]\right) + (1-t)\left(\mathbb{E}_{P^{w_2}}[m(X)] - \mathbb{E}_Q[m(X)]\right)| \\ &\leq t\,|\mathbb{E}_{P^{w_1}}[m(X)] - \mathbb{E}_Q[m(X)]) + (1-t)\,|\mathbb{E}_{P^{w_2}}[m(X)] - \mathbb{E}_Q[m(X)]| \text{ from the triangle inequality} \\ &\leq tI_{\mathcal{M}}(w_1) + (1-t)I_{\mathcal{M}}(w_2) \text{ taking the supremum wrt } m \text{ on each term on the RHS} \end{aligned}$$

so taking the supremum wrt m on the LHS, $I_{\mathcal{M}}(w_t) \leq tI_{\mathcal{M}}(w_1) + (1-t)I_{\mathcal{M}}(w_2)$; thus $I_{\mathcal{M}}$ is convex. As $u \mapsto u^2$ is convex non-decreasing, $I_{\mathcal{M}}^2$ is convex. Also, for any $t \in [0,1], w_1, w_2 \in A$, $m \in \mathcal{M}$, again letting $w_t = tw_1 + (1-t)w_2$,

$$\begin{aligned} &tS(w_1) + (1-t)S(w_2) - S(w_t) \\ &= t\mathbb{E}_P[w_1(X)^2] + (1-t)\mathbb{E}_P[w_2(X)^2] - \mathbb{E}_P[(tw_1(X) + (1-t)w_2(X))^2] \\ &= t\mathbb{E}_P[w_1(X)^2] + (1-t)\mathbb{E}_P[w_2(X)^2] - t^2\mathbb{E}_P[w_1(X)^2] - (1-t)^2\mathbb{E}_P[w_2(X)^2] - 2t(1-t)\mathbb{E}_P[w_1(X)w_2(X)] \\ &= t(1-t)\mathbb{E}_P[(w_1(X) - w_2(X))^2] \end{aligned}$$

which is non-negative, and zero iff $t = 0$, $t = 1$ or $w_1 = w_2$ P_X-a.s.. Thus, S is strictly convex. Thus, the sum of $I_{\mathcal{M}}^2$ and $\sigma^2 \cdot S$, that is J, is strictly convex.

As A is a convex subset of $L_2(P_X)$ and J is strictly convex, there is at most one minimizer of J in A.

Part 2 : There is at least one solution. From e.g. Theorem 2 of `https://www.math.umd.edu/~yanir/742/742-5-6.pdf`, the existence of a minimizer of J in A is guaranteed if A is weakly closed and J is coercive and sequentially weakly lower semi-continuous.

First, we show that A is weakly closed. Let $w_n \in A^{\mathbb{N}}$ weakly converging to some $w_* \in L_2(P_X)$, that is such that

$$\forall h \in L_2(P_X), \;\; \mathbb{E}_P[w_n(X)h(X)] \xrightarrow[n\to\infty]{} \mathbb{E}_P[w_*(X)h(X)].$$

Then taking $h = 1$, we have $1 = \mathbb{E}_P[w_n(X)] \xrightarrow[n\to\infty]{} \mathbb{E}_P[w_*(X)]$, thus $\mathbb{E}_P[w_*(X)] = 1$.

Further, for $k \in \mathbb{N}^*$, let $B_k := \{w_*(X) \leq -\frac{1}{k}\}$. Then, with $h := 1_{B_k}$, as $w_n \geq 0$ P_X-a.s. for each $n \in \mathbb{N}$

$$0 \leq \mathbb{E}_P[h(X)w_n(X)] \xrightarrow[n\to\infty]{} \mathbb{E}_P[w_*(X)h(X)] \leq -\frac{P(B_k)}{k}$$

which leads to $0 \leq \mathbb{E}_P[w_*(X)h(X)] \leq -\frac{P(B_k)}{k}$, which is not contradictory only if $P(B_k) = 0$. Then, as $\{w_*(X) < 0\} = \{\cup_{k\in\mathbb{N}^*} B_k\}$,

$$\begin{aligned} P(w_*(X) < 0) &= P\left(\cup_{k\in\mathbb{N}^*} B_k\right) \\ &\leq \sum_{k\in\mathbb{N}^*} P(B_k) \\ &= 0 \end{aligned}$$

Thus, $w^* \geq 0$ P_X-a.s.. As a result, $w_* \in A$, so A is weakly closed. We note that J is coercive, as S is clearly coercive and $I_{\mathcal{M}}$ is non-negative. What is left to prove in this part is then that J is sequentially weakly lower semi-continuous. Let $w_n \in A^{\mathbb{N}}$ weakly converging to some $w_* \in A$. We want to show that

$$\liminf_{n\to\infty} J(w_n) \geq J(w_*).$$

Indeed,

$$\begin{aligned} \liminf_{n\to\infty} I_{\mathcal{M}}(w_n)^2 &= \liminf_{n\to\infty} \sup_{m\in\mathcal{M}} \left|\mathbb{E}_P[w_n(X)m(X)] - \mathbb{E}_Q[m(X)]\right|^2 \\ &\geq \sup_{m\in\mathcal{M}} \liminf_{n\to\infty} \underbrace{\left|\mathbb{E}_P[w_n(X)m(X)] - \mathbb{E}_Q[m(X)]\right|^2}_{\xrightarrow[n\to\infty]{} \left|\mathbb{E}_P[w_*(X)m(X)] - \mathbb{E}_Q[m(X)]\right|^2 \text{ as m } \in L_2(P_X)} \\ &= \sup_{m\in\mathcal{M}} \left|\mathbb{E}_P[w_*(X)m(X)] - \mathbb{E}_Q[m(X)]\right|^2 \\ &= I_{\mathcal{M}}(w_*)^2 \end{aligned}$$

and by convexity of $u \mapsto u^2$,

$$\forall x,\ w_n(x)^2 \geq w_*(x)^2 + 2w_*(x)\left(w_n(x) - w_*(x)\right)$$

so

$$\begin{aligned} \liminf_{n\to\infty} S(w_n) &= \liminf_{n\to\infty} \mathbb{E}_P[w_n(X)^2] \\ &\geq \liminf_{n\to\infty} \mathbb{E}_P[w_*(X)^2] + 2\Big(\underbrace{\mathbb{E}_P[w_n(X)w_*(X)] - \mathbb{E}_P[w_*(X)^2]}_{\xrightarrow[n\to\infty]{} 0}\Big) \\ &= \mathbb{E}_P[w_*(X)^2] \\ &= S(w_*) \end{aligned}$$

and

$$\begin{aligned} \liminf_{n\to\infty} I_{\mathcal{M}}(w_n)^2 + \sigma^2 S(w_n) &\geq \liminf_{n\to\infty} I_{\mathcal{M}}(w_n)^2 + \liminf_{n\to\infty} \sigma^2 S(w_n) \\ &\geq I_{\mathcal{M}}(w_*)^2 + \sigma^2 S(w_*) \text{ from the above.} \end{aligned}$$

All of this shows that J is sequentially weakly lower semi-continuous, concluding this part of the proof.

Part 3 : Any solution is a function of $\phi(x)$. For any $w \in L_2(P_X)$, let $\bar{w}(z) = \mathbb{E}_P[w(X)|\phi(X) = z]$. If $w \in A$, then $\bar{w}(\phi(.)) \in A$. Indeed, the conditional expectation of any $L_2(P)$ random variable is also $L_2(P)$, so $\bar{w}(\phi(.)) \in L_2(P_X)$. Further, the conditional expectation of any almost surely non-negative random variable is also almost surely non-negative, so $\bar{w}(\phi(.)) \geq 0$ P_X-a.s.. Finally, the tower property shows that

$$\mathbb{E}_P[\bar{w}(\phi(X))] = \mathbb{E}_P[\mathbb{E}_P[w(X)|\phi(X)]] = \mathbb{E}_P[w(X)] = 1.$$

Thus, $\bar{w}(\phi(.)) \in A$. It actually turns out that $J(\bar{w}(\phi(.))) \leq J(w)$, with equality iff $w = \bar{w}(\phi(.))$. This concludes the proof, as a minimizer of J in A has to be a function of $\phi(x)$, as otherwise we can construct a weight function in A that realises a strictly lower objective, which is contradictory.

First,

$$\begin{aligned}\forall g \in \mathcal{G},\ \mathbb{E}_P[w(X)g(\phi(X))] &= \mathbb{E}_P[\mathbb{E}_P[w(X)g(\phi(X))|\phi(X)]] \text{ from the tower property} \\ &= \mathbb{E}_P[\mathbb{E}_P[w(X)|\phi(X)]g(\phi(X))] \\ &= \mathbb{E}_P[\bar{w}(\phi(X))g(\phi(X))]\end{aligned}$$

so $I_{\mathcal{M}}(\bar{w}(\phi(.))) = I_{\mathcal{M}}(w)$. Further,

$$\begin{aligned}S(w) - S(\bar{w}(\phi(.))) &= \mathbb{E}_P[w(X)^2] - \mathbb{E}_P[\mathbb{E}_P[w(X)|\phi(X)]^2] \\ &= \mathbb{E}_P[\mathbb{E}_P[w(X)^2|\phi(X)]] - \mathbb{E}_P[\mathbb{E}_P[w(X)|\phi(X)]^2] \text{ from the tower property} \\ &= \mathbb{E}_P\left[\mathbb{E}_P[w(X)^2|\phi(X)]] - \mathbb{E}_P[w(X)|\phi(X)]^2\right] \\ &= \mathbb{E}_P[\text{Var}(w(X)|\phi(X))] \\ &= \mathbb{E}_P[\mathbb{E}_P[(w(X) - \bar{w}(\phi(X)))^2|\phi(X)]] \\ &= \mathbb{E}_P[(w(X) - \bar{w}(\phi(X)))^2] \text{ from the tower property.}\end{aligned}$$

Taken all together, $J(w) \geq J(\bar{w}(\phi(.)))$ with equality iff $\mathbb{E}_P[(w(X) - \bar{w}(\phi(X)))^2] = 0$, that is $w = \bar{w}(\phi(.))$ P_X-a.s.. This concludes the proof.

F.1.2 Item 2

Let w be an $L_2(P_X)$ weight function such that $w = \bar{w}(\phi(.))$ P_X-a.s. for some $\bar{w}$. Then,

$$\begin{aligned}\text{Chosen weights bias of } w &= \mathbb{E}_{P^w}\left[\mathbb{E}_P[\tilde{Y}|X] - \mathbb{E}_P[\tilde{Y}|\phi(X)]\right] \\ &= \mathbb{E}_P\left[w(X)\mathbb{E}_P[\tilde{Y}|X] - w(X)\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] \\ &= \mathbb{E}_P\left[\bar{w}(\phi(X))\mathbb{E}_P[\tilde{Y}|X] - \bar{w}(\phi(X))\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] \\ &= \mathbb{E}_P\left[\bar{w}(\phi(X))\mathbb{E}_P[\tilde{Y}|X]\right] - \mathbb{E}_P\left[\bar{w}(\phi(X))\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] \\ &= \mathbb{E}_P\left[\mathbb{E}_P[\bar{w}(\phi(X))\tilde{Y}|X]\right] - \mathbb{E}_P\left[E_P[\bar{w}(\phi(X))\tilde{Y}|\phi(X)]\right] \\ &= \mathbb{E}_P[\bar{w}(\phi(X))\tilde{Y}] - E_P[\bar{w}(\phi(X))\tilde{Y}] \text{ from the tower property} \\ &= 0\end{aligned}$$

F.2 PROOF OF PROPOSITION 3.4

Let Σ_Z denote the σ-algebra of the space of values taken by random variable Z.

Let $B \in \Sigma_{\phi(X)}$ such that $P_{\phi(X)}(B) = 0$. Then $0 = P_{\phi(X)}(B) = P_X(\phi^{-1}(B))$ where $\phi^{-1}(B) \in \Sigma_X$ as ϕ is measurable. By Assumption 2.3, $Q_X(\phi^{-1}(B)) = 0$. Then $0 = Q_X(\phi^{-1}(B)) = Q_{\phi(X)}(B)$. Thus, $Q_{\phi(X)}$ is absolutely continuous wrt $P_{\phi(X)}$.

Notably, from the Radon-Nikodym theorem, $\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}$ exists. Then for any $B \in \Sigma_{\phi(X)}$,

$$\begin{aligned}&\mathbb{E}_P\left[\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X)) \cdot 1_B(\phi(X))\right] \\ &= \mathbb{E}_Q[1_B(\phi(X))] \\ &= \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) \cdot 1_B(\phi(X))\right] \text{ by taking the Radon-Nikodym derivative wrt } X \\ &= \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) \cdot 1_B(\phi(X))\middle|\phi(X)\right]\right] \text{ from the tower property} \\ &= \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X)\right] \cdot 1_B(\phi(X))\right]\end{aligned}$$

where all integrals are well-defined as the Radon-Nikodym derivative is measurable and $L_1(P_X)$, and its conditional expectation is also $L_1(P_{\phi(X)})$ as any conditional expectation of any $L_1(P)$ random variable is also $L_1(P)$.

Thus we have shown that $\forall B \in \Sigma_{\phi(X)},\ \int h \cdot 1_B \mathrm{d}P_{\phi(X)} = 0$ where $h(z) = \frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(z) - \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X) = z\right]$. We now show that $h = 0$, which concludes the proof for the first part of the Proposition. Note that h is measurable as any Radon-Nikodym derivative is measurable, and any conditional expectation is measurable. Notably, as $\mathbb{R}_+$ and $\mathbb{R}_-$ are in the Borel σ-algebra, $B_+ = h^{-1}(\mathbb{R}_+)$ and $B_- = h^{-1}(\mathbb{R}_-)$ are in $\Sigma_{\phi(X)}$. Thus,

$$0 = \int_{\mathcal{Z}} h \cdot 1_{B_+} \mathrm{d}P_{\phi(X)} = \int_{\mathcal{Z}} h_+ \mathrm{d}P_{\phi(X)}$$
$$0 = \int_{\mathcal{Z}} h \cdot 1_{B_-} \mathrm{d}P_{\phi(X)} = -\int_{\mathcal{Z}} h_- \mathrm{d}P_{\phi(X)}$$

which implies that $h_+ = 0$ and $h_- = 0$, both $P_{\phi(X)}$-a.s., as these two functions are non-negative. Thus, $h = 0$ $P_{\phi(X)}$-a.s., which concludes the first part of proof.

Now we further assume Assumptions 3.1 and 3.2. Then, we note that the confounding bias is equal to $-\mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\left(\mathbb{E}_P[\tilde{Y}|X] - \mathbb{E}_P[\tilde{Y}|\phi(X)]\right)\right]$. As $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ is now a $L_2(P_X)$ weight function wrt P, and using that $\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}} = \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X) = .\right]$ $P_{\phi(X)}$-a.s., identical computations as in the proof of item 2 in Proposition 3.3 show that $\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(.))$ is also a $L_2(P_X)$ weight function wrt P, while being a function of $\phi(x)$. Applying Proposition 3.3, item 2, to $\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(.))$ leads to $\mathbb{E}_P\left[\frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))\left(\mathbb{E}_P[\tilde{Y}|X] - \mathbb{E}_P[\tilde{Y}|\phi(X)]\right)\right] = 0$. Summing this to the confounding bias leads to

$$\text{Confounding bias} = -\mathbb{E}_P\left[\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - \frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))\right) \cdot \left(\mathbb{E}_P[\tilde{Y}|X] - \mathbb{E}_P[\tilde{Y}|\phi(X)]\right)\right].$$

Finally,

$$\begin{aligned}
&\mathbb{E}_P\left[\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - \frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))\right)\mathbb{E}_P[\tilde{Y}|\phi(X)\right] \\
&= \mathbb{E}_P\left[\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X)\right]\right)\mathbb{E}_P[\tilde{Y}|\phi(X)\right] \text{ from the first part of the Proposition} \\
&= \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] - \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\middle|\phi(X)\right]\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] \\
&= \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] - \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\mathbb{E}_P[\tilde{Y}|\phi(X)]\middle|\phi(X)\right]\right] \\
&= \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] - \mathbb{E}_P\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\mathbb{E}_P[\tilde{Y}|\phi(X)]\right] \text{ from the tower property} \\
&= 0
\end{aligned}$$

Thus,

$$\text{Confounding bias} = -\mathbb{E}_P\left[\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X) - \frac{\mathrm{d}Q_{\phi(X)}}{\mathrm{d}P_{\phi(X)}}(\phi(X))\right) \cdot \mathbb{E}_P[\tilde{Y}|X]\right].$$

F.3 PROOF OF COROLLARY 3.5

Note that from the tower property,

$$\mathbb{E}_P[\tilde{Y}|\phi(X)] = \mathbb{E}_P[\mathbb{E}_P[\tilde{Y}|X, \phi(X)]|\phi(X)] = \mathbb{E}_P[\mathbb{E}_P[\tilde{Y}|X] \mid \phi(X)] \tag{11}$$

From Proposition 3.3, for any w depending on ϕ P_X-a.s., the zero chosen weights bias is zero. Thus,

$$\begin{aligned}
|\text{Bias}_{P,Q}(w)| &\leq \left|\mathbb{E}_{P^w}[\mathbb{E}[\tilde{Y}|\phi(X)]] - \mathbb{E}_Q[\mathbb{E}[\tilde{Y}|\phi(X)]]\right| + |\text{Confounding bias}| \\
&\qquad \text{where } \mathbb{E}_P[\tilde{Y}|x] \in \mathcal{M} \text{ so from Equation 11, } \mathbb{E}_P[\tilde{Y}|\phi(x)] \in \phi(\mathcal{M}, P) \\
&\leq \text{IPM}_{\phi(\mathcal{M},P)}(P^w_{\phi(X)}, Q_{\phi(X)}) + |\text{Confounding bias}| \text{ by definition of an IPM} \\
&\leq \text{IPM}_{\phi(\mathcal{M},P)}(P^w_{\phi(X)}, Q_{\phi(X)}) + ||\tilde{Y}||_{L_2(P)} \cdot \text{BSE}_{P,Q}(\phi) \text{ from Equation 7}
\end{aligned}$$

F.4 PROOF OF COROLLARY 3.6

From Corollary 3.5, for any $\alpha \in \Lambda$,

$$\begin{aligned}
&\text{Bias}^2_{P^\alpha,Q^\alpha}(w^\alpha) \\
&\leq \left(\text{IPM}_{\phi^\alpha(\mathcal{M}^\alpha,P^\alpha)}(P^{\alpha,w^\alpha}_{\phi^\alpha(X)}, Q^\alpha_{\phi^\alpha(X)}) + ||\tilde{Y}||_{L_2(P^\alpha)} \cdot \text{BSE}_{P^\alpha,Q^\alpha}(\phi^\alpha)\right)^2.
\end{aligned}$$

Noting that $\forall a, b, (a+b)^2 \leq 2(a^2+b^2)$ and taking the expectation wrt $p_\Lambda(\alpha)$ gives

$$\frac{1}{2} \cdot \text{Bias}^2_{P^\Lambda,Q^\Lambda}(w^\Lambda) \leq \mathbb{E}_{p_\Lambda(\alpha)}\left[\text{IPM}^2_{\phi^\alpha(\mathcal{M}^\alpha,P^\alpha)}(P^{\alpha,w^\alpha}_{\phi^\alpha(X)}, Q^\alpha_{\phi^\alpha(X)})\right] + \mathbb{E}_{p_\Lambda(\alpha)}\left[||\tilde{Y}||^2_{L_2(P^\alpha)} \cdot \text{BSE}^2_{P^\alpha,Q^\alpha}(\phi^\alpha)\right]$$

Taking $||\tilde{Y}||^2_{L_2(P^\alpha)} \leq \sup_{\alpha\in\Lambda} ||\tilde{Y}||^2_{L_2(P^\alpha)}$ in the expectation with the BSE's leads to the result.

F.5 PROOF OF PROPOSITION B.1

First, let's note two useful properties :

- For any distribution R and random variable Z,

$$\forall x, \;\; \mathbb{E}_R[\mathbb{E}_R[Z|X] \mid \phi(X) = \phi(x)] = \mathbb{E}_R[Z|\phi(X) = \phi(x)]. \tag{12}$$

- For any distributions R and function f,

$$\Big(\exists g,\ \forall x\ R_X\text{-a.s.},\ f(x) = g(\phi(x))\Big) \iff \forall x\ R_X\text{-a.s.},\ f(x) = \mathbb{E}_R[f(X) \mid \phi(X) = \phi(x)]. \tag{13}$$

Proof of a), ATT case: Let $e(x) := P^{\text{data}}(A = 1|X = x)$

ϕ is a balancing score

$\iff \exists g,\ e(x) = g(\phi(x))\ \forall x\ P^{\text{data}}_X$-a.s. from Rosenbaum and Rubin [1983b][1]

$\iff \exists g,\ e(x) = g(\phi(x))\ \forall x\ P^{\text{data}}_{X|A=0}$-a.s. from the overlap assumption

$\iff \exists g,\ \dfrac{\mathrm{d}P^{\text{data}}_{X|A=1}}{\mathrm{d}P^{\text{data}}_{X|A=0}}(x) = g(\phi(x))\ \forall x\ P^{\text{data}}_{X|A=0}$-a.s. as $\dfrac{\mathrm{d}P^{\text{data}}_{X|A=1}}{\mathrm{d}P^{\text{data}}_{X|A=0}}(x)$ is a bijective function of $e(x)$ from Bayes' rule

$\iff \phi(x)$ is a generalized balancing score.

Proof of a), ATE case : we fix $a \in \mathcal{A}$ and work with the following definition [Imbens, 2000] 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 generalized 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,

ϕ is a balancing score

[2]While the original statement in Rosenbaum and Rubin [1983b] is not P^{data}_X-a.s., we note that it can be relaxed to P^{data}_X-a.s. as it pertains to the adjustment formula that involves an expectation wrt P^{data}_X

$\iff P^{\text{data}}(a|x) = P^{\text{data}}(a|\phi(x))\ \forall x\ P_X^{\text{data}}$-a.s.

$\iff P^{\text{data}}(a|x) = \mathbb{E}[P^{\text{data}}(a|X)|\phi(X) = \phi(x)]\ \forall x\ P_X^{\text{data}}$-a.s. using Equation 12 with $Z = 1_{\{A=a\}}$

$\iff \exists g_a,\ P^{\text{data}}(a|x) = g_a(\phi(x))\ \forall x\ P_X^{\text{data}}$-a.s. from Equation 13

$\iff \exists g_a,\ \dfrac{\mathrm{d}P_X^{\text{data}}}{\mathrm{d}P_{X|A=a}^{\text{data}}}(x) = g_a(\phi(x))\ \forall x\ P_X^{\text{data}}$-a.s.

where $\dfrac{\mathrm{d}P_X^{\text{data}}}{\mathrm{d}P_{X|A=a}^{\text{data}}}(x)$ is the true weights and is a bijective function of $P^{\text{data}}(a|x)$ from Bayes' rule

$\iff \exists g_a,\ \dfrac{\mathrm{d}P_X^{\text{data}}}{\mathrm{d}P_{X|A=a}^{\text{data}}}(x) = g_a(\phi(x))\ \forall x\ P_{X|A=a}^{\text{data}}$-a.s. from the overlap assumption

$\iff \phi(x)$ is a generalized balancing score.

Proof of b) : we slightly change the definition of deconfounding scores [D'Amour and Franks, 2021] 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]. To this aim, it is sufficient to show that, in Problem 2.1 applied to estimation of $\mathbb{E}[Y(a)]$, the confounding bias is equal to $\mathbb{E}[\mathbb{E}[Y|\phi(X), A = a]] - \mathbb{E}[Y(a)]$. From the original definition of the confounding bias, this simplifies further to $\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y|X, A = a]]$. This follows from the canonical unconfoundedness, overlap and SUTVA assumptions.

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,

$$\begin{aligned}\forall x\ P_{X|A=a}^{\text{data}}\text{-a.s.}, \mathbb{E}[Y|x, A = a] &:= \mathbb{E}[Y(a)|x] \\ &= \mathbb{E}[Y(a)|x, \phi(x)] \\ &= \mathbb{E}[Y(a)|\phi(x)] \text{ by application of the definition of a prognostic score,}\end{aligned}$$

so $\mathbb{E}[Y|x, A = a]$ is a function of $\phi(x) P_{X|A=a}^{\text{data}}$-a.s. from the overlap assumption, making the latter a generalized prognostic score.

Now assume that $\mathbb{E}[Y|x, A = a]$ itself is a prognostic score, that is $Y(a) \perp\!\!\!\perp X \mid \mathbb{E}[Y|X, A = a]$. Then, $p^{\text{data}}(Y(a)|x) = p^{\text{data}}(Y(a)|\mathbb{E}[Y|x, A = a])\ \forall x\ P_X^{\text{data}}$-a.s.. Let $\phi(X)$ be a generalized prognostic score. Then, there exists a function g_a such that $\mathbb{E}[Y|x, A = a] = g_a(\phi(x))\ \forall x\ P_{X|A=a}^{\text{data}}$-a.s., which can be replaced with P_X^{data}-a.s. from the overlap assumption. In particular, as $p^{\text{data}}(Y(a)|x)$ is already a function of $\mathbb{E}[Y|x, A = a]\ P_X^{\text{data}}$-a.s., it is also a function of $\phi(x)\ P_X^{\text{data}}$-a.s.. So there exists a function h_a such that $p^{\text{data}}(Y(a)|x) = h_a(\phi(x))\ \forall x\ P_X^{\text{data}}$-a.s.. In particular, by application of Equation 13, $p^{\text{data}}(Y(a)|x) = \mathbb{E}[p^{\text{data}}(Y(a)|X)|\phi(X) = \phi(x)]\ \forall x\ P_X^{\text{data}}$-a.s. and by application of Equation 12 to $Z = 1_{\{Y(a)=.\}}$, $p^{\text{data}}(Y(a)|x) = p^{\text{data}}(Y(a)|\phi(x))\ \forall x\ P_X^{\text{data}}$-a.s.. 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. We also use this notation for e) and f).

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

$$\begin{aligned}\forall x\ P_X^{\text{data}}\text{-a.s.}, \mathbb{E}_P[\tilde{Y}|x] &= \text{CATE}(x) \text{ (under the transportability assumption)} \\ &= \mathbb{E}[Y(1) - Y(0)|x] \\ &= \mathbb{E}[Y(1) - Y(0)|x_{-\mathcal{I}}, x_{\mathcal{I}}] \\ &= \mathbb{E}[Y(1) - Y(0)|x_{\mathcal{I}}] \text{ by definition of a heterogeneity set}\end{aligned}$$

where P_X^{data}-a.s. is equivalent to $P_{X|S=1}^{\text{data}}$-a.s. under the support inclusion (i.e. overlap) assumption, so $\mathbb{E}_P[\tilde{Y}|x]$ is a function of $x_{\mathcal{I}}\ P_{X|S=1}^{\text{data}}$-a.s., making the latter a generalized prognostic score.

Proof of e) : If $x_{\mathcal{I}}$ is a sampling set, that is $Y(1), Y(0), S \perp\!\!\!\perp X_{-\mathcal{I}}|X_{\mathcal{I}}$, then

$$\begin{aligned}\forall x P_X^{\text{data}}\text{-a.s.}, \frac{\mathrm{d}P_{X|S=0}^{\text{data}}}{\mathrm{d}P_{X|S=1}^{\text{data}}}(x) &= \frac{p^{\text{data}}(x|S=0)}{p^{\text{data}}(x|S=1)} \\ &= \frac{P^{\text{data}}(S=1)}{P^{\text{data}}(S=0)} \frac{p^{\text{data}}(S=0|x)}{p^{\text{data}}(S=1|x)} \text{ from Bayes' rule}\end{aligned}$$

$$
\begin{aligned}
&= \frac{P^{\text{data}}(S=1)}{P^{\text{data}}(S=0)} \frac{p^{\text{data}}(S=0|x_{\mathcal{I}}, x_{-\mathcal{I}})}{p^{\text{data}}(S=1|x_{\mathcal{I}}, x_{-\mathcal{I}})} \\
&= \frac{P^{\text{data}}(S=1)}{P^{\text{data}}(S=0)} \frac{p^{\text{data}}(S=0|x_{\mathcal{I}})}{p^{\text{data}}(S=1|x_{\mathcal{I}})} \text{ as } x_{\mathcal{I}} \text{ is a sampling set} \\
&= \frac{p^{\text{data}}(x_{\mathcal{I}}|S=0)}{p^{\text{data}}(x_{\mathcal{I}}|S=1)} \\
&= \frac{\mathrm{d}P^{\text{data}}_{X_{\mathcal{I}}|S=0}}{\mathrm{d}P^{\text{data}}_{X_{\mathcal{I}}|S=1}}(x_{\mathcal{I}})
\end{aligned}
$$

thus $\frac{\mathrm{d}P^{\text{data}}_{X|S=0}}{\mathrm{d}P^{\text{data}}_{X|S=1}}(x)$ depends on $x_{\mathcal{I}}$ $\forall x$ P^{data}_X-a.s., which is equivalent to $P^{\text{data}}_{X|S=1}$-a.s. under the support inclusion (i.e. overlap) assumption, and the last two lines illustrate the fact that, in this case, the density ratio wrt X is equal to that wrt the representation a.s. under the source distribution.

Proof of f) : If $x_{\mathcal{I}}$ is a separating set, that is $Y(1) - Y(0) \perp\!\!\!\perp S|X_{\mathcal{I}}$,

$$
\begin{aligned}
P^{\text{data}}\text{-a.s., } \mathbb{E}_P[\tilde{Y}|X_{\mathcal{I}}] &= \mathbb{E}_P[\mathbb{E}_P[\tilde{Y}|X]|X_{\mathcal{I}}] \text{ from Equation 12} \\
&= \mathbb{E}_P[\text{CATE}(X)|X_{\mathcal{I}}] \\
&= \mathbb{E}[\text{CATE}(X)|X_{\mathcal{I}}, S=1] \\
&= \mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X, S=1]|X_{\mathcal{I}}, S=1] \text{ under the transportability assumption} \\
&= \mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X, X_{\mathcal{I}}, S=1]|X_{\mathcal{I}}, S=1] \\
&= \mathbb{E}[Y(1) - Y(0)|X_{\mathcal{I}}, S=1] \text{ under the tower property} \\
&= \mathbb{E}[Y(1) - Y(0)|X_{\mathcal{I}}, S=0] \text{ by definition of a separating set.}
\end{aligned}
$$

where P^{data}-a.s. implies $P^{\text{data}}(.|S=0)-$ a.s., thus

$$
\begin{aligned}
&\text{Confounding bias of } x_{\mathcal{I}} \\
&= \mathbb{E}_Q\Big[\mathbb{E}_P[\tilde{Y}|X_{\mathcal{I}}] - \mathbb{E}_P[\tilde{Y}|X]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[Y(1) - Y(0)|X_{\mathcal{I}}, S=0] - \mathbb{E}_P[\tilde{Y}|X]\Big|S=0\Big] \text{ from the above} \\
&= \mathbb{E}\Big[\mathbb{E}[Y(1) - Y(0)|X_{\mathcal{I}}, S=0] - \text{CATE}(X)\Big|S=0\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[Y(1) - Y(0)|X_{\mathcal{I}}, S=0] - \mathbb{E}[Y(1) - Y(0)|X, S=0]\Big|S=0\Big] \text{ from the transportability assumption} \\
&= \mathbb{E}\Big[\mathbb{E}[Y(1) - Y(0)|X_{\mathcal{I}}, S=0]\Big|S=0\Big] - \mathbb{E}\Big[\mathbb{E}[Y(1) - Y(0)|X, S=0]\Big|S=0\Big] \\
&= \mathbb{E}[Y(1) - Y(0)|S=0] - \mathbb{E}[Y(1) - Y(0)|S=0] \text{ under the tower property} \\
&= 0,
\end{aligned}
$$

so $x_{\mathcal{I}}$ is a generalized deconfounding score.

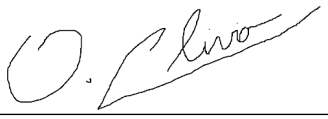
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Towards Representation Learning for Weighting Problems in Design-Based Causal Inference
Publication Status	X Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Clivio, O., Feller, A. & Holmes, C.. (2024). Towards Representation Learning for Weighting Problems in Design-Based Causal Inference. *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, in *Proceedings of Machine Learning Research* 244:856-880

Student Confirmation

Student Name:	Oscar Clivio		
Contribution to the Paper	As the first author, I found the idea, derived all theoretical results, wrote the codebase, conducted all experiments, wrote the first iteration of the paper and took the lead in the writing. Co-authors gave very helpful pointers that notably helped shape the idea, and assisted in checking theoretical and numerical results as well as writing.		
Signature		Date	2025/05/19

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Chris Holmes		
Supervisor comments I agree with the student's comments		
Signature 	Date	2025/06/26

This completed form should be included in the thesis, at the end of the relevant chapter.

4

Deconfounding Scores and Representation Learning for Causal Effect Estimation with Weak Overlap

Deconfounding Scores and Representation Learning for Causal Effect Estimation with Weak Overlap

Oscar Clivio* University of Oxford

Alexander D'Amour* Google DeepMind

Alexander Franks* University of California, Santa Barbara

David Bruns-Smith Stanford University

Chris Holmes University of Oxford, Ellison Institute of Technology

Avi Feller University of California, Berkeley

Abstract

Overlap, also known as positivity, is a key condition for causal treatment effect estimation. Many popular estimators suffer from high variance and become brittle when features differ strongly across treatment groups. This is especially challenging in high dimensions: the curse of dimensionality can make overlap implausible. To address this, we propose a class of feature representations called *deconfounding scores*, which preserve both identification and the target of estimation; the classical propensity and prognostic scores are two special cases. We characterize the problem of finding a representation with better overlap as minimizing an *overlap divergence* under a deconfounding score constraint. We then derive closed-form expressions for a class of deconfounding scores under a broad family of generalized linear models with Gaussian features and show that prognostic scores are overlap-optimal within this class. We conduct extensive experiments to assess this behavior empirically.

1 INTRODUCTION

In observational causal inference, researchers typically emphasize the key role of the ignorability assumption, which assumes that observed features contain all confounders of the relationship between the treatment and the outcome. Equally important—yet less often the focus—is *overlap* (or *positivity*): the distributions of features in the treated and control groups must have common support. Both assumptions are critical for nonparametric identification, enabling flexible confounder adjustment for unbiased estimation of causal estimands.

Even when overlap holds, features can still differ substantially between treatment and control groups, degrading both theoretical guarantees and practical performance of importance weighting and doubly robust estimators (Rothe, 2017; Hong et al., 2020). The problem is compounded in modern settings where adjusting for a large number of features is necessary for ignorability to be plausible (D'Amour et al., 2021). *Representation learning* offers a natural strategy: learn a low-dimensional mapping of the features that preserves unbiased estimation while potentially improving overlap. Two widely-used representations are *balancing scores* (including *propensity scores*) (Rosenbaum and Rubin, 1983) and *prognostic scores* (Hansen, 2008), which capture all feature information predictive of the treatment assignment and the outcome, respectively.

In this paper, we introduce *deconfounding scores*: the complete class of representations that preserve unbiased estimation of the target estimand under ignorability. Balancing and prognostic scores are special cases. Our main contributions are:

- We characterize the class of deconfounding scores and specify the conditions necessary to compute them; this class naturally falls on a continuum between prognostic and balancing scores.
- We introduce an *overlap divergence* that quantifies the lack of overlap with respect to a represen-

*Equal contribution.

Preprint.

tation and show that it controls the semiparametric efficiency bound.

- We establish that (i) representations always improve overlap—as measured by overlap divergence—compared to original covariates, and (ii) those that are less predictive of the treatment assignment exhibit better overlap.
- In the canonical setting with Gaussian features and generalized linear models for the outcome and treatment assignment, we analytically characterize a family of deconfounding scores as lying on a hyperbola whose endpoints correspond to balancing and prognostic scores. For this family, **prognostic scores minimize the overlap divergence while balancing scores maximize it**.
- In simulations, a correctly-specified prognostic score improves performance over raw covariates and most other deconfounding scores. On semi-synthetic datasets, there is always at least one deconfounding score improving over covariates, with the prognostic score yielding the most improvements over covariates.

1.1 Related Work

Inference with Poor Overlap. When the overlap assumption is not satisfied, the conditional average treatment effect (CATE) is no longer nonparametrically identified on points or regions without overlap (Hernán and Robins, 2010). This can be mitigated with the help of assumptions on the outcome or propensity models, by extrapolating the CATE from regions with overlap to those without overlap (Petersen et al., 2012; Nethery et al., 2019), or through *partial identification*, that is, bounds on the CATE on regions without overlap (Manski, 1990; Lee and Weidner, 2021; Khan et al., 2024). An alternative is the use of balancing weights to estimate population-wide treatment effects (Kallus, 2020; Bruns-Smith and Feller, 2022). Even when overlap holds, stricter versions are required for root-n confidence intervals (Khan and Tamer, 2010; Rothe, 2017; Hong et al., 2020) but are again often unrealistic in high dimensions (D'Amour et al., 2021). To tackle this, previous approaches have typically focused on changing the treatment effect estimand to a population with better overlap (Crump et al., 2009; Matsouaka and Zhou, 2020), trimming extreme estimated propensity scores (Stürmer et al., 2010; Chaudhuri and Hill, 2013; Mehrabi and Wager, 2024), or adjusting on representations with better overlap, which we detail next.

Learning Representations as Adjustment Sets. There is a substantial literature on adjusting for representations, i.e., deterministic mappings of covariates (Leacy and Stuart, 2014; Lee and Lee, 2022). Such representations should preserve confounding information; extreme examples are balancing and prognostic scores, with exact definitions in Section 3.1. Those representations are generally not given and need to be estimated; classical approaches learn scalar representations using linear/logistic regression (Hansen, 2008; Leacy and Stuart, 2014) or focus on subsets of covariates (Schneeweiss et al., 2009) while more recent approaches leverage the compositional nature of neural networks to extract multivariate representations (Shalit et al., 2017; Chernozhukov et al., 2022b; Clivio et al., 2022). Errors in estimation can lead to loss of confounding information; such loss is typically either not checked or assumed not to exist (Shalit et al., 2017; Melnychuk et al., 2025), while more recent approaches attempt to quantify or minimize this confounding loss (Johansson et al., 2019; Melnychuk et al., 2023; Clivio et al., 2024). In contrast, *we provide representations with no loss of confounding information*, generalizing balancing or prognostic scores.

Learning Representations to Improve Overlap. Researchers typically adjust for prognostic scores to reduce asymptotic variance (Austin et al., 2007; Schuler et al., 2021); this approach has been referred to as *collaborative* in the Targeted Machine Learning Estimation (TMLE) literature (Benkeser et al., 2020; Rudolph et al., 2023). Typically, overlap in prognostic scores between treated and control groups is considered less stringent than overlap in original features (Luo et al., 2017; D'Amour et al., 2020; Wu and Fukumizu, 2021). In contrast, balancing scores are not used to improve overlap; instead, overlap is improved by removing non-confounding variables that predict the treatment assignment (Rubin, 1997; Wooldridge, 2016). Other approaches, inspired by domain adaptation, minimize an objective that balances a treatment effect regression error against measures of poor overlap, including distributional distances (Shalit et al., 2017; Johansson et al., 2022), support discrepancy measures (Johansson et al., 2019), and conditional outcome posterior variances (Zhang et al., 2020). However, to the best of our knowledge, none of these measures directly connects poor overlap to inferential challenges such as estimator variance. In contrast, our overlap divergence explicitly links the degree of overlap with respect to a representation to the asymptotic variance of estimators adjusting on that representation. Some of these approaches further incorporate inverse propensity weights in the objective, both in the regression error and in the measure of poor overlap (Assaad et al., 2021; Johansson et al., 2022). While this approach can improve treatment effect estimation,

it is orthogonal to finding representations with better overlap. We further compare our approach to these lines of work in Appendix B.1.

2 PRELIMINARIES

Let $(X_i, T_i, Y_i) \overset{\text{i.i.d.}}{\sim} P$ be i.i.d. samples of covariates X, a binary treatment T, and an outcome Y. Denote $P^0 := P(.|T=0)$, $P^1 := P(.|T=1)$, $\pi_1 := P(T=1)$. We assume $0 < \pi_1 < 1$. Denote for any random variable Z and distribution R, R_Z the law of Z in R, $\mathbb{E}_R$ the expectation wrt R, $m_0(Z) := \mathbb{E}_{P^0}[Y|Z]$ and $m_1(Z) := \mathbb{E}_{P^1}[Y|Z]$ the outcome models wrt Z, $\Delta m(Z) = m_1(Z) - m_0(Z)$ their difference, and $e(Z) := p(T=1|Z)$ which is called the propensity score wrt Z. Note that, when relevant, the superscript of a distribution denotes the treatment indicator and its subscript is a random variable.

We focus on estimating the *Average Treatment effect on the Treated* (ATT), $\tau = \mathbb{E}_{P^1}[Y(1) - Y(0)]$, where $Y(1)$ and $Y(0)$ denote the potential outcomes under treatment and control, respectively.[1] Throughout, we assume unconfoundedness, (one-sided) overlap wrt X, and technical assumptions about the outcome Y and the density ratio between control and treatment distributions of X.

Assumption 2.1 *(Unconfoundedness) All potential confounders of the relationship between treatment and potential outcomes are included in covariates X:*

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

Assumption 2.2 *(One-sided overlap) P^1_X is absolutely continuous wrt P^0_X, or equivalently, $e(X) < 1$ P-almost surely.*

Assumption 2.3 *(Square-integrability of $\frac{dP^1_X}{dP^0_X}(X)$) The density ratio $\frac{dP^1_X}{dP^0_X}(X)$ between control and treatment distributions of X is square-integrable in P^0:*

$$\mathbb{E}_{P^0}\left[\left(\frac{dP^1_X}{dP^0_X}(X)\right)^2\right] < \infty.$$

Assumption 2.4 *(Square-integrability of Y) The observed outcome Y is square-integrable in P:*

$$\mathbb{E}_P[Y^2] < \infty.$$

Assumption 2.2 is a generalization of the standard overlap assumption, which states that every unit has some non-zero chance of receiving either treatment condition (or, equivalently, that there are no values of the covariates such that units with these values are either all in treatment or all in control). For this paper, we focus on the one-sided version of this assumption, where every unit has some non-zero probability of having been assigned to control, since our target estimand is the ATT that ignores units which cannot receive the treatment (in contrast, the standard overlap assumption applies when the target estimand is the standard Average Treatment Effect (ATE)). A stricter version of Assumption 2.2 is Assumption 2.3, which is the minimal assumption on overlap ensuring that many expectations in our derivations are well-defined. This is still weaker than the strict overlap assumption of D'Amour et al. (2021) which uniformly bounds the density ratio. Assumption 2.4 is a technical assumption that prohibits pathological outcome distributions; for example, all bounded outcomes satisfy this assumption. Under these assumptions, τ can be identified by adjusting for X, that is, $\tau = \tau_X$ where, for any Z,

$$\tau_Z := \mathbb{E}_{P^1}[Y] - \mathbb{E}_{P^1}[m_0(Z)]. \tag{1}$$

We will consider the properties of statistical estimands τ_Z that adjust for variables Z other than X. The minimal possible asymptotic variance of regular and asymptotically linear (RAL) estimators of τ_Z from samples of $P(Z,T,Y)$ is the *semiparametric efficiency bound* V^Z_{eff} (Tsiatis, 2006) given as (Hahn, 1998)

$$\begin{aligned} V^Z_{\text{eff}} =& \mathbb{E}_P\left[\frac{e(Z)\text{Var}_{P^1}(Y|Z) + (\Delta m(Z) - \tau_Z)^2 e(Z)}{\pi_1^2}\right] \\ &+ \mathbb{E}_{P^0}\left[\frac{\text{Var}_{P^0}(Y|Z)}{1-\pi_1}\left(\frac{dP^1_Z}{dP^0_Z}(Z)\right)^2\right]. \end{aligned}$$

We can see that V^Z_{eff} depends on the magnitude of the density ratio $\frac{dP^1_Z}{dP^0_Z}$; this magnitude describes the strength of overlap between the distributions of Z in the treatment and control groups. When $\frac{dP^1_Z}{dP^0_Z}$ takes large values, even under one-sided overlap wrt Z, RAL estimators of τ_Z have high asymptotic variance. Thus, our goal will be to build representations $\phi(X)$ that improve overlap compared to X while ensuring $\tau = \tau_{\phi(X)}$.

3 DECONFOUNDING SCORES AND OVERLAP DIVERGENCE

3.1 Deconfounding Scores

We now introduce *deconfounding scores*. These are defined as representations $\phi(X)$ such that τ can be identified by adjusting only for $\phi(X)$, that is $\tau = \tau_{\phi(X)}$.

[1] Our results can be readily extended to general covariate shift, full population average treatment effect estimation or transportability (Clivio et al., 2024).

We can therefore view deconfounding scores as preserving the confounding information in X.

A key property is that deconfounding scores can be characterized as representations that introduce zero "confounding bias", expressible in terms of an observable conditional covariance (all proofs in Appendix A).

Lemma 3.1 *For any ϕ, the confounding bias equals*

$$\tau_{\phi(X)} - \tau = \mathbb{E}_{P^0}\left[Cov_{P^0}\left(m_0(X), \frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right)\right].$$

Setting the conditional covariance in Lemma 3.1 to zero serves as a constraint that deconfounding scores must satisfy. It is straightforward to verify that both the propensity score $e(X)$ (the density ratio is a measurable function of $e(X)$) and the control outcome model $m_0(X)$ are deconfounding scores. More generally, any $\phi(X)$ such that $m_0(X)$ is a measurable function of $\phi(X)$ (a *prognostic score*, generalizing Hansen (2008)) or $e(X)$ is a measurable function of $\phi(X)$ (a *balancing score* as introduced in Rosenbaum and Rubin (1983)) is a deconfounding score. Crucially, the constraint also defines a continuum of representations between these two extremes, which we explore in this paper. We further discuss this result in Appendix B.2.

3.2 Overlap Divergence

While deconfounding scores all yield the same ATT estimand τ, they can have different overlap properties. Here, we introduce the *overlap divergence*, which we use to measure the degree of overlap wrt a representation. As we will show shortly, it is closely connected to the semiparametric efficiency bound. The overlap divergence of a random variable Z is defined as

$$\mathcal{O}(Z) := \mathbb{E}_{P^0}\left[\left(\frac{dP_Z^1}{dP_Z^0}(Z)\right)^2\right] = \chi^2(P_Z^1||P_Z^0) + 1,$$

where $\chi^2(P_Z^1||P_Z^0) = \mathbb{E}_{P^0}\left[\left(\frac{dP_Z^1}{dP_Z^0}(Z) - 1\right)^2\right]$ is the χ^2-divergence between P_Z^0 and P_Z^1. $\mathcal{O}(\phi(X))$ quantifies the lack of overlap wrt $\phi(X)$, as it reflects the amplitude of $\frac{dP^1_{\phi(X)}}{dP^0_{\phi(X)}}$. On the one hand, $\mathcal{O}(\phi(X))$ is minimized (and equal to 1) if and only if $P^0_{\phi(X)} = P^1_{\phi(X)}$, which represents perfect overlap; however, note that such representations $\phi(X)$ are typically not deconfounding scores, e.g., when $\phi(X)$ is constant. On the other hand, when the one-sided overlap assumption is not satisfied wrt $\phi(X)$, that is $P^1_{\phi(X)}$ is not absolutely continuous wrt $P^0_{\phi(X)}$, we have $\mathcal{O}(\phi(X)) = \infty$. One-sided overlap wrt $\phi(X)$ does not guarantee that $\mathcal{O}(\phi(X)) < \infty$; for example, $\mathcal{O}(\phi(X))$ can be infinite if $e(\phi(X))$ approaches 1. Note that Assumption 2.3 can equivalently be written as $\mathcal{O}(X) < \infty$.

We now justify our overlap divergence as controlling the semiparametric efficiency bound, as we show next.

Lemma 3.2 *For any representation $\phi(X)$, we have:*

1. *If $Var_{P^0}(Y|X) \geq \sigma^2$ for some $\sigma > 0$, then $V_{eff}^{\phi(X)} \geq \frac{\sigma^2}{1-\pi_1} \cdot \mathcal{O}(\phi(X))$.*
2. *If Y is bounded by some constant $Y_{max} > 0$ then $V_{eff}^{\phi(X)} \leq \frac{5 \cdot Y_{max}^2}{\pi_1} + \frac{Y_{max}^2}{1-\pi_1}\mathcal{O}(\phi(X))$.*

Together, these two bounds show that the overlap divergence is tightly linked to the efficiency bound: Item 1 implies that reducing the overlap divergence is *necessary* for reducing the efficiency bound, while Item 2 implies that it may be *sufficient*. We further discuss this result in Appendix B.3.

4 OPTIMIZING THE OVERLAP DIVERGENCE

We now aim to minimize $\mathcal{O}(\phi(X))$ subject to the constraint that $\phi(X)$ is a deconfounding score. We first establish general properties in the nonparametric case, then show that a prognostic score solves this optimization problem in a Gaussian design setting.

4.1 Nonparametric Case: Representations Always Improve Overlap

We show that representations always lead to better overlap than original covariates and formalize the longstanding intuition that information predictive only of treatment should be excluded.

Lemma 4.1 *The improvement of overlap divergence induced by a representation $\phi(X)$ equals*

$$\mathcal{O}(X) - \mathcal{O}(\phi(X)) = \mathbb{E}_{P^0}\left[Var_{P^0}\left(\frac{dP_X^1}{dP_X^0}(X) \;\middle|\; \phi(X)\right)\right] \geq 0$$

and it upper-bounds the absolute confounding bias as

$$|\tau_{\phi(X)} - \tau| \leq \sqrt{\mathbb{E}_{P^0}\left[Var_{P^0}\left(m_0(X) \mid \phi(X)\right)\right]} \times \sqrt{\mathcal{O}(X) - \mathcal{O}(\phi(X))}.$$

The first part of Lemma 4.1 shows that a representation *always* improves overlap relative to the original covariates. Moreover, the improvement equals a measure of how poorly $\phi(X)$ predicts the density ratio $\frac{dP_X^1}{dP_X^0}(X)$—and thus the propensity score $e(X)$—termed

the *balancing score error* in Clivio et al. (2024). This confirms that variables predictive of treatment but not outcome should be excluded to reduce variance, which is a common intuition throughout the literature and has been highlighted in, e.g., Rubin (1997), Brookhart et al. (2006), Wooldridge (2016) and Colnet et al. (2024). To the best of our knowledge, this is the first mathematical proof of this intuition.

Note that when $\phi(X)$ is not a balancing score, its propensity score $e(\phi(X))$ will differ from the original propensity score $e(X)$. This is desirable, however: as long as $\phi(X)$ is a deconfounding score, unbiased estimation is preserved, and Lemma 4.1 shows that the overlap divergence is strictly reduced.

Finally, the upper bound in Lemma 4.1 suggests that a representation achieving both zero confounding bias and minimal overlap divergence should be a prognostic score. In the next section, we confirm this in a tractable analytical setting.

4.2 Gaussian Design: Prognostic Scores Optimize Overlap

In this section, we explore the properties of deconfounding scores in a simple setting where we can obtain analytical expressions for a family of deconfounding scores. Specifically, we consider a setting where covariates are standard centered Gaussian variables, the treatment assignment and outcome model are generalized linear models (GLMs), and representations are linear. We study two variants. First, consider the case where covariates are Gaussian in P^0.

Assumption 4.2 *$P_X^0 = \mathcal{N}(0, I_d)$, $m_0(x) = m(\alpha' x)$, $\frac{dP_X^1}{dP_X^0}(x) = \frac{h(\beta' x)}{\mathbb{E}_{Z\sim\mathcal{N}(0,1)}[h(Z)]}$, $\phi(x) = \phi_\gamma(x) := \gamma' x$ for α, β, γ unit vectors in $\mathbb{R}^d$ with $\alpha'\beta \in (-1, 1)$, h, m real functions with $h \geq 0$ and $0 < \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[h(Z)] < \infty$.*

Second, we consider the case where covariates are Gaussian in P. This variant is motivated by the fact that it is often impractical to make assumptions on the distribution of covariates in either treatment group and on the density ratio between the treatment groups — instead preferring assumptions on covariates at the whole population level and on the propensity score.

Assumption 4.3 *$P_X = \mathcal{N}(0, I_d)$, $m_0(x) = m(\alpha' x)$, $e(x) = h(\beta' x)$, $\phi(x) = \phi_\gamma(x) := \gamma' x$ for α, β, γ unit vectors in $\mathbb{R}^d$ with $\alpha'\beta \in (-1, 1)$, h, m real functions with $0 \leq h(.) < 1$.*

4.2.1 Analytical Characterization of Deconfounding Scores

We now show that a family of deconfounding scores can be computed in closed form in both settings.

Theorem 4.4 *If either Assumption 4.2 or Assumption 4.3 holds then $\phi_\gamma(X)$ is a deconfounding score for any γ of the form $\gamma = w_1 \frac{\alpha+\beta}{\sqrt{2+2\alpha'\beta}} + w_2 \frac{\alpha-\beta}{\sqrt{2-2\alpha'\beta}} + n$ where $(1+\alpha'\beta)w_1^2 - (1-\alpha'\beta)w_2^2 = 2\alpha'\beta$, $w_1^2 + w_2^2 \leq 1$, n has norm $\sqrt{1 - w_1^2 - w_2^2}$ and is in Null(Span(α, β)). We refer to the set of such γ's as $\mathcal{D}_{\alpha,\beta}$.*

Note that $\mathcal{D}_{\alpha,\beta}$ does not depend on h and m. To interpret this result, we note that the coordinates (w_1, w_2) that yield valid unit-length values of γ trace out two opposite segments of a hyperbola that lies on the subspace spanned by the prognostic score and propensity score coefficient vectors α and β. One segment has endpoints α and β (when $\alpha'\beta \geq 0$) or $-\beta$ (when $\alpha'\beta < 0$), and the other has endpoints $-\alpha$ and $-\beta$ (when $\alpha'\beta \geq 0$) or β (when $\alpha'\beta < 0$). Figure 1 shows the former segment for different values of $\alpha'\beta$.

When $\alpha'\beta \geq 0$, w_2 controls the position of the projection of γ onto Span(α, β) on either branch of the hyperbola. On the branch with endpoints α and β, when we move w_2 along its valid range $\left[-\sqrt{\frac{1-\alpha'\beta}{2}}, \sqrt{\frac{1-\alpha'\beta}{2}}\right]$, setting w_2 to its maximal value implies that $\gamma = \alpha$ so that $\phi_\gamma(X)$ is a prognostic score, whereas setting w_2 to its minimal value implies that $\gamma = \beta$ so that $\phi_\gamma(X)$ is a balancing score. $w_2 = 0$ implies that γ is equiangular to α and β, that is $\alpha'\gamma = \beta'\gamma$. An analogous description can be made for the branch with endpoints $-\alpha$ and $-\beta$. For points on the interior of w_2's range, there is an equivalence class of γ's: the projection of γ onto Span(α, β) is strongly constrained, but the orthogonal component of γ is only constrained to make sure that γ has norm 1. We interpret w_2 as a scalar parameter that controls the similarity of the deconfounding score to the balancing and prognostic scores.

Note that w_1 can be analogously interpreted when $\alpha'\beta \leq 0$. In this case, the segment with prognostic score endpoint α has $-\beta$ as a propensity score endpoint; the other segment has prognostic score and propensity score endpoints $-\alpha$ and β, respectively. Equivalently, we can enforce $\alpha'\beta > 0$ and replace β with $-\beta$ when $\alpha'\beta < 0$; we justify this in Appendix B.4.

4.2.2 Optimality of Prognostic Scores

Given this family of deconfounding scores, we might suspect from Lemma 4.1 that its prognostic scores, which have no explicit dependence on the propensity

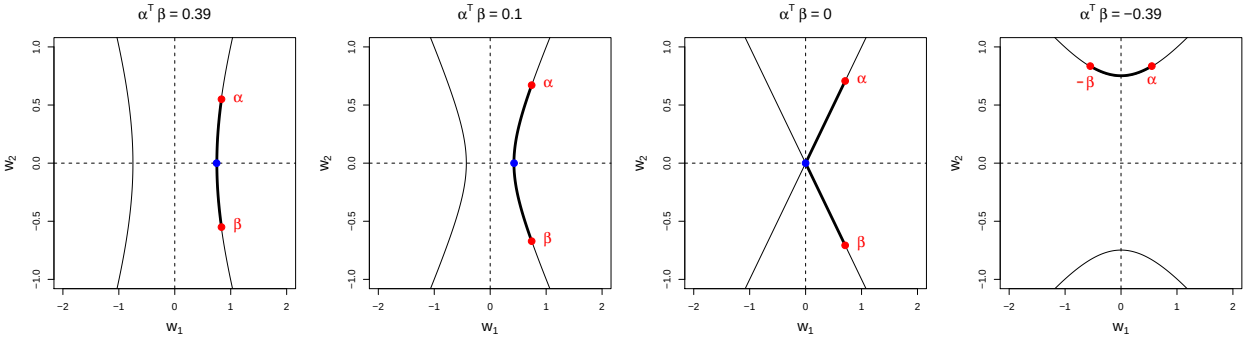


Figure 1: The projection of γ onto the space spanned by α and β lies on a segment of a hyperbola (bold black line) whose endpoints correspond to $\gamma = \alpha$ and $\gamma = \beta$ (when $\alpha'\beta \geq 0$) or to $\gamma = \alpha$ and $\gamma = -\beta$ (when $\alpha'\beta < 0$), shown here, and the opposite segment, not shown here. The orientation of the hyperbola and the endpoints depends on $\alpha'\beta$.

score or the density ratio, would be overlap-optimal. Here, we show that this is in fact the case.

Theorem 4.5 *Defining, for any integer K and for fixed $C > 1$, $0 \leq \lambda < \frac{1}{4}$, $0 \leq \lambda' < \frac{1-4\lambda}{6}$,*

$$\mathcal{H}^K_{C,\lambda} := \{h \mid h \text{ is } K \text{ times differentiable}, \\ \forall k = 0, .., K,\ \forall z,\ |h^{(k)}(z)| \leq Ce^{\lambda z^2}\},$$

$$\mathcal{H}'^K_{C,\lambda,\lambda'} := \{h \mid h \text{ is } K \text{ times differentiable}, \\ \forall z,\ 1 - h(z) \geq \frac{e^{-\lambda' z^2}}{C},\ h(z) \geq 0, \\ \forall k = 1, .., K,\ \forall z,\ |h^{(k)}(z)| \leq Ce^{\lambda z^2}\},$$

we have:

1. *If either (a) Assumption 4.2 holds with $h \in \mathcal{H}^2_{C,\lambda}$, or (b) Assumption 4.3 holds with $h \in \mathcal{H}'^2_{C,\lambda,\lambda'}$, then*
   - *(A) $\mathcal{O}(\phi_\gamma(X))$ is non-decreasing in $|\beta'\gamma|$.*
   - *(B) If $\alpha'\beta \neq 0$, $\mathcal{O}(\phi_\gamma(X))$ is non-decreasing when moving from α to β (when $\alpha'\beta > 0$) or to $-\beta$ (when $\alpha'\beta < 0$), and when moving from $-\alpha$ to $-\beta$ (when $\alpha'\beta > 0$) or to β (when $\alpha'\beta < 0$), on the corresponding portion of $\mathcal{D}_{\alpha,\beta}$; notably, $\gamma = \alpha$ and $\gamma = -\alpha$ are global minimizers of $\mathcal{O}(\phi_\gamma(X))$ on $\mathcal{D}_{\alpha,\beta}$.*
   - *(C) If $\alpha'\beta = 0$, the γ's whose projection onto $\mathrm{Span}(\alpha, \beta)$ belongs to $[-\alpha, \alpha]$ are global optimizers of $\mathcal{O}(\phi_\gamma(X))$ on $\mathcal{D}_{\alpha,\beta}$.*
2. *If either (a) Assumption 4.2 holds with (i) $h \in \mathcal{H}^{K+1}_{C,\lambda}$ for some integer $K \geq 2$ such that $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[h^{(K)}(Z)\right] \neq 0$, or (ii) $h(z) = 1_{\{z \leq z_0\}}$ for some $z_0 \in \mathbb{R}$, or (iii) $h(z) = ReLU(z)$, or if (b) Assumption 4.3 holds with $h \in \mathcal{H}'^{K+1}_{C,\lambda,\lambda'}$ for some integer $K \geq 2$ such that $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[h^{(K)}(Z)\right] \neq 0$, then we obtain the same (A), (B) and (C) as in 1. but where "non-decreasing" is replaced with "increasing" and "global minimizers" with "the only global minimizers".*

Under the assumptions of Theorem 4.5, the overlap divergence of $\phi_\gamma(X)$ is non-decreasing (or strictly increasing) in $|\beta'\gamma|$, which measures the strength of the association between the treatment assignment (parameterized by β) and the representation (parameterized by γ), consistent with Lemma 4.1. In the generic case ($\alpha'\beta \neq 0$), overlap improves monotonically as we move along either segment of the hyperbola from the balancing score toward the prognostic score, strengthening Lemma 4.1. In the degenerate case ($\alpha'\beta = 0$), half of either segment is optimal, but the prognostic scores always remain among the optimizers. We emphasize the key conclusion: *among deconfounding scores in $\mathcal{D}_{\alpha,\beta}$, the prognostic scores yield the best overlap.* We further discuss assumptions in Appendix B.5.

5 EXPERIMENTS

We now assess how our analytical results translate to finite-sample estimation. We evaluate ATT analogues of the outcome regression (Hahn, 1998), IPW (Horvitz and Thompson, 1952), and AIPW (Robins et al., 1994) estimators, replacing covariates X with linear deconfounding scores $\phi_\gamma(X)$ expressed as in the $\mathcal{D}_{\alpha,\beta}$ set of Theorem 4.4. We give further details in Appendix C. The code to reproduce experiments is available at `https://github.com/oscarclivio/deconfounding_scores_paper`.

5.1 Estimators and Inference

We consider canonical ATT estimators: (i) the outcome regression estimator, obtained by plugging the estimated outcome model into Equation 1 ("Regr"), (ii) the IPW estimator ("IPW"), and (iii) the AIPW estimator ("AIPW"). We compare these to the analogous estimators where input features are deconfounding scores $\phi_\gamma(X)$. They are denoted as "Method-γ", where "Method" refers to the base methods "Regr", "IPW", "AIPW" and γ is the coefficient vector of the deconfounding score passed to the base method. We estimate the outcome and propensity models using either LASSO regression or Ridge regression. Regularization parameters for models fit to original features are selected via cross-validation using the `glmnet` R package (Friedman et al., 2010; R Core Team, 2024). We do not use regularization when estimating models wrt one-dimensional deconfounding scores.

To estimate deconfounding scores, we use coefficient vectors $\hat{\alpha} = \hat{\alpha}^1$ and $\hat{\beta} = \text{sign}(\hat{\alpha}^{1\prime}\hat{\beta}^1)\hat{\beta}^1$ where $\hat{\alpha}^1$ and $\hat{\beta}^1$ are the normalized coefficient vectors obtained through the above LASSO or Ridge regression with respect to original features. To ensure $\hat{\alpha}'\hat{\beta} \geq 0$, we replace $\hat{\beta}^1$ with $-\hat{\beta}^1$ whenever $\hat{\alpha}^{1\prime}\hat{\beta}^1 < 0$. These are used as plug-ins in the set $\mathcal{D}_{\hat{\alpha},\hat{\beta}}$ given in Theorem 4.4 to yield estimated linear deconfounding scores $\phi_{\hat{\gamma}}(X)$; we sample one orthogonal component $\hat{n}$ with appropriate normalization in $\text{Null}(\text{Span}(\hat{\alpha}, \hat{\beta}))$ using the `rstiefel` R package (Hoff, 2013). We parameterize $\mathcal{D}_{\hat{\alpha},\hat{\beta}}$ using a normalized version w of w_2, where $w_2 = -\sqrt{\frac{1-\hat{\alpha}'\hat{\beta}}{2}} \times w$. Here, (i) $w = 1$ indicates that $\phi_{\hat{\gamma}}(X) = \hat{\beta}'X$, which is an estimated balancing score; (ii) $w = -1$ means $\phi_{\hat{\gamma}}(X) = \hat{\alpha}'X$, which is an estimated prognostic score; (iii) $w = 0$ gives a $\phi_{\hat{\gamma}}(X)$ that is equiangular to the estimated prognostic and balancing scores on the hyperbola; we denote its coefficient vector by $\hat{\delta}$ and its ground-truth analogue by δ.

5.2 Results on a Simulated Dataset

Using the model given in Assumption 4.3, we set m to the identity and h to the inverse logit function. We generate n i.i.d. triples (X_i, T_i, Y_i) according to

$$
\begin{aligned}
&X \sim \mathcal{N}(0, I_p),\\
&T \sim \text{Bernoulli}(e(X)), \quad e(X) = \text{logit}^{-1}(\beta_0 + s_T X'\beta),\\
&Y \sim \mathcal{N}(\alpha_0 + s_Y X'\alpha + \tau T, 1).
\end{aligned}
$$

Here, s_Y corresponds to the signal-to-noise ratio (SNR) for the outcome model, while s_T controls overlap; higher values of s_T correspond to poorer overlap. α and β are constructed to share the same 20-element support with $\alpha'\beta = 0.75$. We take $n = 500$, $p = 1000$ and $\tau = 0$. We consider high overlap $s_T = 1$ and low overlap $s_T = 4$, as well as $s_Y = 2$ and $s_Y = 5$ with larger values implying a higher SNR. Since the true outcome and propensity models are sparse, LASSO with appropriate variable selection is correctly specified, while Ridge, which does not perform variable selection, is misspecified. We report averages across 100 runs.

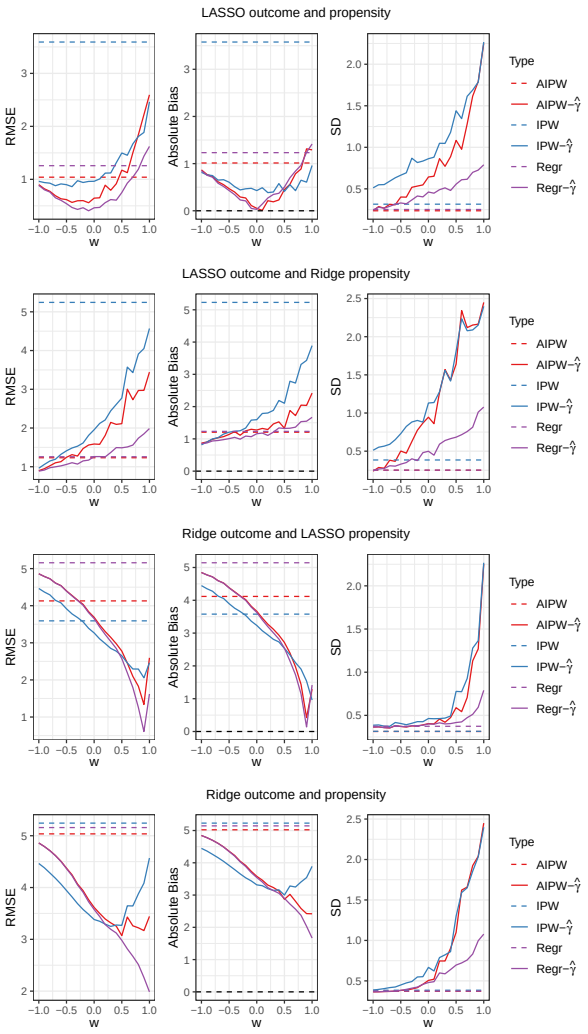


Figure 2: **RMSE, bias and standard deviation for simulated datasets.** Each metric is plotted according to the deconfounding score coordinate parameter w; methods using base covariates are constant. In all plots, $s_T = 4$ (low overlap) and $s_Y = 5$ (high SNR). See Section 5.1 for definitions of names of estimators.

Results on the Overall RMSE. Table 1 reports the root mean squared error (RMSE) of ATT estimates using either the original covariates or three decon-

Table 1: RMSEs on simulated datasets. Green: $\phi_{\hat{\gamma}}(X)$ improves over X for the same base method; Red: it is worse. Best performance underlined. See Section 5.1 for definitions of names of estimators.

Overlap	High		Low	
SNR	Low	High	Low	High
LASSO outcome and propensity				
IPW	0.98	2.45	1.45	3.59
AIPW	0.35	0.39	0.79	1.04
Regr	0.4	0.43	1	1.26
IPW-$\hat{\beta}$	0.66	1.58	1.1	2.46
AIPW-$\hat{\beta}$	0.59	1.39	1.13	2.6
Regr-$\hat{\beta}$	0.47	1.12	0.68	1.62
IPW-$\hat{\delta}$	0.43	1.01	0.5	0.96
AIPW-$\hat{\delta}$	0.39	0.88	0.37	0.65
Regr-$\hat{\delta}$	0.39	0.88	0.31	0.46
IPW-$\hat{\alpha}$	0.27	0.56	0.55	0.96
AIPW-$\hat{\alpha}$	0.22	0.26	0.65	0.9
Regr-$\hat{\alpha}$	0.22	0.25	0.64	0.89
LASSO outcome and Ridge propensity				
IPW	1.17	2.93	2.1	5.24
AIPW	0.39	0.42	0.98	1.23
Regr	0.4	0.43	1	1.26
IPW-$\hat{\beta}$	1.14	2.7	1.92	4.57
AIPW-$\hat{\beta}$	1.15	2.71	1.51	3.44
Regr-$\hat{\beta}$	0.47	1.09	0.84	1.99
IPW-$\hat{\delta}$	0.32	0.71	0.79	1.96
AIPW-$\hat{\delta}$	0.28	0.61	0.74	1.59
Regr-$\hat{\delta}$	0.22	0.45	0.57	1.26
IPW-$\hat{\alpha}$	0.18	0.29	0.55	0.96
AIPW-$\hat{\alpha}$	0.22	0.25	0.65	0.9
Regr-$\hat{\alpha}$	0.22	0.25	0.64	0.89
Ridge outcome and LASSO propensity				
IPW	0.98	2.45	1.45	3.59
AIPW	0.99	2.47	1.66	4.13
Regr	1.12	2.81	2.06	5.16
IPW-$\hat{\beta}$	0.66	1.58	1.1	2.46
AIPW-$\hat{\beta}$	0.61	1.45	1.13	2.6
Regr-$\hat{\beta}$	0.5	1.2	0.68	1.62
IPW-$\hat{\delta}$	0.66	1.64	1.29	3.27
AIPW-$\hat{\delta}$	0.75	1.9	1.45	3.69
Regr-$\hat{\delta}$	0.75	1.9	1.43	3.65
IPW-$\hat{\alpha}$	0.89	2.22	1.79	4.46
AIPW-$\hat{\alpha}$	1.02	2.55	1.94	4.86
Regr-$\hat{\alpha}$	1.02	2.55	1.94	4.86
Ridge outcome and propensity				
IPW	1.17	2.93	2.1	5.24
AIPW	1.09	2.73	2.02	5.04
Regr	1.12	2.81	2.06	5.16
IPW-$\hat{\beta}$	1.14	2.7	1.92	4.57
AIPW-$\hat{\beta}$	1.15	2.71	1.51	3.44
Regr-$\hat{\beta}$	0.47	1.09	0.84	1.99
IPW-$\hat{\delta}$	0.51	1.22	1.37	3.38
AIPW-$\hat{\delta}$	0.55	1.37	1.46	3.62
Regr-$\hat{\delta}$	0.55	1.35	1.43	3.58
IPW-$\hat{\alpha}$	0.88	2.19	1.79	4.46
AIPW-$\hat{\alpha}$	1.01	2.54	1.94	4.86
Regr-$\hat{\alpha}$	1.01	2.54	1.94	4.86

founding scores: the estimated balancing score, prognostic score, and equiangular score. Several patterns emerge. When the outcome model is well-specified, prognostic-score-based estimators always improve over original estimators and often have the lowest RMSE, particularly when the propensity model is misspecified. When the outcome model is misspecified but the propensity model is well-specified, the pattern reverses: balancing-score-based methods dominate. When both models are misspecified, deconfounding scores nearly always outperform raw covariates. Overall, the equiangular score is an attractive middle ground: it benefits from combining information from both models and avoids overfitting to a single poorly-specified score. We detail this next.

Decomposition into Bias and Variance. To explain the results in Table 1, Figure 2 plots the RMSE, absolute bias, and standard deviation (SD) of estimators for $s_T = 4$ and $s_Y = 5$ across deconfounding scores parameterized by $w \in \{-1, -0.9, \ldots, 0.9, 1\}$. The SD generally decreases as w moves from 1 (estimated balancing score) toward -1 (estimated prognostic score), consistent with the theory, although deconfounding-score-based estimators have somewhat higher SDs than those using raw covariates. Since RMSE is dominated by bias in these experiments, the SD patterns are less consequential for overall performance.

The bias patterns are more informative and track model specification closely: (i) when both outcome and propensity models are well-specified, intermediate deconfounding scores exhibit lower bias, likely because they combine information from both models; (ii) when only one model is well-specified, bias is lowest near the corresponding endpoint—the prognostic score when the outcome model is correct, the balancing score when the propensity model is correct—and increases toward the misspecified endpoint; (iii) when both models are misspecified, bias for deconfounding scores is generally lower than for raw covariates.

5.3 Results on Semi-Synthetic Datasets

We now assess these estimators on canonical semi-synthetic datasets: IHDP (Hill, 2011), ACIC 2016 (Dorie et al., 2017) and HC-MNIST (Jesson et al., 2021). Critically, they do not satisfy Assumptions 4.2 and 4.3. IHDP offers 6 different settings to generate the data, ACIC 2016 offers 77 settings, and HC-MNIST offers 1 setting. Thus, we report RMSEs over both settings and runs for each of these datasets. We conduct 100 runs. Results are in Table 2. We note that the best performance is always achieved by a deconfounding score. Performance of each individual type of deconfounding score depends on the dataset and specification of models. Overall, estimated prognos-

tic scores tend to outperform original covariates more frequently than other scores, consistent with the intuition from the theory. However, other scores can offer superior performance depending on the dataset, notably the equiangular score which again proves to be an attractive alternative. We hypothesize that this overall behavior depends on which model (outcome or propensity) is adequately captured by the corresponding learned coefficient vector ($\hat{\alpha}$ or $\hat{\beta}$, respectively) under misspecification, with the equiangular score potentially offering a robust compromise.

6 CONCLUSION

We have introduced deconfounding scores—representations subject to a zero-confounding-bias constraint—and shown that prognostic scores are overlap-optimal—as measured by overlap divergence—within a family of deconfounding scores under Gaussian covariates and generalized linear models. In experiments, there is always at least one deconfounding score improving ATT estimation over raw covariates, with simulations suggesting that performance of deconfounding scores is determined by the correctness of the estimated outcome and propensity models. Importantly, our approach is complementary to existing estimators: deconfounding scores can serve as drop-in replacements for covariates in any treatment effect estimation method, including AIPW (Robins et al., 1994), TMLE (Van der Laan et al., 2011), and double/debiased machine learning (Chernozhukov et al., 2018, 2021).

Two key limitations merit future work. First, our analytical results rely on restrictive assumptions (Gaussian covariates, generalized linear models), and performance depends on correct specification of the outcome and propensity models; analyzing the impact of estimation error in the coefficient vectors is a natural next step. Second, estimating overlap-optimal representations from finite samples in more general settings remains an open problem. One direction would be to generalize classical χ^2-divergence optimization methods (Nguyen et al., 2010; Dieng et al., 2017; Huggins et al., 2020) to non-trivial push-forward measures with a deconfounding score constraint. Additionally, while Lemma 3.2 identifies the squared density ratio as the dominant term in the efficiency bound, other terms may become important when the density ratio is moderate; jointly optimizing the full efficiency bound with respect to the representation is a challenging but valuable direction. We discuss possible extensions more broadly in Appendix B.6.

Table 2: RMSEs on semi-synthetic datasets. Green: $\phi_{\hat{\gamma}}(X)$ improves over X for the same base method; Red: it is worse. Best performance underlined. See Section 5.1 for definitions of names of estimators.

Dataset	IHDP	ACIC2016	HC-MNIST
LASSO outcome and propensity			
IPW	2.35	2.29	0.2
AIPW	2.41	2.09	0.22
Regr	2.41	0.62	0.22
IPW-$\hat{\beta}$	2.46	2.48	0.18
AIPW-$\hat{\beta}$	2.43	2.41	0.22
Regr-$\hat{\beta}$	2.46	0.64	0.14
IPW-$\hat{\delta}$	2.21	1.27	0.49
AIPW-$\hat{\delta}$	2.21	1.41	0.96
Regr-$\hat{\delta}$	2.21	0.97	0.5
IPW-$\hat{\alpha}$	2.44	1.04	0.17
AIPW-$\hat{\alpha}$	2.43	0.72	0.21
Regr-$\hat{\alpha}$	2.41	0.6	0.2
LASSO outcome and Ridge propensity			
IPW	2.38	1.78	0.21
AIPW	2.41	1.61	0.22
Regr	2.41	0.62	0.22
IPW-$\hat{\beta}$	2.53	2.24	0.19
AIPW-$\hat{\beta}$	2.48	2.16	0.22
Regr-$\hat{\beta}$	2.53	0.63	0.14
IPW-$\hat{\delta}$	2.24	1.08	0.48
AIPW-$\hat{\delta}$	2.24	1.21	0.48
Regr-$\hat{\delta}$	2.24	1.03	0.49
IPW-$\hat{\alpha}$	2.44	1.04	0.17
AIPW-$\hat{\alpha}$	2.43	0.72	0.21
Regr-$\hat{\alpha}$	2.41	0.6	0.2
Ridge outcome and LASSO propensity			
IPW	2.35	2.29	0.2
AIPW	2.4	2.04	0.22
Regr	2.4	0.62	0.27
IPW-$\hat{\beta}$	2.46	2.48	0.18
AIPW-$\hat{\beta}$	2.43	2.41	0.22
Regr-$\hat{\beta}$	2.46	0.64	0.14
IPW-$\hat{\delta}$	2.19	1.3	0.67
AIPW-$\hat{\delta}$	2.2	1.36	3.55
Regr-$\hat{\delta}$	2.2	1.02	0.5
IPW-$\hat{\alpha}$	2.45	1.02	0.19
AIPW-$\hat{\alpha}$	2.43	0.69	0.23
Regr-$\hat{\alpha}$	2.42	0.6	0.22
Ridge outcome and propensity			
IPW	2.38	1.78	0.21
AIPW	2.41	1.57	0.23
Regr	2.4	0.62	0.27
IPW-$\hat{\beta}$	2.53	2.24	0.19
AIPW-$\hat{\beta}$	2.48	2.16	0.22
Regr-$\hat{\beta}$	2.53	0.63	0.14
IPW-$\hat{\delta}$	2.23	1.12	0.61
AIPW-$\hat{\delta}$	2.23	1.25	5.66
Regr-$\hat{\delta}$	2.23	1.08	0.52
IPW-$\hat{\alpha}$	2.45	1.02	0.19
AIPW-$\hat{\alpha}$	2.43	0.69	0.23
Regr-$\hat{\alpha}$	2.42	0.6	0.22

Acknowledgements

We sincerely thank anonymous reviewers for valuable feedback. O.C. was supported by Novo Nordisk and the U.K. Engineering and Physical Sciences Research Council through the Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (Project EP/S023151/1). A.D. is an employee of Google and may own stock as a part of a standard compensation package. Al.F. was supported in part by the U.S. National Institutes of Health through Grants 1R01GM144967-01 and 1R03CA211160-01, by the U.S. National Science Foundation through Award 1924205, and by the Chan Zuckerberg Initiative. D.B.-S. and Av.F. were supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R305D200010 and R305D240036, and by the U.S. National Science Foundation through Award 2243822. C.H. was supported by the Alan Turing Institute, the Li Ka Shing Foundation, the Ellison Institute of Technology, the U.K. Engineering and Physical Sciences Research Council through the Bayes4Health grant EP/R018561/1, and U.K. Research and Innovation through the Medical Research Council and the "AI and data science for engineering, health and government (ASG)" programme.

References

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.

Peter C Austin, Paul Grootendorst, and Geoffrey M Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4): 734–753, 2007.

David Benkeser, Weixin Cai, and Mark J. van der Laan. A nonparametric super-efficient estimator of the average treatment effect. *Statistical Science*, 35 (3):484 – 495, 2020. doi: 10.1214/19-STS735. URL https://doi.org/10.1214/19-STS735.

M. Brookhart, S. Schneeweiss, K. Rothman, K. Rothman, R. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163 12:1149–56, 2006. URL https://academic.oup.com/aje/article-pdf/163/12/1149/223159/kwj149.pdf.

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. URL https://proceedings.mlr.press/v151/bruns-smith22a/bruns-smith22a.pdf.

Saraswata Chaudhuri and Jonathan B Hill. Robust estimation for average treatment effects. *Work& ing Paper, Dept. of Economics, University of North Carolina*, 2013.

V. Chernozhukov, Whitney Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018. URL https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf.

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022a. URL https://www.nber.org/system/files/working_papers/w30302/w30302.pdf.

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, 2022b. URL https://proceedings.mlr.press/v162/chernozhukov22a/chernozhukov22a.pdf.

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. URL https://proceedings.mlr.press/v151/clivio22a/clivio22a.pdf.

Oscar Clivio, Avi Feller, and Chris Holmes. Towards representation learning for weighting problems in design-based causal inference. *arXiv preprint arXiv:2409.16407*, 2024. URL https://arxiv.org/pdf/2409.16407.

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023. URL https://arxiv.org/pdf/2303.16008.

Bénédicte Colnet, Julie Josse, Gaël Varoquaux, and Erwan Scornet. Re-weighting the randomized controlled trial for generalization: Finite-sample error

and variable selection. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188(2): 345–372, 05 2024. ISSN 0964-1998.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009. URL https://dash.harvard.edu/bitstream/handle/1/3007645/imbens_addressing.pdf.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates, 2020. URL http://arxiv.org/pdf/1711.02582v4.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via \chi upper bound minimization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/35464c848f410e55a13bb9d78e7fddd0-Paper.pdf.

Vincent Dorie, J. Hill, Uri Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 2017. URL https://doi.org/10.1214/18-sts667.

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. URL https://www.sciencedirect.com/science/article/pii/S0304407620302694.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998. URL https://statweb.rutgers.edu/ztan/material/hahn98.pdf.

Jaroslav Hájek. Comment on a paper by d. basu. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, page 236. Holt, Rinehart and Winston, Toronto, 1971.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008. URL https://scholar.archive.org/work/qk3kv4htq5hbrd6y4c6dsmz5wy/access/wayback/http://www.nyu.edu/gsas/dept/politics/seminars/analogue2007-03.pdf.

Miguel A Hernán and James M Robins. Causal inference, 2010. URL https://grass.upc.edu/en/seminar/presentation-files/causal-inference/chapters-1-i-2/@@download/file/BookHernanRobinsCap1_2.pdf.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Peter D. Hoff. Bayesian analysis of matrix data with rstiefel, 2013. URL https://arxiv.org/abs/1304.3673.

Han Hong, Michael P Leung, and Jessie Li. Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1): 32–47, 2020. URL https://academic.oup.com/ectj/article/23/1/32/5558232.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685, 1952. URL https://e-l.unifi.it/pluginfile.php/808426/mod_resource/content/1/Horvitz-Thompson-1952-jasa.pdf.

Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1792–1802. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/huggins20a.html.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2021. URL http://proceedings.mlr.press/v139/jesson21a/jesson21a.pdf.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016. URL http://proceedings.mlr.press/v48/johansson16.pdf.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR,

2019. URL http://proceedings.mlr.press/v89/johansson19a/johansson19a.pdf.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *The Journal of Machine Learning Research*, 23(1):7489–7538, 2022. URL https://www.jmlr.org/papers/volume23/19-511/19-511.pdf.

Nathan Kallus. Generalized optimal matching methods for causal inference. *The Journal of Machine Learning Research*, 21(1):2300–2353, 2020. URL https://www.jmlr.org/papers/volume21/19-120/19-120.pdf.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: partial identification through smoothness, 2024. URL http://arxiv.org/pdf/2305.11812v2.

Shakeeb Khan and Elie Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010. URL https://economics.uwo.ca/newsletter/misc/2007/khan_nov21.pdf.

Finbarr P Leacy and Elizabeth A Stuart. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine*, 33(20):3488–3508, 2014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3995901/.

Myoung-jae Lee and Sanghyeok Lee. Review and comparison of treatment effect estimators using propensity and prognostic scores. *The International Journal of Biostatistics*, 18:357 – 380, 2022.

Sokbae Lee and Martin Weidner. Bounding treatment effects by pooling limited information across observations. *arXiv preprint arXiv:2111.05243*, 2021. URL https://arxiv.org/pdf/2111.05243.

Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990. ISSN 00028282. URL http://www.jstor.org/stable/2006592.

Roland A Matsouaka and Yunji Zhou. A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights. *arXiv preprint arXiv:2011.01388*, 2020. URL https://arxiv.org/pdf/2011.01388.

Mohammad Mehrabi and Stefan Wager. Off-policy evaluation in markov decision processes under weak distributional overlap, 2024. URL http://arxiv.org/pdf/2402.08201v1.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. *ArXiv*, abs/2311.11321, 2023.

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Orthogonal representation learning for estimating causal quantities. *arXiv preprint arXiv:2502.04274*, 2025.

Erica EM Moodie, Olli Saarela, and David A Stephens. A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, 7(1):e205, 2018. URL https://escholarship.mcgill.ca/downloads/6h440z638.

Rachel C Nethery, Fabrizia Mealli, and Francesca Dominici. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The annals of applied statistics*, 13(2):1242, 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6658123/.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. doi: 10.1109/TIT.2010.2068870.

Donald Bruce Owen. A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, 9(4):389–419, 1980.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4107929/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994. URL https://www.academia.edu/download/42230525/Estimation_of_Regression_Coefficients_Wh20160206-14055-10joi7l.pdf.

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

1983. URL https://academic.oup.com/biomet/article-pdf/70/1/41/662954/70-1-41.pdf.

Christoph Rothe. Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660, 2017. URL https://www.econstor.eu/bitstream/10419/107545/1/dp8758.pdf.

Donald V. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127:757–763, 1997.

Kara E Rudolph, Nicholas T Williams, Elizabeth A Stuart, and Ivan Diaz. Efficiently transporting average treatment effects using a sufficient subset of effect modifiers. *arXiv preprint arXiv:2304.00117*, 2023. URL https://arxiv.org/pdf/2304.00117.

S. Schneeweiss, J. Rassen, R. Glynn, J. Avorn, H. Mogun, and M. Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20: 512–522, 2009. URL https://doi.org/10.1097/ede.0b013e3181a663cc.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, and Charles Fisher. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score, 2021. URL http://arxiv.org/pdf/2012.09935v3.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017. URL http://proceedings.mlr.press/v70/shalit17a/shalit17a.pdf.

Til Stürmer, Kenneth J Rothman, Jerry Avorn, and Robert J Glynn. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American journal of epidemiology*, 172(7):843–854, 2010. URL https://academic.oup.com/aje/article/172/7/843/86816.

F. William Townes. Review of probability distributions for modeling count data, 2020. URL https://arxiv.org/abs/2001.04343.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011. URL https://helios2.mi.parisdescartes.fr/~chambaz/Atelier209/07vanderLaan.pdf.

Jeffrey M Wooldridge. Should instrumental variables be used as matching variables? *Research in Economics*, 70(2):232–237, 2016. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ba7e757d93c15f1eecca73410c16c19594a76942.

Pengzhou (Abel) Wu and K. Fukumizu. Beta-intact-vae: Identifying and estimating causal effects under limited overlap. *ArXiv*, abs/2110.05225, 2021.

Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020. URL http://proceedings.mlr.press/v108/zhang20c/zhang20c.pdf.

Deconfounding Scores and Representation Learning for Causal Effect Estimation with Weak Overlap: Supplementary Materials

A PROOFS

A.1 Proof of Lemma 3.1

For any ϕ, we have

$$
\begin{aligned}
&\tau_{\phi(X)} - \tau \\
&= \tau_{\phi(X)} - \tau_X \text{ by unconfoundedness} \\
&= \mathbb{E}_{P^1}[Y] - \mathbb{E}_{P^1}[m_0(\phi(X))] - (\mathbb{E}_{P^1}[Y] - \mathbb{E}_{P^1}[m_0(X)]) \\
&= \mathbb{E}_{P^1}[\mathbb{E}_{P^0}[Y|X]] - \mathbb{E}_{P^1}[\mathbb{E}_{P^0}[Y|\phi(X)]] \\
&= \mathbb{E}_{P^0}\left[(m_0(X) - \mathbb{E}_{P^0}[m_0(X)|\phi(X)]) \cdot \left(\frac{dP_X^1}{dP_X^0}(X) - \mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]\right)\right] \\
&\quad \text{from Proposition 3.4 of Clivio et al. (2024) and } m_0(\phi(X)) = \mathbb{E}_{P^0}[m_0(X)|\phi(X)] \\
&= \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[(m_0(X) - \mathbb{E}_{P^0}[m_0(X)|\phi(X)]) \cdot \left(\frac{dP_X^1}{dP_X^0}(X) - \mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]\right)\middle|\phi(X)\right]\right] \\
&\quad \text{from the tower property} \\
&= \mathbb{E}_{P^0}\left[\mathrm{Cov}_{P^0}\left(m_0(X), \frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right)\right].
\end{aligned}
$$

A.2 Proof of Lemma 3.2

Let ϕ be a representation. From Theorem 1 of Hahn (1998), $V_{\text{eff}}^{\phi(X)}$ is equal to

$$
\begin{aligned}
\mathbb{E}_P\Big[&\frac{e(\phi(X))\mathrm{Var}_{P^1}(Y|\phi(X))}{\pi_1^2} + \frac{e(\phi(X))^2\mathrm{Var}_{P^0}(Y|\phi(X))}{\pi_1^2(1-e(\phi(X)))} \\
&+ \frac{\left(m_1(\phi(X)) - m_0(\phi(X)) - \tau_{\phi(X)}\right)^2 e(\phi(X))}{\pi_1^2}\Big].
\end{aligned}
$$

We note that the second term of this sum would be equal to the overlap divergence up to a constant if the conditional variance of Y was constant. Thus, we attempt to upper-bound or lower-bound this conditional variance with constants; this will give the result.

From the law of total variance,

$$
\begin{aligned}
\mathrm{Var}_{P^0}(Y|\phi(X)) &= \mathbb{E}_{P^0}[\mathrm{Var}_{P^0}(Y|\phi(X), X)|\phi(X)] + \mathrm{Var}_{P^0}(\mathbb{E}_{P^0}[Y|\phi(X), X]|\phi(X)) \\
&= \mathbb{E}_{P^0}[\mathrm{Var}_{P^0}(Y|X)|\phi(X)] + \mathrm{Var}_{P^0}(m_0(X)|\phi(X)) \\
&\geq \mathbb{E}_{P^0}[\mathrm{Var}_{P^0}(Y|X)|\phi(X)],
\end{aligned}
$$

so if $\mathrm{Var}_{P^0}(Y|X) \geq \sigma^2$ then $\mathrm{Var}_{P^0}(Y|\phi(X)) \geq \sigma^2$ and

$$
V_{\text{eff}}^{\phi(X)} \geq \mathbb{E}_P\left[\frac{e(\phi(X))^2\mathrm{Var}_{P^0}(Y|\phi(X))}{\pi_1^2(1-e(\phi(X)))}\right]
$$

$$
\begin{aligned}
&\geq \sigma^2 \mathbb{E}_P \left[\frac{e(\phi(X))^2}{\pi_1^2 (1 - e(\phi(X)))} \right] \text{ from the above} \\
&= \sigma^2 \mathbb{E}_P \left[\frac{1 - e(\phi(X))}{(1-\pi_1)^2} \left(\frac{dP^1_{\phi(X)}}{dP^0_{\phi(X)}}(\phi(X)) \right)^2 \right] \text{ as } \frac{dP^1_{\phi(X)}}{dP^0_{\phi(X)}}(\phi(X)) = \frac{(1-\pi_1)e(\phi(X))}{\pi_1(1-e(\phi(X)))} \\
&= \frac{\sigma^2}{1-\pi_1} \mathbb{E}_{P^0} \left[\left(\frac{dP^1_{\phi(X)}}{dP^0_{\phi(X)}}(\phi(X)) \right)^2 \right] \text{ as } \frac{dP^0_{\phi(X)}}{dP_{\phi(X)}}(\phi(X)) = \frac{1 - e(\phi(X))}{1-\pi_1} \\
&= \frac{\sigma^2}{1-\pi_1} \mathcal{O}(\phi(X)).
\end{aligned}
$$

Further, if $|Y| \leq Y_{\max}$ for a constant $Y_{\max} > 0$ then any conditional variance of Y is bounded by $Y^2_{\max}$ and any conditional expectation of Y is bounded by $Y_{\max}$, and the propensity score is always bounded by 1, so

$$
\begin{aligned}
\mathbb{E}_P \Big[\frac{e(\phi(X)) \mathrm{Var}_{P^1}(Y|\phi(X))}{\pi_1^2} \Big] &= \mathbb{E}_{P^1} \Big[\frac{\mathrm{Var}_{P^1}(Y|\phi(X))}{\pi_1} \Big] \text{ as } \frac{dP^1_{\phi(X)}}{dP_{\phi(X)}}(\phi(X)) = \frac{e(\phi(X))}{\pi_1} \\
&\leq \frac{Y^2_{\max}}{\pi_1}
\end{aligned}
$$

and, similarly as above,

$$
\mathbb{E}_P \Big[\frac{e(\phi(X))^2 \mathrm{Var}_{P^0}(Y|\phi(X))}{\pi_1^2 (1 - e(\phi(X)))} \Big] \leq \frac{Y^2_{\max}}{1-\pi_1} \mathcal{O}(\phi(X)).
$$

Finally,

$$
\begin{aligned}
&\mathbb{E}_P \Big[\frac{\big(m_1(\phi(X)) - m_0(\phi(X)) - \tau_{\phi(X)}\big)^2 e(\phi(X))}{\pi_1^2} \Big] \\
&= \frac{1}{\pi_1} \mathbb{E}_{P^1} \Big[\big(m_1(\phi(X)) - m_0(\phi(X)) - \tau_{\phi(X)}\big)^2 \Big] \text{ as } \frac{dP^1_{\phi(X)}}{dP_{\phi(X)}}(\phi(X)) = \frac{e(\phi(X))}{\pi_1} \\
&= \frac{1}{\pi_1} \mathrm{Var}_{P^1} \Big(m_1(\phi(X)) - m_0(\phi(X)) \Big) \\
&\leq \frac{1}{\pi_1} \mathbb{E}_{P^1} \Big[\big(m_1(\phi(X)) - m_0(\phi(X))\big)^2 \Big] \\
&\leq \frac{1}{\pi_1} \mathbb{E}_{P^1} \Big[2 \big(m_1(\phi(X))^2 + m_0(\phi(X))^2 \big) \Big] \text{ from } (a-b)^2 \leq 2(a^2+b^2) \ \ \forall a, b \\
&\leq \frac{4 Y^2_{\max}}{\pi_1}.
\end{aligned}
$$

All of this yields

$$
V_{\text{eff}}^{\phi(X)} \leq \frac{5 Y^2_{\max}}{\pi_1} + \frac{Y^2_{\max}}{1-\pi_1} \mathcal{O}(\phi(X)).
$$

A.3 Proof of Lemma 4.1

For any ϕ, we have

$$
\begin{aligned}
&\mathcal{O}(X) - \mathcal{O}(\phi(X)) \\
&= \mathbb{E}_{P^0} \left[\left(\frac{dP^1_X}{dP^0_X}(X) \right)^2 \right] - \mathbb{E}_{P^0} \left[\left(\frac{dP^1_{\phi(X)}}{dP^0_{\phi(X)}}(\phi(X)) \right)^2 \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{P^0}\left[\left(\frac{dP_X^1}{dP_X^0}(X)\right)^2\right] - \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]^2\right] \\
&\quad \text{from Proposition 3.4 of Clivio et al. (2024)} \\
&= \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\left(\frac{dP_X^1}{dP_X^0}(X)\right)^2\middle|\phi(X)\right]\right] - \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]^2\right] \\
&\quad \text{from the tower property} \\
&= \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\left(\frac{dP_X^1}{dP_X^0}(X)\right)^2\middle|\phi(X)\right] - \mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]^2\right] \\
&= \mathbb{E}_{P^0}\left[\mathrm{Var}_{P^0}\left(\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right)\right].
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\left|\tau_{\phi(X)} - \tau\right| \\
&= \left|\mathbb{E}_{P^0}\left[(m_0(X) - \mathbb{E}_{P^0}[m_0(X)|\phi(X)]) \cdot \left(\frac{dP_X^1}{dP_X^0}(X) - \mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]\right)\right]\right| \\
&\quad \text{from the proof of Lemma 3.1} \\
&\leq \sqrt{\mathbb{E}_{P^0}\left[(m_0(X) - \mathbb{E}_{P^0}[m_0(X)|\phi(X)])^2\right]} \sqrt{\mathbb{E}_{P^0}\left[\left(\frac{dP_X^1}{dP_X^0}(X) - \mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]\right)^2\right]} \\
&\quad \text{from the Cauchy-Schwarz inequality} \\
&= \sqrt{\mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[(m_0(X) - \mathbb{E}_{P^0}[m_0(X)|\phi(X)])^2\middle|\phi(X)\right]\right]} \\
&\quad \times \sqrt{\mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\left(\frac{dP_X^1}{dP_X^0}(X) - \mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right]\right)^2\middle|\phi(X)\right]\right]} \\
&\quad \text{from the tower property} \\
&= \sqrt{\mathbb{E}_{P^0}\left[\mathrm{Var}_{P^0}\left(m_0(X)|\phi(X)\right)\right]} \sqrt{\mathbb{E}_{P^0}\left[\mathrm{Var}_{P^0}\left(\frac{dP_X^1}{dP_X^0}(X)\middle|\phi(X)\right)\right]} \\
&= \sqrt{\mathbb{E}_{P^0}\left[\mathrm{Var}_{P^0}\left(m_0(X)|\phi(X)\right)\right]} \sqrt{\mathcal{O}(X) - \mathcal{O}(\phi(X))} \text{ from the above.}
\end{aligned}
$$

A.4 Proof of Theorem 4.4

Preliminary. Assume that R is a distribution such that $R_X = \mathcal{N}(0, I_d)$. We show that for any functions f_1, f_2 and any $\gamma \in \mathcal{D}_{\alpha,\beta}$, then $\mathrm{Cov}_R\left(f_1(\alpha'X), f_2(\beta'X)|\gamma'X\right) = 0$. This result will be used to prove the Theorem by expressing the confounding bias using such a conditional covariance.

Indeed, as $R_X = \mathcal{N}(0, I_d)$, $\mathrm{Cov}_R\left(\alpha'X, \beta'X|\gamma'X\right) = 0$ implies $\mathrm{Cov}_R\left(f_1(\alpha'X), f_2(\beta'X)|\gamma'X\right) = 0$ regardless of f_1, f_2, and $\mathrm{Cov}_R\left(\alpha'X, \beta'X|\gamma'X\right) = \alpha'\beta - \alpha'\gamma\gamma'\beta$ so $\mathrm{Cov}_R\left(\alpha'X, \beta'X|\gamma'X\right)$ being zero is equivalent to

$$
\gamma'\left(\frac{\alpha\beta' + \beta\alpha'}{2}\right)\gamma = \alpha'\beta.
$$

As $\alpha'\beta \notin \{-1, 1\}$ with $||\alpha||_2 = ||\beta||_2 = 1$, α and β are not collinear so $\left(\frac{\alpha\beta' + \beta\alpha'}{2}\right)$ is rank 2 with exactly one positive eigenvalue and one negative eigenvalue. Specifically, the eigenvectors u_1, u_2 and eigenvalues λ_1, λ_2 are given by

$$
u_1 = \frac{(\alpha + \beta)}{\sqrt{2 + 2\alpha'\beta}}, u_2 = \frac{(\alpha - \beta)}{\sqrt{2 - 2\alpha'\beta}}, \tag{2}
$$

$$\lambda_1 = \frac{(\alpha'\beta + 1)}{2}, \lambda_2 = \frac{(\alpha'\beta - 1)}{2}. \tag{3}$$

Then we have a solution characterized by the hyperbola

$$(\alpha'\beta + 1)\left(\frac{(\alpha+\beta)'\gamma}{\sqrt{2+2\alpha'\beta}}\right)^2 - (1-\alpha'\beta)\left(\frac{(\alpha-\beta)'\gamma}{\sqrt{2-2\alpha'\beta}}\right)^2 = 2\alpha'\beta. \tag{4}$$

Simplifying notation by collapsing scalars into w_1 and w_2 yields that for any $\gamma \in \mathcal{D}_{\alpha,\beta}$, we have $\mathrm{Cov}_R\left(f_1(\alpha'X), f_2(\beta'X)|\gamma'X\right) = 0$. We now use this to prove the result, separating by Assumption 4.2 or 4.3.

If Assumption 4.2 Applies. From Lemma 3.1, a representation $\phi_\gamma(X)$ will be a deconfounding score if

$$\begin{aligned} 0 = \tau_{\phi_\gamma(X)} - \tau &= \mathbb{E}_{P^0}\left[\mathrm{Cov}_{P^0}\left(m_0(X), \frac{dP_X^1}{dP_X^0}(X)\middle|\phi_\gamma(X)\right)\right] \\ &= \mathbb{E}_{P^0}\left[\mathrm{Cov}_{P^0}\left(m(\alpha'X), h(\beta'X)|\gamma'X\right)\right] \end{aligned}$$

where $P_X^0 = \mathcal{N}(0, I_d)$ so the Preliminary gives the result.

If Assumption 4.3 Applies. From Proposition 3.4 of Clivio et al. (2024), the confounding bias can be written as

$$\begin{aligned} &\tau_{\phi_\gamma(X)} - \tau \\ &= \mathbb{E}_{P^0}\left[m_0(X)\left(\frac{dP_X^1}{dP_X^0}(X) - \frac{dP^1_{\phi_\gamma(X)}}{dP^0_{\phi_\gamma(X)}}(\phi_\gamma(X))\right)\right] \\ &= \frac{1-\pi_1}{\pi_1}\mathbb{E}_{P^0}\left[m_0(X)\left(\frac{e(X)}{1-e(X)} - \frac{e(\phi_\gamma(X))}{1-e(\phi_\gamma(X))}\right)\right] \\ &= \frac{1}{\pi_1}\mathbb{E}_{P^0}\left[\frac{1-\pi_1}{1-e(X)}m_0(X)\left(e(X) - e(\phi_\gamma(X))\right)\right] \\ &\quad + \frac{1}{\pi_1}\mathbb{E}_{P^0}\left[\frac{1-\pi_1}{1-e(X)}m_0(X)\left(e(\phi_\gamma(X)) - \frac{e(\phi_\gamma(X))(1-e(X))}{1-e(\phi_\gamma(X))}\right)\right] \\ &= \frac{1}{\pi_1}\mathbb{E}_{P}\left[m_0(X)\left(e(X) - e(\phi_\gamma(X))\right)\right] \\ &\quad + \frac{1}{\pi_1}\mathbb{E}_{P}\left[m_0(X)\left(e(\phi_\gamma(X)) - \frac{e(\phi_\gamma(X))(1-e(X))}{1-e(\phi_\gamma(X))}\right)\right] \text{ as } \frac{dP_X}{dP_X^0}(X) = \frac{1-\pi_1}{1-e(X)} \\ &= \frac{1}{\pi_1}\mathbb{E}_{P}\left[m_0(X)\left(e(X) - e(\phi_\gamma(X))\right)\right] \\ &\quad + \frac{1}{\pi_1}\mathbb{E}_{P}\left[\frac{e(\phi_\gamma(X))}{1-e(\phi_\gamma(X))}m_0(X)\Big(1 - e(\phi_\gamma(X)) - (1-e(X))\Big)\right] \\ &= \frac{1}{\pi_1}\mathbb{E}_{P}\left[m_0(X)\left(e(X) - e(\phi_\gamma(X))\right)\right] \\ &\quad + \frac{1}{\pi_1}\mathbb{E}_{P}\left[\frac{e(\phi_\gamma(X))}{1-e(\phi_\gamma(X))}m_0(X)\Big(e(X) - e(\phi_\gamma(X))\Big)\right]. \end{aligned}$$

Further, from the tower property, for any functions f_1, f_2 of X and f_3 of $\phi_\gamma(X)$,

$$\begin{aligned} &\mathbb{E}_P\left[f_3(\phi_\gamma(X))\cdot f_1(X)\cdot\left(f_2(X) - \mathbb{E}_P\left[f_2(X)|\phi_\gamma(X)\right]\right)\right] \\ &= \mathbb{E}_P\left[f_3(\phi_\gamma(X))\cdot\left(f_1(X) - \mathbb{E}_P\left[f_1(X)|\phi_\gamma(X)\right]\right)\cdot\left(f_2(X) - \mathbb{E}_P\left[f_2(X)|\phi_\gamma(X)\right]\right)\right] \\ &= \mathbb{E}_P\left[f_3(\phi_\gamma(X))\cdot\mathbb{E}_P\left[\left(f_1(X) - \mathbb{E}_P\left[f_1(X)|\phi_\gamma(X)\right]\right)\cdot\left(f_2(X) - \mathbb{E}_P\left[f_2(X)|\phi_\gamma(X)\right]\right)\right]|\phi_\gamma(X)\right] \\ &= \mathbb{E}_P\left[f_3(\phi_\gamma(X))\cdot\mathrm{Cov}_P\left(f_1(X), f_2(X)|\phi_\gamma(X)\right)\right] \end{aligned}$$

and, as $e(\phi_\gamma(X)) = \mathbb{E}_P\left[e(X)|\phi_\gamma(X)\right]$, we obtain

$$\tau_{\phi_\gamma(X)} - \tau$$

$$
\begin{aligned}
&= \frac{1}{\pi_1}\mathbb{E}_P\left[\mathrm{Cov}_P\left(m_0(X), e(X)|\phi_\gamma(X)\right)\right] + \frac{1}{\pi_1}\mathbb{E}_P\left[\frac{e(\phi_\gamma(X))}{1-e(\phi_\gamma(X))}\mathrm{Cov}_P\left(m_0(X), e(X)|\phi_\gamma(X)\right)\right] \\
&= \frac{1}{\pi_1}\mathbb{E}_P\left[\mathrm{Cov}_P\left(m(\alpha'X), h(\beta'X)|\gamma'X\right)\right] \\
&\quad + \frac{1}{\pi_1}\mathbb{E}_P\left[\frac{e(\gamma'X)}{1-e(\gamma'X)}\mathrm{Cov}_P\left(m(\alpha'X), h(\beta'X)|\gamma'X\right)\right],
\end{aligned}
$$

so the Preliminary applied to $P_X = \mathcal{N}(0, I_d)$ gives the result.

A.5 Proof of Theorem 4.5

Sketch of the Proof. First, we compute the global minimizers of $\gamma \mapsto |\beta'\gamma|$; then we show that the overlap divergence is a non-decreasing or increasing function of $|\beta'\gamma|$, showing that the former global minimizers of $|\beta'\gamma|$ are also (the unique) global minimizers of the overlap divergence.

A.5.1 Optimal γ's for $|\beta'\gamma|$

We derive them by separating the study depending on the orientation of the hyperbola. Let $\gamma \in \mathcal{D}_{\alpha,\beta}$ with associated w_1, w_2, n.

Assume that $\alpha'\beta \geq 0$. First, note that reparameterizing w_1, w_2 as

$$
\begin{aligned}
w_1 &= \epsilon_a \sqrt{\frac{2\alpha'\beta + (1-\alpha'\beta)w^2}{1+\alpha'\beta}}, \\
w_2 &= \epsilon_b w,
\end{aligned}
$$

with $\epsilon_a, \epsilon_b \in \{-1, 1\}, w \in \left[0, \sqrt{\frac{1-\alpha'\beta}{2}}\right]$, we have

$$
\begin{aligned}
\gamma &= \frac{\epsilon_a\sqrt{2\alpha'\beta + (1-\alpha'\beta)w^2}\cdot(\alpha+\beta)}{\sqrt{2}(1+\alpha'\beta)} + \frac{\epsilon_b w\cdot(\alpha-\beta)}{\sqrt{2(1-\alpha'\beta)}} + \sqrt{\frac{1-\alpha'\beta-2w^2}{1+\alpha'\beta}}\cdot n, \\
\beta'\gamma &= \frac{1}{\sqrt{2}}\left(\epsilon_a\sqrt{2\alpha'\beta + (1-\alpha'\beta)w^2} - \epsilon_b w\sqrt{1-\alpha'\beta}\right),
\end{aligned}
$$

where the term with n has been removed as $\beta' n = 0$ by definition of n.

Assume that $\alpha'\beta \leq 0$. First, note that reparameterizing w_1, w_2 as

$$
\begin{aligned}
w_2 &= \epsilon_a \sqrt{\frac{-2\alpha'\beta + (1+\alpha'\beta)w^2}{1-\alpha'\beta}}, \\
w_1 &= \epsilon_b w,
\end{aligned}
$$

with $\epsilon_a, \epsilon_b \in \{-1, 1\}, w \in \left[0, \sqrt{\frac{1+\alpha'\beta}{2}}\right]$, we have

$$
\begin{aligned}
\gamma &= \frac{\epsilon_a\sqrt{-2\alpha'\beta + (1+\alpha'\beta)w^2}\cdot(\alpha-\beta)}{\sqrt{2}(1-\alpha'\beta)} + \frac{\epsilon_b w\cdot(\alpha+\beta)}{\sqrt{2(1+\alpha'\beta)}} + \sqrt{\frac{1+\alpha'\beta-2w^2}{1-\alpha'\beta}}\cdot n, \\
\beta'\gamma &= \frac{1}{\sqrt{2}}\left(-\epsilon_a\sqrt{-2\alpha'\beta + (1+\alpha'\beta)w^2} + \epsilon_b w\sqrt{1+\alpha'\beta}\right),
\end{aligned}
$$

where the term with n has been removed as $\beta' n = 0$ by definition of n.

Factorization. Thus, noting $\text{sign}(x) = 2 \cdot 1_{\{x \geq 0\}} - 1$ we always have

$$\begin{aligned}\gamma &= \frac{\epsilon_a \sqrt{2|\alpha'\beta| + (1-|\alpha'\beta|)w^2} \cdot (\alpha + \text{sign}(\alpha'\beta)\beta)}{\sqrt{2}(1+|\alpha'\beta|)} \\ &\quad + \frac{\epsilon_b w \cdot (\alpha - \text{sign}(\alpha'\beta)\beta)}{\sqrt{2(1-|\alpha'\beta|)}} + \sqrt{\frac{1-|\alpha'\beta|-2w^2}{1+|\alpha'\beta|}} \cdot n, \\ \beta'\gamma &= \frac{\text{sign}(\alpha'\beta)}{\sqrt{2}} \left(\epsilon_a \sqrt{2|\alpha'\beta| + (1-|\alpha'\beta|)w^2} - \epsilon_b w \sqrt{1-|\alpha'\beta|} \right),\end{aligned}$$

with $\epsilon_a, \epsilon_b \in \{-1, 1\}, w \in \left[0, \sqrt{\frac{1-|\alpha'\beta|}{2}}\right]$. Then,

$$\begin{aligned}(\beta'\gamma)^2 &= \frac{1}{2}\Big(2|\alpha'\beta| + (1-|\alpha'\beta|)w^2 + w^2(1-|\alpha'\beta|) \\ &\qquad - 2\epsilon_a\epsilon_b w\sqrt{1-|\alpha'\beta|}\sqrt{2|\alpha'\beta| + (1-|\alpha'\beta|)w^2}\Big) \\ &= |\alpha'\beta| + z - \epsilon_a\epsilon_b\sqrt{2|\alpha'\beta|z + z^2},\end{aligned}$$

where $z := (1-|\alpha'\beta|)w^2 \in \left[0, \frac{(1-|\alpha'\beta|)^2}{2}\right]$. From there, note that if $\epsilon_a\epsilon_b = -1$ then $(\beta'\gamma)^2$ is an increasing function of z: thus a unique global minimizer in this case is $z = 0$, for which $(\beta'\gamma)^2 = |\alpha'\beta|$. Now assume that $\epsilon_a\epsilon_b = 1$. Then,

$$\frac{\partial(\beta'\gamma)^2}{\partial z} = 1 - \frac{|\alpha'\beta| + z}{\sqrt{2|\alpha'\beta|z + z^2}},$$

whose sign is given by $2|\alpha'\beta|z + z^2 - (|\alpha'\beta| + z)^2 = -|\alpha'\beta|^2$.

Thus, for $\epsilon_a\epsilon_b > 0$, $(\beta'\gamma)^2$ is a decreasing function of z if $\alpha'\beta \neq 0$ and $(\beta'\gamma)^2$ is constant if $\alpha'\beta = 0$. Then one global minimum, and the only one if $\alpha'\beta \neq 0$, is achieved for the maximal value of z, that is $\frac{(1-|\alpha'\beta|)^2}{2}$. For this value of z,

$$\begin{aligned}(\beta'\gamma)^2 &= |\alpha'\beta| + \frac{(1-|\alpha'\beta|)^2}{2} - \frac{1-|\alpha'\beta|}{\sqrt{2}}\sqrt{2|\alpha'\beta| + \frac{(1-|\alpha'\beta|)^2}{2}} \\ &= \frac{1+|\alpha'\beta|^2}{2} - \frac{(1-|\alpha'\beta|)(1+|\alpha'\beta|)}{2} \\ &= |\alpha'\beta|^2\end{aligned}$$

which is lower than $|\alpha'\beta|$. Thus,

- If $\alpha'\beta \neq 0$, $(\beta'\gamma)^2$ is minimized for $\epsilon_a = \epsilon_b =: \epsilon \in \{-1, 1\}$ and $w = \sqrt{\frac{1-|\alpha'\beta|}{2}}$, yielding

$$\begin{aligned}\gamma &= \epsilon\frac{\sqrt{2|\alpha'\beta| + \frac{(1-|\alpha'\beta|)^2}{2}}}{\sqrt{2}(1+|\alpha'\beta|)}(\alpha + \text{sign}(\alpha'\beta)\beta) + \epsilon\sqrt{\frac{1-|\alpha'\beta|}{2}}\frac{\alpha - \text{sign}(\alpha'\beta)\beta}{\sqrt{2(1-|\alpha'\beta|)}} \\ &\qquad \text{where the term in } n \text{ is zero} \\ &= \frac{\epsilon}{2}(\alpha + \text{sign}(\alpha'\beta)\beta) + \frac{\epsilon}{2}(\alpha - \text{sign}(\alpha'\beta)\beta) \\ &= \epsilon\alpha.\end{aligned}$$

- If $\alpha'\beta = 0$, $(\beta'\gamma)^2$ is minimized for $\epsilon_a = \epsilon_b =: \epsilon \in \{-1, 1\}$ and any w, yielding

$$\begin{aligned}\gamma &= \frac{\epsilon}{\sqrt{2}}w(\alpha + \text{sign}(\alpha'\beta)\beta) + \frac{\epsilon}{\sqrt{2}}w(\alpha - \text{sign}(\alpha'\beta)\beta) + \sqrt{1-2w^2}n \\ &= \sqrt{2}\epsilon w\alpha + \sqrt{1-2w^2}n.\end{aligned}$$

  As $\epsilon \in \{-1, 1\}$, and $w \in [0, \frac{1}{\sqrt{2}}]$, we have that $\sqrt{2}\epsilon w\alpha$ spans the entire segment $[-\alpha, \alpha]$.

Conclusion. As $\epsilon_a = 1$ corresponds to the portion of $\mathcal{D}_{\alpha,\beta}$ with endpoints α and $\text{sign}(\alpha'\beta)\beta$, and $\epsilon_a = -1$ to that with endpoints $-\alpha$ and $-\text{sign}(\alpha'\beta)\beta$, we have that

- If $\alpha'\beta \neq 0$, $|\beta'\gamma|$ is decreasing when moving from $\text{sign}(\alpha'\beta)\beta$ to α on the portion of $\mathcal{D}_{\alpha,\beta}$ with endpoints α and $\text{sign}(\alpha'\beta)\beta$ or when moving from $-\text{sign}(\alpha'\beta)\beta$ to $-\alpha$ on the portion of $\mathcal{D}_{\alpha,\beta}$ with endpoints $-\alpha$ and $-\text{sign}(\alpha'\beta)\beta$; notably, $\gamma = \alpha$ and $\gamma = -\alpha$ are the only global minimizers of $|\beta'\gamma|$ on $\mathcal{D}_{\alpha,\beta}$.
- If $\alpha'\beta = 0$, the γ's whose projection on the span of α and β belongs to the segment $[-\alpha, \alpha]$ are the only global optimizers of $|\beta'\gamma|$ on $\mathcal{D}_{\alpha,\beta}$.

Thus, the proof of Theorem 4.5 consists in proving that $\mathcal{O}(\phi_\gamma(X))$ is a non-decreasing or increasing function of $|\beta'\gamma|$.

A.5.2 Proof of 1.(a), 1.(b), 2.(a)(i) and 2.(b)

This part of the proof is done by expressing the overlap divergence in terms of a specific function of $|\beta'\gamma|$ and using the dominated convergence theorem to differentiate this function; the derivative is then shown to be non-negative (or positive).

Expression of the Overlap Divergence in 1.(a) and 2.(a)(i). For simplicity, and without loss of generality, assume that $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}[h(Z)] = 1$. Also, remember that $\beta'X|\gamma'X \sim_{P^0} \mathcal{N}((\beta'\gamma)\gamma'X, 1-(\beta'\gamma)^2)$, so

$$\begin{aligned}\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma'X\right] &= \mathbb{E}_{Z\sim\mathcal{N}((\beta'\gamma)\gamma'X,1-(\beta'\gamma)^2)}\left[h(Z)\right] \\ &= \mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h((\beta'\gamma)\gamma'X + \sqrt{1-(\beta'\gamma)^2}Z')\right]\end{aligned}$$

and, as $\gamma'X \sim_{P^0} \mathcal{N}(0,1)$,

$$\begin{aligned}\mathcal{O}(\phi_\gamma(X)) &= \mathbb{E}_{P^0}\left[\left(\frac{dP^1_{\phi_\gamma(X)}}{dP^0_{\phi_\gamma(X)}}(\phi_\gamma(X))\right)^2\right] \\ &= \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\phi_\gamma(X)\right]^2\right] \text{ from Proposition 3.4 of Clivio et al. (2024)} \\ &= \mathbb{E}_{P^0}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma'X\right]^2\right] \\ &= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma'X = Z\right]^2\right] \\ &= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h((\beta'\gamma)Z + \sqrt{1-(\beta'\gamma)^2}Z')\right]^2\right] \\ &= g(\beta'\gamma)\end{aligned}$$

where $g(u) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(uZ + \sqrt{1-u^2}Z')\right]^2\right]$ for $u \in [-1,1]$.

Expression of the Overlap Divergence in 1.(b) and 2.(b). Again, $\beta'X|\gamma'X \sim_P \mathcal{N}((\beta'\gamma)\gamma'X, 1-(\beta'\gamma)^2)$, so

$$\begin{aligned}e(\phi_\gamma(X)) &= \mathbb{E}_P\left[e(X)|\phi_\gamma(X)\right] \\ &= \mathbb{E}_P\left[h(\beta'X)|\gamma'X\right] \\ &= \mathbb{E}_{Z\sim\mathcal{N}((\beta'\gamma)\gamma'X,1-(\beta'\gamma)^2)}\left[h(Z)\right] \\ &= \mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h((\beta'\gamma)\gamma'X + \sqrt{1-(\beta'\gamma)^2}Z')\right]\end{aligned}$$

and, as $\phi_\gamma(X) = \gamma' X \sim_P \mathcal{N}(0,1)$,

$$\begin{aligned}
\mathcal{O}(\phi_\gamma(X)) &= \mathbb{E}_{P^0}\left[\left(\frac{dP^1_{\phi_\gamma(X)}}{dP^0_{\phi_\gamma(X)}}(\phi_\gamma(X))\right)^2\right] \\
&= \mathbb{E}_{P^0}\left[\frac{(1-\pi_1)^2}{\pi_1^2}\frac{e(\phi_\gamma(X))^2}{(1-e(\phi_\gamma(X)))^2}\right] \\
&= \mathbb{E}_{P}\left[\frac{1-\pi_1}{\pi_1^2}\frac{e(\phi_\gamma(X))^2}{1-e(\phi_\gamma(X))}\right] \\
&= g(\beta'\gamma)
\end{aligned}$$

where $g(u) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\frac{1-\pi_1}{\pi_1^2}\frac{\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(uZ+\sqrt{1-u^2}Z')\right]^2}{1-\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(uZ+\sqrt{1-u^2}Z')\right]}\right]$ for $u \in [-1,1]$.

Factorization of 1.(a), 1.(b), 2.(a)(i) and 2.(b). Thus, in these four settings,

$$\mathcal{O}(\phi_\gamma(X)) = g(\beta'\gamma) \text{ with } g(u) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[f\left(\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right]$$

where $f(t) = f_a(t) := t^2$ for 1.(a) and 2.(a)(i), and $f(t) = f_b(t) := \frac{1-\pi_1}{\pi_1^2}\frac{t^2}{1-t}$ for 1.(b) and 2.(b).

First, note that g is symmetric as for any $u \in [-1,1]$,

$$\begin{aligned}
g(-u) &= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[f\left(\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h((-u)Z+\sqrt{1-(-u)^2}Z')\right]\right)\right] \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[f\left(\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(u(-Z)+\sqrt{1-u^2}Z')\right]\right)\right] \\
&= \mathbb{E}_{Z''\sim\mathcal{N}(0,1)}\left[f\left(\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(uZ''+\sqrt{1-u^2}Z')\right]\right)\right] \\
&\qquad \text{as if } Z\sim\mathcal{N}(0,1) \text{ then } -Z\sim\mathcal{N}(0,1) \\
&= g(u).
\end{aligned}$$

Then, for any $u \in [-1,1], g(u) = g(|u|)$. Thus, we need to show that the restriction of g to $[0,1]$ is non-decreasing, or increasing depending on whether we place ourselves in setups 1.(a) and 1.(b) or setups 2.(a)(i) and 2.(b). We show that it is non-decreasing (resp. increasing) on every interval $[0,u_0]$ where $u_0 \in [0,1)$; then it will be so on $[0,1)$, and the proof of 1.(a) and 1.(b) (resp. 2.(a)(i) and 2.(b)) is then complete if we show that $\forall u \in [0,1], \quad g(u) \leq g(1)$. Indeed, let $u \in [0,1]$,

$$\begin{aligned}
g(u) &= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[f\left(\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right] \\
&\leq \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{Z'\sim\mathcal{N}(0,1)}\left[f(h(uZ+\sqrt{1-u^2}Z'))\right]\right] \\
&\qquad \text{from Jensen's inequality, as } f \text{ is convex} \\
&= \mathbb{E}_{Z,Z'\overset{\text{indep}}{\sim}\mathcal{N}(0,1)}\left[f(h(uZ+\sqrt{1-u^2}Z'))\right] \text{ where } uZ+\sqrt{1-u^2}Z' \sim \mathcal{N}(0,1) \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[f(h(Z))\right] \\
&= g(1)
\end{aligned}$$

which then completes the proof. We now fix $u_0 \in [0,1)$. In the following, unless specified otherwise, $Z, Z' \overset{\text{indep}}{\sim} \mathcal{N}(0,1)$.

Preliminary Bounds. First we prove that for any μ, $\sigma > 0$ and $\bar{\lambda} < \frac{1}{2\sigma^2}$,

$$\mathbb{E}_Z\left[|Z|e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] \leq e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}\left(\frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}} + \frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{4\bar{\lambda}^2\sigma^2\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right).$$

Indeed, as $\forall z, |z| \leq 1 + z^2$, we have

$$\begin{aligned}\mathbb{E}_Z\left[|Z|e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] &\leq \mathbb{E}_Z\left[\left(1+Z^2\right)e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] \\ &= \mathbb{E}_Z\left[e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] + \mathbb{E}_Z\left[Z^2 e^{\bar{\lambda}(\mu+\sigma Z)^2}\right].\end{aligned}$$

From the MGF of an uncentered χ^2 distribution,

$$\mathbb{E}_Z\left[e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] = \frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}} e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}.$$

Then,

$$\begin{aligned}\mathbb{E}_Z\left[Z^2 e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] = \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\left(\frac{X-\mu}{\sigma}\right)^2 e^{\bar{\lambda}X^2}\right] = \frac{1}{\sigma^2}\Big(&\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[X^2 e^{\bar{\lambda}X^2}\right] - 2\mu\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[Xe^{\bar{\lambda}X^2}\right] \\ &+ \mu^2\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[e^{\bar{\lambda}X^2}\right]\Big)\end{aligned}$$

where, again,

$$\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[e^{\bar{\lambda}X^2}\right] = \frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}} e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}$$

and, when $\bar{\lambda} \neq 0$ (the following result trivially holds for $\bar{\lambda} = 0$),

$$\begin{aligned}\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[Xe^{\bar{\lambda}X^2}\right] &= \mathbb{E}_Z\left[(\mu+\sigma Z)e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] \\ &= \frac{1}{2\bar{\lambda}}\mathbb{E}_Z\left[\frac{\partial}{\partial\mu}\left(e^{\bar{\lambda}(\mu+\sigma Z)^2}\right)\right] \\ &= \frac{1}{2\bar{\lambda}}\frac{\partial}{\partial\mu}\left(\mathbb{E}_Z\left[e^{\bar{\lambda}(\mu+\sigma Z)^2}\right]\right) \\ &= \frac{1}{2\bar{\lambda}}\frac{\partial}{\partial\mu}\left(\frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}} e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}\right) \\ &= \frac{\mu}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}\end{aligned}$$

where differentiation and expectation can be exchanged thanks to the dominated convergence theorem, restricting μ to $[-M, M]$ for $M > 0$ so $\frac{\partial}{\partial\mu}\left(e^{\bar{\lambda}(\mu+\sigma Z)^2}\right)$ is bounded by $2\bar{\lambda}(M + \sigma|Z|)e^{\bar{\lambda}(M+\sigma|Z|)^2}$ which is integrable, and finally,

$$\begin{aligned}\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[X^2 e^{\bar{\lambda}X^2}\right] &= \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\frac{\partial}{\partial\bar{\lambda}}\left(e^{\bar{\lambda}X^2}\right)\right] \\ &= \frac{\partial}{\partial\bar{\lambda}}\left(\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[e^{\bar{\lambda}X^2}\right]\right) \\ &= \frac{\partial}{\partial\bar{\lambda}}\left(\frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}} e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}\right) \\ &= \left(\frac{\sigma^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right)\cdot e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}\end{aligned}$$

where differentiation and expectation can be exchanged thanks to the dominated convergence theorem, restricting $\bar{\lambda}$ to $(-\infty, \bar{\lambda}^\star]$ for $\bar{\lambda}^\star < \frac{1}{2\sigma^2}$ so $\frac{\partial}{\partial\bar{\lambda}}\left(e^{\bar{\lambda}X^2}\right)$ is bounded by $X^2 e^{\bar{\lambda}^\star X^2}$ which is integrable. In the end,

$$\mathbb{E}_Z\left[Z^2 e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] = \frac{1}{\sigma^2}\Big(\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[X^2 e^{\bar{\lambda}X^2}\right] - 2\mu\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[Xe^{\bar{\lambda}X^2}\right]$$

$$
\begin{aligned}
&\quad + \mu^2 \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[e^{\bar{\lambda}X^2}\right]\Big) \\
&= \Big(\frac{\sigma^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5} - \frac{2\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} \\
&\quad + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}}\Big)\cdot \frac{e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}}{\sigma^2} \\
&= \left(\frac{\sigma^2-2\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5} + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}}\right)\cdot \frac{e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}}{\sigma^2} \\
&= \left(\frac{\sigma^2-\mu^2-2\bar{\lambda}\sigma^2\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right)\cdot \frac{e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}}{\sigma^2} \\
&= \left(\frac{\left(\sigma^2-\mu^2-2\bar{\lambda}\sigma^2\mu^2\right)\left(1-2\bar{\lambda}\sigma^2\right)}{\sqrt{1-2\bar{\lambda}\sigma^2}^5} + \frac{\mu^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right)\cdot \frac{e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}}{\sigma^2} \\
&= \left(\frac{\sigma^2\left(1-2\bar{\lambda}\sigma^2+4\bar{\lambda}^2\mu^2\sigma^2\right)}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right)\cdot \frac{e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}}{\sigma^2} \\
&= \left(\frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{4\bar{\lambda}^2\mu^2\sigma^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right)\cdot e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}
\end{aligned}
$$

so the final bound is

$$
\begin{aligned}
\mathbb{E}_Z\left[|Z|e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] &\le \mathbb{E}_Z\left[e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] + \mathbb{E}_Z\left[Z^2 e^{\bar{\lambda}(\mu+\sigma Z)^2}\right] \\
&= \left(\frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}} + \frac{1}{\sqrt{1-2\bar{\lambda}\sigma^2}^3} + \frac{4\bar{\lambda}^2\mu^2\sigma^2}{\sqrt{1-2\bar{\lambda}\sigma^2}^5}\right)\cdot e^{\frac{\bar{\lambda}\mu^2}{1-2\bar{\lambda}\sigma^2}}.
\end{aligned}
$$

Further, for $u \in [0, u_0] \subset [0,1)$ and $z \in \mathbb{R}$, assume $\mu = uz$, $\sigma^2 = 1-u^2$, $0 \le \lambda < \frac{1}{4}$, $0 \le \lambda' < \frac{1-4\lambda}{6}$. Then,

$$
e^{\frac{\lambda\mu^2}{1-2\lambda\sigma^2}} = e^{\frac{\lambda u^2 z^2}{1-2\lambda(1-u^2)}} \le e^{\lambda z^2}
$$

which follows from the function $u \to \frac{\lambda u^2 z^2}{1-2\lambda(1-u^2)}$ which is non-decreasing when $0 \le \lambda \le 1/2$, thus bounded above by λz^2. Further, $\frac{1}{\sqrt{1-2\lambda\sigma^2}} \le \sqrt{2}$. In the end,

$$
\begin{aligned}
\mathbb{E}_Z\left[|Z|e^{\lambda(\mu+\sigma Z)^2}\right] &= \mathbb{E}_Z\left[|Z|e^{\lambda(uz+\sqrt{1-u^2}Z)^2}\right] \le \sqrt{2}(3+16\lambda^2 z^2)\cdot e^{\lambda z^2}, \\
\mathbb{E}_Z\left[e^{\lambda(\mu+\sigma Z)^2}\right] &= \mathbb{E}_Z\left[e^{\lambda(uz+\sqrt{1-u^2}Z)^2}\right] \le \sqrt{2}\cdot e^{\lambda z^2}.
\end{aligned}
$$

Thus, as $\mathcal{H}'^2_{C,\lambda,\lambda'} \subset \mathcal{H}^2_{C,\lambda}$, for any $k = 0, 1, 2$,

$$
\begin{aligned}
\mathbb{E}_Z\left[\left|Zh^{(k)}(uz+\sqrt{1-u^2}Z)\right|\right] &\le \mathbb{E}_Z\left[C|Z|e^{\lambda(uz+\sqrt{1-u^2}Z)^2}\right] \le \sqrt{2}(3+16\lambda^2 z^2)C\cdot e^{\lambda z^2}, \\
\mathbb{E}_Z\left[\left|h^{(k)}(uz+\sqrt{1-u^2}Z)\right|\right] &\le \mathbb{E}_Z\left[Ce^{\lambda(uz+\sqrt{1-u^2}Z)^2}\right] \le \sqrt{2}C\cdot e^{\lambda z^2}.
\end{aligned}
$$

Further, if $h \in \mathcal{H}'^2_{C,\lambda,\lambda'}$,

$$
\begin{aligned}
\frac{1}{1-\mathbb{E}_Z\left[h(uz+\sqrt{1-u^2}Z)\right]} &= \frac{1}{\mathbb{E}_Z\left[(1-h)(uz+\sqrt{1-u^2}Z)\right]} \\
&\le \frac{1}{\mathbb{E}_Z\left[\frac{1}{C}e^{-\lambda'(uz+\sqrt{1-u^2}Z)^2}\right]} \\
&= C\sqrt{1+2\lambda'(1-u^2)}e^{\lambda'\frac{u^2z^2}{1+2\lambda'(1-u^2)}}
\end{aligned}
$$

$$\leq \frac{2}{\sqrt{3}} C e^{\lambda' z^2}$$

as $u \mapsto \lambda' \frac{u^2 z^2}{1+2\lambda'(1-u^2)}$ is non-decreasing on $[0,1)$. As $f_b'(t) = \frac{1-\pi_1}{\pi_1^2} \frac{t(2-t)}{(1-t)^2}$, with $|f_b'(t)| \leq \frac{1-\pi_1}{\pi_1^2} \frac{1}{(1-t)^2}$ for $0 \leq t \leq 1$, and $f_b''(t) = \frac{1-\pi_1}{\pi_1^2} \frac{2}{(1-t)^3}$, this yields

$$\left| f_b' \left(\mathbb{E}_Z \left[h(uz + \sqrt{1-u^2} Z) \right] \right) \right| \leq \frac{4}{3} C^2 \frac{1-\pi_1}{\pi_1^2} e^{2\lambda' z^2},$$
$$\left| f_b'' \left(\mathbb{E}_Z \left[h(uz + \sqrt{1-u^2} Z) \right] \right) \right| \leq \frac{16}{\sqrt{3}^3} C^3 \frac{1-\pi_1}{\pi_1^2} e^{3\lambda' z^2},$$

together with the alternative $h \in \mathcal{H}^2_{C,\lambda}$ and $f_a'(t) = 2t$, $f_a''(t) = 2$, this gives

$$\left| f' \left(\mathbb{E}_Z \left[h(uz + \sqrt{1-u^2} Z) \right] \right) \right| \leq \frac{4}{3} C^2 \frac{1-\pi_1}{\pi_1^2} e^{2\lambda' z^2} + 2\sqrt{2} C e^{\lambda z^2},$$
$$\left| f'' \left(\mathbb{E}_Z \left[h(uz + \sqrt{1-u^2} Z) \right] \right) \right| \leq \frac{16}{\sqrt{3}^3} C^3 \frac{1-\pi_1}{\pi_1^2} e^{3\lambda' z^2} + 2.$$

Proof of 1.(a) and 1.(b). First, let us show that for any z,

$$\frac{\partial}{\partial u} \left(\mathbb{E}_{Z'} \left[h(uz + \sqrt{1-u^2} Z') \right] \right) = \mathbb{E}_{Z'} \left[\left(z - \frac{u}{\sqrt{1-u^2}} Z' \right) h'(uz + \sqrt{1-u^2} Z') \right].$$

Indeed

$$\frac{\partial}{\partial u} h(uz + \sqrt{1-u^2} z') = \left(z - \frac{u}{\sqrt{1-u^2}} z' \right) h'(uz + \sqrt{1-u^2} z')$$

which is bounded by $C \cdot (|z| + \frac{|z'|}{\sqrt{1-u_0^2}}) \cdot e^{\lambda(uz + \sqrt{1-u^2} z')^2}$, which is integrable when replacing z' with Z'. Thus, the dominated convergence theorem applies and

$$\begin{aligned} \frac{\partial}{\partial u} \left(\mathbb{E}_{Z'} \left[h(uz + \sqrt{1-u^2} Z') \right] \right) &= \mathbb{E}_{Z'} \left[\frac{\partial}{\partial u} \left(h(uz + \sqrt{1-u^2} Z') \right) \right] \\ &= \mathbb{E}_{Z'} \left[\left(z - \frac{u}{\sqrt{1-u^2}} Z' \right) h'(uz + \sqrt{1-u^2} Z') \right]. \end{aligned}$$

Then, we show that,

$$\begin{aligned} &g'(u) \\ &= \mathbb{E}_Z \left[\mathbb{E}_{Z'} \left[\left(Z - \frac{u}{\sqrt{1-u^2}} Z' \right) h'(uZ + \sqrt{1-u^2} Z') \right] \cdot f' \left(\mathbb{E}_{Z'} \left[h(uZ + \sqrt{1-u^2} Z') \right] \right) \right]. \end{aligned}$$

Indeed, for any z,

$$\begin{aligned} &\frac{\partial}{\partial u} f \left(\mathbb{E}_{Z'} \left[h(uz + \sqrt{1-u^2} Z') \right] \right) \\ &= \frac{\partial}{\partial u} \left(\mathbb{E}_{Z'} \left[h(uz + \sqrt{1-u^2} Z') \right] \right) \cdot f' \left(\mathbb{E}_{Z'} \left[h(uz + \sqrt{1-u^2} Z') \right] \right) \\ &= \mathbb{E}_{Z'} \left[\left(z - \frac{u}{\sqrt{1-u^2}} Z' \right) h'(uz + \sqrt{1-u^2} Z') \right] \cdot f' \left(\mathbb{E}_{Z'} \left[h(uz + \sqrt{1-u^2} Z') \right] \right) \end{aligned}$$

which is bounded by

$$\mathbb{E}_{Z'}\left[\left(|z|+\frac{1}{\sqrt{1-u_0^2}}|Z'|\right)\cdot Ce^{\lambda(uz+\sqrt{1-u^2}Z')^2}\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)$$

from the assumptions

$$\leq C\left(|z|\sqrt{2}e^{\lambda z^2}+\frac{\sqrt{2}}{\sqrt{1-u_0^2}}(3+16\lambda^2z^2)e^{\lambda z^2}\right)\cdot\left(\frac{4}{3}C^2\frac{1-\pi_1}{\pi_1^2}e^{2\lambda' z^2}+2\sqrt{2}Ce^{\lambda z^2}\right)$$

from the preliminary bounds,

which is integrable when replacing z with Z; the dominated convergence theorem gives the result for $g'(u)$. We simplify it further and show it is non-negative. Indeed, $g'(u)$ can be decomposed as

$$\begin{aligned}g'(u) =&\mathbb{E}_Z\left[Z\cdot\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right]\\&-\mathbb{E}_Z\left[\frac{u}{\sqrt{1-u^2}}\cdot\mathbb{E}_{Z'}\left[Z'h'(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right].\end{aligned}$$

Then, note that from Stein's lemma,

$$\forall z,\mathbb{E}_{Z'}\left[Z'h'(uz+\sqrt{1-u^2}Z')\right]=\sqrt{1-u^2}\cdot\mathbb{E}_{Z'}\left[h''(uz+\sqrt{1-u^2}Z')\right].$$

So the second term of the decomposition of $g'(u)$ is equal to

$$-\mathbb{E}_Z\left[u\cdot\mathbb{E}_{Z'}\left[h''(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right].$$

We now turn to the first term of the decomposition. It is equal to

$$\begin{aligned}&\mathbb{E}_Z\left[\lim_{R\to\infty}1_{|Z|\leq R}\cdot Z\cdot\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right]\\&=\lim_{R\to\infty}\mathbb{E}_Z\left[1_{|Z|\leq R}\cdot Z\cdot\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right]\end{aligned}$$

from the dominated convergence theorem as the integrand of the RHS is bounded by $|Z|C\sqrt{2}\cdot e^{\lambda Z^2}\cdot\left(\frac{4}{3}C^2\frac{1-\pi_1}{\pi_1^2}e^{2\lambda' Z^2}+2\sqrt{2}Ce^{\lambda Z^2}\right)$ which is integrable. Further, we have that

$$\begin{aligned}\forall k=0,1,\ \frac{\partial}{\partial z}\mathbb{E}_{Z'}\left[h^{(k)}(uz+\sqrt{1-u^2}Z')\right]&=\mathbb{E}_{Z'}\left[\frac{\partial}{\partial z}h^{(k)}(uz+\sqrt{1-u^2}Z')\right]\\&=\mathbb{E}_{Z'}\left[uh^{(k+1)}(uz+\sqrt{1-u^2}Z')\right]\end{aligned}$$

from the dominated convergence theorem as the integrand in the RHS is bounded by $Cue^{\lambda(uR+\sqrt{1-u^2}|Z'|)^2}$ which is integrable. Note that this is where the use of $1_{|Z|\leq R}$ is needed ; otherwise the integrand in the RHS could not be uniformly (wrt z) bounded by an integrable function! Thus, we obtain that

$$\begin{aligned}&\frac{\partial}{\partial z}\left(\mathbb{E}_{Z'}\left[h'(uz+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\right)\\&=u\cdot\Bigg(\mathbb{E}_{Z'}\left[h''(uz+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\\&\qquad+\mathbb{E}_{Z'}\left[h'(uz+\sqrt{1-u^2}Z')\right]^2 f''\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\Bigg).\end{aligned}$$

Then, integration by parts gives a slight variation of Stein's lemma, as

$$\begin{aligned}&\mathbb{E}_Z\left[1_{|Z|\leq R}\cdot Z\cdot\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right]\\&=\left[-\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\mathbb{E}_{Z'}\left[h'(uz+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\right]_{z=-R}^{z=R}\end{aligned}$$

$$+\mathbb{E}_Z\Big[1_{|Z|\leq R}\cdot u\cdot\Big(\mathbb{E}_{Z'}\left[h''(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)$$
$$+\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]^2 f''\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\Big)\Big].$$

The function of z in the $[\dots]_{z=-R}^{z=R}$ brackets is bounded by $\frac{C}{\sqrt{\pi}}e^{\left(\lambda-\frac{1}{2}\right)z^2}\left(\frac{4}{3}C^2\frac{1-\pi_1}{\pi_1^2}e^{2\lambda' z^2}+2\sqrt{2}Ce^{\lambda z^2}\right)$ which goes to 0 as $z\to\pm\infty$, so these $[\dots]_{z=-R}^{z=R}$ brackets vanish when $R\to\infty$. For the expectation wrt Z, note that the integrand is bounded by

$$uC\sqrt{2}e^{\lambda Z^2}\left(\frac{4}{3}C^2\frac{1-\pi_1}{\pi_1^2}e^{2\lambda' Z^2}+2\sqrt{2}Ce^{2\lambda Z^2}\right)+uC^2\cdot 2e^{2\lambda Z^2}\left(\frac{16}{\sqrt{3}^3}C^3\frac{1-\pi_1}{\pi_1^2}e^{3\lambda' Z^2}+2\right)$$

which is integrable, so the dominated convergence theorem applies and

$$\begin{aligned}
&\lim_{R\to\infty}\mathbb{E}_Z\Big[1_{|Z|\leq R}\cdot u\cdot\Big(\mathbb{E}_{Z'}\left[h''(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\\
&\qquad+\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]^2 f''\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\Big)\Big]\\
&=\mathbb{E}_Z\Big[u\cdot\Big(\mathbb{E}_{Z'}\left[h''(uZ+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\\
&\qquad+\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]^2 f''\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\Big)\Big],
\end{aligned}$$

and the first term of the decomposition of $g'(u)$ is equal to the RHS just above. Thus, summing the first and second terms of the decomposition, the terms involving h'' cancel out and we obtain

$$g'(u)=u\cdot\mathbb{E}_Z\left[\mathbb{E}_{Z'}\left[h'(uZ+\sqrt{1-u^2}Z')\right]^2 f''\left(\mathbb{E}_{Z'}\left[h(uZ+\sqrt{1-u^2}Z')\right]\right)\right]$$

which is non-negative, concluding the proof for this part.

Proof of 2.(a)(i). First note that when $f=f_a$, the above computations can be repeated by replacing h with $h^{(k)}$ for $k=0,\dots,K$ to show that, noting $\forall k, J_k(u):=\mathbb{E}_Z\left[\mathbb{E}_{Z'}\left[h^{(k)}(uZ+\sqrt{1-u^2}Z')\right]^2\right]$, every J_k for $k=0,\dots,K$ is differentiable and, most importantly,

$$\forall k=0,\dots,K-1,\ (J_k)'(u)=2u\cdot J_{k+1}(u).$$

Thus, starting from $g(u)=J_0(u)$, we obtain by recursion that g is K times differentiable and for any $k=0,\dots,K$, $g^{(k)}(u)=\sum_{i=0}^k p_i^k(u)J_i(u)$ where each $p_i^k(u)$ is a polynomial function of degree at most i with non-negative coefficients, and each $p_k^k(u)$ further has a degree exactly k. We note that every J_i and every p_i^k is non-negative, and so is every $p_i^k(.)J_i(.)$, thus every $g^{(k)}$.

Further, $J_K(0)>0$ since $\mathbb{E}_{Z'}\left[h^{(K)}(Z')\right]\neq 0$ by assumption and J_K is continuous as it is differentiable ; thereby J_K is positive on an interval of the form $[0,\epsilon]$ where $\epsilon\in(0,u_0]$. Notably, since p_K^K is of degree exactly K with non-negative coefficients, it is positive on $(0,u_0]$. As a result, $p_K^K(.)J_K(.)$ is positive on $(0,\epsilon]$, thus so is $g^{(K)}$, with $g^{(K)}(0)\geq 0$. This yields $g^{(K-1)}$ being increasing on $[0,\epsilon]$. Remember that it is also non-decreasing on $[\epsilon,u_0]$, since $g^{(K)}$ is generally non-negative. Since $g^{(K-1)}(0)\geq 0$ by non-negativity of $g^{(K-1)}$, we obtain that $\forall u\in(0,\epsilon],\ g^{(K-1)}(u)>g^{(K-1)}(0)\geq 0$ and $\forall u\in[\epsilon,u_0],\ g^{(K-1)}(u)\geq g^{(K-1)}(\epsilon)>g^{(K-1)}(0)\geq 0$, showing that $\forall u\in(0,u_0],\ g^{(K-1)}(u)>0$ and $g^{(K-1)}(0)\geq 0$. Thus, by recursion we obtain for every $k=K-1,\dots,1$, $g^{(k)}$ is generally non-negative while being positive on $(0,u_0]$. When $k=1$, this yields that g is an increasing function on $[0,u_0]$, concluding the proof.

Proof of 2.(b). When $f = f_b$, we still have $f'' \geq 2\frac{1-\pi_1}{\pi_1^2}$ so $g'(u) \geq 2u\frac{1-\pi_1}{\pi_1^2}J_1(u)$, where the RHS is, up to a positive constant, the derivative of J_0. Thus, as $h \in \mathcal{H}'^{K+1}_{C,\lambda,\lambda'} \subset \mathcal{H}^{K+1}_{C,\lambda}$ the proof of 2.(a)(i) can be repeated to show that the derivative of J_0 is generally non-negative while being positive on $(0, u_0]$; it immediately follows that g' is also generally non-negative while being positive on $(0, u_0]$. Thus g is an increasing function on $[0, u_0]$. This concludes the proof of 2.(b).

A.5.3 Proof of 2.(a)(ii)

When $h(z) = 1_{\{z \leq z_0\}}$, the normalizing constant is $\mathbb{E}_Z[h(Z)] = \Phi(z_0)$, where Φ is the CDF of the centered standard Gaussian distribution, and $\frac{dP_X^1}{dP_X^0}(x) = \frac{1_{\{\beta' x \leq z_0\}}}{\Phi(z_0)}$. Then,

$$\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma' X\right] = \mathbb{E}_{Z\sim\mathcal{N}((\beta'\gamma)\gamma' X, 1-(\beta'\gamma)^2)}\left[\frac{1_{\{Z\leq z_0\}}}{\Phi(z_0)}\right] = \frac{\Phi\left(\frac{z_0-(\beta'\gamma)\gamma' X}{\sqrt{1-(\beta'\gamma)^2}}\right)}{\Phi(z_0)}.$$

From Formula 20,010.4 of Owen (1980), we have

$$\forall a, b, \quad \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\Phi(a+bZ)^2\right] = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) - 2T\left(\frac{a}{\sqrt{1+b^2}}, \frac{1}{\sqrt{1+2b^2}}\right)$$

where $T(.,.)$ is Owen's T function. Thus, as $\mathcal{O}(\phi_\gamma(X)) = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma' X = Z\right]^2\right]$ from the above, it is equal, up to a $\frac{1}{\Phi(z_0)^2}$ constant, to the RHS of this equation for $a = \frac{z_0}{\sqrt{1-(\beta'\gamma)^2}}$, $b = \frac{-\beta'\gamma}{\sqrt{1-(\beta'\gamma)^2}}$, leading to

$$\frac{a}{\sqrt{1+b^2}} = \frac{\frac{z_0}{\sqrt{1-(\beta'\gamma)^2}}}{\sqrt{1+\frac{(\beta'\gamma)^2}{1-(\beta'\gamma)^2}}} = z_0,$$
$$\frac{1}{\sqrt{1+2b^2}} = \frac{1}{\sqrt{1+2\frac{(\beta'\gamma)^2}{1-(\beta'\gamma)^2}}} = \sqrt{\frac{1-(\beta'\gamma)^2}{1+(\beta'\gamma)^2}} = \sqrt{\frac{2}{1+(\beta'\gamma)^2}-1}.$$

Thus, in the end,

$$\mathcal{O}(\phi_\gamma(X)) = \frac{\Phi(z_0) - 2T\left(z_0, \sqrt{\frac{2}{1+(\beta'\gamma)^2}-1}\right)}{\Phi(z_0)^2}.$$

As T is increasing in its second argument (Owen, 1980), we obtain that $\mathcal{O}(\phi_\gamma(X))$ is increasing in $|\beta'\gamma|$.

A.5.4 Proof of 2.(a)(iii)

First, noting $\varphi(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$, for any μ and any $\sigma > 0$,

$$\begin{aligned}
\mathbb{E}_{Z\sim\mathcal{N}(\mu,\sigma^2)}[\max(0,Z)] &= \int_0^\infty z\,\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)dz \\
&= \int_{-\mu/\sigma}^\infty (\sigma z' + \mu)\frac{1}{\sqrt{2\pi}}e^{-z'^2/2}dz' \\
&= \sigma\int_{-\mu/\sigma}^\infty z'\,\frac{1}{\sqrt{2\pi}}e^{-z'^2/2}dz' + \mu\int_{-\mu/\sigma}^\infty \frac{1}{\sqrt{2\pi}}e^{-z'^2/2}dz' \\
&= \sigma\,\varphi\left(-\frac{\mu}{\sigma}\right) + \mu\left[1-\Phi\left(-\frac{\mu}{\sigma}\right)\right] \\
&= \sigma\,\varphi\left(\frac{\mu}{\sigma}\right) + \mu\,\Phi\left(\frac{\mu}{\sigma}\right).
\end{aligned}$$

So the normalizing constant $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[h(Z)\right] = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\max(0,Z)\right] = \varphi(0) = \frac{1}{\sqrt{2\pi}}$, so $\frac{dP_X^1}{dP_X^0}(x) = \sqrt{2\pi}\max(0,\beta' x)$. Further,

$$\begin{aligned}
&\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma' X\right] \\
&= \sqrt{2\pi}\mathbb{E}_{Z\sim\mathcal{N}((\beta'\gamma)\gamma' X, 1-(\beta'\gamma)^2)}[\max(0,Z)] \\
&= \sqrt{2\pi}\left(\sqrt{1-(\beta'\gamma)^2}\varphi\left(\frac{(\beta'\gamma)\gamma' X}{\sqrt{1-(\beta'\gamma)^2}}\right) + (\beta'\gamma)\gamma' X\Phi\left(\frac{(\beta'\gamma)\gamma' X}{\sqrt{1-(\beta'\gamma)^2}}\right)\right).
\end{aligned}$$

Thus, noting $u = \beta'\gamma$, and $a = \frac{u}{\sqrt{1-u^2}}$,

$$\begin{aligned}
\mathcal{O}(\phi_\gamma(X)) &= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{P^0}\left[\frac{dP_X^1}{dP_X^0}(X)\middle|\gamma' X = Z\right]^2\right] \\
&= 2\pi\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\left(\sqrt{1-u^2}\varphi\left(\frac{uZ}{\sqrt{1-u^2}}\right) + uZ\Phi\left(\frac{uZ}{\sqrt{1-u^2}}\right)\right)^2\right] \\
&= 2\pi(1-u^2)\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\varphi(aZ)^2\right] \\
&\qquad + 4\pi u\sqrt{1-u^2}\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z\varphi(aZ)\Phi(aZ)\right] \\
&\qquad + 2\pi u^2\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z^2\Phi(aZ)^2\right]
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\varphi(aZ)^2\right] &= \int \frac{1}{2\pi}e^{-a^2z^2}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \\
&= \frac{1}{\sqrt{2\pi}^3}\int e^{-(2a^2+1)z^2/2}dz \\
&= \frac{1}{2\pi\sqrt{2a^2+1}} \\
&= \frac{1}{2\pi\sqrt{\frac{2u^2}{1-u^2}+1}} \\
&= \frac{1}{2\pi}\sqrt{\frac{1-u^2}{1+u^2}}
\end{aligned}$$

and, from Stein's lemma,

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z\varphi(aZ)\Phi(aZ)\right] \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[a\varphi'(aZ)\Phi(aZ) + a\varphi(aZ)\Phi'(aZ)\right] \\
&= -a^2\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z\varphi(aZ)\Phi(aZ)\right] + a\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\varphi(aZ)^2\right]
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z\varphi(aZ)\Phi(aZ)\right] &= \frac{a}{1+a^2}\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\varphi(aZ)^2\right] \\
&= \frac{\frac{u}{\sqrt{1-u^2}}}{1+\frac{u^2}{1-u^2}}\frac{1}{2\pi}\sqrt{\frac{1-u^2}{1+u^2}} \\
&= \frac{u}{2\pi}\frac{1-u^2}{\sqrt{1+u^2}}.
\end{aligned}$$

Thus, again from Stein's lemma, and from Formula 2,010.3 of Owen (1980),

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z^2\Phi(aZ)^2\right] \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\Phi(aZ)^2 + 2aZ\Phi'(aZ)\Phi(aZ)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\Phi(aZ)^2\right] + 2a\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z\varphi(aZ)\Phi(aZ)\right] \\
&= \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{a^2}{1+a^2}\right) + 2a\frac{u}{2\pi}\frac{1-u^2}{\sqrt{1+u^2}} \\
&= \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{\frac{u^2}{1-u^2}}{1+\frac{u^2}{1-u^2}}\right) + 2\frac{u}{\sqrt{1-u^2}}\frac{u}{2\pi}\frac{1-u^2}{\sqrt{1+u^2}} \\
&= \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(u^2\right) + \frac{u^2}{\pi}\sqrt{\frac{1-u^2}{1+u^2}}.
\end{aligned}
$$

In the end,

$$
\begin{aligned}
&\mathcal{O}(\phi_\gamma(X)) \\
&= 2\pi(1-u^2)\frac{1}{2\pi}\sqrt{\frac{1-u^2}{1+u^2}} + 4\pi u\sqrt{1-u^2}\frac{u}{2\pi}\frac{1-u^2}{\sqrt{1+u^2}} \\
&\qquad + 2\pi u^2\left(\frac{1}{4} + \frac{1}{2\pi}\arcsin\left(u^2\right) + \frac{u^2}{\pi}\sqrt{\frac{1-u^2}{1+u^2}}\right) \\
&= \frac{\sqrt{1-u^2}^3}{\sqrt{1+u^2}} + \frac{2u^2\sqrt{1-u^2}^3}{\sqrt{1+u^2}} + \frac{\pi}{2}u^2 + u^2\arcsin(u^2) + \frac{2u^4\sqrt{1-u^2}}{\sqrt{1+u^2}} \\
&= \sqrt{\frac{1-u^2}{1+u^2}}\cdot\left(1-u^2+2u^2(1-u^2)+2u^4\right) + \frac{\pi}{2}u^2 + u^2\arcsin(u^2) \\
&= \sqrt{\frac{1-u^2}{1+u^2}}\cdot\left(1+u^2\right) + \frac{\pi}{2}u^2 + u^2\arcsin(u^2) \\
&= \sqrt{1-w^2} + \frac{\pi}{2}w + w\arcsin(w)
\end{aligned}
$$

where $w := u^2 = (\beta'\gamma)^2$, and $\mathcal{O}(\phi_\gamma(X))$ is an increasing function of w as

$$
\frac{\partial}{\partial w}\mathcal{O}(\phi_\gamma(X)) = \frac{-2w}{2\sqrt{1-w^2}} + \frac{\pi}{2} + \arcsin(w) + \frac{w}{\sqrt{1-w^2}} = \frac{\pi}{2} + \arcsin(w)
$$

which is positive as $w \geq 0$. This concludes the proof.

B FURTHER DISCUSSION ON PREVIOUS WORK AND THEORETICAL RESULTS

B.1 Comparison with Approaches Inspired by Domain Adaptation

When comparing our approach to the related works inspired by domain adaptation, as in Shalit et al. (2017), Johansson et al. (2022) or Zhang et al. (2020), one subtlety is that the appropriate metric for measuring overlap may vary by (1) the target of estimation, and (2) the parametric assumptions one is willing to impose on the outcome process. In our approach, we target the scalar ATT, while the aforementioned related works target the Conditional Average Treatment Effect (CATE), which is a whole function. As a result, while our work leverages the semiparametric efficiency bound of the ATT as a scalar objective, the latter cannot readily be applied to the CATE. Instead, the related works construct bounds under different sets of assumptions. Each of the related works decomposes the MSE of the CATE into a factual, observable part and a counterfactual, unobservable part. The unobservable part is then further upper-bounded by a quantity that measures overlap, but the particular overlap metric depends on parametric restrictions on the unknown counterfactual outcome model. Shalit et al. (2017) and Johansson et al. (2022) assume that the pointwise loss of the counterfactual regressor lives in a restricted function class, so use an IPM to bound the counterfactual loss, while Zhang et al. (2020) assume a posterior measure over the counterfactual regressor, which makes the posterior variance relevant for PAC-Bayes bounds.

Further, note that some works such as Assaad et al. (2021) and Johansson et al. (2022) incorporate inverse probability weights in both the observable part of the MSE of the CATE and the quantity measuring overlap. This can improve the estimation of the CATE as Assaad et al. (2021) shows that introducing such weights reduces the contradiction between the minimizations of these two components, while Johansson et al. (2022) notes that the resulting reweighted overlap measure should be lower than the original one. However, this approach is not directly applicable to our setup where we aim to find representations with better overlap *in the original distribution*. Indeed, the reweighted overlap measure measures the overlap between treated and control representation *reweighted* distributions and minimizing it does not encourage learning representations with better overlap in the original distribution, especially as the weights estimate true inverse propensity weights which render treated and control distributions of both original covariates and *any* representation identical by design.

B.2 Lemma 3.1

Lemma 3.1 gives a "doubly robust" moment that identifies deconfounding scores. This is analogous to the "mixed bias property" in automatic debiased machine learning (AutoDML) for sensitivity analysis (Chernozhukov et al., 2022b), where the bias of the estimator is the correlation of the residuals from an outcome model regression and a Riesz representer regression. In our setting, each of these regressions would be the perfect regression of its corresponding ground-truth model on the representation $\phi(X)$. Chernozhukov et al. (2022a) established a mixed bias property for AutoDML in sensitivity analysis; our Lemma 3.1 is a special case for the ATT and adapted to our broader setup.

B.3 Lemma 3.2

Lemma 3.2 shows that overlap divergence is a proxy for the semiparametric efficiency bound of the estimand obtained by adjusting on the representation; this motivates our specific choice of the χ^2-divergence.

In general, the assumptions on outcomes in Items 1 and 2 are relatively mild and commonly invoked in the literature: the outcome is binary in many applications, e.g. medicine (Colnet et al., 2023), and $\mathrm{Var}_{P^0}(Y|X=x)$ is constant in several popular treatment effect estimation datasets such as IHDP (Hill, 2011) or News (Johansson et al., 2016).

B.4 Theorem 4.4

When $\alpha'\beta < 0$, the deconfounding score with coordinate w_1 on the segment with prognostic score endpoint α in $\mathcal{D}_{\alpha,\beta}$ is equal to the deconfounding score with coordinate w_2 on the segment with prognostic score endpoint α in $\mathcal{D}_{\alpha,-\beta}$. The same statement can be made when replacing α with $-\alpha$ as the prognostic score endpoint. More generally, regardless of the sign of $\alpha'\beta$, Assumptions 4.2 and 4.3 can be expressed equivalently when replacing β with $-\beta$ and h with $h(-.)$, and any γ with coordinates (w_1, w_2) and orthogonal component n in $\mathcal{D}_{\alpha,\beta}$ can be expressed with coordinates (w_2, w_1) and the same orthogonal component n in $\mathcal{D}_{\alpha,-\beta}$.

B.5 Theorem 4.5

The assumptions besides those imposing Gaussian covariates and generalized linear models are not too stringent: the upper bounds in the classes $\mathcal{H}^K_{C,\lambda}$ and $\mathcal{H}'^K_{C,\lambda,\lambda'}$ simply ensure two-factor products of these functions remain integrable for the standard centered Gaussian density measure, and in particular that the overlap divergence of X is finite in Item 1.(a) of Theorem 4.5, in line with Assumption 2.3. The lower bound on $1-h(z)$ in $\mathcal{H}'^K_{C,\lambda,\lambda'}$ can be interpreted as ensuring that $e(X)$ does not converge to 1 too fast when X takes large values, a condition that is reasonable if one wants sufficient overlap wrt X. Notably, it is verified for the classical case $h(z) = \mathrm{logit}^{-1}(z)$ for any $\lambda' > 0$ if one takes sufficiently high C. However it is perhaps the most stringent condition in Theorem 4.5 as one could allow up to $\lambda' < \frac{1}{2}$ to allow a finite overlap divergence in X, while the $\lambda' < \frac{1-4\lambda}{6}$ assumption implies $\lambda' < \frac{1}{6}$. The additional condition of some derivative of h admitting a non-zero expectation for the standard centered Gaussian density measure in Items 2.(a)(i) and 2.(b) of Theorem 4.5 is also not stringent, e.g. it is again verified for $h(z) = \mathrm{logit}^{-1}(z)$. Items 2.(a)(ii) and 2.(a)(iii) show that the conclusions of Theorem 4.5 also hold for at least some non-continuously-differentiable functions; notably, $h = \mathrm{ReLU}$ is a popular link function in the machine learning community.

Further, note that for any K, C, λ, λ', if $h \in \mathcal{H}^K_{C,\lambda}$ then $h(-.) \in \mathcal{H}^K_{C,\lambda}$, if $h \in \mathcal{H}'^K_{C,\lambda,\lambda'}$ then $h(-.) \in \mathcal{H}'^K_{C,\lambda,\lambda'}$, and if $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[h^{(K)}(Z)\right] \neq 0$ then $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[h(-.)^{(K)}(Z)\right] \neq 0$, justifying the operation replacing β with $-\beta$ and h with $h(-.)$ when $\alpha'\beta < 0$ as the assumptions of Items 1.(a), 1.(b), 2.(a)(i) and 2.(b) of Theorem 4.5 will still hold even after this transformation.

B.6 Potential Extensions

While we focused on the ATT, our results can be immediately extended to more general covariate shift, thus to other causal inference problems such as estimation of the classical average treatment effect (ATE) or transportability (Clivio et al., 2024).

Note that our work straightforwardly extends to non-standard Gaussian variables: if $X \sim \mathcal{N}(0, \Sigma)$ then the entire Section 4.2 is valid by replacing X, α, β with $\Sigma^{-1/2}X$, $\Sigma^{1/2}\alpha$, $\Sigma^{1/2}\beta$, respectively. Computing closed-form deconfounding scores with any further relaxations in assumptions will be difficult as our results generally require computing the distribution of $f(X)$ given $\phi(X)$, which to the best of our knowledge is classically known only for linear f, ϕ and Gaussian X. For deconfounding scores, only the equivalence between independence and non-correlation for Gaussian covariates allows GLM forms of f. However, a potential direction would be to assume independent Poisson covariates and $\phi_B(X) = \sum_{i\in B} X_i$ for some $B \subset \{1, \ldots, d\}$, as the distribution of $(X_j)_{j\in B}$ conditional on $\phi_B(X)$ is known to be multinomial with parameters depending on individual Poisson parameters (Townes, 2020). Additionally, if results could be generalized to multivariate linear representations, then the GLM assumption would encompass widely-used neural networks.

We chose the one-dimensional class of representation in this paper because it could be characterized analytically, and so that we could develop intuition for the unbiasedness constraint implied by Lemma 3.1. While representations from causal deep learning (Shalit et al., 2017; Johansson et al., 2022; Zhang et al., 2020) are able to estimate richer sets of prognostic/balancing scores, deconfounding scores are designed to have zero confounding bias and their overlap divergence directly controls the semiparametric efficiency bound. How to bridge these two types of representations remains an open question. In one direction, Assumptions 4.2 and 4.3 could be relaxed to incorporate neural-network outcome and propensity models. In another direction, many of the insights in our paper can be applied to designing regularizers that (1) enforce the zero confounding bias constraint as in Lemma 3.1, and (2) incorporate the overlap divergence from our paper. Note that this proposal actually addresses one of the key questions in Johansson et al. (2022) about how regularization should be chosen for causal inference. Generally, we expect the resulting representations to outperform those from the current causal deep learning literature, as they will be both flexible and suited to preserving unconfoundedness and improving overlap. Again, we view this as important motivation for our own work, as it lays the groundwork for such representations.

C DETAILS ON EXPERIMENTS

ATT Estimators. Here we present the analogs of the IPW (Horvitz and Thompson, 1952) and AIPW (Robins et al., 1994) estimators of the ATT (Moodie et al., 2018); using estimators $\hat{e}(X)$ of $e(X)$ and $\hat{m}_0(X)$ of $m_0(X)$:

$$\hat{\tau}^{ATT}_{IPW} := \frac{\sum_{i=1}^N T_i Y_i}{\sum_{i=1}^N T_i} - \frac{\sum_{i=1}^N (1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} Y_i}{\sum_{i=1}^N T_i} \tag{5}$$

and

$$\hat{\tau}^{ATT}_{AIPW} := \frac{\sum_{i=1}^N T_i (Y_i - \hat{m}_0(X_i))}{\sum_{i=1}^N T_i} - \frac{\sum_{i=1}^N (1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} (Y_i - \hat{m}_0(X_i))}{\sum_{i=1}^N T_i}. \tag{6}$$

Further, applying Hajek normalization (Hájek, 1971) gives the following estimators, which we use in our simulations:

$$\hat{\tau}^{ATT,\text{Hajek}}_{IPW} = \frac{\sum_{i=1}^N T_i Y_i}{\sum_{i=1}^N T_i} - \frac{\sum_{i=1}^N (1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} Y_i}{\sum_{i=1}^N (1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}} \tag{7}$$

and

$$\hat{\tau}^{ATT,\text{Hajek}}_{AIPW} := \frac{\sum_{i=1}^{N} T_i(Y_i - \hat{m}_0(X_i))}{\sum_{i=1}^{N} T_i} - \frac{\sum_{i=1}^{N}(1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(Y_i - \hat{m}_0(X_i))}{\sum_{i=1}^{N}(1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}}. \tag{8}$$

Trimming. To avoid division by zero $1-\hat{e}(X_i)$ in the above estimators, we apply the following transformation to the original propensity score estimator $\hat{e}_O$:

$$\hat{e}(X_i) = \begin{cases} 1-\epsilon & \text{if } 1-\hat{e}_O(X_i) < \epsilon, \\ \hat{e}_O(X_i) & \text{otherwise,} \end{cases}$$

where ϵ is chosen at R's machine epsilon value `.Machine$double.eps`, equal to 2.220446×10^{-16} (R Core Team, 2024).

Generation of α and β in Simulated Datasets. We fix a support $\mathcal{S} \subset \{1,\dots,p\}$ for both α and β. To generate α, we generate $\bar{\alpha}$ as $\forall i \in \mathcal{S}, \bar{\alpha}_i \sim \mathcal{N}(0,1)$, and $\forall i \notin \mathcal{S}, \bar{\alpha}_i = 0$, then retrieve $\alpha = \frac{\bar{\alpha}}{||\bar{\alpha}||}$. We then generate β with support $\mathcal{S}$ and chosen such that $\alpha'\beta = K$ for fixed $K \in (-1,1)$ as follows. Again, we generate u such that $\forall i \in \mathcal{S}, u_i \sim \mathcal{N}(0,1)$, and $\forall i \notin \mathcal{S}, u_i = 0$. We construct v as $v = u - (\alpha' u)\alpha$, which is the canonical Gram-Schmidt vector and is orthogonal to α by construction, and deduce $v_n = \frac{v}{||v||}$. Finally, we take $\beta = K\alpha + \sqrt{1-K^2}v_n$. Throughout simulations, we chose $\mathcal{S} = \{1,\dots,20\}$ and $K = 0.75$. We sampled α and β with the same seed across all simulations; notably the values of α in $\mathcal{S}$ were 0.283645181, −0.073268306, 0.298657804, 0.285773167, 0.093123755, −0.345855281, −0.208545604, −0.066190863, −0.001295241, 0.540057827, 0.171494414, −0.179448398, −0.257750723, −0.065009780, −0.067200315, −0.092420657, 0.056646520, −0.200315350, 0.097849518, −0.277937067 and those of β in $\mathcal{S}$ were 0.15135872, 0.02785895, 0.23680749, 0.36840029, 0.05315568, −0.13631776, 0.08172226, −0.19110589, −0.27014375, 0.38747427, 0.07053042, −0.23927785, −0.27107102, −0.18158571, 0.10532350, 0.17679514, 0.24756079, −0.23010864, 0.32796449, −0.25291885.

Zero Estimated $\hat{\alpha}$ and $\hat{\beta}$. Deconfounding scores are ill-defined when either α or β is zero. However, in this scenario, both adjusting for X and adjusting for an empty variable $Z = \emptyset$ yield the same correct ATT estimand τ, so we conjecture that there is no motivation for even using a deconfounding score. As a result, when we find either $\hat{\alpha}$ or $\hat{\beta}$ to be (near-)zero, we impute deconfounding score estimates to be identical to original covariate estimates. We apply this rule when $||\hat{\alpha}_O||_\infty < 10^{-10}$ or $||\hat{\beta}_O||_\infty < 10^{-10}$, where $\hat{\alpha}_O, \hat{\beta}_O$ are the original unnormalized coefficient vectors obtained from the treatment assignment and outcome model regressions, respectively. This remains a minor phenomenon, however. In simulated datasets, this only took place for 4 out of 16 hyperparameter assignments, and within those for at most 3 out of 100 draws of the data. In ACIC2016 datasets, this only took place for 75 out of 308 hyperparameter assignments, and within them for at most 6 out of 100 draws of the data. IHDP and HC-MNIST were unaffected by this phenomenon. We refer to the `_warnings.txt` files produced by the code in the `results` folder.

Zero-Variance Deconfounding Scores. It remains possible that estimated deconfounding scores have either zero empirical variance, rendering both propensity score and outcome model estimation impossible, or zero control empirical variance, rendering outcome model estimation impossible. In this case, again, we impute estimates for this specific deconfounding score to be identical to original covariate estimates. This phenomenon only took place for one deconfounding score in the entire experiments, the equiangular score in the 67th iteration of ACIC 2016 with setting 47 and LASSO estimators.

Ground-Truth ATT in Semi-Synthetic Datasets. In semi-synthetic datasets, covariates are based on real-world studies but outcome (and most often treatment) models are given and synthetic. Thus, such datasets provide $m_t(X_i)$ for all $i = 1,\dots,N$ and $t = 0,1$ but not the ATT due to the unknown distribution P_X; thus, we set the ground-truth ATT to $\frac{\sum_{i=1}^{N} T_i \Delta m(X_i)}{\sum_{i=1}^{N} T_i}$.

Source and Settings in Datasets. IHDP (Hill, 2011) was implemented according to the `npci` package available at `https://github.com/vdorie/npci`, specifically from the `generateDataForIterInCurrentEnvironment`

function available in the package's IHDP example. For that function, we chose `w = 0.5` and all covariates. We varied the IHDP setting between A, B, C and overlap between lower overlap and higher overlap; this gives 6 settings. ACIC 2016 was implemented using its official implementation of Dorie et al. (2017), available at `https://github.com/vdorie/aciccomp`; its 77 settings are described in Dorie et al. (2017). HC-MNIST (Jesson et al., 2021) was converted from its original Python implementation available at `https://github.com/anndvision/quince/blob/main/quince/library/datasets/hcmnist.py` to R, with default parameters, and $\Gamma^\star = 1$ to remove the influence of unobserved confounders.

Infrastructure. Experiments were conducted on a single CPU, of model name "Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz CPU" with 2 threads per core, 6 cores per socket, and 1 socket.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Prognostic scores and representation learning for causal effect estimation with weak overlap
Publication Status	□ Published □ Accepted for Publication X Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Clivio, O.*, D'Amour, A.*, Franks, A.*, Bruns-Smith D., Holmes, C.C., Feller, A. (2025). Prognostic scores and representation learning for causal effect estimation with weak overlap. In *Submitted*. (* for equal contribution)

Student Confirmation

Student Name:	Oscar Clivio		
Contribution to the Paper	As the first author, I introduced the idea of explicitly optimizing overlap when learning a representation, introduced overlap divergence and derived its properties in both the nonparametric and simplified cases, derived the current version of the confounding bias conditional covariance result and of the hyperbola result, adjusted the codebase and numerical results, wrote the first iteration of the paper and took the lead in the writing. This work was partially done by merging the two following preprints: 1) Clivio, O., Bruns-Smith, D., Feller, A., & Holmes, C. C. (2024). Towards Principled Representation Learning to Improve Overlap in Treatment Effect Estimation. In *9th Causal Inference Workshop at UAI 2024*. 2) D'Amour, A., & Franks, A. (2021). Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*. D'Amour and Franks (2021) introduced deconfounding scores, derived an early version of the confounding bias conditional covariance and hyperbola results, wrote the bulk of the codebase and experimental framework, designed plots with hyperbolas, wrote first versions of the corresponding paragraph and of other paragraphs in the experimental section. Co-authors also assisted in checking theoretical and numerical results as well as writing.		
Signature		Date	2025/05/19

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Chris Holmes		
Supervisor comments I agree with the student's comments		
Signature 	Date	2025/06/26

This completed form should be included in the thesis, at the end of the relevant chapter.

5

Discussion and Conclusion

Contents

5.1 Summary of contributions 120
5.1.1 Chapter 2 120
5.1.2 Chapter 3 120
5.1.3 Chapter 4 121
5.2 Key findings 121
5.2.1 More advanced representation learning for design-based causal inference 122
5.2.2 First steps towards controlling the confounding bias 122
5.2.3 An impossibility result on design-based causal inference and overlap 123
5.3 Limitations 124
5.3.1 Reliance on well-specified models 124
5.3.2 Unknown behaviour of representation-wise estimators 125
5.4 General conclusion 125

In this concluding Chapter, we summarise contributions of the thesis (Section 5.1), highlight key findings from these contributions (Section 5.2), describe limitations spanning the entire thesis (Section 5.3), and finally outline a general conclusion for the thesis (Section 5.4).

5.1 Summary of contributions

5.1.1 Chapter 2

In Chapter 2, we provided an extension of the classical propensity score matching method [Rosenbaum and Rubin, 1983]. This extension builds on the compositional nature of the balancing score property in Rosenbaum and Rubin [1983]: any representation of covariates that completely predicts the propensity score is a valid balancing score, thus a valid adjustment set. As neural networks naturally provide this compositionality, hidden layers in a neural network model for the propensity score can be used as representations. If the neural network model is well-specified and the learnt representation is correct, then the imbalance in the representation space controls the imbalance in the original covariate space, which itself controls the bias of the matching estimator [Ben-Michael et al., 2021, Kallus, 2020]. When the neural network model is misspecified or more generally the resulting representation is not a balancing score, then the imbalance in the representation space controls the imbalance in the covariate space up to a "balancing score error". As the name suggests, the latter measures how much the representation is not a balancing score.

5.1.2 Chapter 3

In Chapter 3, besides generalising the framework of interest from matching to weighting in general, we provided an alternative balancing score error that can be used to learn the representation. More specifically, we identified (absolute) confounding bias as the key measure of misspecification of a given representation that is used as the input feature space for weighting. We demonstrated that it could be written as the covariance of the two residuals of regressing the outcome model and the covariate shift (which is inherent to treatment effect estimation problems) on the representation. The standard deviation of the second residual then gives the new form of the balancing score error. It presents several advantages over the balancing score error of Chapter 2: (i) it inherently controls the confounding bias of the representation; (ii) it is more intuitive, as it measures how much a given representation does not predict the covariate shift's density ratio; and (iii) it can further be bounded using the minimal AutoDML loss over a class of functions,

with the bound becoming an equality if the class of functions is sufficiently rich, yielding a procedure to learn a representation.

5.1.3 Chapter 4

In Chapter 4, we returned to the important question of poor overlap in high dimensions. We showed that an *overlap divergence* objective function measures how poor overlap is while determining the variance of semiparametric estimators adjusting for the representation. This helps formalise learning representations as minimising this objective function under a zero confounding bias. Additionally, the overlap divergence helps us formally prove an important intuition in the literature, which is that the less a representation predicts the treatment assignment, the better its overlap. This means that outcomes are required to provide a zero or minimal confounding bias together with improved overlap. This further raises the question of whether the more a representation predicts the outcome, the better the overlap. We answered positively in a simplified setting, with Gaussian covariates and generalised linear models for the outcome and for the treatment assignment. In this setting, a set of representations with no confounding bias can be explicitly computed in closed form, and can be expressed as a segment of a hyperbola with endpoints being the propensity and prognostic scores. Any other representation without confounding bias can be interpreted as being composed of both scores and being closer to either depending on its position on the hyperbola. We also showed that for such representations, the closer to the prognostic score on the hyperbola, the better the overlap. As a result, the prognostic score endpoint has the best overlap. A key question remains on how to generalise this result to more general data-generating processes.

5.2 Key findings

In this section, we underline the most important findings spanning the different chapters of the thesis.

5.2.1 More advanced representation learning for design-based causal inference

To the best of our knowledge, variable selection and representation learning for treatment effect estimation have mostly focused on incorporating outcome information, in order to improve efficiency in general and overlap in particular. In contrast, in design-based causal inference, only propensity score fitting [Rosenbaum and Rubin, 1983] and the application of sufficient dimension reduction and partial least squares by Ghosh [2011] were used to learn representations of covariates. This thesis fills this gap in two ways. First, Chapter 2 leverages the common practice of fitting a propensity score to provide multivariate representations. Those might contain richer information than the scalar propensity score, as suggested by experiments. Second, Chapter 3 proposes an alternative loss that can be used not only to learn a multivariate representation but also the propensity score itself. This loss more directly targets the confounding bias of the representation, which is more tailored towards the goal of learning a suitable adjustment set. In contrast, classical metrics to assess classification performance of the estimated propensity score, such as the area under the ROC curve, are typically not suited to assess performance of treatment effect estimators adjusting for such a score [Westreich et al., 2011].

5.2.2 First steps towards controlling the confounding bias

Chapters 3 and 4 are important steps towards controlling the confounding bias induced by a representation, which itself controls the bias of estimators adjusting for the representation. In Chapter 3, the AutoDML loss explicitly upper-bounds the absolute confounding bias, and such an upper bound can be directly minimised to obtain a representation. In contrast, to the best of our knowledge, variants of the confounding bias were (i) provided in Johansson et al. [2019] and Curth et al. [2021] without explicitly attempting to estimate, bound or minimise them, (ii) bounded in Melnychuk et al. [2024], without attempting to minimise the bounds. The AutoDML loss might be a loose bound, however, as it only uses treatment assignment information. Chapter 4 explicitly gives closed forms for representations with no confounding bias in a specific case. Estimating such representations or providing tighter operational bounds on the confounding bias in a more

general case remains an area to explore. The result of Chapters 3 (Proposition 3.4) and 4 (Lemma 3.1) that the confounding bias can be written as the covariance between the residuals of regressing the outcome model and the density ratio on the representation can be used towards this aim. It also illustrates previous results in the literature that the prognostic score (outcome model) and the propensity score are suitable adjustment sets, i.e., that they have no confounding bias, as each makes its respective residual zero almost surely. A key intermediary result in Proposition 3.4 of Chapter 3 to establish this form of the confounding bias is that the density ratio in the representation space is the conditional expectation of the original density ratio on the representation in the source distribution. This result might be of independent interest to the broader covariate shift and domain adaptation community.

5.2.3 An impossibility result on design-based causal inference and overlap

While confounding bias determines the bias of estimators that adjust for the representation, controlling their variance remains a question. This was done in Chapter 4 where overlap divergence measures the degree of poor overlap in a representation while determining the efficiency bound of semiparametric estimators adjusting for this representation. Lemma 4.1 of Chapter 4 further showed that the reduction in overlap divergence induced by a representation compared to the baseline covariates is precisely the (squared) balancing score error of Chapter 3. Thus, while Chapter 3 aims at finding representations with *minimal* balancing score error to induce a minimal confounding bias, Chapter 4 states that an ideal representation should have zero or minimal confounding bias together with *maximal* balancing score error. This underlines a fundamental limitation of learning representations in design-based causal inference: in the absence of outcome information, a representation can only be made a suitable adjustment set by making it predict the treatment assignment. In turn, this leads to overlap not improving. As a result, learning representations in design-based causal inference can inherently not help to improve overlap, or will only do so at the expense of confounding bias. Outcome information is required to bypass this apparent trade-off and improve overlap *while* maintaining a zero

confounding bias. That said, Chapters 2 and 3 do demonstrate improved design-based treatment effect estimation performance when learning low-dimensional representations. This could be due to such representations bypassing issues with dimensionality in matching or weighting specifically, e.g., in nearest-neighbour methods, or due to a greater improvement in overlap (or variance more generally) compared to the added confounding bias. Assessing this would require additional investigation.

5.3 Limitations

Besides individual limitations listed in every chapter, we now describe limitations spanning the entire thesis.

5.3.1 Reliance on well-specified models

All research papers in the thesis rely on some form of model specification. Chapter 2 relies on fitting a neural-network propensity score model: apart from the bounds involving non-balancing scores (Proposition 7), theoretical guarantees are available only when the fitted network recovers the true propensity score or, at least, yields a valid balancing score. Chapter 3 partly addresses this by bounding the confounding bias, which can be interpreted as the misspecification of the representation. However, two successive bounds are applied: the balancing score error and the minimal AutoDML loss over a class of functions. If the class of functions is not sufficiently large, thus being in a sense misspecified, then the two bounds will not coincide, and the gap might be large. Additionally, using a representation as input for the balancing weights method induces both a new outcome model (projecting the original one on the representation) and a new class for the integral probability metric class (doing the same projection for all functions in the class, see Equation 3 in the Chapter). As with standard balancing weights, we cannot do better than just assuming that this outcome model belongs to the class. Finally, in Chapter 4, the closed-form hyperbola-like set of deconfounding scores (Theorem 4.4) and the result on overlap for this set (Theorem 4.5) depend on strict assumptions like Gaussianity and generalised linear models. The computation of such deconfounding scores further requires well-specification of the posited generalised linear

models. Thus, future work should aim to reduce dependence on such assumptions; a key step will likely be to find tighter operational bounds for the absolute confounding bias. Despite such issues with well-specification, we observe that using low-dimensional representations improves performance of treatment effect estimators in practice. This leads to the next limitation.

5.3.2 Unknown behaviour of representation-wise estimators

In all research papers, learning representations relies on learning suitable propensity, density ratio or outcome models. In theory, such models could be used directly as treatment effect estimators and even as scalar representations. However, we observed that (i) hidden-layer representations from Chapter 2 performed better than propensity scores extracted from the same neural network and used as scalar representations; (ii) AutoDML-loss-based representations from Chapter 3 performed better than the corresponding density ratio model used directly as weights; and (iii) AIPW, IPW and outcome regression estimators from Chapter 4 were always outperformed by at least one corresponding deconfounding score method. The thesis focused on learning these input representations but did not explore *why* the estimators adjusting for them performed better in treatment effect estimation than the models used to learn the representations. This raises the question of the properties of such estimators, for example their consistency or asymptotic variance. This step will be critical to ensure the deployment of representations for treatment effect estimation in high-stakes domains where better understanding and control of uncertainty are essential.

5.4 General conclusion

In this thesis, we presented three contributions to treatment effect estimation with high-dimensional covariates, all leveraging representation learning. A first contribution extended propensity score matching to learning multivariate representations. A second contribution bypassed traditional propensity score fitting to instead minimize a loss that directly controls the confounding bias induced by the representation. A third contribution

switched to explicitly improving overlap using representations and provided first results in this direction, including an impossibility result and an alternative in a simple setting.

Generally, the thesis advanced representation learning in design-based causal inference and in dealing with poor overlap. It also advanced the control of the confounding bias induced by a representation. Limitations of the thesis that future work might mitigate include the dependence on well-specified models and the focus on the representation learning stage in contrast to the second estimation and inference stage which remains to be explored.

The thesis might be of help in modern industry settings where high-dimensional datasets are prevalent. Improving on limitations would ensure a safe deployment of the thesis's methods as well as more general representation learning for treatment effect estimation in high-stakes environments, such as medicine or finance.

Bibliography

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer, 2001.

C. A. Anderson. Temperature and aggression: Ubiquitous effects of heat on occurrence of human violence. *Psychological Bulletin*, 106(1):74, 1989.

J. Antonelli, M. Cefalu, N. Palmer, and D. Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

S. Assaad, S. Zeng, C. Tao, S. Datta, N. Mehta, R. Henao, F. Li, and L. Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.

S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623, 02 2018. ISSN 1369-7412.

F. Bach. *Learning Theory From First Principles*. MIT Press, 2024.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521 – 547, 2013.

A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014a.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014b.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

E. Ben-Michael, A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021.

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

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer, 1999.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 245–250, 2001.

C. R. Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

J. Bradic, S. Wager, and Y. Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156, 2006.

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

P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

S. Chaudhuri and J. B. Hill. Heavy tail robust estimation and inference for average treatment effects. *Econometric Reviews*, pages 1–43, 2025.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221.

V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic debiased machine learning via riesz regression. *arXiv preprint arXiv:2104.14737*, 2021.

V. Chernozhukov, W. Newey, V. M. Quintas-Martınez, and V. Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022a.

V. Chernozhukov, W. K. Newey, and R. Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022b.

V. Chernozhukov, W. K. Newey, and R. Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25 (3):576–601, 2022c.

C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim. Sensitivity analysis of linear structural causal models. In *International Conference on Machine Learning*, pages 1252–1261. PMLR, 2019.

O. Clivio, F. Falck, B. Lehmann, G. Deligiannidis, and C. Holmes. Neural score matching for high-dimensional causal inference. In G. Camps-Valls, F. J. R. Ruiz, and I. 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.

O. Clivio, A. Feller, and C. Holmes. Towards representation learning for weighting problems in design-based causal inference. In N. Kiyavash and J. M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 856–880. PMLR, 15–19 Jul 2024.

O. Clivio, D. Mahajan, P. Taslakian, S. Magliacane, I. Mitliagkas, V. Zantedeschi, and A. Drouin. Learning to defer for causal discovery with imperfect experts. In *ICLR 2025 Workshop on "Reasoning and Planning for Large Language Models"*, 2025.

O. Clivio, A. D'Amour, A. Franks, D. Bruns-Smith, C. Holmes, and A. Feller. Deconfounding scores and representation learning for causal effect estimation with weak overlap. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*. PMLR, 2–5 May 2026.

B. Colnet, J. Josse, G. Varoquaux, and E. Scornet. Re-weighting the randomized controlled trial for generalization: Finite-sample error and variable selection. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188(2):345–372, 05 2024. ISSN 0964-1998.

R. D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. John Wiley & Sons, 2009.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

A. Curth, C. Lee, and M. van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 34:26740–26753, 2021.

P. Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

I. Degtiar and S. Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

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

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

A. Ertefaie, M. Asgharian, and D. A. Stephens. Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, 6(1):20170010, 2018.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 2008.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7): 761–767, 2011.

D. Ghosh. Propensity score modelling in observational studies using dimension reduction methods. *Statistics & Probability Letters*, 81(7):813–820, 2011.

A. N. Glynn and K. M. Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56, 2010.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

A. Gretton, A. J. Smola, J. Huang, M. Schmittfull, K. M. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning*, pages 131 – 160. MIT Press, Cambridge, MA, 2009. ISBN 978-0-262-17005-5.

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

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

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

M. Hernan and J. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2024. ISBN 9781420076165.

J. Hill and Y.-S. Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

D. A. Hirshberg and S. Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227, 2021.

R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

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

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

M. Huang and S. D. Pimentel. Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, 112(1):asae040, 2025.

J. D. Huling and S. Mak. Energy balancing of covariate distributions. *Journal of Causal Inference*, 12(1):20220029, 2024.

P. Hünermund, B. Louw, and I. Caspi. Double machine learning and automated confounder selection: A cautionary tale. *Journal of Causal Inference*, 11(1):20220078, 2023.

G. W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4): 1129–1179, 2020.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014.

A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

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

F. D. Johansson, D. Sontag, and R. Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

W. B. Johnson, J. Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.

N. Kallus. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020.

N. Kallus. More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534):646–658, 2021.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

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

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

S. Khan and E. Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

N. Kreif and K. DiazOrdaz. Machine learning in policy evaluation: New tools for causal inference. *arXiv preprint arXiv:1903.00402*, 2019.

Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, 2, 1989.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

B. Li. *Sufficient Dimension Reduction: Methods and Applications With R*. Chapman and Hall/CRC, 2018.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

F. Li, H. Lam, and S. Prusty. Robust importance weighting for covariate shift. In *International Conference on Artificial Intelligence and Statistics*, pages 352–362. PMLR, 2020.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

S. Li and Y. Fu. Matching on balanced nonlinear representations for treatment effects estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

W. Luo, Y. Zhu, and D. Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

S. Ma, L. Zhu, Z. Zhang, C.-L. Tsai, and R. J. Carroll. A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *Annals of Statistics*, 47(3):1505, 2019.

X. Ma and J. Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

D. Madras, T. Pitassi, and R. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Two-stage learning to defer with multiple experts. *Advances in Neural Information Processing Systems*, 36:3578–3606, 2023.

R. A. Matsouaka and Y. Zhou. A framework for causal inference in the presence of extreme inverse probability weights: The role of overlap weights. *arXiv preprint arXiv:2011.01388*, 2020.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

E. McMahan. Eating ice cream does not lead to murder: Association, correlation, and causation. *Eye on Psi Chi*, 26(2):31–50, 2021.

M. Mehrabi and S. Wager. Off-policy evaluation in markov decision processes under weak distributional overlap. *arXiv preprint arXiv:2402.08201*, 2024.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

V. Melnychuk, D. Frauen, and S. Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024.

D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, 2010.

E. E. Moodie, O. Saarela, and D. A. Stephens. A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, 7(1):e205, 2018.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538 – 557, 2012.

Y. Ning, P. Sida, and K. Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–688, 1995.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Pearl. Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, 174(11):1223–1227, 2011.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018. OpenAI Technical Report.

A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. 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.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

J. Raymaekers and P. J. Rousseeuw. Fast robust correlation for high-dimensional data. *Technometrics*, 63(2):184–198, 2021.

T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30): 2013, 2013.

P. Rigollet and J.-C. Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

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

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.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

D. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322 – 331, 2005.

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

D. B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808 – 840, 2008.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Y. Sasaki and T. Ura. Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory*, 38(1):66–112, 2022.

S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522, 2009.

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

L. S. Shapley. 17. a value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games, Volume II*, pages 307–318. Princeton University Press, Princeton, 1953. ISBN 9781400881970.

C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

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

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472, 1990.

E. A. Stuart. 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.

T. Stürmer, K. J. Rothman, J. Avorn, and R. J. Glynn. Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution—a simulation study. *American Journal of Epidemiology*, 172(7): 843–854, 2010.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Z. Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Z. Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020.

L. Ter-Minassian, O. Clivio, K. Diazordaz, R. J. Evans, and C. C. Holmes. PWSHAP: A path-wise explanation model for targeted variables. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34054–34089. PMLR, 23–29 Jul 2023.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

T. Wang, M. Morucci, M. U. Awan, Y. Liu, S. Roy, C. Rudin, and A. Volfovsky. Flame: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, 22(31):1–41, 2021.

L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2004.

D. Westreich, S. R. Cole, M. J. Funk, M. A. Brookhart, and T. Stürmer. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety*, 20(3):317–320, 2011.

H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, pages 391–420, 1966.

H. Wold. Soft modelling by latent variables: The non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142, 1975.

J. M. Wooldridge. Should instrumental variables be used as matching variables? *Research in Economics*, 70(2):232–237, 2016.

R. Wyss, M. van der Laan, S. Gruber, X. Shi, H. Lee, S. K. Dutcher, J. C. Nelson, S. Toh, M. Russo, S. V. Wang, et al. Targeted learning with an undersmoothed lasso propensity score model for large-scale covariate adjustment in health-care database studies. *American Journal of Epidemiology*, 193(11):1632–1640, 2024.

L. Yang, V. Shirvaikar, O. Clivio, and F. Falck. A critical review of causal reasoning benchmarks for large language models. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2024.

Y.-L. Yu and C. Szepesvári. Analysis of kernel mean matching under covariate shift. In *International Conference on Machine Learning*, pages 1147–1154. PMLR, 2012.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

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

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

Q. Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965 – 993, 2019.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. ISSN 1369-7412.

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