Causal Inference with Selectively Deconfounded Data

Kyra Gan **Andrew A. Li** **Zachary C. Lipton** **Sridhar Tayur**
Carnegie Mellon University, Pittsburgh, PA 15213
{kyragan,aali1,zlipton,stayur}@cmu.edu

Abstract

Given only data generated by a standard confounding graph with unobserved confounder, the Average Treatment Effect (ATE) is not identifiable. To estimate the ATE, a practitioner must then either (a) collect deconfounded data; (b) run a clinical trial; or (c) elucidate further properties of the causal graph that might render the ATE identifiable. In this paper, we consider the benefit of incorporating a large *confounded* observational dataset (*confounder unobserved*) alongside a small *deconfounded* observational dataset (*confounder revealed*) when estimating the ATE. Our theoretical results suggest that the inclusion of confounded data can significantly reduce the quantity of deconfounded data required to estimate the ATE to within a desired accuracy level. Moreover, in some cases—say, genetics—we could imagine retrospectively selecting samples to deconfound. We demonstrate that by actively selecting these samples based upon the (already observed) treatment and outcome, we can reduce sample complexity further. Our theoretical and empirical results establish that the worst-case relative performance of our approach (vs. a natural benchmark) is bounded while our best-case gains are unbounded. Finally, we demonstrate the benefits of selective deconfounding using a large real-world dataset related to genetic mutation in cancer.

1 Introduction

The fundamental problem in causal inference is to estimate causal effects using *observational* data. This task is particularly motivated by scenarios when experiments are infeasible. While the literature typically addresses a rigid setting in which confounders are either *always* or *never* observed, in many applications we might observe confounders for a *subset* of samples. For example, in healthcare, a particular gene might be suspected to confound the relation between a behavior and a health outcome of interest. Due to the high cost of genetic tests, we might only be able to afford to reveal the value of the genetic confounder for a subset of patients. Note that for a variable such as a genetic mutation, we might observe retrospectively, even after the treatment and outcome have been observed. We call this process of revealing the value of an (initially unobserved) confounder *deconfounding*, and the samples where treatment, outcome, and confounders are all observed *deconfounded data*.

So motivated, this paper addresses the middle ground along the confounded-deconfounded spectrum. Naively, one could estimate the ATE with standard methods using only the deconfounded data. First, we ask: *how much can we improve our ATE estimates by incorporating confounded data over approaches that rely on deconfounded data alone?* Second, motivated by the setting in which our confounders are genetic traits that might be retrospectively observed for cases with known treatments and outcomes, we introduce the problem of *selective deconfounding*—allocating a fixed budget for revealing the confounder based upon observed treatments and outcomes. This prompts our second question: *what is the optimal policy for selecting data to deconfound?* To our knowledge, this is the first paper that focuses on the case where ample (cheaply-acquired) confounded data is available and we can select only few confounded samples to deconfound (expensive).

We address these questions for a standard confounding graph where the treatment and outcome are binary, and the confounder is categorical. First, we propose a simple method for incorporating confounded data that achieves a constant-factor improvement in ATE estimation error. In short, the inclusion of (infinite) confounded data reduces the number of free parameters

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

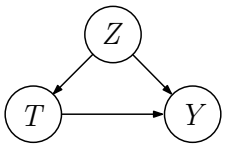

Figure 1: Causal graph with treatment T, outcome Y, and selectively observed confounder Z

to be estimated, improving our estimates of the remaining parameters. Moreover, due to the multiplicative factors in the causal functional, errors in parameter estimates can compound. Thus, our improvements in parameter estimates yield greater benefits in estimating treatment effects. For binary confounders, our numerical results show that on average, over problem instances selected uniformly on the parameter simplex, our method achieves roughly 2.5× improvements in ATE estimation error.

Next, we show that we can reduce error further by actively choosing which samples to deconfound. Our proposed policy for selecting samples dominates reasonable benchmarks. In the worst case, our method requires no more than 2× as many samples as a natural sampling policy and our best-case gains are unbounded. Moreover, our qualitative analysis characterizes those situations most favorable/unfavorable for our method. We extend our work to the scenario where only a finite amount of confounded samples is present, demonstrating our qualitative insights continue to apply (Appendix C). Additionally, we validate our methods using COSMIC (Tate et al., 2019; Cosmic, 2019), a real-world dataset containing cancer types, genetic mutations, and other patient features, showing the practical benefits of our proposed sampling policy. Throughout the paper, we implicitly assume that the *confounded* data was sampled i.i.d. from the target population of interest (but our policy for selecting data to deconfound need not be).

2 Related Work

Causal inference has been studied thoroughly under the ignorability assumption, i.e., no unobserved confounding (Neyman, 1923; Rubin, 1974; Holland, 1986). Some approaches for estimating the ATE under ignorability include inverse propensity score weighting (Rosenbaum and Rubin, 1983; Hirano et al., 2003; McCaffrey et al., 2004), matching (Dehejia and Wahba, 2002), the backdoor adjustment (Pearl, 1995), and targeted learning (Van der Laan and Rose, 2011). Some related papers look to combine various sources of information, for instance from randomized control trials and observational data to estimate the ATE (Stuart et al., 2011; Hartman et al., 2015). Other papers leverage machine learning techniques, such as random forests, for estimating causal effects (Alaa and van der Schaar, 2017; Wager and Athey, 2018).

Some papers investigate ATE estimation with confounded data by leveraging mediators (Pearl, 1995) and proxies (Miao et al., 2018). Others investigate combining confounded observational data with *experimental* data. Kuroki and Pearl (2014) identify graphical structures under which causal effect can be identified. Miao et al. (2018) propose to use two different types of proxies to recover causal effects with one unobserved confounder. Shi et al. (2020) extend the work by Miao et al. (2018) to multiple confounders. However, both methods require knowledge of proxy categories a priori and are not robust under misspecification of proxy categories. Louizos et al. (2017) use variational autoencoders to recover the causal effect under the model where when conditioned on the unobserved confounders, the proxies are independent of treatment and outcome. Pearl (1995) introduces the front-door adjustment, expressing the causal effect as a functional that concerns only the (possibly confounded) treatment and outcome, and an (unconfounded) mediator that transmits the entire effect.

In other work, Bareinboim and Pearl (2013) propose to combine observational and experimental data under distribution shift, learning the treatment effect from the experimental data and transporting it to the confounded observational data to obtain a bias-free estimator for the causal effect. Recently, Kallus et al. (2018) propose a two-step process to remove hidden confounding by incorporating experimental data. Lastly, few papers provide finite sample guarantees for causal inference. Shalit et al. (2017) upper bound the estimation error for a family of algorithms that estimate causal effects under the ignorability assumption.

Unlike most prior work, we (i) address confounded and deconfounded (but not experimental) observational data, (ii) perform finite sample analysis to quantify the relative benefit of additional confounded and deconfounded data towards improving our estimate of the average treatment effect, and (iii) investigate sample-efficient policies for selective deconfounding.

3 Methods and Theory

Let T and Y be random variables denoting the treatment and outcome. We restrict these to be binary, viewing T as an indicator of whether a particular treatment has occurred and Y as an indicator of whether the outcome was *successful*. In this work, we assume the existence of a single (possible) confounder, denoted Z, which can take up to k categorical values (Figure 1). In addition, although we only include one unobserved

confounder in our model, because our variables are categorical, (as shown in Section 3.1) this subsumes scenarios with multiple categorical confounders. Following Pearl's nomenclature (Pearl, 2000), let

$$P(Y = y|\text{do}(T = t)) := \sum_{z\in[k]} P_{Y|T,Z}(y|t, z)P_Z(z).$$

Our goal is to estimate the ATE, which can be expressed via the back-door adjustment in terms of the joint distribution $P_{Y,T,Z}$ on (Y, T, Z), as:

$$\begin{aligned}\text{ATE} &:= P(Y = 1|\text{do}(T = 1)) - P(Y = 1|\text{do}(T = 0)),\\ &= \sum_{z\in[k]} \left(P_{Y|T,Z}(1|1, z) - P_{Y|T,Z}(1|0, z)\right) P_Z(z). \quad (1)\end{aligned}$$

Our key contribution is to analyze and empirically validate methods for estimating the ATE from both *confounded* and *deconfounded* observations. In our setup, the *confounded data* contains n i.i.d. samples from the joint distribution $P_{Y,T}$ (marginalized over the hidden confounder Z), and the *deconfounded data* contains m i.i.d. samples from the full joint distribution $P_{Y,T,Z}$. Thus, the confounded and deconfounded data are (y, t) and (y, t, z) tuples, respectively. Recall that here *deconfounding* means selecting a confounded data point (y, t) and revealing the value of its confounder z. There are two ways that we can obtain m *deconfounded data*, one through collecting m deconfounded data directly without using the confounded data, and the other through revealing the value of the confounder for m confounded data points. Note that given this graph, we cannot exactly calculate the ATE unless we intervene or make further assumptions on the structure of the causal graph. Recall that such interventions or graph structures may not be available (e.g., in the case of genetic mutation). Furthermore, when deconfounded data is scarce and confounded data is comparatively plentiful, we hope to improve our ATE estimates.

3.1 Generalizability of Our Model

First, we note that the use of categorical (even binary) data is well-established in both theory (Bareinboim and Pearl, 2013) and application (Knudson, 2001; Rayner et al., 2016), and not merely a simplifying proxy for continuous data. Next, we show that our model subsumes scenarios with multiple categorical confounders. First, absent additional distributional assumptions, our model captures multiple unobserved confounders by simple concatenation (since we impose no limit on the number of classes) *without loss*. Now, one could make additional assumptions (indeed, a *high*-dimensional setting might necessitate such assumptions) that could render alternative algorithms applicable. However, there exist many applications where (a) the confounder is of moderate dimension; and (b) a practitioner would be dubious of any additional assumption (Bates et al., 2020). Second, although in this scenario we implicitly assume that the set of confounders is either never observed or entirely observed, this is also without loss so long as the costs of revealing each confounders are the same (e.g. the genetic example). Intuitively, because we do not impose any independence assumption on the set of confounder, revealing all confounders offers maximal information on the joint distribution of the confounders. We formalize this statement in Appendix A.1. While for simplicity we focus only on the setting in which our confounder can be retroactively observed, as we show in Appendix A.2 our model can be applied straightforwardly to handle a set of additional pretreatment covariates.

3.2 Infinite Confounded Data

In this subsection, we address the setting where we have an *infinite* amount of confounded data ($n = \infty$), i.e., the marginal distribution $P_{Y,T}$ is known exactly. We leave the analysis on *finite* confounded data to Appendix C.

Deconfounded Data Alone We begin with the baseline approach of using only the deconfounded data. Let $p^z_{yt} = P_{Y,T,Z}(y, t, z)$, and let $\hat{p}^z_{yt}$ be the empirical estimate of p^z_{yt} from the deconfounded data using the Maximum Likelihood Estimator (MLE). Let $\widehat{\text{ATE}}$ be the estimated average treatment effect calculated by plugging $\hat{p}^z_{yt}$'s into Equation (1). In the following theorem, we show a quantity of deconfounded samples m which is sufficient to estimate the ATE to within a desired level of accuracy under the estimation process described above. Let $C = 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$ throughout.

Theorem 1. *(Upper Bound) Using deconfounded data alone,* $P\left(\left|\widehat{\text{ATE}} - \text{ATE}\right| \geq \epsilon\right) < \delta$ *is satisfied if the deconfounded sample size* m *is at least*

$$m_{\text{base}} := \max_{t,z} C\left(\sum_y p^z_{yt}\right)^{-2} = \max_{t,z} \frac{1}{P_{T,Z}(t, z)^2} C.$$

The proof of Theorem 1 (Appendix B.2) relies on an additive decomposition of ATE estimation error in terms of the estimation errors on the p^z_{yt}'s, along with concentration via Hoeffding's inequality. We will contrast Theorem 1 with counterpart methods that use confounded data.

Incorporating Confounded Data Estimating the ATE requires estimating the entire distribution $P_{Y,T,Z}$. To assess the utility of confounded data, we decompose

$P_{Y,T,Z}$ into two components: (i) the confounded distribution $P_{Y,T}$; and (ii) the conditional distributions $P_{Z|Y,T}$. Given infinite confounded data, the confounded distribution $P_{Y,T}$ is known exactly, reducing the number of free parameters in $P_{Y,T,Z}$ by three. The deconfounded data can then be used exclusively to estimate the conditional distributions $P_{Z|Y,T}$. To ease notation, let $a_{yt} = P_{Y,T}(y,t)$, and let $q^z_{yt} = P_{Z=z|Y,T}(y,t)$. Moreover, let $\mathbf{a}{:=}\ (a_{00}, a_{01}, a_{10}, a_{11})$, and let $\mathbf{q}$ denote the vector that contains q^z_{yt} for all values of Y, T and Z.

Hardness of The Problem We first show that for particular choices of the conditional distributions, this estimation problem can be arbitrarily hard for any confounded distribution $\mathbf{a}$. In particular, we show that for every fixed confounded distribution encoded by $\mathbf{a}$ and for any finite amount of deconfounded data m, there exist two conditional distributions encoded by $\mathbf{q}$'s such that we cannot distinguish these two distributions with high probability while their corresponding ATE values are constant away from each other. Let $\text{ATE}_{\mathbf{a}}(\mathbf{q})$ denote the value of the ATE when evaluated under the distributions $\mathbf{a}$ and $\mathbf{q}$. Then, we have

Proposition 1. *For every $\mathbf{a}$, there exists some ϵ, δ such that for any fixed number of deconfounded samples m, we can always construct a pair of $\mathbf{q}$'s, say $\mathbf{q_1}$ and $\mathbf{q_2}$, such that no algorithm can distinguish these two conditional distributions with probability more than $1-\delta$, and their corresponding ATE values are ϵ away: $|\text{ATE}_{\mathbf{a}}(\mathbf{q_1}) - \text{ATE}_{\mathbf{a}}(\mathbf{q_2})| \geq \epsilon$.*

Here, ϵ is a function of the confounded distribution $\mathbf{a}$, and the values of $\mathbf{q_1}$ and $\mathbf{q_2}$ depend on the variable δ and the number of deconfounded samples, m. The proof of Proposition 1 (Appendix B.3) relies on constructing a pair of $\mathbf{q_1}$ and $\mathbf{q_2}$ such that the value of $|\text{ATE}_{\mathbf{a}}(\mathbf{q_1}) - \text{ATE}_{\mathbf{a}}(\mathbf{q_2})|$ is constant for all confounded distribution $\mathbf{a}$ where the entries of $\mathbf{a}$ are strictly positive. In particular, this happens when the entries of the conditional distribution $\mathbf{q}$ approach to 0.

Unless otherwise mentioned, in the rest of the paper, we assume that each entry of the conditional distribution, q^z_{yt}, is bounded within the interval $[\beta, 1-\beta]$, for some small positive constant β. We first provide a lower bound on the sample complexity needed for any algorithm and any confounded distribution:

Theorem 2. *(Lower Bound) For any estimator and sample selection policy, the number of deconfounded samples m needed to achieve $P\left(\left|\widehat{\text{ATE}} - \text{ATE}\right| \geq \epsilon\right) < \delta$ is at least $\Omega(\epsilon^{-2}\log(\delta^{-1}))$.*

The proof of Theorem 2 (Appendix B.4) proceeds by construction. When comparing this lower bound with the upper bounds that we will present later, we observe that our sample complexities are tight up to a constant.

In the rest of the section, we first derive an upper bound of the sample complexity of a natural policy that is analogous to passive sampling (Theorem 3). We then derive the worst-case upper bound over all possible conditional distributions, $P_{Z|Y,T}$, in Corollary 1. Next, we propose two additional sampling policies, one of which enjoys an instance independent guarantee over the worst-case conditional distribution in $P_{Z|Y,T}$. We compare these sampling policies by investigating their sample complexity upper bounds (Theorem 5), worst-case upper bounds (Corollary 2), and lower bounds (Theorem 6). Table 1 summarizes our worst-case upper bound and lower bound results. In addition, we derive a worst-case sample complexity guarantee of our proposed sampling policy in Theorem 4.

Let $\hat{q}^z_{yt}$ be the empirical estimate of q^z_{yt} from the confounded data using the MLE (where m confounded data were deconfounded randomly). Then, we will always calculate $\widehat{\text{ATE}}$ by plugging the a_{yt}'s and $\hat{q}^z_{yt}$'s into Equation (1). The following theorem upper bounds the sample complexity for this estimator (later, we refer to this sampling policy as the *natural* selection policy):

Theorem 3. *(Upper Bound) When incorporating (infinite) confounded data, $P(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta$ is satisfied if the number of deconfounded samples m is at least*

$$m_{\text{nsp}} := \max_{t,z} \frac{C \sum_y a_{yt}}{\left(\sum_y a_{yt} q^z_{yt}\right)^2} = \max_{t,z} \frac{P_T(t)}{P_{T,Z}(t,z)^2} C. \quad (2)$$

The proof of Theorem 3 is included in Appendix B.5. Notably, m_{nsp} is less than m_{base} for any problem instance, highlighting the value of confounded data.

In addition, when $q^z_{yt} \in [\beta, 1-\beta]$, the maximum of Equation (2) over $\mathbf{q}$ is obtained at $q^z_{yt} = \beta$ for some t, z. Let M_{nsp} be the worst-case m_{nsp} over all possible values of $\mathbf{q}$. Since $\min \max_t 1/(\sum_y a_{yt})$ is achieved when $\sum_y a_{yt} = 1/2$, $\max_t 1/(\sum_y a_{yt}) \geq 2$. Thus,

Corollary 1. *(Worst-Case Upper Bound Guarantee)*

$$M_{\text{nsp}} := \max_{\mathbf{q}} m_{\text{nsp}} = \max_t \frac{C}{\beta^2 \sum_y a_{yt}} \geq \frac{2C}{\beta^2}.$$

Sample Selection Policies One important consequence of our procedure for estimating the ATE is that the four conditional distributions are estimated separately: the deconfounded data is partitioned into four groups, one for each $(y,t) \in \{0,1\}^2$, and the empirical measures $\hat{q}^z_{yt}$'s are then calculated separately. This means that the procedure does *not* rely on the fact that the deconfounded data is drawn from the exact distribution $P_{Y,T,Z}$, and in particular, the draws might as well have been made directly from the conditional

	w	M
NSP	$\beta^{-2}C_1 \max_t \left(\frac{a_{1t}(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2}, \frac{a_{0t}(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2} \right)$	$\beta^{-2}C \max_t \frac{1}{\sum_y a_{yt}}$
USP	$4\beta^{-2}C_1 \max_t \left(\frac{a_{1t}^2(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2}, \frac{a_{0t}^2(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2} \right)$	$4\beta^{-2}C \max_t \frac{\sum_y a_{yt}^2}{(\sum_y a_{yt})^2}$
OWSP	$2\beta^{-2}C_1 \max_t \left(\frac{a_{1t}(\sum_y a_{y\bar{t}})^2}{\sum_y a_{yt}}, \frac{a_{0t}(\sum_y a_{y\bar{t}})^2}{\sum_y a_{yt}} \right)$	$2\beta^{-2}C$

Table 1: Comparison between the instance-specific lower bound (w) and the worst-case upper bound (M).

distributions $P_{Z|Y,T}$. Suppose now that we can draw directly from these conditional distributions. This situation may arise when the confounder is fixed (like a genetic trait) and can be observed retrospectively. We now ask, given a budget for selective deconfounding samples, how should we allocate our samples among the four groups ($(y,t) \in \{0,1\}^2$)?

Let $\mathbf{x} = (x_{00}, x_{01}, x_{10}, x_{11})$ denote a selection policy with x_{yt} indicating the proportion of samples allocated to group (y,t), and $\sum_{yt} x_{yt} = 1$. We consider the following three non-adaptive selection policies:

1. **Natural (NSP):** $x_{yt} = a_{yt} = P_{Y,T}(y,t)$—this is similar to drawing from $P_{Y,T,Z}$, and is analogous to passive sampling.

2. **Uniform (USP):** $x_{yt} = 1/4$. Splits samples evenly across all four conditional distributions.

3. **Outcome-weighted (OWSP):** $x_{yt} = \frac{a_{yt}}{2\sum_y a_{yt}} = P_{Y|T}(y|t)/2$. Splits samples evenly across treatment groups ($T = 0$ vs. 1), and within each treatment group, choosing the number of samples to be proportional to the outcome ($Y = 0$ vs. 1).

While the particular form of OWSP appears to be the least intuitive, later we show it was in fact the unique policy that provides an instance independent guarantee when considering the worst-case $\mathbf{q}$'s.

For some fixed ϵ and δ, let μ_{nsp} be the minimum number of samples needed to achieve $P(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta$ under the *natural* selection policy over all estimators. We similarly define μ_{usp} and μ_{owsp}. Then Theorem 3 provides an upper bound on μ_{nsp} by studying the upper bound of a specific estimator. Before we provide an upper bound of the sample complexity of μ_{usp} and μ_{owsp}, we first establish that μ_{nsp} may be significantly worse than μ_{owsp}, but μ_{owsp} is never *much* worse.

Theorem 4. *For any fixed $\epsilon \in [0, 0.5 - 2\beta(1-\beta)]$ and any fixed $\delta < 1$, there exist distributions where $\mu_{\text{owsp}}/\mu_{\text{nsp}}$ is arbitrarily close to zero. In addition, for any estimator and every distribution, $\mu_{\text{owsp}}/\mu_{\text{nsp}} \leq 2$.*

The proof of Theorem 4 (Appendix B.6) proceeds by construction. Note that the upper bound of ϵ in Theorem 4 is not necessary the maximum achievable ϵ. Instead it provides a range where Theorem 4 holds. Next, we provide the upper bounds of μ_{usp} and μ_{owsp} by analyzing our algorithm (analogous to Theorems 1 and 3):

Theorem 5. *(Upper Bound) Under the uniform selection policy, with (infinite) confounded data incorporated, $P(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta$ is satisfied if μ_{usp} is at least*

$$m_{\text{usp}} := \max_{t,z} \frac{C\sum_y 4a_{yt}^2}{\left(\sum_y a_{yt} q_{yt}^z\right)^2} = \max_{t,z} \frac{4\sum_y P_{Y,T}(y,t)^2}{P_{T,Z}(t,z)^2} C.$$

Similarly, for the outcome-weighted selection policy:

$$m_{\text{owsp}} := \max_{t,z} \frac{2C\left(\sum_y a_{yt}\right)^2}{\left(\sum_y a_{yt} q_{yt}^z\right)^2} = \max_{t,z} \frac{2}{P_{Z|T}(z|t)^2} C.$$

The proofs of Theorems 3 and 5 (Appendix B.5), which differ from that of Theorem 1, require a modification to Hoeffding's inequality (Appendix, Lemma 4), which we derive to bound the sample complexity of the weighted sum of two independent random variables. Theorem 5 points to some *additional* advantages of OWSP. First, OWSP has the nice property that the sufficient number of samples, m_{owsp}, does not depend on $P_{Y,T}$. Second, a comparison of the quantities m_{usp} and m_{owsp} suggests that USP is strictly dominated by OWSP, since $4a_{0t}^2 + 4a_{1t}^2 - 2(a_{0t} + a_{1t})^2 = 2(a_{0t} - a_{1t})^2 \geq 0$. We might hope for a similar result by comparing m_{owsp} with m_{nsp} from comparing Theorems 3 and 5, but neither strictly dominates the other. Instead, recall that Theorem 4 shows that μ_{nsp} may be significantly worse than μ_{owsp}, but μ_{owsp} is never *much* worse. Similar to Corollary 1, we now derive an equivalent corollary for Theorem 5 where we consider the worst case over $\mathbf{q}$'s.

Let M_{usp} and M_{owsp} be the maximum values of m_{usp} and m_{owsp}, respectively, over all possible values of $\mathbf{q}$.

Corollary 2. *(Worst-Case Upper Bound Guarantee)*

$$M_{\text{usp}} = \max_t \frac{4C\sum_y a_{yt}^2}{\beta^2(\sum_y a_{yt})^2};\ M_{\text{owsp}} = \max_t \frac{2C}{\beta^2} \leq M_{\text{nsp}}.$$

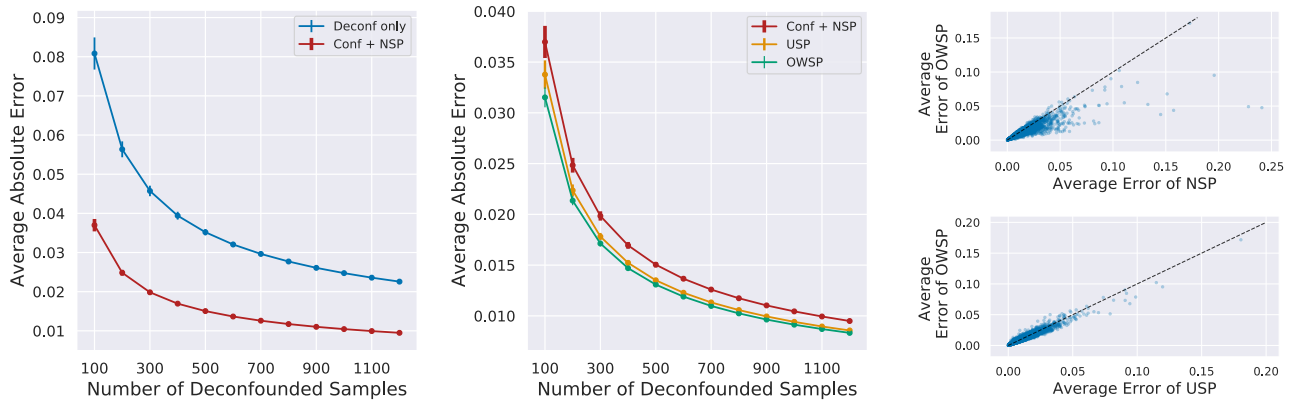

Figure 2: Performance of the four policies over 13,000 distributions $P_{Y,T,Z}$, given infinite confounded data. Left and Middle: averaged error over 13,000 distributions for varying numbers of deconfounded samples. Right: error comparison (each point is a single distribution averaged over 100 replications) for 1,200 deconfounded samples.

First, note that M_{owsp} is independent of the confounded distribution $\mathbf{a}$. Furthermore, from the proof of Theorem 5, we observe that OWSP is the unique policy that makes this upper bound independent of $\mathbf{a}$. When comparing Corollaries 1 and 2, we observe that OWSP always dominates NSP when taking the worst case over $\mathbf{q}$'s. Lastly, we provide the lower bounds of $\mu_{\text{nsp}}, \mu_{\text{usp}}$, and μ_{owsp} that are analogous to Theorem 2:

Theorem 6. *(Lower Bound) For every* $\mathbf{a}$*, there exists a* $\mathbf{q}$ *such that* μ_{nsp} *is at least*

$$w_{\text{nsp}} := \frac{C_1}{\beta^2} \max_t \left(\frac{a_{1t}(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2}, \frac{a_{0t}(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2} \right);$$

similarly for uniform selection policy:

$$w_{\text{usp}} := \frac{C_1}{\beta^2} \max_t \left(4\frac{a_{1t}^2(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2}, 4\frac{a_{0t}^2(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2} \right);$$

similarly for outcome-weighted sample selection policy:

$$w_{\text{owsp}} := \frac{C_1}{\beta^2} \max_t \left(2\frac{a_{1t}(\sum_y a_{y\bar{t}})^2}{\sum_y a_{yt}}, 2\frac{a_{0t}(\sum_y a_{y\bar{t}})^2}{\sum_y a_{yt}} \right),$$

where $\bar{t} = 1 - t$ *and* $C_1 \propto (k\beta - 1)^2 \ln(\delta^{-1})\epsilon^{-2}$.

The proof (Appendix B.7) proceeds by construction. Table 1 summarizes our worst-case upper bounds and instance-specific lower bounds. When comparing the constants C and C_1, we observe that the upper bounds and lower bounds match in k, ϵ, and δ, demonstrating the relative tightness of our analysis.

We have shown the advantages of OWSP *given an infinite amount of confounded data*. However, in practice, the confounded data is finite. In Appendix C, we analyze the sample complexity upper bound of our algorithm under *finite* confounded data. One new issue that arises with finite confounded data is that a sampling policy may not be feasible because there are not enough confounded samples to deconfound. In our experiments, when this happens, we approximate the target sampling policy as closely as is feasible (see Appendix E).

4 Experiments

Since the upper bounds that we derived in Section 3 are not necessarily tight, we first perform synthetic experiments to assess the tightness of our bounds. For the purpose of illustration, we focus on binary confounders Z throughout this section, and denote $q_{yt} = P_{Z=1|Y,T}(y,t)$. We first compare the sampling policies in synthetic experiments on randomly chosen distributions $P_{Y,T,Z}$, measuring both the average and worst-case performance of each sampling policy. We then measure the effect of having finite (vs. infinite) confounded data. Finally, we test the performance of OWSP on real-world data taken from a genetic database, COSMIC, that includes genetic mutations of cancer patients (Tate et al., 2019; Cosmic, 2019). Because this is (to our knowledge) the first paper to investigate the problem of *selective deconfounding*, the methods in described Section 2 are not directly comparable to ours.

4.1 Infinite Confounded Data

Assuming access to infinite confounded data, we experimentally evaluate all four sampling methods for estimating the ATE: using deconfounded data alone, and using confounded data that has been selected according to NSP, USP, and OWSP. Let $\mathbf{a} :=$

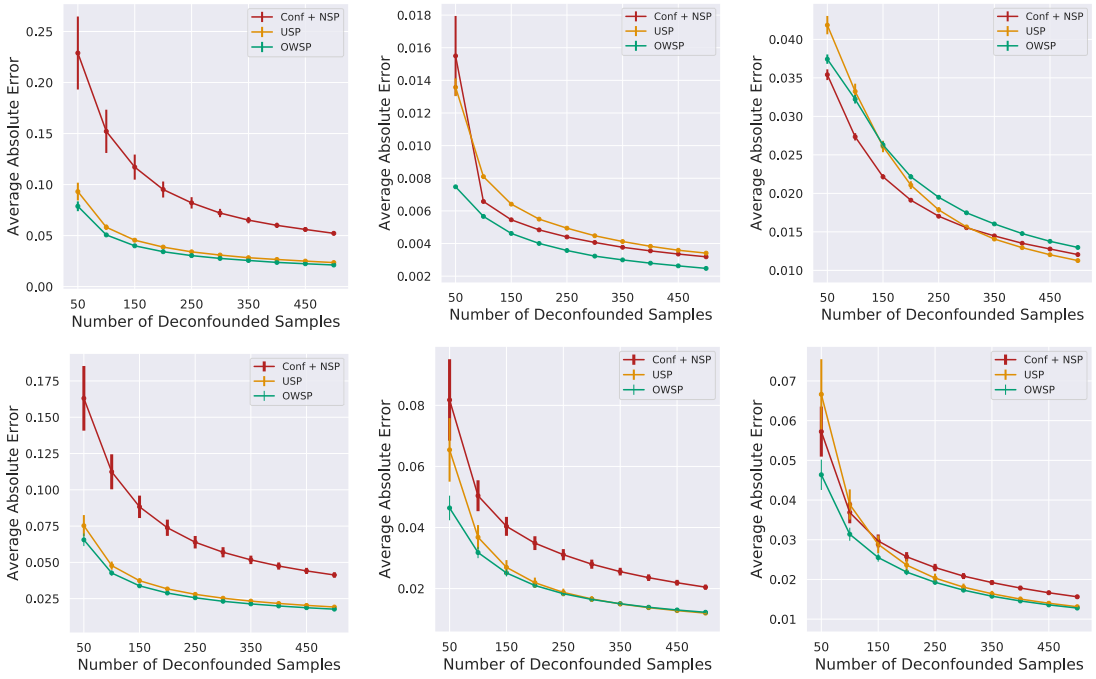

Figure 3: Comparison of selection policies for adversarially chosen instances. Top (left) NSP performs the worst: $\mathbf{a} = (0.9, 0.02, 0.01, 0.07)$ and $\mathbf{q} = (0.9, 0.7, 0.01, 0.3)$. Top (middle) USP performs the worst: $\mathbf{a} = (0.79, 0.01, 0.02, 0.18)$ and $\mathbf{q} = (0.5, 0.01, 0.05, 0.5)$. Top (right) OWSP performs the worst: $\mathbf{a} = (0.5, 0.01, 0.19, 0.3)$ and $\mathbf{q} = (0.05, 0.5, 0.055, 0.4)$. Bottom: the same $\mathbf{a}$'s but averaged over 500 $\mathbf{q}$'s drawn uniformly from $[0, 1]^4$.

$(a_{00}, a_{01}, a_{10}, a_{11})$, and $\mathbf{q} := (q_{00}, q_{01}, q_{10}, q_{11})$, encoding the confounded and conditional distributions, respectively. We evaluate the performance of four methods in terms of the *absolute error*, $|\widehat{\text{ATE}} - \text{ATE}|$. Because the variance of our estimators cannot be analyzed in closed form, we report the variance of the *absolute error* averaged over different instances in terms of the error bar in the figures.

Randomly Generated Instances We first evaluate the four methods over a randomly selected set of distributions. Figure 2 was generated by averaging over 13,000 instances, each with the distribution $P_{Y,T,Z}$ drawn uniformly from the unit 7-Simplex. Every instance consists of 100 replications, each with a random draw of 1,200 deconfounded samples. The absolute error is measured as a function of the number of deconfounded samples in steps of 100 samples. Figure 2 (left) compares the use of deconfounded data along with the incorporation of confounded data selected naturally (as in the comparison of Theorems 1 and 3). It shows that incorporating confounded data yields a significant improvement in estimation error. For example, achieving an absolute error of 0.02 using deconfounded data alone requires more than 1,200 samples on average, while by incorporating confounded data, only 300 samples are required. We observe that the variance of our estimator has decreased dramatically by incorporating infinite confounded data. Having established the value of confounded data, Figure 2 (middle) compares the three selection policies. We find that OWSP outperforms both NSP and USP in terms of both the absolute error and the variance when averaged over joint distributions. To compare the performance of our sampling policies on an instance level, we provide two scatter plots in Figure 2 (right), each containing the 13,000 instances in the left figures and averaged over 100 replications. The number of deconfounded samples is fixed at 1,200. We observe that OWSP outperforms NSP and USP in the majority of instances.

Worst-Case Instances We evaluate the performance of the three selection policies on joint distributions chosen adversarially against each in Figure 3. The three sub-figures (the columns) correspond to instances where NSP, USP, and OWSP perform the worst, respectively, from the left to the right. Each sub-figure is further subdivided: the top contains results for the single adversarial example while the bottom is averaged over 500 $\mathbf{q}$'s sampled uniformly from $[0, 1]^4$. The absolute error is averaged over 10,000 replications in the left figures and over 500 in the right. In all cases, we draw 500 deconfounded samples and measure the absolute error in steps of 50 samples. Figure 3 (left) validates

Figure 4: Experiment on finite confounded data over 13,000 distributions $P_{Y,T,Z}$, each averaged over 100 replications. The number of deconfounded samples is fixed at 100. Left: averaged over the 13,000 distributions. Middle and Right: error comparison at 681 confounded samples.

Corollary 4. We observe that when the distribution of $\mathbf{a}$ is heavily skewed towards $(Y = 0, T = 0)$, OWSP and USP significantly outperform NSP. Figure 3 (middle) shows that USP can underperform NSP, but when averaged over all possible values of $\mathbf{q}$, USP performs better than NSP. Figure 3 (right) shows that OWSP can underperform NSP and USP, but, when compared with the left and middle column, the performance of OWSP is close to that of NSP and USP. When averaged over all possible values of $\mathbf{q}$, OWSP outperforms both. Moreover, OWSP's variance is the lowest across all scenarios. Appendix D provides examples in which each of these joint distributions could appear.

4.2 Finite Confounded Data

Given only n confounded data, we test the performance of the OWSP against NSP and USP. In Figure 4, the absolute error is measured as a function of the number of confounded samples in step sizes that increment in the log scale from 100 to 10,000 while fixing the number of deconfounded samples to 100. Since when we only have 100 confounded samples, the three sampling policies are identical, the error curves corresponding to NSP, USP and OWSP start at the same point on the top left corner in Figure 4 (left). We observe that as the number of confounded samples increases, OWSP quickly outperforms NSP and USP on average, and the gaps between OWSP and the other two selection policies widen. Since we fix the number of deconfounded samples to be 100, 1) all three sampling policies are equivalent when there are only 100 confounded samples in the dataset (i.e., we need to deconfound all 100 confounded samples in all cases), and 2) the average absolute errors of the three selection policies do not converge to 0 in Figure 4. Figure 4 (middle and right) compare the performance of OWSP with that of the NSP and USP, respectively, on an instance level. We observe that OWSP dominates NSP and USP in the majority of instances by both the absolute error and variance. Note that if we fix the number of confounded samples and increase the number of deconfounded samples (with $m \leq n$), we observe that OWSP dominates USP and NSP when the number of deconfounded samples is small. The gap shrinks as the number of deconfounded samples increases. When at $m = n$, all three methods are equivalent.

4.3 Real-World Experiments: COSMIC

Data Previously, we chose the underlying distribution $P_{Y,T,Z}$ uniformly from the unit 7-Simplex. However, real-world problems of interest may not be uniformly distributed. To illustrate the practicality of our methods, we consider a real-world dataset, picking three variables to be the outcome, treatment, and confounder, and artificially hiding the confounder for some samples. Finally, we evaluate our proposed sampling methods under the assumption that we have access to infinitely many confounded samples. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a public database of DNA sequences of tumor samples. It consists of targeted gene-screening panels aggregated and manually curated over 25,000 peer reviewed papers. We focus on the variables: `primary cancer site` and `gene`. Specifically, for 1,350,015 cancer patients, we observe their cancer types, and for a subset of genes, whether or not a mutation was observed in each gene.

Causal Models In our experiments, we designate cancer type as the outcome, a particular mutation as the treatment, and another mutation as the confounder—this setup seems reasonable because it is well known that multiple genetic mutations are correlated with individual cancer types (Knudson, 2001), and that mutations can cause both cancer itself and other mutations. As a concrete example, mutations in the genes that code RNA polymerases (responsible for

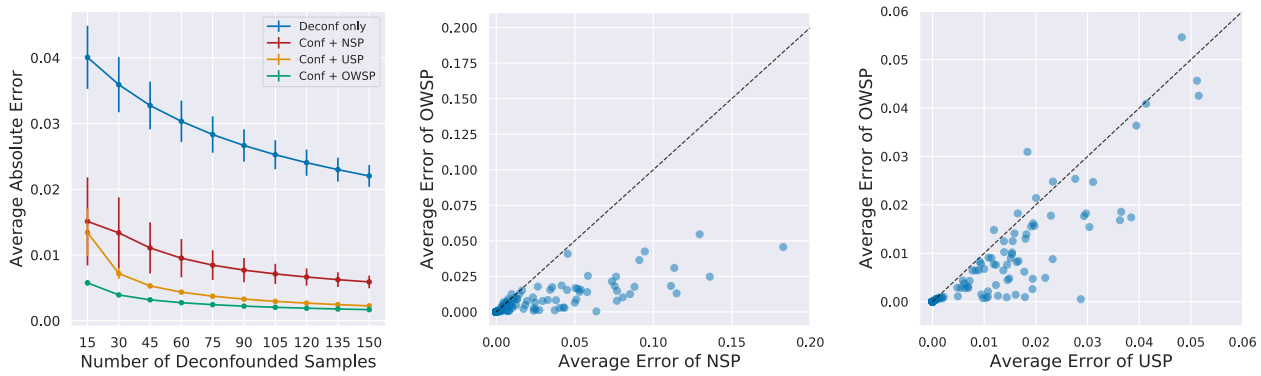

Figure 5: Performance of the four sampling policies on the COSMIC dataset assuming infinite confounded data. 275 unique (cancer, mutation, mutation) combinations were extracted. Left: averaged over 275 instances, and each averaged over 10,000 replications. Middle and Right: error comparison at 45 deconfounded samples.

ensuring the accuracy of replicating RNA) are found to increase the likelihood of both other mutations and certain cancer types (Rayner et al., 2016). The setting where the treatment mutation and cancer outcome are observed and the confounding mutation is unobserved is plausible because it is common that the majority of patients only have a subset of genes sequenced (e.g. from a commercial panel). The top 6 most commonly mutated genes were selected as treatment and confounder candidates. For each combination of a cancer type and two of these genes, we removed patients for whom these genes was not sequenced, and kept all pairs that had at least 40 patients in each of the four treatment-outcome groups (to ensure our deconfounding policies would have enough samples to deconfound). This procedure gave us 275 unique combinations of a cancer (outcome), mutation (treatment), and another mutation (confounder). Since on average, each {cancer, mutation, mutation} tuple contains around 125,883 patients, we took the estimated empirical distribution as the data-generating distribution and applied the ATE formula described in Section 3 to obtain the true ATE. To model the unobserved confounder, we hid the confounding mutation parameter, only revealing it to a sampling policy when it requested a deconfounded sample. We compared the use of deconfounded data along with the incorporation of confounded data under the three sampling selection polices: NSP, USP, and OWSP.

Results Figure 5 (left) was generated with these 275 instances each repeated for 10,000 replications. The absolute error is measured as a function of the number of deconfounded samples in step sizes of 15. First, similar to Figure 2, we observe that incorporating confounded data reduces both the absolute estimation error and the variance of the estimator by a large margin. Note the improvement of OWSP over NSP is larger in this case as compared to that seen in Figure 2. Furthermore, when the number of deconfounded samples is small, OWSP outperforms USP. Note that Figure 5 (left) does not start with 0 because absent any deconfounded data, the estimated ATE is the same for all sampling policies. In Figure 5 (middle, right), we fix the number of deconfounded samples to be 45 and compare the performance of OWSP against that of NSP and USP, respectively. Both figures contain the 275 instances in the left figure, averaged over 10,000 replications. We observe that under this setup, OWSP dominates NSP in all instances, and outperforms USP in the majority of instances.

5 Conclusion

We propose the problem of causal inference with *selectively deconfounded* data. This problem is particularly motivated by the scenarios where interventions on treatment is not available. We theoretically analyze the upper bounds and lower bounds on the amount of deconfounded data required under each sample selection policy. In addition, we theoretically demonstrate that the best-case gain of our proposed policy OWSP is unbounded when compared with NSP while the worst-case relative performance is bounded. We point to several promising directions for potential future research. First, we are currently extending our analysis to the adaptive case using ideas from active learning and combinatorial optimization. Second, we plan to extend our results to more general causal problems, including linear and semi-parametric causal models. Finally, we may extend the idea of selective revelation of information beyond confounders to incorporate mediators and proxies.

Acknowledgments

We wish to thank Sivaraman Balakrishnan and Uri Shalit for their valuable feedback. We would also like to thank Amazon AI, Facebook, Salesforce, the NSF, UPMC, the PwC Center, and DARPA for their support of our research.

References

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

Stephen Bates, Matteo Sesia, Chiara Sabatti, and Emmanuel Candes. Causal inference in genetic trio studies. *arXiv preprint arXiv:2002.09644*, 2020.

Cosmic. Cosmic - catalogue of somatic mutations in cancer, Sep 2019. URL https://cancer.sanger.ac.uk/.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778, 2015.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4): 1161–1189, 2003.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960, 1986.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pages 10888–10897, 2018.

Alfred G Knudson. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1(2):157–162, 2001.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403, 2004.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

Emily Rayner, Inge C van Gool, Claire Palles, Stephen E Kearsey, Tjalling Bosse, Ian Tomlinson, and David N Church. A panoply of errors: polymerase proofreading domain mutations in cancer. *Nature Reviews Cancer*, 2016.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085, 2017.

Xu Shi, Wang Miao, Jennifer C Nelson, and Eric J Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540, 2020.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1): D941–D947, 2019.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.