Doubly Robust Counterfactual Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Much of the causal inference literature is concerned with estimands that can be expressed in closed form, such as the average treatment effect. In this work, we consider instead general estimands that are expressible as the solution to a non-linear optimization problem involving counterfactuals. In particular, we study counterfactual classification as a new decision-making tool for classification under hypothetical (contrary to fact) scenarios. We propose a doubly-robust nonparametric estimator for our counterfactual classifier where we can incorporate flexible constraints. We go on to analyze rates of convergence and provide a closed-form expression for the asymptotic distribution of our estimator. Our analysis shows that the proposed estimator is robust against nuisance model misspecification, and can attain fast $\sqrt{n}$ rates with tractable inference even when using flexible machine learning approaches. Finally, we study our methods via simulation, and apply them for recidivism risk prediction.

1 Introduction

Counterfactual (or potential) outcomes are used to describe how a unit would respond to a specific event or treatment, irrespective of whether the event actually takes place. They are commonly used in causal inference, where we would like to intervene on a factual prior event and then assess the consequences of that change [15, 16, 44].

More recently, counterfactual outcomes (or counterfactuals) have also been proved to be useful to support decision making when there is a need for predicting outcomes under hypothetical interventions. This is commonly referred to as *counterfactual prediction*. Counterfactual prediction can be used to build reliable models for decision support when effective treatments exist and differ across settings, but randomized experiments are not feasible. For example in health care, to inform decision making about treatment initiation physicians often require predicted risks assuming no treatment is given. In this case, simply treating the hypothetical treatment as another baseline covariate will rarely give the correct hypothetical risks because of confounding [56]. It may also be needed when we anticipate that treatment policies will change over time, as is often the case in clinical practice. [see, e.g., 10, 25, for more examples]. However, the problem of counterfactual prediction brings some interesting challenges that do not arise in typical prediction problems as the underlying data that we need to build predictive models are inherently not fully observable. Surprisingly, while the development of modern prediction modeling has greatly enriched the counterfactual-outcome-based causal inference particularly via semi-parametric methods [20], the use of causal inference to improve prediction modeling has received far less attention, although its potential is well acknowledged [10, 45].

In this work, we study counterfactual classification, a special case of counterfactual prediction where our outcome is discrete. Our approach allows investigators to flexibly incorporate various constraints into the models, not only to enhance their predictive performance but also to accommodate a wide range of practical constraints relevant to their classification tasks. This poses both theoretical and practical challenges, because unlike typical causal inference problems, the estimand is not expressible

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

as a closed form functional. We tackle this problem by framing counterfactual classification as a special case of nonlinear stochastic programming with counterfactual components.

1.1 Related Work

Our work lies at the intersection of causal inference and stochastic optimization.

Counterfactual prediction is closely related to conditional average treatment effect estimation in causal inference, which plays a crucial role in precision medicine and individualized policy. Let Y^a denote the counterfactual outcome that would have been observed under treatment or intervention $A = a$, $A \in \{0, 1\}$. Then the conditional average treatment effect (CATE) for subjects with a covariate $X = x$ is defined by $\tau(x) = \mathbb{E}[Y^1 - Y^0 \mid X = x]$. There exists a vast literature on estimating the CATE. These include some important early works assuming that $\tau(x)$ follows some known parametric form [e.g., 43, 51, 53]. But more recently, people have moved toward leveraging flexible nonparametric and machine learning methods [e.g., 1, 3, 22, 23, 27, 29, 38, 55]. A desirable property commonly held in the above CATE estimation methods is that the function $\tau(x)$ may be more structured and simple than its component main effect function $\mathbb{E}[Y^a \mid X = x]$.

In counterfactual prediction, however, we are rather concerned with predictions of Y^a conditional on $X = x$ under a "single" hypothetical intervention $A = a$ possibly with various constraints on our model. Thus, our target parameter is different from the CATE. With no constraints, under some appropriate identification assumptions (e.g., (C1)-(C3) in Section 2), it is equivalent to estimating a standard single-output regression function $\mathbb{E}[Y \mid X, A = a]$ so one could use any regression estimator. This direct modeling or *plug-in* approach has already been used for counterfactual prediction in randomized controlled trials [e.g., 24, 37] or as a part of the CATE estimation methods [e.g., 3, 27]. An issue arises when we are estimating a projection of this function onto a finite-dimensional model, or where we instead want to estimate $\mathbb{E}[Y^a \mid V] = \mathbb{E}\{\mathbb{E}[Y \mid X, A = a] \mid V\}$ for some smaller subset $V \subset X$ (e.g., under runtime confounding [9]), which renders the plug-in approach suboptimal. Moreover, it lacks double robustness, a highly desirable property which provides an additional layer of robustness against model misspecification [5], and it is not clear how to incorporate the given constraints into the modeling process.

On the other hand, we often want to incorporate various constraints into our predictive models. Such constraints are often used for flexible penalization [18] or supplying prior information [13] to enhance model performance and interpretation. They can also be used to mitigate certain biases [7, 14]. Further, depending on the scientific question, practitioners occasionally have some constraints they wish to place on their prediction tasks, such as targeting specific sub-populations, reflecting budget constraints, or accounting for the compositional nature of the data [19, 26, 36].

In our approach, we directly formulate and solve an optimization problem that minimizes counterfactual classification risk, where we can flexibly incorporate various forms of constraints. Constrained optimization problem involving counterfactuals, in particular asymptotic behavior of its optimal solutions, has not been extensively studied yet, with very few exceptions in causal inference [e.g., 4, 28, 32]. Our results are closest to the recent work of [31], who also propose a doubly robust estimator for the optimization problem involving counterfactuals. However, i) they use the mean square error (MSE), not the cross-entropy, ii) they consider linear constraints that were estimated rather than known, and iii) they gave rates for the optimal value but not the solution.

We instead tackle the problem of counterfactual classification from the perspective of stochastic programming. The complication arises since in stochastic programming accurate estimation of the unknown components does not necessarily guarantee accurate estimation of the optimal solutions unless additional regularity conditions are imposed. Two most common approaches in stochastic programming are stochastic approximation (SA) and sample average approximation (SAA) [e.g., 34, 49]. However, the standard SA and SAA approaches are not readily feasible in our setting, since we cannot compute sample moments or stochastic subgradients that involve unobserved counterfactuals. Moreover, they cannot harness efficient estimators for counterfactual components: e.g., doubly-robust or semiparametric estimators with cross-fitting [8, 35]. To address this issue, therefore, more general approaches beyond the SA and SAA settings should be considered [e.g., 46–48] at the expense of stronger assumptions on behavior of the optimal solution and its estimator.

1.2 Contribution

We study counterfactual classification as a new decision-making tool under hypothetical (contrary to fact) scenarios. Based on the semiparametric theory for causal inference, we propose a doubly-robust, nonparametric estimator that can incorporate flexible constraints into the modeling process. Then we go on to analyze rates of convergence and provide a closed-form expression for the asymptotic distribution of our estimator. Our analysis shows that the proposed estimator can attain fast $\sqrt{n}$ rates even when nuisance estimators are based on flexible machine learning tools with much slower rates. We study the finite-sample performance of our estimator via simulation and provide a case based on real data. Importantly, our algorithm and analysis are applicable to other problems in which the estimand is given by the solutions to a general optimization problem involving counterfactuals, where closed-form solutions are not available.

2 Problem and Setup

Suppose that we have access to an i.i.d. sample $(Z_1, ..., Z_n)$ of n tuples $Z = (Y, A, X) \sim \mathbb{P}$ for some distribution $\mathbb{P}$, binary outcome $Y \in \{0, 1\}$, covariates $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$, and binary intervention $A \in \mathcal{A} = \{0, 1\}$. For simplicity, we assume A and Y are binary, but in principle they can be multi-valued. We consider a general setting where only a subset of covariates $V \subseteq X$ can be used for predicting the counterfactual outcome Y^a. This allows for *runtime confounding*, where factors used by decision-makers are recorded in the training data but are not available for prediction (see [9] and references therein). We are concerned with the following constrained optimization problem

$$\begin{aligned} \underset{\beta \in \mathcal{B}}{\text{minimize}} \quad & \mathcal{L}\left(Y^a, \sigma(\beta, \boldsymbol{b}(V))\right) := -\mathbb{E}\left\{Y^a \log \sigma(\beta, \boldsymbol{b}(V)) + (1 - Y^a) \log(1 - \sigma(\beta, \boldsymbol{b}(V)))\right\} \\ \text{subject to} \quad & \beta \in \mathcal{S} := \{\beta \mid g_j(\beta) \leq 0, j \in J\} \end{aligned} \tag{P}$$

for some compact subset $\mathcal{B} \in \mathbb{R}^k$, known C^2-functions $g_j : \mathcal{B} \to \mathbb{R}$, $\sigma : \mathcal{B} \times \mathcal{X} \to (0, 1)$, and the index set $J = \{1, ..., m\}$ for the inequality constraints. Here, σ is the score function and $\boldsymbol{b}(V) = [b_1(V), ..., b_k(V)]^\top$ represents a set of basis functions for V (e.g., power series, truncated series, kernel, etc.). $\mathcal{L}\left(Y^a, \sigma(\beta, \boldsymbol{b}(V))\right)$ is our classification risk based on the cross-entropy. $\mathcal{S}$ consists of deterministic inequality constraints[1] and can be used for penalization or other practical restrictions described in Section 1. Let β^* denote an optimal solution in (P). β^* is our optimal model coefficients that minimize the counterfactual classification risk under the given constraints.

Classification risk and score function. Our classification risk $\mathcal{L}(Y^a, \sigma(\beta, \boldsymbol{b}(V)))$ is defined by the expected cross entropy loss between Y^a and $\sigma(\beta, \boldsymbol{b}(V))$. In order to estimate β^*, we first need to estimate this classification risk. Since it involves counterfactuals, the classification risk per se cannot be estimated from observed data without identification, which will be addressed shortly. The score function $\sigma(\beta, \boldsymbol{b}(V))$ depends on the classification technique. For simplicity, in this paper we focus on logistic regression, which is probably the most widely-used choice for classification. Hence with the given basis expansion $\boldsymbol{b}(\cdot)$, our score function is defined by $\sigma(\beta^\top \boldsymbol{b}(V))$ where $\sigma(\cdot)$ is the sigmoid function. Note that in this case, $\mathcal{L}(Y^a, \sigma(\beta^\top \boldsymbol{b}(V)))$ is guaranteed to be strictly convex. It should be noted that, however, one may consider employing other more flexible techniques such as neural networks, as long as we use a smooth, finite-dimensional parametric score function that satisfies the required regularity assumptions used to derive asymptotic results later in Section 4. Importantly in (P), we do not assume anything about the true functional relationship between Y^a and β; β is rather the parameter of the best-fitting function of the form $\sigma(\beta^\top \boldsymbol{b}(V))$. In this sense, our approach is nonparametric.

Identification. To estimate the counterfactual quantity $\mathcal{L}(Y^a, \sigma(\beta^\top \boldsymbol{b}(V)))$ from the observed sample $(Z_1, ..., Z_n)$, it must be expressed in terms of the observational data distribution $\mathbb{P}$. This can be accomplished via the following standard causal assumptions [e.g., 17, Chapter 12]:

- (C1) *Consistency*: $Y = Y^a$ if $A = a$
- (C2) *No unmeasured confounding*: $A \perp\!\!\!\perp Y^a \mid X$
- (C3) *Positivity*: $\mathbb{P}(A = a|X) > 0$ a.s.

[1] Equality constraint can be always expressed by a pair of inequality constraints.

(C1) - (C3) will be assumed throughout this paper. Under these assumptions, our classification risk is identified as

$$\mathcal{L}(\beta) = -\mathbb{E}\left\{\mathbb{E}\left[Y \mid X, A=a\right]\log\sigma(\beta^\top \boldsymbol{b}(V)) + (1-\mathbb{E}\left[Y \mid X, A=a\right])\log(1-\sigma(\beta^\top \boldsymbol{b}(V)))\right\}, \tag{1}$$

where we write $\mathcal{L}(\beta) \equiv \mathcal{L}(Y^a, \sigma(\beta^\top \boldsymbol{b}(V)))$ with the sigmoid function σ. Even though we develop our estimator under the above set of causal assumptions, one may extend our methods to other identification strategies as well (e.g., instrumental variables, mediation), since our approach is based on the analysis of stochastic programming with generic unknown objective functions (See Appendix B).

Notation. Here we give some notation used throughout this paper. For any fixed vector v, we let $\|v\|_2$ denote the Euclidean norm (or L_2-norm). Let $\mathbb{P}_n$ denote the empirical measure over $(Z_1, ..., Z_n)$. Given a sample operator h (e.g., an estimated function), we let $\mathbb{P}$ denote the conditional expectation over a new independent observation Z, as in $\mathbb{P}(h) = \mathbb{P}\{h(Z)\} = \int h(z)d\mathbb{P}(z)$. Then we use $\|h\|_{2,\mathbb{P}}$ to denote the $L_2(\mathbb{P})$ norm of h defined by $\|h\|_{2,\mathbb{P}} = \left[\mathbb{P}(h^2)\right]^{\frac{1}{2}} = \left[\int h(z)^2 d\mathbb{P}(z)\right]^{\frac{1}{2}}$. Further, we let $\mathsf{s}^*(P)$ denote the set of optimal solutions of an optimization program P, i.e., $\beta^* \in \mathsf{s}^*(\mathsf{P})$, and define $\mathsf{dist}(x, S) = \inf\{\|x-y\|_2 : y \in S\}$ to denote the distance from a point x to a set S.

3 Estimation Algorithm

Since (P) is not directly solvable, we need to find an approximating program of the "true" program (P). To this end, we shall first discuss the problem of obtaining estimates for the identified classification risk (1). To simplify notation, we first introduce the following nuisance functions

$$\pi_a(X) = \mathbb{P}[A=a \mid X],$$
$$\mu_a(X) = \mathbb{E}[Y \mid X, A=a],$$

and let $\widehat{\pi}_a, \widehat{\mu}_a$ be their estimators. π_a, μ_a are the propensity score and outcome regression functions, respectively.

A natural estimator for (1) is given by

$$\widehat{\mathcal{L}}(\beta) = -\mathbb{P}_n\left\{\widehat{\mu}_a(X)\log\sigma(\beta^\top \boldsymbol{b}(V)) + (1-\widehat{\mu}_a(X))\log(1-\sigma(\beta^\top \boldsymbol{b}(V)))\right\}, \tag{2}$$

where we simply plug-in the regression estimates $\widehat{\mu}_a$ into the empirical average of (1). Here, we propose a more efficient estimator based on the semiparametric approach in causal inference. Let

$$\varphi_a(Z;\eta) = \frac{\mathbb{1}(A=a)}{\pi_a(X)}\left\{Y - \mu_A(X)\right\} + \mu_a(X),$$

denote the uncentered efficient influence function for the parameter $\mathbb{E}\left\{\mathbb{E}[Y \mid X, A=a]\right\}$, where a set of the nuisance functions is defined by $\eta = \{\pi_a(X), \mu_a(X)\}$. Then it can be deduced that for an arbitrary fixed real-valued function $h : \mathcal{X} \to \mathbb{R}$, the uncentered efficient influence function for the parameter $\psi_a := \mathbb{E}\left\{\mathbb{E}[Y \mid X, A=a]h(X)\right\}$ is given by $\varphi_a(Z;\eta)h(X)$.

Now we provide an influence-function-based semiparametric estimator for ψ_a [21]. Following [8, 22, 42, 57], we propose to use *sample splitting* to allow for arbitrarily complex nuisance estimators $\widehat{\eta}$. Specifically, we split the data into K disjoint groups, each with the size of n/K ideally, by drawing variables $(B_1, ..., B_n)$ independent of the data, with $B_i = b$ indicating that subject i was split into group $b \in \{1, ..., K\}$. Then the semiparametric estimator for ψ_a based on the efficient influence function and sample splitting is given by

$$\widehat{\psi}_a = \frac{1}{K}\sum_{b=1}^{K}\mathbb{P}_n^b\left\{\varphi_a(Z;\widehat{\eta}_{-b})h(X)\right\} \equiv \mathbb{P}_n\left\{\varphi_a(Z;\widehat{\eta}_{-B_K})h(X)\right\}, \tag{3}$$

where we let $\mathbb{P}_n^b$ denote empirical averages only over the set of units $\{i : B_i = b\}$ in the group b and let $\widehat{\eta}_{-B_K}$ denote the nuisance estimator constructed only using those units $\{i : B_i \neq b\}$. Under weak regularity conditions, this semiparametric estimator attains the efficiency bound with the double robustness property, thus allows us to employ flexible machine learning methods while achieving the $\sqrt{n}$-rate of convergence and valid inference under weak conditions (See Lemma A.1 in the appendix

Algorithm 1: Doubly robust estimator for counterfactual classification

input: $\boldsymbol{b}(\cdot), K$
Draw $(B_1, ..., B_n)$ with $B_i \in \{1, ..., K\}$
for $b = 1, ..., K$ **do**
  Let $D_0 = \{Z_i : B_i \neq b\}$ and $D_1 = \{Z_i : B_i = b\}$
  Obtain $\widehat{\eta}_{-b}$ by constructing $\widehat{\pi}_a, \widehat{\mu}_a$ on D_0
  $M_{1,b}(\beta) \leftarrow$ empirical average of $\varphi_a(Z; \widehat{\eta}_{-b}) \log \sigma(\beta^\top \boldsymbol{b}(V))$ over D_1
  $M_{0,b}(\beta) \leftarrow$ empirical average of $(1 - \varphi_a(Z; \widehat{\eta}_{-b})) \log(1 - \sigma(\beta^\top \boldsymbol{b}(V)))$ over D_1
$\widehat{\mathcal{L}}(\beta) \leftarrow \sum_{b=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(B_i = b) \right\} (M_{1,b}(\beta) + M_{0,b}(\beta))$
solve $(\widehat{\mathsf{P}})$ with $\widehat{\mathcal{L}}(\beta)$

for the formal statement). If one is willing to rely on appropriate empirical process conditions (e.g., Donsker-type or low entropy conditions [52]), then η can be estimated on the same sample without sample splitting. However this would limit the flexibility of the nuisance estimators.

The classification risk $\mathcal{L}(\beta)$ is a sum of two functionals, each of which is in the form of ψ_a, Thus, for each β, we propose to estimate the classification risk using (3) as follows

$$\widehat{\mathcal{L}}(\beta) = -\mathbb{P}_n \left\{ \varphi_a(Z; \widehat{\eta}_{-B_K}) \log \sigma(\beta^\top \boldsymbol{b}(V)) + (1 - \varphi_a(Z; \widehat{\eta}_{-B_K})) \log(1 - \sigma(\beta^\top \boldsymbol{b}(V))) \right\}. \quad (4)$$

Now that we have proposed the efficient method to estimate the counterfactual component of the true program (P), in what follows we provide an approximating program which we aim to actually solve by substituting $\widehat{\mathcal{L}}(\beta)$ for $\mathcal{L}(\beta)$

$$\begin{aligned} \underset{\beta \in \mathcal{B}}{\text{minimize}} \quad & \widehat{\mathcal{L}}(\beta) \\ \text{subject to} \quad & \beta \in \mathcal{S}. \end{aligned} \quad (\widehat{\mathsf{P}})$$

$(\widehat{\mathsf{P}})$ is a smooth nonlinear optimization problem whose objective function depends on data. Note that unlike (P), $(\widehat{\mathsf{P}})$ is not guaranteed to be convex even if $\mathcal{S}$ is convex. Substantial progress has been made in general smooth nonlinear optimization in recent decades [6, 41], and a number of efficient global optimization algorithms are available via open-source libraries (e.g., NLOPT). Let $\widehat{\beta} \in \mathsf{s}^*(\widehat{\mathsf{P}})$. Then $\widehat{\beta}$ is our estimator for β^*. We summarize our algorithm detailing how to compute the estimator $\widehat{\beta}$ in Algorithm 1.

4 Asymptotic Analysis

This section is devoted to analyzing the rates of convergence and asymptotic distribution for the estimated optimal solution $\widehat{\beta}$. First, we introduce the following assumptions for our counterfactual component estimator $\widehat{\mathcal{L}}$.

(A1) $\mathbb{P}(\widehat{\pi}_a \in [\epsilon, 1 - \epsilon]) = 1$ for some $\epsilon > 0$

(A2) $\|\widehat{\mu}_a - \mu_a\|_{2,\mathbb{P}} = o_{\mathbb{P}}(1)$ or $\|\widehat{\pi}_a - \pi_a\|_{2,\mathbb{P}} = o_{\mathbb{P}}(1)$

(A3) $\|\widehat{\pi}_a - \pi_a\|_{2,\mathbb{P}} \|\widehat{\mu}_a - \mu_a\|_{2,\mathbb{P}} = o_{\mathbb{P}}(n^{-\frac{1}{2}})$

Assumptions (A1) - (A3) are commonly used in semiparametric estimation in the causal inference literature [20]. Next, for a feasible point $\bar{\beta} \in \mathcal{S}$ we define the active index set.

Definition 4.1 (Active set)**.** *For $\bar{\beta} \in \mathcal{S}$, we define the active index set J_0 by*

$$J_0(\bar{\beta}) = \{1 \leq j \leq m \mid g_j(\bar{\beta}) = 0\}.$$

Then we introduce the following technical condition on g_j.

(B1) For each $\beta^* \in \mathsf{s}^*(\mathsf{P})$,

$$d^\top \nabla^2_\beta g_j(\beta^*) d \geq 0 \quad \forall d \in \{d \mid \nabla_\beta g_j(\beta^*) = 0, j \in J_0(\bar{\beta})\}.$$

Assumption (B1) holds, for example, if each g_j is locally convex around β^*. In what follows, based on the result of [46], we characterize the rates of convergence for $\widehat{\beta}$ in terms of the nuisance estimation error under relatively weak conditions.

Theorem 4.1 (Rate of Convergence)**.** *Assume that (A1), (A2), and (B1), hold. Then*

$$\textit{dist}\left(\widehat{\beta}, \mathsf{s}^*(\mathsf{P})\right) = O_{\mathbb{P}}\left(\|\widehat{\pi}_a - \pi_a\|_{2,\mathbb{P}}\|\widehat{\mu}_a - \mu_a\|_{2,\mathbb{P}}\right).$$

Hence, if we further assume the nonparametric condition (A3), we obtain

$$\textit{dist}\left(\widehat{\beta}, \mathsf{s}^*(\mathsf{P})\right) = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

Theorem 4.1 indicates that double robustness is possible for our estimator, and thereby $\sqrt{n}$ rates are attainable even when each of the nuisance regression functions is estimated at much slower rates (e.g., $n^{-1/4}$ rates for each). Since $\mathcal{L}$ is continuously differentiable with bounded derivative, the consistency of the optimal value naturally follows by the result of Theorem 4.1 and the continuous mapping theorem. More specifically, in the following corollary, we show that the same rates are attained for the optimal value as well under the same conditions.

Corollary 4.1 (Rate of Convergence for Optimal Value)**.** *Suppose (A1), (A2), (A3), (B1) hold and let v^* and $\widehat{v}$ be the optimal values corresponding to $\beta^* \in \mathsf{s}^*(\mathsf{P})$ and $\widehat{\beta}$, respectively. Then we have $|\widehat{v} - v^*| = O_{\mathbb{P}}\left(\|\widehat{\pi}_a - \pi_a\|_{2,\mathbb{P}}\|\widehat{\mu}_a - \mu_a\|_{2,\mathbb{P}}\right)$.*

In order to conduct statistical inference, it is also desirable to characterize the asymptotic distribution of $\widehat{\beta}$. This requires stronger assumptions and more delicate analysis [46]. Asymptotic properties of optimal solutions in stochastic programming are typically studied based on the generalization of the delta method for directionally differentiable mappings [e.g., 47–49]. Asymptotic normality is of particular interest since without asymptotic normality, consistency of the bootstrap is no longer guaranteed for the solution estimators [12].

We start with giving additional definitions of some popular regularity conditions with respect to (P).

Definition 4.2 (LICQ)**.** *Linear independence constraint qualification (LICQ) is satisfied at $\bar{\beta} \in \mathcal{S}$ if the vectors $\nabla_\beta g_j(\bar{\beta})$, $j \in J_0(\bar{\beta})$ are linearly independent.*

Definition 4.3 (SC)**.** *Let $L(\beta, \gamma)$ be the Lagrangian. Strict Complementarity (SC) is satisfied at $\bar{\beta} \in \mathcal{S}$ if, with multipliers $\bar{\gamma}_j \geq 0$, $j \in J_0(\bar{\beta})$, the Karush-Kuhn-Tucker (KKT) condition*

$$\nabla_\beta L(\bar{\beta}, \bar{\gamma}) := \nabla_\beta \mathcal{L}(\bar{\beta}) + \sum_{j \in J_0(\bar{\beta})} \bar{\gamma}_j \nabla_\beta g_j(\bar{\beta}) = 0,$$

is satisfied such that $\bar{\gamma}_j > 0, \forall j \in J_0(\bar{\beta})$.

LICQ is arguably one of the most widely-used constraint qualifications that admit the first-order necessary conditions. SC means that if the j-th inequality constraint is active then the corresponding dual variable is strictly positive, so exactly one of them is zero for each $1 \leq j \leq m$. SC is widely used in the optimization literature, particularly in the context of parametric optimization [e.g., 49, 50]. We further require uniqueness of the optimal solution in (P).

(B2) Program (P) has a unique optimal solution β^* (i.e., $\mathsf{s}^*(\mathsf{P}) \equiv \{\beta^*\}$ is singleton).

Note that under (B2) if LICQ holds at β^*, then the corresponding multipliers are determined uniquely [54]. In the next theorem, we provide a closed-form expression for the asymptotic distribution of $\widehat{\beta}$.

Theorem 4.2 (Asymptotic Distribution)**.** *Assume that (A1) - (A3), (B1), and (B2) hold, and that LICQ and SC hold at β^* with the corresponding multipliers γ^*. Then*

$$n^{-\frac{1}{2}}\left(\widehat{\beta} - \beta^*\right) = \begin{bmatrix} \nabla^2_\beta L(\beta^*, \gamma^*) & \mathsf{B} \\ \mathsf{B}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}^\top \Upsilon + o_{\mathbb{P}}(1)$$

for some $k \times |J_0(\beta^*)|$ matrix B and random variable Υ such that

$$\Upsilon \xrightarrow{d} N\left(0, \textit{var}\left(\varphi_a(Z;\eta)h_1(V,\beta^*) + \{1 - \varphi_a(Z;\eta)\}h_0(V,\beta^*)\right)\right),$$

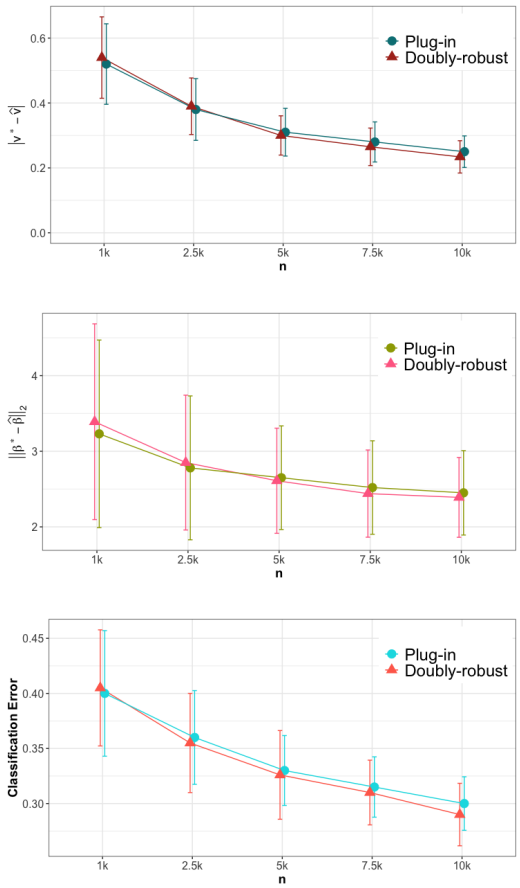

Figure 1: With correct X

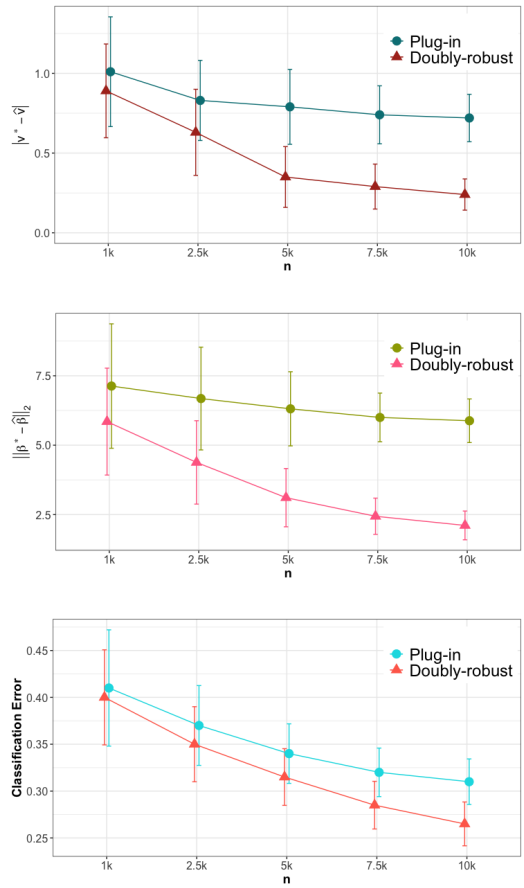

Figure 2: With distorted X

where

$$\mathsf{B} = \left[\nabla_\beta g_j(\beta^*)^\top,\, j \in J_0(\beta^*)\right],$$
$$h_1(V,\beta) = \frac{1}{\log \sigma(\beta^\top \boldsymbol{b}(V))} \boldsymbol{b}(V)\sigma(\beta^\top \boldsymbol{b}(V))\{1-\sigma(\beta^\top \boldsymbol{b}(V))\},$$
$$h_0(V,\beta) = -\frac{1}{\log(1-\sigma(\beta^\top \boldsymbol{b}(V)))} \boldsymbol{b}(V)\sigma(\beta^\top \boldsymbol{b}(V))\{1-\sigma(\beta^\top \boldsymbol{b}(V))\}.$$

The above theorem gives explicit conditions under which $\widehat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal. We harness the classical results of [47] that use an expansion of $\widehat{\beta}$ in terms of an auxiliary parametric program. To show asymptotic normality of $\widehat{\beta}$, linearity of the directional derivative of optimal solutions in the parametric program is required. We have accomplished this based on an appropriate form of the implicit function theorem [11]. This is in contrast to [31] that relied on the structure of the smooth, closed-form solution estimator that enables direct use the delta method. Lastly, it is worth noting that our results in this section are extendable to a more general constrained nonlinear optimization setting where the objective function involves counterfactuals (see Lemmas B.1, B.2 in the appendix).

5 Simulation and Case Study

5.1 Simulation

We explore the finite sample properties of our estimators in the simulated dataset where we aim to empirically demonstrate the double-robustness property described in Section 3. Our data generation

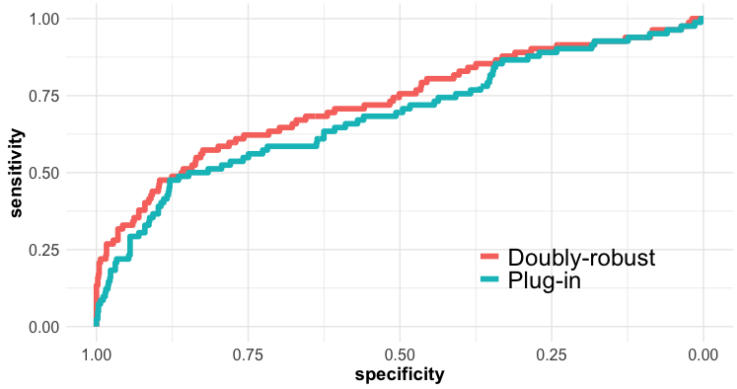

Figure 3: ROC curves

Method	AUC	Accuracy
Plug-in	0.692	0.64
Doubly-Robust	**0.718**	**0.68**
Raw COMPAS Score	0.688	0.65

Table 1: AUC and classification accuracy

process is as follows:

$$
\begin{gathered}
V \equiv X = (X_1, ..., X_6) \sim N(0, I), \\
\pi_a(X) = \text{expit}(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4 + 0.05X_5 + 0.05X_6), \\
Y = A\mathbb{1}\left\{X_1 + 2X_2 - 2X_3 - X_4 + X_5 + \varepsilon > 0\right\} + (1-A)\mathbb{1}\left\{X_1 + 2X_2 - 2X_3 - X_4 + X_6 + \varepsilon < 0\right\}, \\
\varepsilon \sim N(0, 1).
\end{gathered}
$$

Our classification target is Y^1. For $\boldsymbol{b}(X)$, we use X, X^2 and their pairwise products. We assume that we have box constraints for our solution: $|\beta_j^*| \leq 1$, $j = 1, ..., k$. Since there exist no other natural baselines, we compare our methods to the plug-in method where we use (2) for our approximating program $\widehat{\mathsf{P}}$. For nuisance estimation we use the cross-validation-based Super Learner ensemble via the SUPERLEARNER R package to combine generalized additive models, multivariate adaptive regression splines, and random forests. We use sample splitting as described in Algorithm 1 with $K = 2$ splits. We further consider two versions of each of our estimators, based on the correct and distorted X, where the distorted values are only used to estimate the outcome regression μ_a. The distortion is caused by a transformation $X \mapsto (X_1X_3X_6, X_2^2, X_4/(1 + \exp(X_5)), \exp(X_5/2))$.

To solve $\widehat{\mathsf{P}}$, we first use the StoGo algorithm [39] via the NLOPTR R package as it has shown the best performance in terms of accuracy in the survey study of [33]. After running the StoGo, we then use the global optimum as a starting point for the BOBYQA local optimization algorithm [40] to further polish the optimum to a greater accuracy. We use sample sizes $n = 1k, 2.5k, 5k, 7.5k, 10k$ and repeat the simulation 100 times for each n. Then we compute the average of $|v^* - \widehat{v}|$ and $\|\beta^* - \widehat{\beta}\|_2$. Using the estimated counterfactual predictor, we also compute the classification error on an independent sample with the equal sample size. Standard error bars are presented around each point. The results with the correct and distorted X are presented in Figures 1 and 2, respectively.

With the correct X, it appears that the proposed estimator performs as well or slightly better than the plug-in methods. However, in Figure 2 when $\widehat{\mu}_a$ is constructed based on the distorted X, the proposed estimator gives substantially smaller errors in general and improves better with n. This is indicative of the fact that the proposed estimator has the doubly-robust, second-order multiplicative bias, thus supports our theoretical results in Section 4.

5.2 Case Study: COMPAS Dataset

Next we apply our method in recidivism risk prediction using the COMPAS dataset [2]. This dataset was originally designed to assess the COMPAS recidivism risk scores, and has been utilized for studying machine bias in the context of algorithmic fairness [2]. More recently, the dataset has been reanalyzed in the framework of counterfactual outcomes [30–32]. Here, we rather focus purely on predictive purpose. We let an intervention A represent pretrial release, coded as $A = 0$ if defendants are released and $A = 1$ if they are incarcerated, following methodology suggested by [32] [3]. We aim to classify the binary counterfactual outcome Y^0 that indicates whether a defendant is rearrested within two years, should the defendant be released pretrial. We use the dataset for two-year recidivism records with five covariates: age, sex, the number of prior arrests, charge degree, race. We consider three races: 3175 Black, 2103 White, 509 Hispanic. We split the data ($n = 5787$) randomly into two groups: train set with 3000 instances and test set with the rest. Other model settings remain the same as our simulation in the previous subsection, including the box constraints.

Figure 3 and Table 1 show that the proposed doubly-robust method achieves slightly higher ROC AUC and classification accuracy than the plug-in, as well as the results based on the raw COMPAS risk scores. We conjecture that the relatively incremental improvements are in part due to violation of the causal assumptions [30].

6 Discussion

In this paper we studied the problem of counterfactual classification under arbitrary smooth constraints, and proposed a doubly-robust estimator which leverages flexible machine learning tools. Our theoretical framework is not only limited to counterfactual classification, but can be applied to other settings where the estimand is the optimal solution of a smooth nonlinear programming problem with a counterfactual objective function.

We conclude with some potential limitations of our methods, and ways in which our work could be generalized. First, owing to the stochastic nature of $\widehat{\mathcal{L}}$ it is possible that we end up solving a non-convex program. Despite many recent advances, still non-convex problems are usually more difficult than convex ones due to high variance and slow computing time. To address this, one may instead consider finding the closest convex approximation of the function $\widehat{\mathcal{L}}$ and solving the corresponding convex program. This can be justified by the fact that $\widehat{\mathcal{L}} \xrightarrow{p} \mathcal{L}$. Second, we only considered the situations where the feasible set is deterministic and fixed. However, sometimes it may be useful to consider the general case where the feasible set can be different between the true and the approximating programs; this corresponds to the case where g_j's are to be estimated as well. This can be particularly helpful when incorporating (counterfactual) fairness constraints [31, 32]. Dealing with the varying feasible set with general nonlinear constraints is a complicated task and requires even stronger assumptions [47]. A promising direction for future research is to generalize our framework to the case of a varying feasible set. Lastly, although we showed that the counterfactual objective function is estimated efficiently via $\widehat{\mathcal{L}}$, it is not necessarily clear whether we can deduce efficiency argument for the solution estimator $\widehat{\beta}$ due to the inherent complexity of the optimal solution mapping in the presence of constraints. We conjecture that one may show that the semiparametric efficiency bound can be also attained for $\widehat{\beta}$ possibly under slightly stronger regularity assumptions, but we leave this for future work.

[2] `https://github.com/propublica/compas-analysis`

[3] The dataset itself does not include information whether defendants were released pretrial, but it includes dates in and out of jail. So we set the treatment A to 0 if defendants left jail within three days of being arrested, and 1 otherwise, as Florida state law generally requires individuals to be brought before a judge for a bail hearing within 2 days of arrest [32, Section 6.2].

References

[1] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[3] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[4] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

[5] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[6] Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 3, 2020.

[7] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pages 71–80. IEEE, 2013.

[8] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, May 2017.

[9] Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. Counterfactual predictions under runtime confounding. In *Advances in neural information processing systems*, volume 33, pages 4150–4162, 2020.

[10] Barbra A Dickerman and Miguel A Hernán. Counterfactual prediction is not only for causal inference. *European Journal of Epidemiology*, 35(7):615–617, 2020.

[11] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

[12] Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019.

[13] Brian R Gaines, Juhyun Kim, and Hua Zhou. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018.

[14] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[15] Marc Höfler. Causal inference based on counterfactuals. *BMC medical research methodology*, 5(1):1–12, 2005.

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

[17] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[18] Gareth M James, Courtney Paulson, and Paat Rusmevichientong. Penalized and constrained regression. *Unpublished manuscript, http://www-bcf. usc. edu/gareth/research/Research. html*, 2013.

[19] Gareth M James, Courtney Paulson, and Paat Rusmevichientong. Penalized and constrained optimization: an application to high-dimensional website advertising. *Journal of the American Statistical Association*, 2019.

[20] Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

[21] Edward H Kennedy. Semiparametric theory. *arXiv preprint arXiv:1709.06418*, 2017.

[22] Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[23] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

[24] Junlong Li, Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, Andrea Callegaro, Benjamin Dizier, Bart Spiessens, Fernando Ulloa-Montoya, and Lee-Jen Wei. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. *Biometrics*, 72(3):877–887, 2016.

[25] Lijing Lin, Matthew Sperrin, David A Jenkins, Glen P Martin, and Niels Peek. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnostic and prognostic research*, 5(1):1–16, 2021.

[26] Jiarui Lu, Pixu Shi, and Hongzhe Li. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics*, 75(1):235–244, 2019.

[27] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

[28] Alexander R Luedtke and Mark J van der Laan. Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics*, 12(1):283–303, 2016.

[29] Alexander R Luedtke and Mark J van der Laan. Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics*, 12(1):305–332, 2016.

[30] Alan Mishler. Modeling risk and achieving algorithmic fairness using potential outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 555–556, 2019.

[31] Alan Mishler and Edward Kennedy. Fade: Fair double ensemble learning for observable and counterfactual outcomes. *arXiv preprint arXiv:2109.00173*, 2021.

[32] Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

[33] Katharine M Mullen. Continuous global optimization in r. *Journal of Statistical Software*, 60: 1–45, 2014.

[34] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

[35] Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

[36] Sam Nguyen, Ryan Chan, Jose Cadena, Braden Soper, Paul Kiszka, Lucas Womack, Mark Work, Joan M Duggan, Steven T Haller, Jennifer A Hanrahan, et al. Budget constrained machine learning for early prediction of adverse outcomes for covid-19 patients. *Scientific Reports*, 11 (1):1–14, 2021.

[37] Tri-Long Nguyen, Gary S Collins, Paul Landais, and Yannick Le Manach. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—an illustration with the international stroke trial. *Journal of clinical epidemiology*, 125:47–56, 2020.

[38] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[39] Vladimir I Norkin, Georg Ch Pflug, and Andrzej Ruszczyński. A branch and bound method for stochastic global optimization. *Mathematical programming*, 83(1):425–450, 1998.

[40] Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26, 2009.

[41] Tamás Rapcsák. *Smooth nonlinear optimization in Rn*, volume 19. Springer Science & Business Media, 2013.

[42] James Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

[43] James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

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

[45] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems*, 30:1697–1708, 2017.

[46] Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186, 1991.

[47] Alexander Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

[48] Alexander Shapiro. Statistical inference of stochastic optimization problems. In *Probabilistic constrained optimization*, pages 282–307. Springer, 2000.

[49] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[50] Georg Still. Lectures on parametric optimization: An introduction. *Optimization Online*, 2018.

[51] Mark J Van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

[52] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[53] Stijn Vansteelandt and Marshall Joffe. Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

[54] Gerd Wachsmuth. On licq and the uniqueness of lagrange multipliers. *Operations Research Letters*, 41(1):78–80, 2013.

[55] 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.

[56] Daniel Westreich and Sander Greenland. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298, 2013.

[57] Wenjing Zheng and Mark J Van Der Laan. Asymptotic theory for cross-valiyeard targeted maximum likelihood estimation. *Working Paper 273*, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work focuses on the methodological development and is not tied to particular applications. Also, our methods do not suggest any particular ways to quantify fairness. Hence discussion about potential negative societal impacts is not applicable to ours.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section B in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] But we are planning on making our algorithm as an independent function and posting it on Github.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We have only used the public data and the open-source packages in R (see Section 5).
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]