Statistical Inference Under Constrained Selection Bias

Santiago Cortes-Gomez [1] **Mateo Dulce** [2] **Carlos Patino** [3] **Bryan Wilder** [1]

Abstract

Large-scale datasets are increasingly being used to inform decision making. While this effort aims to ground policy in real-world evidence, challenges have arisen as selection bias and other forms of distribution shifts often plague observational data. Previous attempts to provide robust inference have given guarantees depending on a user-specified *amount* of possible distribution shift (e.g., the maximum KL divergence between the observed and target distributions). However, decision makers will often have additional knowledge about the target distribution which constrains the *kind* of possible shifts. To leverage such information, we propose a framework that enables statistical inference in the presence of selection bias which obeys user-specified constraints in the form of functions whose expectation is known under the target distribution. The output is high-probability bounds on the value of an estimand for the target distribution. Hence, our method leverages domain knowledge in order to partially identify a wide class of estimands. We analyze the computational and statistical properties of methods to estimate these bounds and show that our method can produce informative bounds on a variety of simulated and semisynthetic tasks, as well as in a real-world use case.

1. Introduction

Decision-makers in the public and private sectors increasingly rely on machine learning or statistical models built on top of large-scale datasets in order to inform policy, operational decisions, individualized treatment rules, and more. However, these administrative datasets are typically purely observational, meaning that they are not carefully designed to sample from a true distribution of interest. Consequently, such efforts have been hindered by sampling biases and other distributional shifts between the observed data and the target distribution. Such selection biases have presented severe issues for past algorithmic systems (Chouldechova et al., 2018; Gianfrancesco et al., 2018), policy analysis (Harron et al., 2017; Knox et al., 2020), epidemiological studies (Jensen et al., 2015; Haneuse & Daniels, 2016), among others.

As a motivating example, decision-makers in public health want to target preventative interventions to mitigate the impact of a COVID-19 outbreak. To achieve this aim, they may wish to estimate the risk factors contributing to a particular adverse outcome related to COVID-19, e.g. hospitalizations. Nevertheless, they typically only will have data from individuals who engaged with some specific services in the past, such as patients who visited the healthcare system or insurance claims. Furthermore, it is a known phenomenon that individuals with a low income background will seek care primarily when facing severe diseases, thus increasing observed hospitalizations for this group. Hence, the selection bias present in the data will be driven by some latent factors (for instance, high healthcare costs) and the outcome itself; thus creating a problem similar to confounding when estimating the relationship between these two quantities.

We consider the task of accurately estimating some functional of the ground-truth distribution using samples from an observed, potentially shifted, distribution. For instance, our goal might be to estimate the expectation of a measured covariate or the treatment effect from an intervention. In general, such inference is intractable without assumptions on the relationship between the observed and target distributions. The simplest setting is where we can directly observe some samples from the target distribution. However, we are interested in scenarios where this is not the case, for example, the one introduced in the previous paragraph; given that the data is conditioned among other things on the outcome, it presents the insidious problem that it will never be representative of the population as whole. Absent target-distribution samples, other frameworks such as those related to distributionally robust optimization (DRO) (Gupta & Rothenhäusler, 2021; Bertsimas et al., 2022; Yadlowsky et al., 2022; Rothenhäusler & Bühlmann, 2022) or sensitivity analysis (Robins et al., 2000; Tan, 2006; Ding &

[*]Equal contribution [1]Department of Machine Learning, Carnegie Mellon University [2]Department of Statistics, Carnegie Mellon University [3]Factored AI. Correspondence to: Santiago Cortes-Gomez <scortesg@cs.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

VanderWeele, 2016) allow users to specify the total *magnitude* of distribution shift allowed. Such magnitude is usually modeled by imposing a limit on the maximum distance (e.g., KL or χ^2 divergence) between the target distribution and the observed one. However, since the target distribution (or samples of it) is unavailable, this assumption cannot be empirically verified, and often cannot even be set in an informed manner. Our goal is to provide provable guarantees *without introducing untestable assumptions on the magnitude of shift*, by using only observable quantities to set bounds on the possible difference in distributions.

Intuitively, this may be possible because decision makers often have aggregate knowledge about the target distribution. For instance, a policymaker may know, via census data, the distribution of demographic characteristics such as age, race, or income from the population as a whole. Moreover, in the public health setting, serosurveys may provide ground-truth estimates of exposure to infectious diseases in specific locations or population groups (Havers et al., 2020). This aggregate information is typically not sufficient by itself for the original task because it fails to measure the key outcome or covariates of interest (e.g., knowing the demographic distribution of a population is by itself unhelpful for estimating a patient's risk of hospitalization). However, it imposes constraints on the nature of the distribution shift between the observed distribution and the underlying population: any valid shift must respect these known quantities.

Concretely, our contributions are as follows:

- We introduce a framework that allows user-specified constraints on distribution shift via known expectations. Our framework incorporates such external information into an optimization program whose value gives valid bounds on a statistic of interest using samples from an observed distribution. As a result, we are able to provide estimates by adding more observables in lieu of assumptions that are not empirically testable (e.g., the KL divergence between the target and sample distributions).
- We analyze the statistical properties of estimating these bounds using a sample average approximation for the optimization problem. We show that our estimators are asymptotically normal, allowing us to provide valid confidence intervals.
- We extend our framework to accommodate estimands without a closed form (e.g., a regression coefficient or an estimated model parameter), provide statistical guarantees for this setting, and propose computational approaches to solve the resulting optimization problem.
- We perform experiments on synthetic and semi-synthetic data to test our methods. The results empirically confirm that our framework provides valid bounds for the target estimand and allows effective use of domain knowledge: incorporation of more informative constraints produces tighter bounds. All the code for our experiments, together with a guide on how to replicate each one of them, is available at `https://github.com/secg5/inference_contrained_distribution_shift`.
- We showcase a real-world case of study of assessing disparities in COVID-19 disease burden using a large-scale insurance claims dataset where our method allows us to produce policy-relevant results under the presence of selection bias.

Additional related work. Our work is broadly related to the literature on partial identification, which spans statistics, economics, epidemiology, and computer science (Manski, 2003; Tamer, 2010; Ho & Rosen, 2015; Molinari, 2020). Interest in partial identification has grown recently in the machine learning community, particularly in the causal inference setting. For example, (Hu et al., 2021; Balazadeh Meresht et al., 2022) consider partial identification of treatment effects for a known causal graph, extending the classic framework of (Balke & Pearl, 1997) to incorporate generative models. (Guo et al., 2022) consider the setting where covariates are subject to a user-specified level of noise. Perhaps closer to our setting is (Padh et al., 2023). They consider methods for incorporating domain knowledge into partial identification problems. However, in their setting, domain knowledge takes the form of functional form restrictions for the treatment effect (e.g., smoothness or number of inflection points), rather than constraints on shift between observed and target distributions. Similar ideas in the context of prediction are explored by Bertail et al. (2021); they provide finite sample guarantees for the empirical risk minimization problem under the assumption that the learned parameter is part of a well behaved parametric class of functions. However, their assumption shares the same problem as the current methods in the literature: it is not empirically verifiable. Our work differs from the existing ML literature both in that we are concerned with inference problems broadly (not restricted to treatment effects), and in that we provide a means to impose externally-known constraints on shifts such as selection bias. Related issues have recently been considered in the epidemiology and biostatistics literature, motivated by the growing use of biobank-style datasets with known selection biases (Tudball et al., 2019; Horowitz & Lee, 2022). Our work differs in that we explicitly consider the algorithmic properties of computing the resulting bounds, and in our analysis of estimators which themselves require fitting a model.

2. Problem formulation

Let X be a random variable over a space $\mathcal{X}$, with distribution $\mathbb{P}$ in a population of interest. Our goal is to estimate a real-valued functional $f(\mathbb{P})$. One prominent example is estimating the expectation $f(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}[h(X)]$ for some function h, but we will consider a range of examples for f. If we observe iid samples from $\mathbb{P}$, estimating $f(\mathbb{P})$ is a standard inference task. For instance, we may use the plug-in estimate $f(\hat{\mathbb{P}}_N)$, (where $\hat{\mathbb{P}}_N$ is the empirical distribution) or any number of other strategies. However, we consider a setting where we instead observe samples drawn iid from a different distribution $\mathbb{Q}$. For illustration purposes, recall the example mentioned in the introduction: policymakers are interested in information relative to the population as a whole ($f(\mathbb{P})$) but they only have access to observational data from the hospitals or insurance claims, i.e., a sample from $\mathbb{Q}$.

We assume that both $\mathbb{P}$ and $\mathbb{Q}$ have densities p and q, respectively, and that $q(X) > 0$ whenever $p(X) > 0$. More formally, we require that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ so that the ground truth density ratio $\theta_0(X) = \frac{p(X)}{q(X)}$ is well defined. Intuitively, inference on $\mathbb{P}$ is impossible if some portions of the distribution can never be observed. Accordingly, similar overlap assumptions are near-universal in causal inference (Imbens, 2004), domain adaptation (Byrd & Lipton, 2019), selection bias (Rosenbaum & Rubin, 1983), distributionally robust optimization (Namkoong & Duchi, 2016), etc. We expect this assumption to be satisfied in many application domains because selection bias is not strong enough to make some observations literally impossible. Our running COVID-19 example illustrates this: a patient with any covariates X can in principle seek care (and some such patients do), even if they have worse access than patients with a different set of covariates X'.

With slight abuse of notation, we let $\theta\mathbb{P}$ denote a distribution with density $\theta(X)p(X)$, so under our definition of $\theta_0(X)$, $\theta_0\mathbb{Q} = \mathbb{P}$. In the case of estimating an expectation, we have that $\mathbb{E}_{X\sim\mathbb{P}}[h(X)] = \mathbb{E}_{X\sim\mathbb{Q}}[\theta_0(X)h(X)]$. If θ_0 was known, estimating $f(\mathbb{P})$ using samples from $\mathbb{Q}$ would be easily accomplished using standard methods (importance sampling). However, our focus is on the setting where θ_0 is *not* known, and we don't have access to samples from $\mathbb{P}$ from which to estimate it.

Without any information about the relationship between $\mathbb{P}$ and $\mathbb{Q}$ this is a hopeless task. However, in many settings of importance *some* information is available, in the form of auxiliary functions whose true expectation under $\mathbb{P}$ is known. In policy settings, this may come from census data, or well-design population surveys that estimate the ground-truth prevalence of specific health conditions (c.f. (Murray & Lopez, 2013; Johnson et al., 2014; Havers et al., 2020)). Concretely, suppose we are able to observe $\mathbb{E}_{\mathbb{P}}[g_j(X)]$ for a collection of functions $g_1, \ldots, g_m$. Given that $\theta_0\mathbb{Q} = \mathbb{P}$, we have that θ_0 must satisfy

$$\mathbb{E}_{\mathbb{Q}}[\theta_0(X)g_j(X)] = \mathbb{E}_{\mathbb{P}}[g_j(X)], \quad j = 1, \ldots, m.$$

Let Θ to be the set of density ratios θ which satisfy the above constraints, all of which are consistent with our observations. In general, Θ will not be a singleton and so $f(\mathbb{P})$ will not be point-identified. However, it is *partially* identified with the following bounds:

Proposition 2.1.

$$f(\mathbb{P}) \in \left(\min_{\theta\in\Theta} f(\theta\mathbb{Q}), \quad \max_{\theta\in\Theta} f(\theta\mathbb{Q})\right),$$

and these bounds are tight.

In the following, we provide computationally and statistically efficient methods for estimating these upper and lower bounds, each of which are defined via an optimization problem over θ. Θ encapsulates the constraints on potential distribution shifts that are known in a particular domain, allowing an analyst to translate additional domain knowledge into tighter identification of the estimand. Intuitively, such identification will be close to point-wise when the constraints are informative enough about the behaviour of the estimand of interest under the true distribution. On the other hand, the bounds will be vacuous if there is no constraint on the joint distribution, when knowledge about one covariate does not inform the others. In general, we expect our method to be useful when some such joint information is available, and we provide several experimental examples of policy-relevant questions using both Census data as well as a real application to Covid-19 insurance claims data. Additionally, we compute explicit non-trivial bounds for an example in the Appendix.

Our aim is to use the observed sample $X_1, \ldots, X_N$ drawn iid from $\mathbb{Q}$ to estimate the value of the optimization problems defining our lower and upper bounds. Note that in order to do so, we have to tackle two different types of uncertainty. First, the uncertainty derived from partial identification; even if $\mathbb{Q}$ was known exactly, the population level quantities $\min_{\theta\in\Theta} / \max_{\theta\in\Theta} f(\theta\mathbb{Q})$ do not exactly identify $f(\mathbb{P})$. Second, finite-sample uncertainty from the estimation of the population level quantity $f(\theta\mathbb{Q})$ using a sample $X_1, \ldots, X_N$ from $\mathbb{Q}$: the statistical uncertainty from taking $f(\theta\hat{\mathbb{Q}}_N)$ as an approximation of $f(\theta\mathbb{Q})$. We will provide confidence bounds for $f(\mathbb{P})$ which account for both partial identification and finite-sample uncertainty in estimation. Throughout, we will assume for convenience that the optimization problems defining each side of the bound have a unique solution (if this is not satisfied, we could, e.g., modify the objective function to select a minimum-norm solution).

3. Methods

We now proceed to estimate the aforementioned bounds. Proposition 2.1 reduces the problem of computing such bounds to solving an optimization program. Throughout, we consider the problem of estimating the lower bound (i.e., solving the minimization problem), as the upper bound is symmetric. To start, we impose no restrictions on $\theta(X)$ by representing $\theta(X)$ as an arbitrary function. This yields the following optimization problem:

$$\begin{aligned} &\min_{\theta} f(\theta\mathbb{Q}) \\ &\theta(X) \geq 0 \quad \forall X, \\ &\mathbb{E}_{X\sim\mathbb{Q}}[\theta(X)g_j(X)] = \mathbb{E}_{\mathbb{P}}[g_j(X)] \quad j = 1, \ldots, m. \end{aligned} \tag{1}$$

where the constraint $\theta(X) \geq 0$ ensures that θ is a valid density ratio. However, neither $f(\theta\mathbb{Q})$ nor the population level restrictions can be observed directly. Nevertheless, having access to a sample from $\mathbb{Q}$, it is only natural to use plug-in estimators to estimate the unavailable quantities from the previous program. This yields the sample optimization problem:

$$\begin{aligned} &\min_{\theta} f(\theta\hat{\mathbb{Q}}_N) \\ &\theta(X_i) \geq 0 \quad \forall i = 1, \ldots, N, \\ &\frac{1}{N}\sum_{i=1}^{N} \theta(X_i)g_j(X_i) = \mathbb{E}_{\mathbb{P}}[g_j(X)] \quad j = 1, \ldots, m \end{aligned} \tag{2}$$

where $\hat{\mathbb{Q}}_N$ is the empirical distribution of $\mathbb{Q}$. We refer to the first problem (1) as the **population optimization problem** ($\min_\theta f(\theta\mathbb{Q})$) and to the second one (2) as the **plug-in approximation** ($\min_\theta f(\theta\hat{\mathbb{Q}}_N)$).

Let ν be the optimal value of the population problem and $\hat{\nu}_N$ the optimal value of the plug-in approximation problem. Suppose $\sqrt{n}(\hat{\nu}_N - \nu)$ converges in distribution to a normally distributed random variable, such that, it has mean zero and variance that can be estimated from the data. If this is indeed the case, then the standard procedure to derive valid confidence intervals for ν can be applied. Furthermore, as corollary from Proposition 2.1, we can produce a high probability lower bound for $f(\mathbb{P})$. As the maximization problem is symmetric, we can use the same argument to upper bound $f(\mathbb{P})$ with high probability.

We now turn to analyze several scenarios where this recipe can be applied. We use the theory of *sample average approximation* in optimization (Shapiro, 1991; Shapiro et al., 2021) to describe the conditions to guarantee, both convergence and asymptotic normality, for the statistic $\sqrt{n}(\hat{\nu}_N - \nu)$. However, before developing this mathematical scaffolding, it is necessary to make one more assumption: we will assume a differentiable and finite-dimensional parametrization of $\theta(X)$; such parametrization will be denoted by $\theta_\alpha(X)$ for some $\alpha \in \mathbb{R}^d$. Note that this class is still quite expressive, for instance, a standard basis for smooth functions (e.g., a set of polynomial basis functions) allows us to represent any smooth function in this framework. Alternatively, an analyst could use a highly expressive class such as neural networks (Hornik et al., 1989).

3.1. Convex estimands

We start with the case where $f(\theta_\alpha\mathbb{Q})$ is a convex function of α. The most prominent case where this holds is when f is the expectation of some function h, in which case $f(\theta_\alpha\hat{\mathbb{Q}}_N) = \frac{1}{N}\sum_{i=1}^{N}\theta_\alpha(X_i)h(X_i)$ is linear in α, and will be convex in α when $\alpha \mapsto \theta_\alpha$ is convex. Another example is when f is a conditional expectation, conditioned on some event $X \in C$. Then, we have

$$\begin{aligned} f(\theta_\alpha\mathbb{Q}) &= E_{X\sim\theta_\alpha\mathbb{Q}}[h(X)|X \in C], \\ f(\theta_\alpha\hat{\mathbb{Q}}_N) &= \frac{\sum_{i=1}^{N} 1[X_i \in C]\cdot\theta_\alpha(X_i)h(X_i)}{\sum_{i=1}^{N}\theta_\alpha(X_i)1[X_i \in C]}. \end{aligned}$$

The plug-in approximation of f is no longer linear in θ_α because θ_α determines both the numerator and the denominator. However, it can be reformulated as a linear function using the standard Charnes-Cooper transformation (Zionts, 1968) and thus can be computed by the means of a linear program. It is worth pointing out that, being linear programs, the previous two examples can be efficiently computed.

When f is a convex function of α, we can leverage standard results in sample average approximation programs (Shapiro et al., 2021) to show asymptotic normality in our setting. Let λ_j be the dual variable associated with constraint j, and λ_j^* be the optimal value of λ_j in the population problem. Similarly, let α^* be the population-optimal value of θ_α. Then, we have

Proposition 3.1. *Let $f(\theta_\alpha\hat{\mathbb{Q}}_N)$ be convex in α. Moreover, assume $\alpha \in S$ for S compact and that the Slater's condition holds. Then, $\sqrt{N}(\nu - \hat{\nu}_N) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. In particular, if $f(\theta_\alpha\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[h(x)\theta_\alpha]$, then $\sigma^2 = Var[\theta_{\alpha^*}(X)h(X) + \sum_{j=1}^{M}\lambda_j^*(\theta_{\alpha^*}(X)g_j(X) - \mathbb{E}_{\mathbb{P}}[g_j(X)])]$.*

We can then produce confidence intervals by estimating this variance via the sample estimates $\hat{\theta}_N$ and $\hat{\lambda}_{N,j}$ (see e.g. (Shapiro et al., 2021) Eqs. 5.183 and 5.172). In order to simplify estimation of this variance, it may be recommended to use a form of sample splitting (at the expense of a slower convergence rate), where disjoint sets of samples are used to estimate the objective function and each constraint in the sample problem. This eliminates the covariance terms in the expression for σ above. Alternatively, if computational power is not a constraint, it may be simpler in practice to estimate the variance using the bootstrap.

3.2. General estimands

When the estimand $f(\theta_\alpha \mathbb{Q})$ is a non-convex function of α, obtaining provably optimal solutions for the plug-in approximation problem is in general not feasible. However, we can still obtain locally optimal solutions (as is common for other partial identification settings (Hu et al., 2021; Balazadeh Meresht et al., 2022; Horowitz & Lee, 2022; Padh et al., 2023)), although the statistical properties of the plug-in estimator $f(\theta_\alpha \hat{\mathbb{Q}}_N)$ may become more complex, as we illustrate in the two following examples.

Example 1: Average treatment effect estimation. Consider the setting where $\mathbb{P}$ is a distribution over tuples (X, Y, A), where X is a covariate vector, Y is an outcome, and A is a binary treatment indicator variable. For simplicity, we consider the case where the outcome Y is also binary. Researchers are often interested in estimating the average treatment effect. Under standard identifying assumptions (most prominently that $A \perp Y^{A=a}|X$), this is done by the means of the estimand:

$$f(\mathbb{P}) = \int_X [p(Y|A=1, X) - p(Y|A=0, X)]\, d\mathbb{P}(X)$$

Now, consider the setting where we observe samples from a distribution different than $\mathbb{P}$, resulting in a density ratio $\theta(X, Y, A)$. Computing the appropriate marginals and conditionals of $\theta(X, Y, A)p(X, Y, A)$, and substituting these for p in the above expression, gives an objective that is non-convex in terms of θ, but which is nonetheless differentiable, enabling gradient-based optimization. However, we will still have to estimate the nuisance functions $p(Y = 1|A = a, X)$ and the statistical properties of the resulting bounds will depend on how these are estimated (Coston et al., 2022).

Example 2: Coefficients of parametric models. Suppose that a researcher is interested in interpreting the estimated coefficient of a parametric model, e.g., linear or generalized linear models as commonly used in a variety of applied settings. For example, an applied researcher may estimate the odds ratio for an outcome given some exposure using a logistic model and wishes to obtain bounds on this parametric odds ratio under potential selection bias or other distribution shifts.

Bounding M-estimators: To provide one way of addressing these (and other) examples, we consider the general challenge of partially identifying quantities produced by an M-estimator. M-estimators are those which estimate a parameter β via minimizing the expected value of a function m, i.e., $f(\mathbb{P}) = \arg\min_\beta \mathbb{E}_{X\sim\mathbb{P}}[m(X, \beta)]$, widely used across many areas of statistics and machine learning (Van der Vaart, 2000). One prominent example is when m is the negative log likelihood, resulting in a maximum likelihood estimate of β. We are interested in producing bounds for some real-valued function h of β. For example, h may be the value of a single coordinate of β if we are interested in bounding a specific model coefficient that will be interpreted, or h may be the treatment effect functional described above. This results in the optimization problems:

$$\begin{aligned}
&\min_\alpha h(\beta(\theta_\alpha)) \\
&\beta(\theta_\alpha) = \arg\min_\beta \mathbb{E}\left[\theta_\alpha(X) m(X, \beta)\right] \\
&\theta_\alpha(X) \geq 0 \\
&\mathbb{E}_{X\sim\mathbb{Q}}[\theta_\alpha g_j(X)] = \mathbb{E}_{\mathbb{P}}[g_j(X)] \quad j = 1, \ldots, m
\end{aligned}$$

$$\begin{aligned}
&\min_\alpha h(\hat{\beta}_N(\theta_\alpha)) \\
&\hat{\beta}_N(\theta) = \arg\min_\beta \frac{1}{N}\sum_{i=1}^N \theta_\alpha(X_i) m(X_i, \beta) \\
&\theta_\alpha(X_i) \geq 0 \quad i = 1, \ldots, N \\
&\frac{1}{N}\sum_{i=1}^N \theta_\alpha(X_i) g_j(X_i) = \mathbb{E}_{\mathbb{P}}[g_j(X)] \quad j = 1, \ldots, m.
\end{aligned}$$

Since $h(\beta(\theta_\alpha))$ is not convex in general, Slater's condition is not a strong enough regularity condition to guarantee asymptotic normality of the estimated optimal value. Therefore, we require to impose a more general constraint qualification. Specifically, we will assume the standard *Mangasarian-Fromovitz Constraint Qualification* (Kyparisis, 1985) for the non-convex case. This regularity condition is sufficient to ensure that each minimizer of the population problem has a unique λ_j^* associated in the dual. Hence, under the MFCQ regularity condition and if Σ^* is the covariance matrix of the M-estimator at the optimal α we prove the following asymptotic normality result:

Proposition 3.2. *Assume that m is twice differentiable and locally convex in β. Assume as well that $||\nabla_\alpha \theta_\alpha(x)||_2 \leq \kappa_\theta(x)$, $||\nabla_\beta m(x, \beta)||_2 \leq \kappa_m(x)$, $||\nabla_\beta h(\beta)||_2 \leq \kappa_h$. Then, if α lies in a compact set S, $h(\alpha)$ and $g_i(x)$ are differentiable, then $\sqrt{N}(\hat{\nu}_N - \nu) \to N(0, \sigma^2)$. Using sample splitting, the variance can be approximated by $\nabla h(\beta^*(\alpha^*))^T \Sigma^* \nabla h(\beta^*(\alpha^*)) + Var(\sum_{i=j}^m \lambda_j^* \theta_{\alpha^*}(x) g_j(x))$.*

The proof of Proposition 3.2 requires tools from empirical process theory in order to study the asymptotic properties of the optimal value of the plug-in approximation problem. To sketch the main idea, let Y_n be the vector of functions for the plug-in approximation problem and μ be the one for the population problem (where the first coordinate in each is the objective function and the rest are the constraints). Let ψ be the function $\psi(\mu) = \min_{\theta_\alpha \in \Theta} f(\theta_\alpha \mathbb{Q})$. If the empirical process $\sqrt{n}(Y_n - \mu)$ converges in distribution to an object, then a functional version of the delta method can be applied

to obtain our result. This strategy requires two conditions. First, that a limiting object exists (in distribution) for the random vector $\sqrt{n}(Y_n - \mu)$, which we accomplish by showing that the empirical process belongs to a Donsker class. Second, that such a limiting object, with specified variance, still exists after applying the function ψ. We show this, subject to an appropriate constraint qualification, by applying a generalized version of the continuous mapping theorem (as in (Shapiro, 1991)). The whole proof and explicit expressions for σ^2 can be found in the Appendix.

3.3. Computational approach

As long as $f(\theta_\alpha)$ has well defined subgradients, the plug-in approximation problem can be solved efficiently by projected gradient descent. Even for programs that include an M-estimator, this is easily accomplished in autodifferentiation frameworks where we can use differentiable optimization (Amos & Kolter, 2017; Agrawal et al., 2019) or meta-learning style methods (Finn et al., 2017; Bertinetto et al., 2019; Lee et al., 2019) to implement the $\arg\min$ defining β in a manner which supports automatic backpropagation. In general, we can only obtain locally (instead of globally) optimal solutions for non-convex problems. However, we observe experimentally that the values obtained are nearly identical across many random restarts of the optimization, suggesting good empirical performance.

4. Experiments

We conduct experiments to show how our method allows users to specify domain knowledge in order to obtain informative bounds on the estimand of interest. We simulate inference for a range of different $f(\mathbb{P})$ across various scenarios by testing our method with different choices of constraints and datasets. In addition, we present a real-world use case.

In each experiment, we start with samples from a ground truth distribution $\mathbb{P}$ and then simulate the observed distribution $\mathbb{Q}$ using sampling probabilities which depend on the covariates (i.e., simulating selection bias). This ensures that the ground-truth value $f(\mathbb{P})$ is known (to high precision), allowing us to verify if our bounds contain the true value. We consider two classes of estimands. First, estimating the conditional mean $\mathbb{E}_{\mathbb{P}}[Y|A = 1]$ for an outcome Y and covariate A. Second, estimating the coefficient of a linear regression model as an example of the m-estimation setting from above. To further investigate the stability of our estimand, every experiment was run across 5 different random restarts of the projected gradient descent algorithm.

We show how the amount of specified domain knowledge can result in tighter bounds along two axes. First, we vary the parametric form for θ, ranging from an arbitrary function of X (i.e., a separate parameter for each discrete covariate stratum) to one that imposes specific assumptions (e.g., separability across specific groups of covariates, or that the covariates which the selection probabilities depend on are known). Second, we vary the number and informativeness of the constraints $\{g_j\}$ used to form the set Θ; modeling the ability of users to impose an increasing degree of constraints on possible distribution shifts. The Appendix shows experiments with additional kinds of constraints, e.g., on the sign of the covariance between pairs of variables.

The closest baselines to compare our methods against are sensitivity analyses based on distributionally robust optimization (DRO). However, DRO requires the distance between distributions under some divergence (the unobservable parameter ρ) as input. Conversely, our method requires observable aggregate data to provide partial identification, making a head-to-head comparison difficult. In order to provide baselines for comparison, we implement DRO on two artificial settings where the maximum distance ρ is provided. The first baseline (*omniscient*) optimistically assumes that the χ^2 divergence (ρ) between $\hat{\mathbb{Q}}_n$ and $\mathbb{P}$ is known. This gives an "unfair advantage" to DRO because not only is a valid value of ρ provided but, furthermore, it is the smallest of such feasible values (which is not actually observable when $\mathbb{P}$ is unknown). Hence, we also propose a second scenario (*observable*) where the tightest possible value of ρ is inferred from the same data used by our method. In this scenario ρ is estimated by solving for $\max_{\theta\in\Theta} D_{\chi^2}(\theta\hat{\mathbb{Q}}_n, \hat{\mathbb{Q}}_n)$. Thus, the solution of the program will be the radius of the smallest ball centered at $\hat{\mathbb{Q}}_n$ that contains all distributions that satisfy all the observable constraints. Of course the previous program is not convex in θ; hence we settle for the solution obtained from a saddle point obtained via gradient descent (which errs in favor of DRO by selecting a smaller value of ρ than the one that may be justified by the data). We now proceed to summarize how these experiments were instantiated in each setting, providing detailed information in the Appendix.

Synthetic data experiments. For the first set of experiments, we simulate a distribution over binary variables $X = (Y, Y_2, A, X_1, X_2) \sim \mathbb{P}$ used to evaluate previous causal inference methods (Kennedy et al., 2019). We add a selection bias scenario to this process by simulating an indicator variable $R \sim Ber(logit^{-1}(X_1 - X_2))$. The observed distribution $\mathbb{Q}$ consists of those samples for which $R = 1$. In this domain, we consider the task of estimating bounds for $\mathbb{E}_{\mathbb{P}}[Y|A = 1]$ setting restrictions on $\mathbb{E}_{\mathbb{Q}}[\theta Y X_2]$ (i.e., $\mathbb{E}_{\mathbb{P}}[Y X_2]$ is known). A total of six experiments were run. In the first three experiments, we vary the functional form for θ while leaving the constraints defining Θ fixed. In the first experiment (*unrestricted*), θ is an arbitrary function of X. In the second experiment (*separable)*, we specify $\theta(X) := \theta_1(A) + \theta_2(X_1, X_2)$ where θ_1 is an arbitrary

function of A and θ_2 is an arbitrary function of X_1 and X_2. Finally, in the third experiment (*targeted*), we fix $\theta(X) = \theta(X_1, X_2)$, i.e., θ is a function only of the variables which determine R, thus simulating the scenario where the variables driving selection bias are known (even if the exact selection probabilities are not).

For experiments four to six, we vary Θ instead by adding more informative constraints in each successive experiment. As our intention is to isolate the effect of adding more constraints, the parametric form of θ is set as flexible as possible, i.e., as an arbitrary function of X (the *unrestricted* experiment above). In experiment four, we constrain two of the four parameters of the joint distribution of Y_2 and X_2 via constraints on $\mathbb{E}_{\mathbb{Q}}[\theta Y_2 X_2]$ and $\mathbb{E}_{\mathbb{Q}}[\theta Y_2(1 - X_2)]$. In experiment five, we add constraints on the remaining components of the joint distribution $\mathbb{E}_{\mathbb{Q}}[\theta(1 - Y_2)X_2]$ and $\mathbb{E}_{\mathbb{Q}}[\theta(1 - Y_2)(1 - X_2)]$. Experiment six, add constraints on the outcome ($\mathbb{E}_{\mathbb{P}}[YX_2]$) as well. Throughout, we will refer to these experimental setups as *(partial) race + income*, *(full) race + income* and *race + income + outcome* respectively (the names were chosen to keep consistency with the semi-synthetic experiments).

Semi-synthetic data experiments. We use the Folktables dataset (Ding et al., 2021) which provides an interface for the data from the US Census American Community Survey from 2014 to 2019. Based on the ACSEmployment task suggested in (Ding et al., 2021), we consider a binary outcome variable Y indicating whether or not a person was employed at the time of the survey. We are interested in predicting the female unemployment rate (e.g., motivated by a policymaker who seeks to track or intervene on gender employment gaps (Albanesi & Şahin, 2018; Mate et al., 2022)). The variables driving the selection bias are citizenship and veteran status, as these variables can contribute to self-selection in a potential social services database.

We first consider estimating $\mathbb{E}_{\mathbb{P}}[Y|A = 1]$ where A indicates the sex of an individual. We conduct six experiments that exactly parallel the construction of the synthetic experiments, with the variable for income level playing the role of Y_2 and race/ethnicity playing the role of X_2 (see Appendix for details). This simulates a setting where information about other outcomes correlated with the variables of interest is leveraged for partial identification. In the last experiment *race + income + outcome*, the additional constrain $\mathbb{E}_{\mathbb{Q}}[\theta Y X_2]$ simulates a scenario where information on the outcome of interest is available with respect to a different demographic grouping than our desired estimand (race/ethnicity instead of gender).

We then consider estimating the coefficient of the indicator variable $A = 1$ in a linear regression model of Y on 7 covariates. This second setting is an example of the M-estimator framework from above. We run the experiments *(partial) race + income*, *(full) race + income* and *race + income + outcome* using the settings of the $E_{\mathbb{P}}[Y|A = 1]$ case. Details for each experiment are shown in the Appendix.

Real world case study. As our running example already exposed, evidence based policy formulation is permeated by the selection bias inherent from healthcare data. In particular, addressing this bias when predicting COVID-19 related outcomes becomes of paramount importance (Joynt Maddox et al., 2022; Jacobson et al., 2021; Saatci et al., 2021) to produce reliable estimands required on the decision-making process done by policy officials.

We use real-world data[1][2] to study disparities in hospitalization risk for US COVID-19 patients across race/ethnicity groups. However, instead of directly estimating the racial disparities among outcomes, we applied our method to provide bounds for the *relative excess risk of hospitalization* for non-white patients after conditioning on clinically relevant covariates (a set of five comorbidities and age). Formally, our estimand is $f(\mathbb{P}) = \sum_x [P(Y = 1|X = x, Race) - P(Y = 1|X = x, Race = White)]P(X = x|Race)/P(Y = 1|Race = White, X)$, where Y is whether a patient was hospitalized. As mentioned in the introduction, selection bias is particularly challenging for these data, because it may depend jointly on both the outcome and main covariate of interest. We model selection based on race, income, and hospitalization status to account for this potential interaction between disease severity and healthcare access in ascertainment. As constraints g, we use the total hospitalizations per racial group reported by the CDC and the total number of estimated infections per group from CDC serology studies.

Results. Figures 1 and 2 show our main results. Each figure plots the bounds output by our method for each experiment described above. In each plot, we also show the true value of the estimand $f(\mathbb{P})$, the naive estimate using samples from $\mathbb{Q}$ (without any attempt to account for selection bias) and the bounds produced by the two DRO baselines. In each plot (1 and 2), accounting for the different results across multiple bootstrap replicates, there is a box plot at the endpoints of every bound as well as a band to represent $f(\mathbb{Q})$.

The results of our experiments are consistent across all six scenarios, showing how the estimated bounds vary depending on the constraints imposed and on the functional form assumed for θ. Concordant with the theory, the bounds always contain the true value of the estimand. Additionally, as expected, when more external information is available the bounds become narrower. When additional constraints are added, it is possible to obtain narrow bounds for the esti-

[1]We use data from five million de-identified medical claims from COVID-19 patients.

[2]Medical claims provide the services that patients received, their diagnosis, and demographic attributes.

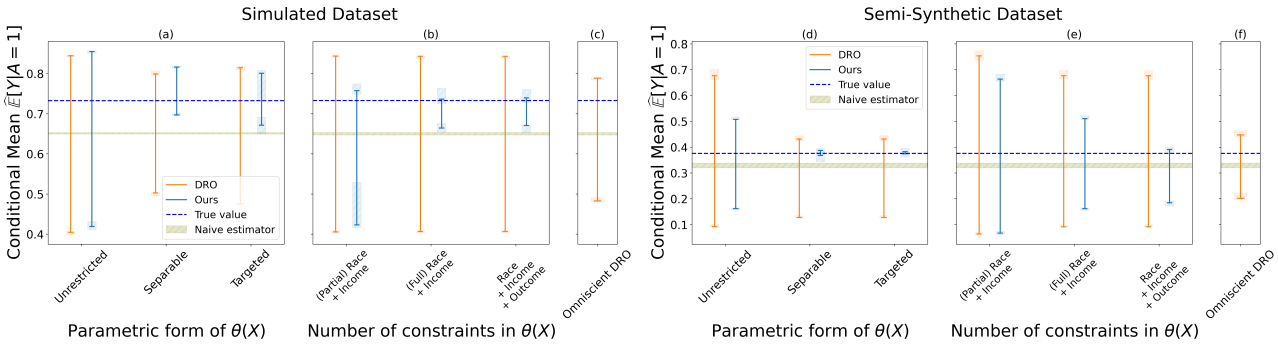

Figure 1. Partial identification of a conditional mean in simulated and semi-synthetic experiments using our proposed method vs. *observed* DRO. The experiments include several parameterizations of $\theta(X)$ and different sets of constraints Θ.

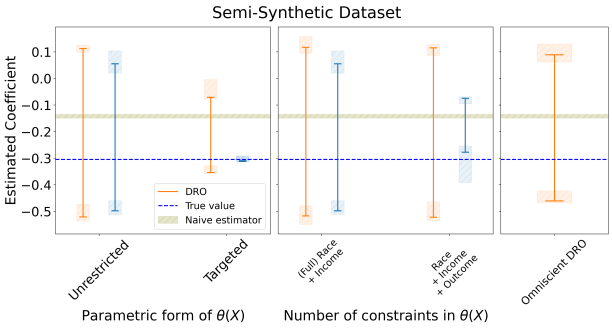

Figure 2. Estimated value of $\beta(\theta)$ with our proposed method for the semi-synthetic dataset. The first two images are results for different parameterizations of $\theta(X)$ and results for different sets of constraints Θ, respectively. In both cases, incorporating more external information leads to narrower bounds.

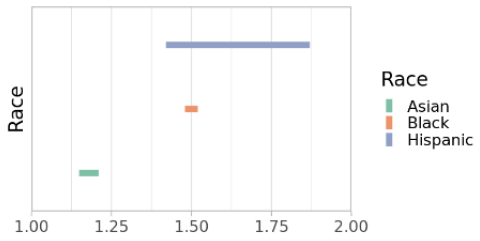

Figure 3. Estimated relative risk for hospitalization compared to White patients. Our method certifies that there is an increased risk of hospitalization due to COVID for Asian, Black, and Hispanic populations when compared to the White population. For the Black and Hispanic population, our method of partial identification is close to a point identification.

mand that rule out the naive estimates. Our method Pareto-improves over the observed-data DRO baseline across all settings. Interestingly, in the majority of settings, our method also produces tighter intervals than the *omniscient* DRO baseline. Thus, it can be highlighted the advantage of more nuanced descriptions of uncertainty: even in the case the true value of the divergence is known, moment constraints may provide a tighter bound than a divergence metric. We also find that in the *targeted* experiment where the variables driving selection bias are known, our method often provides close to point identification, indicating that this is a particularly powerful form of domain knowledge when available. These results demonstrate that our framework allows users to translate common forms of domain knowledge into robust and informative statistical inferences.

COVID-19 application results. Figure 3 summarizes the results from the real-world case study. Since the true underlying target distribution is unknown for real data, the question is whether our method produces informative conclusions (i.e., meaningful bounds), as opposed to comparison to a known 'ground truth'. We observe that the bounds produced by our method are highly informative and support several policy-relevant conclusions. First, the relative risk of hospitalization is > 1 for every race/ethnicity group, allowing us to certify that Black, Asian and Hispanic patients have a higher risk of hospitalization compared to White patients. That is, our method is able to rule out the concern that results are overly distorted by selection bias based on healthcare access for mild cases. Second, the bounds are **close to point identification** for Black and Asian patients while highlighting greater sensitivity to potential selection biases for Hispanic patients. This provides actionable feedback for policy makers, for example to focus data collection efforts on high-quality studies in low-income Hispanic populations if more precise estimates are needed. Meanwhile, policy makers can be confident in precise estimates of the value of targeting primary care interventions for Black patients (e.g., improving access to antivirals to prevent hospitaliza-

tion). This showcases that our methods produce informative and theoretically-guaranteed intervals for a policy-relevant question.

Impact Statement

We anticipate that our methods will have a positive social impact via increasing the robustness of statistical and machine learning methods to distribution shifts such as selection bias. This is particularly consequential for studying the equity impacts of policies and systems (as marginalized populations are typically less likely to be represented in the dataset). However, our methods do come with important limitations, in that the bounds will only be as good as the domain knowledge used to specify the constraints. In particular, it is up to the user to specify the correct target distribution, select informative constraint functions, and ensure that these functions can be accurately estimated on the target distribution (which may be difficult if even the nominal target distribution is subject to measurement errors or other complications). Negative consequences are possible if incorrect domain knowledge is used as the input, resulting in an incorrect inference which is improperly certified as "robust". In short, the framework we propose provides a means to translate domain expertise about potential mechanisms for data biases into rigorous inferences. Our hope is that it will serve to enable collaborations between domain experts and quantitative researchers towards this goal.

Acknowledgements

This work was supported in part by the AI2050 program at Schmidt Futures (Grant G-22-64474).

References

Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., and Kolter, J. Z. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Albanesi, S. and Şahin, A. The gender unemployment gap. *Review of Economic Dynamics*, 30:47–67, 2018.

Amos, B. and Kolter, J. Z. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pp. 136–145. PMLR, 2017.

Balazadeh Meresht, V., Syrgkanis, V., and Krishnan, R. G. Partial identification of treatment effects with implicit generative models. *Advances in Neural Information Processing Systems*, 35:22816–22829, 2022.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Bertail, P., Clémençon, S., Guyonvarch, Y., and Noiry, N. Learning from biased data: A semi-parametric approach. In *International Conference on Machine Learning*, pp. 803–812. PMLR, 2021.

Bertinetto, L., Henriques, J. F., Torr, P., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Bertsimas, D., Imai, K., and Li, M. L. Distributionally robust causal inference with observational data. *arXiv preprint arXiv:2210.08326*, 2022.

Billingsley, P. *Convergence of probability measures*. John Wiley & Sons, 2013.

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pp. 872–881. PMLR, 2019.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148. PMLR, 2018.

Coston, A., Dulce Rubio, M., and Kennedy, E. H. *AI for Social Impact*, chapter Randomized experiments. 2022. https://ai4sibook.org/.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Ding, P. and VanderWeele, T. J. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3): 368, 2016.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

Guo, W., Yin, M., Wang, Y., and Jordan, M. Partial identification with noisy covariates: A robust optimization approach. In *Conference on Causal Learning and Reasoning*, pp. 318–335. PMLR, 2022.

Gupta, S. and Rothenhäusler, D. The s-value: evaluating stability with respect to distributional shifts. *arXiv preprint arXiv:2105.03067*, 2021.

Haneuse, S. and Daniels, M. A general framework for considering selection bias in ehr-based studies: what data are observed and why? *EGEMs*, 4(1), 2016.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. Challenges in administrative data linkage for research. *Big data & society*, 4 (2):2053951717745678, 2017.

Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., et al. Seroprevalence of antibodies to sars-cov-2 in 10 sites in the united states, march 23-may 12, 2020. *JAMA internal medicine*, 180(12):1576–1586, 2020.

Ho, K. and Rosen, A. M. Partial identification in applied research: benefits and challenges. Technical report, National Bureau of Economic Research, 2015.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Horowitz, J. L. and Lee, S. Inference in a class of optimization problems: Confidence regions and finite sample bounds on errors in coverage probabilities. *Journal of Business & Economic Statistics*, pp. 1–12, 2022.

Hu, Y., Wu, Y., Zhang, L., and Wu, X. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12104–12112, 2021.

Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Jacobson, M., Chang, T. Y., Shah, M., Pramanik, R., and Shah, S. B. Racial and ethnic disparities in sars-cov-2 testing and covid-19 outcomes in a medicaid managed care cohort. *American Journal of Preventive Medicine*, 61(5):644–651, 2021.

Jensen, E. T., Cook, S. F., Allen, J. K., Logie, J., Brookhart, M. A., Kappelman, M. D., and Dellon, E. S. Enrollment factors and bias of disease prevalence estimates in administrative claims data. *Annals of epidemiology*, 25(7): 519–525, 2015.

Johnson, C. L., Dohrmann, S. M., Burt, V. L., and Mohadjer, L. K. *National health and nutrition examination survey: sample design, 2011-2014*. Number 2014. US Department of Health and Human Services, Centers for Disease Control and . . . , 2014.

Joynt Maddox, K. E., Reidhead, M., Grotzinger, J., McBride, T., Mody, A., Nagasako, E., Ross, W., Steensma, J. T., and Barker, A. R. Understanding contributors to racial and ethnic inequities in covid-19 incidence and mortality rates. *Plos one*, 17(1):e0260262, 2022.

Kennedy, E. H., Kangovi, S., and Mitra, N. Estimating scaled treatment effects with multiple outcomes. *Statistical methods in medical research*, 28(4):1094–1104, 2019.

Knox, D., Lowe, W., and Mummolo, J. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020.

Kyparisis, J. On uniqueness of kuhn-tucker multipliers in nonlinear programming. *Mathematical Programming*, 32: 242–246, 1985.

Lee, K., Maji, S., Ravichandran, A., and Soatto, S. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10657–10665, 2019.

Manski, C. F. *Partial identification of probability distributions*, volume 5. Springer, 2003.

Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., and Tambe, M. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12017–12025, 2022.

Molinari, F. Microeconometrics with partial identification. *Handbook of econometrics*, 7:355–486, 2020.

Murray, C. J. and Lopez, A. D. Measuring the global burden of disease. *New England Journal of Medicine*, 369(5): 448–457, 2013.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., and Kilbertus, N. Stochastic causal programming for bounding treatment effects. *Conference on Causal Learning and Reasoning*, 2023.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS*, 116:1–94, 2000.

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

Rothenhäusler, D. and Bühlmann, P. Distributionally robust and generalizable inference. *arXiv preprint arXiv:2209.09352*, 2022.

Saatci, D., Ranger, T. A., Garriga, C., Clift, A. K., Zaccardi, F., San Tan, P., Patone, M., Coupland, C., Harnden, A., Griffin, S. J., et al. Association between race and covid-19 outcomes among 2.6 million children in england. *JAMA pediatrics*, 175(9):928–938, 2021.

Shapiro, A. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30:169–186, 1991.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Tamer, E. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.

Tan, Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Tudball, M., Hughes, R., Tilling, K., Bowden, J., and Zhao, Q. Sample-constrained partial identification with application to selection bias. *arXiv preprint arXiv:1906.10159*, 2019.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615, 2022.

Zionts, S. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 15(3):449–451, 1968.

A. Proofs

A.1. Proof proposition 2.1

This holds by construction since $\mathbb{P} = \theta\mathbb{Q}$ for some $\theta \in \Theta$. To see that these bounds are tight, suppose *wlog* that we have a conjectured lower bound $\ell > \min_{\theta\in\Theta} f(\theta\mathbb{Q})$. Then, there would be some θ consistent with all the constraints in Θ for which $f(\theta\mathbb{Q}) < \ell$, implying that there is a target distribution $\mathbb{P}$ consistent with all constraints for which ℓ is not a valid lower bound.

A.2. Proof proposition 3.1

This follows directly from Theorem 5.11 of (Shapiro et al., 2021), since the objective and constraints are convex and we have assumed that the population problem has a unique solution.

A.3. Proof proposition 3.2

We want to show that the empirical process $\sqrt{N}(\hat{\nu}_N - \nu)$ converges to a Gaussian process when $N \to \infty$, where $\widehat{\nu}_N$ is the value of the plug-in approximation program estimated from a sample of size N. To prove this, we will use the delta method in Theorem 2.1 of (Shapiro, 1991) following the Hadamard differentiability of the outer problem $\min_\alpha f(\alpha, \xi)$. The later condition is shown to hold by invoking Theorem 3.6 in (Shapiro, 1991). However, to apply such theorem, the *Mangasarian-Fromovitz Constraint Qualification* has to hold, and the approximation vector has to have a limiting process in $C(S)$; the first requirement is satisfied by assumption, and the second is shown in the following.

Assumption 1. $\theta(x) = \theta_\alpha(x)$ has some finite dimensional parameterization $\alpha \in S \subset R^d$.

Consider the function $f(\alpha, \xi) = h(\beta^*(\theta_\alpha, \xi))$ where

$$\beta^*(\theta_\alpha, \xi) = \arg\min_\beta \mathbb{E}_\xi[\theta_\alpha(x_i)m(\beta, x_i)].$$

and ξ is the distribution that the expectation is taken with respect to, e.g., the realized empirical distribution on a set of samples $x_1, \ldots, x_N$ or the limiting distribution $\mathbb{Q}$.

Assumption 2. The functions f and g_i are differentiable on S.

Assumption 3. β^* is always a strict local minimizer for any θ and realization ξ so that the Hessian matrix $H(\beta) = \nabla^2_\beta\left[\mathbb{E}_\xi[\theta_\alpha(x_i)m(\beta, x_i)]\right]$ is positive definite. In particular, we assume that the smallest eigenvalue satisfies $\lambda_{\text{min}}(H(\beta)) > \gamma_\xi$ for some $\gamma_\xi > 0$ which may depend on the realization of the randomness ξ, but not on θ.

Step 1. Our goal will be to show that f is Lipschitz as a function of α for any fixed ξ, which we will accomplish by showing that $||\nabla_\alpha f||$ is bounded. We use $||\cdot||_2$ for matrices to denote the spectral norm, or operator 2-norm, and for vectors to denote the Euclidean norm. We assume that $||\nabla_\alpha\theta_\alpha(x)||_2$, $||\nabla_\beta m(x, \beta)||_2$, and $||\nabla_\beta h(\beta^*)||_2$ all satisfy bounds which in the case of the first two are allowed to depend on x:

$$||\nabla_\alpha\theta_\alpha(x)||_2 \leq \kappa_\theta(x), \quad ||\nabla_\beta m(x, \beta)||_2 \leq \kappa_m(x), \quad ||\nabla_\beta h(\beta)||_2 \leq \kappa_h.$$

That is, they are each Lipschitz in their respective parameter for a fixed realization x, but with Lipschitz constant which is allowed to depend on x. While we present these assumptions in terms of the Euclidean norm for convenience, since we only aim to prove that the Lipschitz constant of f is bounded (our results do not depend on its particular value), bounds on the gradients above in any norm suffice (using the equivalence of all norms in finite dimension).

In what follows we fix ξ and drop it from the notation. We first apply the chain rule to obtain that $\nabla_\alpha f = [D_\alpha\beta^*(\alpha)]\nabla_\beta h(\beta^*(\alpha))$ where $D_\alpha\beta^*(\alpha)$ is the Jacobian of β^* with respect to α. To calculate this term, we apply the implicit function theorem to the first-order optimality condition

$$\nabla_\beta\mathbb{E}[\theta_\alpha(x)m(\beta, x)] = 0$$

and obtain

$$D_\alpha\beta^*(\alpha) = H(\beta^*)^{-1}\nabla_\alpha\nabla_\beta\mathbb{E}[\theta_\alpha(x)m(\beta^*, x)]$$

where we remark that H is guaranteed to be invertible by our assumption that it is positive definite in a neighborhood of β^* and hence the implicit function theorem applies.

Next, note that

$$\nabla_\alpha \nabla_\beta \mathbb{E}[\theta_\alpha(x) m(\beta^*, x)] = \mathbb{E}[\nabla_\alpha \theta_\alpha(x) [\nabla_\beta m(x, \beta)]^T]$$

and

$$||\mathbb{E}[\nabla_\alpha \theta_\alpha(x) [\nabla_\beta m(x, \beta)]^T]||_2 \leq \mathbb{E}[||\nabla_\alpha \theta_\alpha(x) [\nabla_\beta m(x, \beta)]^T||_2]$$

by Jensen's inequality. Using the Lipschitz assumptions on θ and m combined with the definition of the spectral norm we have that the outer product satisfies

$$||\nabla_\alpha \theta_\alpha(x) [\nabla_\beta m(x, \beta)]^T||_2 \leq \kappa_\theta(x) \kappa_m(x)$$

and hence

$$||\mathbb{E}[\nabla_\alpha \theta_\alpha(x) [\nabla_\beta m(x, \beta)]^T]||_2 \leq \mathbb{E}[\kappa_\theta(x) \kappa_m(x)].$$

Since H has bounded minimum eigenvalue, H^{-1} has bounded spectral norm as well: $||H^{-1}||_2 \leq \frac{1}{\gamma}$. Finally, using the Lipschitz assumption on h and again applying the sub-multiplicative property of the spectral norm, we have

$$||\nabla_\alpha f||_2 \leq \frac{1}{\gamma} \mathbb{E}[\kappa_\theta(x) \kappa_m(x)] \kappa_h.$$

Stepping back to re-introduce the random function ξ, we have that the Lipschitz constant of f with respect to α is $\frac{1}{\gamma_\xi} \mathbb{E}_\xi[\kappa_\theta(x) \kappa_m(x)] \kappa_h$.

Step 2. Now we are going to proceed to prove that the restrictions are also Lipchitz. In order to do that, we are going to proceed using the same strategy as before and show that the restrictions also have bounded gradients. Let us denote the restrictions as $r_i(\alpha)$. Note that

$$\nabla_\alpha r_i = \frac{1}{n} g_i(X) \nabla_\alpha \theta(X)$$

Thus it must be true that:

$$||\nabla_\alpha r_i|| = \frac{1}{n} |g_i(X)| ||\nabla_\alpha \theta(X)||$$

As g is continuous in a compact set it must be bounded by a finite number M. Then by hypothesis:

$$||\nabla_\alpha r_i|| \leq k_\theta(x) M$$

And again by Jensen's inequality:

$$||\nabla_\alpha r_i|| \leq \mathbb{E}[k_\theta(x)] M$$

Step 3. Since f and all r_i are Lipschitz on a finite-dimensional parameter, this implies that they belong to a Donsker class (Van der Vaart, 2000) and so we are guaranteed to a limiting stochastic process for each r_i and for f marginally. Furthermore, as $C(S)$ is separable, by theorem 3.2 of (Billingsley, 2013), the vector $(f, r_1, .., r_n)$ has a limiting process under the product measure.

Step 4. Thus, as $\alpha \in S$ and because of the constraint qualification assumed for the problem, we may apply Theorem 3.6 of (Shapiro, 1991) to guarantee Hadamard differentiability of the outer problem $\min_\alpha f(\alpha, \xi)$. Hence, the delta method in Theorem 2.1 of (Shapiro, 1991) can be applied to guarantee asymptotic normality and thus the result follows.

Step 5 In order to explicitly compute the variance of $\sqrt{N}(\hat{\nu}_N - \nu)$, let

$$Y_n = \begin{bmatrix} h(\hat{\beta}_N(\alpha)) \\ \frac{1}{n} \sum_i \theta_\alpha(X_i) g_1(X_i) \\ \vdots \\ \frac{1}{n} \sum_i \theta_\alpha(X_i) g_n(X_i) \end{bmatrix}$$

and

$$\mu = \begin{bmatrix} h(\beta^*(\alpha)) \\ \mathbb{E}[\theta_\alpha(X)g_1(X)] \\ \vdots \\ \mathbb{E}[\theta_\alpha(X)g_m(X)]. \end{bmatrix}$$

Then, if $g(f) = \min_{\alpha\in\Theta} f(\alpha)$, again by theorem 2.1 (Shapiro, 1991)

$$\sqrt{n}(g(Y_n) - g(\mu)) \to_D g'_\mu(Z)$$

Where Z is the limiting object that we proved the empirical process has. As a consequence of theorem 3.6 of (Shapiro, 1991), if the set of minimizers is a singleton, $g'_\mu(\sqrt{n}(Y_n - \mu)) = \sqrt{n}[h(\hat{\beta}_n(\alpha_0)) - h(\beta^*(\alpha_0))] + \sqrt{n}\sum_{j=1}^m \lambda_j[\frac{1}{n}\sum_i \theta_{\alpha_0}(X_i)g_j(X_i) - \mathbb{E}_{\mathbb{Q}}[\theta_{\alpha_0}(X)g_j(x)]]$. The term $[h(\beta^*(\alpha_0)) - h(\hat{\beta}_n(\alpha_0))]$ is normal as it is a M-estimator itself. Hence $g'_\mu(\sqrt{n}(Y_n - \mu))$ is the sum of Normally distributed random variables and therefore,

$$\sqrt{n}(\nu_n - \nu) \to N\left(0, Var\left(\sqrt{n}\left(h(\beta^*(\alpha_0)) - h(\hat{\beta}_n(\alpha_0)) - \sum_{j=1}^m \lambda_j \frac{1}{n}\left[\sum_{i=1}^n \theta_{\alpha_0}(X_i)g_j(X_i)\right]\right)\right)\right).$$

Using sample splitting and the delta method on $h(\beta^*)$, the variance can be approximated by $\nabla h(\beta^*(\alpha_0))^T\Sigma_0\nabla h(\beta^*(\alpha_0)) + Var(\sum_{i=j}^m \lambda_j^*\theta_{\alpha_0}(x)g_j(x))$, completing the proof.

B. Experiments

B.1. Constraint derivation

In some cases, we may be able to replace the constrain $\theta(X) \geq 0$ with a tighter constraint. For example, consider the case of *selection bias*, where individual samples from $\mathbb{P}$ select into $\mathbb{Q}$ based on X. Formally, we model this via an indicator variable R, which is 1 if a unit is observed in the sample and 0 otherwise. Then, $q(x) = p(x|R = 1)$, and via Bayes theorem we have $\theta_0(X) = \frac{\Pr(R=1)}{\Pr(R=1|X)}$. Since $\Pr(R = 1|X) \leq 1$, we are guaranteed that $\theta_0(X) \geq \Pr(R = 1)$. In many cases, the marginal $\Pr(R = 1)$ is easily observable because we know the total size of the population that appears in our sample relative to the true population (e.g., a government may know the fraction of people in a city who are enrolled in a program). This allows us to replace the constraint $\theta(X) \geq 0$ with the tighter constraint $\theta(X) \geq \Pr(R = 1)$. This was indeed the case in all the experiments presented in the article.

B.2. Synthetic data experiments details

Inspired by data models used in the causal inference literature (Kennedy et al., 2019), the distribution $X = (Y, Y_2, A, X_1, X_2) \sim \mathbb{P}$ is given by the following model:

$$\begin{aligned} X_1 &\sim Multinomial(3, 0.5, 0.3) \\ X_2 &\sim Ber(0.4) \\ A &\sim Ber(logit^{-1}(X_2 - X_1)) \\ Y &\sim Ber(logit^{-1}(2A - X_1 + X_2)) \\ Y_2 &\sim Ber(logit^{-1}((X_1 + X_2)/2 - A) \end{aligned}$$

The observed distribution $\mathbb{Q}$ is given by simulating selection bias via an indicator variable:

$$R \sim Ber(logit^{-1}(X_1 - X_2))$$

Naturally, the sample from $\mathbb{Q}$ are all those samples for which $R = 1$. The set Θ used in the experiments was:

- **Unrestricted**, **separable** and **Targeted** : $S := \{\mathbb{E}_{X\sim\mathbb{P}}[Y_2X_2] = c_1, \mathbb{E}_{X\sim\mathbb{P}}[X_2(1 - Y_2)] = c_2, \mathbb{E}_{X\sim\mathbb{P}}[(1 - X_2)Y_2] = c_3, \mathbb{E}_{X\sim\mathbb{P}}[(1 - X_2)(1 - Y_2)] = c_4\}$.

- **(partial) race + income :**, Θ was $S \setminus \{\mathbb{E}_{X\sim\mathbb{P}}[Y_2X_2] = c_1, \mathbb{E}_{X\sim\mathbb{P}}[(X_2-1)Y_2] = c_2\}$.
- **(full) race + income :** Θ was S.
- **race + income + Outcome :** Θ was $S \cup \{\mathbb{E}_{X\sim\mathbb{P}}[YX_2] = c_5, \mathbb{E}_{X\sim\mathbb{P}}[Y(1-X_2)] = c_6\}$.

Each experiment was run 5 times over different bootstrap replicates of the sample of distribution $\mathbb{P}$. The experimental results are summarized in table 1.

Table 1. Synthetic data experimental results. True conditional mean: 0.733.

		Lower bound		Upper bound	
		mean	std	mean	std
Separable	DRO	0.475	0.004	0.814	0.003
	Ours	0.671	0.016	0.786	0.033
Targeted	DRO	0.501	0.004	0.800	0.004
	Ours	0.697	0.003	0.816	0.002
Unrestricted	DRO	0.403	0.004	0.844	0.002
	Ours	0.420	0.008	0.855	0.002
(Full) Race + Income	DRO	0.406	0.002	0.842	0.002
	Ours	0.665	0.009	0.743	0.016
(Partial) Race + Income	DRO	0.405	0.002	0.843	0.002
	Ours	0.449	0.052	0.761	0.011
Race + Income + Outcome	DRO	0.406	0.002	0.842	0.002
	Ours	0.665	0.008	0.744	0.014
DRO Omniscient		0.484	0.004	0.788	0.002

B.3. Semi-synthetic data

For the semi-synthetic experiments we used the Folkstables package (Ding et al., 2021) which provides an interface for curated US Census data. We use the ACSEmployment task, where Y is whether or not a person is employed, and the variable of interest A is the sex of an individual. The rest of the covariates come from a one-hot encode of the features listed below. The last level of each variable is dropped to avoid an unidentifiable model, thus obtaining a 15-dimensional representation for every entry in the dataset.

1. **Citizenship status**: 0: Born in the U.S, 1: Born in Puerto Rico, Guam, the U.S. Virgin Islands, or the Northern Marianas, 2: Born abroad of American parent(s), 3: U.S. citizen by naturalization, 4: Not a citizen of the U.S.
2. **Military service**: 0: Is or was in active duty, 1: Never served in the military.
3. **Nativity**: 0: Native, 1: Foreign-born.
4. **disability**: 0: Not having any disability, 1: Having a Hearing, vision, or cognitive disability.
5. **Income**: 0: Personal income over 50000 USD a year , 1: Personal income below 50000 USD a year.
6. **Race**: 0:self-identifying as not white , 1:self-identifying as white

The observed distribution $\mathbb{Q}$ is given by simulating selection bias via an indicator variable:

$$R \sim Ber(logit^{-1}(X_1 - X_2))$$

Again, the sample from $\mathbb{Q}$ are all those samples for which $R = 1$.

Let Y be unemployment status, A sex, Y_2 income and X_2 be race. The set Θ used in the first setting of experiments (estimating $\mathbb{E}_{\mathbb{P}}[Y|A = 1]$) is:

- **Unrestricted**, **separable** and **Targeted** : $S := \{\mathbb{E}_{X\sim\mathbb{P}}[Y_2X_2] = c_1, \mathbb{E}_{X\sim\mathbb{P}}[X_2(1-Y_2)] = c_2, \mathbb{E}_{X\sim\mathbb{P}}[(1-X_2)Y_2] = c_3, \mathbb{E}_{X\sim\mathbb{P}}[(1-X_2)(1-Y_2)] = c_4\}$.
- **(partial) race + income :** Θ was $S \setminus \{\mathbb{E}_{X\sim\mathbb{P}}[Y_2X_2] = c_1, \mathbb{E}_{X\sim\mathbb{P}}[(X_2-1)Y_2] = c_2\}$.
- **(full) race + income :** Θ was S.
- **race + income + Outcome :** Θ was $S \cup \{\mathbb{E}_{X\sim\mathbb{P}}[YX_2] = c_5, \mathbb{E}_{X\sim\mathbb{P}}[Y(1-X_2)] = c_6\}$.

Each experiment was run 5 times over different bootstrap replicates of the sample of distribution $\mathbb{P}$. The experimental results are summarized in table 2.

Table 2. Semi-synthetic data experimental results. True conditional mean: 0.376.

		Lower bound		Upper bound	
		mean	std	mean	std
Unrestricted	DRO	0.132	0.055	0.602	0.107
	Ours	0.161	0.003	0.509	0.004
Separable	DRO	0.129	0.004	0.433	0.008
	Ours	0.371	0.007	0.383	0.009
Targeted	DRO	0.129	0.004	0.433	0.008
	Ours	0.366	0.015	0.385	0.010
(Partial) Race + Income	ours	0.067	0.004	0.666	0.012
	DRO	0.064	0.004	0.753	0.015
(Full) Race + Income	ours	0.160	0.004	0.513	0.006
	DRO	0.093	0.003	0.678	0.014
Race + Income + Outcome	ours	0.184	0.008	0.389	0.007
	DRO	0.093	0.003	0.677	0.013
DRO Omniscient		0.208	0.012	0.451	0.01

The set Θ used in the second setting of experiments (estimating the coefficient β of the indicator variable A in a linear regression model) is:

- **Unrestricted** and **targeted**: $S := \{\mathbb{E}_{X\sim\mathbb{P}}[Y_2X_2] = c_1, \mathbb{E}_{X\sim\mathbb{P}}[X_2(1-Y_2)] = c_2, \mathbb{E}_{X\sim\mathbb{P}}[(1-X_2)Y_2] = c_3, \mathbb{E}_{X\sim\mathbb{P}}[(1-X_2)(1-Y_2)] = c_4\}$.
- **(Full) Race + income :**, Θ was $S \setminus \{\mathbb{E}_{X\sim\mathbb{P}}[Y_2X_2] = c_1, \mathbb{E}_{X\sim\mathbb{P}}[Y_2(X_2-1)] = c_2\}$.
- **Race + income + Outcome :** Θ was S.

Each experiment was run 5 times over different bootstrap replicates of the sample of distribution $\mathbb{P}$. The experimental results are summarized in table 3.

B.4. Logistic regression model

We run one additional experiment to estimate the coefficient β of the indicator variable A. However, instead of being the coefficient of a linear model, it is now the coefficient of a logistic regression. Only the **unrestricted** experiment was run. The experiment was run 5 times. The data reported is the average value and standard deviation for the 5 outputs obtained from the experiment. The results are summarized in figure 4 and table 4.

B.5. Covariance-like restrictions

We run one additional experiment to estimate the coefficient β of the indicator variable A. However, instead of being the coefficient of a linear model, it is now the coefficient of a logistic regression. The experiment was run 5 times. The data reported is the average value and standard deviation for the 5 outputs obtained from the experiment. The results are in figure

Table 3. Semi-synthetic data experimental results. True conditional mean: -0.305.

		Lower bound		Upper bound	
		mean	std	mean	std
Unrestricted	DRO	-0.513	0.026	0.112	0.011
	Ours	-0.492	0.022	0.062	0.036
Separable	DRO	-0.345	0.013	-0.055	0.033
	Ours	-0.305	0.010	-0.303	0.009
(Full) Race + Income	DRO	-0.492	0.022	0.062	0.036
	Ours	-0.515	0.028	0.123	0.025
Race + Income + Outcome	DRO	-0.509	0.029	0.110	0.017
	Ours	-0.304	0.057	-0.079	0.011

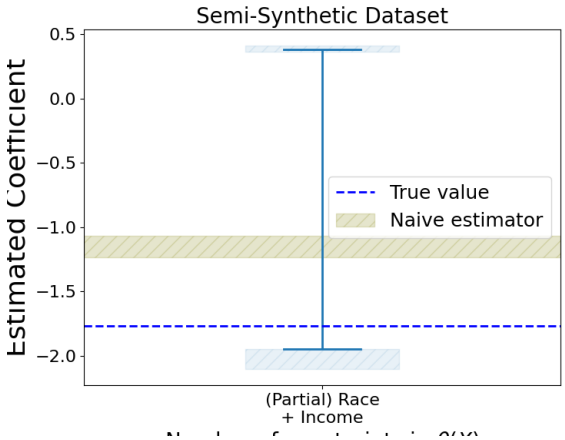

Figure 4. Bounds outputted by our method for the coefficient β of the indicator variable A .

Table 4. Semi-Synthetic data experimental results. True conditional mean: -1.769.

	Lower bound		Upper bound	
	mean	std	mean	std
Unrestricted	-1.998	0.069	0.382	0.019

5. For the results in figure 5, we constrained the covariance for a set of variables to be positive. Specifically, we took the pair of variables with the largest positive covariance—being a US citizen and being a native—and restricted that covariance to positive during the optimization.

Overall, we experimented with two types of restrictions on the covariance: restricting the sign of the covariance or the exact covariance value. The two types of constraints had similar results. We also tested restricting on multiple feature pairs but didn't see any significant differences in the bounds.

B.6. Guide to run the experiments

In the **experiments_metadata** folder are JSON files with the exact hyperparameters used in all the experiments presented in the paper. To execute a particular set of experiments related to the conditional mean, use the `inference.py` script. For example, the following command generates the bounds with different parametric forms for the simulated dataset:

```
python inference.py experiments_metadata/sim_theta_ours.json
```

To recreate the results of the experiments that involve regression models, use the `inference_non_closed_form.py`

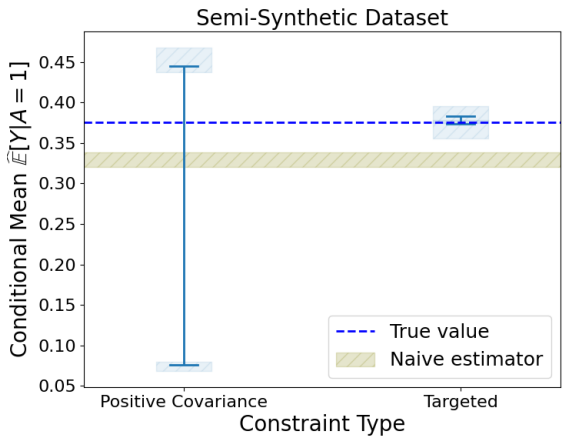

Figure 5. Bounds outputted by our method for the coefficient β of the indicator variable A .

script:

```
python inference_non_closed_form.py experiments_metadata/logistic.json
```

We will provide a detailed table with the JSON files that generate the data of the plots we included.

Each successful run of the two commands will produce a folder containing a data frame with the results and several plots. The folder's name is a timestamp denoting the experiment's end time. To reproduce the plots from the paper, use the notebook in the folder named **plotting**.

The directory hierarchy is the following

```
/ (root)
├── README.md
├── experiment_artifacts
├── optimization
│   ├── experiments_metadata
│   ├── inference.py
│   └── inference_non_closed_form.py
└── plotting
    ├── paper_plots.ipynb
    └── utils.py
```

B.7. Explicit bound Computation

It can be analytically verified that the bounds are non-vacuous for the synthetic experiment called *targeted*. Here, $f(\theta\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[\theta(x_1, x_2)Y|A = 1]$ and $\{\mathbb{E}_{\mathbb{Q}}[\theta(x_1, x_2)YX_2] = \mathbb{E}_{\mathbb{P}}[YX_2],\ \ \mathbb{E}_{\mathbb{Q}}[\theta(x_1, x_2)Y(1 - X_2)] = \mathbb{E}_{\mathbb{P}}[Y(1 - X_2)]\} \subset \Theta$. For simplicity, hereinafter we will denote the observed values $\mathbb{E}_{\mathbb{P}}[YX_2]$ and $\mathbb{E}_{\mathbb{P}}[Y(1 - X_2)]$ as c_1 and c_2, respectively. As all the variables are binary, the trivial bounds on $f(\theta\mathbb{Q})$ are 0 and 1. However this is not the case here. Lets begin by noting that,

$$
\begin{aligned}
f(\theta\mathbb{Q}) &= \mathbb{E}_{\mathbb{Q}}[\theta Y|A = 1] \\
&= \frac{1}{\mathbb{Q}(A = 1)} \sum_{(x_1,x_2,A=1,y)} \theta(x_1, x_2)y\mathbb{Q}(y, x_1, x_2, A = 1) \\
&= \frac{1}{\mathbb{Q}(A = 1)} \sum_{(x_1,x_2,A=1,Y=1)} \theta(x_1, x_2)\mathbb{Q}(y, x_1, x_2, a)
\end{aligned}
$$

Similarly,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\theta(x_1,x_2)YX_2] &= c_1 \\ &= \sum_{(x_1,x_2,a,y)} \theta(x_1,x_2)yx_2\mathbb{Q}(y,x_1,x_2,a) \\ &= \sum_{(x_1,X_2=1,a,Y=1)} \theta(x_1,x_2)\mathbb{Q}(y,x_1,x_2,a)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\theta(x_1,x_2)Y(1-X_2)] &= c_2 \\ &= \sum_{(x_1,x_2,a,y)} \theta(x_1,x_2)y(1-x_2)\mathbb{Q}(y,x_1,x_2,a) \\ &= \sum_{(x_1,X_2=0,a,Y=1)} \theta(x_1,x_2)\mathbb{Q}(y,x_1,x_2,a)\end{aligned}$$

Remember that all variables are binary and all the terms on the sum are non-negative. Additionally we can ensure $\theta(X) \geq P(R=1)$ (see section B of the Appendix), therefore:

$$\begin{aligned}f(\theta\mathbb{Q}) &= \frac{1}{\mathbb{Q}(A=1)} \sum_{(x_1,x_2,A=1,Y=1)} \theta(x_1,x_2)\mathbb{Q}(y,x_1,x_2,a) \\ &= \frac{1}{\mathbb{Q}(A=1)} \left(\sum_{(x_1,X_2=0,A=1,Y=1)} \theta(x_1,x_2)\mathbb{Q}(y,x_1,x_2,a) + \sum_{(x_1,X_2=1,A=1,Y=1)} \theta(x_1,x_2)\mathbb{Q}(y,x_1,x_2,a) \right) \\ &\leq \frac{1}{\mathbb{Q}(A=1)} (c_1+c_2)\end{aligned}$$

On the other hand,

$$\begin{aligned}f(\theta\mathbb{Q}) &= \frac{1}{\mathbb{Q}(A=1)} \sum_{(x_1,x_2,A=1,Y=1)} \theta(x_1,x_2)\mathbb{Q}(y,x_1,x_2,a) \\ f(\theta\mathbb{Q}) &\geq \frac{1}{\mathbb{Q}(A=1)} \sum_{(x_1,x_2,A=1,Y=1)} P(R=1)\mathbb{Q}(y,x_1,x_2,a) \\ f(\theta\mathbb{Q}) &\geq \frac{\mathbb{P}(R=1)}{\mathbb{Q}(A=1)} \sum_{(x_1,x_2,A=1,Y=1)} \mathbb{Q}(y,x_1,x_2,a)\end{aligned}$$

$\frac{\mathbb{P}(R=1)}{\mathbb{Q}(A=1)} \sum_{(x_1,x_2,A=1,Y=1)} \mathbb{Q}(y,x_1,x_2,a)$ is non-zero, because $\mathbb{P}(R=1) > 0$ is non zero by definition (There exist a sample) and $\sum_{(x_1,x_2,A=1,Y=1)} Q(y,x_1,x_2,a)$ is also non zero under the mild assumption of the sample having datum of both outcomes conditioned on the ascertainment.

Hence in summary, the min/max over the constraints satisfy:

$$f(\theta\mathbb{Q}) \leq \frac{1}{\mathbb{Q}(A=1)} (c_1+c_2) < 1, \qquad f(\theta\mathbb{Q}) \geq \frac{\mathbb{P}(R=1)}{\mathbb{Q}(A=1)} \sum_{(x_1,x_2,A=1,Y=1)} \mathbb{Q}(y,x_1,x_2,a), > 0,$$

Where all terms are observable and $f(\theta\mathbb{Q})$ will, for many settings of their values, be bounded away from 0 and 1. To conclude, note as well that the bounds were computed using a proper subset of all the constrains (Θ), thus it follows that even tighter bounds can be derived.