Journal of Econometrics 221 (2021) 644–654

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Overlap in observational studies with high-dimensional covariates☆

Alexander D'Amour*[,1], Peng Ding, Avi Feller, Lihua Lei[2], Jasjeet Sekhon

UC Berkeley Department of Statistics, Evans Hall, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 10 July 2019
Received in revised form 10 July 2019
Accepted 5 October 2019
Available online 21 August 2020

Keywords:
Causal inference
Overlap
Information theory
Curse of dimensionality

ABSTRACT

Estimating causal effects under exogeneity hinges on two key assumptions: unconfoundedness and overlap. Researchers often argue that unconfoundedness is more plausible when more covariates are included in the analysis. Less discussed is the fact that covariate overlap is more difficult to satisfy in this setting. In this paper, we explore the implications of overlap in observational studies with high-dimensional covariates and formalize curse-of-dimensionality argument, suggesting that these assumptions are stronger than investigators likely realize. Our key innovation is to explore how strict overlap restricts global discrepancies between the covariate distributions in the treated and control populations. Exploiting results from information theory, we derive explicit bounds on the average imbalance in covariate means under strict overlap and show that these bounds become more restrictive as the dimension grows large. We discuss how these implications interact with assumptions and procedures commonly deployed in observational causal inference, including sparsity and trimming.

© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Accompanying the rapid growth in administrative databases and online platforms, there has been a push to extend methods for estimating causal effects under exogeneity to settings with high-dimensional covariates (Belloni et al., 2014; Farrell, 2015; Athey et al., 2018). These studies typically require a pair of identifying assumptions (Rosenbaum and Rubin, 1983; Imbens, 2004): *unconfoundedness*, also known as selection on observables, in which the treatment assignment mechanism depends only on observed covariates; and *overlap*, also known as positivity or common support, in which all units have a non-zero probability of assignment to each treatment condition.

A key argument for high-dimensional observational studies is that unconfoundedness is more plausible when the analyst adjusts for more covariates (Rosenbaum, 2002; Rubin, 2009). Setting aside notable counter-examples to this argument (Pearl, 2011; Wooldridge, 2016), the intuition is straightforward to state: the richer the set of covariates, the

☆ We thank Xiaohong Chen, Skip Hirshberg, Elie Tamer, participants at the Atlantic Causal Inference Conference, Stanford University, and Yale University for helpful comments and discussions. PD thanks the National Science Foundation, Grant DMS 1713152 and 1945136. JS thanks Office of Office of Naval Research (ONR) Grants N00014-17-1-2176 and N00014-15-1-2367.

* Corresponding author.
E-mail addresses: alexdamour@google.com (A. D'Amour), pengdingpku@berkeley.edu (P. Ding), afeller@berkeley.edu (A. Feller), lihualei@stanford.edu (L. Lei), sekhon@berkeley.edu (J. Sekhon).

1 Alexander D'Amour is now a Research Scientist at Google Research, Cambridge, MA USA; this work was completed while he was a Neyman Visiting Assistant Professor at UC Berkeley.

2 Lihua Lei is now a postdoctoral fellow in the Department of Statistics at Stanford University, Stanford, CA USA. This work was completed while he was a graduate student in the Department of Statistics at UC Berkeley.

https://doi.org/10.1016/j.jeconom.2019.10.014
0304-4076/© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

more likely that unmeasured confounding variables become measured confounding variables. This intuition, however, has the opposite implications for overlap: the richer the set of covariates, the closer these covariates come to perfectly predicting treatment assignment for at least some subgroups.

We formalize this curse of dimensionality argument and demonstrate that there are strong implications of overlap when there are many covariates. In particular, we focus on the strict overlap assumption, which asserts that the propensity score is bounded away from zero and one with probability one. While this appears to be a local constraint, we show that strict overlap implies global restrictions on the discrepancy between the covariate distributions in the treated and control populations. To do so, we re-frame strict overlap as bounding a likelihood ratio, which is a well-studied problem in information theory (Hellman and Cover, 1970). Adapting results from Rukhin (1997), we derive explicit bounds on various types of covariate imbalance, and show that these bounds become more restrictive as the dimension of the covariates grows. For example, we show that as the dimension of the covariates grows, strict overlap implies that the covariates must either be highly correlated, or that their means must become arbitrarily close to balance on average. To put these results into context, we discuss how the implications of strict overlap intersect with common modeling assumptions, and how our results inform the common practice of trimming in high-dimensional contexts.

We contribute to a growing literature on the critical role of overlap in observational settings. In the context of semiparametric estimators, several papers show that the convergence rate critically depends on the level of overlap (Khan and Tamer, 2010; Hong et al., 2018; Ma and Wang, 2019); see Busso et al. (2014) for relevant simulation evidence. Recognizing this, one common approach is to trim units that have extreme values of the propensity score (Dehejia and Wahba, 1999; Crump et al., 2009; Petersen et al., 2012; Yang and Ding, 2018). An alternative is to instead propose estimators and inference methods that have additional robustness to overlap violations (Chen et al., 2008; van der Laan and Rose, 2011; Chaudhuri and Hill, 2014; Rothe, 2017; Armstrong and Kolesár, 2018; Sasaki and Ura, 2018). Finally, our results are especially relevant for recent efforts to incorporate machine learning into estimating causal effects, partly to exploit rich covariates (see Chernozhukov et al., 2019; Athey and Imbens, 2019, for recent reviews). On the one hand, by using machine learning to perform covariate adjustment, these methods can achieve parametric convergence rates under extremely weak nonparametric modeling assumptions. On the other hand, the cost of this nonparametric flexibility is that these methods are highly sensitive to poor overlap. Thus, understanding the implications of overlap with high-dimensional covariates is therefore critical across many open research areas.

The paper proceeds as follows. Section 2 sets up the problem and defines key notation. Section 3 gives the main results on implications of strict overlap. Section 4 discusses the role of assumptions on the outcome model, such as sparsity, as well as trimming. Section 5 offers some discussion. In separate work, we address possible remedies and methodologies for assessing overlap in this setting, but believe that characterizing the implications of overlap remains of independent interest.

2. Preliminaries

We focus on an observational study with a binary treatment. For each sampled unit i, $(Y_i(0), Y_i(1))$ are potential outcomes, T_i is the treatment indicator, and X_i is the set of covariates. Let $\{(Y_i(0), Y_i(1)), T_i, X_i\}_{i=1}^n$ be independently and identically distributed according to a superpopulation probability measure P. We drop the i subscript when discussing population stochastic properties of these quantities. We observe triples (Y^{obs}, T, X) where $Y^{\text{obs}} = (1-T)Y(0)+TY(1)$. We would like to estimate the average treatment effect

$$\tau^{\text{ATE}} = \mathrm{E}\{Y(1) - Y(0)\},$$

though our results immediately extend to other estimands like the Average Treatment Effect on the Treated.

The standard approach in observational studies is to argue that identification is plausible conditional on a possibly large set of covariates (Rosenbaum and Rubin, 1983; Imbens, 2004). Specifically, the investigator chooses a set of p covariates $X_{1:p} \subset X$, and assumes unconfoundedness.

Assumption 1 (*Unconfoundedness*). $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X_{1:p}$.

Assumption 1 ensures

$$\begin{aligned}\tau^{\text{ATE}} &= \mathrm{E}\Big[\mathrm{E}\{Y(1) \mid X_{1:p}\} - \mathrm{E}\{Y(0) \mid X_{1:p}\}\Big] \\ &= \mathrm{E}\Big[\mathrm{E}\{Y^{\text{obs}} \mid T = 1, X_{1:p}\} - \mathrm{E}\{Y^{\text{obs}} \mid T = 0, X_{1:p}\}\Big]. \end{aligned} \tag{1}$$

Importantly, the conditional expectations in (1) are non-parametrically identifiable only if the following population overlap assumption is satisfied. Let $e(X_{1:p}) = P(T = 1 \mid X_{1:p})$ be the propensity score.

Assumption 2 (*Population Overlap*). $0 < e(X_{1:p}) < 1$ with probability 1.

Assumption 2 is sufficient for non-parametric identification of τ^{ATE}, but is not sufficient for efficient semiparametric estimation of τ^{ATE}, a fact we discuss in further detail in the next section. For this reason, investigators typically invoke a stronger variant of Assumption 2 (e.g., Hirano et al., 2003; Khan and Tamer, 2010), which we call the strict overlap assumption with bound η.

Assumption 3 (*Strict Overlap*)**.** For some constant $\eta \in (0, 0.5)$, $\eta \leq e(X_{1:p}) \leq 1 - \eta$ with probability 1.

Strict overlap is integral across a range of settings. Without any restrictions on the outcome distribution, strict overlap is a necessary condition for the existence of regular semiparametric estimators of τ^{ATE} that are uniformly $n^{1/2}$-consistent over a nonparametric model family (Khan and Tamer, 2010). This necessity may not hold if other conditions, e.g., conditional moment conditions and smoothness conditions, are imposed on the potential outcomes (e.g., Chen et al., 2008; Hirshberg and Wager, 2017; Ma and Wang, 2019). Technically, we can relax Assumption 3, but this will involve non-standard asymptotic analyses (e.g., Hong et al., 2018; Ma and Wang, 2019) and it is difficult, if not impossible, to conduct uniform inference on τ^{ATE} (e.g. Khan and Nekipelov, 2013). Nevertheless, a large body of literature assumes strict overlap, even in the presence of outcome restrictions, as it facilitates theoretical analysis; see, for example, van der Laan and Rose (2011) and Chernozhukov et al. (2019). Moreover, as Khan and Nekipelov (2013, Section 4.1) observe, it is not clear how to conduct uniform inference without strict overlap conditions, except in corner cases. Indeed, Khan and Nekipelov (2013) prove that neither bootstrap inference nor pivotal inference is asymptotically valid without this assumption. Ma and Wang (2019) shed some light on the possibility of uniform inference under assumptions on tail behaviors of inverse propensity scores though they do not provide a complete recipe. In general, a lack of uniform inference is problematic in practice, even if we can characterize the limiting behavior for every data generating distribution in a model, because the correct choice of inferential procedure will depend on the unknown truth. See, e.g., Romano and Wolf (1999), Andrews and Cheng (2012, 2013), and Chen et al. (2011) for discussion in other contexts.

3. Implications of strict overlap

3.1. Framework

In this section, we show that strict overlap restricts the overall discrepancy between the treated and control covariate measures, and that this restriction becomes more binding as the dimension p increases. Formally, we write the control and treatment measures for covariates, for all p, as:

$$P_0(X_{1:p} \in A) := P(X_{1:p} \in A \mid T = 0),$$
$$P_1(X_{1:p} \in A) := P(X_{1:p} \in A \mid T = 1).$$

For the remainder of the paper, we will assume that the marginal probability that any unit is assigned to treatment, $\pi := P(T = 1)$, is bounded by $\eta \leq \pi \leq 1 - \eta$. With a slight abuse of notation, we define the marginal probability measure on covariates, implied by the superpopulation distribution, as $P = \pi P_1 + (1 - \pi)P_0$, a mixture of the condition-specific probability measures P_0 and P_1.

We write the densities of P_1 and P_0 with respect to the dominating measure P as $\mathrm{d}P_1/\mathrm{d}P$ and $\mathrm{d}P_0/\mathrm{d}P$. We write the marginal probability measures of finite-dimensional covariate sets $X_{1:p}$ as $P_0(X_{1:p})$ and $P_1(X_{1:p})$, and the marginal densities as $\mathrm{d}P_1/\mathrm{d}P(X_{1:p})$ and $\mathrm{d}P_0/\mathrm{d}P(X_{1:p})$. When discussing density ratios, we will omit the dominating measure $\mathrm{d}P$.

By Bayes' Theorem, Assumption 3 is equivalent to the following bound on the density ratio between P_1 and P_0, which we will refer to as a likelihood ratio:

$$b_{\min} \leq \frac{\mathrm{d}P_1(X_{1:p})}{\mathrm{d}P_0(X_{1:p})} \leq b_{\max}, \tag{2}$$

where

$$b_{\min} := \frac{1-\pi}{\pi}\frac{\eta}{1-\eta}, \qquad b_{\max} := \frac{1-\pi}{\pi}\frac{1-\eta}{\eta}. \tag{3}$$

Implications of bounded likelihood ratios are well-studied in information theory (Hellman and Cover, 1970; Rukhin, 1993, 1997). Each of the results that follow are applications of a theorem due to Rukhin (1997), which relates likelihood ratio bounds of the form (2) to upper bounds on certain divergences measuring the discrepancy between the distributions $P_0(X_{1:p})$ and $P_1(X_{1:p})$. We include an adaptation of Rukhin's theorem in the Appendix, as Theorem 2. We also derive additional implications of this result in the Appendix.

In the subsequent, we explore the implications of Assumption 3 when there are many covariates. To do so, we set up an analytical framework in which the covariate sequence X is a stochastic process $(X_{(k)})_{k>0}$. For any single problem, the investigator selects a finite set of covariates $X_{1:p}$ from the infinite pool of covariates $(X_{(k)})_{k>0}$. Importantly, this framework includes no notion of sample size because we are examining the population-level implications of an assumption about the population measure P. Our results are independent of the number of samples that an investigator might draw from this population.

Remark 1 (*Strict Overlap and Gaussian Covariates*)**.** While we focus on the implications of strict overlap in high dimensions, this assumption also has surprising implications in low dimensions. For example, if X is one-dimensional and follows a Gaussian distribution under both P_0 and P_1, strict overlap implies that $P_0 = P_1$, or that the covariate is perfectly balanced. This is because if $P_0 \neq P_1$, the log-density ratio $\log \mathrm{d}P_0/\mathrm{d}P_1(X)$ diverges for values of X with large magnitude, implying

that $e(X)$ can be arbitrarily close to 0 or 1 with positive probability. Similar results can be derived when $X_{1:p}$ is multi-dimensional Gaussian. Thus, for Gaussianly distributed covariates, the implications of strict overlap are so strong that they are uninteresting. For this reason, we do not give any examples of the implications of the strict overlap assumption when the covariates are Gaussian.

3.2. Strict overlap implies bounded mean discrepancy

We now use these bounds to derive concrete implications of strict overlap. Here, we show that strict overlap implies a strong restriction on the discrepancy between the means of $P_0(X_{1:p})$ and $P_1(X_{1:p})$. In particular, when p is large, strict overlap implies that either the covariates are highly correlated under both P_0 and P_1, or the average discrepancy in means across covariates is small.

We represent the expectations and covariance matrices of $X_{1:p}$ under P_0 and P_1 as follows:

$$\mu_{0,1:p} \coloneqq (\mu_{0,(1)}, \ldots, \mu_{0,(p)}) \coloneqq \mathrm{E}_{P_0}(X_{1:p}), \qquad \Sigma_{0,1:p} \coloneqq \mathrm{var}_{P_0}(X_{1:p}),$$
$$\mu_{1,1:p} \coloneqq (\mu_{1,(1)}, \ldots, \mu_{1,(p)}) \coloneqq \mathrm{E}_{P_1}(X_{1:p}), \qquad \Sigma_{1,1:p} \coloneqq \mathrm{var}_{P_1}(X_{1:p}).$$

We use $\|\cdot\|$ to denote the Euclidean norm of a vector, and $\|\cdot\|_{\mathrm{op}}$ to denote the operator norm of a matrix.

Theorem 1. *Assumption 3 implies*

$$\|\mu_{0,1:p} - \mu_{1,1:p}\| \leq \min\left\{\|\Sigma_{0,1:p}\|_{\mathrm{op}}^{1/2} \cdot B_{\chi^2(1\|0)}^{1/2}, \quad \|\Sigma_{1,1:p}\|_{\mathrm{op}}^{1/2} \cdot B_{\chi^2(0\|1)}^{1/2}\right\}, \tag{4}$$

where $b_{\min}$ and $b_{\max}$ are defined in (3), and

$$B_{\chi^2(1\|0)} = (1 - b_{\min})(b_{\max} - 1), \qquad B_{\chi^2(0\|1)} = (1 - b_{\max}^{-1})(b_{\min}^{-1} - 1)$$

are free of p.

The proof is included in the Appendix. Theorem 1 has strong implications when p is large. These implications become apparent when we examine how much each covariate mean can differ, on average, under (4).

Corollary 1. *Assumption 3 implies*

$$p^{-1} \sum_{k=1}^{p} \left|\mu_{0,(k)} - \mu_{1,(k)}\right| \leq p^{-1/2} \min\left\{\|\Sigma_{0,1:p}\|_{\mathrm{op}}^{1/2} \cdot B_{\chi^2(1\|0)}^{1/2}, \quad \|\Sigma_{1,1:p}\|_{\mathrm{op}}^{1/2} \cdot B_{\chi^2(0\|1)}^{1/2}\right\}. \tag{5}$$

The mean discrepancy bounds in Theorem 1 and Corollary 1 depend on the operator norms of the covariance matrices $\Sigma_{0,1:p}$ and $\Sigma_{1,1:p}$. The operator norm is equal to the largest eigenvalue of the covariance matrix and is a proxy for the degree to which the covariates $X_{1:p}$ are correlated. In particular, the operator norm is large relative to the dimension p if and only if a large proportion of the variance in $X_{1:p}$ is contained in a low-dimensional projection of $X_{1:p}$. For example, in the cases where the components of $X_{1:p}$ are independent, or where $X_{1:p}$ are samples from a stationary ergodic process, the operator norm scales like a constant in p. On the other hand, in the case where the variance in $X_{1:p}$ is dominated by a low-dimensional latent factor model, the operator norm scales linearly in p. We treat these examples precisely in the Appendix.

Corollary 1 establishes that strict overlap implies that the average mean discrepancy across covariates is not too large relative to the operator norms of the covariance matrices $\Sigma_{0,1:p}$, and $\Sigma_{1,1:p}$. When p is large, these implications are strong. To explore this, let $(X_{(k)})_{k>0}$ be a sequence of covariates such that for each p, $X_{1:p} \subset (X_{(k)})_{k>0}$. When the smaller operator norm $\min(\|\Sigma_{0,1:p}\|_{\mathrm{op}}, \|\Sigma_{1,1:p}\|_{\mathrm{op}})$ grows more slowly than p, the bound in (5) converges to zero, implying that the covariate means are, on average, arbitrarily close to balance. On the other hand, for the bound to remain non-zero as p grows large, both operator norms must grow at the same rate as p. This is a strong restriction on the covariance structure; it implies that all but a vanishing proportion of the variance in $X_{1:p}$ concentrates in a finite-dimensional subspace under both P_0 and P_1.

Remark 2. Theorem 1 bounds the mean discrepancy of $X_{1:p}$, which is a special case of a bound on any functional discrepancy of the form $\left|\mathrm{E}_{P_0}\{g(X_{1:p})\} - \mathrm{E}_{P_1}\{g(X_{1:p})\}\right|$ for any function $g : \mathbb{R}^p \mapsto \mathbb{R}$ that is measurable and square-integrable under P_0 or P_1. This result is of independent interest, and is included in the Appendix.

3.3. Strict overlap restricts general distinguishability

In addition to bounds on mean discrepancies, strict overlap also implies restrictions on more general discrepancies between $P_0(X_{1:p})$ and $P_1(X_{1:p})$. In this section, we present two additional results showing that strict overlap restricts how well the covariate distributions can be distinguished from each other.

First, we show that Assumption 3 restricts the extent to which $P_0(X_{1:p})$ can be distinguished from $P_1(X_{1:p})$ by any classifier or statistical test. Let $\phi(X_{1:p})$ be a classifier that maps from the covariate support $\mathcal{X}_{1:p}$ to $\{0, 1\}$. We have the following upper bound on the accuracy of any classifier $\phi(X_{1:p})$ when Assumption 3 holds.

Proposition 1. *Let $\phi(X_{1:p})$ be an arbitrary classifier of $P_0(X_{1:p})$ against $P_1(X_{1:p})$. Assumption 3 implies the following upper bound on the accuracy of $\phi(X_{1:p})$:*

$$P(\phi(X_{1:p}) = T) \leq 1 - \eta.$$

Proof. Let

$$\tilde{\phi}(X_{1:p}) = I\{e(X_{1:p}) \geq 0.5\} \tag{6}$$

be the Bayes optimal classifier. The probability of a correct decision from the Bayes optimal classifier is

$$\begin{aligned} P(\tilde{\phi}(X_{1:p}) = T) &= \mathrm{E}\left[P\{\tilde{\phi}(X_{1:p}) = T \mid e(X_{1:p})\}\right] \\ &= \mathrm{E}\left[I\{e(X_{1:p}) \geq 0.5\}e(X_{1:p}) + I\{e(X_{1:p}) < 0.5\}\{1 - e(X_{1:p})\}\right] \\ &= \mathrm{E}\left[\max\left\{e(X_{1:p}), 1 - e(X_{1:p})\right\}\right]. \end{aligned}$$

Assumption 3 immediately implies $P(\tilde{\phi}(X_{1:p}) = T) \leq 1 - \eta$. The conclusion follows because the Bayes optimal classifier $\tilde{\phi}(X_{1:p})$ has the highest accuracy among all classifiers based on the covariate set $X_{1:p}$ (Devroye et al., 1996, Theorem 2.1). □

Asymptotically, by Proposition 1, strict overlap implies that there exists no consistent classifier of P_0 against P_1 in the large-p limit.

Definition 1. A classifier $\phi(X_{1:p})$ is *p-consistent* if and only if $P(\phi(X_{1:p}) = T) \to 1$ as p grows to infinity.

Corollary 2 (*No Consistent Classifier*)**.** *Let $(X_{(k)})_{k>0}$ be a sequence of covariates, and for each p, let $X_{1:p}$ be a finite subset. If Assumption 3 holds as p grows large, there exists no p-consistent test of P_0 against P_1.*

We can characterize the relationship between the dimension p and the distinguishability of $P_0(X_{1:p})$ from $P_1(X_{1:p})$ non-asymptotically by examining the Kullback–Leibler (KL) divergence. The following result is a special case of Theorem 2, included in the Appendix.

Proposition 2 (*KL Divergence Bound*)**.** *Assumption 3 implies*

$$\mathrm{KL}(P_1(X_{1:p}) \| P_0(X_{1:p})) \leq B_{\mathrm{KL}(1\|0)}, \qquad \mathrm{KL}(P_0(X_{1:p}) \| P_1(X_{1:p})) \leq B_{\mathrm{KL}(0\|1)},$$

where

$$\begin{aligned} B_{\mathrm{KL}(1\|0)} &:= \frac{(1 - b_{\min})b_{\max}\log b_{\max} + (b_{\max} - 1)b_{\min}\log b_{\min}}{b_{\max} - b_{\min}}, \\ B_{\mathrm{KL}(0\|1)} &:= -\frac{(1 - b_{\min})\log b_{\max} + (b_{\max} - 1)\log b_{\min}}{b_{\max} - b_{\min}} \end{aligned}$$

are free of p, with $b_{\min}$ and $b_{\max}$ defined in (3).

In the case of balanced treatment assignment with $\pi = 0.5$, $B_{\mathrm{KL}(1\|0)}$ and $B_{\mathrm{KL}(0\|1)}$ have a simple form:

$$B_{\mathrm{KL}(1\|0)} = B_{\mathrm{KL}(0\|1)} = (1 - 2\eta)\left|\log\frac{\eta}{1 - \eta}\right|.$$

Proposition 2 becomes more restrictive for larger values of p. This follows because neither bound in Proposition 2 depends on p, while the KL divergence is free to grow in p. In particular, by the so-called chain rule, the KL divergence can be expanded into a summation of p non-negative terms (Cover and Thomas, 2005, Theorem 2.5.3):

$$\mathrm{KL}(P_1(X_{1:p}) \parallel P_0(X_{1:p})) = \sum_{k=1}^{p} \mathrm{E}_{P_1}\left\{\mathrm{KL}(P_1(X_{(k)} \mid X_{1:k-1}) \parallel P_0(X_{(k)} \mid X_{1:k-1}))\right\}. \tag{7}$$

Each term in (7) is the expected KL divergence between the conditional distributions of the kth covariate $X_{(k)}$ under P_0 and P_1, after conditioning on all previous covariates $X_{1:k-1}$. Thus, each term corresponds to the discriminating information added by $X_{(k)}$, beyond the information contained in $X_{1:k-1}$. In the large-p limit, strict overlap implies that the average unique discriminating information contained in each covariate $X_{(k)}$ converges to zero.

Corollary 3. *Let $(X_{(k)})_{k>0}$ be a sequence of covariates, and for each p, let $X_{1:p}$ be a finite subset of $(X_{(k)})_{k>0}$. As p grows large, Assumption 3 implies*

$$p^{-1}\sum_{k=1}^{p} \mathrm{E}_{P_1}\left\{\mathrm{KL}(P_1(X_{(k)} \mid X_{1:k-1}) \parallel P_0(X_{(k)} \mid X_{1:k-1}))\right\} = O(p^{-1}), \tag{8}$$

and likewise for the KL divergence evaluated in the opposite direction.

By Corollary 3, strict overlap implies that, on average, the conditional distributions of each covariate $X_{(k)}$, given all previous covariates $X_{1:k-1}$, are arbitrarily close to balance. In the special case where the covariates $X_{(k)}$ are mutually independent under both P_0 and P_1, Corollary 3 implies that, on average, the marginal treated and control distributions for each covariate $X_{(k)}$ are arbitrarily close to balance.

4. Strict overlap and modeling assumptions

4.1. Treatment models: Strict overlap with fewer implications

In this section, we discuss how the implications of strict overlap align with common modeling assumptions about the assignment mechanism. We show that certain modeling assumptions already impose many of the constraints that strict overlap implies. Thus, if one is willing to accept these modeling assumptions, strict overlap has fewer unique implications.

We will focus specifically on the class of modeling assumptions that assert that the propensity score $e(X_{1:p})$ is only a function of a sufficient summary of the covariates $b(X_{1:p})$. In this case, overlap in the summary $b(X_{1:p})$ implies overlap in the full set of covariates $X_{1:p}$. Models in this class include sparse models and latent variable models.

Assumption 4 (*Sufficient Condition for Strict Overlap*)**.** There exists some function of the covariates $b(X_{1:p})$ satisfying the following two conditions:

$$X_{1:p} \perp\!\!\!\perp T \mid b(X_{1:p}),$$
$$\eta \leq e_b(X_{1:p}) \leq 1 - \eta,$$

where $e_b(X_{1:p}) \coloneqq P(T = 1 \mid b(X_{1:p}))$.

Here, the variable $b(X_{1:p})$ is a *balancing score* as in Rosenbaum and Rubin (1983). The propensity score is the coarsest balancing score in the sense that there exists some $h(\cdot)$ such that $e(X_{1:p}) = h(b(X_{1:p}))$. Thus, $b(X_{1:p})$ is a sufficient summary of the covariates $X_{1:p}$ for the treatment assignment T, and overlap in $b(X_{1:p})$ is a sufficient condition for overlap in the entire covariate set $X_{1:p}$.

Proposition 3 (*Sufficient Condition Statement*)**.** *Assumption 4 implies Assumption 3.*

Proof. The conclusion follows from $e(X_{1:p}) = P(T = 1 \mid X_{1:p}) = \mathrm{E}\{P(T = 1 \mid X_{1:p}, b(X_{1:p})) \mid X_{1:p}\} = \mathrm{E}\{P(T = 1 \mid b(X_{1:p})) \mid X_{1:p}\} = \mathrm{E}\{e_b(X_{1:p}) \mid X_{1:p}\} = e_b(X_{1:p})$. □

Assumption 4 has some trivial specifications, which are useful examples. At one extreme, we may specify that $b(X_{1:p}) = e(X_{1:p})$. In this case, Assumption 4 is vacuous: there are no restrictions on the form of the propensity score; and strict overlap overall is equivalent to strict overlap with respect to $b(X_{1:p})$. At the other extreme, we may specify $b(X_{1:p})$ to be a constant, i.e., we assume that the data were generated from a randomized trial. In this case, the overlap condition in Assumption 4 holds automatically.

Of particular interest are restrictions on $b(X_{1:p})$ between these two extremes, such as the sparse propensity score model in Example 1. Such restrictions trade off stronger modeling assumptions on the propensity score $e(X_{1:p})$ with weaker implications of strict overlap.[3]

Example 1 (*Sparse Propensity Score*)**.** Consider a study where the propensity score is sparse in the covariate set $X_{1:p}$, so that for some subset of covariates $X_{1:s} \subset X_{1:p}$ with $s < p$,

$$e(X_{1:p}) = e(X_{1:s}).$$

This implies

$$X_{1:p} \perp\!\!\!\perp T \mid X_{1:s},$$

and $e(X_{1:s})$ is a balancing score. In this case, strict overlap in the lower-dimensional $X_{1:s}$ implies strict overlap for $X_{1:p}$. Belloni et al. (2013) and Farrell (2015) propose a specification similar to this, with an "approximately sparse" specification for the propensity score. The approximately sparse specification in these papers is broader than the model defined here, but has similar implications for overlap.

Example 2 (*Latent Variable Model for Propensity Score*)**.** Consider a study where the treatment assignment mechanism is only a function of some possibly multivariate latent variable U, such that

$$X_{1:p} \perp\!\!\!\perp T \mid U.$$

[3] These specifications exclude cases such as deterministic treatment rules or treatment assignment in a Regression Discontinuity Design: even when the covariates are high-dimensional, the information they contain about the treatment assignment is upper bounded by the information contained in $b(X_{1:p})$.

For example, such a structure exists when treatment is assigned only as a function of a latent class or latent factor. In that case, the projection of $e(U) := P(T = 1 \mid U)$ onto $X_{1:p}$ is a balancing score:[4]

$$X_{1:p} \perp\!\!\!\perp T \mid b_U(X_{1:p}), \tag{9}$$

where $b_U(X_{1:p}) := \mathrm{E}\{e(U) \mid X_{1:p}\}$. Due to (9), strict overlap in the latent variable U implies strict overlap in $b_U(X_{1:p})$, which implies strict overlap in $X_{1:p}$ by Proposition 3. Athey et al. (2018) propose a specification similar to this in their simulations, in which the propensity score is dense with respect to observable covariates but can be specified simply in terms of a latent class.

4.2. Outcome models: Identification and estimation with weaker overlap

The average treatment effect can be identified and estimated under weaker overlap conditions if one is willing to make structural assumptions about the data generating process. For example, if one assumes that the conditional expectations of outcomes $\mathrm{E}[Y(0) \mid X_{1:p}]$ and $\mathrm{E}[Y(1) \mid X_{1:p}]$ belong to a restricted class, Hansen (2008) established that τ^{ATE} can be estimated under Assumption 1 and the following assumption.

Assumption 5 (*Prognostic Identification*)**.** There exists some function $r(X_{1:p})$ satisfying the following two conditions

$$(Y(0), Y(1)) \perp\!\!\!\perp X_{1:p} \mid r(X_{1:p}), \tag{10}$$

$$\eta \leq e_r(X_{1:p}) \leq 1 - \eta, \tag{11}$$

where $e_r(X_{1:p}) := P(T = 1 \mid r(X_{1:p}))$.

Modifying Hansen (2008)'s nomenclature slightly, we call $r(X_{1:p})$ a prognostic score. The assumption of strict overlap in a prognostic score $r(X_{1:p})$ in (11) is never more stringent than Assumption 3 with the same η.[5] van der Laan and Gruber (2010) and Luo et al. (2017) propose methodology designed to exploit this sort of structure.

One can also weaken overlap requirements by imposing modeling assumptions on the outcome process via the conditional average treatment effect $\tau(X_{1:p}) := \mathrm{E}[Y(1) - Y(0) \mid X_{1:p}]$. If $\tau(X_{1:p})$ is assumed constant, for example, in the case of the partial linear model (Belloni et al., 2014; Farrell, 2015), then estimation of τ^{ATE} only requires that strict overlap holds with positive probability, rather than with probability 1.

Assumption 6 (*Strict Overlap with Positive Probability*)**.** For some $\delta > 0$,

$$P(\eta \leq e(X_{1:p}) \leq 1 - \eta) > \delta.$$

Here, Assumption 6 is sufficient because the constant treatment effect assumption justifies extrapolation from subpopulations where the treatment effect can be estimated to other subpopulations for which strict overlap may fail. The constant treatment effect assumption can also be used to justify trimming strategies, which we turn to next.

4.3. Trimming

When Assumption 3 does not hold, one can still estimate an average treatment effect within a subpopulation in which strict overlap does hold. This motivates the common practice of trimming, where the investigator drops observations in regions without overlap (Dehejia and Wahba, 1999; Crump et al., 2009; Petersen et al., 2012; Yang and Ding, 2018). In general, trimming changes the estimand unless additional structure, such as a constant treatment effect, is imposed on the conditional treatment effect surface $\tau(X_{1:p})$.[6]

Our results suggest that trimming may need to be employed more often when the covariate dimension p is large, especially in cases where overlap violations result from small imbalances accumulated over many dimensions. In these cases, trimming procedures may have undesirable properties for the same reason that strict overlap does not hold. For example, in high dimensions, one may need to trim a large proportion of units to achieve desirable overlap in the new target subpopulation. The proportion of units that can be retained under a trimming policy designed to achieve overlap bound $\tilde{\eta}$ is related to the accuracy of the Bayes optimal classifier in (6) by the following proposition.

Proposition 4. *For an overlap bound $\tilde{\eta} \in (0, 1/2)$, we have*

$$P\left(\tilde{\eta} \leq e(X_{1:p}) \leq 1 - \tilde{\eta}\right) \leq \left[1 - P\left(\tilde{\phi}(X_{1:p}) = T\right)\right] / \tilde{\eta}.$$

---

[4] The scalar $b_U(X_{1:p}) := \mathrm{E}\{e(U) \mid X_{1:p}\}$ is a balancing score because it is equal to the propensity score $e(X_{1:p}) := P(T = 1 \mid X_{1:p})$. Specifically, $e(X_{1:p}) = P(T = 1 \mid X_{1:p}) = \mathrm{E}\{P(T = 1 \mid X_{1:p}, U) \mid X_{1:p}\} = \mathrm{E}\{P(T = 1 \mid U) \mid X_{1:p}\} = \mathrm{E}\{e(U) \mid X_{1:p}\} = b_U(X_{1:p})$.

[5] By the law of iterated expectations, if $\eta \leq e(X_{1:p}) \leq 1 - \eta$, then $e_r(X_{1:p}) = P(T = 1 \mid r(X_{1:p})) = \mathrm{E}\{P(T = 1 \mid X_{1:p}, r(X_{1:p})) \mid r(X_{1:p})\} = \mathrm{E}\{P(T = 1 \mid X_{1:p}) \mid r(X_{1:p})\} \in [\eta, 1 - \eta]$.

[6] An alternative strategy is to estimate a weighted average of the conditional treatment effect, e.g., $\tau^w = \mathrm{E}\{w(X_{1:p})\tau(X_{1:p})\}$ with $w(X_{1:p}) \propto e(X_{1:p})\{1 - e(X_{1:p})\}$ (Crump et al., 2006; Li et al., 2018). This estimand downweights the units with propensity scores close to zero and one, and can be viewed as a smooth version of trimming. We anticipate that our argument extends to this weighting case as well.

Proof. Define the event $\mathcal{A} := \{\tilde{\eta} \leq e(X_{1:p}) \leq 1 - \tilde{\eta}\}$. The conclusion follows from

$$P(\tilde{\phi}(X_{1:p}) \neq T) \geq P(\mathcal{A})P(\tilde{\phi}(X_{1:p}) \neq T \mid \mathcal{A}) \geq P(\mathcal{A})\tilde{\eta}. \quad \square$$

When large covariate sets $X_{1:p}$ enable units to be more accurately classified in treatment and control, the probability that a unit has an acceptable propensity score becomes small. In this case, a trimming procedure must throw away a large proportion of the sample. In the large-p limit, if the Bayes optimal classifier $\tilde{\phi}(X_{1:p})$ is consistent in the sense of Definition 1, then the expected proportion of the sample that must be discarded to achieve any $\tilde{\eta}$ approaches 1.

5. Discussion

In this paper, we have shown that the strict overlap assumption has strong implications in settings with high-dimensional covariates. In particular, we show that the strict overlap assumption implies that the information distinguishing the treated and control covariate distributions must remain fixed — even as the dimension of the covariates grows. This results in binding, population-level restrictions on the data-generating process. Importantly, techniques such as regularization do not avoid these restrictions, though they are often necessary for estimation with high-dimensional covariates.

Our results suggest that overlap assumptions should be carefully considered when adjusting for rich covariates. First, strict overlap is a testable assumption in the sense that, for any fixed bound η, one can construct finite-sample exact tests (Lei et al., 2020). We explore this in separate work and suggest that such empirical validation should be standard practice in these settings. In addition, in cases where the unconfoundedness assumption is violated, overlap appears to play a key role in bias amplification phenomena that result from adjusting for covariates, such as instruments, that are highly predictive of treatment assignment but not of the outcome (Myers et al., 2011; Pearl, 2010; Ding et al., 2017). As the dimensionality increases, appropriately accounting for these complications is important both from a population and finite-sample perspective.

Appendix A. Strict overlap implies bounded f-divergences

Here, we adapt a theorem from information theory, due to Rukhin (1997), to derive general implications of strict overlap. The theorem states that a likelihood ratio bound of the form (2) implies upper bounds on f-divergences between P_0 and P_1. f-divergences are a family of discrepancy measures between probability distributions defined in terms of a convex function f (Csiszár, 1963; Ali and Silvey, 1966; Liese and Vajda, 2006). Formally, the f-divergence from some probability measure Q_0 to another Q_1 is defined as

$$D_f(Q_1(X_{1:p}) \parallel Q_0(X_{1:p})) := \mathrm{E}_{Q_0}\left[f\left(\frac{\mathrm{d}Q_1(X_{1:p})}{\mathrm{d}Q_0(X_{1:p})}\right)\right],$$

f-divergences are non-negative, achieve a minimum when $Q_0 = Q_1$, and are, in general, asymmetric in their arguments. Common examples of f-divergences include the KL divergence, with $f(t) = t \log t$, and the χ^2- or Pearson divergence, with $f(t) = (t-1)^2$. Here, we restate Rukhin (1997)'s theorem in terms of strict overlap and the bounds defined in (2).

Theorem 2. *Let D_f be an f-divergence such that f is convex and has a minimum at 1. Assumption 3 implies*

$$D_f(P_1(X_{1:p}) \parallel P_0(X_{1:p})) \leq \frac{b_{\max} - 1}{b_{\max} - b_{\min}} f(b_{\min}) + \frac{1 - b_{\min}}{b_{\max} - b_{\min}} f(b_{\max}), \tag{A.1}$$

$$D_f(P_0(X_{1:p}) \parallel P_1(X_{1:p})) \leq \frac{b_{\min}^{-1} - 1}{b_{\min}^{-1} - b_{\max}^{-1}} f(b_{\max}^{-1}) + \frac{1 - b_{\max}^{-1}}{b_{\min}^{-1} - b_{\max}^{-1}} f(b_{\min}^{-1}). \tag{A.2}$$

Proof. Theorem 2.1 of Rukhin (1997) shows that the likelihood ratio bound in (2) implies the bounds in (A.1) and (A.2) when f has a minimum at 1 and is "bowl-shaped", i.e., non-increasing on $(0, 1)$ and non-decreasing on $(1, \infty)$. The "bowl-shaped" constraint is satisfied because f is convex. $\square$

Appendix B. Proof of Theorem 1

B.1. Strict overlap implies bounded functional discrepancy

The proof of Theorem 1 follows from several steps, each of which is of independent interest.

Here, we apply Theorem 2 to show that strict overlap implies an upper bound on functional discrepancies of the form

$$\left|\mathrm{E}_{P_0}\{g(X_{1:p})\} - \mathrm{E}_{P_1}\{g(X_{1:p})\}\right| \tag{B.1}$$

for any function $g : \mathbb{R}^p \mapsto \mathbb{R}$ that is measurable under P_0 and P_1. This result plays a key role in the proof of Theorem 1, but is general enough to be of independent interest.

We establish this bound by applying Theorem 2 to the special case of the χ^2-divergence

$$\chi^2(Q_1(X_{1:p}) \,\|\, Q_0(X_{1:p})) := \mathrm{E}_{Q_0}\left[\left(\frac{\mathrm{d}Q_1(X_{1:p})}{\mathrm{d}Q_0(X_{1:p})} - 1\right)^2\right].$$

Strict overlap implies the following bound on the χ^2-divergence.

Corollary 4. *Assumption 3 implies*

$$\chi^2(P_1(X_{1:p}) \,\|\, P_0(X_{1:p})) \le B_{\chi^2(1\|0)}, \qquad \chi^2(P_0(X_{1:p}) \,\|\, P_1(X_{1:p})) \le B_{\chi^2(0\|1)}, \tag{B.2}$$

where

$$B_{\chi^2(1\|0)} := (1 - b_{\min})(b_{\max} - 1), \qquad B_{\chi^2(0\|1)} := (1 - b_{\max}^{-1})(b_{\min}^{-1} - 1).$$

In the case of balanced treatment assignment with $\pi = 0.5$, $B_{\chi^2(1\|0)}$ and $B_{\chi^2(0\|1)}$ have a simple form: $B_{\chi^2(1\|0)} = B_{\chi^2(0\|1)} = \{\eta(1-\eta)\}^{-1} - 4$.

We now apply Corollary 4 to show that strict overlap implies an explicit upper bound on functional discrepancies of form (B.1). Below we let $\|g\|_{P,q} := \left[\mathrm{E}_P\{|g|^q\}\right]^{1/q}$ denote the q-norm of the function g under measure P.

Corollary 5. *Assumption 3 implies*

$$\left|\mathrm{E}_{P_1}[g(X_{1:P})] - \mathrm{E}_{P_0}[g(X_{1:p})]\right| \le \min\left\{\mathrm{var}_{P_0}^{1/2}(g(X_{1:p})) \cdot B_{\chi^2(1\|0)}^{1/2}, \mathrm{var}_{P_1}^{1/2}(g(X_{1:p})) \cdot B_{\chi^2(0\|1)}^{1/2}\right\}. \tag{B.3}$$

Proof. By the Cauchy–Schwarz inequality,

$$|\mathrm{E}_{P_1}[g(X_{1:p})] - \mathrm{E}_{P_0}[g(X_{1:p})]| = \left|\mathrm{E}_{P_0}\left[(g(X_{1:p}) - C) \cdot \left(\frac{\mathrm{d}P_1(X_{1:p})}{\mathrm{d}P_0(X_{1:p})} - 1\right)\right]\right| \tag{B.4}$$

$$\le \|g(X_{1:p}) - C\|_{P_0,2} \cdot \sqrt{\chi^2(P_1(X_{1:p}) \,\|\, P_0(X_{1:p}))}, \tag{B.5}$$

for any finite constant C. A similar bound holds with respect to the χ^2-divergence evaluated in the opposite direction.

Let $C = \mathrm{E}_{P_0}[g(X_{1:p})]$ then apply (B.5) and Corollary 4. Do the same for $C = \mathrm{E}_{P_1}[g(X_{1:p})]$.

Corollary 5 remains valid even when $\mathrm{var}_{P_0}(g(X_{1:p})) = \mathrm{var}_{P_1}(g(X_{1:p})) = \infty$; in this case, inequality (B.3) holds automatically. □

B.2. Proof of Theorem 1

Theorem 1 is a special case of Corollary 5. In particular, let $g(X_{1:p}) := a'X_{1:p}$, where $a := (\mu_{1,1:p} - \mu_{0,1:p})/\|\mu_{1,1:p} - \mu_{0,1:p}\|$ is a vector of unit length, and apply Corollary 5. $\mathrm{var}_{P_0}(a'(X_{1:p} - \mu_{0,1:p}))$ is upper-bounded by $\|\Sigma_{0,1:p}\|_{\mathrm{op}}$ by definition, and likewise for P_1. The result follows.

Appendix C. Other implications of strict overlap

The decomposition in (B.4) can be used to construct additional upper bounds on the mean discrepancy in g using Hölder's inequality in combination with upper bounds on χ^α-divergences (Vajda, 1973). These bounds give a tighter bound in terms of η, but are functions of higher-order moments of $g(X_{1:p})$. Formally, χ^α-divergences are a class of divergences that generalize the χ^2-divergence (Vajda, 1973):

$$\chi^\alpha(P_1(X_{1:p}) \,\|\, P_0(X_{1:p})) := \mathrm{E}_{P_0}\left[\left|\frac{\mathrm{d}P_1(X_{1:p})}{\mathrm{d}P_0(X_{1:p})} - 1\right|^\alpha\right], \quad (\alpha \ge 1).$$

The χ^α divergence in the opposite direction is obtained by switching the roles of P_0 and P_1.

Theorem 2.1 of Rukhin (1997) implies that, under strict overlap with bound η,

$$\chi^\alpha(P_0(X_{1:p})\|P_1(X_{1:p})) \le B_{\chi^\alpha(0\|1)}, \qquad \chi^\alpha(P_1(X_{1:p})\|P_0(X_{1:p})) \le B_{\chi^\alpha(1\|0)},$$

where

$$B_{\chi^\alpha(0\|1)} := (b_{\max} - 1)(1 - b_{\min})\frac{(1 - b_{\min})^{\alpha-1} + (b_{\max} - 1)^{\alpha-1}}{b_{\max} - b_{\min}},$$

$$B_{\chi^\alpha(1\|0)} := (b_{\min}^{-1} - 1)(1 - b_{\max}^{-1})\frac{(1 - b_{\max}^{-1})^{\alpha-1} + (b_{\min}^{-1} - 1)^{\alpha-1}}{b_{\min}^{-1} - b_{\max}^{-1}}.$$

Applying Hölder's inequality to (B.4), we obtain

$$|\mathrm{E}_{P_1}\{g(X_{1:p})\} - \mathrm{E}_{P_0}\{g(X_{1:p})\}| \leq \min\left\{ \|g(X_{1:p}) - C\|_{P_0,q_\alpha} \cdot B_{\chi^\alpha(1\|0)}^{1/\alpha}, \|g(X_{1:p}) - C\|_{P_1,q_\alpha} \cdot B_{\chi^\alpha(0\|1)}^{1/\alpha} \right\},$$

where $q_\alpha := \alpha/(\alpha-1)$ is the Hölder conjugate of α. Setting $C = \mathrm{E}_{P_0}[g(X_{1:p})]$ establishes a relationship between the q_αth central moment of $g(X_{1:p})$ under P_0 and the functional discrepancy between P_0 and P_1. For small values of η, this bound scales as $\eta^{-1/\alpha}$, whereas (B.3) scales as $\eta^{-1/2}$.

Appendix D. Operator norm

The behavior of the bounds in Theorem 1 and Corollary 1 depends on the operator norm of the covariance matrix under P_0 and P_1. Heuristically, this operator norm is large whenever there is high correlation between the covariates $X_{1:p}$ under the corresponding probability measure. Thus, these bounds on mean imbalance become more restrictive as the dimension grows. Because all points in this discussion apply equally to $\Sigma_{0,1:p}$ and $\Sigma_{1,1:p}$, we will refer to a generic covariance matrix $\Sigma_{1:p}$, which can be taken to be either $\Sigma_{0,1:p}$ or $\Sigma_{1,1:p}$.

In this section, we give several examples of covariance structures and the behavior of their corresponding operator norm as p grows large. In the first two examples, the operator norm is of constant order; in the third example, the growth rate of the operator norm can vary from $O(1)$ to $O(p)$.

Example 3 (*Independent Case*)**.** When the components of $(X_{(k)})_{k>0}$ are independent, with component-wise variance given by σ_k^2, $\|\Sigma_{1:p}\|_{\mathrm{op}} = \max_{1\leq k\leq p} \sigma_k^2$. Thus, if the covariate-wise variances are bounded, the operator norm is $O(1)$.

Example 4 (*Stationary Covariance Case*)**.** When $(X_{(k)})_{k>0}$ is a stationary ergodic process with spectral density bounded by M, $\|\Sigma_{1:p}\|_{\mathrm{op}} \leq M$ (Bickel and Levina, 2004). For example, when $(X_{(k)})_{k>0}$ is an MA(1) process with parameter θ, it has a banded covariance matrix so that all elements on the diagonal $\sigma_{k,k} = \sigma^2$ and all elements on the first off-diagonal $\sigma_{k,k\pm1} = \theta$. In this case, the spectral density is upper bounded by $\sigma^2(1+\theta)^2/(2\pi)$, so the operator norm is $O(1)$.

Example 5 (*Restricted Rank Case*)**.** If $(X_{(k)})_{k>0}$ has component-wise variances given by σ_k^2 and $\Sigma_{1:p}$ has rank s_p, then $\|\Sigma_{1:p}\|_{\mathrm{op}} \geq s_p^{-1}\sum_{k=1}^p \sigma_k^2$, because the maximum eigenvalue of $\Sigma_{1:p}$ must be larger than the average of its non-zero eigenvalues. Thus, if $s_p = s$ is constant in p and the component-wise variances are bounded away from 0 and ∞, the operator norm is $O(p)$. In the special case where $s = 1$, the covariates are perfectly correlated. On the other hand, if s_p is a non-decreasing function of p, then the operator norm grows as $O(p/s_p)$.

Each example shows that if the covariates $X_{1:p}$ are not too correlated, so that $\|\Sigma_{1:p}\|_{\mathrm{op}} = o(p)$, strict overlap implies that the mean absolute discrepancy in (5) converges to zero, and the covariate means approach balance, on average, as p grows large.

References

Ali, S.M., Silvey, S.D., 1966. A general class of coefficients of divergence of one distribution from another. J. R. Stat. Soc. Ser. B Stat. Methodol. 28 (1), 131–142.

Andrews, D.W., Cheng, X., 2012. Estimation and inference with weak, semi-strong, and strong identification. Econometrica 80 (5), 2153–2211.

Andrews, D.W., Cheng, X., 2013. Maximum likelihood estimation and uniform inference with sporadic identification failure. J. Econometrics 173 (1), 36–56.

Armstrong, T.B., Kolesár, M., 2018. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. arXiv preprint arXiv:1712.04594.

Athey, S., Imbens, G., 2019. Machine learning methods economists should know about. Annu. Rev. Econ. 11, 685–725.

Athey, S., Imbens, G.W., Wager, S., 2018. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. J. R. Stat. Soc. Ser. B Stat. Methodol. 80, 597–623.

Belloni, A., Chernozhukov, V., Hansen, C., 2013. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81 (2), 608–650.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 28 (2), 29–50.

Bickel, P.J., Levina, E., 2004. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10 (6), 989–1010.

Busso, M., DiNardo, J., McCrary, J., 2014. New evidence on the finite sample properties of propensity score reweighting and matching estimators. Rev. Econ. Stat. 96 (5), 885–897.

Chaudhuri, S., Hill, J.B., 2014. Heavy Tail Robust Estimation and Inference for Average Treatment Effects. Tech. rep., Working Paper.

Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Statist. 36, 808–843.

Chen, X., Ponomareva, M., Tamer, E., 2011. Likelihood inference in finite mixture models with applications to experimental data. In: Cowles Working Paper.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2019. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21, C1–C68.

Cover, T.M., Thomas, J.A., 2005. Entropy, relative entropy, and mutual information. In: Elements of Information Theory, no. x. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 13–55, chap. 2.

Crump, R., Hotz, V.J., Imbens, G., Mitnik, O., 2006. Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand. National Bureau of Economic Research, Cambridge, Mass., USA.

Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96, 187–199.
Csiszár, I., 1963. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. 8, 95–108.
Dehejia, R.H., Wahba, S., 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. J. Amer. Statist. Assoc. 94, 1053–1062.
Devroye, L., Györfi, L., Lugosi, G., 1996. The Bayes Error. In: Stochastic Modelling and Applied Probability, vol. 31, Springer New York, New York, NY, pp. 9–20.
Ding, P., Vanderweele, T.J., Robins, J.M., 2017. Instrumental variables as bias amplifiers with general outcome and confounding. Biometrika 104 (2), 291–302.
Farrell, M.H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econometrics 189 (1), 1–23.
Hansen, B.B., 2008. The prognostic analogue of the propensity score. Biometrika 95 (2), 481–488.
Hellman, M.E., Cover, T.M., 1970. Learning with finite memory. Ann. Math. Stat. 41 (3), 765–782.
Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.
Hirshberg, D.A., Wager, S., 2017. Augmented minimax linear estimation. arXiv preprint arXiv:1712.00038.
Hong, H., Leung, M.P., Li, J., 2018. Inference on finite population treatment effects under limited overlap. Available at SSRN 3128546.
Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev. Econ. Stat. 86 (1), 4–29.
Khan, S., Nekipelov, D., 2013. On uniform inference in nonlinear models with endogeneity. In: Economic Research Initiatives at Duke (ERID) Working Paper.
Khan, S., Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation. Econometrica 78 (6), 2021–2042.
Lei, L., D'Amour, A., Ding, P., Feller, A., Sekhon, J., 2020. Model-Free Assessment of Population Overlap in Observational Studies. Tech. rep..
Li, F., Morgan, K.L., Zaslavsky, A.M., 2018. Balancing covariates via propensity score weighting. J. Amer. Statist. Assoc. 113 (521), 390–400.
Liese, F., Vajda, I., 2006. On divergences and informations in statistics and information theory. IEEE Trans. Inform. Theory 52 (10), 4394–4412.
Luo, W., Zhu, Y., Ghosh, D., 2017. On estimating regression-based causal effects using sufficient dimension reduction. Biometrika 104 (1), 51–65.
Ma, X., Wang, J., 2019. Robust inference using inverse probability weighting. J. Amer. Statist. Assoc. (in press).
Myers, J.A., Rassen, J.A., Gagne, J.J., Huybrechts, K.F., Schneeweiss, S., Rothman, K.J., Joffe, M.M., Glynn, R.J., 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am. J. Epidemiol. 174 (11), 1213–1222.
Pearl, J., 2010. On a class of bias-amplifying variables that endanger effect estimates. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010). pp. 425–432.
Pearl, J., 2011. Invited commentary: Understanding bias amplification. Am. J. Epidemiol. 174 (11), 1223–1227.
Petersen, M.L., Porter, K.E., Gruber, S., Wang, Y., van der Laan, M.J., 2012. Diagnosing and responding to violations in the positivity assumption. Stat. Methods Med. Res. 21 (1), 31–54.
Romano, J.P., Wolf, M., 1999. Subsampling inference for the mean in the heavy-tailed case. Metrika 50 (1), 55–69.
Rosenbaum, P.R., 2002. Observational Studies. Springer, New York, NY.
Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rothe, C., 2017. Robust confidence intervals for average treatment effects under limited overlap. Econometrica 85 (2), 645–660.
Rubin, D.B., 2009. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? Stat. Med. 28 (9), 1420–1423.
Rukhin, A.L., 1993. Lower bound on the error probability for families with bounded likelihood ratios. Proc. Amer. Math. Soc. 119 (4), 1307.
Rukhin, A.L., 1997. Information-type divergence when the likelihood ratios are bounded. Appl. Math. 24 (4), 415–423.
Sasaki, Y., Ura, T., 2018. Inference for moments of ratios with robustness against large trimming bias and unknown convergence rate. arXiv preprint arXiv:1709.00981.
Vajda, I., 1973. χ^{α}-Divergence and generalized Fisher's information. In: Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes. pp. 873–886.
van der Laan, M.J., Gruber, S., 2010. Collaborative double robust targeted maximum likelihood estimation. Int. J. Biostat. 6, 17.
van der Laan, M.J., Rose, S., 2011. Targeted Learning. Springer, New York, NY.
Wooldridge, J.M., 2016. Should instrumental variables be used as matching variables? Res. Econ. 70 (2), 232–237.
Yang, S., Ding, P., 2018. Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. Biometrika 105, 487–493.