Realizing Fairness in an Unfair World: A Causal Approach to Fair Predictive Modeling via Penalized Maximum Likelihood Estimation

Biostatistics Master's Project

Antonella Basso

Brown University School of Public Health, 2023

Abstract

In light of the ongoing social injustices catalyzed and compounded by automated decision making, the primary aim of this research is to deliver an intuitive causal-based method for fair predictive modeling that appeals to path-specific effects (PSEs) and a penalty-based approach to maximum likelihood estimation. Specifically, we propose an adaptation to the work of Nabi, Malinsky, and Shpitser (2022), that simultaneously simplifies the (constrained) optimization problem, produces more explainable results, and holds decision makers accountable for the consequences of predictions. In addition to demonstrating the proposed methodology on a simulated biased dataset with direct discrimination, we discuss important limitations of this approach and provide recommendations for maximizing social impact. Further, we outline some possible extensions to this work that we believe can substantially strengthen its performance and broaden its applicability. More generally, we seek to show how grounding the fairness problem within a counterfactual reasoning framework allows us to formalize discrimination in a way that is most consistent with rational thought, as well as the moral and legal principles of equity and justice that govern society.

I. Introduction

In the age of information, where society's dependence on predictive, data-driven algorithms has grown, attention must be paid to the ethical implications of our models. That is, when automated predictions lead to decisions that directly impact people's lives, limiting the bias within our systems is pivotal to ensuring justice and promoting equity. For this reason, the research space of algorithmic fairness has gained significant traction within the statistical, computer, and data science communities, prompting a wide range of metrics for quantifying existing notions of fairness in attempts to reduce algorithmic discrimination against marginalized groups and individuals.[1]

Motivated by the urgent need to develop data mining and machine learning techniques that are

[1]As is common in the fairness literature, we use the legal terms "sensitive" or "protected" attributes to refer to personal characteristics that may be used to discriminate against particular groups or individuals, and hence require protection from unfair treatment in various contexts, such as employment, education, housing, and lending. Sensitive attributes that are typically subject to legal protections under human rights and anti-discrimination laws in many countries include, but are not limited to, "race", "ethnicity", "gender", "sexual orientation", "age", "disability", and "religion" (Bonchi et al. 2017).

Table 1: Statistical Non-Discrimination Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$
$\mathrm{E}[\hat{Y}\|A]$	$\mathrm{E}[\hat{Y}\|Y, A]$	$\mathrm{E}[Y\|\hat{Y}, A]$

not only accurate but maximally fair, we follow the works of Nabi and Shpitser (2017) and Nabi, Malinsky, and Shpitser (2022) to show how one may utilize the causal notion of path-specific effects (PSEs) to identify the degree of discrimination in a dataset, and subsequently generate new predictions according to an optimized semi-parametric form of its observed likelihood subject to fairness constraints. Specifically, we build on this work, proposing a more straightforward approach to optimization that appeals to a redefined objective function with an L1 penalty on the PSE as a way of overcoming the need to reparameterize the data likelihood as well as the assumption that the bounds of fair prediction are known a priori. In addition to turning the focal constrained optimization task into a simpler unconstrained problem that supports nonlinear forms of the PSE, our approach fosters a more transparent and flexible environment for decision making, empowering users to understand the trade-offs between accuracy and fairness, and hence establish optimality contextually via empirical analysis. Finally, we implement our proposed method on a simulated biased dataset to demonstrate a simple application to direct discrimination, while discussing ways in which it may be modified to satisfy other contextual aims.

In what follows, we outline some of the most recognized notions of fairness and fairness criteria that exist in the literature and serve as the primary building blocks for many of the data mining and machine learning strategies we see today. Moreover, we provide an overview of causally-motivated approaches to algorithmic fairness, primarily focusing on the intuition behind this mode of reasoning, as well as the ways in which it can strengthen computational developments in fairness research. Lastly, in addition to commenting on potential limitations of our approach—including those general to the causal, or counterfactual reasoning, framework—we consider some promising extensions to this work in hopes of inspiring future research to combat algorithmic injustice.

II. Background

Statistical non-discrimination or fairness criteria, which aim to define the absence of discrimination—often by equalizing some group-dependent statistical quantity across groups—are formally expressed as properties of the joint distribution of a sensitive attribute A, an outcome Y, covariates X, and a classifier or predictor $\hat{Y}$ (Barocas, Hardt, and Narayanan 2018). That is, non-discrimination, in the context of prediction problems and automated decision making, is generally understood as the "probability of a particular prediction occurring given that some prior conditions hold" (Loftus, Russell, et al. 2018). While several criteria have been proposed in this regard as a way of capturing different intuitions about what it means for predictions to be fair, Barocas, Hardt, and Narayanan (2018) "simplify the landscape" of these constraints by providing three conditions that synthesize their aims. For binary outcomes specifically, these include equalized (i) acceptance rates $\mathrm{P}(\hat{Y} = 1)$, (ii) error rates $\mathrm{P}(\hat{Y} = 0|Y = 1)$ and $\mathrm{P}(\hat{Y} = 1|Y = 0)$, and outcome frequency $\mathrm{P}(Y = 1|R = r)$ for a given risk score value $R = r(x) = \mathrm{P}(Y = 1|X = x)$ across groups. For standard prediction problems, assuming any type of response variable, these criteria generalize to equalizing expected predictive values $\mathrm{E}[\hat{Y}]$, model error $\mathrm{E}[\hat{Y}|Y]$, and outcome frequency $\mathrm{E}[Y|R = r]$ for $r(x) = \mathrm{E}[Y|X = x]$ across groups, where R may be substituted for $\hat{Y}$. It should be noted however, that the second criterion, which uses Y as a stand-in for merit, is not applicable in cases where the target variable used for training is biased (Barocas, Hardt, and Narayanan 2018). These criteria can be expressed as score functions using independence statements as shown in Table 1 above.

As is most often the case, discriminatory or "unfair" algorithmic predictions stem from historically biased data. Put simply, when an algorithm is trained to pick up on data patterns that themselves reflect a history of discrimination against stigmatized groups, it is guaranteed to replicate and compound such biases when predicting on unobserved data. In turn, this leads to discriminatory decisions that perpetuate injustice through the inconspicuous production and reproduction of biased data. An abundance of data mining and machine learning strategies have been proposed in response to such adverse feedback loops between algorithms and society, each driven by distinct computational realizations of fairness and unique research endeavors. Together, they epitomize the interdisciplinary pursuit for algorithmic justice; seeking to subject predictive algorithms to the pre-processing, in-processing, and/or post-processing mechanisms that effectuate outcomes we can reasonably accept as fair.

Pre-processing strategies are specifically designed to debias the "train(ing)" data, that is, modify the data fed into the algorithm such that it no longer bears any underlying elements of discrimination (Mehrabi et al. 2022). For instance, Kamiran and Calders (2012) study the problem of learning a classifier that maximizes accuracy without exhibiting unlawful discrimination towards sensitive attributes in predictions by assessing the overall ability to remove discrimination from the training data between (i) suppressing sensitive attributes, (ii) changing class labels, and (iii) re-weighing or re-sampling the data. Conversely, Mancuhan and Clifton (2014) propose a more elaborate approach to discrimination detection and prevention that first measures the effect of a sensitive attribute class on a subset of the data using a probability distribution estimated via a Bayesian network, and subsequently implements a classification method to correct for discovered discrimination in the data that excludes class labels from prediction.

In-processing methods on the other hand, strive to alter the learning algorithm itself to guide unbiased predictions, often in compliance with specific standards of fairness (Mehrabi et al. 2022). With a focus on adversarial learning for example, Zhang, Lemoine, and Mitchell (2018) offer a framework for training unbiased machine learning models, which involves maximizing a given predictor's ability to predict an outcome while minimizing an adversary's ability to predict a sensitive attribute so as to mitigate unwanted biases and obtain results that are more consistent with various notions of fairness. In contrast, post-processing techniques aim to evaluate and correct predictions made on unobserved "test(ing)" data according to given fairness metric(s) (Mehrabi et al. 2022). For example, Mehrabi et al. (2021) design a post-processing bias mitigation strategy that uses attributions derived from their proposed attention-based framework to reduce the attention weights, and hence the effects, of features that lead to unfair outcomes.

Like Kamishima et al. (2012), our proposed method employs regularization as an in-processing approach to fair predictive modeling, but differs in being grounded within a causal view of fairness. Causal inference, like other areas in statistics, has guided a number of processing schemes in offering an intuitive way to reason about fairness and, consequently, a promising computational framework for promoting it algorithmically, as we discuss in the coming section. This is the case for both pre-processing and post-processing techniques[2], as well as in-processing methods similar to the one we propose. We note that although this work specifically focuses on an in-processing approach to training fair predictive models, we do not posit that such a mechanism would outperform its pre- and post-processing counterparts, were they to exist. Instead, we contend that PSEs and the role they play within a counterfactual view of fairness, as we later show, provide a solid means for producing fair predictions, which so happens to best fit within, rather than prior to or following, the algorithmic learning process. We do however argue that a combined processing approach can only aid the detection and removal of bias and ought to be considered in practice for additional protection against potentially unfair decisions. Hence, it is possible, but remains to be shown, that implementing more than one processing strategy is in fact more effective in actualizing algorithmic fairness.

[2]See Zhang, Lemoine, and Mitchell (2018), and Altman, Wood, and Vayena (2018) for examples.

III. Causality

Discrimination detection and removal is the primary focus of most pre-processing strategies for algorithmic fairness that exist within the data mining community. The main drawback to the majority of these approaches however, lies in their understanding of fairness as an *associative*, rather than as a *causal*, matter—thus, rendering their claims about discrimination valid on account of mere correlation between sensitive attributes and unfair outcomes. Yet, discrimination, as it is understood within the legal system, is a causal phenomenon. This is implicit in its definition—any occurrence in which (i) a group is treated "less favorably" compared to others, or (ii) a higher proportion of non-members complies with a qualifying criterion, according to current legislation—which "requires plaintiffs to demonstrate a causal connection between the challenged outcome and a protected status characteristic" (Bonchi et al. 2017). That is, discrimination claims legally require proof of whether group membership is tied to unfavorable treatment or outcome, which is tantamount to answering the counterfactual question: "How would one's outcome differ if they had been a member of another group?" And, while correlation-based evidence may suffice in demonstrating the presence of discrimination; in line with the well known fact that correlation is not causation; it cannot guarantee the truth as it does not "properly address the causal question" posed by the court (Bonchi et al. 2017).

Similarly, although it is possible for both associative and causal approaches to coincide in their mathematical formulations and yield identical results, Kusner et al. argue that we have something more to gain from making explicit the causal assumptions that underlie fairness. Specifically, in addition to removing ambiguity from correlation-based methods, overtly stating these causal assumptions allows us to utilize the information we have available on other existing factors in the data to minimize their influence such that we can identify the root cause(s) of discrimination in prediction (Loftus, Russell, et al. 2018). To illustrate this benefit and demonstrate how associative methods are inadequate for justifying discrimination claims, several works turn to the famous UC Berkeley graduate admissions case from 1973 involving a gender bias lawsuit against the university, which exemplifies what is known as Simpson's paradox—a statistical phenomenon wherein an observed correlative trend "disappears or reverses when the same data is disaggregated into its underlying subgroups" (Mehrabi et al. 2022).

Specifically, this case reflects a scenario wherein correlation-based evidence of admissions decisions was used to argue that men were more likely than women to be admitted into the graduate school—an indication of potential discrimination against applicants who identify as women. However, when the data was disaggregated and analyzed by department, it was shown that women in fact had equal, and in some cases even slightly higher, chances of being admitted compared to men. The reason being that women tended to apply to more competitive departments, while men tended to apply to departments with much higher admission rates (Qureshi et al. 2019). Hence, department choice "mediates" the relationship between gender and admittance decision in such a way that adjusting for it reveals that there is in fact no *direct* relationship, and in turn no forbidden form of gender bias. That is, once department choice has been accounted for, gender no longer has any influence over admission outcomes, thus ruling out illegal discrimination by gender (Barocas, Hardt, and Narayanan 2018).

In this way, "causal discrimination discovery provides a solid reasoning framework to detect biases contained in the training data, which may be inherited by the learned decision model", so as to prevent black-box systems whose predictions are not only ridden with discrimination, but are unexplainable (Qureshi et al. 2019). Moreover, causal approaches to discrimination discovery beget more reliable and defensible fairness analyses in practice in that they, unlike more widespread correlation-based frameworks that disregard confounding factors and mediators, allow for sound causal effect estimation.

In addition to debiasing algorithmic predictions through discrimination detection and removal, an abundance of machine learning strategies have been proposed in the literature in an effort to satisfy

particular notions of fairness—often with ample debate over which is most successful in promoting fair algorithmic outcomes, primarily on account of their inability to satisfy competing accounts of fairness concurrently and the failure to guarantee them indefinitely. Being guided by what we take to be a computational and more confined view of the fairness problem, many such (in-processing) techniques surrender either to the impossible task of jointly meeting several existing fairness criteria, or to devising a single fairness criterion that will encompass all of its known dimensions, thereby losing sight of the true contextual nature of fairness and its more substantive realization in decision making.

Although popular statistical notions of fairness are not necessarily in conflict with a causal reasoning framework, as shown by Kusner et al., their use as a basis for developing fair predictive models and other such processing methods limits their applicability in that they target specific aspects of outcome fairness without considering the very causes of unfairness that facilitate the dynamic demands of justice. Thus, grounding the fairness problem in causality provides a means for mathematizing our innate understanding of fairness, which in turn, allows us to overcome several key issues that encumber other proposed schemes—all which can be tied to their inherent foci on particular notions of fairness that are discordant with the underlying causal nature of discrimination and often "lead to misleading and undesirable outcomes" (Loftus, Russell, et al. 2018). Precisely, causal inference allows us to target specific forms of illegal discrimination by identifying source(s) of unfairness that bias predictions through explicit statements about the causal relationships in the data, thereby prompting an intuitive framework for algorithmic fairness that not only begets more straightforward solutions, but is conceptually more consistent with the moral and legal principles of equity and justice that govern society.

Several, and increasingly more recent, works have adopted a causal understanding of (un)fairness and have thus turned to the principles of causal inference to help guide specific conceptual and methodological frameworks for algorithmic fairness. Propensity score analysis (PSA)—a statistical tool for estimating causal effects from observational data subject to measured confounding—is one such route that has been proposed as a causal approach to discovering and removing (direct) discrimination from biased data when training or designing fair models. For example, Calders et al. (2013) implement propensity-score-based stratification as a data mining strategy for linear regression that controls for explainable group-level differences so as to isolate the (unjust) discriminatory effects in the data. Similarly, Qureshi et al. (2019) introduce a method for (individual) discrimination discovery that appeals to propensity score weighting as a way of achieving balance between group-wise covariate distributions "that eliminates the effect of observed confounding factors".

While several causally-motivated mechanisms for promoting algorithmic fairness have been proposed, only one formal "definition" and criterion for algorithmic fairness currently exists that synthesizes the benefits of causal reasoning that we've discussed. Originally proposed by Kusner et al., *counterfactual fairness* asserts that for some protected attribute A a predictor $\hat{Y}$ of outcome Y in a causal model—containing the set of observed variables V, the set of structural equations F that correspond[3] to each element of V, and a set of unobserved latent variables U not caused by any factor in V—is counterfactually fair if whenever $A = a$ and observed covariates $X = x$,

$$\mathrm{P}(\hat{Y}_{A\leftarrow a}(U) = y | X = x, A = a) = \mathrm{P}(\hat{Y}_{A\leftarrow a'}(U) = y | X = x, A = a),$$

for all y and any possible value a' of A. Barocas, Hardt, and Narayanan (2018) introduce this condition as *counterfactual demographic parity* for binary predictors due to its similarities to *conditional demographic parity*, stating counterfactual fairness is achieved when

$$\mathrm{E}[\hat{Y} | X = x, A = a] = \mathrm{E}[\hat{Y} | X = x, A = a'],$$

for all feature settings x and groups a, a'. Under this simplified formulation of counterfactual fairness, authors point out that the most straightforward way of preventing discrimination is to use predictors

[3]Namely, $F = \{f_1, f_2, ..., f_n\}$ such that for each $f_i \in F$ and $V_i \in V$, $V_i = f_i(pa_i, U_{pa_i})$, where $pa_i \subseteq V \backslash \{V_i\}$ and $U_{pa_i} \subseteq U$ denote the observed and unobserved parents of V_i respectively.

that are independent of the sensitive attribute, namely, restricting $\hat{Y}$ to non-descendants of A in the corresponding causal graph (Barocas, Hardt, and Narayanan 2018).

This conceptualization of fairness captures the intuition that a particular decision is fair if an individual's outcome in the actual world is the same as it would have been had they belonged to a different social group. That is, if group membership has no causal effect on the outcome (Loftus, Kusner, et al. 2018). Mathematically, following the equations above, such an analysis of counterfactual quantities corresponds to performing an *intervention* or "hypothetical action" on A by substituting $A = f(P, U)$ with a particular $A = a$ in a *structural causal model* $\mathcal{M}$ and assessing the consequent change in the outcome. Using Pearl's *do*-calculus to formalize interventions, we say that A has no causal effect on $\hat{Y}$ if performing an intervention on A results in no immediate change on the outcome when all other covariates X are held fixed. Depending on the causal effect measure of interest, this could amount to

$$\mathrm{E}[\hat{Y}|X = x, do(A = a)] - \mathrm{E}[\hat{Y}|X = x, do(A = a')] = 0,$$

$$\text{or} \quad \frac{\mathrm{E}[\hat{Y}|X = x, do(A = a)]}{\mathrm{E}[\hat{Y}|X = x, do(A = a')]} = 1,$$

where "$do(A = a)$ expresses an intervention on A at level a" (Loftus, Kusner, et al. 2018). In this way, counterfactual analysis allows us to mathematize discrimination as it is understood within the legal system.

IV. Path-Specific Effects

A common extension to counterfactual fairness, is the focus on *path-specific effects* (PSEs), which are often used in mediation analysis to explain the ways, both direct and indirect, in which a given variable affects an outcome (Nabi, Malinsky, and Shpitser 2022). Understanding a causal model as a set of relationships among factors enables us to visualize the network as a causal diagram, typically a *directed acyclic graph* (DAG), in which nodes and arrows are used to represent variables and their corresponding causal relationships. Hence, PSEs provide a framework for decomposing the *total effect* of a variable on an outcome into direct and indirect "components associated with particular causal pathways" (Nabi, Malinsky, and Shpitser 2022). That is, the effect a predictor has directly on the outcome as well as that mediated by another variable(s), namely, through a specific path or set of paths in a causal DAG. These causal parameters are often referred to as the *natural direct effect* (NDE) and the *natural indirect effect* (NIE), respectively. Although, both are simply specific types of PSEs—effects that operate through given pathway(s) in the model, while blocking all other paths between the exposure and the outcome.

In the context of fairness, we are thus interested in specific paths that reflect part of the total effect of some group membership A on an outcome Y. For example, suppose we want to train a model to aid in hiring decisions for a recent job opening at a company and that acceptance decision Y is modeled as a function of religious affiliation A and level of education M of the applicant, where A is presumed to be a personal trait that is either possessed or not possessed by individuals, per the example given in (Barocas, Hardt, and Narayanan 2018). Let us assume for demonstration purposes that (i) previous applicant data reflects correlation between A and M, which expert knowledge suggests could be related to the idea that one's decision to pursue higher education is partially influenced by their (religious) beliefs and/or social peer groups; and (ii) the fact that A may influence choices regarding one's appearance and/or social practices, which invites the potential for direct discrimination in its being observable to employers, leading us to assume a DAG similar to the one shown in Figure 1, where X is a set of baseline confounders. Specifically, while a direct effect of A on Y along path $A \to Y$ would amount to direct discrimination, an indirect effect of A on Y along path $A \to M \to Y$ would reflect indirect discrimination via mediator M. Hence, maximal fairness in this setting is achieved when both direct and indirect effects are 0. For this reason, we can think

of any pathway connecting A to Y as a potential conductor of unfairness in that it is subject to transmitting a non-zero discriminatory effect. Our task then becomes to constrain such paths along which discrimination may be passed on to the outcome so as to train a model whose predictions are both optimally accurate and maximally fair. The appeal of path-specificity for algorithmic fairness lies not just in the fact that we are no longer restricted to using non-descendants of A to predict Y, but also in that it provides a solid framework for understanding the mechanisms that drive discrimination as direct and indirect causes that can be targeted more explicitly.

We note that this approach to fairness requires that the PSEs of interest be *identifiable*. That is, the requirement that our causal parameters be a function $g(\cdot)$ of the data distribution $p(V)$, as induced by the causal model. As stated in Nabi, Malinsky, and Shpitser (2022), interventional distributions in the structural causal model of a DAG, such as $p(Y(a, M(a')))$ in potential outcome notation, may be estimated via *standardization* or the *parametric g-formula* under given identifiability conditions[4]. Where $Y(a, M(a'))$ gives the value of Y when A is set to a and M is set to the value it would have attained had A been set to a', $p(Y(a, M(a')))$ is thus given by

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

Following the same notation, the NDE is then defined as $\mathrm{E}[Y(a, M(a'))] - \mathrm{E}[Y(a')]$, while the NIE is given by $\mathrm{E}[Y(a)] - \mathrm{E}[Y(a, M(a'))]$ in the expectation-difference scale. In turn, a positive effect of any kind would indicate the presence of discrimination against group $A = a'$ in terms of Y. That is, in our hiring decisions example, discrimination against applicants with religious affiliation, i.e., individuals for whom $A = 1$, would hence be characterized by a positive direct and/or indirect effect when $a = 0$ and $a' = 1$ under the given definitions.

V. Methodology

With an understanding of path-specificity as a decomposition of discriminatory effects within the data, which tie group membership to the outcome, into direct and indirect components, we tackle the problem of fair prediction from the perspective of maximum likelihood estimation (MLE), adapting the work of Nabi, Malinsky, and Shpitser (2022), which introduces a method for training fair predictive models appealing to this notion of path-specific effects to prevent discriminatory predictions. Specifically, given an observed data distribution $p(Y, Z)$, authors propose to transform the inference problem on $p(Y, Z)$, or what they call the "unfair world", into one on another distribution $p^*(Y, Z)$, which they refer to as the "fair world"; a distribution that is minimally close to $p(Y, Z)$, while also having the property that the discriminatory effect is bounded by some pre-determined (ϵ_l, ϵ_u).

While Nabi, Malinsky, and Shpitser (2022) implement a semi-parametric modeling approach with an imposed constraint on a particular discriminatory effect—what we call an "unfair PSE"—that involves a complex reparameterization of the data likelihood for constrained optimization over the entire distribution, we choose to pursue a penalty-based method maintaining the semi-parametric modeling framework, not just for simplicity, but for making human judgment a necessary element of modeling fairly and effectively. That is, rather than imposing a specific fairness constraint on the discriminatory effect, we take the route of penalizing the unfair PSE as an added term to the main objective function, thereby overcoming the need to reparameterize the likelihood and know the fairness bounds a priori. This way, our approach also gives decision makers more flexibility to define optimality contextually via empirical analysis. Moreover, with our aims to shrink discriminatory effects to 0, we decide to impose an L1 penalty on the unfair PSE.

Assuming a likelihood objective function $\mathcal{L}_{Y,Z}(\mathcal{D}; \beta)$ parameterized by the maximum likelihood estimates β for some dataset $\mathcal{D} = \{Y_i, Z_i, i = 1, 2, ..., n\}$ drawn from the unfair world $p(Y, Z)$, we

[4]See Pearl (2012a) and Lange, Rasmussen, and Thygesen (2014) for additional information on identifiability conditions for natural effects.

propose to approximate $p^*(Y, Z)$ by solving the following unconstrained optimization problem of the penalized likelihood,

$$\hat{\beta}_{\text{opt}} = \arg\max_{\beta} \mathcal{L}_{Y,Z}(\mathcal{D}; \beta) + \lambda_{\text{opt}} \Big| \hat{g}(p(Y, Z; \beta)) \Big|,$$

where $\hat{g}(p(Y, Z; \beta))$ is the unfair PSE estimator and λ_{opt} denotes the optimal value of the tuning parameter. Generally, we should consider the optimal value of λ as being the smallest value to yield maximum likelihood estimates $\hat{\beta}_{\text{opt}}$ for which the target discriminatory effect is minimally close to 0. However, we let λ, and hence $\hat{\beta}$, be variable quantities to allow decision makers to observe the trade-offs between accuracy and fairness that correspond to changes in the tuning parameter as a way to encourage more appropriate modeling choices for given settings. Focusing specifically on the negative log-likelihood for practicality, the optimization problem above translates to,

$$\hat{\beta}_{\text{opt}} = \arg\min_{\beta} \; -\log(\mathcal{L}_{Y,Z}(\mathcal{D}; \beta)) + \lambda_{\text{opt}} \Big| \hat{g}(p(Y, Z; \beta)) \Big|.$$

As mentioned in the previous section, we appeal to the g-formula to compute the PSE of interest. In the case that we are interested in the NDE for penalizing direct discrimination against individuals with sensitive attribute $A = a'$, we would have

$$g(p(Y, Z; \beta)) = \mathrm{E}[Y(a, M(a'))] - \mathrm{E}[Y(a')],$$

where $Z = \{A, M, X\}$ and $p(Y(a, M(a'))) = \sum_{X,M} p(Y|a, M, X)p(M|a', X)p(X)$, for outcome Y, sensitive attribute A, mediator M, and baseline confounder X, as depicted in Figure 1. We note that while we model counterfactual densities $p(Y|a, M, X)$ and $p(M|a', X)$ parametrically, the g-formula does not require parametric modeling of confounder distributions $p(X)$ so long as variable(s) X are neither time-varying nor affected by A (Hernan and Robins 2020). Hence, following Nabi, Malinsky, and Shpitser (2022), we model $p(X)$ nonparametrically, resulting in a semi-parametric model of causal parameters. As Hernan and Robins (2020) explain, this approach allows us to approximate the cumulative distribution function (CDF) of X nonparametrically by averaging over its observed values, i.e., its *empirical distribution*, rather than having to integrate over the confounder distribution directly to identify the PSE when X is continuous. As shown below, this simply amounts to placing a weight of $1/n$ over every point x_i in the g-formula. In turn, we can define the NDE as do Nabi and Shpitser (2017), which corresponds to an MLE plug-in estimator for the causal mediation formula given by Pearl (2012b). That is,

$$\hat{g}(p(Y, Z; \beta)) = \frac{1}{n} \cdot \left[\sum_{i=1}^{n} \sum_{M} \big(\mathrm{E}[Y|A = a, M, X_i] - \mathrm{E}[Y|A = a', M, X_i] \big) \cdot p(M|A = a', X_i) \right].$$

This technique is especially helpful when dealing with messy high-dimensional data that reflects multiple observed confounders and/or covariate levels (Hernan and Robins 2020). Particularly, Pearl (2012a) provides a useful guide for identifying direct and indirect causal effects under various conditions and structures of the data for which our approach may be easily adapted to accommodate new scenarios.

Given that constraints on the PSEs may "involve nonlinear and complicated functionals of the observed data distribution", Nabi, Malinsky, and Shpitser (2022) propose to reparameterize the semi-parametric likelihood to allow for more efficient constrained optimization in these settings. Moreover, they propose to constrain the entire data likelihood, rather than averaging over the constrained variables, thus including the unspecified empirical weights $p_i = p(X_i = x_i)$ as parameters to be optimized heuristically. However, while authors claim that this combined approach leads to improved performance, the heuristic procedure used to obtain both causal and empirical weight parameters is susceptible to bias and may not always yield an optimal solution that can be used in practice. Conversely, by shifting complexity to the objective function, our proposed penalty-based

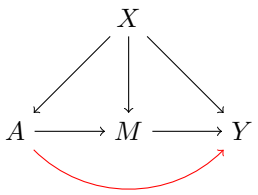

Figure 1: Example DAG for Direct Discrimination

method avoids having to reparameterize the likelihood for optimization and relaxes the assumption that fairness bounds are known a priori, all while allowing PSEs to take on nonlinear functional forms of the data distribution within a simpler unconstrained optimization framework. Moreover, our proposed penalty-based method, which integrates the regularized PSE into the semi-parametric likelihood, gives decision makers an opportunity to visualize the costs of fairness in response to changes in the tuning parameter for the purpose of encouraging well-informed, explainable choices that simultaneously satisfy given contextual objectives and adhere to principles of justice.

VI. Application to Direct Discrimination with Simulated Data

In a typical real-world setting, we might start by positing a reasonable structural causal model and graph, following exploratory data analysis and prior subject-matter knowledge of relevant features. Within a standard generalized linear modeling framework, it may then be of interest to undergo model selection procedures to better capture variable relationships as they appear in the actual world. Maybe X is tied to A and Y, but not to M, or perhaps the interaction between A and M is particularly significant for predicting Y. In such cases it's important to engage in traditional parametric modeling practices to protect against model misspecification and ensure our results are as accurate as possible.

Given that we are simulating our own biased data, we specify our unconstrained model parameters beforehand to have complete control over the discriminatory effects for demonstration purposes. Specifically, for this simple example, we generate data $\mathcal{D} = \{X, A, M, Y\}$ according to the DAG shown in Figure 1, with $n = 1,000$ observations drawn from $p(Y, Z)$, as induced by the following structural causal model,

$$
\begin{aligned}
X &= f_X(\varepsilon_X), \\
A &= f_A(X, \varepsilon_A), \\
M &= f_M(X, A, \varepsilon_M), \\
Y &= f_Y(X, A, M, \varepsilon_Y),
\end{aligned}
$$

where Y is our continuous outcome variable and $Z = \{X, A, M\}$ includes continuous baseline confounder X, binary sensitive attribute A, and binary mediator M.

Assuming the following distributions, we initialize our target parameters β according to the values given in Table 2 so as to generate our data such that a traditional generalized linear modeling procedure would yield comparable maximum likelihood estimates.

$$
\begin{aligned}
X &\sim \mathrm{N}(0, 1) \\
A &\sim \mathrm{Bin}(n, p_{a|x}) \\
M &\sim \mathrm{Bin}(n, p_{m|x,a}) \\
Y &= p_{y|x,a,m} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, 1)
\end{aligned}
$$

Table 2: Initial Parameter Values for Unfair Data

$\beta_{a,0}$	$\beta_{a,x}$	$\beta_{m,0}$	$\beta_{m,x}$	$\beta_{m,a}$	$\beta_{y,0}$	$\beta_{y,x}$	$\beta_{y,a}$	$\beta_{y,m}$
0.5	0.5	0.5	1	0.5	1	1	5	1

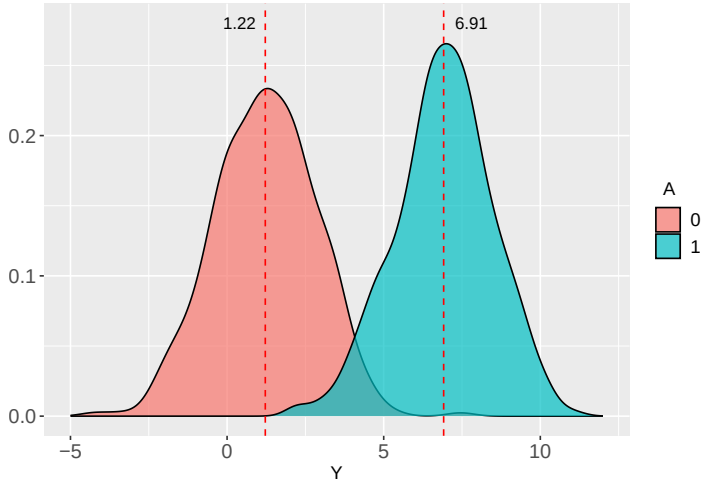

Figure 2: Unfair World Outcome Distributions by Group

$$p_{a|x} = f_A(X) = \frac{e^{(\beta_{a,0}+\beta_{a,x}X)}}{1+e^{(\beta_{a,0}+\beta_{a,x}X)}}, \quad \log\left(\frac{p_{a|x}}{1-p_{a|x}}\right) = \text{logit}(p_{a|x}) = \beta_{a,0}+\beta_{a,x}X$$

$$p_{m|x,a} = f_M(X,A) = \frac{e^{(\beta_{m,0}+\beta_{m,x}X+\beta_{m,a}A)}}{1+e^{(\beta_{m,0}+\beta_{m,x}X+\beta_{m,a}A)}}, \quad \log\left(\frac{p_{m|x,a}}{1-p_{m|x,a}}\right) = \text{logit}(p_{m|x,a}) = \beta_{m,0}+\beta_{m,x}X+\beta_{m,a}A$$

$$p_{y|x,a,m} = f_Y(X,A,M) = \beta_{y,0}+\beta_{y,x}X+\beta_{y,a}A+\beta_{y,m}M$$

Given that A, M, and Y each assume a simple main effects model of parental nodes, the unfair PSE that quantifies direct discrimination is simply the coefficient $\beta_{y,a}$. That is, the direct effect of group membership A on our outcome Y, which can be interpreted as the expected increase in Y for an individual in the positive sensitive attribute class (i.e., $A = 1$) compared to an individual in the negative class (i.e., $A = 0$), adjusting for X and M in the model. This can partially be observed in Figure 2, which shows the difference in unadjusted groups-specific sampled outcome distributions. As expected, Y values are more than 5 units higher for those in the positive sensitive attribute class compared to those in the negative class—specifically, $\text{E}[Y|A=1] - \text{E}[Y|A=0] \approx 6.91 - 1.22 = 5.69$. While we know that approximately 5 units of the difference in outcome means between groups results from direct discrimination alone, we assume that additional group-level variation in Y comes from baseline differences in X and indirect influence. Simple regression adjustment for covariates confirms this in showing that $\text{E}[Y|A=1,M,X] - \text{E}[Y|A=0,M,X] \approx 4.98$. In the context of statistical non-discrimination criteria, since $\text{E}[Y|A=1,M,X] \neq \text{E}[Y|A=0,M,X]$, it follows that outcomes between groups are illegally unfair on average. Moreover, assuming that higher values of Y yield increased individual benefits, this distribution of the data indicates that individuals in the positive class have an unfair advantage solely attributed to their status of A. In a fair world, as we later show, both distributions should be approximately equal conditional on all other features.

Defining our initial optimization objective as the negative log-likelihood, we now have

$$\hat{A} = \frac{e^{(\beta_{a,0}+\beta_{a,x}X)}}{1+e^{(\beta_{a,0}+\beta_{a,x}X)}},$$

$$\hat{M} = \frac{e^{(\beta_{m,0}+\beta_{m,x}X+\beta_{m,a}A)}}{1+e^{(\beta_{m,0}+\beta_{m,x}X+\beta_{m,a}A)}},$$

$$\hat{Y} = \beta_{y,0} + \beta_{y,x}X + \beta_{y,a}A + \beta_{y,m}M,$$

$$f(a;\hat{A}) = a \cdot \hat{A} + (1-a)\cdot(1-\hat{A}),$$

$$f(m;\hat{M}) = m \cdot \hat{M} + (1-m)\cdot(1-\hat{M}),$$

$$f(y;\hat{Y}) = \frac{e^{-\frac{1}{2}(y-\hat{Y})^2}}{\sqrt{2\pi}},$$

$$-\log \mathcal{L}_{Y,Z}(\mathcal{D};\beta) = -\sum_{i=1}^{n} \log(f(a_i;\hat{A})) + \log(f(m_i;\hat{M})) + \log(f(y_i;\hat{Y})) + \log(1/n),$$

where the fact that $\hat{A}$, $\hat{M}$, and $\hat{Y}$ are themselves functions of β allows us to write the likelihood as being parameterized by β. Using the mediation formula, as previously discussed, we define our desired PSE estimator for direct discrimination as

$$\hat{g}(p(Y,Z;\beta)) = \frac{1}{n}\cdot\left[\sum_{i=1}^{n}\sum_{M\in\{0,1\}}\left(\mathrm{E}[Y|A^{a=1},M,X_i] - \mathrm{E}[Y|A^{a=0},M,X_i]\right)\cdot p(M|A^{a=0},X_i)\right].$$

In turn, we define our new objective function by adding the absolute value of the unfair PSE given by $\hat{g}(p(Y,Z;\beta))$ above as a penalty term to the negative log-likelihood with tuning parameter λ. Specifically, for $i = 0, 1, 2, ..., 150$, we solve the following unconstrained (nonlinear) optimization problem, setting $\lambda_i = 10i$, to obtain the corresponding set of causal parameter estimates $\hat{\beta}_i$, the negative log-likelihood, and the NDE at the i^{th} iteration.

$$\hat{\beta}_i = \arg\min_{\beta} \; -\log(\mathcal{L}_{Y,Z}(\mathcal{D};\beta)) + \lambda_i\left|\hat{g}(p(Y,Z;\beta))\right|$$

With optimization being carried out via the R package `nloptr`, we opt for the "Constrained Optimization BY Linear Approximations" algorithm or "COBYLA"—the only NLopt-adapted algorithm for local derivative-free optimization that supports nonlinear inequality and equality constraints as well as bound-constrained or unconstrained problems. Moreover, we specify stopping criteria `"xtol_rel"=1.0e-8` and `"maxeval"=10000`, which pertain to having reached the specified relative change in parameters and having exceeded the maximum number of function evaluations, respectively. Having minimized the objective for each λ_i, we construct the five plots given by Figures 3 and 4 to observe changes in the negative log-likelihood, mean squared error, and NDE as we increase the value of λ.

Noticeably, setting $\lambda_0 = 0$ results in the unpenalized maximum likelihood estimates corresponding to the unfair world distribution for which the negative log-likelihood and mean squared error are as small as possible and the NDE is approximately 5. Meanwhile, increasing values of λ yield fair world approximations that correspond to exponentially large values of the negative log-likelihood and NDEs that decrease at a constant rate of approximately $5/1090 \approx 0.0046$ per unit increase in λ or $5/109 \approx 0.046$ for each λ_{i+1}. Assuming, for our purposes, that the set of optimal parameters $\hat{\beta}_i$ correspond to the first λ_i for which the PSE is minimally close to 0, we let $\lambda_{\text{opt}} = \lambda_{109} = 1,090$ in accordance with three graphs in Figure 4. Hence, $\hat{\beta}_{\text{opt}} = \hat{\beta}_{109}$, the values for which are given in Table 3.

Table 3: Optimal Parameter Values for Fair Data

$\beta_{a,0}$	$\beta_{a,x}$	$\beta_{m,0}$	$\beta_{m,x}$	$\beta_{m,a}$	$\beta_{y,0}$	$\beta_{y,x}$	$\beta_{y,a}$	$\beta_{y,m}$
0.5263	0.5899	0.5667	0.9167	0.3496	3.7603	1.5124	0.0006	1.5683

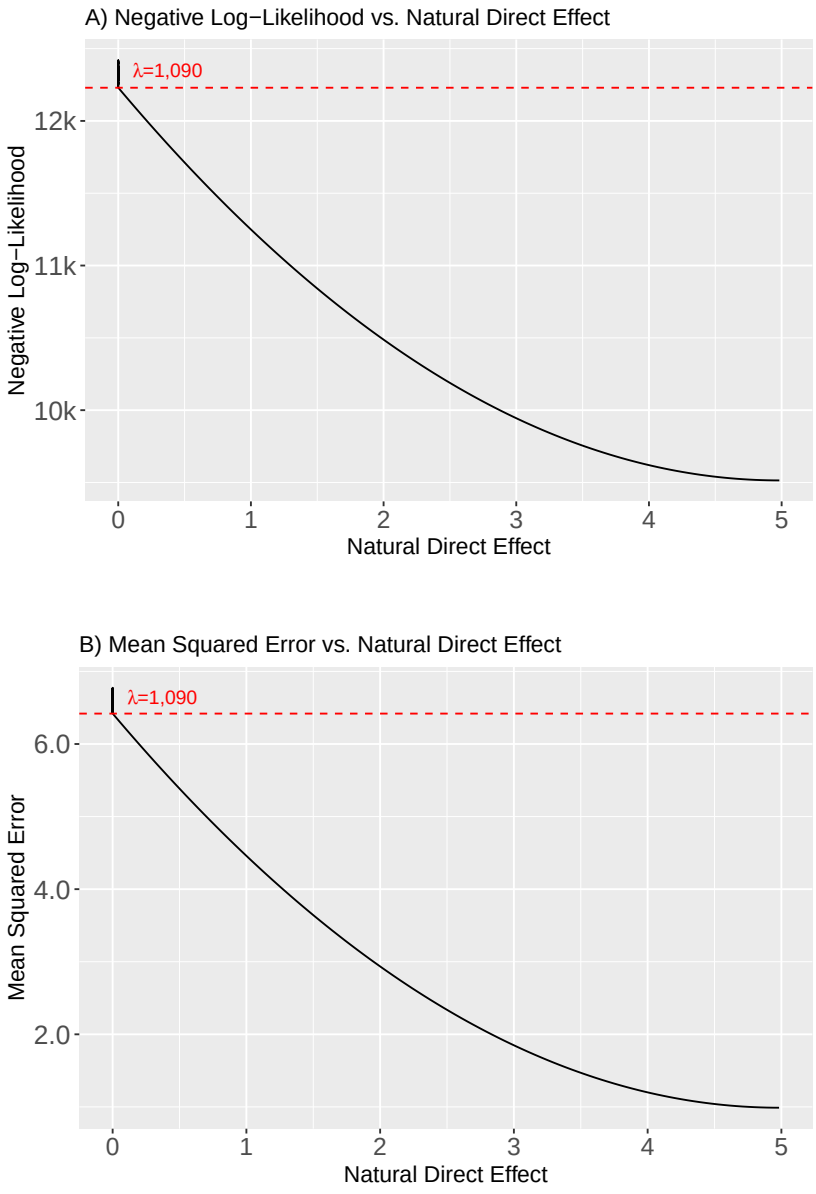

Figure 3: Negative Log-Likelihood & Mean Squared Error vs. Natural Direct Effect

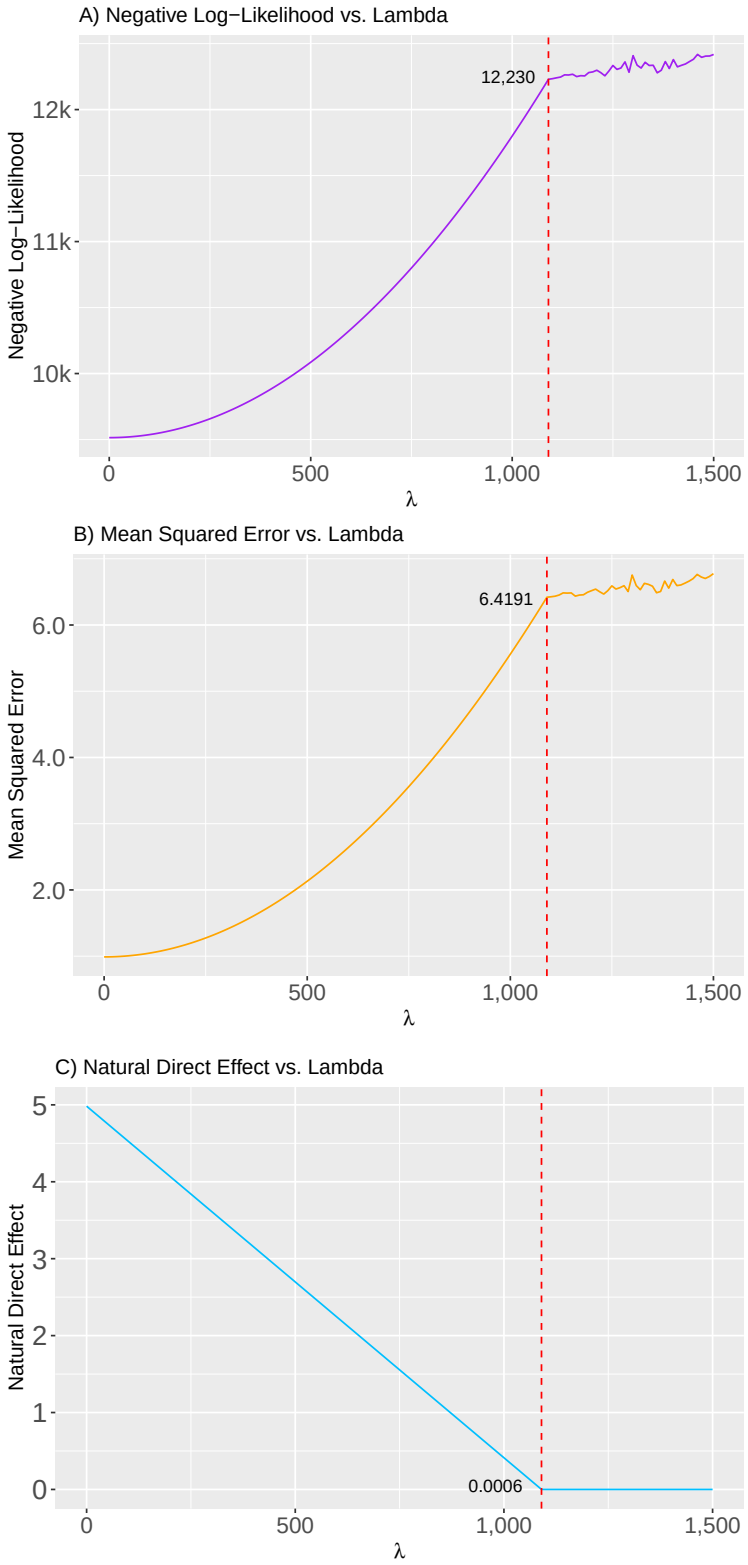

Figure 4: Negative Log-Likelihood, Mean Squared Error, & Natural Direct Effect for Varying Lambda

Table 4: Regression Adjusted Estimates of Discriminatory Effect

	Estimate	Std. Error	t-value	Pr(>|t|)
Adjusted	-0.01577	0.06933	-0.227	0.82
Unadjusted	1.00370	0.14020	7.159	0.00

With our established set of optimal parameters, we proceed with modeling $p^*(\hat{Y}, \hat{A}, \hat{M}, X)$ as an approximation to the hypothetical fair world distribution $p^*(Y, Z)$ by sampling $n = 1,000$ new observations for $\{\hat{A}, \hat{M}\}$ to obtain a vector of fair predictions according to $\hat{Y}$. Figures 5 and 6 roughly give the corresponding group-wise unadjusted and conditional predictive distributions by group. Specifically, we notice that

$$\mathrm{E}[\hat{Y}|A = 1] - \mathrm{E}[\hat{Y}|A = 0] \approx 5.15 - 4.15 = 1,$$

$$\mathrm{E}[\hat{Y}|A = 1, M, X] - \mathrm{E}[\hat{Y}|A = 0, M, X] \approx 4.58 - 4.39 = 0.19.$$

We note that despite being close to 0, the latter causal effect, estimated according to Figure 5, is biased due to our concise partitioning of X for visualization purposes. Unadjusted regression, as shown in Table 4, confirms that we can expect a one-unit difference in prediction means between groups when additional factors are not considered. Conversely, regression adjustment for covariates demonstrates that there is in fact no statistically significant influence of A on $\hat{Y}$ directly. Thus, it follows that the predictor $\hat{Y}$ is counterfactually fair with respect to direct path $A \to Y$ for groups in A on average. That is, since $\mathrm{E}[\hat{Y}|A = 1, M, X] \cong \mathrm{E}[\hat{Y}|A = 0, M, X]$, new predictions no longer reflect any illegal forms of discrimination.

On a broader level, we should also aim to protect groups and individuals against indirect forms of discrimination in automated decision making, despite not being explicitly required by law. That is, when discrimination is systemic in nature and/or mediated by additional factors, we have, if not a legal obligation, a moral duty to construct predictive models that actively fight injustice wherever present. In what follows, we outline some possible extensions to our proposed method, in addition to mitigating implicit bias, that make this work increasingly more applicable.

VII. Extensions

Against the backdrop of counterfactual fairness, mediation analysis and path-specificity offer an unmatched tool for dissecting, quantifying, and formally targeting the various facets of injustice that plague data-driven, predictive algorithms. What sets this particular framework apart is its ability to provide intuitive and causally-sound realizations of variable relationships so as to uncover the hidden mechanisms that perpetuate unfairness. As previously mentioned, systemic forms of discrimination can (and should) be scrutinized in this way and explicitly addressed when training predictive models. Like the methods introduced in Nabi, Malinsky, and Shpitser (2022), our approach may be easily adapted to remove such implicit biases by redefining the unfair PSE via g-computation. Specifically, Pearl's mediation formula can be extended to target unique indirect paths, thus allowing us to express NIEs as functionals of the likelihood as we saw previously in the case of the NDEs. Paired with solid background knowledge, these techniques allow us to gain a more robust understanding of *how* discrimination operates within our data through the lens of broader social, political, and historical contexts.

Although our worked example consisted of continuous baseline and target features as well as binary mediator and protected attribute variables, our approach will readily accommodate for all types of covariate and outcome structures in addition to non-binary sensitive attribute classes. More than allowing for our data to assume various forms however, our methods can also be modified to incorporate the element of intersectionality. Precisely, when discrimination pertains not to one but to

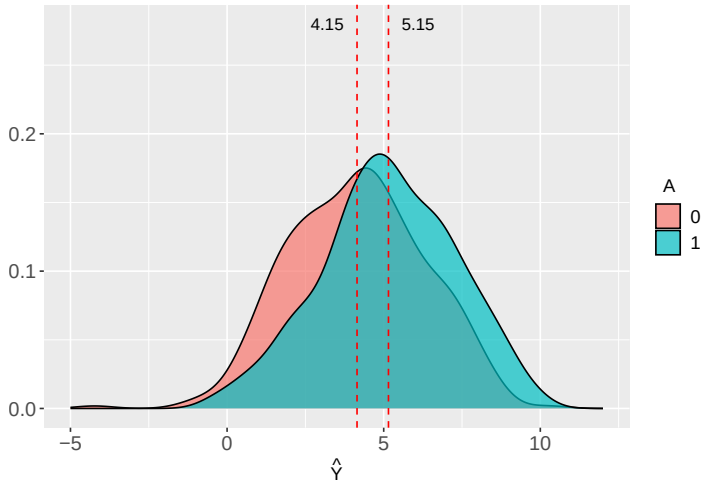

Figure 5: Fair World Prediction Distributions by Group

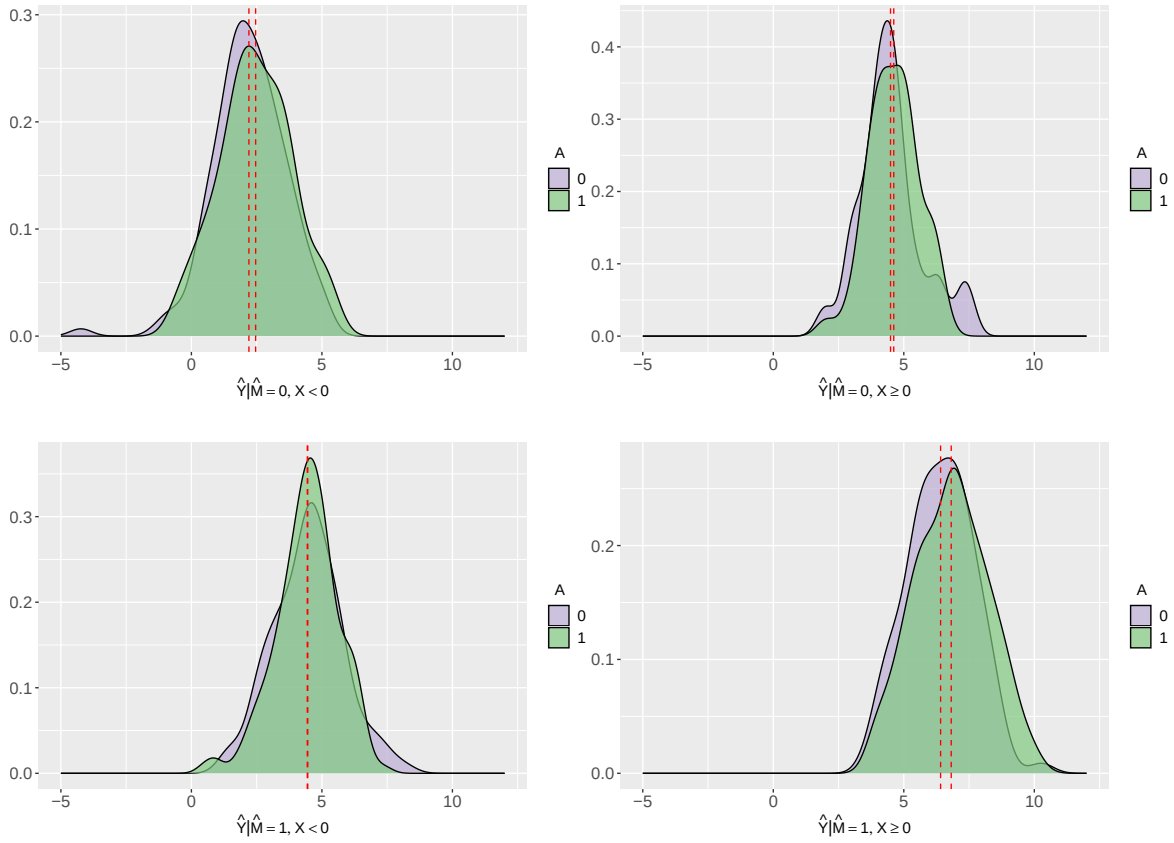

Figure 6: Stratified Fair World Prediction Distributions by Group

multiple sensitive attributes in tandem, this approach can be used to target such "unique forms of disadvantage that members of multiple protected categories may experience" (Barocas, Hardt, and Narayanan 2018). As an example, suppose that our data contains two binary sensitive features $A_1 = \{0, 1\}$ and $A_2 = \{0, 2\}$ with discrimination present for a subset of individuals with $A_1 = 0$ and $A_2 = 0$. Depending on the nature of the protected attribute relationship and the context of the problem, a straightforward way of addressing unfairness of this kind may be to treat these features as a single variable A with multiple categories. For instance, we could let $A = A_1 + A_2$ such that each value in $A = \{0, 1, 2, 3\}$ corresponds to one of four unique combinations of sensitive attribute values. It should be noted however that a counterfactually fair predictor for a combined set of attributes does not guarantee fair predictions between groups of a single protected variable. Similarly, achieving counterfactual fairness for a single sensitive feature along a given set of paths could result in further disadvantage towards intersectional subgroups. Therefore, meticulous data analysis, together with careful selection of relevant features and a prior understanding of variable relationships, is critical for fair and accurate causal modeling—a fortiori when what it means for outcomes to be fair, and for *whom*, is itself subject to change.

Despite the necessary assumption of (sequential) ignorability in causal effect estimation, the potential for unmeasured confounding can never be entirely ruled out when using observational data. This is especially true in the context of fairness, where effects are viewed in light of sensitive attributes, many of them socially constructed, that bear complex historical ties with various facets of people's lives. For this reason, it's important to consider additional measures to assess the robustness of our results and/or reduce susceptibility to unobserved confounding of the protected variable and outcome relationship that may lead to biased discriminatory effect estimates in practice. For example, Nabi, Malinsky, and Shpitser (2022) discuss a graphical condition under which we may still derive causal parameters in the case where we have unmeasured (but perhaps known) confounders. Specifically, authors demonstrate how failing to satisfy what is known as the *recanting witness criterion*, allows us to nonparametrically identify PSEs via the *edge g-formula* under given assumptions. While such a condition for PSE identifiability in the presence of unobserved confounding broadens the range of cases to which our methods may be applied, it may however be of additional interest to allow for sensitivity analyses so as to ensure our estimates remain as unbiased as possible—especially in situations where the recanting witness criterion may not apply. In this regard, Kilbertus et al. (2019) discuss the use sensitivity analysis to minimize the potential for biased causal effect estimates due to model misspecification, precisely through unmeasured confounding. In particular, authors formalize the notion of unmeasured confounding as non-zero correlations between error variables, providing techniques to assess their impact in causal non-linear additive noise models (ANMs). Though applied within an alternative modeling scheme, this tool provides a promising avenue for extending and solidifying our approach in the immediate future. To more explicitly target the inescapable causal dilemma of unmeasured confounding, Helwegen, Louizos, and Forré (2020) propose a similar approach to training fair predictive models, that like the proposed work of Loftus, Kusner, et al. (2018), assumes latent representations of the unobserved variables. Specifically, authors employ recent advances in variational inference to better address the issue of unobserved confounding, appealing to what is known as a causal effect variational autoencoder (CEVAE) to more efficiently estimate PSEs in these restricted settings. Not only does such a method formally account for unmeasured confounding, but it also has the advantage of not requiring direct computation of PSEs, which can become increasingly difficult in more complicated scenarios. While we believe in and stand by the benefits of our approach, the latent variable framework offers a promising direction for future research on fair predictive modeling that is worth exploring—particularly with regards to explainability of solutions and the ability to hold decision makers accountable for their consequences.

VIII. Limitations & Considerations

While our approach as a whole lays claim to several strengths, it also bears a handful of limitations we ought to consider in order to ensure its responsible and effective application in practice. First, the adopted counterfactual path-specificity framework relies heavily on strong and often untestable identifiability assumptions, including that of (sequential) ignorability. Specifically, as is typical in mediation analysis, requiring that unfair PSEs be indentifiable, by extension calls for us to assume the absence of unmeasured confounding—a condition that is both impossible to prove and difficult to satisfy in practice without running the risk of producing biased discriminatory effect estimates. While the extensions previously discussed may lessen the sensitivity to unmeasured confounding, they do not ultimately guarantee unbiased prediction. On the one hand, these conditions set a higher standard for fairness, which more than requiring decision makers to have significant prior knowledge of variable interactions, could limit the situations to which these methods may apply. On the other hand, they bring caution, awareness, and understanding into the decision making process, encouraging users to make careful and justifiable modeling choices in consideration of those affected by decisions and the ways in which they may be affected.

Second, as discussed by Hernan and Robins (2020), semi-parametric modeling with standardization is not entirely robust to model misspecification. That is, our semi-parametric approach and use of the parametric g-formula for discriminatory effect computation requires that the outcome model conditional on covariates be correct. Otherwise, such an estimator for the unfair PSE is highly prone to bias. While this is of primary concern in the face of time-varying confounders and parsimonious outcome regression models, such as the linear main effects model assumed for Y in our example, we may get a sense for the extent to which our model suffers from misspecification by comparing standardized effect estimates to those obtained using an IP weighted estimator, or by implementing a bootstrapping technique to compute discriminatory effect standard errors (Hernan and Robins 2020). Alternatively, sensitivity to model and graph misspecification can be minimized by adopting a doubly-robust estimator, which only requires that either model for the outcome or sensitive attribute be correct. Although, for our purposes, such an estimator would need to be capable of being written as a functional of the semi-parametric likelihood to be compatible with the optimization objective. Ultimately, given that some degree of model misspecification is to be expected no matter the situation, it goes without saying that subject-matter judgement and "healthy skepticism" are critical aspects to any and all causal inferences.

While these inescapable constraints underlie the counterfactual PSE paradigm as a whole, there are additional drawbacks specific to our adaptation of Nabi, Malinsky, and Shpitser (2022) that are also worth addressing. For instance, our penalized maximum likelihood estimation approach to identify contextually-optimal parameters relies on recursive optimization for different values of λ, which comes at high computational costs. In particular, repeated (and potentially nonlinear) optimization, even in the rather simple unconstrained setting, can become an increasingly time-consuming and resource-intensive process depending on the number of iterations, dimensions of the data, and the software used for optimization. Moreover, despite providing insight into the trade-offs between accuracy and fairness, the fact that our approach requires users to define optimality based on their prediction goals is itself subject to personal or subjective biases. To address the first concern, which we acknowledge has no straightforward solution, we argue that whenever fairness is central to the objective, computational cost should not play such a critical role in choosing an adequate training method(s). Although computational efficiency and speed are highly desirable properties in ML, unlike utility and explainability, they are not necessary conditions for fairness. For this reason, their absence, within reason, does not in itself constitute sufficient grounds for rejecting the use of this approach—especially when the degree to which decisions impact people's lives demands that decision makers be transparent about their choices and that equity remain at the forefront of the prediction task. Thus, while we should aim to minimize computational complexity wherever possible, we must not let such losses in practicality outweigh the benefits of effectively protecting groups and

individuals from the harms of discrimination.

Importantly, human engagement is also integral to achieving fairness in automated decision making. Given that algorithms have no understanding of the moral or legal implications of discrimination, it is crucial to incorporate the element of human judgement and intervention into the automated decision making process—not only to prevent the transfer of hidden biases from the data to prediction, but to bring forth a level of accountability on behalf of users for the decisions made thereafter. Aside from making users the sole determinants of optimal parameters, this could take the additional form of involving experts in the development and testing of algorithms for example, and/or providing avenues for individuals to appeal or challenge automated decisions, as well as conducting ongoing monitoring of our models to identify and address any biases that may arise. As mentioned however, we must also recognize that human judgement and intervention are not immune to bias themselves. That is, people's own prejudices can influence their perceptions of what is optimal in certain situations, even though they may not be consciously aware of it nor have the capacity to mitigate these biases. As such, we note that the act of visualizing the trade-offs between accuracy and fairness to define optimality in itself makes it more difficult for such biases to affect decisions in that it puts users in a position where they must be able to justify their choices and explain how these choices beget outcomes that individuals can reasonably accept as fair. It is nonetheless important however, to train predictive models in a way that balances the benefits of automation with the need for human oversight, as well as to involve multiple parties in the continual evaluation and improvement of such systems to ensure they are achieving optimal outcomes for **all** stakeholders.

IX. Conclusion

As we approach an increasingly automated world, powered by data imprinted with our nation's legacies of oppression and corrupt institutions, it is vital to our collective functioning as free and equal citizens of a democratic society that we make the conscious effort to prevent the production and reproduction of (discriminatory) bias within our systems—especially those which serve as tools for consequential decision making. It is in this light that we decided to tackle the project of fair prediction from the angle of causality, proposing an intuitive method for optimally training predictive models that will satisfy the standards of accuracy and fairness assumed in given contexts. In particular, we followed the work of Nabi, Malinsky, and Shpitser (2022) to present a parallel approach that appeals to path-specific counterfactual fairness and simplifies the optimization problem by penalizing, rather than constraining, the unfair PSE. By proposing to modify the objective function directly to include an L1 penalty on the PSE as an added term, we showed how one may transform what was previously a constrained MLE problem into an unconstrained optimization task. In turn, this adaptation not only relaxes the assumption that fairness bounds are known a priori, but it avoids complex reparameterizations of the semi-parametric likelihood, while still allowing for potential nonlinear forms of the PSE.

Additionally, and perhaps more importantly, our approach offers a more flexible environment for decision making by encouraging users to define optimality contextually via empirical analysis of the trade-offs between accuracy and fairness, thereby also placing greater responsibility at the hands of decision makers and establishing a basis for more explainable outcomes by extension. This implicit requirement of increased human involvement in the decision making process is a key advantage to our approach. Precisely in that it challenges decision makers to interact with the model to make explainable choices that reflect not just their contextual aims, but the moral and legal requirements of fairness. Even when no optimal solution exists, our method tacitly holds users accountable for the outcomes of prediction, as they must be prepared to gauge the thresholds of accuracy and fairness to impose necessary judgement on potential parameters. As discussed in the previous section, while facilitating such an interface often comes with its own set of costs, we argue that the sociotechnical nature of algorithmic decision making and its life-altering downstream effects should provide sufficient reason to enforce this level of user interaction.

Moreover, we maintain that in order to capture the dynamism of fairness in automation, we must have a way to encode our knowledge of the social, political, and historical processes that shape our daily lives, and a language to talk about "what might be or what might have been"—the more qualitative aspects of prediction that also force us to ask ourselves what actions and decisions mean for groups and individuals in the context of equity and social justice. While we acknowledge the unavoidable limitations of causal inference, we argue further that counterfactual reasoning is currently the best and perhaps the only way to bridge this theoretical gap. As we've pointed out throughout this work, path-specificity and mediation analysis specifically, provide a means for better understanding the complex processes that drive various forms of unfairness and the mathematics to target them explicitly. However, it bears repeating that counterfactual discrimination reasoning is not yet a complete work. In addition to the points already discussed, several objections[5] to its sensitive attribute semantics and use as a formal framework for algorithmic fairness challenge its robustness and ability to capture and represent discrimination in a way that ensures desirable solutions. But, rather than discrediting its efficacy and potential, these critiques highlight important areas of improvement, providing valuable insight into where we ought to direct future research.

In this regard, we encourage the engagement of both technical and philosophical work in this space, including the aforementioned extensions to our proposed methods, as well as the pursuit of novel approaches to fairness in automation. That is, bearing in mind that any formalized scheme, its underlying assumptions, and the language it provides us to discuss concepts like "fairness", should always be approached with "a healthy level of skepticism". Ultimately, we hope that this work will inspire future discussions on the matter and motivate similar research endeavors by demonstrating how causality, or counterfatuality, when carefully used as a basis for reasoning about fairness, offers a promising methodological framework for explicitizing and manifesting the true demands of justice.

> "Counterfactuals are the building blocks of moral behavior as well as scientific thought. The ability to reflect on one's past actions and envision alternative scenarios is the basis of free will and social responsibility."
>
> — *Judea Pearl, "The Book of Why"*

[5]See Issa Kohler-Hausmann (2019) for some relevant objections and Howe et al. (2022) for important recommendations to consider.

References

Altman, Micah, Alexandra Wood, and Effy Vayena. 2018. "A Harm-Reduction Framework for Algorithmic Fairness." *IEEE Security & Privacy* 16 (3): 34–45. https://doi.org/10.1109/MSP.2018.2701149.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2018. "Fairness and Machine Learning Limitations and Opportunities." In.

Bonchi, Francesco, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. 2017. "Exposing the Probabilistic Causal Structure of Discrimination." *International Journal of Data Science and Analytics* 3 (February). https://doi.org/10.1007/s41060-016-0040-z.

Calders, Toon, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. "Controlling Attribute Effect in Linear Regression." In *2013 IEEE 13th International Conference on Data Mining*, 71–80. https://doi.org/10.1109/ICDM.2013.114.

Helwegen, Rik, Christos Louizos, and Patrick Forré. 2020. "Improving Fair Predictions Using Variational Inference In Causal Models." arXiv. https://doi.org/10.48550/arXiv.2008.10880.

Hernan, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Howe, Chanelle J, Zinzi D Bailey, Julia R Raifman, and John W Jackson. 2022. "Recommendations for Using Causal Diagrams to Study Racial Health Disparities." *American Journal of Epidemiology* 191 (12): 1981–89. https://doi.org/10.1093/aje/kwac140.

Issa Kohler-Hausmann. 2019. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination." *Northwestern University Law Review* 113 (5): 1163–1228.

Kamiran, Faisal, and Toon Calders. 2012. "Data Preprocessing Techniques for Classification Without Discrimination." *Knowledge and Information Systems* 33 (1): 1–33. https://doi.org/10.1007/s10115-011-0463-8.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. "Fairness-Aware Classifier with Prejudice Remover Regularizer." In *Machine Learning and Knowledge Discovery in Databases*, edited by Peter A. Flach, Tijl De Bie, and Nello Cristianini, 35–50. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33486-3_3.

Kilbertus, Niki, Philip J. Ball, Matt J. Kusner, Adrian Weller, and Ricardo Silva. 2019. "The Sensitivity of Counterfactual Fairness to Unmeasured Confounding." arXiv. https://doi.org/10.48550/arXiv.1907.01040.

Lange, Theis, Mette Rasmussen, and Lau Caspar Thygesen. 2014. "Assessing Natural Direct and Indirect Effects Through Multiple Pathways." *American Journal of Epidemiology* 179 (4): 513–18. https://doi.org/10.1093/aje/kwt270.

Loftus, Joshua R., Matt J. Kusner, Chris Russell, and Ricardo Silva. 2018. "Counterfactual Fairness." arXiv. https://doi.org/10.48550/arXiv.1703.06856.

Loftus, Joshua R., Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. "Causal Reasoning for Algorithmic Fairness." arXiv. https://doi.org/10.48550/arXiv.1805.05859.

Mancuhan, Koray, and Chris Clifton. 2014. "Combating Discrimination Using Bayesian Networks." *Artificial Intelligence and Law* 22 (2): 211–38. https://doi.org/10.1007/s10506-014-9156-4.

Mehrabi, Ninareh, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2021. "Attributing Fair Decisions with Attention Interventions." arXiv. https://doi.org/10.48550/arXiv.2109.03952.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. "A Survey on Bias and Fairness in Machine Learning." arXiv. https://doi.org/10.48550/arXiv.1908.09635.

Nabi, Razieh, Daniel Malinsky, and Ilya Shpitser. 2022. "Optimal Training of Fair Predictive Models." In *Proceedings of the First Conference on Causal Learning and Reasoning*, 594–617. PMLR.

Nabi, Razieh, and Ilya Shpitser. 2017. "Fair Inference On Outcomes," May. https://doi.org/10.48550/arXiv.1705.10378.

Pearl, Judea. 2012a. "Interpretable Conditions for Identifying Direct and Indirect Effects:" Fort Belvoir, VA: Defense Technical Information Center. https://doi.org/10.21236/ADA564093.

———. 2012b. "The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms." *Prevention Science* 13 (4): 426–36. https://doi.org/10.1007/s11121-011-0270-1.

Qureshi, Bilal, Faisal Kamiran, Asim Karim, Salvatore Ruggieri, and Dino Pedreschi. 2019. "Causal Inference for Social Discrimination Reasoning." arXiv. https://doi.org/10.48550/arXiv.1608.03735.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. "Mitigating Unwanted Biases with Adversarial Learning." arXiv. https://doi.org/10.48550/arXiv.1801.07593.