Fair Risk Minimization under Causal Path-Specific Effect Constraints

Razieh Nabi[1] **David Benkeser**[1]

[1]Departartment of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA

Abstract

This paper introduces a comprehensive framework for deriving and estimating fair optimal predictions in machine learning, grounded in causal and counterfactual path-specific effects as constraints. We detail the theoretical foundations of our approach, and provide closed-form solutions for constrained optimization within prevalent risk frameworks, including mean squared error and cross-entropy risks. These solutions conceptualize the fair risk minimizer as a nuanced adjustment to the unconstrained minimizer, influenced by the magnitude of the constraint, its canonical gradient, and the variance of this gradient. Additionally, we propose flexible semiparametric estimation strategies for these nuisance components, tailored to diverse model specifications. Such flexibility is essential for accurately implementing fairness adjustments across varied contexts. This work advances the discourse on algorithmic fairness by seamlessly integrating complex causal considerations into model training, thus providing optimal strategies for implementing fair models in real-world applications. **The full paper is on arXiv under the same title.**

INTRODUCTION

The discourse on fairness in machine learning encompasses various definitions, each shaped by distinct ethical considerations and operational implications [Mitchell et al., 2021, Barocas et al., 2023]. Counterfactual and causal reasoning frameworks are crucial for quantifying algorithmic fairness Zhang et al. [2017], Kusner et al. [2017], Zhang and Bareinboim [2018], Chiappa [2019], Nabi et al. [2019]. Our work builds prediction functions that satisfy constraints based on path-specific causal effects Nabi and Shpitser [2018], Nabi et al. [2022]. By leveraging causal inference, constrained optimization, and semiparametric statistics, we aim to develop an optimal predictive model by nullifying a specified path-specific effect. We frame fair optimal predictions as the outcomes of a penalized risk function, integrating fairness principles into the optimization process [Donini et al., 2018, Chamon et al., 2022, Nabi et al., 2024].

Our contributions to fair machine learning include introducing closed-form solutions for deriving optimal predictions subject to causal fairness constraints under mean squared error and cross-entropy frameworks. These solutions transform an unconstrained risk minimizer into a constrained one by incorporating the constraint, its gradient, and the variance of the gradient. We also provide a flexible semiparametric estimation strategy for the required fairness adjustments, accommodating diverse model specifications and enhancing the adaptability and robustness of fairness interventions across various data-generating contexts. This comprehensive framework equips practitioners with tools to effectively integrate fairness into predictive modeling, promoting societal equity and justice in algorithmic decisions.

PROBLEM FORMULATION

Consider the observed datum $O = (S, X, Y)$, with S indicating a sensitive attribute, X denoting other covariates, and Y being the outcome variable. Let $Z = (S, X)$. To facilitate our discourse, we adopt the framework of directed acyclic graphs (DAGs) to describe causal relationships among variables. The causal model implied by a DAG is often described by a set of nonparametric structural equation models with independent error terms (NPSEM-IE) [Pearl, 2009].

We define ψ_0 as the unconstrained minimizer of a relevant risk function $R_{P_0}(\psi_0)$, that is $\psi_0(z) = \text{argmin}_{\psi \in \Psi} \; R_{P_0}(\psi)$. Such risks are often formulated as the expectation of a loss function $L(\psi)$, $R_{P_0}(\psi) = \int L(\psi)(o) dP_0(o)$. We consider learning ψ_0 while adhering to a pre-defined fairness constraint, which requires that a real-valued functional parameter of ψ_0 is set to zero or is otherwise bounded. Let $\Theta_{P_0}(\psi)$ denote a user-selected constraint. The constrained functional parameter, denoted by ψ_0^*, is defined as $\psi_0^* = \text{argmin}_{\psi \in \Psi, \Theta_{P_0}(\psi)=0} R_{P_0}(\psi)$.

CLOSED-FORM SOLUTION

Nabi et al. [2024] proposed to construct a *constraint-specific path*, indexed by the Lagrangian multiplier $\lambda \in \mathbb{R}$, through the unconstrained parameter that would yield a solution to estimation of the constrained functional parameter. Any given point on this path, denoted by $\psi_{0,\lambda}$, is the minimizer to the Lagrangian problem: $\psi_{0,\lambda} = \text{argmin}_{\psi \in \Psi} R_{P_0}(\psi) + \lambda\Theta_{P_0}(\lambda)$. The authors proved that, for any given datum o and $\forall\lambda \in \mathbb{R}$, the constraint-specific path satisfies:

$$D_{R,P_0}(\psi_{0,\lambda})(o) + \lambda D_{\Theta,P_0}(\psi_{0,\lambda})(o) = 0 \, , \tag{C1}$$

where D_{R,P_0} and D_{Θ,P_0} denote the *canonical gradient*s of the risk function and the constraint functional.

We adopt the causal perspective on fairness described by Nabi and Shpitser [2018], Nabi et al. [2022] and provide closed-form solutions for fair optimal predictions. Due to page limits, we present results under MSE risk only.

Theorem 1 (Mean Squared Error risk) *Let* $\psi_0^*(z) = \text{argmin}_{\psi\in\Psi,\Theta_{\Delta,P_0}(\psi)=0} P_0 L(\psi)$*, with* $L(\psi)$ *representing the L2 loss and* $\Theta_{\Delta,P_0}(\psi)$ *denoting the identified functional for a pre-defined unfair path-specific effect. The conjunction of condition* (C1) *and* $\Theta_{\Delta,P_0}(\psi_0^*) = 0$ *necessitates*

$$\psi_0^*(z) = \psi_0(z) - \Theta_{\Delta,P_0}(\psi_0)\frac{D_{\Theta_\Delta,P_0}(z)}{\sigma^2(D_{\Theta_\Delta,P_0})} \, , \tag{1}$$

where $D_{\Theta_\Delta,P_0}(z)$ *is the constraint gradient, and* $\sigma^2(D_{\Theta_\Delta,P_0}) = \int D^2_{\Theta_\Delta,P_0}(z)dP_0(z)$*.*

DISCUSSION

Equation (1) implies that the fair risk minimizer ψ_0^* can be viewed as an adjustment to the unconstrained risk minimizer ψ_0. This adjustment is characterized by three components: $\Theta_{\Delta,P_0}(\psi_0)$, D_{Θ_Δ,P_0}, and $\sigma^2(D_{\Theta_\Delta,P_0})$. Each component is interpretable in its own right: (i) *Magnitude of systematic disparities*: The parameter $\Theta_{\Delta,P_0}(\psi_0)$ represents the magnitude of systematic disparities linked to S under sampling from P_0 – the larger the underlying disparities, the larger the adjustment that must be made to ψ_0; (ii) *Adjustment where it matters most*: The gradient of the constraint D_{Θ_Δ,P_0} can be viewed as the direction in the model space for ψ that leads to the largest change in the constraint. Thus, the constrained minimizer seeks to minimize changes made to ψ_0 by making the largest adjustments to ψ_0 in regions where these adjustments maximally impact the value of the constraint; (iii) *Ability to impact fairness through adjustment*: The variance of the constraint gradient, $\sigma^2(D_{\Theta_\Delta,P_0})$, indicates how adjustments in ψ_0 can impact the constraint. Large variance implies regions in the covariate space where the gradient is steeper, meaning minor changes to ψ_0 would result in comparably large changes in the value of the constraint. In this case, we only need make relatively minor adjustments ψ_0 to satisfy the fairness constraint. On the other hand, if the variability in the gradient of the constraint is small, then adjustments to ψ_0 must be made approximately uniformly across the covariate space, as all values of Z have approximately the same influence on the value of the constraint.

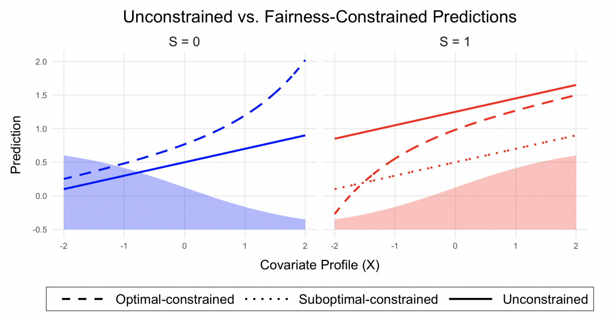

Figure 1: Predictions for the $S = 0$ group (left, blue) and $S = 1$ group (right, red). The optimal prediction function ψ_0 (solid line) disadvantages the $S = 0$ group with lower predictions. Using $\psi_0(0, X)$ for both groups (dash/dotted line) meets the fairness constraint but results in large errors for the $S = 1$ group. These errors are minimized by the optimal constrained prediction function ψ_0^* (dashed line).

As an example, consider a causal model where nullifying the average effect of S on Y is the constraint of interest. Suppose that X is a univariate standard normal random variable and $P_0(S = 1|X = x) = \text{expit}(x)$. Thus, higher values of x are associated with the $S = 1$ class. Outcomes are generated according to a linear mean function, $\psi_0(s, x) = 0.5 + 0.2x + 0.75s$ (Figure 1, solid lines), implying the ATE of $S = 1$ vs. $S = 0$ is $\Theta_{\Delta,P_0}(\psi_0) = 0.75$. If higher predicted outcomes confer an advantage, then the positive value of $\Theta_{\Delta,P_0}(\psi_0)$ implies that the $S = 0$ group would be on average disadvantaged by predicting from ψ_0. A simple solution to nullify the ATE in this example is to use $\psi_0(0, x)$ to predict for an individual with $X = x$, irrespective of their observed value of S (Figure 1, dotted line). However, this approach is suboptimal since it introduces a population-level bias, $\mathbb{E}\{\psi_0(0, X) - \psi_0(S, X)\} \neq 0$, and individual-level predictions suffer due to the relatively large differences between $\psi(1, X)$ and $\psi(0, X)$ for the relatively large number of observations with $S = 1$ and higher X values, evident in Figure 1. The optimal predictions, based on the gradient of the constraint, are shown as dashed lines. These predictions differ minimally from the unconstrained ones in the most supported regions of X, with larger differences occurring only in less common X values, yet these still effectively satisfy the constraint.

We extend our discussions to risk minimization under any identifiable path-specific effect within the NPSEM-IE framework, considering both MSE and cross-entropy risks. We also introduce a flexible semiparametric framework for estimating nuisance parameters essential for fairness adjustments. For details, see our submission on arXiv.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2022.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30. PMLR, 2017.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 2021.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. PMLR, 2019.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. In *Conference on Causal Learning and Reasoning*, pages 594–617. PMLR, 2022.

Razieh Nabi, Nima S Hejazi, Mark J. van der Laan, and David Benkeser. Statistical learning for constrained functional parameters in infinite-dimensional models with applications in fair machine learning. *arXiv preprint arXiv:2404.09847*, 2024.

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

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Association for the Advancement of Artificial Intelligence*, 2017.