Working Draft
arXiv: arXiv:0000.0000

OVERLAP IN HIGH DIMENSIONAL OBSERVATIONAL STUDIES
WORKING DRAFT; REVISED: 10/23/2017

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

University of California, Berkeley

Causal inference in observational settings typically rests on a pair of identifying assumptions: (1) unconfoundedness and (2) covariate overlap, also known as positivity or common support. Investigators 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 high-dimensional causal inference, arguing that this assumption is stronger than investigators typically realize. Our main innovation is to frame (strict) overlap in terms of bounds on a likelihood ratio, which allows us to leverage and expand on existing results from information theory. In particular, we show that strict overlap bounds discriminating information (e.g., KL divergence) between the covariate distributions in the treated and control populations. We then use these results to derive explicit bounds on the average imbalance in covariate means under a range of assumptions on the covariate distributions. Importantly, these bounds grow tighter as the dimension grows large, and, in some cases, converge to zero. We also address how restrictions on the treatment assignment and outcome processes can weaken the implications of overlap, but at the cost of stronger requirements for unconfoundedness. These results are particularly relevant for recent proposals to extend regular semiparametric estimators, which rely heavily on overlap, to high-dimensional observational studies.

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 [25, 29]. Although there are high-profile counter-examples, notably M-bias [23], the intuition behind this assertion is that the richer the covariate set, the more plausible it is that the covariates are finer than any confounding variables, and thus the more plausible it is that the covariates can explain away variation due to confounding.

Keywords and phrases: Causal inference, High dimensional covariates, Overlap

On the other hand, the plausibility of overlap, a second identifying assumption stated in Assumption 2, has largely been ignored in high-dimensional settings. The overlap assumption asserts that all units have some probability of being assigned to each treatment condition. This is essential for identifying causal effects without relying on strong modeling assumptions and extrapolation. Unfortunately, the same intuition that makes unconfoundedness more plausible with high-dimensional covariates also suggests that overlap is less plausible with high-dimensional covariates: the richer the covariates, the more likely they can perfectly explain treatment assignment for at least some subgroups of units. This can be a particularly pressing concern when drawing causal inferences from administrative databases, where the covariates are collected for the express purpose of explaining why a particular treatment assignment decision was made. Consider for example example, recent efforts to estimate causal effects from observational electronic health record data [32].

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 explore the implications of overlap in high-dimensional causal inference, arguing that this is assumption much stronger in high dimensions than it is in low dimensions. In particular, we focus on the *strict* overlap assumption (Assumption 4), which asserts that the propensity score is bounded away from 0 and 1 with probability 1. Our main innovation is to frame overlap in terms of bounds on the likelihood ratio that contrasts the covariate distributions in the treated and control populations. Strict overlap imposes a bound on this likelihood ratio that remains fixed even as the dimension of the covariates grows. Because the likelihood ratio is nondecreasing in the dimension of the covariates, this constraint has a number of implications that become stronger in higher dimensions.

In our main results, we build on results from the intersection of information theory and statistics to show that strict overlap induces upper bounds on a family of information divergences that measure discrepancies between the the treated and control covariate distributions. These bounds directly

restrict the imbalance between the covariate distributions, implying that, to satisfy strict overlap in high dimensions, the difference in covariate distributions cannot be too far from the balance achieved in randomized trials.

To make these information divergence bounds concrete, we apply them to derive explicit bounds on the average imbalance in covariate means under a range of assumptions on the covariate distributions. In addition to non-asymptotic bounds, in many cases, we can make statements of the following form,

$$\frac{1}{p}\sum_{k=1}^{p}\left|\mathrm{E}_{P_1}[X^{(k)}] - \mathrm{E}_{P_0}[X^{(k)}]\right| = O(p^{-1/2}),$$

where each $X^{(k)}$ is a covariate dimension and P_0 and P_1 are the covariate probability measures in the treated and control populations. Here, the left-hand side expression represents the average mean discrepancy across covariates. In cases where a result of this type holds, strict overlap implies that the average imbalance in covariate means converges to zero.

These results establish that overlap cannot be taken for granted in high-dimensional settings, and that additional justification, based on the structure of the study at hand, may be necessary to argue that overlap holds in a given observational study. One such justification is that the treatment assignment mechanism or outcome process in a particular study has a fundamentally low-dimensional relationship with the covariates. In such a study, the overlap assumptions necessary for causal identification can be easier to satisfy. However, these assumptions can also make the unconfoundedness assumption more difficult to satisfy. Taken together, our results highlight the fact that adjusting for high-dimensional covariates does not necessarily make causal identification more plausible.

Our results have particular implications for the application of regular semiparametric estimators of the ATE in high dimensions. Many estimators in this class have recently been proposed or modified to operate in high-dimensional settings, including Targeted Maximum Likelihood Estimation (TMLE) [35], Double/Debiased Machine Learning (DML) [8], and double selection methods (DS) [4, 12]. These estimators are *regular* in that they are $n^{1/2}$-consistent and asymptotically normal along any sequence of parametric models that approach the true data-generating process. This general validity comes at a price: the strict overlap assumption is a necessary condition for this desirable behavior.

The paper proceeds as follows. In Section 2, we present the setting and define our analytical framework. We present our main results in the next two sections. In Section 3, we present bounds that the strict overlap as-

sumption imposes on the discriminating information between the treated and control covariate distributions, paying special attention to the KL and χ^2-divergences, and the error rate of any test that attempts to discriminate between the covariate distributions. In Section 4, we apply the results from Section 3 to derive bounds on the mean discrepancy between the treated and control covariate distributions. In Section 5 we discuss restrictions on the treatment assignment or outcome processes that can make overlap assumptions easier to satisfy. We conclude with a discussion of possible future research directions in Section 6.

2. Framework and Preliminaries.

2.1. *Overlap.* 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 [26]. Specifically, the investigator chooses a set of p covariates $X_{1:p} \subset X$, and assumes the relation $(*)$:

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

The functional in (1) identifies τ^{ATE} under the following two assumptions. First, 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.

Second, the conditional expectations in (1) are non-parametrically identifiable only 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. 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 $(X^{(k)})_{k>0}$. This framing makes explicit the investigator's role in choosing a particular subset of covariates $X_{1:p}$ among all observable covariates $(X^{(k)})_{k>0}$ to include in an analysis. 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 [4, 5, 12].

We also depart from the standard treatment of high-dimensional causal inference problems by treating the covariates generatively. 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 $0 < \alpha < 1$. 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).

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

Assumption 3 states that the overlap assumption is a restriction on the likelihood ratio between P_0 and P_1. It also hints at the fact that the probability measures P_0 and P_1 cannot be too distinct as p grows large; if they were, then the likelihood ratio would diverge, as it does in large-sample hypothesis testing settings.

2.2. *Strict Overlap.* While investigators typically invoke the overlap assumption in Assumption 2 for identification, they instead invoke a stronger overlap assumption to characterize the behavior of ATE estimators. For the remainder of the paper, we will focus on this stronger variant of the overlap assumption that we call the *strict overlap assumption*.

ASSUMPTION 4 (Strict Overlap, Propensity Score Form). For some constant $\eta \in (0, 0.5)$,

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

We call η the *bound* of the strict overlap assumption.

Strict overlap is a necessary condition for regular semiparametric estimators of the ATE to be $n^{1/2}$-consistent [20]. Specifically, all regular semiparametric estimators have asymptotic variance proportional to the following variance lower bound, known as the semiparametric efficiency bound [14, 10]:

$$V^{\text{eff}} = \mathrm{E}\left[\frac{\mathrm{var}(Y(1) \mid X_{1:p})}{e(X_{1:p})} + \frac{\mathrm{var}(Y(0) \mid X_{1:p})}{1 - e(X_{1:p})} + (\tau(X_{1:p}) - \tau^{ATE})^2\right]. \tag{3}$$

Since the propensity score appears in the denominator, these these fractions are only bounded if strict overlap holds. If it does not, all regular semiparametric estimators will converge to the ATE at a rate slower than $O(n^{-1/2})$.

The strict overlap assumption can also be represented in likelihood ratio form.

ASSUMPTION 5 (Strict Overlap, Likelihood Ratio Form). For some $\eta \in (0, 0.5)$, for each $A \in \sigma(X_{1:p})$ such that $P(A) > 0$,

$$\frac{\eta}{1-\eta} < \frac{\alpha}{1-\alpha}\frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} < \frac{1-\eta}{\eta} \quad \text{w.p. 1.} \tag{4}$$

Assumption 5 makes clear that the strict overlap assumption is a *bounded likelihood ratio* assumption. This representation establishes a relationship between the strict overlap assumption and the literature on testing and estimation in contexts where the likelihood ratio is bounded [16, 30, 31]. We make use of many of these results to state implications of strict overlap.

REMARK 1. From the definition (4) we can conclude that $\eta \leq \min\{\alpha, 1-\alpha\}$. Otherwise, if $\eta > \alpha$ and $\alpha \leq \frac{1}{2}$, then $\frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} > 1 + \frac{\eta-\alpha}{\alpha(1-\eta)}$ almost surely. But this would imply that $\int dP_1 = \int \frac{dP_1}{dP_0} dP_0 > 1$, which contradicts the fact that dP_1 is a probability density. An implication of this fact is that the strict

overlap assumption is harder to satisfy when the treatment group have very different sizes.

REMARK 2 (Gaussian Incompatibility with Strict Overlap). While we focus on the implications of strict overlap in high dimension, 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-likelihood ratio diverges for values of X with large magnitude. Similar results can be derived when $X_{1:p}$ is multi-dimensional. 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 Gaussianly distributed. This implication can be weakened if the covariates $X_{1:p}$ are restricted to lie on some compact subset of $\mathbb{R}^p$, but still have density proportional to a Gaussian distribution under both P_0 and P_1. This case is subsumed by the case of multivariate sub-exponential covariates, which we consider in Section 4.2.

3. Strict Overlap Bounds Discriminating Information. We now present our primary results that strict overlap bounds certain measures of discrepancy between the covariate measures $P_0(X_{1:p})$ and $P_1(X_{1:p})$. We show that strict overlap imposes an upper bound on discrepancies measured by the family of f-divergences. We then give specific bounds for two f-divergences, the KL divergence and the χ^2-divergence. Finally, we show that strict overlap imposes a lower bound on the error of any hypothesis test that attempts to distinguish between $P_0(X_{1:p})$ and $P_1(X_{1:p})$.

3.1. *f-Divergence Bounds.* 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 f-divergences, also known as ϕ-divergences [31]. f-divergences are a family of discrepancy measures between probability distributions defined in terms of a non-negative convex function f [11, 2, 21]. Formally, the f-divergence from some probability measure Q_0 to another Q_1 is defined as

$$D_f(Q_1(X_{1:p}) \| Q_0(X_{1:p})) := \mathrm{E}_{Q_0}\left[f\left(\frac{dQ_1(X_{1:p})}{dQ_0(X_{1:p})} \right) \right], \tag{5}$$

where dQ_0 and dQ_1 are densities defined with respect to any common dominating measure, say, $Q = \frac{1}{2}Q_0 + \frac{1}{2}Q_1$. f-divergences are non-negative and achieve a minimum when $Q_0 = Q_1$, and are, in general, asymmetric in their arguments.

Two well-known f-divergences are the KL divergence and the χ^2-divergence (also known as the Pearson divergence), defined as follows.

$$KL(Q_1(X_{1:p})\|Q_0(X_{1:p})) := \mathrm{E}_{Q_1}\left[\log \frac{dQ_1(X_{1:p})}{dQ_0(X_{1:p})}\right] \tag{6}$$

$$\chi^2(Q_1(X_{1:p})\|Q_0(X_{1:p})) := \mathrm{E}_{Q_0}\left[\left(\frac{dQ_1(X_{1:p})}{dQ_0(X_{1:p})} - 1\right)^2\right]. \tag{7}$$

These divergences correspond to the convex functions $f_{KL}(x) = x \log x$ and $f_{\chi^2} = (x-1)^2$.

Theorem 1 shows that strict overlap bounds any f-divergence between $P_0(X_{1:p})$ and $P_1(X_{1:p})$. This bound does not depend on the dimension p. In the sections that follow, we will show that this implies the bound becomes more restrictive the larger the dimension p.

For this theorem and the following discussion, it is useful to define notation for the lower and upper bound on the likelihood ratio implied by Assumption 5:

$$b_{\min} := \frac{1-\alpha}{\alpha}\frac{\eta}{1-\eta} < \frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} < \frac{1-\alpha}{\alpha}\frac{1-\eta}{\eta} =: b_{\max}. \tag{8}$$

Given the constraint on the likelihood ratio in Assumption 5 and the positivity and convexity of f, it should be clear that under strict overlap, any f-divergence is upper-bounded by $\max\{f(b_{\max}), f(b_{\min})\}$. Rukhin [31] derived tighter bounds, which we present in Theorem 1.

THEOREM 1 (f-Divergence Bound; Adaptation of Rukhin [31] Theorem 2.1). The strict overlap assumption with bound η implies that for any f-divergence D_f,

$$D_f(P_1(X_{1:p})\|P_0(X_{1:p})) < \frac{b_{\max}-1}{b_{\max}-b_{\min}} f(b_{\min}) + \frac{1-b_{\min}}{b_{\max}-b_{\min}} f(b_{\max}) \tag{9}$$

PROOF. By the definition in Assumption 5, the likelihood ratio $\frac{dP_0(X_{1:p})}{dP_1(X_{1:p})}$ is bounded by $b_{\min}$ and $b_{\max}$. Theorem 2.1 of Rukhin [31] gives the bound in (9). □

Rukhin [31] showed that the bound in (9) is tight. A similar result can be obtained for f-divergences in the opposite direction by switching the roles of P_0 and P_1. This entails swapping $(b_{\min}^{-1}, b_{\max}^{-1})$ for $(b_{\max}, b_{\min})$.

Using this result, we can state bounds on the KL divergence and the χ^2-divergence, each of which gives unique insight into the the strength of the strict overlap assumption as the dimensionality of $X_{1:p}$ increases.

3.2. *KL Divergence Bound.* Here, we use Theorem 1 to show that strict overlap imposes an upper bound on the KL divergence between P_0 and P_1. We use the chain rule that decomposes the overall KL divergence into a summation across dimensions to formally state that the strict overlap assumption becomes more restrictive as the dimension p of the covariates $X_{1:p}$ increases. In particular, we show that when p is large, the strict overlap condition implies that, on average, each covariate $X^{(k)}$ contains very little unique information discriminating between P_0 and P_1.

Proposition 1 shows that strict overlap bounds the KL divergence between $P_0(X_{1:p})$ and $P_1(X_{1:p})$ in both directions.

PROPOSITION 1 (KL Divergence Bound). The strict overlap assumption with bound η implies the following two inequalities

(10)
$$KL(P_1(X_{1:p})\|P_0(X_{1:p})) < \frac{(1-b_{\min})b_{\max}\log b_{\max} + (b_{\max}-1)b_{\min}\log b_{\min}}{b_{\max}-b_{\min}},$$

(11)
$$KL(P_0(X_{1:p})\|P_1(X_{1:p})) < -\frac{(1-b_{\min})\log b_{\max} + (b_{\max}-1)\log b_{\min}}{b_{\max}-b_{\min}}.$$

We will denote the upper bound in (10) as $B_{KL(1\|0)}$ and the upper bound in (11) as $B_{KL(0\|1)}$. In this case, we can obtain a loose upper bound on $B_{KL(1\|0)}$ and $B_{KL(0\|1)}$ that is more interpretable. This bound follows almost immediately from the statement of Assumption 5.

(12)
$$\max\left\{B_{KL(1\|0)}, B_{KL(0\|1)}\right\} \leq \left|\log\frac{\eta}{1-\eta}\right| + \left|\log\frac{\alpha}{1-\alpha}\right|$$

The first term of this loose 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, $\alpha = 0.5$. In this case, $B_{KL(1\|0)}$ and $B_{KL(0\|1)}$ also take a simple form that highlights the slack in the loose bound in (12):

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

Importantly, neither bound in Proposition 1 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 [9, Theorem 2.5.3]:

(13)
$$KL(P_1(X_{1:p})\|P_0(X_{1:p})) = \sum_{k=1}^{p} \mathrm{E}_{P_1} KL(P_1(X^{(k)} \mid X_{1:k-1})\|P_0(X^{(k)} \mid X_{1:k-1})).$$

Thus, the bound in Proposition 1 becomes more restrictive as p increases.

We can also use (13) 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 (13) is the expected KL divergence between the conditional distributions of the covariate $X^{(k)}$ under P_0 and P_1. Thus, Proposition 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

(14)
$$\frac{1}{p}\sum_{k=1}^{p} \mathrm{E}_{P_1} KL(P_1(X^{(k)} \mid X_{1:k-1})\|P_0(X^{(k)} \mid X_{1:k-1})) = O(p^{-1}),$$

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

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

3.3. *χ^2-Divergence Bound.* Here, we use Theorem 1 to show that strict overlap imposes an upper bound on the χ^2-divergence between $P_0(X_{1:p})$ and $P_1(X_{1:p})$. We can use this result to bound arbitrary discrepancies of the form:

(15)
$$|\mathrm{E}_{P_0}[g(X_{1:p})] - \mathrm{E}_{P_1}[g(X_{1:p})]|$$

for any measurable function $f : \mathbb{R}^p \mapsto \mathbb{R}$ that is square-integrable under P_0 and P_1. We apply this result in the following subsection to bound the mean discrepancy between $P_0(X_{1:p})$ and $P_1(X_{1:p})$.

Recall the definition of the χ^2-divergence in (7), and the definitions of $b_{\min}$ and $b_{\max}$ in (8). Applying Theorem 1, strict overlap imposes the following upper bounds on the χ^2-divergences between $P_0(X_{1:p})$ and $P_1(X_{1:p})$.

PROPOSITION 2. The strict overlap assumption with bound η implies the following two inequalities

$$\chi^2(P_1(X_{1:p})\|P_0(X_{1:p})) < (1 - b_{\min})(b_{\max} - 1) \tag{16}$$

$$\chi^2(P_0(X_{1:p})\|P_1(X_{1:p})) < (1 - b_{\max}^{-1})(b_{\min}^{-1} - 1). \tag{17}$$

We denote the bound in (16) as $B_{\chi^2(1\|0)}$ and the bound in (17) as $B_{\chi^2(0\|1)}$.

The χ^2-divergence is an important discrepancy measure because it appears in upper bounds on functional discrepancies of the form (15) derived by the Cauchy-Schwarz inequality. In particular, the expression in (15) has the following upper bound

(18)
$$|\mathrm{E}_{P_1} g(X_{1:p}) - \mathrm{E}_{P_0} g(X_{1:p})| = \left|\mathrm{E}_{P_0}\left[(g(X_{1:p}) - C) \cdot \left(\frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} - 1\right)\right]\right|$$

$$\leq \|g(X_{1:p}) - C\|_{P_0,2} \cdot \sqrt{\chi^2(P_1(X_{1:p})\|P_0(X_{1:p}))}, \tag{19}$$

for any finite constant C, and where $\|g\|_{P,q} := \mathrm{E}_P[|g|^q]^{1/q}$ denotes the q-norm of the function g under measure P. A similar bound holds with respect to the χ^2-divergence evaluated in the opposite direction.

Combining this observation with Proposition 2, we obtain the following explicit upper bound on functional discrepancies between P_0 and P_1.

THEOREM 2. The strict overlap assumption with bound η implies

$$|\mathrm{E}_{P_1} g(X_{1:P}) - \mathrm{E}_{P_0} g(X_{1:p})| \leq \min\left\{\sqrt{\mathrm{var}_{P_0}(g(X_{1:p}))} \cdot \sqrt{B_{\chi^2(1\|0)}}, \sqrt{\mathrm{var}_{P_1}(g(X_{1:p}))} \cdot \sqrt{B_{\chi^2(0\|1)}}\right\}. \tag{20}$$

and likewise for P_0 and P_1 switched.

PROOF. Let $C = \mathrm{E}_{P_0} g(X_{1:p})$ and apply (19) and Proposition 2. Do the same for $C = \mathrm{E}_{P_1} g(X_{1:p})$. □

REMARK 3 (Generalization to χ^α-Divergences). The decomposition in (18) can be used to construct additional upper bounds on the mean discrepancy in g using Hölder's inequality in combination with χ^α-divergences [33]. These bounds give a tighter bound in terms of η, but are functions of higher-order moments of $g(X_{1:p})$. We give full details of these generalized results in the appendix.

3.4. *Test Error Lower Bound.* To close this section, we present an information bound stated in a slightly different way: we show that strict overlap implies a lower bound on the error probability of any hypothesis test of the form

$$H_0 : X_{1:p} \sim P_0; \quad H_A : X_{1:p} \sim P_1. \tag{21}$$

This corresponds to an intuitive interpretation of strict overlap: when strict overlap holds, one should not be able to guess with a low error rate whether any unit was assigned to treatment or control on the basis of its covariate vector $X_{1:p}$. This result is closely related to the information divergence results stated above although its statement is superficially different. We discuss this connection after presenting the result.

Our approach in this section is similar to approaches that appear in the high-dimensional testing literature, for example, Addario-Berry et al. [1]. 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.

$$\begin{aligned} \delta_\phi &:= \max\{\text{size}, 1 - \text{power}\} \\ &:= \max\{P(\phi(S(X_{1:p})) = 1 \mid T = 0), P(\phi(S(X_{1:p})) = 0 \mid T = 1)\}. \end{aligned} \tag{22}$$

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 $\delta_\phi \leq \eta$.

PROOF. Let $\mathcal{E} := \{\phi(S(X_{1:p})) \neq T\}$, or the event that the test ϕ makes an error. By construction,

$$P(\mathcal{E}) = (1 - \alpha) \cdot \text{size} + \alpha \cdot (1 - \text{power}) \leq \delta_\phi.$$

By Fano's inequality [9, Corollary of Theorem 2.10.1]:

$$h(P(\mathcal{E})) \geq H(T \mid X_{1:p}), \tag{23}$$

where $h(x) := -x \log x - (1-x)\log(1-x)$ and $H(T \mid X_{1:p})$ is the conditional entropy of T given $X_{1:p}$, defined as

$$\begin{aligned} H(T \mid X_{1:p}) &= \mathrm{E}[-e(X_{1:p}) \log(e(X_{1:p})) - (1 - e(X_{1:p})) \log(1 - e(X_{1:p}))] \\ &= h(e(X_{1:p})). \end{aligned}$$

Note that h is concave, with a maximum at $x = 0.5$, and is symmetric about $x = 0.5$. Thus, by Assumption 4,

$$h(\eta) \leq H(T \mid X_{1:p}) \leq h(P(\mathcal{E})) \Longrightarrow \eta \leq P(\mathcal{E}) \leq \delta_\phi.$$

□

Theorem 3 is closely related to the f-divergence bounds stated earlier in this section. In particular, the probability of error $P(\mathcal{E})$ can itself be expressed as an f-divergence, and bounded using Theorem 1 [31, 30]. The bound in Theorem 3 derived from Fano's inequality is tighter; in particular, it is not a function of $\alpha := P(T = 1)$.

For additional intuition, we can also re-state Theorem 3 from the perspective of classification. For any number of features p, strict overlap implies there exists no classifier that can classify units into $T = 1$ or $T = 0$ with misclassification probability less than η. This framing immediately suggests empirical tests of whether the test bound in Theorem 3 is satisfied; we discuss this in more detail in Section 6.1.

Finally, Theorem 3 is a useful proof device for deriving concrete implications of overlap. In particular, we use this to convert any test of P_0 against P_1, regardless of whether that test is optimal, into an upper bound on the discrepancy between parameters of P_0 and P_1 identified by the test statistic $S(X_{1:p})$. We can also test this error bound empirically, a topic we leave to future work, but which we describe broadly in Section 6.1.

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 $X_{1:p}$. In particular, let $\phi(S(X_{1:p}))$ be *p-consistent* if and only if $\delta_\phi \to_P 0$ as p grows large. In the large-p limit, strict overlap holds only if there exists no p-consistent test.

COROLLARY 2 (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 p-consistent test of P_0 against P_1.

4. Applications to Mean Discrepancy Bounds. We now apply the results from the previous section to show that strict overlap bounds the mean discrepancy between covariate distributions $P_0(X_{1:p})$ and $P_1(X_{1:p})$. In other words, strict overlap restricts covariate imbalance, the primary observable property that separates observational studies from randomized trials.

We give two bounds on mean imbalance on this section. The first is a bound that applies generally, whenever the covariance of $X_{1:p}$ has a finite operator norm under P_0 or P_1. We derive this bound by applying Theorem 2. The second is a tighter bound that applies in the case where $X_{1:p}$ is multivariate sub-exponential under both P_0 and P_1. We derive this bound by applying Theorem 3.

Throughout this section, we use the following notation for the expectations and covariance matrices of $X_{1:p}$ under P_0 and P_1:

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

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

4.1. *General Mean Discrepancy Bound.*

4.1.1. *Bound.* Here, we apply Theorem 2 to derive an upper bound on the mean discrepancy between $P_0(X_{1:p})$ and $P_1(X_{1:p})$. Under weak distributional assumptions, this upper bound becomes more restrictive as the dimension of $X_{1:p}$ grows.

COROLLARY 3. Strict overlap with bound η implies

$$\|\mu_{0,1:p} - \mu_{1,1:p}\| \leq \min\Big\{\|\Sigma_{0,1:p}\|_{op}^{1/2} \cdot \sqrt{B_{\chi^2(1\|0)}}, \; \|\Sigma_{1,1:p}\|_{op}^{1/2} \cdot \sqrt{B_{\chi^2(0\|1)}}\Big\}. \tag{24}$$

PROOF. Let $g(X_{1:p}) := a'(X_{1:p} - \mu_0)$, where $a := \frac{\mu_{1,1:p} - \mu_{0,1:p}}{\|\mu_{1,1:p} - \mu_{0,1:p}\|}$, and apply Theorem 2. $\mathrm{var}_{P_0}(a'(X_{1:p} - \mu_0))$ is upper-bounded by $\|\Sigma_{0,1:p}\|_{op}$ by definition, and the result follows. □

To more easily represent the relationship between dimension and the upper bound on mean discrepancy in Corollary 3, we convert the Euclidean distance bound to an upper bound on mean absolute discrepancy between the covariate means.

COROLLARY 4. Under the conditions of Corollary 3,

$$\frac{1}{p}\sum_{i=1}^{p}\left|\mu_0^{(k)}-\mu_1^{(k)}\right| \leq p^{-1/2}\cdot\min\left\{\|\Sigma_{0,1:p}\|_{op}^{1/2}\cdot\sqrt{B_{\chi^2(1\|0)}},\ \|\Sigma_{1,1:p}\|_{op}^{1/2}\cdot\sqrt{B_{\chi^2(0\|1)}}\right\}. \tag{25}$$

We can characterize the conditions under which Corollary 4 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 such that for each p $X_{1:p} \subset (X^{(k)})_{k>0}$. Then the bound in Corollary 4 behaves as

$$\frac{1}{p}\sum_{i=1}^{p}\left|\mu_0^{(k)}-\mu_1^{(k)}\right| \leq O\left(p^{-1/2}\cdot\min\left\{\|\Sigma_{0,1:p}\|_{op}^{1/2},\|,\Sigma_{1,1:p}\|_{op}^{1/2}\right\}\right) \tag{26}$$

and converges to zero when $\min\left\{\|\Sigma_{1,0:p}\|_{op}^{1/2},\|,\Sigma_{1,1:p}\|_{op}^{1/2}\right\} = o(p^{1/2})$. Thus, if the minimum operator norm does not grow too fast, the constraint in Corollary 4 becomes more binding as p grows large .

REMARK 4 (Generalization to Other Functions f). Theorem 2 may be useful for bounding a number of other distributional discrepancies. For example, discrepancies of the form (15) are used to define families of Maximum Mean Discrepancies, which include discrepancies such as the Wasserstein distance as special cases [13].

4.1.2. *Operator Norm.* Because it plays a central role in the bound, we briefly discuss the growth rate of the operator norm. Mathematically, the operator norm of a covariance matrix $\|\Sigma_{1:p}\|_{op}$ is the largest eigenvalue of the covariance matrix $\Sigma_{1:p}$, so this is a condition on how quickly these eigenvalues grow with p. Geometrically, the square root of the operator norm of a covariance matrix $\|\Sigma_{1:p}\|_{op}^{1/2}$ of some random vector $X_{1:p}$ is the maximum standard deviation of any one-dimensional linear projection of $X_{1:p}$. Thus,

the operator norm grows more slowly than p if the number of non-degenerate orthogonal projections of $X_{1:p}$ grows with p.

We now give several examples of covariance structures and the behavior of their corresponding operator norm. 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 2 (Independent Case). When the components of $(X^{(k)})_{k>0}$ are independent, with component-wise variance given by σ_k^2, $\|\Sigma_{1:p}\|_{op} = \max_k \sigma_k^2$. 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 $f < M$, $\|\Sigma_{1:p}\|_{op} \leq M$ [6]. 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 (Restricted Rank Case). If $(X^{(k)})_{k>0}$ has component-wise variances given by σ_k^2 and a number $s_p \leq p$ of factors, so that $\Sigma_{1,1:p}$ has rank s_p, then $\|\Sigma_{1:p}\|_{op} \geq s_p^{-1} \sum_k \sigma_k^2$, 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 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 $O(p/s_p)$.

Thus, if the covariates $X_{1:p}$ are not too correlated, so that $\|\Sigma_{1:p}\|_{op} = o(p)$, strict overlap implies that the mean absolute discrepancy in (26) converges to zero, and the covariate means approach balance, on average, as p grows large.

4.2. *Sub-Exponential Mean Discrepancy Bound.* Here, we combine Theorem 3 with distributional assumptions about $X_{1:p}$ to derive tighter bounds on mean discrepancy than those obtained in Corollary 3.

In particular, we consider the case where $X_{1:p}$ is *multivariate sub-exponential* under both P_0 and P_1 [7, 18]. 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 $a \in \mathbb{R}^p$,

$$\mathrm{E}_{P_0} \exp(a'[X_{1:p} - \mu_{0,1:p}]) \leq \exp\left(\frac{\|a\|^2 \sigma_p^2}{2}\right) \quad \text{for} \quad \|a\| \leq \frac{1}{b_p}, \tag{27}$$

and likewise for P_1 and $\mu_{1,1:p}$. A consequence of (27) is that all one-dimensional linear projections of multivariate sub-exponential random variables have tails that can be upper bounded by Gaussian or exponential random variables. These bounds can be used to construct tight, non-asymptotic concentration inequalities that are useful for constructing statistical tests.

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. Multivariate sub-Gaussian random variables, which have been invoked in previous discussions of covariate distributions in the high-dimensional setting [3], are a special case of multivariate sub-exponential random variables.

We apply Theorem 3 to a test based on tail bounds to obtain the following result.

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

$$\|\mu_{0,1:p} - \mu_{1,1:p}\| \leq \begin{cases} \sqrt{8\sigma_p^2 \log \frac{1}{\eta}} & \text{if } \sigma_p^2/b_p^2 > -2\log\eta \\ 4b_p \log \frac{1}{\eta} & \text{if } \sigma_p^2/b_p^2 \leq -2\log\eta. \end{cases} \tag{28}$$

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 $S(X_{1:p}) = a'(X_{1:p} - \mu_0)$, 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, $S(X_{1:p})$ is sub-exponential with parameters (σ_p^2, b_p). Define a test that rejects P_0 whenever $S(X_{1:p}) > \Delta_p/2$, i.e., when the projection of $X_{1:p} - \mu_0$ is closer to μ_1 than it is to μ_0, and accepts P_0 otherwise.

The probability of rejecting under P_0 is

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

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

Denote by $\Delta^*_{p,\eta}$ the value of Δ_p that sets $\delta = \eta$ in each case. Solving for $\Delta^*_{p,\eta}$ gives the result. Further, $\Delta^*_{p,\eta} < \frac{2\sigma_p^2}{b_p p^{1/2}}$ if and only if $\sigma_p^2/b_p^2 > -2\log\eta$. □

The bounds obtained in Theorem 4 have tighter scaling in η than those in Corollary 3; the bounds in (28) scale in $\log^{1/2}\frac{1}{\eta}$ and $\log\frac{1}{\eta}$, whereas the bound in (24) scales in $\eta^{-1/2}$.

REMARK 5. One could obtain tighter bounds on the mean discrepancy in this case by specifying separate sub-exponential parameters (σ_p^2, b_p) for each of P_0 and P_1. We have chosen to present the simpler, looser bound for clearer exposition.

As with Corollary 3, we can translate Theorem 4 into a bound on the absolute difference in means.

COROLLARY 5. In the same setting as Theorem 4, strict overlap with bound η implies that

$$\frac{1}{p}\sum_{k=1}^{p}\left|\mu_0^{(k)} - \mu_1^{(k)}\right| \leq \begin{cases} \sigma_p p^{-1/2}\sqrt{8\log\frac{1}{\eta}} & \text{if } \sigma_p^2/b_p^2 > -2\log\eta \\ b_p p^{-1/2} 4\log\frac{1}{\eta} & \text{if } \sigma_p^2/b_p^2 \leq -2\log\eta. \end{cases} \tag{29}$$

We can analyze Corollary 5 asymptotically in the same way that we did with Corollary 4. Let $(X^{(k)})_{k>0}$ 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

$$\frac{1}{p}\sum_{k=1}^{p}\left|\mu_0^{(k)} - \mu_1^{(k)}\right| \leq O(\max\{\sigma_p, b_p\}p^{-1/2}) \tag{30}$$

and converges to zero when $\max\{\sigma_p, b_p\} = o(p^{1/2})$. Once again, the case where the upper bound in (30) approaches zero is of particular interest because it implies that, for large p, the covariate means are, on average, arbitrarily close to balance.

The asymptotic behavior of the bound on Corollary 5 is determined by the scaling of the sub-exponential parameters (σ_p^2, b_p). σ_p and b_p are the standard deviations of the Gaussian and exponential random variables used to upper bound the tails of all one-dimensional projections of $X_{1:p}$; in this way, they are analogous to the operator norm discussed in Section 4.1.2, but

they are associated with the bounding distribution rather than the covariate distribution itself. As in the cases considered in Examples 2–4, 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.

5. Overlap and Low-Dimensional Structure Assumptions. So far, we have shown that strict overlap is a more restrictive assumption in high-dimensional settings than it is in low-dimensional settings. These results imply that strict overlap cannot be taken for granted in high dimensions, and that additional justification may be necessary to argue that overlap holds in a high-dimensional study.

We now explore one set of assumptions that make overlap conditions easier to satisfy in a high-dimensional setting, namely, that the treatment assignment mechanism or outcome process in a given study has a fundamentally low-dimensional relationship with the covariate sequence $(X^{(k)})_{k>0}$. In particular, we formalize the intuition that, under this restriction, identification only requires overlap with respect to this low-dimensional structure. However, we also highlight a key tension: while overlap in this low-dimensional structure has weaker implications, unconfoundedness in this setting is a stronger assumption.

We consider two special cases that have been proposed in the literature. First, we consider the case where the propensity score $e(X_{1:p})$ can be written as a function of a low-dimensional variable, regardless of the dimension of $X_{1:p}$. In this case, strict overlap in $X_{1:p}$ has weaker implications for covariate balance in high dimensions. Second, we consider the case where the conditional expectation of the potential outcomes $\mathrm{E}[(Y(0), Y(1)) \mid X_{1:p}]$ can be characterized in terms of a similar low-dimensional variable. In this case, the ATE can be identified under a weaker overlap assumption.

These results highlight the role of modeling restrictions for identification. Typically, investigators assume such low-dimensional structure in order to establish *estimability* of the propensity score $e(X_{1:p})$ or the outcome model $\mathrm{E}[Y^{obs} \mid T, X_{1:p}]$. While true, we argue that these assumptions play a more fundamental role in identification. Thus, even if data-splitting, ensemble learning, or machine learning methods are employed to make $e(X_{1:p})$ or $\mathrm{E}[Y^{obs} \mid T, X_{1:p}]$ estimable in a wider range of settings, assumptions of low-dimensional structure may still be necessary to argue that the strict overlap assumption holds in a high-dimensional setting.

5.1. *Low-Dimensional Treatment Assignment.* We now explore the case where the treatment assignment mechanism is fundamentally low-dimensional.

For this discussion, it is useful to define a single treatment assignment mechanism associated with the entire covariate sequence $(X^{(k)})_{k>0}$. This treatment assignment mechanism remains fixed regardless of the covariate set $X_{1:p} \subset (X^{(k)})_{k>0}$ included in the analysis. A natural characterization of this treatment assignment mechanism is the *limiting propensity score*.

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

$$e(X_{1:\infty}) := P(T = 1 \mid \sigma(X_{1:\infty})). \tag{31}$$

The limiting propensity score is the best prediction of treatment assignment that can be constructed from the covariate sequence $(X^{(k)})_{k>0}$. Importantly, 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 $X_{1:p}$ (see the proof for Proposition 3). As we show next, when Assumption 6 holds, strict overlap has weaker implications.

ASSUMPTION 6 (Sufficient Condition for Strict Overlap).

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

$$X_{1:p} \perp\!\!\!\perp T \mid B \quad \forall X_{1:p}. \tag{32}$$

2. Strict overlap holds with respect to B:

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

PROPOSITION 3 (Sufficient Condition Statement). Assumption 6 implies, for all p, strict overlap in $X_{1:p}$ with bound η.

PROOF. B is finer than $e(X_{1:p})$ for any p. Strict overlap in a variable implies strict overlap in all functions of that variable, so strict overlap in B implies strict overlap in $e(X_{1:p})$, which implies overlap in $X_{1:p}$. □

We call a variable B that satisfies (32) a *balancing variable*. Assumption 6 is in fact a restriction on the limiting propensity score, because the limiting

propensity score is coarser than any balancing variable B, in that there exists some function h such that

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

Assumption 6 has some trivial specifications, which are useful examples. At one extreme, we may specify that $B = e(X_{1:\infty})$. In this case, Assumption 6 is vacuous: this puts no restrictions on the form of the limiting propensity score and strict overlap with respect to B is equivalent to strict overlap holding in the large-p limit. 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 5 below. Such restrictions trade off stronger modeling assumptions on the limiting propensity score $e(X_{1:\infty})$ with weaker implications of strict overlap. Importantly, these specifications 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.

EXAMPLE 5 (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 $X_{1:s}$,

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

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 [4] and Farrell [12] 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 6 (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 or latent factor model. In this case,

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

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 [3] 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.

In the appendix, we give a detailed example of a case where the low-dimensional structure described in Example 6 can make the mean discrepancy bounds presented in Section 4 are not binding.

5.2. *Low-Dimensional Outcome Process.*

5.2.1. *Identification.* We now turn to the case where the conditional expectation of the potential outcomes $\mathrm{E}[(Y(0), Y(1)) \mid X_{1:p}]$ has low-dimensional structure for all p. In this setting, the ATE is identifiable under a weaker overlap condition than Assumption 2. We make this precise in a slight modification of the setup in Hansen [15], who established that, for a given covariate set $X_{1:p}$, the ATE is identified under the following assumption.

ASSUMPTION 7 (Prognostic Identification).

1. Unconfoundedness holds with respect to $X_{1:p}$ (Assumption 1).
2. There exists some function $r(X_{1:p})$ such that

$$(Y(0), Y(1)) \perp\!\!\!\perp X_{1:p} \mid r(X_{1:p}). \tag{33}$$

3. Overlap is satisfied with respect to $r(X_{1:p})$.

We modify the nomenclature of Hansen [15] slightly and call $r(X_{1:p})$ the *prognostic score*.[1] Petersen et al. [24] and Luo, Zhu and Ghosh [22] proposed similar identification conditions. When assumption 7 holds, the ATE can be identified by the following functional

$$\tau^{ATE} = \mathrm{E}[\mathrm{E}[Y^{obs} \mid T = 1, r(X_{1:p})] - \mathrm{E}[Y^{obs} \mid T = 0, r(X_{1:p})]]. \tag{34}$$

If strict overlap holds with respect to $r(X_{1:p})$, a $n^{1/2}$-consistent *non-regular* semiparametric estimator of (34) exists. We discuss some examples of these non-regular semiparametric estimators at the end of this section.

We now define the *limiting prognostic score*, which fixes a single outcome process for any covariate set $X_{1:p} \subset (X^{(k)})_{k>0}$.

[1] Hansen used the "prognostic score" to refer to the expectation of $Y(0)$ alone given covariates; here, we use it to refer to the expectation of both $Y(0)$ and $Y(1)$. In Hansen's terminology, this would be the union of the prognostic score and the "effect modifier".

DEFINITION 2 (Limiting prognostic score). In the same context as Definition 1, the limiting prognostic score is defined as

$$r(X_{1:\infty}) := \mathrm{E}[(Y(0), Y(1)) \mid T, \sigma(X_{1:\infty})]. \tag{35}$$

The limiting prognostic score is the best prediction of the potential outcomes that can be constructed from the covariate sequence $(X^{(k)})_{k>0}$. The limiting prognostic score $r(X_{1:\infty})$ is finer than the prognostic score $r(X_{1:p})$ for any finite dimension p. By placing restrictions on the limiting prognostic score, we can define settings in which there is overlap in the prognostic score $r(X_{1:p})$ even for large p.

ASSUMPTION 8 (Sufficient Condition for ATE Identification).

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

$$(Y(0), Y(1)) \perp\!\!\!\perp X_{1:p} \mid R \quad \forall X_{1:p}. \tag{36}$$

2. Strict overlap holds with respect to R:

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

PROPOSITION 4 (Sufficient Condition Statement). Assumption 8 implies that there is overlap in the prognostic score $r(X_{1:p})$ for any p.

PROOF. R is finer than $r(X_{1:p})$ for all p. □

We call a variable R that satisfies (36) a *prognostic variable*. Assumption 6 is a restriction on the limiting prognostic score, in that there exists some function h such that

$$r(X_{1:\infty}) = h(R).$$

The assumption of strict overlap in the the full covariates $X_{1:p}$ is at least as strong as the assumption of strict overlap in the prognostic score $r(X_{1:p})$, and is stronger whenever $r(X_{1:p})$ is not finer than the propensity score $e(X_{1:p})$. Thus, imposing restrictions on the outcome process weakens the necessary strict overlap condition for the existence of a $n^{1/2}$-consistent estimator of the ATE.

5.2.2. *Estimation.* Given a prognostic score, one can specify a non-regular semiparametric estimator for the ATE with respect to the prognostic score $r(X_{1:p})$, as opposed to the full covariates $X_{1:p}$. Such an estimator is subject to a semiparametric efficiency bound parameterized in terms of $e_r(X_{1:p}) = P(T = 1 \mid r(X_{1:p}))$ instead of $e(X_{1:p})$ [22]:

(37)
$$V_r^{\text{eff}} = \mathrm{E}\left[\frac{\mathrm{var}(Y(1) \mid r(X_{1:p}))}{e_r(X_{1:p})} + \frac{\mathrm{var}(Y(0) \mid r(X_{1:p}))}{1 - e_r(X_{1:p})} + (\tau(r(X_{1:p})) - \tau^{ATE})^2\right].$$

In cases where strict overlap holds with respect to $r(X_{1:p})$, but not with respect to $X_{1:p}$, no $n^{1/2}$-consistent *regular* semiparametric estimator exists, and this non-regular estimator is *super-efficient*, meaning it has lower asymptotic variance than the semiparametric efficiency bound for regular estimators [34, 22].

The caveat is that such a non-regular estimator only enjoys uniform $n^{1/2}$-consistency guarantees within the non-parametric family of data-generating processes where

$$\mathrm{E}[(Y(0), Y(1)) \mid X_{1:p}] = \mathrm{E}[(Y(0), Y(1)) \mid r(X_{1:p})].$$

In particular, the non-regular estimator can have unbounded bias if the reduction $r(X_{1:p})$ is misspecified and is not in fact a prognostic score.

Several authors have proposed non-regular estimators based on outcome model restrictions. Luo, Zhu and Ghosh [22] 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. van der Laan and Gruber [34] proposed the Collaborative TMLE (C-TMLE) method, which uses some of the machinery of semiparametric estimation, and can generate valid estimates of the ATE even if there is no overlap in $X_{1:p}$ but there is overlap in the prognostic score. Hill and Su [17] suggested using BART (implicitly assuming the outcome process to be estimable by BART) to identify areas of "common causal support", or overlap in sufficient statistics of the outcome model. 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 [25, 19]. Roy et al. [27] 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 covariate reductions that may further weaken overlap conditions in Section 6.2.

5.3. *Implications for Unconfoundedness.* So far, we have shown that low-dimensional structure in the limiting propensity score or limiting prognostic score can make overlap conditions that are necessary for causal identification easier to satisfy. At the same time, these process restrictions complicate the unconfoundedness assumption. In particular, when the relationship between the covariate sequence $(X^{(k)})_{k>0}$ and the treatment assignment mechanism or outcome process is low-dimensional, unconfoundedness is a stronger assumption. We make this statement precise in the following proposition.

PROPOSITION 5 (Low-Dimensional Confounding).

1. Suppose that Assumption 6 holds with respect to some B and unconfoundedness (Assumption 1) holds with respect to some covariate set $X_{1:p} \subset (X^{(k)})_{k>0}$. Then unconfoundendess holds with respect to a variable $h_p(B)$ alone, which is coarser than B.
2. Suppose that Assumption 8 holds with respect to some R and unconfoundedness (Assumption 1) holds with respect to some covariate set $X_{1:p} \subset (X^{(k)})_{k>0}$. Then unconfoundendess holds with respect to a variable $h_p(R)$ alone, which is coarser than R.

PROOF. For the first statement, by the argument in the proof of Proposition 3, for any $X_{1:p}$, there exists a function h_p such that the propensity score $e(X_{1:p}) = h_p(B)$. If unconfoundedness holds for any $X_{1:p}$, then it also holds for $e(X_{1:p})$ [26].

For the second statement, we can make the same argument with respect to the prognostic score $r(X_{1:p})$. □

Thus, there is a tradeoff between the strength of overlap and unconfoundedness assumptions. Qualitatively, the assumptions in this section that weaken the overlap requirements for causal identification do so by restricting how informative the covariate sequence $(X^{(k)})_{k>0}$ is about the treatment assignment and potential outcomes in the study. Under this limitation, it becomes harder to argue that, despite this information restriction, the covariates are informative enough to explain away all confounding in the study. Likewise, in settings where high-dimensional adjustment is necessary to eliminate confounding, there may not exist a low-dimensional structure on the treatment assignment mechanism that makes the overlap assumption plausible.

This tension highlights the fact that adjusting for high-dimensional covariates does not necessarily make causal identification more plausible. In high-dimensional studies, careful justification regarding the structures of the

treatment assignment or outcome process are necessary to argue that overlap and unconfoundedness are simultaneously plausible.

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

6.1. *Future Work for Testing Overlap.* One of the primary implications of our results is that the strict overlap assumption has significant empirical content. Unlike the unconfoundedness assumption, the strict overlap assumption can be tested. In light of this, we suggest that higher importance should be put on verifying that overlap holds in high dimensions. Each of the bounds derived in Sections 3 and 4 have testable implications in finite samples, the clearest being the mean discrepancy bounds in Corollary 3 and Theorem 4. The test error bound in Theorem 3 can also be tested directly: the out-of-sample classification error of any classification procedure trained to classify units into "treated" and "control" is an estimate of an upper bound on the test error.

Assumptions about low-dimensional structure in the treatment assignment mechanism (Assumption 6) also have testable implications. In particular, when B is a balancing variable the following equality holds for each set A such that $P(A) > 0$:

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

In cases where B is estimable, this hypothesis could be tested using nonparametric two-sample testing methods such as kernel MMD [13], modified to use balancing weights.

6.2. *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 assumptions about the structure of confounding in a given observational study. Such methods could construct a generalization of propensity scores and prognostic scores that satisfy

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid d(X_{1:p}) \tag{38}$$

whenever unconfoundedness holds with respect to $X_{1:p}$. If strict overlap holds with respect to $d(X_{1:p})$, the ATE could be identified by the functional

$$\tau^{ATE} = \mathrm{E}[\mathrm{E}[Y^{obs} \mid T = 1, d(X_{1:p})] - \mathrm{E}[Y^{obs} \mid T = 0, d(X_{1:p})]]. \tag{39}$$

Despite reducing the covariates $X_{1:p}$, overlap in a deconfounding score $d(X_{1:p})$ is not always a weaker assumption than overlap in $X_{1:p}$. In particular, no balancing score $b(X_{1:p})$ satisfying

$$T \perp\!\!\!\perp X_{1:p} \mid b(X_{1:p}), \tag{40}$$

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 [26].

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 [28].

In this vein, 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 (38), 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 (38) is a ternary relationship between three sets of random variables, instead of a binary relationship between two random variables as in (40).

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: GENERALIZED FUNCTIONAL DISCREPANCY BOUNDS USING χ^α DIVERGENCES

This section gives details of the claim made in Remark 3, that one can obtain additional bounds on the functional discrepancy

$$|\mathrm{E}_{P_1} g(X_{1:p}) - \mathrm{E}_{P_0} g(X_{1:p})|$$

by bounding the χ^α divergences between P_0 and P_1.

Formally, χ^α-divergences are a class of divergences that generalize the χ^2-divergence [33]:

$$\chi^\alpha(P_0(X_{1:p}) \| P_1(X_{1:p})) := \mathrm{E}_{P_0}\left[\left|\frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} - 1\right|^\alpha\right] \quad \text{for } \alpha \geq 1. \tag{41}$$

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 [31] implies that, under strict overlap with bound η,

$$\chi^\alpha(P_0(X_{1:p}) \| P_1(X_{1:p})) \leq (b_{\max} - 1)(1 - b_{\min}) \frac{(1 - b_{\min})^{\alpha-1} + (b_{\max} - 1)^{\alpha-1}}{b_{\max} - b_{\min}}$$

$$\chi^\alpha(P_1(X_{1:p}) \| P_0(X_{1:p})) \leq (b_{\min}^{-1} - 1)(1 - b_{\max}^{-1}) \frac{(1 - b_{\max}^{-1})^{\alpha-1} + (b_{\min}^{-1} - 1)^{\alpha-1}}{b_{\min}^{-1} - b_{\max}^{-1}}.$$

We denote these bounds as $B_{\chi^\alpha(0\|1)}$ and $B_{\chi^\alpha(1\|0)}$, respectively.

Applying Hölder's inequality to (18), we obtain

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

where $q_\alpha := \frac{\alpha}{\alpha - 1}$ 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. This bound scales as $\eta^{-1/\alpha}$, whereas (20) scales as $\eta^{-1/2}$.

APPENDIX B: WEAKENED MEAN-DISCREPANCY IMPLICATIONS

In the following example, the latent variable B induces strong correlation in the covariates $X_{1:p}$ so that the mean discrepancy bound from Corollary 3 is not binding as p increases.

EXAMPLE 7 (Mixture exception to mean difference bounds). In the setting of Example 6, suppose 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 $\mu_{1,1:p}$.

REFERENCES

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

[2] Ali, S. M. and Silvey, S. D. (1966). A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)* **28** 131–142.

[3] 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.

[4] 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.

[5] 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.

[6] Bickel, P. J. and 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** 989–1010.

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

[8] 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.

[9] 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.

[10] 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.

[11] Csiszár, I. (1963). Eine informationstheoretische {U}ngleichung und ihre anwendung auf den {B}eweis der ergodizität von {M}arkoffschen {K}etten. *Publ. Math. Inst. Hungar. Acad.* **8** 95–108.

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

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

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

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

[16] Hellman, M. E. and Cover, T. M. (1970). Learning with Finite Memory. *The Annals of Mathematical Statistics* **41** 765–782.

[17] 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.

[18] 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.

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

[20] Khan, S. and Tamer, E. (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica* **78** 2021–2042.
[21] Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory* **52** 4394–4412.
[22] Luo, W., Zhu, Y. and Ghosh, D. (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika* **104** 51–65.
[23] Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge University Press, New York, NY, USA.
[24] Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* **21** 31–54.
[25] Rosenbaum, P. R. (2002). *Observational Studies. Springer Series in Statistics.* Springer New York, New York, NY.
[26] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **1083** 41–55.
[27] Roy, S., Rudin, C., Volfovsky, A. and Wang, T. (2017). FLAME: A Fast Large-scale Almost Matching Exactly Approach to Causal Inference. 1–23.
[28] Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2** 808–840.
[29] 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.
[30] Rukhin, A. L. (1993). Lower Bound on the Error Probability for Families with Bounded Likelihood Ratios. *Proceedings of the American Mathematical Society* **119** 1307.
[31] Rukhin, A. L. (1997). Information-type divergence when the likelihood ratios are bounded. *Applicationes Mathematicae* **24** 415–423.
[32] 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**.
[33] Vajda, I. (1973). $\chi\hat{}\{\alpha\}$-divergence and generalized Fisher's information. *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes* 873–886.
[34] van der Laan, M. J. and Gruber, S. (2010). Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics* **6**.
[35] van der Laan, M. J. and Rose, S. (2011). *Targeted Learning. Springer Series in Statistics.* Springer New York, New York, NY.