Optimal Training of Fair Predictive Models

Razieh Nabi, Daniel Malinsky, Ilya Shpitser
Department of Computer Science,
Johns Hopkins University, Baltimore, MD, USA
(*rnabi*@, *malinsky*@, *ilyas*@cs.)jhu.edu

Abstract

Recently there has been sustained interest in modifying prediction algorithms to satisfy fairness constraints. These constraints are typically complex nonlinear functionals of the observed data distribution. Focusing on the causal constraints proposed by Nabi and Shpitser (2018), we introduce new theoretical results and optimization techniques to make model training easier and more accurate. Specifically, we show how to reparameterize the observed data likelihood such that fairness constraints correspond directly to parameters that appear in the likelihood, transforming a complex constrained optimization objective into a simple optimization problem with box constraints. We also exploit methods from empirical likelihood theory in statistics to improve predictive performance, without requiring parametric models for high-dimensional feature vectors.

1 INTRODUCTION

Predictive models trained on imperfect data are increasingly being used in socially-impactful settings. Predictions (such as risk scores) have been used to inform high-stakes decisions in criminal justice (Perry et al., 2013), healthcare (Kappen et al., 2018), and finance (Khandani et al., 2010). While automation may bring many potential benefits – such as speed and accuracy – it is also fraught with risks. Predictive models introduce two dangers in particular: the illusion of objectivity and violation of fairness norms. Predictive models may appear to be "neutral," since humans are less involved and because they are products of a seemingly impartial optimization process. How-

ever, predictive models are trained on data that reflects the structural inequities, historical disparities, and other imperfections of our society. Often data includes sensitive attributes (e.g., race, gender, age, disability status), or proxies for such attributes. A particular worry in the context of data-driven decision-making is "perpetuating injustice," which occurs when unfair dependence between sensitive features and outcomes is maintained, introduced, or reinforced by automated tools.

We study how to construct fair predictive models by correcting for the unfair causal dependence of predicted outcomes on sensitive features. We work with the proposed fairness criteria in Nabi and Shpitser (2018), where the authors propose that fair prediction requires imposing hard constraints on the prediction problem in the form of restricting certain causal path-specific effects. Impermissible pathways are user-specified and context-specific, hence require input from policymakers, legal experts, or the general public. Some alternative but also causally-motivated constrained prediction methods are proposed in Chiappa (2019); Kusner et al. (2017a) and Zhang and Bareinboim (2018). For a survey and discussion of distinct fairness criteria (both causal and associative) see Mitchell et al. (2018).

We advance the state of the art in two ways. First, we give a novel reparameterization of the observed data likelihood in which unfair path-specific effects appear directly as parameters. This allows us to greatly simplify the constrained optimization problem, which has previously required complex or inefficient algorithms. Second, we demonstrate how tools from the empirical likelihood literature (Owen, 2001) can be readily adapted to construct hybrid (semi-parametric) observed data likelihoods that satisfy given fairness criteria. With this approach, the entire likelihood is constrained, rather than only part of the likelihood as in past proposals (Nabi and Shpitser, 2018). As a result, we use the data more efficiently and achieve better performance. Finally, we show how both innovations

may be combined into a single procedure.

As a guiding example, we consider a setting such as automated hiring, in which we want to predict job success from applicant data. We have historical data on job success, resumes, and demographics, as well as new individuals for which we only see resumes and demographics for whom we would like to estimate a risk score with our predictive model. This may be considered a variant of semi-supervised learning or prediction with missing labels on a subset of the population. We aim to estimate those scores subject to path-specific fairness constraints. In order to describe the various components of this proposal, we must review some background on causal inference, path-specific effects, and constrained prediction.

2 CAUSAL INFERENCE AND A CAUSAL APPROACH TO FAIRNESS

Causal inference is concerned with quantities which describe the consequences of interventions. Causal models are often represented graphically, e.g. by directed acyclic graphs (DAGs). We will use capital letters (V) to denote sets of random variables as well as corresponding vertices in graphs and lowercase letters (v) to denote values or assignments to those random variables. A DAG consists of a set of vertices V connected by directed edges ($V_i \rightarrow V_j$ for some $\{V_i, V_j\} \subseteq V$) such that there are no cycles. The set $\mathrm{pa}_{\mathcal{G}}(V_i) \equiv \{V_j \in V \mid V_j \rightarrow V_i\}$ denotes the parents of V_i in DAG $\mathcal{G}$. $\mathfrak{X}_A$ denotes the statespace of $A \subseteq V$.

A causal model of a DAG $\mathcal{G}$ is a set of distributions defined on potential outcomes (a.k.a. counterfactuals). For example, we consider distributions $p(V(a))$ subject to some restrictions, where $V(a)$ represents the value of V had all variables in $\mathrm{pa}_{\mathcal{G}}(V)$ been set, possibly contrary to fact, to value a. In this paper, we assume Pearl's *functional model* (Pearl, 2009) for a DAG $\mathcal{G}$ which stipulates that the sets of potential outcome variables $\left\{ \{V_i(a_i) \mid a_i \in \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V_i)}\} \mid V_i \in V \right\}$ are mutually independent. All other counterfactuals may be defined using *recursive substitution*. For any $A \subseteq V \setminus \{V_i\}$,

$$V_i(a) \equiv V_i(a_{\mathrm{pa}_{\mathcal{G}}(V_i) \cap A}, \{V_j(a) : V_j \in \mathrm{pa}_{\mathcal{G}}(V_i) \setminus A\}),$$

where $\{V_j(a) : V_j \in \mathrm{pa}_{\mathcal{G}}(V_i) \setminus A\}$ is taken to mean the (recursively defined) set of counterfactuals associated with variables in $\mathrm{pa}_{\mathcal{G}}(V_i) \setminus A$, had A been set to a. Equivalently, Pearl's model may be described by a system of nonparametric structural equations with independent errors.

A causal parameter is said to be *identified* in a causal model if it is a function of the observed data distribution $p(V)$. In the functional model of a DAG $\mathcal{G}$ (as well as some weaker causal models), all interventional distributions $p(V(a))$, for any $A \subseteq V$, are identified by the *extended g-formula*:

$$p(V(a)) = \prod_{V_i \in V} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i)) \big|_{A=a}.$$

For example, consider the DAG in Fig. 1(a). $Y(a)$ is defined to be $Y(a, M(a, X), X)$ by recursive substitution and its distribution is identified as $\sum_{X,M} p(Y|a, M, X) p(M|a, X) p(X)$. The mean difference between $Y(a)$ and $Y(a')$ for some treatment value a of interest and reference value a' is $\mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$ and quantifies the *average causal effect* of treatment A on the outcome Y.

2.1 Mediation Analysis and Path-Specific Effects

An important goal in causal inference is to understand the mechanisms by which some treatment A influences some outcome Y. A common framework for studying mechanisms is *mediation analysis* which seeks to decompose the effect of A on Y into the *direct effect* and the *indirect effect* mediated by a third variable, or more generally into components associated with particular causal pathways. As an example, the direct effect of A on Y in Fig. 1(a) corresponds to the effect along the edge $A \rightarrow Y$ and the indirect effect corresponds to the effect along the path $A \rightarrow M \rightarrow Y$, mediated by M.

In the potential outcome notation, the direct and indirect effects can be defined using nested counterfactuals such as $Y(a, M(a'))$ for $a, a' \in \mathfrak{X}_A$, which denotes the value of Y when A is set to a while M is set to whatever value it would have attained had A been set to a'. Given $p(a, M(a'))$, the *natural direct effect* (NDE) (on the expectation difference scale) is defined as $\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')]$, and the *natural indirect effect* (NID) is defined as $\mathbb{E}[Y(a)] - \mathbb{E}[Y(a, M(a'))]$. Under certain identification assumptions discussed by (Pearl, 2001), the distribution of $Y(a, M(a'))$ (and thereby direct and indirect effects) can be nonparametrically identified from observed data by the following formula:

$$p(Y(a, M(a'))) = \sum_{X,M} p(Y \mid a, X, M) \, p(M \mid a', X) \, p(X).$$

More generally, when there are multiple pathways from A to Y one may define various *path-specific effects* (PSEs). In this case, effect along a particular path will be obtained by comparing two potential outcomes, one where for the selected paths all nodes behave as if $A = a$, and along all other paths nodes behave as if $A = a'$.

PSEs are defined by means of nested, path-specific potential outcomes. Fix a set of treatment variables A, and a subset of *proper causal paths* π from any element in A. A proper causal path only intersects A at the source node. Next, pick a pair of value sets a and a' for elements in A. For any $V_i \in V$, define the potential outcome $V_i(\pi, a, a')$ by setting A to a for the purposes of paths in π, and to a' for the purposes of proper causal paths from A to Y not in π. Formally, the definition is as follows, for any $V_i \in V$, $V_i(\pi, a, a') \equiv a$ if $V_i \in A$, otherwise

$$V_i(\pi, a, a') \equiv V_i\Big(\{V_j(\pi, a, a') \mid V_j \in \mathrm{pa}_{\mathcal{G}}^{\pi}(V_i)\},$$
$$\{V_j(a') \mid V_j \in \mathrm{pa}_{\mathcal{G}}^{\overline{\pi}}(V_i)\}\Big) \quad (1)$$

where $V_j(a') \equiv a'$ if $V_j \in A$, and given by recursive substitution otherwise, $\mathrm{pa}_{\mathcal{G}}^{\pi}(V_i)$ is the set of parents of V_i along an edge which is a part of a path in π, and $\mathrm{pa}_{\mathcal{G}}^{\overline{\pi}}(V_i)$ is the set of all other parents of V_i.

A counterfactual $V_i(\pi, a, a')$ is said to be *edge inconsistent* if counterfactuals of the form $V_j(a_k, \ldots)$ and $V_j(a'_k, \ldots)$ occur in $V_i(\pi, a, a')$, otherwise it is said to be *edge consistent*. It is well known that a joint distribution $p(V(\pi, a, a'))$ containing an edge-inconsistent counterfactual $V_i(\pi, a, a')$ is not identified in the functional causal model (nor weaker causal models) with a corresponding graphical criterion on π and $\mathcal{G}(V)$ called the 'recanting witness' (Shpitser, 2013; Shpitser and Tchetgen Tchetgen, 2016). Under some assumptions, PSEs are nonparametrically identified by means of the *edge g-formula* described in Shpitser and Tchetgen Tchetgen (2016).

As an example, consider the DAG in Fig. 1(b). The PSE of A on Y along the paths $\pi = \{A \to Y, A \to L \to Y\}$ is encoded by a counterfactual contrast of the form $Y(\pi, a, a') = Y(a, M(a'), L(a, M(a')))$. This counterfactual density is identified by the edge g-formula as follows:

$$p(Y(a, M(a'), L(a, M(a')))) =$$
$$\sum_{X, M, L} p(Y \mid a, X, M)\, p(L \mid a, M, X)\, p(M \mid a', X)\, p(X).$$

For more details on PSEs, see Shpitser (2013) and Shpitser and Sherman (2018), and Nabi et al. (2018).

2.2 Algorithmic Fairness via Constraining Path-Specific Effects

There has been a growing interest in the issue of fairness in machine learning (Pedreshi et al., 2008; Feldman et al., 2015; Hardt et al., 2016; Kamiran et al., 2013; Corbett-Davies et al., 2017; Jabbari et al., 2017; Kusner et al., 2017b; Zhang and Bareinboim, 2018; Zhang et al., 2017). In this paper, we adopt the causal notion of fairness

described in Nabi and Shpitser (2018) and Nabi et al. (2019), where unfairness corresponds to the presence of undesirable or impermissble path-specific effects of sensitive attributes on outcomes – a view which generalizes an example discussed in Pearl (2009). We provide a brief summary of their perspective on fairness in the following without defending it for lack of space; see Nabi and Shpitser for more details.

Consider an observed data distribution $p(Y, Z)$ induced by a causal model, where Y is an outcome and $Z = \{X, A, M\}$ includes all baseline factors X, sensitive features A, and post-treatment pre-outcome mediators M. Context and background ethical considerations pick out some path-specific effect of the sensitive feature A on the outcome Y as unfair; we assume this effect is identified as a functional $g(p(Y, Z))$. Fix upper and lower bounds ϵ_l, ϵ_u for the PSE, representing a tolerable range. The most relevant bounds in practice are $\epsilon_l = \epsilon_u = 0$ or approximately zero. Nabi and Shpitser propose to transform the inference problem on $p(Y, Z)$, the "unfair world," into an inference problem on another distribution $p^*(Y, Z)$, called the "fair world," which is close in the sense of minimal KL-divergence to $p(Y, Z)$ while also having the property that the PSE lies within (ϵ_l, ϵ_u).

Given a dataset $\mathcal{D} = \{(Y_i, Z_i), i = 1, \ldots, n\}$ drawn from $p(Y, Z)$, a likelihood function $\mathcal{L}(\mathcal{D}; \alpha)$, an estimator $\widehat{g}(\mathcal{D})$ of the unfair PSE, and bounds ϵ_l, ϵ_u, Nabi and Shpitser suggest to approximate $p^*(Y, Z)$ by solving the following constrained maximum likelihood problem:

$$\widehat{\alpha} = \arg\max_{\alpha} \ \mathcal{L}_{Y,Z}(\mathcal{D}; \alpha),$$
$$\text{subject to} \ \ \epsilon_l \le \widehat{g}(\mathcal{D}) \le \epsilon_u. \quad (2)$$

Having approximated the fair world $p^*(Y, Z; \widehat{\alpha})$ in this way, Nabi and Shpitser point out a key difficulty for using these estimated parameters to predict outcomes for new instances (e.g., new job applicants). A new set of observations Z is not sampled from the "fair world" $p^*(Z)$ but from "unfair world" $p(Z)$. Nabi and Shpitser propose to map new instances from p to p^* and to use the result for predicting Y with model parameters $\widehat{\alpha}$. They assume Z can be partitioned into Z_1 and Z_2 such that $p^*(Y, Z) = p^*(Y, Z_1|Z_2)p(Z_2)$. In other words, variables in Z_2 are shared between p and p^*: $p^*(Z_2) = p(Z_2)$ but $p^*(Z_1|Z_2) \ne p(Z_1|Z_2)$. Z_1 typically corresponds to variables that appear in the estimator $\widehat{g}(\mathcal{D})$. There is no obvious principled way of knowing exactly what values of Z_1 the "fair version" of the new instance would attain. Consequently, all such possible values are averaged out, weighted appropriately by how likely they are according to the estimated p^*. This entails predicting Y as the expected value $\mathbb{E}^*[Y|Z_2]$ (with respect to the distribution $\sum_{Z_1} p^*(Y, Z_1|Z_2)$).
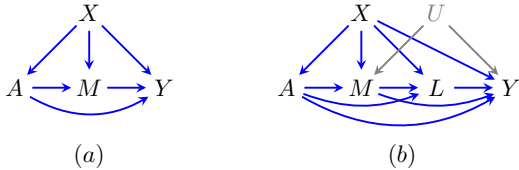
Figure 1: **(a)** A simple causal DAG, with treatment A, outcome Y, baseline variables X, and a mediator M. **(b)** A causal graph with two mediators M and L and unmeasured confounders captured in U.

Next, we explain some limitations of the inference procedure described here and present our main contributions to address these limitations.

3 FAIR PREDICTIVE MODELS IN A BATCH SETTING

Prediction problems in machine learning are typically tackled from the perspective of nonparametric risk minimization and the "train-and-test" framework. Here, we instead take the perspective of maximum likelihood and missing data, i.e., we treat unknown outcomes as missing values which we hope to impute in a way that is consistent with our specified likelihood for the entire data set. Our motivation for doing so is the nature of our constrained prediction problem. Specifically, our causal constraints contain "nuisance" components (conditional expectations and conditional distributions derived from the observed data distribution) which must be modeled correctly to ensure the causal effects are reliably estimated. In the subsequent prediction step, we should predict in a way that is consistent with what has already been modeled – or else we fail to exploit all the information we have already committed to in the constraint estimation step. We chose the maximum likelihood framework as the most natural and simplest approach to accomplish this. Alternative methods for coherently combining nuisance estimation with nonparametric risk minimization are left to future work.

Unlike Nabi and Shpitser (2018), we consider a batch prediction setting – this allows us to avoid the inefficient averaging described in the previous section. In our case, historical data (of sample size n_1) consists of observations on $\{X, A, M, Y\}$ and new instances (of size n_2) comprise a set of observations with just $\{X, A, M\}$. The outcome labels for new instances are missing data which we aim to predict, subject to fairness constraints. Instead of training our constrained model on historical data alone, we train on the combination of historical data and new instances. This seems complicated since the observed

data likelihood for the combined data set includes some complete rows and some partially incomplete rows. However, we can borrow ideas from the literature on missing data to accomplish this task. Specifically, we can impute missing outcomes ("labels") using appropriate functions of observed data. In this paper we assume the labels are *missing at random* (MAR) (Little and Rubin, 2002). However, our methods extend to any identifiable missing not at random (MNAR) model. Let the random variable R denote the missingness status of the outcome variable Y for each instance. That is, $R = 1$ for all rows in the historical data (since Y is observed) and $R = 0$ for all rows in the new instances. Then the observed data likelihood is $\prod_{i=1}^{n=n_1+n_2} p(X_i, A_i, M_i) \ p(Y_i|X_i, A_i, M_i)^{R_i}$.

This likelihood function describes the probability of the entire data set, though only uses Y values from historical data. We can then maximize the likelihood subject to the specified path-specific constraints, and associate predicted values $\hat{Y}_{new}$ to the new instances. Note that the setting where new instances arrive sequentially one-at-a-time is a special case of this general setup, which would require retraining on the full combined data after the arrival of each instance. Though this is computationally more intensive than the proposal in Nabi and Shpitser (2018) (where they only train once), it will deliver significantly more accurate predictions because it uses all available information. We will elaborate on this point in Section 4.

The approach to fair prediction outlined in Nabi and Shpitser (2018) suffers from two problems: one general and one specific to our setting here. First, their approach requires solving a computationally challenging constrained optimization problem. Likelihood functions are not in general convex and the constraints on path-specific effects involve nonlinear and complicated functionals of the observed data distribution. This makes the proposed constrained optimization a daunting task that relies on complex optimization software (or computationally expensive methods such as rejection sampling), which does not always find high quality local optima. Second, Nabi and Shpitser propose to constrain only part of the likelihood. Specifically they do not constrain the density $p(X)$ over the baseline features (since this is high-dimensional and thus inplausible to model accurately in their parametric approach). The baseline density is instead estimated by placing $1/n$ mass at every observed data point. This is sub-optimal in the specific setting we consider, where we do not need to average over constrained variables. Constraining a larger part of the joint should lead to a fair world distribution KL-closer to the observed distribution, which leads to better predictive performance as long

as the likelihood is correctly specified. This intuition is formalized in the following result:

Theorem 1 *Let $p(Z)$ denote the observed data distribution, $M_1 = \{p_1^*(Z) = \arg\max_{q(Z)} D_{KL}(p\|q) \ s.t. \ \epsilon_l \leq g(q(Z)) \leq \epsilon_u, \ q(Z_1) = p(Z_1)\}$, and $M_2 = \{p_2^*(Z) = \arg\max_{q(Z)} D_{KL}(p\|q) \ s.t. \ \epsilon_l \leq g(q(Z)) \leq \epsilon_u, \ q(Z_2) = p(Z_2)\}$. If $Z_2 \subseteq Z_1 \subseteq Z$, then $D_{KL}(p\|p_2^*) \leq D_{KL}(p\|p_1^*)$.*

In other words, if a larger part of the joint is being constrained in M_2 compared to M_1, then $p_2^*(Z)$ is at least as close to $p(Z)$ as $p_1^*(Z)$.

To address the first difficulty, we provide a novel reparameterization of the observed data likelihood such that the causal parameter corresponding to the unfair PSE appears directly in the likelihood. This approach generalizes previous work on reparameterizations implied by structural nested models (Robins, 1999; Tchetgen Tchetgen and Shpitser, 2014) to apply to a wide class of PSEs. With such a reparameterization, the MLE with a constrained PSE simply corresponds to maximum likelihood inference in a submodel where a certain likelihood parameter is set to 0. This type of inference can be implemented with standard software.

To address the second difficulty, we propose an approach to constraining the density $p(X)$. An alternative to fully parametric modeling is to consider nonparametric representations of $p(X)$. It is well known that the nonparametric maximum likelihood estimate of any $p(X)$ given a set of i.i.d draws is the empirical distribution which places mass $1/n$ at every observed point. Empirical likelihood methods have been developed for settings where the nonparametric and parametric (hybrid) likelihood must be maximized subject to moment constraints (Owen, 2001). We describe below how these methods may be adapted to our setting.

Finally, we show how both the reparameterization method and the empirical likelihood method can be combined to yield a constrained optimization method that maximizes a semi-parametric (hybrid reparameterized) likelihood using standard software.

4 EFFICIENT APPROXIMATION OF FAIR WORLDS

4.1 Imposing Fairness Constraints With Reparameterized Likelihoods

In this section, we describe how to reparameterize the observed data likelihood in terms of causal parameters that correspond to the effect of A on Y along cer-

tain causal pathways. The results presented in the following theorem greatly simplifies the constrained optimization problem shown in (2) in settings where the PSE includes the direct influence of A on Y. This is due to the fact that the constrained parameter, corresponding to the PSE of interest, now appears as a single coefficient in the outcome regression model.

Theorem 2 *Assume the observed data distribution $p(Y, Z)$ is induced by a causal model, where $Z = \{X, A, M\}$ includes pre-treatment measures X, treatment A, and post-treatment pre-outcome mediators M. Let $p(Y(\pi, a, a'))$ denote the potential outcome distribution that corresponds to the effect of A on Y along proper causal paths in π, where π includes the direct influence of A on Y, and let $p(Y_0(\pi, a, a'))$ denote the identifying functional for $p(Y(\pi, a, a'))$ obtained from the edge-formula in (1), where the term $p(Y|Z)$ is evaluated at $\{Z \setminus A\} = 0$. Then $\mathbb{E}[Y|Z]$ can be written as follows:*

$$\mathbb{E}[Y|Z] = f(Z) - (\mathbb{E}[Y(\pi, a, a')] - \mathbb{E}[Y_0(\pi, a, a')]) + \phi(A),$$

where $f(Z) := \mathbb{E}[Y|Z] - \mathbb{E}[Y|A, \{Z \setminus A\} = 0]$ and $\phi(A) = w_0 + w_a A$. Furthermore, w_a corresponds to π-specific effect of A on Y.

To illustrate the above reparameterization, consider the graph in Fig. 1(b), discussed in Nabi and Shpitser (2018) and Chiappa (2019). Assume the direct path and the paths through M of A on Y are the impermissible pathways (depicted with green edges). The corresponding PSE is encoded by a counterfactual contrast with respect to $Y(a, M(a), L(a', M(a)))$. The reparameterization in Theorem 2 amounts to:

$$\mathbb{E}[Y|Z] = f(Z) - \sum_{Z \setminus A} \big\{ f(Z) \times p(L|M, X, A = 0) \times$$
$$p(M|X, A = 1) \times p(X) \big\} + w_0 + w_a A, \quad (3)$$

where w_a represents the PSE of interest; see the appendix for more details. A special case of this reparameterization when π includes only the direct edge $A \rightarrow Y$ is implicit in the work of Tchetgen Tchetgen and Shpitser (2014).

Under linearity assumptions, the PSE of interest in Fig. 1(b) has a simple form. Assume the data generating process in Fig. 1(b) is the same as the one given in display (2) of Chiappa (2019), where PSE = $\theta_a^y + \theta_m^y \theta_a^m + \theta_l^y \theta_m^l \theta_a^m$. In this case, our reparameterization takes the following form:

$$\mathbb{E}[Y|X, A, M, L] = \underbrace{\left(\theta_x^y X + \theta_m^y M + \theta_l^y L \right)}_{f(Z)} -$$

$$\underbrace{\left(\left(\theta_0^m \theta_m^y + (\theta_0^l + \theta_m^l \theta_0^m) \theta_l^y \right) + (\theta_m^y \theta_a^m + \theta_l^y \theta_m^l \theta_a^m) A \right)}_{\sum_{Z \setminus A} \left\{ f(Z) p(L|M, X, A=0) p(M|X, A=1) p(X) \right\}} +$$

$$\underbrace{\left(\theta_0^y + (\theta_0^m \theta_m^y + (\theta_0^l + \theta_m^l \theta_0^m) \theta_l^y) \right)}_{w_0} + \underbrace{\left(\theta_a^y + \theta_m^y \theta_a^m + \theta_l^y \theta_m^l \theta_a^m \right)}_{w_a \equiv \text{PSE}} A$$

In order to move away from the linear setting and exploit more flexible techniques, Chiappa (2019) makes assumptions on the latent variables. However, such assumptions are often hard to verify in practice. In contrast, our result is entirely nonparametric and does not rely on any assumptions beyond what is encoded in the causal DAG.

By Theorem 2, the constrained optimization problem in eq. (2) simplifies significantly to the following optimization problem:

$$\widehat{\alpha} = \arg\max_{\alpha} \ \mathcal{L}_{Y,Z}(\mathcal{D}; \alpha) \quad \text{subject to} \ \ w_a = 0. \quad (4)$$

In the prediction setting, i.e., finding optimal parameters for $\mathbb{E}[Y|Z; \alpha_y]$, this amounts to an unconstrained maximum likelihood problem with outcome regression taking the specific form:

$$\mathbb{E}[Y|Z; \alpha_y] = \quad (5)$$
$$f(Z; \alpha_f) - \sum_{X,M} f(Z; \alpha_f) \ p(M|A = 0, X; \alpha_m) \ p(X) + w_0,$$

where $f(Z) := \mathbb{E}[Y|Z] - \mathbb{E}[Y|A, X = 0, M = 0]$ and is parameterized by α_f. In practice, for each X_i in the data $p(X_i)$ is replaced with its empirical approximation $1/n$, since a parametric specification of $p(X)$ is not feasible. In next section, we explain how $p(X)$ can be incorporated into the constrained optimization problem using empirical likelihood methods.

4.2 Imposing Fairness Constraints With Hybrid Likelihoods

In light of Theorem 1, we are interested in constraining the nonparameteric form of $p(X)$. Following work in Owen (2001), we use hybrid/semi-parametric empirical likelihood methods to estimate $p(X)$ nonparametrically which is a novel idea in the fairness setting. First, according to Theorem 1, constraining $p(X)$ would bring our learned distribution closer to the observed (unfair) distribution, and hence results in improvement of model performance, as we demonstrate in our simulations. Second, $p(X)$ is often a high dimensional object that is difficult to estimate due to the curse of dimensionality. For simplicity of presentation, we focus on the DAG in Fig. 1(a), and the constraint represented by the NDE, although the methods we describe generalize without difficulty to arbitrary causal models and constraints represented by arbitrary PSEs.

Let $(X_i, A_i, M_i, Y_i), i = 1, \ldots, n$ be independent random vectors with common distribution $p = p(X) \times p(A, M, Y|X)$. We assume a known parametric form for $p(A, M, Y|X)$, and leave $p(X)$ unrestricted. Assuming the unfair effect is NDE, the only constraint on the observed distribution is for NDE to be zero. Let $p(Y|M, A, X), p(M|A, X), p(A|X)$ be parameterized by $\alpha_y, \alpha_m, \alpha_a$, respectively. The direct effect can

then be identified by $\mathbb{E}_x[m(X; \alpha)]$, where

$$m(X; \alpha) = \sum_M \Big\{ \mathbb{E}[Y|A = 1, M, X; \alpha_y] - \quad (6)$$
$$\mathbb{E}[Y|A = 0, M, X; \alpha_y] \Big\} p(M|A = 0, X; \alpha_m).$$

The *profile empirical likelihood ratio* parameters $(\{p_i, \widehat{\alpha}\}^{opt})$ are then given by

$$\arg\max_{p_i, \alpha} \prod_{i=1}^{n} p_i p(Y_i|M_i, A_i, X_i; \alpha_y) p(M_i|A_i, X_i; \alpha_m) p(A|X_i; \alpha_a)$$
$$\text{such that} \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \ m(X_i; \alpha) = 0 \quad (7)$$

The above optimization problem involves a semi-parametric *hybrid* likelihood (Owen, 2001), that contains both nonparametric and parametric terms. In order to solve the above optimization problem (formulated on both α and p_i parameters), we can apply the Lagrange multiplier method and solve its dual form (formulated on both α and the Lagrange multipliers); see the appendix for more details. Empirical likelihood methods provide a natural extension to imposing constraints on arbitrary PSEs, since these can be written in the form of $\mathbb{E}_x[m(X; \alpha)]$ for some $m(\cdot)$.

If outcomes are missing at random, the NDE is identified by $\mathbb{E}_x[m(X; \alpha)]$, where

$$m(X; \alpha) = \sum_M \Big\{ \mathbb{E}[Y|A = 1, M, X, R = 1; \alpha_y] - $$
$$\mathbb{E}[Y|A = 0, M, X, R = 1; \alpha_y] \Big\} p(M|A = 0, X; \alpha_m).$$

The resulting functional is then used in the profile empirical likelihood in (7).

Unlike the standard unconstrained prediction setting where it is common to use nonparametric methods to estimate an arbitrary regression function, our task requires a combination of prediction and estimation of the relevant causal parameter (constraint). Estimating the causal parameter requires estimating certain nuisance components (like $E[Y|Z]$ in eq. 6) which we choose to do parametrically in part because we desire certain frequentist properties, namely fast rates of convergence. More fundamentally, the empirical likelihood optimization problem in (7) finds optimal parameter values α, where α appears also in the constraint $\sum_i p_i m(X_i; p_i, \alpha) = 0$. That is, the structure of the empirical likelihood optimization problem requires that Y and M models are specified parametrically. Though some combination of nonparametric risk minimization and empirical likelihood would be an interesting extension, how to accomplish this is an open question.

4.3 Imposing Fairness Constraints With Hybrid Reparameterized Likelihoods

In Section 4.1, we reformulated the constrained optimization problem of interest by rewriting the likelihood in terms of the parameters we were interested in constraining, and directly setting those parameters to zero. However, we did not place any constraints on $p(X)$. In Section 4.2, we used hybrid likelihoods to constrain a nonparametric estimate of $p(X)$, but did not provide a convenient reparameterization of the likelihood in terms of relevant parameters. In this section we describe an approach to optimizing a *hybrid reparameterized likelihood* that combines the advantages of both proposals. This allows us to constrain the entire likelihood and do so with standard maximum likelihood software, since the constraint we must satisfy directly corresponds to a parameter in the hybrid likelihood.

For simplicity of presentation, we again focus on the constraining the NDE, although the methods we describe generalize without difficulty to arbitrary constraints represented by arbitrary PSEs. The direct effect can then be estimated by $\mathbb{E}_x[m(X;\alpha)]$, where $m(X;\alpha)$ is given in (6), and $\mathbb{E}[Y|A, M, X; \alpha_y]$ is given in (5). Assuming $p(X_i = x_i) = p_i$ as in (5), $m(X;\alpha)$ will be a function of p_is as well. The *profile empirical likelihood ratio* $(\{p_i, \widehat{\alpha}\}^{opt})$ in this setting is then given by

$$\arg\max_{p_i,\alpha}\prod_{i=1}^{n}p_i p(Y_i|M_i,A_i,X_i;\alpha_y)p(M_i|A_i,X_i;\alpha_m)p(A|X_i;\alpha_a)$$

$$\text{such that} \quad \sum_{i=1}^{n}p_i = 1, \quad \sum_{i=1}^{n}p_i\, m(X_i;p_i,\alpha) = 0 \quad (8)$$

Unlike the constrained optimization problem in (7), it is not straightforward to find the dual form of the optimization problem in (8), which is a standard approach for solving such problems in the empirical likelihood literature. The reason is that p_i appears in multiple places in the constraint corresponding to setting PSE to zero; that is $m(X;\alpha)$ is now a function of both α and p_i's. As an alternative, we provide a heuristic approach for optimizing (8) via an iterative procedure that starts with initialization of α and p_is, and at the kth iteration updates the values for α^k and p_i^ks by treating $m(X_i;p_i,\ \alpha)$ as a function of $m(X_i;p_i^{k-1},\alpha)$. The procedure terminates when the difference between the two updates is sufficiently small. In Algorithm 1, we provide a detailed description of our proposed iterative procedure to address this issue, which behaves well in experiments.

5 EXPERIMENTS

Given Theorem 1, the accuracy of the prediction procedure depends on what components of $p(Z,Y;\alpha)$ are

---

Algorithm 1 HYBRID REPARAMETERIZED LIKELIHOOD

Input: $\mathcal{D} = \{X_i, A_i, M_i, Y_i\}, i = 1, \ldots, n$ and specification of a PSE of the form $\mathbb{E}_X[m(X;\alpha)]$.

Output: $\widehat{\alpha}, \widehat{p}_i$ by solving

$$\mathcal{R}(p_i, \alpha) = \max \sum_{i=1}^{n}\Big(\log p_i + \log p(Y_i, M_i, A_i|X_i;\alpha)\Big)$$

$$\sum_{i=1}^{n}p_i\, m\big(X_i,;p_i,\alpha\big) = 0, \ \sum_{i=1}^{n}p_i = 1.$$

1: Pick starting values for $p_i^{(1)}$ and $\alpha^{(1)}$.
2: At k^{th} iteration, given fixed $p_i^{(k-1)}$ and $\alpha^{(k-1)}$, estimate the following (in order)

   I. $m\Big(X_i; \{p_i^{(k-1)}\}, \alpha^{(k-1)}\Big)$

   II. λ by solving $\quad \sum_{i=1}^{n}\frac{m(X_i;\theta)}{1+\lambda\ m(X_i;\theta)} = 0$, which is a monotone function in λ.

   III. $p_i^{(k)}$ using $p_i^{(k)} = \frac{1}{n}\frac{1}{1+\lambda m(X_i;\theta)}, \forall i = 1, \ldots, n,$

   IV. $\alpha^{(k)}$ by maximizing the following

$$\widehat{\alpha}^{(k)} = \arg\max_{\alpha}\ \mathcal{L}_{Y,M,A|X}(\mathcal{D};\alpha)$$

$$\text{subject to}\ \ w_a = 0,$$

   where $\mathbb{E}[Y|X, A, M; \alpha_y] = w_0 + f(Z;\alpha_f) - \sum_{i=1}^{n}\{\sum_m f(Z_i;\alpha_f)p(M|A = 0, X_i;\alpha_m)\}p_i^{(k)}$ and $f(Z) := \mathbb{E}[Y|X, A, M] - \mathbb{E}[Y|X = 0, A, M = 0].$
3: Repeat Step (2) until convergence.

---

constrained, and following Nabi and Shpitser (2018) this depends on the chosen estimator $\widehat{g}(\mathcal{D})$. Here, we illustrate this dependence via experiments by considering four consistent estimators of the NDE presented in Tchetgen Tchetgen and Shpitser (2012) (assuming the model shown in Fig. 1(a) is correct). We fit models $E[Y|A, M, X; \alpha_y]$, $p(M|A, X; \alpha_m)$, and $p(A|X; \alpha_a)$ by maximum likelihood. The first estimator (G-formula), is the MLE plug-in estimator and uses Y and M models to estimate NDE. The second one is the inverse probability weighted (IPW) estimator that uses A and M models. The "mixed" estimator uses the A and Y models, and the augmented IPW estimator (AIPW) uses all three models. See the appendix for details on these estimators.

We generated a sample of size $6,000$ using the data generating process described in the appendix. We approximate the fair world, p^*, by constraining MLE given in Section 2. We estimated the NDE using the four methods described above and evaluated the performance of the approximated p^* for each case. In Table 1 we show the estimated NDE with respect to p^*, the log likelihood, KL-divergence between p^* and

Table 1: Comparing different versions of p^* estimated by constraining different parts of the likelihood.

Method	Estimator	Direct Effect	Log $\mathcal{L}$	$D_{KL}(p||p^*)$	MSE
Constrained MLE	G-formula	0.000	-53359	0.4220	3.5045
	IPW	0.000	-51695	0.1146	5.0409
	Mixed	0.000	-52754	0.2993	2.7505
	AIPW	0.000	-51630	0.0935	4.9545
Unconstrained MLE	G-formula	2.235	-51160	0.0018	0.9331

Table 2: Evaluating different p^* estimation methods by KL divergence and predictive accuracy (MSE).

Likelihood Method	Direct Effect	$D_{KL}(p||p^*)$	MSE
M_0: Unconstrained MLE	2.1946	0.000	1.002
M_1: Constrained MLE (sec. 2.2)	0.000	0.415	3.546
M_2: Reparameterized MLE (sec. 4.1)	0.000	0.282	3.410
M_3: Hybrid MLE (sec. 4.2)	0.034	0.352	1.165
M_4: Hybrid reparameterized MLE (sec. 4.3)	0.019	0.284	1.541

p, and the mean squared error between the observed outcomes and the predicted ones. We contrast these results with the unconstrained prediction model. Unconstrained MLE is KL-closest to the true distribution and yields the lowest MSE, as expected. However, it suffers from being unfair: NDE = 2.235. AIPW produces the second closest approximation to the true distribution while being fair. However, the MSE under AIPW is relatively large, since new instances are being mapped and more information are averaged out from the predictions. The approximated fair distributions under the other three estimators are KL-farther from the true distribution, and the accuracy of prediction varies, underscoring how the performance of the learned prediction model depends strongly on what part of the information is being averaged out and what estimator is being used.

Next, we illustrate that even in simple settings our proposed methods for solving constrained maximum likelihood problems considerably outperform the existing method described in Nabi and Shpitser (2018). We will use continuous outcomes for simplicity, but our results are not substantially affected if outcomes are discrete. We generated synthetic data ($n = 7000$ with 20% missing outcomes) according to the causal model shown in Fig. 1(a), where A, M are binary and X, Y are continuous variables. The model specification details are reported in Appendix D and the code is attached to the submission. For illustration purposes, we assume that the direct effect of sensitive feature A on outcome Y is unfair and estimate it via the g-formula. We approximate the fair world, p^*, by constrained MLE using the three methods described in Section 4 and contrast them with the constrained MLE described in Section 2 as well as unconstrained MLE. We evaluated the performance of all five methods by

computing the direct effect with respect to p^*, KL-divergence between p^* and p, and the mean squared error between the observed and predicted outcomes.

Results are displayed in Table 2 (averaged over 20 repetitions). We see that all three proposed methods achieve an approximatation to the fair distribution p^* KL-closer to the true unfair distribution p, compared to standard constrained MLE. Using the reparameterized MLE by itself requires averaging over the constrained covariates as in Nabi and Shpitser (2018), so there is only minimal improvement in prediction accuracy (measured by MSE). However, the last two methods involve prediction in batch mode as described above – that is, use all information in the data – and so can achieve substantial improvements in prediction accuracy.

6 CONCLUSION

Imposing hard fairness constraints on predictive models involves a balance of parametric modeling, nonparametric methods, and constrained optimization. In this paper we have proposed two innovations to make the problem easier and make predictions more accurate: a reparameterization of the likelihood such that nonlinear constraints appear explictly as likelihood parameters constrained to be zero and an incorporation of techniques from empirical likelihood theory to make the constrained distribution closer to the unconstrained unfair distribution. Our simulations show that even in a relatively simple setting, we can improve significantly on prior proposals, achieving prediction performance comparable to unconstrained (unfair) maximum likelihood, particularly with the hybrid approach. Though we focus primarily on the path-specific fairness constraints proposed

in Nabi and Shpitser (2018), the ideas presented here should be applicable more broadly to fair prediction proposals that require imposing constraints on predictive models. At this stage, our method which combines reparameterization with hybrid likelihood is somewhat heuristic; in future work, we hope to develop an approach for optimizing EL weights and likelihood parameters jointly without the need for iteration.

References

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances In Neural Information Processing Systems*, pages 3315–3323, 2016.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, , and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of International Conference on Machine Learning*, 2017.

Faisal Kamiran, Indre Zliobaite, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, 2013.

Teus H Kappen, Wilton A van Klei, Leo van Wolfswinkel, Cor J Kalkman, Yvonne Vergouwe, and Karel GM Moons. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research*, 2(1):11, 2018.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine learning algorithms. *Journal of Banking & Finance*, 34:2767–2787, 2010.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017a.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances In Neural Information Processing Systems*, 2017b.

Rodrick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002.

Shira Mitchell, Eric Potash, and Solon Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Razieh Nabi, Phyllis Kanki, and Ilya Shpitser. Estimation of personalized effects associated with causal pathways. In *Proceedings of the Thirty Fourth Conference on Uncertainty in Artificial Intelligence (UAI-34th)*. AUAI Press, 2018.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Art Owen. *Empirical Likelihood*. Chapman & Hall, 2001.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2001.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

Walter L. Perry, Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood. Predictive policing: The role of crime forecasting in law enforcement operations. RAND Corporation, `http://www.rand.org/pubs/research_reports/RR233.html`, 2013.

Jing Qin. *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer, 2017.

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*. NY: Springer-Verlag, 1999.

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science (Rumelhart special issue)*, 37: 1011–1035, 2013.

Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence*, 2018.

Ilya Shpitser and Eric J. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466, 2016.

Eric J. Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 2012.

Eric J. Tchetgen Tchetgen and Ilya Shpitser. Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864, 2014.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Association for the Advancement of Artificial Intelligence*, 2018.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3929–3935, 2017.

APPENDIX

In **Appendix A**, we provide additional details for the direct effect reparameterization example (illustrating Theorem 2) discussed in the main paper. In **Appendix B**, we provide a brief overview of empirical likelihood methods and some additional theoretical details useful for understanding our proposed hybrid likelihood approach. In **Appendix C**, we state the statistical modeling assumptions we made in our simulation experiments. In **Appendix D**, we give some relevant details for the first simulation reported in the main paper. **Appendix E** contains proofs of our theorems. For a clearer presentation of materials in this supplement, we use a one-column format.

A. Reparameterized Likelihood Example: Additional Details

Consider the DAG in Fig. 1(a), and assume the natural direct effect is the unfair PSE we wish to constrain to be 0. Theorem 2 leads to the following reparameterization of the regression function:

$$\mathbb{E}[Y \mid X, A, M] = \underbrace{\mathbb{E}[Y \mid X, A, M] - \mathbb{E}[Y \mid X = 0, A, M = 0]}_{f(X,A,M)}$$
$$- \sum_{X,M} f(X, A, M) \; p(M \mid A = 0, X) \; p(X)$$
$$+ \underbrace{\sum_{X,M} \mathbb{E}[Y \mid X, A, M] \; p(M \mid A = 0, C) \; p(X).}_{\phi(A) = w_0 + w_a A}$$

The coefficient w_a corresponds to the direct effect, since

$$\text{NDE} = \sum_{X,M} \Big\{ \mathbb{E}[Y \mid X, A = 1, M] - \mathbb{E}[Y \mid X, A = 0, M] \Big\} \; p(M \mid A = 0, X) \; p(X)$$
$$= \sum_{X,M} \mathbb{E}[Y \mid X, A = 1, M] \; p(M \mid A = 0, X) \; p(X) - \sum_{X,M} \mathbb{E}[Y \mid X, A = 0, M] \; p(M \mid A = 0, X) \; p(X)$$
$$= \phi(A = 1) - \phi(A = 0)$$
$$= w_a.$$

The observed data likelihood is given by

$$\mathcal{L}_{Y,M,A,X}(\mathcal{D}; \alpha) = \prod_{i=1}^{n} p(Y_i \mid M_i, A_i, X_i; \alpha_y) \; p(M_i \mid A_i, X_i; \alpha_m) \; p(A_i \mid X_i; \alpha_a) \; p(X_i),$$

where $p(Y \mid M, A, X; \alpha_y)$ has mean

$$\mathbb{E}[Y \mid X, A, M; \alpha_y] = f(X, A, M; \alpha_f) - \sum_{x,m} f(X, A, M; \alpha_f) \; p(M \mid A = 0, X; \alpha_m) \; p(X) + w_0.$$

The constrained optimization problem in eq. (2) then simplifies to the following optimization problem:

$$\arg \max_{\alpha} \quad \mathcal{L}_{Y,M,A,X}(\mathcal{D}; \alpha)$$
$$\text{subject to} \quad w_a = 0.$$

B. Hybrid Likelihood: Overview and Details

Empirical Likelihood

We briefly review empirical likelihood methods, described in detail in Owen (2001). Let $X_1, \ldots, X_n$ be independent random vectors with common distribution F_0. Let F be any CDF, where $F(x) = p(X \leq x)$, and F_n be the empirical distribution. Suppose that we are interested in F through $\theta = T(F)$, where T is a real-valued function of the distribution. The true unknown parameter is $\theta_0 = T(F_0)$. Proceeding by analogy to parametric MLE, the

non-parametric MLE of θ is $\hat{\theta} = T(F_n)$. The nonparametric likelihood ratio, $R(F) = \frac{\mathcal{L}(F)}{\mathcal{L}(F_n)}$, is used as a basis for hypothesis testing and deriving confidence intervals. The **profile likelihood ratio** function is defined as

$$\mathcal{R}(\theta) = \sup \big\{ R(F) \mid T(F) = \theta, F \in \mathcal{F} \big\},$$

where $\mathcal{F}$ denotes the set of all distributions on $\mathbb{R}$.

Often, $\theta \equiv \theta(F)$ is the solution to an estimating equation of the form $\mathbb{E}[m(X, \theta)] = 0$. A natural estimator for θ is produced by solving the empirical estimating equation $\frac{1}{n} \sum_{i=1}^{n} m(X_i, \widehat{\theta}) = 0$. Assuming $p_i = f(X = x_i)$ for $i = 1, \ldots, n$, the **profile empirical likelihood ratio** function of θ is defined as

$$\mathcal{R}(\theta) = \max \Big\{ \prod_{i=1}^{n} n p_i \quad \text{such that} \quad \sum_{i=1}^{n} p_i\, m(X_i, \theta) = 0,\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1 \Big\}. \tag{9}$$

Since maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood, the profile empirical likelihood ratio is rewritten in terms of log likelihood as follows.

$$\mathcal{R}(\theta) = \max \Big\{ \sum_{i=1}^{n} \log p_i \quad \text{such that} \quad \sum_{i=1}^{n} p_i\, m(X_i, \theta) = 0,\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1 \Big\}. \tag{10}$$

In order to solve the above optimization problem, we can apply the Lagrange multiplier method.

$$T(\{p_i\}, \lambda, \lambda_1) = \sum_{i=1}^{n} \log p_i + \lambda_1 \big(\sum_{i=1}^{n} p_i - 1 \big) - n\lambda \sum_{i=1}^{n} p_i\, m(X_i; \theta),$$

where λ, λ_1 are the Lagrange multipliers. We take the derivative of $T(\{p_i\}, \lambda, \lambda_1)$, with respect to the p_i's, and set them to zero. Solving the system of equations reveals that $\lambda_1 = -n$, and

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda m(X_i; \theta)}, \forall i = 1, \ldots, n, \tag{11}$$

where λ is the solution to

$$\sum_{i=1}^{n} \frac{m(X_i; \theta)}{1 + \lambda\, m(X_i; \theta)} = 0, \tag{12}$$

which is a monotone function in λ. Maximizing the profile empirical log-likelihood ration in (10) is equivalent to maximizing the following (substituting p_i from (11) into (10)):

$$l(\theta) = -\sum_{i=1}^{n} \log(1 + \lambda\, m(X_i; \theta)) - n \log n. \tag{13}$$

Maximizing $l(\theta)$ over a small set of parameters θ, is a much simpler optimization problem than maximizing (10) over n unknowns. Equation 13 is known as the dual representation of 10. See Owen (2001) for more details.

Hybrid Likelihood

Now, consider independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. Suppose that all n observations are independent, and that we have a correctly specified parametric model for $p(Y|X; \theta_y)$ but $p(X)$ is unspecified. Let $p_i = p(X = x_i)$. A natural approach for estimating θ_y and the p_is is to form a **hybrid** likelihood that is nonparametric in the distribution of X_i but is parametric in the conditional distribution of $Y_i|X_i$:

$$\mathcal{L}(\mathcal{D}; \{p_i\}, \theta) = \prod_{i=1}^{n} p_i\, p(Y_i|X_i; \theta).$$

Suppose we are interested in parameter θ through the estimating equation $\mathbb{E}[m(X, Y; \theta)] = 0$. Hence, the equivalent form of (10) for the profile hybrid likelihood ratio function is as follows:

$$\mathcal{R}(\theta) = \max \Big\{ \sum_{i=1}^{n} \Big(\log p_i + \log p(Y_i|X_i; \theta) \Big) \quad \text{such that} \quad \sum_{i=1}^{n} p_i\, m(X_i, Y_i; \theta) = 0,\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1 \Big\}. \tag{14}$$

Similar to the empirical likelihood, we can apply the Lagrange multiplier method to solve the above optimization problem. For more details, see Owen (2001) and Qin (2017).

C. Simulation details

Here we report the precise parameter settings used in our simulation studies. We trained our models on a batch size of $7,000$ using the following data generating process, where outcome Y is treated as missing on 20% of the data. Mean squared errors in Tables 1 and 2 are computed only on the missing portion of the outcome Y.

$$
\begin{aligned}
X &\sim \mathcal{N}(0,1) \\
\operatorname{logit}(p(A=1|X)) &\sim 0.5 + 0.5X \\
\operatorname{logit}(p(M=1|A,X)) &\sim 0.5 + X + 0.5A - AX \\
Y &= 1 + X + 2A - 2AX + M + 3XM + AM + XAM + \mathcal{N}(0,1)
\end{aligned}
\tag{15}
$$

D. Details on Estimation Strategies

Given Theorem 1, the accuracy of the prediction procedure will depend on what parts of $p(Z,Y;\alpha)$ are constrained, and following Nabi and Shpitser (2018) this depends on the estimator $\widehat{g}(\mathcal{D})$. Here, we define several consistent estimators of the NDE (assuming the model shown in Fig. 1(a) is correct) presented in Tchetgen Tchetgen and Shpitser (2012).

G-formula: The first estimator is the MLE plug in estimator, where we use the Y and M models to estimate NDE. We fit models $\mathbb{E}[Y|A,M,X;\alpha_y]$ and $p(M|A,X;\alpha_m)$ by maximum likelihood, and use the following formula:

$$
\mathbb{P}_n\left(\sum_m \left(\mathbb{E}[Y_i \mid A=1, X_i, M; \widehat{\alpha}_y] - \mathbb{E}[Y_i \mid A=0, X_i, M; \widehat{\alpha}_y] \right) p(M \mid A=0, X_i; \widehat{\alpha}_m) \right).
\tag{16}
$$

Since solving (2) using (16) entails constraining $\mathbb{E}[Y|A,M,X]$ and $p(M|A,X)$, classifying a new instance entails using $\mathbb{E}[Y|A,X] = \sum_M \mathbb{E}[Y|A,M,X] \, p(M|A,X)$.

Inverse probability weighting (IPW): The second estimator is the IPW estimator where we use the A and M models to estimate NDE. We can fit the models $p(A|X;\alpha_a)$ and $p(M|A,X;\alpha_m)$ by MLE, and use the following weighted empirical average as our estimate of the NDE:

$$
\mathbb{P}_n\left(\frac{\mathbb{I}(A_i=1)}{p(A_i=1|X_i;\widehat{\alpha}_a)} \cdot \frac{p(X_i|A=0,X_i;\widehat{\alpha}_m)}{p(M_i|A=1,X_i;\widehat{\alpha}_m)} \, Y_i - \frac{\mathbb{I}(A_i=0)}{p(A_i=0|X_i;\widehat{\alpha}_a)} \, Y_i \right).
\tag{17}
$$

Since solving the constrained MLE problem using this estimator entails only restricting parameters of A and M models, predicting a new instance is done using $\mathbb{E}[Y|X] = \sum_{A,M} \mathbb{E}[Y|A,M,X] \, p(M|A,X) \, p(A|X)$.

Mixed approach: The third way of computing the NDE is using A and Y models. In this estimator, we fit the models $p(A|X;\alpha_a)$ and $\mathbb{E}[Y|A,M,X;\alpha_y]$ by MLE, as usual, and combine the edge G-formula and IPW in the following way:

$$
\mathbb{P}_n\left(\frac{\mathbb{I}(A_i=0)}{p(A_i=0|X_i;\widehat{\alpha}_a)} \, \mathbb{E}[Y_i|A=1,M_i,X_i;\widehat{\alpha}_y] - \mathbb{E}[Y_i|A=0,M_i,X_i;\widehat{\alpha}_y] \right),
\tag{18}
$$

Since solving the constrained MLE problem using this estimator entails only restricting parameters of A and Y models, predicting a new instance is done using $\mathbb{E}[Y|M,X] = \sum_A \mathbb{E}[Y|A,M,X] \, \frac{p(M|A,X) \cdot p(A|X)}{\sum_A p(M|A,X) \cdot p(A|X)}$.

Augmented inverse probability weighting (AIPW): The final estimator uses all three models, as follows:

$$
\begin{aligned}
\mathbb{P}_n\bigg(&\frac{\mathbb{I}(A_i=1)}{p(A_i=1|X_i;\widehat{\alpha}_a)} \frac{p(M_i \mid A=0,X_i;\widehat{\alpha}_m)}{p(M_i|A=1,X_i;\widehat{\alpha}_m)} \Big\{ Y_i - \mathbb{E}[Y_i|A=1,M_i,X_i;\widehat{\alpha}_y] \Big\} \\
&+ \frac{\mathbb{I}(A_i=0)}{p(A_i=0|X_i)} \Big\{ \mathbb{E}[Y_i|A=1,M_i,X_i;\widehat{\alpha}_y] - \eta(1,0,X_i) \Big\} + \eta(1,0,X_i) \\
&- \frac{\mathbb{I}(A_i=0)}{p(A_i=0|X_i;\widehat{\alpha}_a)} \Big\{ Y_i - \eta(0,0,X_i) \Big\} + \eta(0,0,X_i) \bigg),
\end{aligned}
\tag{19}
$$

with $\eta(a,a',X) \equiv \sum_M \mathbb{E}[Y|a,M,X] p(M|a',X)$. Since the models of A, M, and Y are all constrained with this estimator, predicting Y for a new instance is via $\mathbb{E}[Y|X] = \sum_{A,M} \mathbb{E}[Y|A,M,X] \, p(M|A,X) \, p(A|X)$.

E. Proofs

Theorem 1 *Let $p(Z)$ denote the observed data distribution, $M_1 = \{p_1^*(Z) = \arg\max_{q(Z)} D_{KL}(p||q) \ s.t. \ \epsilon_l \leq g(q(Z)) \leq \epsilon_u, \ q(Z_1) = p(Z_1)\}$, and $M_2 = \{p_2^*(Z) = \arg\max_{q(Z)} D_{KL}(p||q) \ s.t. \ \epsilon_l \leq g(q(Z)) \leq \epsilon_u, \ q(Z_2) = p(Z_2)\}$. If $Z_2 \subseteq Z_1 \subseteq Z$, then $D_{KL}(p||p_2^*) \leq D_{KL}(p||p_1^*)$. In other words, if more densities are being constrained in M_2 compared to M_1, then $p_2^*(Z)$ is closer to $p(Z)$ than $p_1^*(Z)$.*

Proof: M_1 is a submodel of M_2, hence maximizing the likelihood under model M_1 yields a likelihood that is less than or equal to the one under model M_2: $\max \mathcal{L}_{M_1}(\mathcal{D}) \leq \max \mathcal{L}_{M_2}(\mathcal{D})$. Maximizing the likelihood of observed data with respect to the model parameters is equivalent to minimizing KL-divergence between the likelihood and the true distribution of the data Wasserman (2013). Consequently, KL-divergence between p^* and p is smaller in M_2 compared to M_1, i.e $D_{KL}(p||p_2^*) \leq D_{KL}(p||p_1^*)$. $\qquad\square$

Theorem 2 *Assume the observed data distribution $p(Y, Z)$ is induced by a causal model, where $Z = \{X, A, M\}$ includes pre-treatment measures X, treatment A, and post-treatment pre-outcome mediators M. Let $p(Y(\pi, a, a'))$ denote the potential outcome distribution that corresponds to the effect of A on Y along proper causal paths in π, where π includes the direct influence of A on Y, and let $p(Y_0(\pi, a, a'))$ denote the identifying functional for $p(Y(\pi, a, a'))$ obtained from the edge-formula in (1), where the term $p(Y|Z)$ is evaluated at $\{Z \setminus A\} = 0$. Then $\mathbb{E}[Y|Z]$ can be written as follows:*

$$\mathbb{E}[Y|Z] = f(Z) - (\mathbb{E}[Y(\pi, a, a')] - \mathbb{E}[Y_0(\pi, a, a')]) + \phi(A),$$

where $f(Z) := \mathbb{E}[Y|Z] - \mathbb{E}[Y|A, \{Z \setminus A\} = 0]$ and $\phi(A) = w_0 + w_a A$. Furthermore, w_a corresponds to π-specific effect of A on Y.

Proof: By letting $\phi(A = a) = \mathbb{E}[Y(\pi, a, a')]$, it suffices to show that $\mathbb{E}[Y_0(\pi, a, a')] = \mathbb{E}[Y|A, \{Z \setminus A\} = 0]$. Given the identification result for edge-consistent counterfactuals in Shpitser and Tchetgen Tchetgen (2016), we can write the identification functional as follows.

$$\mathbb{E}[Y_0(\pi, a, a')] = \sum_{V \in \mathfrak{X}_V \setminus \{A, Y\}} \mathbb{E}[Y|A = a, \{Z \setminus A\} = 0] \times h(V \in \mathfrak{X}_V \setminus Y),$$

where $h(V \in \mathfrak{X}_V \setminus Y)$ is a function of all variables excluding Y. Note that h, does not include any density where A appears on the LHS of the conditioning bar. Therefore, we have:

$$\mathbb{E}[Y_0(\pi, a, a')] = \mathbb{E}[Y|A = a, \{Z \setminus A\} = 0] \times \sum_{V \in \mathfrak{X}_V \setminus \{A, Y\}} h(V \in \mathfrak{X}_V \setminus Y)$$

$$= \mathbb{E}[Y|A = a, \{Z \setminus A\} = 0].$$

$\qquad\square$