Learning Optimal Fair Policies

Razieh Nabi [1] **Daniel Malinsky** [1] **Ilya Shpitser** [1]

Abstract

Systematic discriminatory biases present in our society influence the way data is collected and stored, the way variables are defined, and the way scientific findings are put into practice as policy. Automated decision procedures and learning algorithms applied to such data may serve to perpetuate existing injustice or unfairness in our society. In this paper, we consider how to make optimal but fair decisions, which "break the cycle of injustice" by correcting for the unfair dependence of both decisions and outcomes on sensitive features (e.g., variables that correspond to gender, race, disability, or other protected attributes). We use methods from causal inference and constrained optimization to learn optimal policies in a way that addresses multiple potential biases which afflict data analysis in sensitive contexts, extending the approach of Nabi & Shpitser (2018). Our proposal comes equipped with the theoretical guarantee that the chosen fair policy will induce a joint distribution for new instances that satisfies given fairness constraints. We illustrate our approach with both synthetic data and real criminal justice data.

1. Introduction

Making optimal and adaptive intervention decisions in the face of uncertainty is a central task in precision medicine, computational social science, and artificial intelligence. In healthcare, the problem of learning optimal policies is studied under the heading of *dynamic treatment regimes* (Chakraborty & Moodie, 2013). The same problem is called *reinforcement learning* in artificial intelligence (Sutton & Barto, 1998), and *optimal stochastic control* (Bertsekas & Tsitsiklis, 1996) in engineering and signal processing. In all of these cases, a policy (a function of historical data to some space of possible actions, or a sequence of such functions) is chosen to maximize some pre-specified outcome quantity, which might be abstractly considered a *utility* (or *reward* in reinforcement learning). Increasingly, ideas from optimal policy learning are being applied in new contexts. In some areas, particularly socially-impactful settings like criminal justice, social welfare policy, hiring, and personal finance, it is essential that automated decisions respect principles of fairness since the relevant data sets include potentially sensitive attributes (e.g., race, gender, age, disability status) and/or features highly correlated with such attributes, so ignoring fairness considerations may have socially unacceptable consequences. A particular worry in the context of automated sequential decision making is "perpetuating injustice," i.e., when maximizing utility maintains, reinforces, or even introduces unfair dependence between sensitive features, decisions, and outcomes. Though there has been growing interest in the issues 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., 2017; Zhang & Bareinboim, 2018; Mitchell & Shalden, 2018; Zhang et al., 2017), so far methods for optimal policy learning subject to fairness constraints have not been well-explored.

As a motivating example, we consider a simplified model for a children's welfare screening program, recently discussed in (Chouldechova et al., 2018; Hurley, 2018). A hotline for child abuse and neglect receives many thousands of calls a year, and call screeners must decide on the basis of calculated risk estimates what action to take in response to any given call, e.g., whether or not to follow up with an in-person visit from a caseworker. The idea is that only cases with substantial potential risk to the child's welfare should be prioritized. The information used to determine the calculated risk level and thereby the agency's action includes potentially sensitive features, such as race and gender, as well as a myriad of other factors such as perhaps whether family members receive public assistance, have an incarceration history, record of drug use, and so on. Though many of these factors may be predictive of subsequent negative outcomes for the children, there is a legitimate worry that both risk calculations and policy choices based on them may depend on sensitive features in inappropriate ways, and thereby lead to unfair racial disparities in the distribution

[1] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. Correspondence to: Razieh Nabi <rnabi@jhu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

of families investigated, and perhaps separated, by child protective services.

Learning high-quality policies that satisfy fairness constraints is difficult due to the fact that multiple sources of bias may occur in the problem simultaneously. One kind of bias, which we call *retrospective bias*, has its origin in the historical data used as input to the policy learning procedure. This data may reflect various systematic disparities and discriminatory historical practices in our society, including prior decisions themselves based on poor data. Algorithms trained on such data can maintain these inequities. Furthermore, decision making algorithms may suffer from what we call *prospective* sources of bias. For instance, suppose the functional form of the chosen decision rule explicitly depends on sensitive features in inappropriate ways. In that case, making decisions based on the new decision rule may perpetuate existing disparities or even introduce disparities that were previously absent. Avoiding this sort of bias may involve imposing non-trivial restrictions on the policy learning procedure. Finally, learning high-quality policies from observational data requires dealing with *confounding bias*, where associations between decision and reward cannot be used directly to assess decision quality due to the presence of confounding variables, as well as *statistical bias* due to the reliance on misspecified statistical models. Policy learning algorithms that respect fairness constraints must address all of these sources of bias.

In this paper, we use tools from mediation analysis and causal inference to formalize fairness criteria as constraints on certain impermissible causal pathways from sensitive features to actions or outcomes (Nabi & Shpitser, 2018). Moreover, we describe how all the aforementioned biases can be addressed by a novel combination of methods from causal inference, constrained optimization, and semiparametric statistics. Our main theoretical result illustrates in what sense enacting fair policies can "break the cycle of injustice": we show how to learn policies such that the joint distribution induced by these policies (in conjunction with reward/utility mechanisms outside the policy-maker's control) will satisfy specified fairness constraints while remaining "close" to the generating distribution. To our knowledge, this paper constitutes the first attempt to integrate algorithmic fairness and policy learning with the possible exception of Jabbari et al. (2017), which addressed what we call prospective bias in the context of Markov Decision Processes.

To precisely describe our approach, we must introduce some necessary concepts and tools from causal inference and policy learning. Then, we summarize the perspective on algorithmic fairness in prediction problems from Nabi & Shpitser (2018), and adapt this framework to learning optimal fair policies. We illustrate our proposal via experiments on synthetic and real data.

2. Notation and preliminaries

Consider a multi-stage decision problem with K pre-specified decision points, indexed by $k = 1, \ldots, K$. Let Y denote the final outcome of interest and A_k denote the action made (treatment administered) at decision point k with the finite state space of $\mathcal{A}_k$. Let X denote the available information prior to the first decision, and Y_k denote the information collected between decisions k and $k+1$, $(Y \equiv Y_K)$. $\overline{A}_k$ represents all treatments administered from time 1 to k; likewise for $\overline{Y}_k$. We combine the treatment and covariate history up to treatment decision A_k into a history vector H_k. The state space of H_k is denoted by $\mathcal{H}_k$. Note that while our proposal in this paper applies to arbitrary state spaces, we present examples with continuous outcomes and binary decisions for simplicity.

The goal of policy learning is to find policies that map vectors in $\mathcal{H}_k$ to values in $\mathcal{A}_k$ (for all k) that maximize the expected value of outcome Y. In offline settings, where exploration by direct experimentation is impossible, finding such policies requires reasoning counterfactually, as is common in causal inference. The value of Y under an assignment of value a to variable A is called a *potential outcome* variable, denoted $Y(a)$. In causal inference, quantities of interest are defined as functions of potential outcomes (also called counterfactuals). Estimating these functions from observational data is a challenging task, and requires assumptions linking potential outcomes to the data actually observed. Our assumptions can be formally represented using *causal graphs*. In a directed acyclic graph (DAG), nodes correspond to random variables, and directed edges represent direct causal relationships. As an example, consider the single treatment causal graph of Fig. 1(a). X is a direct cause of A, and A is both a direct cause of Y as well as an indirect cause of Y through M. A variable like M which lies on a causal pathway from A to Y is called a *mediator*. For more details on causal graphical models see, e.g., Spirtes et al. (2001) and Pearl (2009). In what follows let Z denote the full vector of observed variables in the causal model, e.g., $Z = (Y, M, A, X)$ in our Fig. 1(a).

A causal parameter is said to be *identified* in a causal model if it is a function of the observed data distribution $p(Z)$. In causal DAGs, distributions of potential outcomes are identified by the *g-formula*. For background on general identification theory, see Shpitser (2018). As an example, the distribution of $Y(a)$ in the DAG in Fig. 1(a) is identified by $\sum_{X,M} p(Y|a, M, X)p(M|a, X)p(X)$. Note that some causal parameters may be identified even in causal models with hidden ("latent") variables, typically represented by acyclic directed mixed graphs (ADMGs) (Shpitser, 2018). Though we do not apply our methods to hidden variable models here, the general approach and many of the specific learning strategies we propose are applicable in contexts with hidden variables, so long as the relevant parameters are identified.

In our sequential setting, $Y(\overline{a}_K)$ represents the response Y had the fixed treatment assignment strategy $\overline{A}_K = \overline{a}_K$ been followed, possibly contrary to fact. The contrast $\mathbb{E}[Y(\overline{a}_K)] - \mathbb{E}[Y(\overline{a}'_K)]$, where $\overline{a}_K$ is the treatment history of interest and $\overline{a}'_K$ is the reference treatment history, quantifies the *average causal effect* of $\overline{a}_K$ on the outcome Y.

Mediation and path-specific analysis

One way to understand the mechanisms by which treatments influence outcomes is via mediation analysis. The simplest type of mediation analysis decomposes the causal effect of A on Y into a *direct effect* and an *indirect effect* mediated by a third variable. Consider the graph in Fig 1(a): the direct effect corresponds to the path $A \rightarrow Y$, and indirect effect corresponds to the path through M: $A \rightarrow M \rightarrow Y$. In the potential outcome notation, the direct and indirect effects can be defined using nested counterfactuals such as $Y(a, M(a'))$ for $a, a' \in \mathcal{A}$, 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'. 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), which under some assumptions may be nonparametrically identified by means of the *edge g-formula* provided in Shpitser & Tchetgen Tchetgen (2016). We define a number of PSEs relevant for our examples below. For a general definition, see Shpitser (2013).

Policy counterfactuals and policy learning

Let $f_A = \{f_{A_1}, \ldots, f_{A_K}\}$ be a sequence of decision rules. At the kth decision point, the kth rule f_{A_k} maps the available information prior to the kth treatment decision H_k to treatment decision a_k, i.e. $f_{A_k} : \mathcal{H}_k \mapsto \mathcal{A}_k$. Given f_A we define the counterfactual response of Y had A been assigned according to f_A, or $Y(f_A)$, by the following recursive definition (cf. Robins, 2004; Richardson & Robins, 2013):

$$Y\big(\{f_{A_k}(H_k(f_A)) : A_k \in \mathrm{pa}_{\mathcal{G}}(Y) \cap A\}, \{\mathrm{pa}_{\mathcal{G}}(Y) \setminus A\}(f_A)\big).$$

In words: the potential outcome Y had any parent of Y that is in A been set to f_A in response to counterfactual history H_k up to k, where this history behaves as if A were set to f_A *and* any parent of Y that is not in A, behaves as if A were set to f_A.

Under a causal model associated with the DAG $\mathcal{G}$, the distribution $p(Y(f_A))$, is identified by the following generalization of the g-formula:

$$\sum_{Z \setminus \{Y,A\}} \prod_{V \in Z \setminus A} p(V | \{f_{A_k}(H_k) : A_k \in \mathrm{pa}_{\mathcal{G}}(V) \cap A\}, \mathrm{pa}_{\mathcal{G}}(V) \setminus A).$$

As an example, $Y(a = f_A(X))$ in Fig. 1(a) is defined as $Y(a = f_A(X), M(a = f_A(X), X), X)$, and its distribution is identified as $\sum_{x,m} p(Y|a = f_A(x), M = m, X = x)p(M|a = f_A(x), X = x)p(X = x)$.

Given an identified response to a fixed set of policies f_A, we consider search for the optimal policy set f^*_A, defined to be one that maximizes $\mathbb{E}[Y(f_A)]$. Since $Y(f_A)$ is a counterfactual quantity, validating the found set of policies is difficult given only retrospective data, with statistical bias due to model misspecification being a particular worry. This stands in contrast with online policy learning problems in reinforcement learning, where new data under any policy may be generated and validation is therefore automatic. Partly in response to this issue, a set of orthogonal methods for policy learning have been developed that model different parts of the observed data likelihood function. Q-learning, value search, and g-estimation are common methods used in dynamic treatment regimes literature for learning optimal policies (Chakraborty & Moodie, 2013). We defer detailed descriptions to later in the paper and the supplement.

3. From fair prediction to fair policies

Nabi & Shpitser (2018) argue that fair inference for prediction requires imposing hard constraints on the prediction problem, in the form of restricting certain path-specific effects. We adapt this approach to optimal sequential decision-making. A feature of this approach is that the relevant restrictions are user-specified and context-specific; thus we will generally require input from policymakers, legal experts, bioethicists, or the general public in applied settings. Which pathways may be considered impermissible depends on the domain and the semantics of the variables involved. We do not defend this perspective on fairness here for lack of space; please see Nabi & Shpitser (2018) for more details.

We summarize the proposal from Nabi & Shpitser (2018) with a brief example, inspired by the aforementioned child welfare case. Consider a simple causal model for this scenario, shown in Fig. 1(b). Hotline operators receive thousands of calls per year, and must decide on an action A for each call, e.g., whether or not to send a caseworker. These decisions are made on the basis of a (high-dimensional) vectors of covariates X and M, as well as possibly sensitive features S, such as race. M consists of mediators of the effect of S on A. Y_1 corresponds to an indicator for whether the child is separated from their family by child protective services, and Y_2 corresponds to child hospitalization (presumably attributed to domestic abuse or neglect). The observed joint distribution generated by this causal model

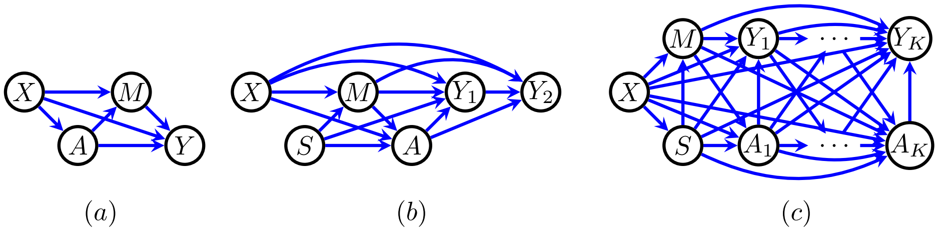

Figure 1. (a) A simple causal DAG, with a single treatment A, a single outcome Y, a vector X of baseline variables, and a single mediator M. (b) A causal DAG corresponding to our (simplified) child welfare example with baseline factors X, sensitive feature S, action A, vector of mediators (including e.g. socioeconomic variables, histories of drug treatment) M, an indicator Y_1 of whether a child is separated from their parents, and an indicator of child hospitalization Y_2. (d) A multistage decision problem, which corresponds to a complete DAG over vertices $X, S, M, A_1, Y_1, \cdots, A_K, Y_K$.

would be $p(Y_1, Y_2, A, M, S, X)$. The proposal from Nabi & Shpitser (2018) is that fairness corresponds to the impermissibility of certain path-specific effects, and so fair inference requires decisions to be made from a counterfactual distribution $p^*(Y_1, Y_2, A, M, S, X)$ which is "nearby" to p (in the sense of minimal Kullback-Leibler divergence) but where these PSEs are constrained to be zero. They call p^* the distribution generated by a "fair world."

Multiple fairness concerns have been raised by experts and advocates in discussions of the child protection decision-making process (Chouldechova et al., 2018; Hurley, 2018). For example, it is clearly impermissible that race has any direct effect on the decision made by the hotline screener, i.e., that all else being held fixed, members from one group have a higher probability of being surveilled by the agency. However, it is perhaps permissible that race has an indirect effect via some mediated pathway, e.g., if race is associated with some behaviors or features which themselves ought to be taken into consideration by hotline staffers, because they are predictive of abuse. If that's true, then $S \rightarrow A$ would be labeled an impermissible pathway whereas $S \rightarrow M \rightarrow A$ (for some M) would be permissible. Similarly, it would be unacceptable if race had an effect on whether children are separated from their families; arguably both the direct pathway $S \rightarrow Y_1$ and indirect pathway though hotline decisions $S \rightarrow A \rightarrow Y_1$ should be considered impermissible. Rather than defend any particular choice of path-specific constraints, we note that the framework outlined in Nabi & Shpitser (2018) can flexibly accommodate any set of given constraints, as long as the PSEs are identifiable from the observed distribution.

3.1. Inference in a nearby "fair world"

We now describe the specifics of the proposal. We assume the data is generated according to some (known) causal model, with observed data distribution $p(\cdot)$, and that we can characterize the fair world by a fair distribution $p^*(\cdot)$ where some set of pre-specified PSEs are constrained to be zero, or within a tolerance range. Without loss of generality we can assume the utility variable Y is some deterministic function of Y_1 and Y_2 (i.e., $Y \equiv u(Y_1, Y_2)$) and thus use Y in place of Y_1 and Y_2 in what follows. Then $Z = (Y, X, S, M, A)$ in our child welfare example. For the purposes of illustration, assume the following two PSEs are impermissible: PSE^{sa}, corresponding to the direct effect of S on A and defined as $\mathbb{E}[A(s, M(s'))] - \mathbb{E}[A(s')]$, and PSE^{sy}, corresponding to the effect of S on Y along the edge $S \rightarrow Y$, and the path $S \rightarrow A \rightarrow Y$ and defined as $\mathbb{E}[Y(s, A(s, M(s')), M(s'))] - \mathbb{E}[Y(s')]$.

If the PSEs are identified under the considered causal model, they can be written as functions of the observed distribution. For example, the unfair PSE of the sensitive feature S on outcome Y in our child welfare example may be written as a functional PSE$^{sy} = g_1(Z) \equiv g_1(p(Y, X, S, M, A))$. Similarly the unfair PSE of S on A is PSE$^{sa} = g_2(Z) \equiv g_2(p(Y, X, S, M, A))$. Generally, given a set of identified PSEs $g_j(Z)\ \forall j \in \{1, ..., J\}$ and corresponding tolerated lower/upper bounds $\epsilon_j^-, \epsilon_j^+$, the fair distribution $p^*(Z)$ is defined:

$$p^*(Z) \equiv \arg\min_q \ D_{KL}(p||q)$$
$$\text{subject to} \ \ \epsilon_j^- \leq g_j(Z) \leq \epsilon_j^+, \ \ \forall j \in \{1, ..., J\}, \qquad (1)$$

where D_{KL} is the KL-divergence and J is the number of constraints.[1] In finite sample settings, Nabi & Shpitser (2018) propose solving the following constrained maximum likelihood problem:

$$\widehat{\alpha} = \arg\max_\alpha \ \mathcal{L}(Z; \alpha)$$
$$\text{subject to} \ \ \epsilon_j^- \leq \widehat{g}_j(Z) \leq \epsilon_j^+, \ \ \forall j \in \{1, ..., J\}, \qquad (2)$$

where $\widehat{g}_j(Z)$ are estimators for the chosen PSEs and $\mathcal{L}(Z; \alpha)$ is the likelihood function. The most relevant bounds in practice are $\epsilon_j^- = \epsilon_j^+ = 0$.

[1] Note that in our examples J will typically be $K + 1$, i.e., one constraint for the S to Y paths and one constraint for each set of paths from S to A_k. We allow for J constraints in general to accommodate more complex settings (e.g., where there are multiple sensitive features, multiple outcomes, or a different set of pathways are constrained).

Given an approximation of p^* learned in this way, Nabi & Shpitser (2018) transform regression problems originally defined on p into regression problems on p^*. In other words, instead of learning a regression function $\mathbb{E}[Y \mid M, A, S, X]$ on the observed data distribution $p(Z)$, they approximate the fair distribution $p^*(Z)$ by constrained maximum likelihood and classify new instances using the constrained model. As we discuss in more detail later, Nabi & Shpitser (2018) choose to average over certain variables to accomodate the fact that new instances are generated from p rather than p^*.

3.2. Fair decision-making

In the sequential decision setting, there are multiple complications. In particular, we aim to learn high-quality policies while simultaneously making sure that the joint distribution induced by the policy satisfies our fairness criteria, potentially involving constraints on multiple causal pathways. This problem must be solved in settings where distributions of some variables, such as outcomes, are not under the policy-maker's control. Finally, we must show that if the learned policy is adapted to new instances (drawn from the original observed distribution) in the right way, then these new instances combined with the learned policy, constrained variables, and variables outside our control, together form a joint distribution where our fairness criteria remain satisfied.

Consider a K-stage decision problem given by a DAG where every vertex pair is connected, and with vertices in a topological order $X, S, M, A_1, Y_1, \ldots, A_K, Y_K$. See Fig. 1(c). Note that the setting where S can be assumed exogenous is a special case of this model with missing edge between X and S. Though we only assume a single set of permissible mediators M here, at the expense of some added cumbersome notation all of the following can be extended to the case where there are distinct sets of mediators $M_1, \ldots, M_K$ preceding every decision point. (We extend the results below to that setting in the Supplement.) We will consider the following PSEs as inadmissible: PSE^{sy}, representing the effect of S on Y along all paths *other than* the paths of the form $S \to M \to \ldots \to Y$; and PSE^{sa_k}, representing the effect of S on A_k along all paths *other than* the paths of the form $S \to M \to \ldots \to A_k$. That is, we consider *only* pathways connecting S and A_k or Y through the allowed mediators M to be fair. In this model, these PSEs are identified by (Shpitser, 2013):

$$\begin{aligned}\text{PSE}^{sy} &= \mathbb{E}[Y(s, M(s'))] - \mathbb{E}[Y(s')] \\ &= \sum_{X,M} \{\mathbb{E}[Y|s, M, X] - \mathbb{E}[Y|s', M, X]\} p(M|s', X) p(X)\end{aligned}$$

$$\begin{aligned}\text{PSE}^{sa_k} &= \mathbb{E}[A_k(s, M(s'))] - \mathbb{E}[A_k(s')] \\ &= \sum_{X,M} \{\mathbb{E}[A_k|s, M, X] - \mathbb{E}[A_k|s', M, X]\} p(M|s', X) p(X)\end{aligned}$$

Numerous approaches for estimating and constraining these identified PSEs are possible. In this paper, we restrict our attention to semiparametric estimators, which model only a part of the likelihood function while leaving the rest completely unrestricted. Estimators of this sort share some advantages with parametric methods (e.g., often being uniformly consistent at favorable rates), but do not require specification of the full probability model. Specifically, we use estimators based on the following result:

Theorem 1 *Assume S is binary. Under the causal model above, the following are consistent estimators of PSE^{sy} and PSE^{sa_k}, assuming all models are correctly specified:*

$$\widehat{g}^{sy}(Z) = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)} \right\} Y_n \tag{3}$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)} \right\} A_{kn} \tag{4}$$

These inverse probability weighted (IPW) estimators use models for M and S. Thus, we can approximate p^* by constraining only the M and S models, i.e., obtaining estimates $\hat{\alpha}_m$ and $\hat{\alpha}_s$ of the parameters α_m and α_s in $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$ by solving (2). The outcomes Y_k and decisions A_k are left unconstrained. This is subtle and important, since it enables us to choose our optimal decision rules f_A^* without restriction of the policy space and allows the mechanism determining outcomes Y_k (based on decisions A_k and history H_k) to remain outside the control of the policy-maker. Consequently, we can show that implementing this procedure guarantees that the joint distribution over all variables Z induced by 1) the constrained M and S models, 2) the conditional distributions for A_k given H_k implied by the optimal policy choice, and 3) *any* choice of $p(Y_k|A_k, H_k)$ will (at the population-level) satisfy the specified fairness constraints. We prove the following result in the Supplement:

Theorem 2 *Consider the K-stage decision problem described by the DAG in Fig. 1(c). Let $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$ be the constrained models chosen to satisfy PSE$^{sy} = 0$ and PSE$^{sa_k} = 0$. Let $\tilde{p}(Z)$ be the joint distribution induced by $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$, and where all other distributions in the factorization are unrestricted. That is,*

$$\tilde{p}(Z) \equiv p(X) p^*(S|X; \alpha_s) p^*(M|S, X; \alpha_m) \times \prod_{k=1}^{K} p(A_k|H_k) p(Y_k|A_k, H_k).$$

Then the functionals PSE^{sy} and PSE^{sa_i} taken w.r.t. $\tilde{p}(Z)$ are also zero.

This theorem implies that any approach for learning policies based on $\tilde{p}(Z)$ addresses both retrospective bias (since the fairness criterion violation present in $p(Z)$ is absent in $\tilde{p}(Z)$) and prospective bias (since the criterion holds in $\tilde{p}(Z)$ for any choice of policy on A_k inducing $p(A_k|H_k)$). As we discuss in detail in the next section, modified policy learning based on $\tilde{p}(Z)$ requires special treatment of the constrained variables S and M. New instances (e.g., new calls to the child protection hotline) will be drawn from the unfair distribution p, not $\tilde{p}$. So, the enacted policy cannot use empirically observed values of S or M. In what follows, our approach is to either average over S and M (following Nabi & Shpitser (2018)), or resample observations of S and M from the constrained models.

4. Estimation of optimal policies in the fair world

In the following, we describe several strategies for learning optimal policies, and our modifications to these strategies based on the above fairness considerations.

4.1. Q-learning

In Q-learning, the optimal policy is chosen to optimize a sequence of counterfactual expectations called Q-functions. These are defined recursively in terms of value functions $V_k(\cdot)$ as follows:

$$\begin{aligned} Q_K(H_K, A_K) &= \mathbb{E}[Y_K(A_K) \mid H_K], \\ V_K(H_K) &= \max_{a_K} Q_K(H_K, a_K), \end{aligned} \tag{5}$$

and for $k = K-1, \ldots, 1$

$$\begin{aligned} Q_k(H_k, A_k) &= \mathbb{E}[V_{k+1}(H_{k+1}, A_k) \mid H_k], \\ V_k(H_k) &= \max_{a_k} Q_k(H_k, a_k). \end{aligned} \tag{6}$$

Assuming $Q_k(H_k, A_k)$ is parameterized by β_k, the optimal policy at each stage may be easily derived from Q-functions as $f^*_{A_k}(H_k) = \arg\max_{a_k} Q_k(H_k, a_k; \widehat{\beta}_k)$. Q-functions are recursively defined regression models where outcomes are value functions, and features are histories up to the current decision point. Thus, parameters β_k ($k = 1, \ldots, K$) of all Q-functions may be learned recursively by maximum likelihood methods applied to regression at stage k, given that the value function at stage $k+1$ was already computed for every row. See Chakraborty & Moodie (2013) for more details.

Note that at each stage k, the identity $Q_k(H_k, A_k) = \mathbb{E}[V_{k+1}(H_{k+1}, A_k) \mid H_k] = \mathbb{E}[V_{k+1}(H_{k+1}) \mid A_k, H_k]$ only holds under our causal model if the *entire past* H_k is conditioned on. In particular, $\mathbb{E}[V_{k+1}(H_{k+1}, A_k) \mid H_k \setminus \{M, S\}] \neq \mathbb{E}[V_{k+1}(H_{k+1}) \mid A_k, H_k \setminus \{M, S\}]$. To see a simple example of this, note that $Y_K(a_1)$ is not independent of A_1 conditional on just X in Fig. 1(c), due to the presence of the path $Y_K \leftarrow M \rightarrow A_1$; however the indepenence does hold conditional on the entire $H_1 = \{X, S, M\}$ (Richardson & Robins, 2013).

In a fair policy learning setting, though $\{M, S\}$ may be in H_k, we cannot condition on values of M, S to learn fair policies since these values were drawn from p rather than p^*. There are multiple ways of addressing this issue. One approach is to modify the procedure to obtain optimal policies that condition on all history *other than* $\{M, S\}$. We first learn Q_ks using (5) and (6). We then provide the following modified definition of Q-functions defined directly on p^*:

$$\begin{aligned} &Q^*_k(H_k \setminus \{M, S\}, A_k; \beta_k) = \\ &\frac{1}{Z} \sum_{m,s} Q_k(H_k, A_k; \beta_k) \prod_{i=1}^{k} p(A_i | H_i \setminus \{M, S\}, m, s). \\ &\prod_{i=2}^{k-1} p(M_i | A_i, H_i \setminus \{M, S\}, m, s)\; p^*(m, s | X), \end{aligned}$$

for $k = K, \ldots, 1$,

$$\begin{aligned} Z = \sum_{m,s} p^*(m, s|X) &\prod_{i=1}^{k} p(A_i | H_i \setminus \{M, S\}, m, s) \\ &\prod_{i=2}^{k-1} p(M_i | A_i, H_i \setminus \{M, S\}, m, s). \end{aligned}$$

The optimal fair policy at each stage is then derived from Q^*-functions as $f^*_{A_k}(H_k) = \arg\max_{a_k} Q^*{}_k(H_k \setminus \{M, S\}, a_k; \widehat{\beta}_k)$.

As an alternative approach, we can compute the original Q-functions defined in (5) and (6) with respect to $p^*(Z)$ by ignoring the observed values M_n and S_n for the nth individual and replacing them with samples drawn from $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$. Then, in (5) and (6), the history at the kth stage, H_k, gets replaced with $H^*_k = \{H_k \setminus \{M, S\}, M^*, S^*\}$.

4.2. Value search

It may be of interest to estimate the optimal policy within a restricted class $\mathcal{F}$. One approach to learning the optimal policy within $\mathcal{F}$ is to directly search for the optimal $f^{*,\mathcal{F}}_A \equiv \arg\max_{f_A \in \mathcal{F}} \mathbb{E}[Y(f_A)]$, which is known as *value search*.

The expected response to an arbitrary policy $\phi = \mathbb{E}[Y(f_A)]$, for $f_A \in \mathcal{F}$ can be estimated in a number of ways. Often $\widehat{\phi}$ takes the form of a solution to some estimating equation $\mathbb{E}[h(\phi)] = 0$ solved empirically given samples from $p(Z)$. A simple estimator for ϕ that uses only the treatment assignment model $\pi(H_k; \psi) \equiv p(A_k = 1|H_k)$ is the IPW estimator that solves the following estimating equation:

$$\mathbb{E}\left[\prod_{k=1}^{K} \left\{C_{f_{A_k}} / \pi_{f_{A_k}}(H_k; \widehat{\psi})\right\} \times Y - \phi\right] = 0, \tag{7}$$

where $C_{f_{A_k}} \equiv \mathbb{I}(A_k = f_{A_k}(H_k))$, $\pi_{f_{A_k}}(H_k; \psi) \equiv \pi(H_k; \psi) f_{A_k}(H_k) + (1 - \pi(H_k; \psi))(1 - f_{A_k}(H_k))$, the expectation is evaluated empirically, and $\widehat{\psi}$ is fit by maximum likelihood.

Finding fair policies via value search involves solving the same problem with respect to $p^*(Z)$ instead. Given known models $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$, we may consider two approaches. The first one involves solving a modified estimating equation of the form

$$\mathbb{E}^*[h(\phi)] \equiv$$
$$\mathbb{E}\left[\sum_{m,s} \mathbb{E}[h(\phi)|M,S,X] p^*(M|S,X;\alpha_m) p^*(S|X;\alpha_s)\right] = 0$$

with respect to $p^*(Z \setminus \{M,S\})$. The alternative is to solve the original estimating equation $\mathbb{E}[h(\phi)] = 0$ with respect to $p^*(Z)$ by replacing observed values M_n and S_n for the nth individual with sampled values M_n^* and S_n^* drawn from $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$. In both approaches, the optimal fair policy at each stage is then derived by replacing the history at the kth stage, H_k, with $H_k^* = \{H_k \setminus \{M,S\}, M^*, S^*\}$. Given constrained models $p^*(M|S,X;\alpha_m)$, and $p^*(S|X;\alpha_s)$ representing $p^*(Z)$, we can perform value search by solving the given estimating equation empirically on a dataset where every row x_n, s_n, m_n in the data is replaced with I rows x_n, s_{ni}^*, m_{ni}^* for $i = 1, \ldots I$, with m_{ni}^* and s_{ni}^* drawn from $p^*(M|S,x_n;\alpha_m)$ and $p^*(S|x_n;\alpha_s)$, respectively.

4.3. G-estimation

A third method for estimating policies is to directly model the counterfactual contrasts known as *optimal blip-to-zero functions* and then learn these functions by a method called g-estimation (Robins, 2004). In the interest of space, we defer a full description of blip-to-zero functions and g-estimation to the Supplement, where we also present some results for our implementation of fair g-estimation.

4.4. Tradeoffs and treatment of constrained variables

We've proposed to constrain the M and S models to satisfy given fairness constraints. Since empirically observed values of M and S are sampled from p rather than p^* (or $\tilde{p}$), our approach requires resampling or averaging over these features. The choice of models to constrain involves a tradeoff. The more models are constrained, the closer the KL distance between p and p^*, but the more features have to be resampled or averaged out; that is, some information on new instances is "lost." Alternative approaches may constrain fewer or different models in the likelihood (for example, we could have elected to constrain the Y model instead of S). However, the benefit of our approach here is that we can guarantee, with outcomes Y outside the policy-maker's control, that the induced joint distribution will satisfy the given fairness constraints (by Theorem 2), whereas alternative procedures which aim to avoid averaging or resampling will typically have no such guarantees. Another alternative that avoids averaging over variables altogether is to consider likelihood parameterizations where the absence of a given PSE directly corresponds to setting some variation-independent likelihood parameter for the Y model to zero (cf. Chiappa (2019)). While such a parameterization is possible for linear structural equation models, it is an open problem in general for arbitrary PSEs and nonlinear settings. Developing novel, general-purpose alternatives that transfer observed distributions to their "fair versions," while avoiding resampling and averaging, is an open problem left to future work.

5. Experiments

Synthetic data

We generated synthetic data for a two-stage decision problem according to the causal model shown in Fig. 1(c) ($K = 2$), where all variables are binary except for the continuous response utility $Y \equiv Y_2$. Details on the specific models used are reported in the Supplement. We generated a dataset of size $5,000$, with 100 bootstrap replications, where the sensitive variable S is randomly assigned and where S is chosen to be an informative covariate in estimating Y.

We use estimators in Theorem 1 to compute PSE^{sy}, PSE^{sa_1}, and PSE^{sa_2} which entail using M and S models. In this setting, the PSE^{sy} is 1.918 (on the mean scale) and is restricted to lie between -0.1 and 0.1. The PSE^{sa_1} is 0.718, and PSE^{sa_2} is 0.921 (on the odds ratio scale) and both are restricted to lie between 0.95 and 1.05. We only constrain M and S models to approximate p^* and fit these two models by maximizing the constrained likelihood using the R package `nloptr`. The parameters in all other models were estimated by maximizing the likelihood.

Optimal fair polices along with optimal (unfair) policies were estimated using the two techniques described in Section 4 (where we used the "averaging" approach in both cases). We evaluated the performance of both techniques by comparing the population-level response under fair policies versus unfair policies. One would expect the unfair policies to lead to higher expected outcomes compared to fair policies since satisfying fairness constraints requires sacrificing some policy effectiveness. The expected outcomes under unfair polices obtained from Q-learning and value search were 7.219 ± 0.005 and 7.622 ± 0.265, respectively. The values dropped to 6.104 ± 0.006 and 6.272 ± 0.133 under fair polices, as expected. In addition, both fair and unfair optimal polices had higher expected outcomes than the observed population-level outcome, using both methods. In our simulations, the population outcome under observed policies was 4.82 ± 0.007. Some additional results are reported in the Supplement.

Application to the COMPAS dataset

COMPAS is a criminal justice risk assessment tool created by the company Northpointe that has been used across the US to determine whether to release or detain a defendant before their trial. Each pretrial defendant receives several COMPAS scores based on factors including but not limited to demographics, criminal history, family history, and social status. Among these scores, we are primarily interested in the "risk of recidivism." We use the data made available by Propublica and described in Angwin et al. (2016). The COMPAS risk score for each defendant ranges from 1 to 10, with 10 being the highest risk. In addition to this score (A), the data also includes records on defendants age ($X_1 \in X$), gender ($X_2 \in X$), race (S), prior convictions (M), and whether or not recidivism occurred in a span of two years (R). We limited our attention to the cohort consisting of African-Americans and Caucasians, and to individuals who either had not been arrested for a new offense or who had recidivated within two years. Our sample size is 5278. All variables were binarized including the COMPAS score, which we treat as an indicator of a binary decision to incarcerate versus release (pretrial) "high risk" individuals, i.e., we assume those with score ≥ 7 were incarcerated. In this data, 28.9% of individuals had scores ≥ 7.

Since the data does not include any variable that corresponds to utility, and there is no uncontroversial definition of what function one should optimize, we define a heuristic utility function from the data as follows. We assume there is some (social, economic, and human) cost, i.e., negative utility, associated with incarceration (deciding $A = 1$), and that there is some cost to releasing individuals who go on to reoffend (i.e., for whom $A = 0$ and $R = 1$). Also, there is positive utility associated with releasing individuals who do not go on to recidivate (i.e., for whom $A = 0$ and $R = 0$). A crucial feature of any realistic utility function is how to balance these relative costs, e.g., how much (if any) "worse" it is to release an individual who goes on to reoffend than to incarcerate them. To model these considerations we define utility $Y \equiv (1 - A) \times \{-\theta R + (1 - R)\} - A$. The utility function is thus parameterized by θ, which quantifies how much "worse" is the case where individuals are released and reoffend as compared with the other two possibilities which are treated symmetrically. We emphasize that this utility function is a heuristic we use to illustrate our optimal policy learning method, and that a realistic utility function would be much more complicated (possibly depending also on factors not recorded in the available data).

We apply our proposed Q-learning procedure to optimize $\mathbb{E}[Y]$, assuming $K = 1$ and exogenous S. The fair policy constrains the $S \rightarrow A$ and $S \rightarrow Y$ pathways. We describe details of our implementation as well as additional results in the Supplement. The proportion of individuals incarcerated ($A = 1$) is a function of θ, which we plot in Fig. 2 stratified by racial group. See the Supplement for results on *overall* incerceration rates, which also vary among the policies. The region of particular interest is between $\theta = 2$ and 3, where fair and unrestricted optimal policies differ and both recommend lower-than-observed overall incarceration rates (see Supplement). For most θ values, the fair policy recommends a decision rule which narrows the racial gap in incarceration rates as compared with the unrestricted policy, though does not eliminate this gap entirely. (Constraining the causal effects of race through mediator M would go further in eliminating this gap.) In regions where $\theta > 3$, both optimal policies in fact recommend higher-than-observed overall incarceration rates but a narrower racial gap, particularly for the fair policy. Comparing fair and unconstrained policy learning on this data serves to simultaneously illustrate how the proposed methods can be applied to real problems and how the choice of utility function is not innocuous.

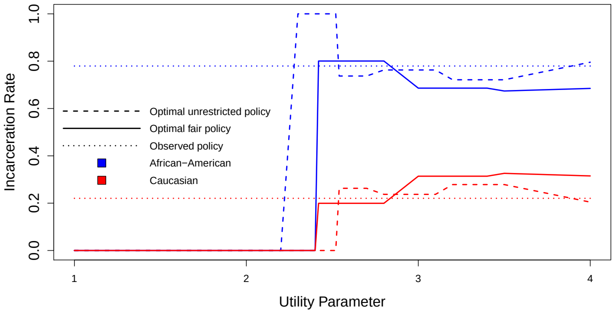

Figure 2. Group-level incarceration rates for the COMPAS data as a function of the utility parameter θ.

6. Conclusion

We have extended a formalization of algorithmic fairness from Nabi & Shpitser (2018) to the setting of learning optimal policies under fairness constraints. We show how to constrain a set of statistical models and learn a policy such that subsequent decision making given new observations from the "unfair world" induces high-quality outcomes while satisfying the specified fairness constraints in the induced joint distribution. In this sense, our approach can be said to "break the cycle of injustice" in decision-making. We investigated the performance of our proposals on synthetic and real data, where in the latter case we have supplemented the data with a heuristic utility function. In future work, we hope to develop and implement more sophisticated constrained optimization methods, to use information as efficiently as possible while satisfying the desired theoretical guarantee, and to explore nonparametric techniques for complex settings where the likelihood is not known.

Acknowledgments

This project is sponsored in part by the National Institutes of Health grant R01 AI127271-01 A1 and the Office of Naval Research grant N00014-18-1-2760.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Propublica, 2016.

Bertsekas, D. P. and Tsitsiklis, J. *Neuro-Dynamic Programming*. Athena Publishing, 1996.

Chakraborty, B. and Moodie, E. E. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. New York: Springer-Verlag, 2013.

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

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances In Neural Information Processing Systems*, pp. 3315–3323, 2016.

Hurley, D. Can an algorithm tell when kids are in danger? The New York Times, 2018.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., , and Roth, A. Fairness in reinforcement learning. In *Proceedings of International Conference on Machine Learning*, 2017.

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

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. In *Advances In Neural Information Processing Systems*, 2017.

Mitchell, S. and Shalden, J. Reflections on quantitative fairness. https://speak-statistics-to-power.github.io/fairness/, 2018.

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

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

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

Pedreshi, D., Ruggieri, S., and Turini, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–568, 2008.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Preprint: http://www.csss.washington.edu/Papers/wp128.pdf, 2013.

Robins, J. M. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, pp. 189–326, 2004.

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

Shpitser, I. Identification in graphical causal models. In *Handbook of Graphical Models*. CRC Press, 2018.

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

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2001.

Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*. MIT press, 1998.

Zhang, J. and Bareinboim, E. 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.

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