Working Draft
arXiv: arXiv:0000.0000

OVERLAP IN HIGH DIMENSIONAL OBSERVATIONAL STUDIES

By Alexander D'Amour, Peng Ding, Avi Feller and Jasjeet Sekhon

University of California, Berkeley

A key advantage of observational studies with high-dimensional covariates is that the unconfoundedness assumption is often more plausible than in low-dimensional settings. Less discussed is the fact that overlap in covariate distributions (a.k.a., positivity or common support) in the population becomes less plausible with high-dimensional covariates. We show that the overlap assumption in high dimensions is stronger than most investigators realize. In particular, overlap implies bounds on the information in the covariates that discriminates between the covariate distributions in the treated and control populations. These bounds more restrictive the higher the covariate dimension. Under some distributional assumptions, this implies an explicit constraint on the imbalance in covariate means, and in many cases, this bound converges to zero as the dimension grows large. These results are particularly relevant to regular semiparametric estimators of the average treatment effect (ATE), which have recently been adapted for high-dimensional settings, and which rely heavily on the overlap assumption. Given the strength of the overlap assumption in high dimensions, we suggest that (i) tests that can used to validate the overlap assumption and (ii) covariate reduction techniques that can weaken the overlap assumption should be developed alongside high-dimensional estimation methods.

1. Introduction. In causal inference, there has been a recent move to using high-dimensional covariates in observational studies. A central motivation is that unconfoundedness, a key identifying assumption, appears to be more plausible when the analysis conditions on a larger set of covariates [20, 24]. Although there are high-profile counter-examples, notably M-bias [18], the intuition behind this assertion is that the richer the covariate set, the better it can explain away variation due to confounding.

On the other hand, overlap, a second identifying assumption, has largely been ignored. Unfortunately, the same intuition that makes unconfoundedness more plausible with high-dimensional covariates suggests that overlap is less plausible with high-dimensional covariates: the richer the covariate set,

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35
Keywords and phrases: sample, LaTeX 2$_\varepsilon$

the better it can separate treated units from control units when treatment assignment was not explicitly randomized. Consider the following example, inspired by recent efforts to estimate causal effects from observational electronic health record data [26].

EXAMPLE 1 (Deterministic Protocol). Suppose that data are collected where treatment assignment decisions are made by agents using a deterministic rule that varies by every agent. In this case, if the covariates contains all of the inputs that went into the decision including indicators for every agent, then treatment is deterministic conditional on the covariates and overlap in those covariates is not satisfied. For example, in the case of electronic health records, this can occur when each doctor follows a particular deterministic medical protocol given a patient's medical history, and the medical history and doctor's identity are included in the data.

In this paper, we formalize the notion that overlap becomes a stronger assumption the higher the dimension of the covariates. In particular, we show that overlap implies substantive constraints on the discrepancy between the covariate distributions of the treated and control populations, and that these constraints become more restrictive in higher dimensions. The main results of the paper show that, in general, overlap implies bounds on the information that discriminates between the treated and control covariate distributions, and that in special cases, overlap implies bounds in the distance between the means of these distributions.

The central insight underlying our results is that distributional discrepancies accumulate in dimension in the same way that information accumulates in sample size. From this perspective, we should be surprised if overlap between high-dimensional distributions holds in the same way that we are surprised if a no test can distinguish between two distinct hypotheses in large-sample settings. The accumulation of information *across* covariate dimensions, in fact, underlies a number of counterinuitive properties of high-dimensional random vectors [30]. This perspective also suggests that intuitions derived from low-dimensional settings, e.g., acceptable degrees of covariate imbalance in each dimension, do not translate to high-dimensional settings.

The paper proceeds as follows. In Section 2, we present the setting and defines our analytical framework. We present our main results in the next two sections. In Section 3 we show that overlap implies an upper bound on the KL divergence between the treated and control covariate distributions; we then apply this bound to the case of Gaussian covariates. In Section 4 we show that overlap implies a lower bound on the error rate of any test

that attempts to discriminate between the covariate distributions; we then apply this bound to the case of sub-exponential covariates. In Section 5 we discuss assumptions about the treatment assignment mechanism that can weaken the overlap assumption. In Section 6, we discuss implications for high-dimensional estimators, particularly regular semiparametric estimators, which rely heavily on the overlap assumption, and discuss strategies for covariate reduction. We conclude with a discussion of possible future research directions in Section 7.

2. Framework and Preliminaries.

2.1. *Notation and Setup.* 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 a sequence of covariates. Let $\{(Y_i(0), Y_i(1)), T_i, X_i\}_{i=1}^n$ be iid, drawn from a superpopulation. Because of this iid sampling, we drop the i subscript when discussing stochastic properties of these quantities. We observe triples (Y^{obs}, T, X) where $Y^{obs} = (1-T)Y(0) + TY(1)$.

We would like to estimate the average treatment effect

$$\tau^{ATE} = \mathrm{E}[Y(1) - Y(0)].$$

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

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

This approach rests on a pair of assumptions. The relation $(*)$ holds if the following unconfoundedness assumption is satisfied.

Assumption 1 (Unconfoundedness).

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X_{1:p}. \tag{2}$$

This assumption is slightly stronger than is necessary for (*) to hold, but is the most common variant because it allows similar conditioning arguments to estimate more general estimands.

Meanwhile, the conditional expectations in (1) are only identifiable if the following population overlap assumption is satisfied.

ASSUMPTION 2 (Population Overlap). Letting $e(X_{1:p}) = P(T = 1 \mid X_{1:p})$ be the propensity score

$$0 < e(X_{1:p}) < 1 \quad \text{with probability (w. p.) } 1.$$

In this paper, we explore the implications of Assumption 2 when there are many covariates. This setting often arises when an investigator determines that many covariates are necessary for Assumption 1 to be plausible.

2.2. *Analytical Framework.* To explore the role of high dimensional covariates in causal inference, 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 $(X^{(k)})_{k>0}$. This framing makes explicit connections to several results concerning absolute continuity of stochastic processes and concentration of high-dimensional random vectors. However, it differs from the more common setup of causal inference in high dimensions, which considers a sequence of distinct problems indexed by increasing dimension p [3, 4, 9].

We also depart from the standard treatment of high-dimensional causal inference problems by treating the covariates generatively, rather than conditioning on them. Formally, we define control and treatment measures for covariates, for all p, for all A such that $P(A) > 0$:

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

In addition, define $\alpha = P(T = 1)$ as the marginal probability that any unit is assigned to treatment. For the remainder of the paper, we will assume that $\alpha > 0$. The relationship between the overall probability measure P and the condition-specific probability measures P_0 and P_1 is given by the mixture

$$P = \alpha P_1 + (1 - \alpha) P_0.$$

Finally, we write the densities of P_1 and P_0 with respect to the dominating measure P as $dP_1(X_{1:p})$ and $dP_0(X_{1:p})$.

With this formalism, we restate the overlap assumption in likelihood ratio form.

ASSUMPTION 3 (Overlap, Likelihood Ratio Form).

$$0 < \frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} < \infty \quad \text{w.p. } 1.$$

For the remainder of the paper, we will focus on a stricter variant of the overlap assumption that is stronger than necessary for identification, but is usually made to support estimation results and to guarantee asymptotic rates of convergence. For example, this assumption (or something similar) is needed to guarantee that the semiparametric efficiency bound for estimating the ATE is finite.

ASSUMPTION 4 (Strict Overlap, Likelihood Ratio Form). For some $\eta \in (0, 0.5)$, for each $A \in \sigma(X_{1:p})$ such that $P(A) > 0$,

$$\frac{\eta}{1-\eta} < \frac{\alpha}{1-\alpha}\frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} < \frac{1-\eta}{\eta} \quad \text{w.p. 1.} \tag{3}$$

We call η the *bound* of the strict overlap assumption. This assumption is usually stated in terms of the propensity score.

ASSUMPTION 5 (Strict Overlap, Propensity Score Form). For some constant $\eta \in [0, 0.5)$,

$$\eta < e(X_{1:p}) < 1 - \eta \quad \text{w.p. 1.}$$

In the next two sections, we show that Assumption 5 has strong implications for high dimensional covariates.

3. Strict Overlap Upper Bounds KL Divergence. This section is the first of two sections that present our main results. In both sections, we begin with an information bound, or a general result showing that strict overlap implies that there is not too much information that discriminates between the covariate probability measures $P_0(X_{1:p})$ and $P_1(X_{1:p})$. These information bounds are fully general, and apply for any probability measures P_0 and P_1. We then apply each information bound to a case where the covariate measures P_0 and P_1 satisfy certain distributional restrictions, and show that in these cases, strict overlap bounds the imbalance in covariate means.

In this section, the information bound states that strict overlap implies an upper bound on the KL divergence between $P_0(X_{1:p})$ and $P_1(X_{1:p})$ in both directions. We apply this result to the case where the covariates $X_{1:p}$ have a multivariate Gaussian distribution under both P_0 and P_1.

3.1. *Information Bound: KL Divergence.* We now show that strict overlap bounds the discriminating information between $P_0(X_{1:p})$ and $P_1(X_{1:p})$

as measured by KL divergence. Formally, the KL divergence from some probability measure Q_0 to another Q_1 is defined as

$$KL(Q_0(X_{1:p}) \| Q_1(X_{1:p})) := \mathrm{E}_{Q_0}\left[\log \frac{dQ_0(X_{1:p})}{dQ_1(X_{1:p})}\right], \tag{4}$$

where dQ_0 and dQ_1 are densities defined with respect to any common dominating measure, say, $Q = \frac{1}{2}Q_0 + \frac{1}{2}Q_1$. KL divergence is non-negative, and is zero if and only if $Q(X_{1:p}) = Q'(X_{1:p})$; for this reason, it is often interpreted as a measurement of the discrepancy between $Q(X_{1:p})$ and $Q'(X_{1:p})$. KL divergence is not symmetric in its arguments.

Theorem 1 shows that strict overlap bounds the KL divergence between $P_0(X_{1:p})$ and $P_1(X_{1:p})$ (in both directions). This bound does not depend on the dimension p and therefore becomes tighter as p grows.

THEOREM 1. The strict overlap assumption with bound η implies the following statement, and the corresponding statement when P_0 and P_1 are switched:

$$KL(P_0(X_{1:p}) \| P_1(X_{1:p})) < \left|\log \frac{\eta}{1-\eta}\right| + \left|\log \frac{\alpha}{1-\alpha}\right|. \tag{5}$$

This theorem follows almost immediately from the statement of Assumption 4. The first term of the bound shows that the information bound approaches 0 as the strict overlap bound η approaches 0.5, which would imply that T is randomly assigned. The second term vanishes when treatment assignment is balanced ($\alpha = 0.5$).

Importantly, neither term in the bound depends on p, in contrast to KL divergence, which can only grow in p. In particular, the KL divergence can be expanded into a summation of p non-negative terms [7, Theorem 2.5.3]:

$$KL(P_0(X_{1:p}) \| P_1(X_{1:p})) = \sum_{k=1}^{p} \mathrm{E}_{P_0} KL(P_0(X^{(k)} \mid X_{1:k-1}) \| P_1(X^{(k)} \mid X_{1:k-1})). \tag{6}$$

Thus, the bound in Theorem 1 becomes increasingly tighter as p increases.

We can also use (6) to assess the discriminating information in the kth covariate, $X^{(k)}$, after conditioning on all previous covariates $X_{1:k-1}$. Specifically, each term in (6) is the expected KL divergence between the conditional distributions of the covariate $X^{(k)}$ under P_0 and P_1. Thus, Theorem 1 also bounds the average unique discriminating information contained in each covariate $X^{(k)}$; this upper bound converges to zero as p grows large.

COROLLARY 1. 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, strict overlap with fixed bound η implies

$$\frac{1}{p}\sum_{k=1}^{p} \mathrm{E}_{P_0} KL(P_0(X^{(k)} \mid X_{1:k-1}) \| P_1(X^{(k)} \mid X_{1:k-1})) \to 0. \tag{7}$$

Thus, strict overlap implies that, on average, the discrepancy between conditional distributions of distributions must vanish as p grows large. In the special case where the covariates $X^{(k)}$ are mutually independent under both P_0 and P_1, Corollary 1 implies that, on average, the marginal treated and control distributions for the covariates $X^{(k)}$ are arbitrarily close to balance.

3.2. *Application: Bound on Mean Imbalance in Gaussian Case.* Here, we apply Theorem 1 to show that strict overlap implies explicit upper bounds on discrepancies between covariate means in the treated and control populations when the covariates $X_{1:p}$ are multivariate Gaussian under P_0 and P_1:

$$X_{1:p} \mid T = 1 \sim N(\mu_{1,1:p}, \Sigma_{1,1:p}) \quad \text{and} \quad X_{1:p} \mid T = 0 \sim N(\mu_{0,1:p}, \Sigma_{0,1:p}). \tag{8}$$

In this section, we use $\|\cdot\|$ to denote the Euclidean norm.

In this case, the KL divergence has the following form [17]:

$$\begin{aligned} &KL(P_1(X_{1:p}) \| P_0(X_{1:p})) \\ &\quad = \frac{1}{2}\left[(\mu_{0,1:p} - \mu_{1,1:p})' {\Sigma_{1,1:p}}^{-1} (\mu_{0,1:p} - \mu_{1,1:p}) - \log \frac{|\Sigma_{0,1:p}|}{|\Sigma_{1,1:p}|} + \left(\mathrm{tr}(\Sigma_{1,1:p}^{-1}\Sigma_{0,1:p}) - p\right)\right]. \end{aligned}$$

In combination with Theorem 1, direct manipulation of this expression yields the following constraint on the discrepancy between $\mu_{0,1:p}$ and $\mu_{1,1:p}$. For simplicity, we will only consider the KL divergence in one direction. The following results remain valid when the roles of $(\mu_{1,1:p}, \Sigma_{1,1:p})$ and $(\mu_{0,1:p}, \Sigma_{0,1:p})$ are reversed.

THEOREM 2 (Gaussian Mean Mahalanobis Distance Bound). Assume (8). Then, strict overlap with bound η implies that the Mahalanobis distance with respect to $\Sigma_{1,1:p}$ between the means $\mu_{0,1:p}$ and $\mu_{1,1:p}$ is bounded by

$$\left\|\Sigma_{1,1:p}^{-1/2}(\mu_{0,1:p} - \mu_{1,1:p})\right\| \leq \sqrt{2\left(\left|\log\frac{\eta}{1-\eta}\right| + \left|\log\frac{\alpha}{1-\alpha}\right|\right)}. \tag{9}$$

To more easily represent the relationship between dimension and the upper bound on mean discrepancy in Theorem 2, we convert the Mahalanobis distance bound to an upper bound on mean absolute discrepancy between the covariate means.

COROLLARY 2. Let $\|\Sigma_{1,1:p}^{1/2}\|_{op}$ be the operator norm (i.e., largest singular value) of $\Sigma_{1,1:p}^{1/2}$.

$$\frac{1}{p}\sum_{i=1}^{p}\left|\mu_0^{(k)}-\mu_1^{(k)}\right| \leq \|\Sigma_{1,1:p}^{1/2}\|_{op}p^{-1/2}\sqrt{2\left(\left|\log\frac{\eta}{1-\eta}\right|+\left|\log\frac{\alpha}{1-\alpha}\right|\right)}.$$

Figure 1 gives a visual illustration of how this constraint manifests in high dimensions. When the difference in means in each dimension is fixed as the dimension grows, the covariate distributions become clearly separable.

We can characterize the conditions under which Corollary 2 becomes more restrictive as p increases. A particularly important case is the setting in which the bound converges to zero, which occurs if the largest eigenvalue of $\Sigma_{1,1:p}$ does not grow too fast. In that case, strict overlap implies that, for large p, the mean absolute distance in covariate means converges to zero.

To explore this, let $(X^{(k)})_{k>0}$ be a sequence of covariates so that, for any p, a finite covariate set $X_{1:p}$ is multivariate Gaussian under P_0 and P_1. Then the bound in Corollary 2 behaves as

$$\frac{1}{p}\sum_{i=1}^{p}\left|\mu_0^{(k)}-\mu_1^{(k)}\right| \leq O(\|\Sigma_{1,1:p}^{1/2}\|_{op}p^{-1/2}) \tag{10}$$

and converges to zero when $\|\Sigma_{1,1:p}^{1/2}\|_{op} = o(p^{1/2})$.

By definition, the operator norm $\|\Sigma_{1,1:p}^{1/2}\|_{op}$ is the square root of the largest eigenvalue of $\Sigma_{1,1:p}$. In the context of Gaussian data, $\|\Sigma_{1,1:p}^{1/2}\|_{op}$ is the maximum standard deviation of any one-dimensional projection of $X_{1:p}$. Thus, the operator norm grows more slowly than $p^{1/2}$ if the number of non-degenerate orthogonal projections of $X_{1:p}$ grows with p.

We give several examples of covariance structures and the behavior of their corresponding operator norm. We begin with two examples where the operator norm is of constant order, and one where the operator norm grows as $O(p^{1/2})$.

EXAMPLE 2 (Independent Case). When the components of $(X^{(k)})_{k>0}$ are independent, with component-wise variance given by σ_k^2, $\|\Sigma_{1,1:p}^{1/2}\|_{op} =$

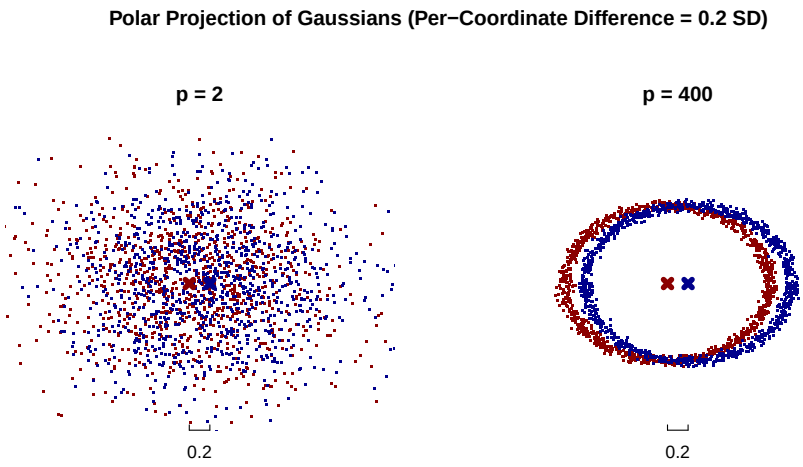


Fig 1: Representation of draws from pairs of high-dimensional Gaussian distributions in low and high dimensions, in the style of [30]. In both cases, the per-covariate discrepancy $p^{1/2}\|\mu_{1,1:p} - \mu_{0,1:p}\| = 0.2$. Each point represents a p-dimensional vector, and each heavy x represent a distribution mean. The coordinates of each point are determined by the normalized distance from the point to its mean ($p^{1/2}\|X_{1:p} - \mu\|$) and the angle from the mean to the point in an arbitrary 2-dimensional plane containing the means of the distributions. This plot preserves the distances of points from their respective means and the distance between the means of the distributions. Distances between points are not preserved. The plot emphasizes that in high dimensions, each distribution approaches a uniform distribution on a shell separated from the mean, and the thickness of the shell relative to the distance between means becomes small if non-negligible deviations are allowed in each dimension.

$\max_k \sigma_k$. Thus, if the covariate-wise variances are bounded, the operator norm is $O(1)$.

EXAMPLE 3 (Stationary Covariance Case). When $(X^{(k)})_{k>0}$ is a stationary ergodic process with bounded spectral density $f < M$, $\|\Sigma_{1,1:p}^{1/2}\|_{op} \leq M^{1/2}$ (Bickel and Levina 2007, 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 $\sigma_{k,k} = \sigma^2$ and $\sigma_{k,k\pm1} = \theta$. In this case, the spectral density is upper bounded by $\frac{\sigma^2}{2\pi}(1+\theta)^2$, so the operator norm is $O(1)$.

EXAMPLE 4 (Fixed Rank Case). If $(X^{(k)})_{k>0}$ has component-wise variances given by σ_k^2 and a fixed number s of factors, so that $\Sigma_{1,1:p}$ has fixed rank s, then $\|\Sigma_{1,1:p}^{1/2}\|_{op} \geq \left(s^{-1}\sum_k \sigma_k^2\right)^{1/2}$. Thus, if the component-wise variance are bounded, the operator norm is $O(p^{1/2})$. In the special case where $s = 1$, the covariates are perfectly correlated.

Thus, if the covariates $X_{1:p}$ are not too correlated, so that $\|\Sigma_{1,1:p}^{1/2}\|_{op} = o(p^{1/2})$, strict overlap implies that the mean absolute discrepancy in (10) converges to zero, and the covariate means approach balance, on average, as p grows large.

A similar result holds for discrepancies in the eigenvalues and eigenvectors of the covariance matrices $\Sigma_{0,1:p}$ and $\Sigma_{1,1:p}$. We reserve this results for the appendix.

4. Strict Overlap Lower Bounds Test Error. In this section, the information bound we derive states that strict overlap implies a lower bound on the error probability of any hypothesis test of the following form:

$$H_0 : X_{1:p} \sim P_0; \quad H_A : X_{1:p} \sim P_1. \tag{11}$$

This corresponds to an intuitive interpretation of strict overlap: when strict overlap holds, one should not be able to guess with a high degree of certainty whether any unit was assigned to treatment or control on the basis of its covariate vector $X_{1:p}$. This error probability bound can be applied more widely than the KL divergence bound in Theorem 1, specifically, in cases where the explicit form of the KL divergence between P_0 and P_1 is not available. In this section, we apply this error probability bound to the case where $X_{1:p}$ are multivariate sub-exponential under both P_0 and P_1, and derive a mean discrepancy bound similar to the one derived in the Gaussian case in Section 3.2. Our approach in this section is similar to approaches that appear in the high-dimensional testing literature, for example, Addario-Berry et al. [1].

4.1. *Information Bound: Test Error.* Formally, let $\phi : \mathbb{R} \mapsto \{0, 1\}$ be a testing procedure of the point hypothesis P_0 against the alternative P_1 (or, equivalently, $T = 0$ against $T = 1$) based on statistic $S(X_{1:p})$. Let $\phi(S(X_{1:p}))$ output 0 if it accepts the hypothesis that $T = 0$ and output 1 if it accepts that hypothesis that $T = 1$. We upper bound the error probability of the test as the maximum of the Type I and Type II errors of the test ϕ.

$$\begin{aligned} \delta_\phi &:= \max\{\text{size}, 1 - \text{power}\} \\ &:= \max\{P(\phi(S(X_{1:p})) = 1 \mid T = 0), P(\phi(S(X_{1:p})) = 0 \mid T = 1)\}. \end{aligned} \tag{12}$$

With this formalism, we state our result.

THEOREM 3 (No test with low error probability). The strict overlap assumption with bound η implies that, for any p, there exists no testing procedure ϕ of $P_0(X_{1:p})$ against $P_1(X_{1:p})$ such that $\delta_\phi < \eta^*$, where

$$\eta^* := \text{logistic}(\text{logit}(\min\{\alpha, 1 - \alpha\}) + \text{logit}(\eta)).$$

REMARK 1. When $\alpha = 0.5$, $\eta^* = \eta$. In all other cases, $\eta^* < \eta$. Here, logit is the log-odds function, and logistic is its inverse.

REMARK 2. For ease of interpretation, the theorem could be restated as follows. Taking p to be the "sample size" of the covariate vector $X_{1:p}$, the strict overlap assumption with bound η implies that any test ϕ with Type I error fixed at η cannot have power greater than $1 - \eta$, regardless of the sample size p. For large p, the power of the test remains bounded away from 1 only if H_A and H_0 are close in all parameters identified by the test statistic $S(X_{1:p})$.

REMARK 3. The lower bound on test error in Theorem 3 is related to but distinct from KL divergence bound stated in Theorem 1. In particular, while the KL divergence result constrains the expectation of the log of the probability ratio in (3), the test error bound constrains the tail area of that ratio.

The proof of Theorem 3 is in the appendix. Theorem 3 is a useful proof device for deriving concrete implications of overlap. In particular, it can be used to convert any test of P_0 against P_1, whether or not it is optimal, into an upper bound on the discrepancy between parameters of P_0 and P_1 identified by the test statistic $S(X_{1:p})$. Moreover, Theorem 3 can be applied in cases where tail bounds on P_0 and P_1 are available, but explicit parametric forms

are not. This generality allows us to derive mean discrepancy bounds for multivariate sub-exponential covariate distributions later in this section.

Like Theorem 1, the implications of Theorem 3 also grow more restrictive as p increases. This is because, for a fixed covariate sequence $(X^{(k)})_{k>0}$, the error probability of an optimal test in terms of δ_ϕ is non-increasing in the dimension of $X_{1:p}$. We can characterize how the constraints behave asymptotically.

Call a test $\phi(S_\phi(X_{1:p}))$ is *consistent* if and only if $\delta_\phi \to_P 0$ as p grows large. Because the lower bound on error probability in Theorem 3 holds even for large p, there can be no consistent test of P_0 against P_1 if strict overlap holds asymptotically.

COROLLARY 3 (No Consistent Test). Let $(X^{(k)})$ be a sequence of covariates, and for each p, let $X_{1:p}$ a finite sub-sequence. If strict overlap with fixed bound η holds as p grows large, there exists no consistent test of P_0 against P_1.

Corollary 3 states how strict overlap restricts the discriminating information contained in $X_{1:p}$ in the limit, but does not explicitly characterize how this constraint becomes more binding for large but finite p. This is difficult to do in general, but with distributional assumptions, we can characterize how Theorem 3 constrains this accumulation along the entire sequence of covariate sets $X_{1:p}$ as p grows. We demonstrate this approach in the context of sub-exponential covariates.

4.2. *Application: Bound on Mean Imbalance in Sub-Exponential Case.* Here, we apply Theorems 3 to show that strict overlap implies an explicit upper bound on covariate mean discrepancies for a more general class of covariate generating processes than we considered in Section 3.2. As in Section 3.2, we will use the following notation for covariate means,

$$\mu_{0_{1:p}} := \mathrm{E}_{P_0}[X_{1:p}] \quad \mu_{1_{1:p}} := \mathrm{E}_{P_1}[X_{1:p}],$$

and we will use $\|\cdot\|$ to denote the Euclidean norm. In this section, we consider the case where $X_{1:p}$ is *multivariate sub-exponential* under both P_0 and P_1 [5, 14]. We say a covariate set $X_{1:p}$ is multivariate sub-exponential if all one-dimensional projections of $X_{1:p}$ are sub-exponential. Formally, we assume that there exist some finite constants σ_p^2 and b_p, such that for all $a \in \mathbb{R}^p$,

$$\mathrm{E}_{P_0} \exp(a'[X_{1:p} - \mu_{0,1:p}]) \leq \exp\left(\frac{\|a\|^2\sigma_p^2}{2}\right) \quad \text{for} \quad \|a\| \leq \frac{1}{b_p},$$

and likewise for P_1 and $\mu_{1,1:p}$. The projections of multivariate sub-exponential random variables admit tight, non-asymptotic concentration inequalities because they have tails that can be upper bounded by Gaussian or exponential random variables. Examples of multivariate sub-exponential random variables include multivariate Gaussian and multivariate Laplace random variables, as well as multivariate random variables that are bounded in all dimensions.

This following result is an application of Theorem 3 to tests constructed using sub-exponential tail bounds.

THEOREM 4 (Sub-exponential Mean Distance Bound). Let $X_{1:p}$ be multivariate sub-exponential with parameters (σ_p^2, b_p) under both P_0 and P_1.

Strict overlap with bound η implies that

$$\|\mu_{0,1:p} - \mu_{1,1:p}\| \leq \begin{cases} \sqrt{8\sigma_p^2 \log \frac{1}{\eta^*}} & \text{if } \sigma_p^2/b_p^2 > -2\log\eta^* \\ 4b_p \log \frac{1}{\eta^*} & \text{if } \sigma_p^2/b_p^2 \leq -2\log\eta^*. \end{cases} \tag{13}$$

As with the previous result, Theorem 4 can be translated into a bound on the absolute difference in means.

COROLLARY 4. In the same setting as Theorem 4, strict overlap with bound η implies that

$$\frac{1}{p}\sum_{k=1}^{p}\left|\mu_0^{(k)} - \mu_1^{(k)}\right| \leq \begin{cases} \sigma_p p^{-1/2}\sqrt{8\log \frac{1}{\eta^*}} & \text{if } \sigma_p^2/b_p^2 > -2\log\eta^* \\ b_p p^{-1/2} 4\log \frac{1}{\eta^*} & \text{if } \sigma_p^2/b_p^2 \leq -2\log\eta^*. \end{cases} \tag{14}$$

Corollary 4 can be paired with an asymptotic regime to derive asymptotic implications. Let $(X^{(k)})$ be a sequence of covariates so that, for any p, a finite covariate set $X_{1:p}$ is multivariate sub-exponential under P_0 and P_1 with parameters (σ_p^2, b_p). Then the bound on mean absolute deviation in means behaves as

$$\frac{1}{p}\sum_{k=1}^{p}\left|\mu_0^{(k)} - \mu_1^{(k)}\right| \leq O(\max\{\sigma_p, b_p\}p^{-1/2}) \tag{15}$$

and converges to zero when $\max\{\sigma_p, b_p\} = o(p^{1/2})$.

As in the Gaussian case, the case where the upper bound in (15) approaches zero is of particular interest, because it implies that for large p, the vast majority of the covariate means are arbitrarily close to balance.

The asymptotic behavior of the bound on Corollary 4 is determined by the scaling of the sub-exponential parameters (σ_p^2, b_p). These are analogous

to the operator norm in the Gaussian case; in fact, in the special case where $X_{1:p}$ is Gaussian, $\sigma_p = \|\Sigma_{1:p}^{1/2}\|_{op}$. In the general multivariate sub-exponential case, σ_p and b_p are the standard deviations of the Gaussian and exponential random variables used to upper bound the tails of the one-dimensional projections of $X_{1:p}$. σ_p and b_p are of constant order when the individual covariates $X^{(k)}$ are independent, but grow when the covariates are correlated so that an increasing proportion of the total variance in $X_{1:p}$ aligns with a particular projection. Thus, as in the Gaussian case, if the components of the sequence $(X^{(k)})_{k>0}$ are not too correlated, the upper bound on the mean absolute discrepancy in (15) approaches zero, and the strict overlap assumption implies ha as p grows large, most covariate means approach balance.

4.3. *Exceptions to Mean Bounds.* Sub-exponential random variables contain a large class of realistic data generating processes for covariates, so the mean discrepancy bounds derived in this section apply in a wide variety of circumstances. However, many common covariate generating processes also fall outside of the sub-exponential class. Mixture distributions are a clear case where the sub-exponential assumption is not met and the mean discrepancy bounds to not apply. We discuss such an exception in the following example.

EXAMPLE 5 (Mixture exception to mean difference bounds). Consider a case where each unit is a member of a latent class, say b_1 or b_2, denoted by a categorical variable B with two levels $\{b_1, b_2\}$, and that

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

Suppose that the distribution of B differs under treatment and control,

$$P_0(B = b_1) = \pi_0 \quad P_1(B = b_1) = \pi_1,$$

and the conditional distribution of covariates given latent class also differs by class membership

$$P_{b_1}(X_{1:p}) = P(X_{1:p} \mid B = b_1) \quad P_{b_2}(X_{1:p}) = P(X_{1:p} \mid B = b_2).$$

Then $P_0(X_{1:p})$ and $P_1(X_{1:p})$ are mixtures of $P_{b_1}(X_{1:p})$ and $P_{b_2}(X_{1:p})$,

$$\begin{aligned} P_0(X_{1:p}) &= \pi_0 P_{b_1}(X_{1:p}) + (1 - \pi_0) P_{b_2}(X_{1:p}) \\ P_1(X_{1:p}) &= \pi_1 P_{b_1}(X_{1:p}) + (1 - \pi_1) P_{b_2}(X_{1:p}). \end{aligned}$$

Note that strict overlap holds if and only if

$$\frac{\eta}{1-\eta} < \frac{\pi_1}{\pi_0} < \frac{1-\eta}{\eta}.$$

In this case, the difference between control and treated means $\mu_{0,1:p}$ and $\mu_{1,1:p}$ is a function of the difference between the class means $\mu_{b_1,1:p}$ and $\mu_{b_2,1:p}$. In particular,

$$\|\mu_{0,1:p} - \mu_{1,1:p}\| = |\pi_0 - \pi_1| \cdot \|\mu_{b_1,1:p} - \mu_{b_2,1:p}\|.$$

Thus, even if the strict overlap assumption holds, if the discrepancy $\|\mu_{b_1,1:p} - \mu_{b_2,1:p}\|$ is not bounded as p grows large, then neither is the discrepancy between $\mu_{0,1:p}$ and $\mu_{1,1:p}$.

This example can be extended to be a special case of Example 7, which appears in the next section. It is closely related a simulation setting employed in Athey, Imbens and Wager [2].

5. Treatment Assignment Models for Weakening Overlap. Estimation in high-dimensional settings is fundamentally challenging, so making guarantees about estimation methodology requires trading off strengths of various assumptions to characterize settings that are likely to appear in practice. The main results of this paper suggest that overlap should factor prominently in these trade-offs. In this section, we turn to discussing trade-offs between the overlap assumption and modeling assumptions on the treatment assignment mechanism. Specifically, we focus on the modeling assumption that the treatment assignment mechanism is fundamentally low-dimensional, which defines a setting in which the overlap assumption is relatively weak. We give several examples of the assumption and relate them to settings that have been invoked in the literature. Finally, we note that this assumption implies a strong unconfoundedness assumption, which may be of particular concern because weakening this assumption is often the motivation for applying high-dimensional methodology.

5.1. *Low-Dimensional Treatment Assignment and Strict Overlap.* For this discussion, it is useful to relate treatment assignment to the entire covariate sequence $(X^{(k)})_{k>0}$, rather than to the particular covariate set $X_{1:p}$ included in the analysis. A natural choice is the *limiting propensity score*.

DEFINITION 1 (Limiting propensity score). Let $(X^{(k)})_{k>0}$ be a sequence of covariates, and let $\sigma(X_{1:\infty}) \in \mathcal{F}$ be the sigma algebra induced by the

sequence. The limiting propensity score is defined equivalently as

$$\begin{aligned} e(X_{1:\infty}) &:= \lim_{p\to\infty} e(X_{1:p}) \\ &:= P(T = 1 \mid \sigma(X_{1:\infty})). \end{aligned} \tag{16}$$

The limiting propensity score is the closest representation of the true assignment mechanism that can be constructed from the covariate sequence $(X^{(k)})_{k>0}$. Strict overlap with respect to the limiting propensity score

$$\eta < e(X_{1:\infty}) < 1 - \eta$$

implies strict overlap with respect to $e(X_{1:p})$ for any subset $X_{1:p}$. Similarly, imposing modeling assumptions on the limiting propensity score $e(X_{1:\infty})$ also imposes modeling assumptions on $e(X_{1:p})$ for any subset $X_{1:p}$. As we show next, imposing such assumptions can weaken strict overlap.

Modeling assumptions about the limiting propensity score can be used to define settings in which strict overlap is a weak assumption for any $X_{1:p} \subset (X^{(k)})_{k>0}$, even for large p. In particular, modeling assumptions of the following form characterize such settings.

ASSUMPTION 6 (Sufficient Condition for Strict Overlap).

1. There exists some low-dimensional variable B following a particular specification such that, for some f,

$$e(X_{1:\infty}) = f(B).$$

2. Strict overlap holds with respect to B:

$$\eta < P(T = 1 \mid B) < 1 - \eta.$$

The first condition in Assumption 6 implies that B is a balancing variable, satisfying

$$X_{1:p} \perp\!\!\!\perp T \mid B \quad \forall X_{1:p} \subset (X^{(k)})_{k>0}. \tag{17}$$

Assumption 6 has some trivial specifications, but these can be enlightening as examples. At one extreme, we may specify that $B = e(X_{1:\infty}$ and Assumption 6 is vacuous: this puts no restrictions on the form of the limiting propensity score and the strict overlap assumption with respect to B is equivalent to the original strict overlap assumption with respect to $X_{1:p}$. At the other extreme, we may specify B 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 6 holds automatically.

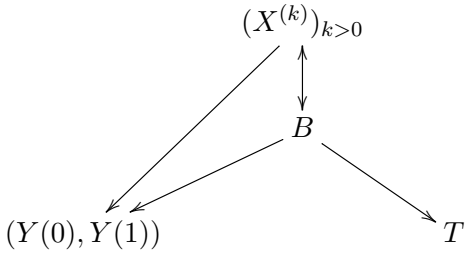
Of particular interest are restrictions on B between these two extremes, such as the sparse propensity score model in Example 6 below. Such restrictions trade off stronger modeling assumptions on the limiting propensity score $e(X_{1:\infty})$ with weaker requirements on strict overlap with respect to low-dimensional B.

Finally, any restrictions on $e(X_{1:\infty})$ exclude cases such as Example 1, in which incorporating more information about the decision process can make the treatment assignment mechanism arbitrarily close to deterministic.

5.2. *Low-Dimensional Treatment Assignment and Unconfoundedness.* While Assumption 6 weakens the strict overlap assumption, it complicates the unconfoundedness assumption. Specifically, if unconfoundedness is to hold under the assumption that B is a balancing variable, then unconfoundedness must hold conditional on B alone, regardless of the dimensionality of $X_{1:p}$.

PROPOSITION 1. Suppose that B is a balancing variable with respect to covariate sequence $(X^{(k)})_{k>0}$. Then unconfoundendess holds with respect to some covariate set $X_{1:p}$ only if unconfoundedness holds with respect to B.

PROOF. This can be shown graphically. Any balancing variable B satisfies the following graphical structure.



If all paths from T to $(Y(0), Y(1))$ are intercepted by some covariate set $X_{1:p} \subset (X^{(k)})_{k>0}$, then by the definition of a balancing random variable, they are also intercepted by B. □

The argument for Proposition 1 follows from the same argument for adjusting for a balancing score rather than the full-dimensional covariates $X_{1:p}$ in the standard setting.

Thus, researchers face a tradeoff between overlap and unconfoundedness when specifying B. On the one hand, strict overlap is typically more plausible

with respect to low-dimensional B versus high-dimensional $X_{1:p}$. On the other hand, unconfoundedness is typically more plausible conditional on high-dimensional $X_{1:p}$ versus low-dimensional B.

5.3. *Examples.* We now give two examples of such models from the literature.

EXAMPLE 6 (Sparse Propensity Score). Consider a study where the limiting propensity score is sparse in the covariate sequence, in the sense that for some subset of covariates $X_{1:s} \subset (X^{(k)})_{k>0}$, and for any covariate set that satisfies $X_{1:s} \subset X_{1:p}$,

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

Stated another way, for any choice of $X_{1:p}$, whether or not it contains X_s,

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

and $X_{1:s}$ is a balancing variable. In this case, overlap in the finite-dimensional $X_{1:s}$ implies overlap for any choice of $X_{1:p} \subset (X^{(k)})_{k>0}$.

Belloni, Chernozhukov and Hansen [3] and Farrell [9] 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 7 (Latent Variable Model for Propensity Score). Consider a study where the limiting propensity score is only a function of some low-dimensional latent covariates $B_{1:s}$, for example, when the propensity score follows a latent class model of the type presented in Example 5, or a latent factor model. In this case,

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

and $B_{1:s}$ is a balancing variable. Thus, overlap in the latent covariates $B_{1:s}$ implies overlap for any choice of $X_{1:p}$.

Athey, Imbens and Wager [2] 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.

6. Implications for High-Dimensional Estimators. In this section, we discuss what our results imply for semiparametric ATE estimators which have recently been adapted to in high-dimensional settings. For these estimators, the strict overlap assumption is necessary for identification, consistency, and asymptotic normality. We then turn to a discussion of covariate reduction strategies that can be used in combination with semiparametric estimation to obtain well-behaved estimators under weaker overlap conditions.

6.1. *Regular Semiparametric Estimators.* In high dimensional settings, semiparametric approaches are essential for estimating the ATE at a parametric rate without making explicit parametric assumptions about the outcome model $\mathrm{E}[Y^{obs} \mid T, X_{1:p}]$ [19]. Recently, several semiparametric estimators for the ATE have been proposed or modified to operate in high-dimensional settings, including Targeted Maximum Likelihood Estimation (TMLE) [29, 28], Double/Debiased Machine Learning (DML) [6], and double selection methods (DS) [3, 9]. These estimators are *regular* semiparametric estimators because they generate consistent, asymptotically normal estimates uniformly over a non-parametric set of outcome models.

The strict overlap assumption is a necessary condition for the desirable behavior of regular semiparametric estimators. Specifically, all regular semiparametric estimators are subject to a lower bound on their variance known as the semiparametric efficiency bound [11, 8]. The form of the bound makes the necessity of the overlap assumption apparent:

$$V^{eff} = \mathrm{E}\left[\frac{\mathrm{var}(Y(1) \mid X_{1:p})}{e(X_{1:p})} + \frac{\mathrm{var}(Y(0) \mid X_{1:p})}{1 - e(X_{1:p})} + (\tau(X_{1:p}) - \tau^{ATE})^2\right]. \tag{18}$$

In this bound, the propensity score appears in the denominator, hence these fractions are unbounded unless the strict overlap assumption holds for some η, that is, $\eta < e(X_{1:p}) < 1 - \eta$ with probability 1.

Our results suggest that the new, high-dimensional settings that general estimation procedures like TMLE and DML unlock are also the settings in which the strict overlap assumption is the strongest. For example, both TMLE and DML are designed to work with any base procedure that can estimate the propensity score $e(X_{1:p})$ and the outcome model $\mathrm{E}[Y^{obs} \mid T, X_{1:p}]$ at a fast-enough rate, expanding the class of high-dimensional settings in which the ATE can be estimated semiperimetrically beyond, e.g., the approximately sparse settings to which DS estimators are confined. Furthermore, DML incorporates a sample-splitting sub-routine, which expands this

class to include high-dimensional propensity score and outcome models of even greater complexity. Overlap is not impossible in these settings, but it is much more likely to be the limiting factor than in more traditional low-dimensional settings. Put simply, in high-dimensional settings, overlap should be a first-order concern.

Thus, as progress continues in high-dimensional estimation, complementary techniques for validating or weakening the overlap assumption should become a priority. To this end, we discuss some potential directions for testing the overlap assumption in Section 7.1. We discuss covariate reduction strategies for estimating the ATE under a weaker overlap assumption next.

6.2. *Covariate Reduction.* Concerns about the plausibility of overlap in high dimensions raise the question of whether the ATE can be estimated with weaker unconfoundedness and overlap assumptions than the sufficient conditions given in Asssumptions 1 and 2. One approach is to reduce the covariates $X_{1:p}$ with a preprocessing function $d(\cdot)$, so that overlap is more plausible in the reduced covariates $d(X_{1:p})$. Here, we outline some requirements for such a covariate reduction scheme.

In principle, the ATE can be identified under weaker conditions than the union of Assumptions 1 and 2.

ASSUMPTION 7 (Functional Identification). Given a set of covariates $X_{1:p}$ such that Assumption 1 is satisfied, there exists some function $d(X_{1:p})$ such that

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid d(X_{1:p}) \tag{19}$$

and overlap are satisfied with respect to $d(X_{1:p})$.

We call a covariate reduction $d(X_{1:p})$ that satisfies (19) a *deconfounding score*. Assumption 7 implies that the ATE can be identified by adjusting for a deconfounding score,

$$\tau^{ATE} = \mathrm{E}[\mathrm{E}[Y^{obs} \mid T = 1, d(X_{1:p})] - \mathrm{E}[Y^{obs} \mid T = 0.d(X_{1:p})]]. \tag{20}$$

Given a deconfounding score, one can specify a non-regular semiparametric estimator for the ATE with respect to the reduced covariates $d(X_{1:p})$. Such an estimator is subject to a semiparametric efficiency bound parameterized in terms of $e(d(X_{1:p})) = P(T = 1 \mid d(X_{1:p}))$ instead of $e(X_{1:p})$. Thus, if strict overlap holds with respect to $d(X_{1:p})$, but not with respect to $X_{1:p}$, the non-regular estimator can be *super-efficient*, meaning it has lower asymptotic variance than the semiparametric efficiency bound for regular estimators [27, 16].

6.3. *Overlap-Weakening Reductions Require Outcome Information.* Not all deconfounding scores weaken the overlap assumption. In particular, no balancing score $b(X_{1:p})$ satisfying

$$T \perp\!\!\!\perp X_{1:p} \mid b(X_{1:p}), \tag{21}$$

including the propensity score $e(X_{1:p})$, can weaken the overlap assumption. A deconfounding score $d(X_{1:p})$ implies a weaker overlap condition only if it discards some of the discriminating information in the unreduced covariates $X_{1:p}$, but balancing score are sufficient for T, and thus retain all discriminating information [21].

Importantly, the constraint that $d(X_{1:p})$ not be a balancing score implies that covariate reductions aimed at weakening the overlap assumption require information about the outcome process. This differs from the case where $d(X_{1:p})$ is a balancing score; in that case the score can be characterized by the sufficiency relationship $T \perp\!\!\!\perp X_{1:p} \mid d(X_{1:p})$, which only involves the T and $X_{1:p}$. When $d(X_{1:p})$ is not a balancing score, $d(X_{1:p})$ must be insufficient for T, and checking the condition that $(Y(0), Y(1)) \perp\!\!\!\perp T \mid d(X_{1:p})$ requires a characterization of the outcome model. In practice, this implies that workflows in which the outcomes are completely masked during a "design" or covariate reduction phase may not be appropriate in high dimensions, and that data splitting may be necessary to minimize moral hazard when constructing covariate reductions that incorporate outcome information Rubin [23].

Several techniques have incorporated or hinted at using non-balancing deconfounding scores. Most of these define the covariate reduction $d(X_{1:p})$ as sufficient statistics for the outcome model $\mathrm{E}[Y^{obs} \mid T, X_{1:p}]$, which can be viewed as prognostic scores in the sense of Hansen [12]. Luo, Zhu and Ghosh [16] proposed methodology most explicitly tailored to this goal, using a sufficient dimension reduction method to identify a linear projection of $X_{1:p}$ that is sufficient for the outcome model. Hill and Su [13] suggested using BART to identify areas of "common causal support", or overlap in sufficient statistics of the outcome model. van der Laan and Gruber [27] proposed the Collaborative TMLE (C-TMLE) method, which uses some of the machinery of semiparametric estimation, but can generate valid estimates of the ATE even if there is no overlap in $X_{1:p}$, but there is overlap in a prognostic score. Practical advice in the propensity score estimation and matching literatures also attempts to capture this idea by recommending that only variables that are related to treatment be incorporated into propensity score models or matching objectives [20, 15]. Roy et al. [22] formalized this approach by specifying a matching procedure that prioritizes covariates based on their

ability to predict outcomes.

We discuss a more general class of deconfounding scores that need not be sufficient for outcome or treatment assignment in Section 7.2.

7. Discussion and Future Work. In this section, we discuss some directions for future work.

7.1. *Future Work for Testing Overlap.* First, we suggest that higher importance should be put on verifying that overlap holds in high dimensions. The mean discrepancy bounds in Theorems 2 and 4 can be tested directly using existing tail bounds on the norms of multivariate sub-exponential random variables [14, 25]. Tests that more directly test KL divergence bound in Theorem 1 would be a useful extension. Similarly, Theorem 3 could be operationalized using novel two sample testing methods that search a family of functions for a test statistic that provides a maximal contrast between treated and control distributions [10].

A simpler, but related testing problem could also be posed for verifying that treatment assignment models satisfy the balancing variable criterion in Assumption 6. In particular, if B is a balancing variable, then it is the case that

$$P_0(X_{1:p}) = \int_{\mathcal{B}} P_1(X_{1:p} \mid B) dP_0(B).$$

In cases where B is estimable, this hypothesis could be tested using two-sample testing methods such as the kernel method Gretton [10], modified to use balancing weights.

7.2. *Future Work for Covariate Reduction.* Second, we suggest a focus on covariate reduction techniques that accept a tradeoff between the high-dimensional overlap assumption and modeling assumptions about deconfounding scores.

In particular, we see a need for methodology that constructs deconfounding scores that are sufficient for neither treatment assignment nor outcomes. Overlap assumptions with respect to these insufficient deconfounding scores are weaker than assuming overlap in propensity scores or prognostic scores. Thus, these scores provide identification even when both treatment assignment and outcome processes are complex, but the structure of confounding is relatively simple.

Exploring the space of deconfounding scores may require novel methodology because the constraint (19), which enforces that the deconfounding score retain unconfoundedness is notably more difficult to work with than the regression constraints that define balancing scores and prognostic scores.

This is because (19) is a ternary relationship between three sets of random variables, instead of a binary relationship between two random variables as in (21).

There is reason to believe that in the context of a particular problem it would be useful to estimate multiple deconfounding scores $d(X_{1:p})$ and to employ an estimation strategy that combines them. Much of the success of doubly robust estimation, particularly in achieving parametric rates of convergence despite employing nonparametric estimation, stems from constructing a doubly robust score function that incorporates two models that would, on their own, be sufficient to identify the ATE. Similar gains in convergence rates may be possible if one can estimate multiple non-redundant deconfounding scores $d(X_{1:p})$.

APPENDIX A: ADDITIONAL GAUSSIAN RESULT

Strict overlap also has implications for covariances.

THEOREM 5 (Gaussian Eigenvalue Near-Balance). Suppose that $X_{1:p}$ is Gaussian under both P_0 and P_1, and that $\mu_{1,1:p} = \mu_{0,1:p} = \mu_{1:p}$ for all p. Suppose that $\Sigma_{0,1:p}$ and $\Sigma_{1,1:p}$ share the same eigenvectors for all p, and let $(\lambda_0^{(k)})_{k=1:p}$ and $(\lambda_1^{(k)})_{k=1:p}$ be the corresponding sets of eigenvalues. Let $(\xi^{(k)})_{k=1:p} = (\lambda_0^{(k)}/\lambda_1^{(k)})_{k=1:p}$ be the ratios of eigenvalues.

Strict overlap with bound η implies, for $h(x) = x - \log x - 1$,

$$\sum_{k=1}^{p} h(\xi^{(k)}) \leq 2\left(\left|\log\frac{\eta}{1-\eta}\right| + \left|\log\frac{\alpha}{1-\alpha}\right|\right). \tag{22}$$

PROOF. Representations of the traces and the determinants of $\Sigma_{1,1:p}$ and $\Sigma_{0,1:p}$ in terms eigenvalues yields the following expression for the KL divergence:

$$KL(P_0\|P_1) = \frac{1}{2}\left[-\sum_{k=1}^{p}\log\frac{\lambda_0^{(k)}}{\lambda_1^{(k)}} + \sum_{i=1}^{p}\left(\frac{\lambda_0^{(k)}}{\lambda_1^{(k)}} - 1\right)\right].$$

□

The function $h(x) = x - \log x - 1$ is non-negative, equal to zero if and only if $x = 1$, and approximately quadratic in the neighborhood of $x = 1$. Thus, the summation grows for each dimension in which the eigenvalues do not match exactly.

APPENDIX B: TEST ERROR PROOFS

PROOF OF THEOREM 3. The strict overlap assumption implies that, for any A such that $P(A) > 0$,

$$\frac{\eta}{1-\eta} < \frac{\alpha}{1-\alpha}\frac{P_1(X_{1:p} \in A)}{P_0(X_{1:p} \in A)} < \frac{1-\eta}{\eta}.$$

This implies, for $\eta^* := \text{logistic}(\text{logit}(\min\{\alpha, 1-\alpha\}) + \text{logit}(\eta))$,

$$\frac{\eta^*}{1-\eta^*} < \frac{P_1(X_{1:p} \in A)}{P_0(X_{1:p} \in A)} < \frac{1-\eta^*}{\eta^*}. \tag{23}$$

For a test ϕ of P_0 against P_1, and test statistic $S(X_{1:p})$, let A_ϕ be the acceptance region for P_0 and $R_\phi = A_\phi^C$ be the rejection region. For readability, for any probability measure P, we write $P(A_\phi)$ and $P(R_\phi)$ as shorthand for $P(S(X_{1:p}) \in A_\phi)$ and $P(S(X_{1:p}) \in R_\phi)$.

Suppose ϕ has error probability η^*, so $\eta^* = \max(P_1(A_\phi), P_0(R_\phi))$. Then there are two cases.

Case 1: $\eta^* = P_1(A_\phi) > P_0(R_\phi)$. This implies

$$\frac{P_1(A_\phi)}{P_0(A_\phi)} = \frac{P_1(A_\phi)}{1 - P_0(R_\phi)} < \frac{\eta^*}{1-\eta^*}. \tag{24}$$

Case 2: $\eta^* = P_0(R_\phi) > P_1(A_\phi)$

$$\frac{P_1(R_\phi)}{P_0(R_\phi)} = \frac{1 - P_1(A_\phi)}{P_0(R_\phi)} > \frac{1-\eta^*}{\eta^*}. \tag{25}$$

Both cases contradict (23). □

PROOF OF THEOREM 4. Let $\Delta_p = \|\mu_{1,1:p} - \mu_{0,1:p}\|$. Define $a = \frac{\mu_0 - \mu_1}{\|\mu_0 - \mu_1\|}$ and test statistic $T(X_{1:p}) = a'(X_{1:p} - \mu_0)$, or the projection $X_{1:p} - \mu_0$ onto the vector $\mu_{0,1:p} - \mu_{1,1:p}$. Under P_0 and P_1, $T(X_{1:p})$ is sub-exponential with parameters (σ_p^2, b_p). Define a test that rejects P_0 whenever $T(X_{1:p}) > \Delta_p/2$, i.e., when $X_{1:p}$ is closer to μ_1 than it is to μ_0, and accepts P_0 otherwise.

The probability of rejecting under P_0 is

$$\delta := P_0(T(X_{1:p}) > \Delta_p/2) \leq \begin{cases} \exp\left(-\frac{\Delta_p^2}{8\sigma_p^2}\right) & \text{for } 0 \leq \Delta_p < \frac{2\sigma_p^2}{b_p} \\ \exp\left(-\frac{\Delta_p}{4b_p}\right) & \text{for } \Delta_p > \frac{2\sigma_p^2}{b_p} \end{cases}$$

and likewise for accepting under P_1, so δ is the error probability of the test. By Theorem 3, strict overlap implies $\delta < \eta^*$.

Solving for Δ_p in each case gives the bound in the result. Further, $\Delta_{p,\eta^*}^* < \frac{2\sigma_p^2}{b_p p^{1/2}}$ if and only if $\sigma_p^2/b_p > -2\log\eta^*$. □

REFERENCES

[1] Addario-Berry, L., Broutin, N., Devroye, L. and Lugosi, G. (2010). On combinatorial testing problems. *Annals of Statistics* **38** 3063–3092.

[2] Athey, S., Imbens, G. W. and Wager, S. (2016). Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. **August** 1–28.

[3] Belloni, A., Chernozhukov, V. and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* **81** 608–650.

[4] Belloni, A., Chernozhukov, V. and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives* **28** 29–50.

[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*. Oxford University Press.

[6] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2016). Double/Debiased Machine Learning for Treatment and Causal Parameters.

[7] Cover, T. M. and Thomas, J. A. (2005). Entropy, Relative Entropy, and Mutual Information. In *Elements of Information Theory* **x** 2, 13–55. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[8] Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199.

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

[10] Gretton, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research* **13** 723–773.

[11] Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica* **66** 315.

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

[13] Hill, J. and Su, Y. S. (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Annals of Applied Statistics* **7** 1386–1420.

[14] Hsu, D., Kakade, S. M. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* **17** 1–6.

[15] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[16] Luo, W., Zhu, Y. and Ghosh, D. (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika* **104** 51–65.

[17] Nielsen, F. and Nock, R. (2009). Clustering Multivariate Normal Distributions. In *Emerging Trends in Visual Computing: LIX Fall Colloquium, ETVC 2008, Palaiseau, France, November 18-20, 2008. Revised Invited Papers* (F. Nielsen, ed.) 164–174. Springer Berlin Heidelberg, Berlin, Heidelberg.

[18] Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge University Press, New York, NY, USA.

[19] Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine* **16** 285–319.

[20] Rosenbaum, P. R. (2002). *Observational Studies. Springer Series in Statistics.* Springer New York, New York, NY.
[21] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **1083** 41–55.
[22] Roy, S., Rudin, C., Volfovsky, A. and Wang, T. (2017). FLAME: A Fast Large-scale Almost Matching Exactly Approach to Causal Inference. 1–23.
[23] Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2** 808–840.
[24] Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine* **28** 1420–1423.
[25] Spokoiny, V. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics* **22** 100–113.
[26] Stuart, E. A., DuGoff, E., Abrams, M., Salkever, D. and Steinwachs, D. (2013). Estimating causal effects in observational studies using Electronic Health Data: Challenges and (some) solutions. *EGEMS (Washington, DC)* **1**.
[27] van der Laan, M. J. and Gruber, S. (2010). Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics* **6**.
[28] van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology* **6**.
[29] van der Laan, M. J. and Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics* **2**.
[30] Vershynin, R. (2017). *High-Dimensional Probability.* Cambridge University Press (to appear).