Towards Principled Representation Learning to Improve Overlap in Treatment Effect Estimation

Oscar Clivio[1] **David Bruns-Smith**[2] **Avi Feller**[2] **Chris Holmes**[1]

[1]University of Oxford
[2]University of California, Berkeley

Abstract

A common approach to mitigate undesirable effects of poor overlap is to use well-crafted representations of covariates as adjustment sets. In this abstract, we motivate quantifying the overlap induced by a representation using the χ^2-divergence, show that the overlap improvement under this metric is precisely how much the representation does not predict the propensity score, which confirms intuitions in previous work, and discuss next steps.

1 INTRODUCTION

1.1 PROBLEMS WITH POOR OVERLAP

A critical assumption for identification of the average treatment effect (ATE) is *overlap*, i.e. that the propensity score, the probability of receiving the treatment given covariates, is different from 0 and 1 almost surely. Indeed, the conditional average treatment effect (CATE) is not identified on any covariate value with propensity score 0 or 1, and can only be estimated on such values using extrapolation [Nethery et al., 2019, Pfister and Bühlmann, 2024, Khan et al., 2024]. Even when overlap is verified, estimation of the ATE might remain difficult when the propensity score is not bounded away from 0 or 1, even though identification of the CATE for any covariate value, thus of the full ATE, is possible. First, estimators involving an inverse propensity score such as IPW or doubly-robust estimators might exhibit outsize errors [Petersen et al., 2012, Li et al., 2019, D'Amour and Franks, 2021]. More generally, estimators might exhibit slow convergence or confidence intervals based on root-n consistency might not be available [Rothe, 2017, D'Amour et al., 2021, Hong et al., 2020]. This is problematic as assuming that the propensity score is bounded away from 0 or 1 might not even be realistic, notably in high dimensions [D'Amour et al., 2021].

To mitigate issues with poor overlap, one set of approaches is to *move the goalposts*, that is change the estimand to the ATE on a (weighted) population where the propensity score is further away from 0 or 1 [Crump et al., 2009, Matsouaka and Zhou, 2020]. However, not only do such methods target a *different* estimand than the original ATE but the gap between these two estimands can also be large, especially when covariates are high-dimensional [Petersen et al., 2012, D'Amour et al., 2021]. In contrast, another set of approaches changes not the population but instead the adjustment set to a *representation* of covariates, that is their image through a given mapping [Luo et al., 2017, D'Amour and Franks, 2021, Wu and Fukumizu, 2022, Breitholtz et al., 2023]. Indeed, this representation will be expected to have a propensity score further away from 0 or 1 compared to initial covariates, improving overlap, and verify unconfoundedness, preserving the original ATE as an estimand.

While such approaches have demonstrated better performance compared to adjusting on original covariates, they still rely on stringent model well-specification assumptions, and it is unclear how exactly they mitigate poor overlap and how this translates into improved performance. In the following, we provide our first results towards *principled* learning of *flexible* representations that mitigate poor overlap to improve performance : we show that the lack of overlap can be quantified using a *misoverlap* measure using χ^2-divergences, and improvement of this misoverlap when using a representation instead of original covariates is precisely given by how much the representation does not predict treatment assignment. This notably confirms previous intuitions in the literature.

2 FORMALISATION, NAILING THE OBJECTIVE

Let us note covariates as X, the binary treatment as A, each potential outcome as $Y(a)$ for $a \in \mathcal{A} := \{0, 1\}$, the observed outcome as Y. Assume we observe i.i.d samples

$(X_i, A_i, Y_i) \sim P$. For a random variable Z and a treatment value $a \in \mathcal{A}$, let

$$\mathbb{E}_a[.] = \mathbb{E}[.|A = a], \;\; \tau(Z) := \mathbb{E}_1[Y|Z] - \mathbb{E}_0[Y|Z],$$
$$\bar{\tau}^Z := \mathbb{E}[\tau(Z)], \;\; \sigma_a^2(Z) := \text{Var}(Y|A = a, Z),$$
$$e_a(Z) := P(A = a|Z), \;\; p(a) := P(A = a),$$

We make the canonical assumptions of unconfoundedness wrt X, SUTVA and overlap wrt X. Notably, we place ourselves in a *weak overlap* regime, where the propensity score is always different but not bounded away from 0 and 1. A common metric for assessing the impact of the proximity of $e_a(\phi(X))$ to 0 or 1 on the variability of estimators of $\bar{\tau}^\phi := \bar{\tau}^{\phi(X)}$ is the *efficiency bound* of $\bar{\tau}^\phi$ [Hahn, 1998, Crump et al., 2009], which gives the lowest possible variance of any regular and asymptotically linear (RAL) semi-parametric estimator of $\bar{\tau}^\phi$ using the sample $(\phi(X_i), A_i, Y_i)$, where X has been replaced with $\phi(X)$. It is given by

$$V_{\text{eff}}(\bar{\tau}^\phi) = \mathbb{E}\left[\sum_{a \in \mathcal{A}} \frac{\sigma_a^2(\phi(X))}{e_a(\phi(X))} + \left(\tau(\phi(X)) - \bar{\tau}^\phi\right)^2\right].$$

We see that it can be very large or even infinite when either propensity score $e_a(\phi(X))$ is not bounded away from 0. We further make the following technical assumption.

Assumption 2.1. $\exists \underline{\sigma}^2 > 0, \forall a \in \mathcal{A}, \sigma_a^2(X) \geq \underline{\sigma}^2$

Assumption 2.1 is not stringent as $\sigma_a^2(x)$ has experimentally been shown to be large in areas with small $e_a(x)$ [Hill and Su, 2013, Zhang et al., 2020], i.e. those which might already make the efficiency bound large. Then, $V_{\text{eff}}(\bar{\tau}^\phi)$ can be lower bounded, yielding our objective (all proofs in the Supplementary).

Proposition 2.2. *Under Assumption 2.1,*

$$V_{\mathit{eff}}(\bar{\tau}^\phi) \geq \underline{\sigma}^2 \mathcal{O}(\phi) \textit{ where } \mathcal{O}(\phi) := \sum_{a \in \mathcal{A}} \frac{\chi_a^2(\phi) + 1}{p(a)}$$

where each $\chi_a^2(\phi) := \mathbb{E}_a\left[\left(\frac{p(\phi(X))}{p(\phi(X)|a)} - 1\right)^2\right]$ *is the* χ^2-*divergence between* $P(\phi(X)|a)$ *and* $P(\phi(X))$.

This justifies finding ϕ minimizing $\mathcal{O}(\phi)$, which we refer to as the *misoverlap* of ϕ. It is minimal when $\phi(X) \perp\!\!\!\perp A$; however such ϕ might destroy confounding information contained in X. In general, $\bar{\tau}^X$ is equal to the ATE, which $\bar{\tau}^\phi$ is not necessarily. Thus, we are looking for ϕ minimizing both $\mathcal{O}(\phi)$ and $\text{CE}(\phi) := |\bar{\tau}^X - \bar{\tau}^\phi|$, where the latter term can be understood as a measure of how much ϕ preserves unconfoundedness, or a *confounding error*.

3 THE IMPROVEMENT IN OVERLAP IS A BALANCING SCORE ERROR

In previous work [Luo et al., 2017, D'Amour and Franks, 2021, Wu and Fukumizu, 2022], using strong assumptions on the data generating process, the representation ϕ was chosen to have a zero confounding error by design and improve overlap by predicting the outcome rather than the treatment assignment. Indeed, as a ϕ predicting the treatment assignment perfectly would leave the propensity scores untouched, it has been posited that a ϕ should incorporate outcome information to make propensity scores less "extreme" [D'Amour and Franks, 2021]. We formalize this next : we show that how badly ϕ predicts treatment assignment *is exactly* how much it improves overlap according to our misoverlap.

Proposition 3.1. *if* $\mathcal{O}(X) := \mathcal{O}(\mathit{Id}) < \infty$*, the reduction of misoverlap from* X *to* $\phi(X)$*,* $\mathcal{O}(X) - \mathcal{O}(\phi)$*, is equal to*

$$\sum_a p(a)\mathbb{E}_a\left[\left(\frac{1}{e_a(X)} - \mathbb{E}_a\left[\frac{1}{e_a(X)}\middle|\phi(X)\right]\right)^2\right]$$

We see that this term is a *squared balancing score error* in the sense that (i) it is always non-negative, (ii) it is zero iff the propensity score $e_1(X)$ is a function of $\phi(X)$ a.s., i.e. if $\phi(X)$ is a balancing score [Rosenbaum and Rubin, 1983], (iii) it generally is a weighted sum of the mean squared errors between each treatment-wise inverse propensity score (itself a bijection of the propensity score $e_1(X)$) and its best predictor from $\phi(X)$. Thus, the improvement in overlap from X to $\phi(X)$ is always non-negative, confirming the intuition that a representation always improves overlap, and at the same time given by how badly $\phi(X)$ predicts the propensity score $e_1(X)$, confirming another intuition.

We can allow for more flexible representations by minimizing the confounding error instead of enforcing it at zero exactly. Clivio et al. [2024] show that the confounding error is bounded by the balancing score error, thus minimize the latter. Our analysis shows that this is not suited for improving overlap ; instead we argue for minimizing the confounding error while *maximizing* the balancing score error. Further, while we focused on the *variance* of estimators of the representation-wise estimand $\bar{\tau}^\phi$, it would also be desirable to characterize their *bias* wrt that estimand. Thus, future work should find representations optimally balancing three terms : the difference between the original and representation-wise estimands (confounding error), the bias of an estimator of the representation-wise estimand, and the variance of this estimator (controlled by overlap).

Acknowledgements

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

References

Adam Breitholtz, Anton Matsson, and Fredrik D Johansson. Unsupervised domain adaptation by learning using privileged information. *arXiv preprint arXiv:2303.09350*, 2023.

Oscar Clivio, Avi Feller, and Christopher C Holmes. Towards representation learning for weighting problems in design-based causal inference. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

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.

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

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.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

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.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: partial identification through smoothness. *arXiv preprint arXiv:2305.11812*, 2024.

Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, 188(1):250–257, 2019.

Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

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.

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.

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.

Niklas Pfister and Peter Bühlmann. Extrapolation-aware nonparametric statistical inference. *arXiv preprint arXiv:2402.09758*, 2024.

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.

Christoph Rothe. Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660, 2017.

Pengzhou Abel Wu and Kenji Fukumizu. β-intact-VAE: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2022.

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.

Towards Principled Representation Learning to Improve Overlap in Treatment Effect Estimation (Supplementary Material)

Oscar Clivio[1] **David Bruns-Smith**[2] **Avi Feller**[2] **Chris Holmes**[1]

[1]University of Oxford
[2]University of California, Berkeley

A PROOF OF PROPOSITION 2.2

For any treatment value a and random variable, let $\tau_a(Z) = \mathbb{E}_a[Y|Z]$ we have

$$\begin{aligned}
\sigma_a^2(\phi(X)) &= \text{Var}(Y|A=a,\phi(X)) \\
&= \mathbb{E}_a\left[(Y-\tau_a(\phi(X)))^2 \,\middle|\, \phi(X)\right] \\
&= \mathbb{E}_a\left[\mathbb{E}_a\left[(Y-\tau_a(\phi(X)))^2 \,\middle|\, X,\phi(X)\right] \,\middle|\, \phi(X)\right] \\
&\qquad \text{from the tower property} \\
&= \mathbb{E}_a\left[\mathbb{E}_a\left[(Y-\tau_a(\phi(X)))^2 \,\middle|\, X\right] \,\middle|\, \phi(X)\right] \\
&\qquad \text{as knowledge of } X \text{ implies knowledge of } \phi(X) \\
&\geq \mathbb{E}_a\left[\mathbb{E}_a\left[(Y-\tau_a(X))^2 \,\middle|\, X\right] \,\middle|\, \phi(X)\right] \\
&\qquad \text{as } \tau_a(X) = \mathbb{E}_a[Y|X] \\
&= \mathbb{E}_a\left[\sigma_a^2(X) \,\middle|\, \phi(X)\right] \\
&\qquad \text{by definition of } \sigma_a^2(X) \\
&\geq \mathbb{E}\left[\underline{\sigma}^2 \,\middle|\, \phi(X)\right] \\
&\qquad \text{by Assumption 2.1} \\
&= \underline{\sigma}^2.
\end{aligned}$$

Then, it is clear that $V_{\text{eff}}(\bar{\tau}^\phi) \geq \underline{\sigma}^2 \sum_{a\in\mathcal{A}} \mathbb{E}_a\left[\frac{1}{e_a(\phi(X))}\right]$. For any $a \in \mathcal{A}$, noting $Z = \phi(X)$,

$$\begin{aligned}
p(a)\cdot\mathbb{E}_a\left[\frac{1}{e_a(\phi(X))}\right] &= \mathbb{E}\left[\frac{p(a)}{p(A=a|Z)}\right] \\
&= \mathbb{E}\left[\frac{p(Z)}{p(Z|A=a)}\right] \text{ from Bayes' rule} \\
&= \int \frac{p(z)}{p(z|A=a)}\cdot p(z)dz \\
&= \int \frac{p(z)}{p(z|A=a)}\cdot\frac{p(z)}{p(z|A=a)}\cdot p(z|A=a)dz \\
&= \mathbb{E}_a\left[\left(\frac{p(Z)}{p(Z|A=a)}\right)^2\right]
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_a\left[\left(\frac{p(Z)}{p(Z|A=a)} - 1\right)^2\right] + 1 \text{ as } \frac{p(Z)}{p(Z|A=a)} \text{ is a density ratio so } \mathbb{E}_a\left[\frac{p(Z)}{p(Z|A=a)}\right] = 1 \\
&= \chi^2_a(\phi) + 1
\end{aligned}$$

which gives the result. □

B PROOF OF PROPOSITION 3.1

From Proposition 3.4 of Clivio et al. [2024],

$$\frac{p(\phi(X))}{p(\phi(X)|a)} = \mathbb{E}_a\left[\frac{p(X)}{p(X|a)}\middle|\phi(X)\right] \quad P(.|a)\text{-a.s.} \tag{1}$$

and from the above,

$$\chi^2_a(\phi) = \mathbb{E}_a\left[\left(\frac{p(\phi(X))}{p(\phi(X)|a)}\right)^2\right] - 1 \tag{2}$$

so

$$\begin{aligned}
\mathcal{O}(X) - \mathcal{O}(\phi) &= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \left(\chi^2_a(X) - \chi^2_a(\phi)\right) \text{ where } \chi^2_a(X) := \chi^2_a(\text{Id}) \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \left(\mathbb{E}_a\left[\left(\frac{p(X)}{p(X|a)}\right)^2\right] - \mathbb{E}_a\left[\left(\frac{p(\phi(X))}{p(\phi(X)|a)}\right)^2\right]\right) \text{ from Equation 2} \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \left(\mathbb{E}_a\left[\mathbb{E}_a\left[\left(\frac{p(X)}{p(X|a)}\right)^2 \middle| \phi(X)\right]\right] - \mathbb{E}_a\left[\left(\frac{p(\phi(X))}{p(\phi(X)|a)}\right)^2\right]\right) \text{ from the tower property} \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \left(\mathbb{E}_a\left[\mathbb{E}_a\left[\left(\frac{p(X)}{p(X|a)}\right)^2 \middle| \phi(X)\right]\right] - \mathbb{E}_a\left[\mathbb{E}_a\left[\frac{p(X)}{p(X|a)}\middle|\phi(X)\right]^2\right]\right) \text{ from Equation 1} \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \mathbb{E}_a\left[\mathbb{E}_a\left[\left(\frac{p(X)}{p(X|a)}\right)^2 \middle| \phi(X)\right] - \mathbb{E}_a\left[\frac{p(X)}{p(X|a)}\middle|\phi(X)\right]^2\right] \text{ from the linearity of the expectation} \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \mathbb{E}_a\left[\text{Var}\left(\frac{p(X)}{p(X|a)} \middle| \phi(X), A=a\right)\right] \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \mathbb{E}_a\left[\mathbb{E}_a\left[\left(\frac{p(X)}{p(X|a)} - \mathbb{E}_a\left[\frac{p(X)}{p(X|a)}\middle|\phi(X)\right]\right)^2 \middle| \phi(X)\right]\right] \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \mathbb{E}_a\left[\left(\frac{p(X)}{p(X|a)} - \mathbb{E}_a\left[\frac{p(X)}{p(X|a)}\middle|\phi(X)\right]\right)^2\right] \text{ from the tower property} \\
&= \sum_{a\in\mathcal{A}} \frac{1}{p(a)} \cdot \mathbb{E}_a\left[\left(\frac{p(a)}{p(a|X)} - \mathbb{E}_a\left[\frac{p(a)}{p(a|X)}\middle|\phi(X)\right]\right)^2\right] \text{ from Bayes' rule} \\
&= \sum_{a\in\mathcal{A}} p(a)\mathbb{E}_a\left[\left(\frac{1}{e_a(X)} - \mathbb{E}_a\left[\frac{1}{e_a(X)}\middle|\phi(X)\right]\right)^2\right],
\end{aligned}$$

which proves the result. Assumption 2.1 simply ensures that all expectations are well-defined □