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.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

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_X^1 is absolutely continuous wrt P_X^0, or equivalently, $e(X) < 1$ P-almost surely.*

Assumption 2.3 *(Square-integrability of $\frac{dP_X^1}{dP_X^0}(X)$) The density ratio $\frac{dP_X^1}{dP_X^0}(X)$ between control and treatment distributions of X is square-integrable in P^0:*

$$\mathbb{E}_{P^0}\left[\left(\frac{dP_X^1}{dP_X^0}(X)\right)^2\right] < \infty.$$

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_{eff}^Z (Tsiatis, 2006) given as (Hahn, 1998)

$$V_{\text{eff}}^Z = \mathbb{E}_P\left[\frac{e(Z)\text{Var}_{P^1}(Y|Z) + (\Delta m(Z) - \tau_Z)^2 e(Z)}{\pi_1^2}\right] + \mathbb{E}_{P^0}\left[\frac{\text{Var}_{P^0}(Y|Z)}{1-\pi_1}\left(\frac{dP_Z^1}{dP_Z^0}(Z)\right)^2\right].$$

We can see that V_{eff}^Z depends on the magnitude of the density ratio $\frac{dP_Z^1}{dP_Z^0}$; this magnitude describes the strength of overlap between the distributions of Z in the treatment and control groups. When $\frac{dP_Z^1}{dP_Z^0}$ 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

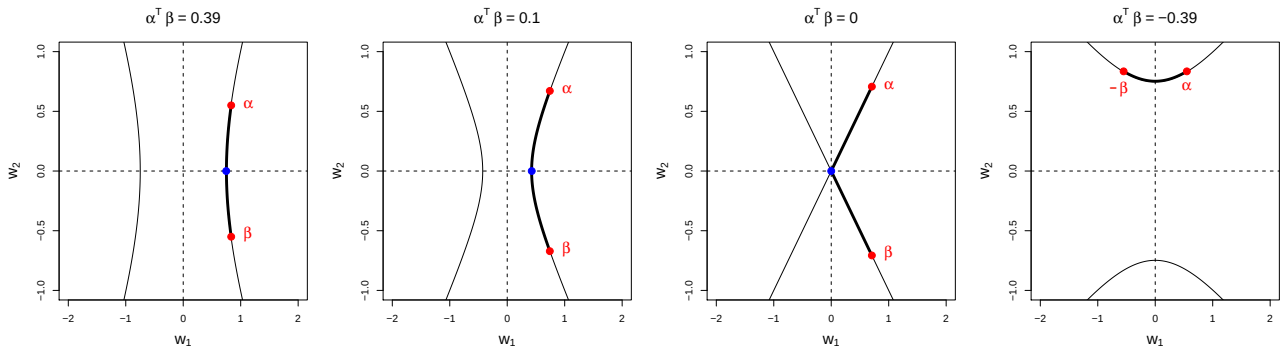

Figure 1: The projection of γ onto the space spanned by α and β lies on a segment of a hyperbola (bold black line) whose endpoints correspond to $\gamma = \alpha$ and $\gamma = \beta$ (when $\alpha'\beta \geq 0$) or to $\gamma = \alpha$ and $\gamma = -\beta$ (when $\alpha'\beta < 0$), shown here, and 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 Null(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.

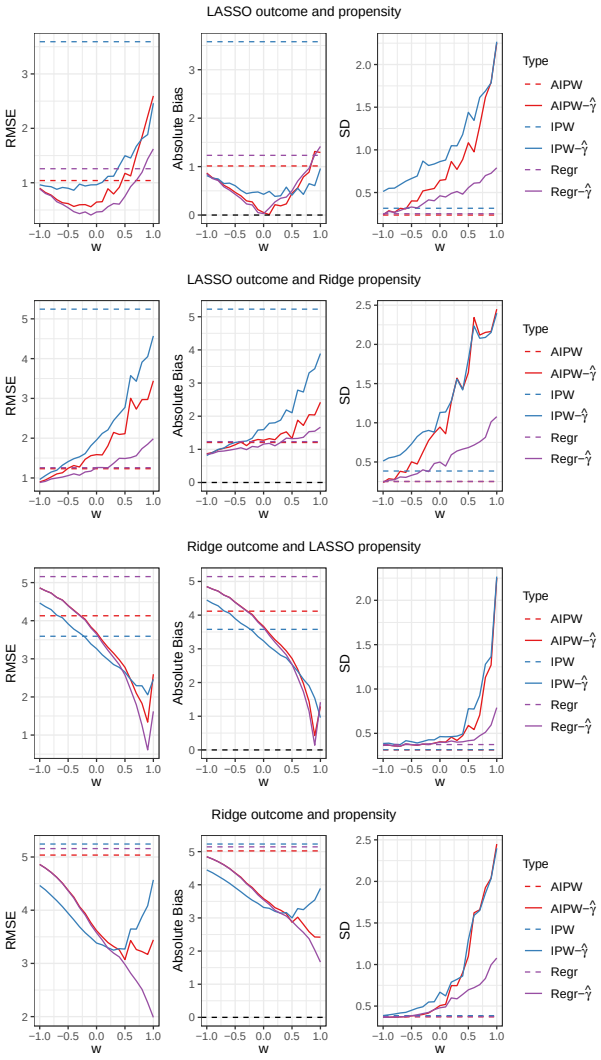

Figure 2: **RMSE, bias and standard deviation for simulated datasets.** Each metric is plotted according to the deconfounding score coordinate parameter w; methods using base covariates are constant. In all plots, $s_T = 4$ (low overlap) and $s_Y = 5$ (high SNR). See Section 5.1 for definitions of names 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. *Working 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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). No: measures of uncertainty were not included due to a lack of space.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. Yes

(b) The license information of the assets, if applicable. Not Applicable

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

(d) Information about consent from data providers/curators. Not Applicable

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

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,

$$
\begin{aligned}
&\mathbb{E}_P \Big[\frac{\big(m_1(\phi(X)) - m_0(\phi(X)) - \tau_{\phi(X)}\big)^2 e(\phi(X))}{\pi_1^2} \Big] \\
&= \frac{1}{\pi_1} \mathbb{E}_{P^1} \Big[\big(m_1(\phi(X)) - m_0(\phi(X)) - \tau_{\phi(X)}\big)^2 \Big] \text{ as } \frac{dP^1_{\phi(X)}}{dP_{\phi(X)}}(\phi(X)) = \frac{e(\phi(X))}{\pi_1} \\
&= \frac{1}{\pi_1} \mathrm{Var}_{P^1} \Big(m_1(\phi(X)) - m_0(\phi(X)) \Big) \\
&\leq \frac{1}{\pi_1} \mathbb{E}_{P^1} \Big[\big(m_1(\phi(X)) - m_0(\phi(X))\big)^2 \Big] \\
&\leq \frac{1}{\pi_1} \mathbb{E}_{P^1} \Big[2 \big(m_1(\phi(X))^2 + m_0(\phi(X))^2 \big) \Big] \text{ from } (a-b)^2 \leq 2(a^2+b^2) \quad \forall a, b \\
&\leq \frac{4 Y^2_{\max}}{\pi_1}.
\end{aligned}
$$

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 $\phi_\gamma(X)$,

$$\begin{aligned} &\mathbb{E}_P\left[f_3(\phi_\gamma(X)) \cdot f_1(X) \cdot (f_2(X) - \mathbb{E}_P\left[f_2(X)|\phi_\gamma(X)\right])\right] \\ &= \mathbb{E}_P\left[f_3(\phi_\gamma(X)) \cdot (f_1(X) - \mathbb{E}_P\left[f_1(X)|\phi_\gamma(X)\right]) \cdot (f_2(X) - \mathbb{E}_P\left[f_2(X)|\phi_\gamma(X)\right])\right] \\ &= \mathbb{E}_P\left[f_3(\phi_\gamma(X)) \cdot \mathbb{E}_P\left[(f_1(X) - \mathbb{E}_P\left[f_1(X)|\phi_\gamma(X)\right]) \cdot (f_2(X) - \mathbb{E}_P\left[f_2(X)|\phi_\gamma(X)\right])\right]|\phi_\gamma(X)\right] \\ &= \mathbb{E}_P\left[f_3(\phi_\gamma(X)) \cdot \mathrm{Cov}_P\left(f_1(X), f_2(X)|\phi_\gamma(X)\right)\right] \end{aligned}$$

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 $\forall u \in [0,1],\ \ g(u) \leq g(1)$. 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

$$\begin{aligned}&\frac{\partial}{\partial z}\left(\mathbb{E}_{Z'}\left[h'(uz+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\right)\\&=u\cdot\Bigg(\mathbb{E}_{Z'}\left[h''(uz+\sqrt{1-u^2}Z')\right]\cdot f'\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\\&\qquad+\mathbb{E}_{Z'}\left[h'(uz+\sqrt{1-u^2}Z')\right]^2f''\left(\mathbb{E}_{Z'}\left[h(uz+\sqrt{1-u^2}Z')\right]\right)\Bigg).\end{aligned}$$

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 $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}[h(Z)] = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\max(0,Z)] = \varphi(0) = \frac{1}{\sqrt{2\pi}}$, 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

$$\hat{\tau}_{AIPW}^{ATT,\text{Hajek}} := \frac{\sum_{i=1}^{N} T_i(Y_i - \hat{m}_0(X_i))}{\sum_{i=1}^{N} T_i} - \frac{\sum_{i=1}^{N}(1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(Y_i - \hat{m}_0(X_i))}{\sum_{i=1}^{N}(1-T_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}}. \tag{8}$$

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.