When Can We Achieve Small Error in Observational Causal Inference?

David Bruns-Smith [1]

Abstract

We explore the conditions necessary to guarantee sharp upper bounds on the mean squared error when estimating mean counterfactual outcomes from observational data. In particular, we analyze the large family of designed-based weighting estimators which include balancing weights and matching. Beginning from the bias-variance decomposition, we argue that assumptions have to be made about the outcome function in order to choose a high performance estimator. For a theoretical framework, we use integral probability metrics and ϕ-divergences to analyze the bias-variance trade-off. Finally, we consider conditions under which our mean squared error bounds are robust to failure of our assumptions.

1. Introduction

Randomization is the linchpin of robust causal inference. Even observational studies are often validated by reference to a smaller randomized experiment. But sometimes, either due to prohibitive cost or ethical concerns, randomization is not possible. For example, it would be unethical for researchers to randomly assign participants to smoke cigarettes. Likewise, the Federal Reserve will not randomly vary interest rates to obtain robust estimates for the impact of monetary policy. This paper asks: what must we assume to obtain estimates of average causal effects with small mean squared error from observational data-sets?

We limit ourselves to the so-called "design-based" approach (Rubin et al., 2008); i.e. we observe the characteristics of units selected into treated and control groups, but we have to choose an estimator without examining the outcomes. The idea is to treat the data as having come from an experiment where the assignment rules have been lost and must be reconstructed, which may prevent mistaking spurious correlation for causality. The building blocks considered here could potentially benefit from judicious use of outcome data, but we leave that for future work.

By far the most common approach to estimating an average treatment effect is to reweight the control units to look like the treated units and vice-versa. This family of estimators includes as special cases: inverse probability weighting (IPW), balancing weights, and matching. We study the assumptions required for these estimators to have a small mean squared error (MSE).

The first is a standard assumption, called "ignorability" or "no unobserved confounding." Under this assumption, the relationship between covariates and outcomes, which I will call the *outcome function*, is the same regardless of treatment status. However, without any further assumptions on the outcome function, it is only possible to achieve bounded bias if we can perfectly reweight the treatment groups to look like one another. Unfortunately, this will typically entail very large weights. The resulting high variance means a heavy penalty for the MSE even if we assume that the covariate distributions have common support.

This motivates an additional assumption which has been explored in the balancing weights literature (Hainmueller, 2012; Zubizarreta, 2015; Wang & Zubizarreta, 2020; Kallus, 2020). The second assumption asserts that the outcome function belongs to some function class. At one extreme, we might assume that the outcome function is linear. To be more modest, we might assume the outcome function belongs to a very general class, like bounded, or Lipschitz functions. Restricting the outcome function makes it possible to have finite bias even when the treatment groups are not perfectly reweighted or when the covariate distributions do not have common support.

Work at the intersection of machine learning and causal inference (Kallus, 2020) has illustrated how to turn an assumption on the outcome function into a bound on the bias using integral probability metrics (IPMs). In this paper, we build on this technical machinery to make three new contributions for robustly controlling the MSE. First, we leverage a relationship between IPMs and information-theoretic divergences to characterize the bias-variance trade-off of

[1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. Correspondence to: David Bruns-Smith <bruns-smith@berkeley.edu>.

Workshop on the Neglected Assumptions in Causal Inference (NACI) at the 38^{th} International Conference on Machine Learning, 2021

reweighting estimators. Second, we use tools from robust statistics to bound the bias even when the assumption on the outcome function is incorrect. Finally, we explore how these techniques can bound the MSE when the ignorability assumption is violated.

2. Problem Setup

Let $X \in \mathcal{X}$ denote the covariates and $T \in \{0, 1\}$ denote treatment status. Under SUTVA, we assume that there exist random variables $Y_0, Y_1 \in \mathbb{R}$, called potential outcomes, but that we only observe the factual outcomes, $Y = (1 - T)Y_0 + TY_1$. For simplicity, assume we are interested in the potential outcome Y_0. We observe Y_0 for the control population, but we do not observe Y_0 for the treated population and our goal is to estimate the missing mean: $\mathbb{E}[Y_0|T = 1]$. A key assumption for weighting estimators is "ignorability", which requires the relationship between covariates and outcomes to be the same across treatment groups:

Assumption 1. *For any $x \in \mathcal{X}$:*

$$\begin{aligned} P(Y_0|X = x) &= P(Y_0|X = x, T = 1) \\ &= P(Y_0|X = x, T = 0). \end{aligned}$$

Given Assumption 1, we can estimate the missing mean by $\mathbb{E}[w(X)Y_0|T = 0]$ for some weights $w(X)$ that are a function of the observed covariates.

2.1. Mean Squared Error

Denote the conditional mean of Y_0 as $m_0(x)$ and the conditional variance of Y_0 as $\sigma_0^2(x)$. To simplify notation, let p be the joint distribution of Y_0 and X for the control units and let q be the joint distribution of Y_0 and X for the treated units. Then we can expand the MSE with the standard bias-variance decomposition:

$$\begin{aligned} \text{MSE}(w) &= \mathbb{E}_p[(w(X)Y_0 - \mathbb{E}_q[Y_0])^2] \\ &= (\mathbb{E}_p[w(X)Y_0] - \mathbb{E}_q[Y_0])^2 + \text{Var}_p[w(X)Y_0] \\ &= (\mathbb{E}_p[w(X)m_0(X)] - \mathbb{E}_q[m_0(X)])^2 \quad (1) \\ &\quad + \mathbb{E}_p[w(X)^2\sigma_0^2(X)]. \quad (2) \end{aligned}$$

The MSE depends on two terms: first, the imbalance of the mean of the outcome function m_0 between the reweighted control distribution and the treated distribution, and second the size of the weights. The exact variance depends on the magnitude and distribution of $\sigma_0^2(X)$. For simplicity, in this paper we will assume that $\sigma_0^2(x) = 1, \forall x$.

In practice, we do not know m_0. If we cannot make any assumptions at all about m_0, then the only $w(x)$ that guarantees finite bias is the density ratio, $q(x)/p(x)$. In this case, the bias term (1) is zero for any m_0. Estimating the density ratio and then using the estimate for weights results is called inverse probability weighting (IPW). In practice the variance of these weights is usually very large. Furthermore, the density ratio can be challenging to estimate and under minor misspecification the results can be erratic as for the IPW specification in (Kang et al., 2007). If p and q do not have common support then the density ratio does not exist, and there are no weights that can guarantee finite bias.

3. Assumptions on the Outcome Function

Intuitively, we would like to regularize our estimator, i.e. choose smaller weights to decrease the variance in exchange for an increase in bias. As mentioned above, if m_0 is completely unrestricted, then this isn't possible; there always exists an m_0 that can make our bias arbitrarily large if the covariate distributions are not perfectly balanced. But in practice, we often expect the relationship between covariates and outcomes to have some additional structure. For example, if we assume that m_0 is linear, then we are guaranteed that the bias is no greater than δ as long as

$$\|\mathbb{E}_p[w(X)X] - \mathbb{E}_q[X]\| \leq \delta.$$

This is the motivation for the stable balancing weights estimator in (Zubizarreta, 2015). For a given δ we can achieve bounded bias (up to scaling of the outcomes) by balancing the covariate means while achieving small MSE by making the variance of the weights as small as possible.

Most of the time, the outcome function will not be linear. Fortunately, a similar strategy works for leveraging *any* prior knowledge about m_0. Stated generally, we make the following assumption:

Assumption 2. *For some class of functions $\mathcal{F}$, $m_0 \in \mathcal{F}$.*

Many choices of $\mathcal{F}$ in Assumption 2 are quite general and easy to justify with domain knowledge. Some examples, ignoring the scaling of the outcomes:

$$\begin{aligned} \text{Bounded functions: } \mathcal{F}_\infty &:= \{f : \|f\|_\infty \leq 1\} \\ \text{Lipschitz functions: } \mathcal{F}_{\text{Lip(c)}} &:= \{f : \|f\|_{\text{Lip(c)}} \leq 1\} \\ \text{RKHS functions: } \mathcal{F}_\mathcal{H} &:= \{f : \|f\|_\mathcal{H} \leq 1\} \end{aligned}$$

where $\|\cdot\|_{\text{Lip(c)}}$ denotes the Lipschitz constant with respect to a metric c and $\|\cdot\|_\mathcal{H}$ denotes the norm in some Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$. As above, we can find minimum variance weights that balance all functions in $\mathcal{F}$ to level $\delta > 0$. This gives rise to the following optimization problem:

$$\min_{w \geq 0} \ \mathbb{E}_p[w(X)^2] \quad (3)$$

$$\text{such that } \mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)] \leq \delta, \forall f \in \mathcal{F}. \quad (4)$$

If the reduction in variance compared to the density ratio is greater than the bias δ, then the weights from this optimization problem can result in a substantially smaller mean

squared error. In fact, even if the covariate distributions do not have common support, there may still exist finite weights that can balance the outcome function class to a finite (but necessarily non-zero) δ.

3.1. Integral Probability Metrics

Existing work (Kallus, 2020) has leveraged distances between probability distributions called integral probability metrics (IPMs) to analyze the balance constraints in the optimization problem above. To ease exposition, assume that $w(X) \geq 0$ and $\mathbb{E}_p[w(X)] = 1$. In this case, $r(x) \coloneqq w(x)p(x)$ is itself a density. Assume $m_0 \in \mathcal{F}$. The bias term (1) is upper-bounded by the worst case difference in means over the class $\mathcal{F}$ between the distributions r and q. An IPM measures exactly this kind of worst-case discrepancy, defined for two distributions μ and ν, and a function class $\mathcal{F}$ as:

$$\text{IPM}_{\mathcal{F}}(\mu, \nu) \coloneqq \sup_{f \in \mathcal{F}} \{\mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(X)]\}. \quad (5)$$

Therefore, under Assumption 2, finding weights with bias less than δ is equivalent to balancing the means of all functions in $\mathcal{F}$ to level δ, which is equivalent to finding a distribution r such that $\text{IPM}_{\mathcal{F}}(r, q) \leq \delta$.

IPMs are a theoretically and computationally useful concept for finding weights that satisfy the constraints (4). At first glance, it might sound impossible to find weights that balance large classes like all bounded functions or all Lipschitz functions. It turns out that the IPMs corresponding to many natural function classes are equivalent to well-known distances between probability distributions. For example, the IPM for bounded functions is equivalent to the total variation (TV) distance, the IPM for Lipschitz functions is equivalent to the Wasserstein distance, and the IPM corresponding to an RKHS is equivalent to the Maximum Mean Discrepancy (MMD). Therefore, if we want to solve the balancing weights problem over Lipschitz functions, we need to find minimum variance weights such that the induced distribution r is within Wasserstein distance δ of the target distribution q. There are many computational strategies for tackling similar problems in the distributionally robust optimization literature - for example, see (Esfahani & Kuhn, 2018; Yu et al., 2021).

Note that another parallel line of work (Shalit et al., 2017; Johansson et al., 2020), uses assumptions on m_0 and the corresponding IPMs to learn better representations for estimating the conditional expectation and then computing the counterfactual mean as a domain adaptation problem. This is a different approach, but since many of the theoretical concepts are similar, maybe there will be opportunities to combine these ideas in the future.

In this paper, we introduce two new ways of theoretically analyzing Assumption 2 using IPMs. First, we'll explore when Assumption 2 can substantially decrease the MSE by combining IPMs and information-theroetic divergences to characterize the bias-variance tradeoff. Second, we will use existing results on IPMs and moment conditions from robust statistics to maintain a finite upper bound on the bias even when Assumption 2 is violated.

4. The Bias-Variance Tradeoff

While IPMs are a useful starting place for bounding the bias, the MSE also depends on the variance. Part of the motivation for introducing an assumption on m_0 was that it becomes possible to find weights with non-zero but finite bias; but that's only useful if in exchange we achieve a large decrease in the variance. The key questions for theoretical analysis are then: (1) for a fixed $\mathcal{F}$, as we increase δ in (4), at what rate does the variance of the weights decrease? and (2) for two difference function classes $\mathcal{F}$ and $\mathcal{G}$, which achieves smaller variance for the same level of δ? As a motivating example, (Kallus, 2020) finds empirically that assuming m_0 lies in a norm ball of an RKHS can result in substantively smaller weights than a Lipchitz assumption, presumably because Lipchitz functions are a more complicated class and therefore more challenging to balance.

4.1. ϕ-Divergences

As our first novel contribution, we propose a theoretical framework for formally analyzing these tradeoffs. We begin by recognizing that the variance term (2) takes the form of an information-theroetic divergence. To see this, recall that we defined $r(x) \coloneqq w(x)p(x)$, and so the variance term can be rewritten as $\mathbb{E}_p[(r(X)/p(X))^2]$, the expected value of the squared density ratio between r and p. This is a special case of a class of divergences called ϕ-divergences, defined for two distributions, μ and ν, and a convex function ϕ as:

$$D_\phi(\mu||\nu) \coloneqq \mathbb{E}_\nu \left[\phi \left(\frac{\mu(X)}{\nu(X)} \right) \right]. \quad (6)$$

Special cases of ϕ-divergences include the Kullback-Leibler divergence, the χ^2-divergence, and the total variation (TV) distance. Our variance term is a ϕ-divergence for $\phi(z) = z^2$. This happens to be the Renyi divergence with $\alpha = 2$, and so we will denote it $D_2(r||p)$. See (Sriperumbudur et al., 2009; Agrawal & Horel, 2020) for further examples of ϕ-divergences and a discussion of their connection with IPMs.

We've now arrived at a straightforward interpretation for weighting estimators: we want to find a distribution r which is close to q (in the sense of an IPM) but not too far from p (in the sense of a ϕ-divergence):

Proposition 1. *If $m_0 \in \mathcal{F}$ and given weights $w > 0, \mathbb{E}_p[w] = 1$, which define a distribution r with density*

$r(x) = w(x)p(x)$, then the MSE of the corresponding weighting estimator is bounded by:

$$MSE(w) \leq IPM_{\mathcal{F}}(q, r) + D_2(r||p).$$

On one extreme, if $w = q/p$, then $r = q$. The bias is zero and our mean-squared error is entirely determined by $D_2(q||p)$. On the other extreme, if $w = 1$, then $r = p$, and the error is entirely determined by $\text{IPM}_{\mathcal{F}}(q, p)$. The optimal w is likely between these two extremes.

As alluded to above, we do not need an overlap condition to bound the error. The only distributions r that correspond to finite weights are absolutely continuous with respect to p. If overlap is violated, i.e. q is not absolutely continuous with respect to p, then it is not possible to choose $r = q$ and $D_2(q||p) = \infty$. But for our purposes, all this means is that there is non-zero minimum bias. Depending on the divergence of the weights, it might still be worth accepting even more bias to further reduce the variance. In this case, the overlap violation has no impact at all on the sharpest achievable MSE bound. See Figure 1 for an illustration. Note, that if we do not make Assumption 2, then this is no true, because if m_0 can be arbitrary, the density ratio is the only w that can guarantee finite bias.

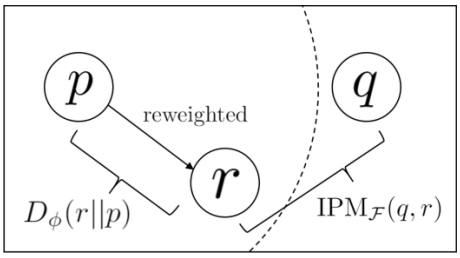

Figure 1. An illustration of the MSE with an overlap violation. The distributions - represented as circles - to the left of the dotted curve are absolutely continuous with respect to p. p can still be reweighted to a new distribution r which trades off the "distance" (as measured by a ϕ-divergence) between p and r and the "distance" (as measured by an IPM) between r and q.

4.2. Variational Representations of ϕ-Divergences

So why does recognizing that the variance term is a ϕ-divergence make it possible to characterize the bias-variance tradeoff? The key is that ϕ-diveregences admit variational representations which have special theoretical properties under Assumption 2. For exposition, we will change the objective slightly. Consider using the KL divergence, which corresponds to $\phi(z) = z \log z$,

$$D_{\text{KL}}(\mu||\nu) \coloneqq \mathbb{E}_\nu \left[\frac{\mu(X)}{\nu(X)} \log \frac{\mu(X)}{\nu(X)} \right] = \mathbb{E}_\mu \left[\log \frac{\mu(X)}{\nu(X)} \right].$$

In fact, this measure of dispersion of the weights is commonly used in the causal inference literature, either explicitly (Hainmueller, 2012) or implicitly (Tan, 2020). For the moment, pretend that the variance of the weights is $D_{\text{KL}}(r||p)$ instead of $D_2(r||p)$.

The KL divergence can be written in the following variational form using results from convex duality (Nguyen et al., 2010; Birrell et al., 2020b):

$$D_{\text{KL}}(\mu||\nu) = \sup_{f>0} \{\mathbb{E}_\mu[\log f(X)] - \mathbb{E}_\nu[f(X)] + 1\}. \quad (7)$$

Note that this form is very nearly the worst case difference in means between μ and ν over all measurable functions $f > 0$ as in the definition of an IPM. In fact, there's a natural relationship between the two definitions. Consider the variational formulation above with the supremum restricted to a function class $\mathcal{F}$. From (Dupuis & Mao, 2019; Birrell et al., 2020a) we have the following result:

$$D^{\mathcal{F}}_{\text{KL}}(\mu||\nu) \coloneqq \sup_{f \in \mathcal{F}} \{\mathbb{E}_\mu[\log f(X)] - \mathbb{E}_\nu[f(X)] + 1\} \quad (8)$$

$$= \inf_{r \in \mathcal{P}(\mathcal{X})} \{\text{IPM}_{\mathcal{F}}(\mu, r) + D_{\text{KL}}(r||\nu)\} \quad (9)$$

where $\mathcal{P}(\mathcal{X})$ is the set of probability distributions over $\mathcal{X}$. There are some requirements on $\mathcal{F}$ for the second equality to hold (see above references), but all three function classes mentioned above satisfy these requirements. This is a very powerful result and applications to GANs have been explored recently in (Song & Ermon, 2020), although it appears that their formulation was developed independently of (Dupuis & Mao, 2019; Birrell et al., 2020a).

One thing stands out right away: the infimum in (9) is nearly exactly the tightest possible MSE bound provided by Proposition 1. By maximizing the variational objective between p and q over a function class that we believe contains m_0, we automatically trade-off the bias and dispersion of the weights. This result also provides some theoretical insights into how relaxing the function balance constraint δ can result in smaller dispersion.

Consider the original variational form in (7). When the supremum is finite, the function f^* that maximizes this objective is the density ratio, $f^*(x) = \mu(x)/\nu(x)$. This corresponds exactly to the case we mentioned previously where $w(x) = q(x)/p(x)$, and therefore $r = q$, there is zero bias, and the MSE is determined by the divergence between q and p. This is yet another way to see that if we cannot make any assumptions at all about the function m_0, then we should completely reweight the control group to look like the treated group and this is the best we can do.

Now assume that m_0 is Lipschitz. We know (again, ignoring the fact that the ϕ-divergences are not the same) that the best achievable MSE is upper-bounded by $D^{\mathcal{F}_{\text{Lip}}}_{\text{KL}}(q||p)$. Because the variational objective is convex, the function $f^*(x)$ that achieves the supremum over $\mathcal{F}_{\text{Lip}}$ will be the

Lipschitz function that is closest to $q(x)/p(x)$. In this case, the best achievable mean squared error is upper-bounded by $\mathbb{E}_\mu[\log f^*(X)] - \mathbb{E}_\nu[f^*(X)] + 1$. In fact, if the density ratio is itself Lipschitz, then the optimum occurs at $r = p$, i.e. it is not worth accepting increased bias to decrease the dispersion of the weights. Generally, the MSE is bounded by the projection of the density ratio onto the function class used in Assumption 2, a promising avenue for data-dependent theoretical results.

The results above largely go through for $D_2(r||p)$, the original variance term. In fact, for any ϕ-divergence, we have a corresponding variational representation (Birrell et al., 2020a):

$$D_\phi(\mu||\nu) = \sup_f \{\mathbb{E}_\nu[f(X)] - \mathbb{E}_\mu[\phi^*(f(X))]\} \qquad (10)$$

where ϕ^* denotes the conjugate dual of ϕ and the supremum is over all measurable functions. Likewise when we restrict f to a function class $\mathcal{F}$, given some regularity conditions on ϕ and $\mathcal{F}$, we can write $D_\phi^{\mathcal{F}}$ as an infimal convolution as in (9). The analysis for D_2 is still in progress. In particular, the unrestricted optimizer f^* for D_2 is closely related to, but not exactly the density ratio.

5. Robustness

Above, we argued that Assumption 2 is crucial for achieving a small MSE because we can only tradeoff bias and variance if we are willing to assume that the outcome function, m_0, lies within some function class. If we choose a more restrictive function class, then we can potentially use much smaller weights because we have to balance fewer functions to satisfy the constraints (4).

What if we balance the bias and variance based on a particular function class $\mathcal{F}$, but it turns out that $m_0 \in \mathcal{G} \neq \mathcal{F}$? For example, if we assume the outcomes are linear in X, but the true outcome function is highly nonlinear, then we might worry that the true bias could be much larger than we assumed. We would like to retain a guarantee that the bias over functions in $\mathcal{G}$ is still finite even though Assumption 2 is violated.

Our second novel contribution restates this issue in terms of IPMs, which then allows us to apply recent results from robust statistics. Specifically, let's say we assumed that $m_0 \in \mathcal{F}$ and we found a reweighted distribution r such that our worst-case discrepancy over $\mathcal{F}$ is no greater than δ. In other words, we found r such that $\text{IPM}_\mathcal{F}(q, r) \leq \delta$. If it turns out that in actuality, $m_0 \in \mathcal{G} \neq \mathcal{F}$, then the true bias is bounded by $\text{IPM}_\mathcal{G}(q, r)$. Therefore, we can still guarantee control of the MSE if $\text{IPM}_\mathcal{F}(q, r) \leq \delta \implies \text{IPM}_\mathcal{G}(q, r) \leq \rho$ for some $\rho < \infty$.

It turns out that almost exactly the same question regularly arises in the robust statistics literature. For example, (Zhu et al., 2019) begins by studying the problem of mean estimation subject to an TV perturbation. A key step includes finding conditions under which the so-called "modulus of continuity" is bounded. This amounts to finding some minimal moment condition under which, for two distributions μ and ν:

$$TV(\mu, \nu) \leq \delta \implies \|\mathbb{E}_\mu[X] - \mathbb{E}_\nu[X]\| \leq \rho \qquad (11)$$

Notice, that this would be exactly the robustness condition we need in the case where we assumed that m_0 was bounded, but it turned out that m_0 was linear. One important insight is that if we do not make any additional assumptions on μ and ν, then we can never guarantee (11) with finite ρ. This is because ν could be equal to μ but with an ϵ fraction of probability mass moved arbitrarily far away, resulting in arbitrarily different means.

However, with very minimal moment conditions on μ and ν, we can guarantee (11) with known dependence of ρ on δ. For example, by applying Theorem 3.2 of (Zhu et al., 2019), we immediately get the following two results. Let w be weights inducing a distribution r that solve the optimization problem (3) for $\mathcal{F}_\infty$ and $0 < \delta < 0.499$. Let $\mathcal{X} = \mathbb{R}^d$ and assume $m_0 = \beta^T X$ with $\|\beta\| = 1$, where we ignore scaling of the outputs for convenience.

Proposition 2. *If r and q have bounded covariance, then we have the following upper bound on the bias:*

$$\|\mathbb{E}_p[w(X)m_0(X)] - \mathbb{E}_q[m_0(X)]\| \leq \rho(\delta),$$

where $\rho(\delta) = O(\sqrt{\delta})$.

Proposition 3. *If r and q are sub-Gaussian, then we have the following upper bound on the bias:*

$$\|\mathbb{E}_p[w(X)m_0(X)] - \mathbb{E}_q[m_0(X)]\| \leq \rho(\delta),$$

where $\rho(\delta) = O(\delta\sqrt{\log(1/\delta)})$.

These examples are illustrative but unrealistic; it's unlikely that we would assume that m_0 was bounded but then be surprised to learn that m_0 was actually exactly linear. We would like to provide a bound for when $m_0 \in \mathcal{G}$ for many classes $\mathcal{G}$. Fortunately, many problems in robust statistics also take this form. For example, robust covariance estimation requires bounding the worst-case mean discrepancy over functions of the form $(v^T X)^2$ for all $\|v\| = 1$. Therefore, conditions exist to bound the worst-case for any (symmetric) function class $\mathcal{G}$. For example, if an Orlicz norm condition (like bounded kth moment or sub-Gaussianity) applies to $f(X)$ for every function in $\mathcal{F}_{\text{Lip(c)}}$ or $\mathcal{F}_\mathcal{H}$ then we can apply Lemma E.2 of (Zhu et al., 2019) to bound the bias with very similar dependence of ρ on δ.

Further results exist for when we originally balance $\mathcal{F}_{\text{Lip(c)}}$ instead of $\mathcal{F}_\infty$. It would be interesting to extend these results to RKHS's; for example, bounding the MMD between spaces corresponding to different kernels.

6. Unobserved Confounders

We've argued that in order to achieve small MSE with reweighting estimators, we require not just ignorability (Assumption 1) but also some structure on the outcome function (Assumption 2). So far, we've shown that with additional moment assumptions (like bounded covariance) we can still achieve small bias if we initially assumed that Assumption 2 holds for $\mathcal{F}_\infty$ but it turned out that this assumption was wrong. Next, we will consider what happens when ignorability is violated instead.

In real-world observational data, there probably exists an unobserved variable U which is correlated both with the selection into T and the outcomes Y_0. In this case, $P(Y_0|X, T = 0) \neq P(Y_0|X, T = 1) \neq P(Y_0|X)$. WLOG, consider the case where $\mathbb{E}[Y_0|X, U, T = 0] = \mathbb{E}[Y_0|X, U, T = 1] = m_0(X, U)$. Our estimand is the mean under the true joint distribution, $q(X, U) \coloneqq P(X, U|T = 1)$. Consider the weighting estimator where we do not observe U and only reweight by the marginal distribution of the X's: $w(X) = q(X)/p(X)$. Then,

$$\mathbb{E}_p[w(X)m_0(X, U)] = \int\int m_0(x, u)p(u|x)q(x)dxdu.$$

The reweighted distribution $\hat{q}(x, u) \coloneqq p(u|x)q(x)$ does not equal the target distribution $q(x, u) = q(u|x)q(x)$, even though we've attempted to fully reweight. In fact, $q(x, u)$ is unknown and therefore the bias $\text{IPM}_\mathcal{F}(q, \hat{q})$ is also unknown, even if we assume $m_0 \in \mathcal{F}$. But, the MSE for this particular choice of w still takes exactly the same form as before:

$$\text{MSE}(w) \leq \text{IPM}_\mathcal{F}(q, \hat{q}) + \mathbb{E}_p[w(X)^2] \tag{12}$$

with $\text{IPM}_\mathcal{F}(q, \hat{q})$ being the only unknown term. Our goal in choosing w remains the same, the only difference is that now we don't observe the true q, only $\hat{q}$.

Without further assumptions, the unknown bias term, $\text{IPM}_\mathcal{F}(q, \hat{q})$, might be unbounded. However, we can use the same robust statistics tools outlined above to perform sensitivity analysis. We can assume that $\text{IPM}_\mathcal{F}(q, \hat{q}) \leq \epsilon$ and then try to find moment conditions that will result in reasonable control of the MSE for all feasible q. Now that q is unobserved, this is more or less exactly the approach used in (Zhu et al., 2019) for adversarial TV and Wasserstein distance perturbations. Finally, we can optimize over all choices of w given the worst-case realization of q, resulting in a minimax bound on the MSE for a fixed level of ϵ.

7. Conclusion

To summarize: if we want to estimate the counterfactual mean outcome for the treated units, we can reweight the outcomes of the control units. If there are no unobserved confounding variables, then using the density ratio of the covariate distributions as the weights will result in an unbiased estimator, but with potentially large variance. Controlling the bias-variance trade-off for the MSE requires making some assumptions about the outcome function, $m_0(x)$, and if we assume that m_0 lies within some function class then we can potentially find weights with much smaller variance. We propose a theoretical framework for analyzing this trade-off using IPMs and variational representations of ϕ-divergences. We then suggest that techniques and moment conditions from robust statistics provide a blueprint for maintaining control of the error even when our assumptions fail.

References

Agrawal, R. and Horel, T. Optimal bounds between f-divergences and integral probability metrics. In *International Conference on Machine Learning*, pp. 115–124. PMLR, 2020.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Rey-Bellet, L. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics. *arXiv preprint arXiv:2011.05953*, 2020a.

Birrell, J., Katsoulakis, M. A., and Pantazis, Y. Optimizing variational representations of divergences and accelerating their statistical estimation. *arXiv preprint arXiv:2006.08781*, 2020b.

Dupuis, P. and Mao, Y. Formulation and properties of a divergence used to compare probability measures without absolute continuity. *arXiv preprint arXiv:1911.07422*, 2019.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, pp. 25–46, 2012.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Kallus, N. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62): 1–54, 2020.

Kang, J. D., Schafer, J. L., et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Rubin, D. B. et al. For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3):808–840, 2008.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Song, J. and Ermon, S. Bridging the gap between f-gans and wasserstein gans. In *International Conference on Machine Learning*, pp. 9078–9087. PMLR, 2020.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Tan, Z. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020.

Wang, Y. and Zubizarreta, J. R. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

Yu, Y., Lin, T., Mazumdar, E., and Jordan, M. I. Fast distributionally robust learning with variance reduced min-max optimization. *arXiv preprint arXiv:2104.13326*, 2021.

Zhu, B., Jiao, J., and Steinhardt, J. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

Zubizarreta, J. R. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.