Fair Participation via Sequential Policies

Reilly Raab[1], **Ross Boczar**[2], **Maryam Fazel**[2], **Yang Liu**[1]

[1]Department of Computer Science and Engineering, University of California, Santa Cruz
[2]Department of Electrical and Computer Engineering, University of Washington
reilly@ucsc.edu, rjboczar@uw.edu, mfazel@uw.edu, yangliu@ucsc.edu

Abstract

Leading approaches to algorithmic fairness and policy-induced distribution shift are often misaligned with long-term objectives in sequential settings. We aim to correct these shortcomings by ensuring that both the objective and fairness constraints account for policy-induced distribution shift. First, we motivate this problem using an example in which individuals subject to algorithmic predictions modulate their willingness to participate with the policy maker. Fairness in this example is measured by the variance of group participation rates. Next, we develop a method for solving the resulting constrained, non-linear optimization problem and prove that this method converges to a fair, locally optimal policy given first-order information. Finally, we experimentally validate our claims in a semi-synthetic setting.

Introduction

Organizations using historical data to optimize a policy which impacts human populations encounter two significant risks: bias and distribution shift. These hazards can interact to amplify social disparity over time as the policy-maker and population mutually adapt to each other. As organizations adopt data-driven methods for socially consequential tasks, such dynamics threaten equitable access to health care, justice, employment, housing, public services, education, credit, and privacy (Crawford and Calo 2016; Chaney, Stewart, and Engelhardt 2018; Ensign et al. 2018; Fuster et al. 2018; Hao 2020; Metz and Satariano 2020; Newton 2021; Schwartz et al. 2022).

Recent research on **long-term fairness** has highlighted the need to account for policy-induced distribution shift when seeking to mitigate disparity (Coate and Loury 1993; Heidari, Nanda, and Gummadi 2019; Wen, Bastani, and Topcu 2019; Mouzannar, Ohannessian, and Srebro 2019; D'Amour et al. 2020; Zhang et al. 2020; Liu et al. 2020; Morik et al. 2020; Raab and Liu 2021; Ge et al. 2021). In particular, it is now well-established that strategies that fail to account for policy-induced distribution shift can actively increase disparity and loss over time (Figure 1).

Despite this growing recognition, it is generally difficult to model or predict distribution shift caused by the reaction of human populations to novel policies. Without robust models or existing data, policies that elicit a population response must be deployed sequentially and learned **online**, i.e., based on information local to the currently deployed policy.

In this paper, we consider long-term fairness as a constrained, generally nonconvex optimization problem that must be solved by using **sequential policies**. In particular, we study the setting in which each policy induces the distribution of individuals it serves. That is, in terms of policy (which we denote by its parameters θ), an objective $\mathcal{L}$, and a disparity measure $\mathcal{H}$, we consider the problem

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(\theta, \mathcal{D}(\theta)) \\ \text{subject to} \quad & \mathcal{H}(\theta, \mathcal{D}(\theta)) \leq 0\,. \end{aligned} \tag{1}$$

Critically, we assume that the data distribution $\mathcal{D}$ describing the population is an (a priori *unknown*) function of the policy θ. Treating the distribution as a function of policy in this way is the key assumption of "performative prediction" as a setting (Perdomo et al. 2021) for a single learner (and extended to multiple learners in (Narang et al. 2022)), to which we add fairness constraints. We additionally add convex constraints to the decision variable θ in *Fair Participation*.

To solve Problem 1 using an iterative algorithm (that generates a sequence of policies), we adapt tools from nonlinear optimization that guarantee convergence to a fair, local optimum (Theorem 1). Our strategy is to rely on first-order information regarding currently deployed policy (Assumption 1) and to assume that $\mathcal{H}$ may be appropriately chosen without affecting the feasible set of policies (Assumption 2). We contrast our algorithm, both conceptually (See *Algorithm and Asymptotic Convergence*) and through experiments (See *Experiments*), to alternative sequential policies featured in prior literature.

Importantly, because $\mathcal{H}$ accounts for distribution shift via $\mathcal{D}(\theta)$, our formulation is inclusive of disparity measures $\mathcal{H}$ that are **intrinsic to the distribution** (i.e., defined independently of the policy). That is, we can choose forms of $\mathcal{H}$ for Problem 1 that depend only on the second argument. For example, we are free to choose $\mathcal{H}$ representing the fraction of individuals that choose to use an algorithmic service, which would be independent of policy in a myopic formulation, which assumes static distributions. We explore this difference more carefully in *A Failure Mode of Myopia*.

To be concrete, we now specialize our discussion to the context of participation dynamics: Imagine an online service that algorithmically recommends user-generated content (e.g., fixed-length videos) to other users based on inferred preferences. The firm seeks to maximize user engagement (e.g., the total daily number of videos watched on the platform) in order to drive advertising revenue, while users will only engage with the service if a sufficiently high percentage of recommended videos are personally interesting. Suppose that the firm optimizes for the predicted engagement of its current or induced user-base. In either case, the resulting recommendations and induced user-base may be biased towards large groups of users with similar interests or towards users that can be strongly engaged (e.g., unsupervised small children or those vulnerable to disinformation). For our hypothetical content-recommendation service, the resulting dynamics may encourage the proliferation of content targeting over-represented users, a decline in the diversity of advertisers targeting the platform, and user interactions that devolve into "echo chambers". To mitigate such risks, the firm may constrain its policies to maintain a diverse user-base, using a fairness constraint that is fundamentally distribution shift-aware.

Related Work

Prior work has highlighted importance of *dynamics* for algorithmic fairness by characterizing the outcomes induced by myopic definitions of fairness. For example, early work by Coate and Loury (1993) identified the potential for nominally-fair hiring practices to result in amplified underlying inequalities when populations adapt rationally. Closely related models explored by Mouzannar, Ohannessian, and Srebro (2019); Liu et al. (2020); Raab and Liu (2021) reveal similar failure modes, and the cited papers propose corrections based on the model of population response.

More recent work has explored a broad array of potential solutions to the misalignment of short-term and long-term fairness. For example, Morik et al. (2020) identify interventions for myopic optimization by adopting a feedback control mechanism, and Yin et al. (2023) adapt online learning methods to provide probabilistic bounds on cumulative regret and disparity. More generally, numerous authors have considered methods based on reinforcement learning (Wen, Bastani, and Topcu 2019; Liu et al. 2021; Ge et al. 2021), though theoretical guarantees are more difficult to establish with this approach.

Broadly, when the reaction of populations to policy are not known in closed form but decisions can have socially deleterious consequences, it remains an open problem to improve sample efficiency and strengthen the theoretical guarantees of online policies. To further progress towards this end, we claim that our formulation of Problem 1 is amenable to the reward-state-action setup of reinforcement learning, but with added structure (actions always induce the same state) that makes it approachable with more direct tools from nonlinear optimization, with more familiar performance guarantees.

In addition to general treatments of long-term fairness, we further highlight prior work that specifically addresses group participation dynamics: Hashimoto et al. (2018) study the potential amplification of participation disparity with a population model based on monotonic rates of retention and new arrivals in each group. Unlike our work, the intervention proposed is based on controlling the worst-case group-specific loss rather than more general and explict fairness definitions such as the equality of group participation rates. In realistic settings, such an intervention may be difficult to justify to stakeholders, since it involves modifying the optimization objective without making fairness targets explicit.

Zhang et al. (2019) also study fairness with participation dynamics based on retention rates that vary with model accuracy: The authors demonstrate that myopic interventions can amplify disparity and call for fairness criteria that account for policy-induced distribution shift, but the proposed method assumes full knowledge of dynamics to explicitly construct the set of feasible policies prior to optimization. Finally, Dean et al. (2022) consider a general family of risk-reducing dynamics governing user participation and policy-updates, but focus on the interaction of users with multiple firms rather than a single algorithmic policy, and characterize the equilibria and stability of the dynamics without fairness constraints.

Setting

Let us formalize the specific setting we consider in this paper. Throughout our discussion, we adopt discrete-time semantics, wherein the parameters θ^t and distribution $\mathcal{D}^t = \mathcal{D}(\theta^t)$ evolve in time $t \in \{1, 2, ...\}$ through repeated interactions between the firm and the population of (potential) users. Hereafter, let us write

$$\mathcal{L}(\theta) \coloneqq \mathcal{L}\big(\theta, \mathcal{D}(\theta)\big); \quad \mathcal{L}^t \coloneqq \mathcal{L}(\theta^t); \tag{2a}$$

$$\mathcal{H}(\theta) \coloneqq \mathcal{H}\big(\theta, \mathcal{D}(\theta)\big); \quad \mathcal{H}^t \coloneqq \mathcal{H}(\theta^t); \tag{2b}$$

$$\nabla\mathcal{L}^t \coloneqq \frac{\partial\mathcal{L}\big(\theta^t, \mathcal{D}^t\big)}{\partial\theta} + \sum_i \frac{\partial\mathcal{D}_i(\theta^t)}{\partial\theta}\frac{\partial\mathcal{L}\big(\theta^t, \mathcal{D}^t\big)}{\partial\mathcal{D}_i}, \tag{2c}$$

$$\nabla\mathcal{H}^t \coloneqq \frac{\partial\mathcal{H}\big(\theta^t, \mathcal{D}^t\big)}{\partial\theta} + \sum_i \frac{\partial\mathcal{D}_i(\theta^t)}{\partial\theta}\frac{\partial\mathcal{H}\big(\theta^t, \mathcal{D}^t\big)}{\partial\mathcal{D}_i}, \tag{2d}$$

where ∇ denotes the gradient with respect to θ accounting for all arguments, and each $\mathcal{D}_i$ is a vector component given an orthonormal basis in the Hilbert space[1] of distributions.

Assumption 1 (Gradients of Deployed Policy)**.** *At each time t, the firm is able to observe $\nabla\mathcal{L}^t$ and $\nabla\mathcal{H}^t$, i.e., the firm has knowledge of the first-order dependence of $\mathcal{L}$ and $\mathcal{H}$ on θ at the currently deployed policy θ^t.*

Assumption 1 is reasonable when $\mathcal{L}$ corresponds to empirical risk and $\mathcal{H}$ measures disparities between sets in the population, as in Problem 3: With small, random perturbations to policies over the set of individuals in the population, first-order statistics can provide an estimate of the local dependence of $\mathcal{L}$ and $\mathcal{H}$ on θ (i.e., via (conditional) correlations between policy perturbations and outcomes). Conceptually, the firm can estimate gradients from A/B testing.

[1]We notate this Hilbert space as finite-dimensional, for clarity, though we do not require this assumption elsewhere.

θ	parameter value.
$\mathcal{D}$	distribution of users.
$\mathcal{L}$	the objective function (total loss).
$\mathcal{H}$	disparity/fairness constraint function.
g	discrete group index.
ℓ_g	average loss for group g.
ρ_g	participation rate for group g.
$\mathcal{A}$	set of achievable losses.
f_g	map from ℓ_g to ρ_g.

Table 1: Choice of notation

Assumption 2 (Properties of Disparity)**.** *$\mathcal{H}$ is an invex function; that is, every critical point of $\mathcal{H}$ is a global minimum.*

Assumption 2 is easily satisfied by choosing a suitable $\mathcal{H}$ that maintains the required zero-level set (e.g., as in Equation (4)).

Fair Participation

For a setting with group-based participation dynamics, we will characterize policies by group-specific losses ℓ_g indexed by demographic group $g \in [k]$. We replace the decision variable θ with the decision variable $\boldsymbol{\ell} \in \mathcal{A}$, such that $\nabla\mathcal{L}$ and $\nabla\mathcal{H}$ denote gradients with respect to $\boldsymbol{\ell}$. For this setting, we choose to measure disparity by the variance of **group participation rates** ρ_g with g, where each group modulates its participation rate in response to the current value of ℓ_g. The problem we consider is to minimize the average loss, weighted by participation across the population, while limiting the divergence of participation rates across groups (Problem 3).

$$\begin{aligned} \underset{\boldsymbol{\ell}\in\mathcal{A}}{\text{minimize}} \quad & \mathcal{L}(\boldsymbol{\ell}) \coloneqq \left\langle \boldsymbol{\ell}, \boldsymbol{\rho} \right\rangle \\ \text{subject to} \quad & \operatorname*{Var}_g[\rho_g] \coloneqq \frac{1}{k}\sum_{g=1}^{k} \|\rho_g - \bar{\rho}\|_2^2 \leq \varepsilon, \\ & \boldsymbol{\rho} = f(\boldsymbol{\ell}), \qquad \bar{\rho} \coloneqq \frac{1}{k}\sum_{g=1}^{k} \rho_g. \end{aligned} \tag{3}$$

Problem 3 is an instance of Problem 1 that summarizes the distribution (in terms of $\boldsymbol{\rho}$), augmented by constraints to the decision variable, $\boldsymbol{\ell} \in \mathcal{A}$. In this instance, we have chosen to constrain a measure of disparity that is *intrinsic to the distribution* (Equation (4)): This fundamentally requires that the firm anticipates distribution shift in reaction to policy for any attempt to constrain it to be meaningful.

To clarify what we mean by *participation*, we consider a population of fixed size, wherein each user in the population voluntarily chooses to interact with the firm as a function of the expected loss for their group (Assumption 4). We define participation rate as the fraction of *prospective* users in each group that choose to interact with the firm. We additionally assume that the firm can always select from the same profile of average group-specific losses (Assumption 3).

Assumption 3 (Achievable Group Losses)**.** *Independent of the distribution of participating agents, at each time t, the firm is able to select from a set of group-specific losses $(\ell_1^t, \ell_2^t, ..., \ell_k^t) \in \mathcal{A}$, where we assume that $\mathcal{A}$ is convex. We omit the t index for clarity where convenient.*

The existence of a set of achievable losses in Assumption 3 may be guaranteed when the profile of active users in each group does not depend on the participation rate in the group, i.e., when the users of each group are treated as independently and identically distributed according to a static, group-specific distribution. The convexity of this set is well justified by the ability of the firm to adopt mixed policies; that is, for any two loss vectors $a, b \in \mathcal{A}$, we assume that the firm is free to deploy a stochastic mixture of the policies that resulted in a and b, implying that $\mathcal{A}$ is closed under convex combinations.

Assumption 4 (Participation rates)**.** *For each group g, the absolute participation rate, which we denote as ρ_g for $g \in [k]$, can be written as a strictly decreasing, differentiable function of ℓ_g. That is, for each g we have $\rho_g = f_g(\ell_g)$ such that $\frac{\mathrm{d}}{\mathrm{d}\ell_g} f_g < 0$.*

Intuitively, Assumption 4 states that the firm will lose users from group g if its loss on that group increases. Where appropriate, we write the (implicitly g-indexed) vector expressions $\boldsymbol{\ell}, \boldsymbol{\rho} = f(\boldsymbol{\ell})$, $\boldsymbol{\ell} = f^{-1}(\boldsymbol{\rho})$, and $f'(\boldsymbol{\ell}) = \mathrm{vec}(\{f_i'(\ell_i)\})$.

Without loss of generality, we fix zero loss for each group to correspond to zero participation with Assumption 5:

Assumption 5. *To restrict interpretations of Problem 3 to situations in which the firm has incentives to realize high participation rates (as opposed to eliminating users that are universally costly), we assume that*

$$\boldsymbol{\ell} \preceq 0 \;\; \textit{and} \;\; f(0) = 0.$$

Finally, we identify the measure of disparity $\mathcal{H}$ in Problem 3 as

$$\mathcal{H}(\boldsymbol{\ell}) = \operatorname*{Var}_g[f_g(\ell_g)] - \varepsilon \leq 0. \tag{4}$$

We confirm that this example satisfies Assumption 2 in the Technical Appendix[2].

Standard Approaches

We contrast the algorithm presented in *Algorithm and Asymptotic Convergence* for solving Problems 1 and 3 with the most commonly considered approach, which involves a sequence of instantaneously fair or instantaneously optimal policy deployments that ignore distribution shift. This approach is referred to as (Fairness-Constrained) "Repeated Risk Minimization" (RRM). In our setting, where $\mathcal{H}$ represents an inherent property of the distribution that is independent of policy without accounting for distribution shift, RRM remains ignorant of the fairness constraint. We therefore formalize RRM as a **sequential quadratic program** equivalent to projected gradient decent on average loss for step size $\eta > 0$:

$$\boldsymbol{\ell}^{t+1} = \operatorname*{arg\,min}_{\boldsymbol{\ell}\in\mathcal{A}} \quad \left\langle \boldsymbol{\ell}, \boldsymbol{\rho}^t \right\rangle + \frac{1}{2\eta}(\boldsymbol{\ell} - \boldsymbol{\ell}^t)^2. \tag{RRM}$$

[2]The Technical Appendix and a link to the code repository associated with this paper will be made available on arXiv.

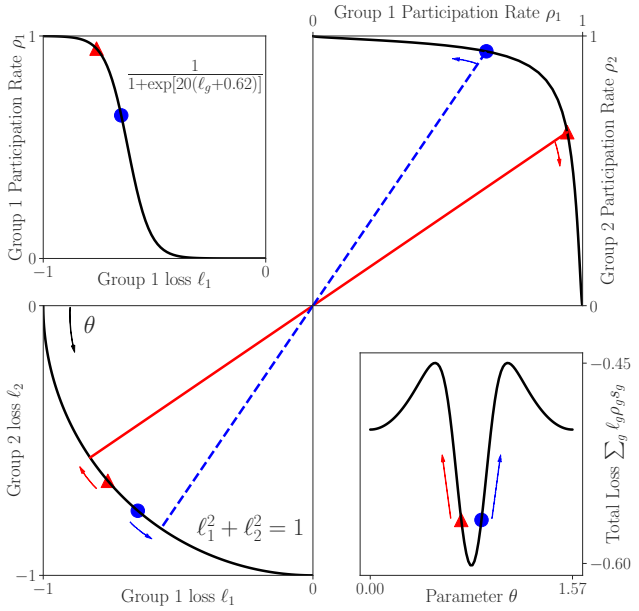

Figure 1: Failures of RRM. In this figure, we consider two policies, represented by the red triangle and the blue circle, which correspond to combinations of achievable group-specific losses. In this example, the set $\mathcal{A}$ is defined by a quadrant of the unit disk (lower left). These group-losses induce corresponding group participation rates (upper right) via $\rho_1 = f(\ell_1)$ (upper left). Because total loss has the form of an inner product, we may interpret $\boldsymbol{\rho}$ as vector in the dual space of $\boldsymbol{\ell}$, where we wish to increase the relative alignment of corresponding $\boldsymbol{\rho}$ and $\boldsymbol{\ell}$ vectors. For this figure, we see that alternative values of $\boldsymbol{\ell}$ would reduce loss *if* $\boldsymbol{\rho}$ *were fixed* (i.e., under the assumption made by RRM). Due to decision-induced distribution shift, however, the true form of the performative loss (lower right) indicates that the policies selected by RRM will actually increase loss (and disparity).

RRM is the typical benchmark policy for describing how myopic fairness interventions, which do not account for policy-induced distribution shift, can yield deleterious results. Nonetheless, RRM is also frequently considered for our primary task (Problem 1) (Perdomo et al. 2021).

Despite its popularity, a gap exists between RRM and a fair, locally optimal performative policy: RRM can actively increase both loss and disparity, even when deployed near fair, local minima of the objective function. We illustrate this fact with an intuitive explanation inspired by the geometric interpretation of our objective, in Figure 1. We emphasize that this example is realistic: In *Experiments*, we provide semi-synthetic examples with the same fundamental issues based on real-world prediction tasks using US Census data (Ding et al. 2021) data on movie preferences (Harper and Konstan 2015) (Figures 2 to 4).

As a comparison to RRM, we consider an algorithm that correctly account for policy-induced distribution shift in the objective, but fails to account for the same in disparity. We refer to this as "Myopic Projected Gradient" (MPG), which we also formalize as a sequential quadratic program:

$$\boldsymbol{\ell}^{t+1} = \arg\min_{\boldsymbol{\ell}\in\mathcal{A}} \quad \langle \boldsymbol{\ell}, \nabla\mathcal{L}^t \rangle + \frac{1}{2\eta}(\boldsymbol{\ell} - \boldsymbol{\ell}^t)^2, \quad \text{(MPG)}$$

where we recall that $\nabla\mathcal{L}^t$ denotes a gradient with respect to $\boldsymbol{\ell}$ in our target setting. While MPG is well-constructed to optimize the objective function, we show in *Experiments* that failure to account for the effect of policy-induced distribution shift on disparity measures that depend solely on the distribution often results in violated constraints in practice.

A Failure Mode of Myopia

Let us examine Equation (2d) more carefully and decompose $\nabla\mathcal{H}^t$ as follows:

$$\nabla\mathcal{H}^t := \underbrace{\frac{\partial\mathcal{H}(\theta^t, \mathcal{D}^t)}{\partial\theta}}_{u} + \underbrace{\sum_i \frac{\partial\mathcal{D}_i(\theta^t)}{\partial\theta}\frac{\partial\mathcal{H}(\theta^t, \mathcal{D}^t)}{\partial\mathcal{D}_i}}_{v},$$

The myopic assumption of Equation (RRM) asserts that $\mathcal{D}$ is constant, and so only considers the first term, u (i.e., assumes $v = 0$). When $\langle u, u + v\rangle < 0$ (i.e., when the myopic approximation of the gradient opposes the true gradient), a myopic first-order method will be *misaligned* with the true direction of decreasing disparity $u + v$. For this reason, explicitly accounting for the second term in Equation (2d) is needed to guarantee reductions of disparity with small updates. Moreover, when $\mathcal{H}$ depends only on $\mathcal{D}(\theta)$, u is identically zero, and myopic assumptions about (the absence of) policy-induced distribution shift render us blind to attempt to reduce disparity, just as RRM and MPG do when considering fair participation.

Algorithm and Asymptotic Convergence

To address Problems 1 and 3, we propose a method related to Fletcher's smooth exact penalty function (Fletcher 1973; Conn, Gould, and Toint 2000). For futher background, we also refer the reader to Nocedal and Wright (1999). Our method involves solving a sequential quadratic program parameterized by step size $\eta > 0$ and a scale factor $\alpha > 0$. We refer to this method as "Constrained Projected Gradient" (CPG), shown below, recalling that $\nabla\mathcal{L}^t = \nabla_{\boldsymbol{\ell}}\mathcal{L}(\boldsymbol{\ell})\big|_{\boldsymbol{\ell}=\boldsymbol{\ell}^t}$ and $\nabla\mathcal{H}^t = \nabla_{\boldsymbol{\ell}}\mathcal{L}(\boldsymbol{\ell})\big|_{\boldsymbol{\ell}=\boldsymbol{\ell}^t}$.

$$\begin{aligned} \boldsymbol{\ell}^{t+1} = \arg\min_{\boldsymbol{\ell}\in\mathcal{A}} \quad & \langle \boldsymbol{\ell}, \nabla\mathcal{L}^t \rangle + \frac{1}{2\eta}(\boldsymbol{\ell} - \boldsymbol{\ell}^t)^2. \\ \text{subject to} \quad & \langle \boldsymbol{\ell} - \boldsymbol{\ell}^t, -\nabla\mathcal{H}^t \rangle \geq \alpha\mathcal{H}^t. \end{aligned} \quad \text{(CPG)}$$

Assumption 6 (Feasibility)**.** *The fairness constraint is feasible. That is,* $\exists\boldsymbol{\ell}^\star \in \mathcal{A}$ *such that* $\mathcal{H}(\boldsymbol{\ell}^\star) \leq 0$*. Furthermore, the subproblem in Equation* (CPG) *is feasible at each* t*.*

The second stipulation of Assumption 6 eliminates the possibility that, for example, $\nabla\mathcal{H}^t = 0$ and $\mathcal{H}^t > 0$.

Theorem 1 (Asymptotic Convergence)**.** *Subject to Assumptions 1, 2 and 6, as* $(t \to \infty)$*, CPG (Equation* (CPG)*) converges to a feasible local optimum of the objective when the step size* η *is sufficiently small.*

Proof Sketch. Our proof relies on establishing that CPG first achieves fairness, then converges to a critical point of the objective function. First, note that $\mathcal{H}^t > 0 \implies \langle \boldsymbol{\ell}^{t+1} - \boldsymbol{\ell}^t, -\nabla\mathcal{H}^t \rangle > 0$, subject to the constraints imposed by $\mathcal{A}$, by the fairness constraint of Equation (CPG) and Assumption 6. That is, when the current policy is *unfair*, the algorithm makes progress towards fairness by decreasing disparity. Second, we show that $\mathcal{H}^t \leq 0 \implies \langle \boldsymbol{\ell}^{t+1} - \boldsymbol{\ell}^t, -\nabla\mathcal{L} \rangle > 0$. That is, once the current policy is *fair*, the algorithm decreases loss. This second fact follows from the fact that $\mathcal{H}^t \leq 0$ implies that the sign of $\langle \boldsymbol{\ell}^{t+1} - \boldsymbol{\ell}^t, -\nabla\mathcal{H} \rangle$ is unconstrained, and minimization of the objective will naturally ensure $\langle \boldsymbol{\ell}^{t+1} - \boldsymbol{\ell}^t, \nabla\mathcal{L} \rangle < 0$ subject to the constraints of $\mathcal{A}$. We provide a rigorous proof in the Technical Appendix.

We briefly outline how CPG relates to Fletcher's smooth exact penalty function subject to convex constraints. Fletcher's penalty method (Fletcher 1973) is outlined by Conn, Gould, and Toint (2000, Sec. 14.6) as a surrogate objective or "merit" function that we seek to minimize in order to solve the constrained optimization problem (Problem 1), which we notate in terms of general, unconstrained decision variable θ:

$$\Phi(\theta) = \mathcal{L}(\theta) + \lambda(\theta)\mathcal{H}(\theta), \tag{5}$$

where, omitting explicit dependence on θ,

$$\lambda(\theta) = \operatorname*{arg\,min}_{z \geq 0} \quad z\mathcal{H} - \frac{1}{2\sigma}\big(\nabla\mathcal{L} + z\nabla\mathcal{H}\big)^2 \tag{6a}$$

$$= \max\left(0, \frac{\sigma\mathcal{H} - \langle \nabla\mathcal{L}, \nabla\mathcal{H} \rangle}{\langle \nabla\mathcal{H}, \nabla\mathcal{H} \rangle}\right). \tag{6b}$$

The standard approach is to minimize $\Phi(\theta)$; however, the potential non-differentiability of $\lambda(\theta)$ can be problematic (Conn, Gould, and Toint 2000). To avoid this potential issue, we interpret $\lambda(\theta)$ as a local estimate for the optimal dual variable $\nu^\star$ in the Lagrangian Λ associated with Problem 1:

$$\Lambda(\theta, \nu) \coloneqq \mathcal{L}(\theta) + \nu\mathcal{H}(\theta)\,. \tag{7}$$

To see this, observe that the Karush–Kuhn–Tucker conditions guarantee that, at an optimal solution (denoted by $\star$),

$$\nabla(\mathcal{L}^\star + \nu^\star\mathcal{H}^\star) = 0 \implies \nu^\star = \Psi(\theta^\star), \text{ where}$$
$$\Psi(\theta) \coloneqq -\frac{\langle \nabla\mathcal{L}(\theta), \nabla\mathcal{H}(\theta) \rangle}{\langle \nabla\mathcal{H}(\theta), \nabla\mathcal{H}(\theta) \rangle}\,. \tag{8}$$

In particular, the estimate $\lambda(\theta)$ given by Equation (6b) corresponds to a gradient step taken with respect to ν in the *maximization* of the Lagrangian Λ away from an initial guess $\Psi(\theta)$.

Thus, interpreting Equation (6b) as an estimate of the dual variable in the Lagrangian, we propose to sequentially update θ with a gradient update in the *minimization* of the Lagrangian Λ using the fixed estimate $\nu = \lambda(\theta^t)$:

$$\theta^{t+1} - \theta^t = -\eta\nabla\Lambda^t \tag{9}$$

$$\nabla\Lambda^t = \nabla\mathcal{L}^t + \lambda(\theta^t)\nabla\mathcal{H}^t. \tag{10}$$

Substituting Equation (6b) into Equation (10) and taking an inner product with $\nabla\mathcal{H}$ results in the relation

$$\langle \nabla\Lambda^t, \nabla\mathcal{H}^t \rangle = \max\Big(\langle \nabla\mathcal{L}^t, \nabla\mathcal{H}^t \rangle, \sigma\mathcal{H}^t\Big).$$

Substituting $\nabla\Lambda^t$ as in Equation (9) then implies the constraint

$$\langle \nabla\Lambda, \nabla\mathcal{H} \rangle = \frac{1}{\eta}\langle \theta^{t+1} - \theta^t, -\nabla\mathcal{H} \rangle \geq \sigma\mathcal{H}. \tag{11}$$

Finally, let $\alpha = \sigma\eta$. Without additional constraints, CPG updates θ by performing gradient descent on $\mathcal{L}$ subject to the constraint of Equation (11):

$$\begin{aligned} \theta^{t+1} = \operatorname*{arg\,min}_{\theta} \quad & \langle \theta, \nabla\mathcal{L}^t \rangle + \frac{1}{2\eta}(\theta - \theta^t)^2. \\ \text{subject to} \quad & \langle \theta - \theta^t, -\nabla\mathcal{H}^t \rangle \geq \alpha\mathcal{H}^t. \end{aligned} \tag{12}$$

Furthermore, if we can optimize over $\boldsymbol{\ell}$ directly, we may add constraints (e.g., $\boldsymbol{\ell} \in \mathcal{A}$) to Equation (CPG) while retaining convexity of the resulting subproblem.

Special Case: Concave Participation Rates

We consider the special case in which $f(\boldsymbol{\ell})$ is concave, in addition to monotonically decreasing. In this case, we can rewrite the objective in Problem 3 in terms of f^{-1} and treat $\boldsymbol{\rho}$ as the decision variable for a concave minimization problem.

Assumption 7. *$f_g(\ell_g)$ is a concave function of ℓ_g for all g.*

Additionally, we assume that we can arbitrarily degrade the loss vector:

Assumption 8. *The set $\mathcal{A}$ is "Pareto-closed", i.e.*

$$\boldsymbol{\ell} \in \mathcal{A},\ \boldsymbol{\ell} \leq \boldsymbol{\ell}' \implies \boldsymbol{\ell}' \in \mathcal{A},$$

where the inequality is taken componentwise.

Assumption 4 implies that f_g is invertible. With Assumption 7, f_g^{-1} is concave.

Proposition 1. *Under Assumptions 3, 4, 7 and 8, the set $f(\mathcal{A})$ is convex.*

Proof. Recall that $\mathcal{A}$ is a convex set of achievable losses $\boldsymbol{\ell}$ (Assumption 3). Given any two points $\boldsymbol{\rho}^0, \boldsymbol{\rho}^1 \in f(\mathcal{A})$, we wish to show that the point $\boldsymbol{\rho}^\alpha \coloneqq \alpha\boldsymbol{\rho}^1 + (1-\alpha)\boldsymbol{\rho}^0 \in f(\mathcal{A})$, for any $\alpha \in [0, 1]$.

We see that the corresponding losses $\boldsymbol{\ell}^0, \boldsymbol{\ell}^1$ lie in $\mathcal{A}$ by construction. Thus, by the above assumptions,

$$f^{-1}(\boldsymbol{\rho}^\alpha) \geq \alpha\boldsymbol{\ell}^0 + (1-\alpha)\boldsymbol{\ell}^1 \in \mathcal{A},$$

so $\boldsymbol{\rho}^\alpha \in f(\mathcal{A})$, and $f(\mathcal{A})$ is convex. □

We can express the "inverse" optimization problem to Problem 3 as

$$\begin{aligned} \operatorname*{minimize}_{\boldsymbol{\rho} \in f(\mathcal{A})} \quad & \langle \boldsymbol{\rho}, f^{-1}(\boldsymbol{\rho}) \rangle \\ \text{subject to} \quad & \mathcal{H}(\boldsymbol{\rho}) \leq 0\,. \end{aligned} \tag{13}$$

Problem 13 has a convex constraint and a concave objective, since

$$\frac{\partial^2}{\partial\rho_g^2}\rho_g f_g^{-1}(\rho_g) = \left(f_g^{-1}(\rho_g) + \rho_g\frac{\partial^2}{\partial\rho_g^2}f_g^{-1}(\rho_g)\right) \leq 0.$$

This problem is still nonconvex, but minimizing a concave function over a convex set allows for a simple algorithm with a very simple convergence proof: We can show that a simple iterative linearization and minimization procedure over our constraint set converges to a local minimum of the loss. The details, convergence proof and an explicit example for our choice of $\mathcal{H}$ is described in the Technical Appendix.

Experiments

We evaluate CPG against RRM and MPG in multiple semi-synthetic settings.

Datasets

Our settings derive from binary classification tasks on the American Community Survey Public Use Microdata Sample (ACS PUMS) dataset[3], as introduced by Ding et al. (2021), for specific US states in 2018, or a recommendation task on movie preferences using data (MovieLens) collected by Harper and Konstan (2015). Each task gives samples of joint feature (X), label ($Y \in \{0, 1\}$), and group ($G \in [k]$) variables, the joint distribution of which we summarize by writing $\mathcal{S}$.

Model Class and Achievable Losses

For each task, we first define a set of achievable losses $\mathcal{A}$. We generate $n = 100$ different binary classifiers and record the vector of group-specific losses $\boldsymbol{\ell}$ achieved by the predicted labels $\hat{Y}$ for each classifier, where we define ℓ_g as the negative binary prediction accuracy conditioned on group g:

$$\ell_g = -\mathop{\mathrm{E}}_{X,Y,G\sim\mathcal{S}}\left[\hat{Y} = Y \mid G = g\right]. \tag{14}$$

The set of achievable loss vectors $\mathcal{A}$ for the task is defined by the convex hull of these samples $\mathcal{A} = \mathrm{Hull}(\{\boldsymbol{\ell}_i : i \in [n]\})$.

In our experiments, we consider logistic classifiers trained on different weighted logistic loss functions for the binary classification task. That, for each classifier $i \in [n]$, we sample a vector of objective function term weights $\boldsymbol{\beta}$ uniformly at random from the $(k-1)$-simplex ($\sum_{g=1}^{k} \beta_g = 1$) and solve the regularized logistic classification task

$$\begin{aligned}
&\min_{\boldsymbol{w}} \quad \sum_{g=1}^{k} l_g(\boldsymbol{w})\beta_g + \frac{1}{2}|\boldsymbol{w}|^2; \\
&l_g(\boldsymbol{w}) = \mathop{\mathrm{E}}_{X,Y,G\sim\mathcal{S}}\left[H_Y(h_{\boldsymbol{w}}(X)) \mid G = g\right]; \\
&H_p(q) = -p\log q - (1-p)\log(1-q); \\
&h_{\boldsymbol{w}}(X) = \frac{1}{1 + e^{-\langle X, \boldsymbol{w}\rangle}},
\end{aligned} \tag{15}$$

using the limited-memory method of Broyden, Fletcher, Goldfarb, and Shanno (LBFGS) (Liu and Nocedal 1989), as implemented by scikit-learn (Pedregosa et al. 2011).

Synthetic Distribution Shift

We model the function f that maps group loss to group participation rate as a reversed logistic function parameterized by bias and sensitivity parameters $b \in (-1, 0)$ and $s > 0$, respectively, and clipped to the interval $[0, 1]$. That is, we model f as

$$\begin{aligned}
f(x) &= \max\left[0, \min[1, g(x)]\right]; \\
g(x) &= \frac{1}{1 + e^{s(x-b)}}; \quad x \in [-1, 0].
\end{aligned} \tag{16}$$

This same function is used in the upper-left panel of Figure 1 with parameters ($s = 20, b = -0.62$).

[3]https://github.com/socialfoundations/folktables

Hyperparameters

We use a learning rate that decays as a harmonic series:

$$\eta^t = \eta^1/t; \quad t \in \{1, 2, ...\}. \tag{17}$$

All experiments follow the same decay schedule and run for the same number of steps (i.e., 30), but the initial learning rate η^1 is equal to half of the diameter of $\mathcal{A}$. Each experiments run in less than 60 seconds on a typical laptop CPU.

We set initial conditions $\boldsymbol{\ell}^0$, the participation function parameters (b, s) (See *Synthetic Distribution Shift*), and the fairness slack parameter ε (refer to Problem 3) to demonstrate qualitatively diverse simulation outcomes among our included results.

Results

In all experiments, CPG achieved a feasible local optimum of the objective, while RRM and MPG did so only rarely. In (Figures 2 to 4), we highlight a few examples of our experimental results on the following tasks:

- **Income**: "Income" task of Ding et al. (2021) with groups redefined to coincide with the binary classification label and restricted to data from Alabama.
- **MovieLens**: From user age, occupation, and gender, predict whether this user exhibits a stronger-than-median preference for "mystery" rather than "adventure" films, using zero-one loss and targeting equal user rates across gender. The data for this task comes from Harper and Konstan (2015).
- **IncomeThree**: "Income" task of Ding et al. (2021) with groups expanded to three divisions of income (below \$60K, between \$60K and \$120K, above \$120K) and restricted to data from Alabama.

As our algorithms are deterministic, we do not consider multiple runs for the same setting and leave characterizations of the robustness of these algorithms in terms of different hyperparameters to future work (see Ethical Statement).

In Figures 2 to 4, the first pane visualizes the set of achievable losses, $\mathcal{A}$, and the samples used to generate it, where the axes correspond to each group's loss. The second pane visualizes the corresponding set of achievable participation rates $\boldsymbol{\rho}$ with axes corresponding to each group's participation rate. The last pane plots total loss and disparity vs. timestep for all three methods in the given setting. In Figures 2 and 3, an additional pane demonstrates the non-convexity of the total loss and disparity surfaces along a curve corresponding to all $\boldsymbol{\ell} \in \mathcal{A}$ which maximize distance from the origin, with angle from the x-axis parameterized by ϕ. In all figures, a distinct marker represents each method (i.e., RRM, MPG, CPG) and their shared initialization across all panes.

Conclusion

We intend our work to push beyond the chorus of literature that highlights the failure of myopic fairness interventions: We tractably address the problem of long-term fairness by incorporating fairness constraints into the "performative" setting (Perdomo et al. 2021), which explicitly accounts for policy-induced distribution shift (Problem 1). We

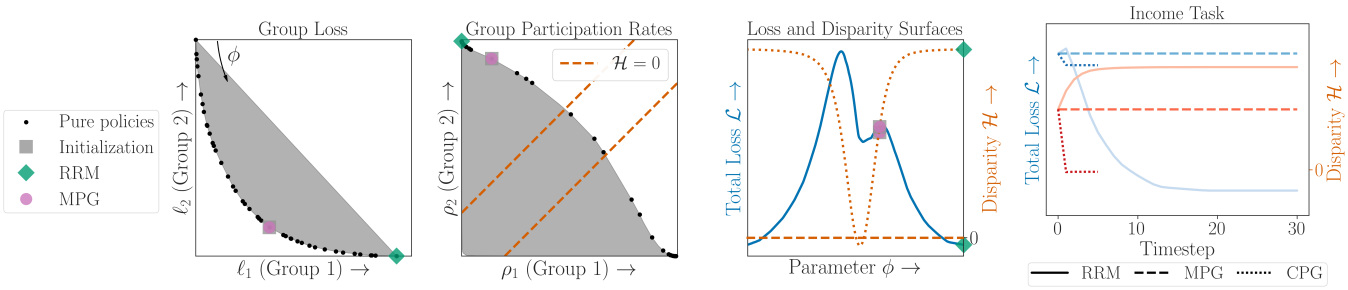

Figure 2: Income task. This setting has a highly non-convex loss surface, shown in the third pane, and demonstrates a situation in which CPG converges to the unique solution, MPG gets stuck in an unfair local minimum of the utility function, and RRM diverges to the highest disparity.

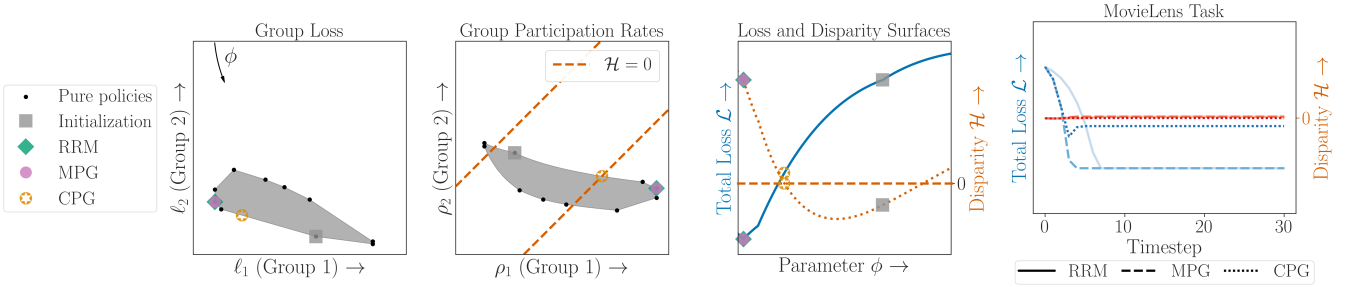

Figure 3: MovieLens Task. This particular setting demonstrates application of our method to a recommendation task. Only our proposed method, CPG, satisfies the fairness constraint, finding a solution to Problem 1 at the boundary of the feasible set, where $\langle \mathcal{L}, \mathcal{H} \rangle < 0$. Both RRM and MPG locally optimize local utility at the expense of fairness.

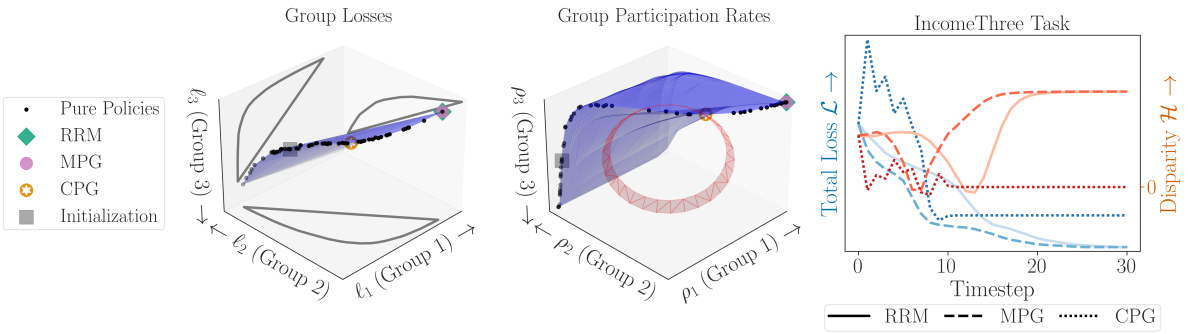

Figure 4: IncomeThree Task. This graphic is intended to showcase that Problem 3 and CPG are not restricted to only two groups, and in general allow very large numbers of groups, though the resulting sets of achievable group losses and participation rates are difficult to visualize. Inspecting the time-series, we see that CPG increases disparity at intermediate timesteps, despite starting outside of the feasible (fair) set; we attribute this to the large initial step size, which is not guaranteed to eliminate non-linear behaviors of $\mathcal{L}$ and $\mathcal{H}$ within the linear approximation trust-region.

show that this problem is amenable to tools derived from non-linear optimization by proposing an algorithm (CPG) which asymptotically solves the problem when restricted to local, first-order information by generating a sequence of policies (Theorem 1).

To provide a concrete setting for Problem 1 and CPG, we have specialized our discussion for when group-specific losses achieved by policy induce group-specific participation rates, thus extending the literature on fair participation dynamics. We compare CPG to standard, myopic approaches (RRM, MPG) in this setting with semi-synthetic experiments informed by real-world classification tasks and show that CPG consistently solves the problem where the myopic baselines fail (See *Experiments*).

Ethical Statement

Because our work focuses on socially consequential applications of machine learning and optimization, it is important to highlight its limitations.

First, it is important to note that asymptotic convergence to fairness (Theorem 1) does not imply fairness at every time step: In practice, the violations of fairness prior to convergence may remain large over many time steps. For real-world settings, this may be unacceptable, and alternative formulations of the problem based on *cumulative* fairness violations or reinforcement learning (See *Related Work*) may be more appropriate.

Second, Assumption 1 also introduces caveats: In realistic settings, the gradients $\nabla\mathcal{L}$ and $\nabla\mathcal{H}$ cannot be observed directly, but must be estimated from samples that introduce additional noise. While noisy gradient estimates, when unbiased, are still practically useful for gradient-based algorithms such as stochastic gradient descent, the additional noise may cause local increases in disparity and loss which may translate to socially deleterious outcomes.

Acknowledgments

This work is partially supported by the National Science Foundation (NSF) under grants IIS-2143895, IIS-2040800, CCF-2023495, CCF-231277, TRIPODS II DMS-2023166, CCF-2007036, and CCF-2212261.

References

Chaney, A. J.; Stewart, B. M.; and Engelhardt, B. E. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 224–232. ACM.

Coate, S.; and Loury, G. C. 1993. Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, 1220–1240.

Conn, A. R.; Gould, N. I.; and Toint, P. L. 2000. *Trust region methods*. SIAM.

Crawford, K.; and Calo, R. 2016. There is a blind spot in AI Research. *Nature News*, 538(7625): 311.

D'Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

Dean, S.; Curmei, M.; Ratliff, L. J.; Morgenstern, J.; and Fazel, M. 2022. Multi-learner risk reduction under endogenous participation dynamics. *arXiv preprint arXiv:2206.02667*.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, 6478–6490. Curran Associates, Inc.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway Feedback Loops in Predictive Policing. In *Conference of Fairness, Accountability, and Transparency*.

Fletcher, R. 1973. An exact penalty function for nonlinear programming with inequalities. *Mathematical Programming*, 5: 129–150.

Fuster, A.; Goldsmith-Pinkham, P.; Ramadorai, T.; and Walther, A. 2018. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Effects of Machine Learning on Credit Markets*.

Ge, Y.; Liu, S.; Gao, R.; Xian, Y.; Li, Y.; Zhao, X.; Pei, C.; Sun, F.; Ge, J.; Ou, W.; et al. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, 445–453.

Hao, K. 2020. The Coming War on the Hidden Algorithms that Trap People in Poverty. *MIT Technology Review*.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938. PMLR.

Heidari, H.; Nanda, V.; and Gummadi, K. P. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. *the International Conference on Machine Learning (ICML)*.

Liu, D. C.; and Nocedal, J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3): 503–528.

Liu, L. T.; Wilson, A.; Haghtalab, N.; Kalai, A. T.; Borgs, C.; and Chayes, J. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 381–391.

Liu, W.; Liu, F.; Tang, R.; Liao, B.; Chen, G.; and Heng, P. A. 2021. Balancing accuracy and fairness for interactive recommendation with reinforcement learning. *arXiv preprint arXiv:2106.13386*.

Metz, C.; and Satariano, A. 2020. An Algorithm That Grants Freedom, or Takes It Away. *The New York Times*.

Morik, M.; Singh, A.; Hong, J.; and Joachims, T. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 429–438.

Mouzannar, H.; Ohannessian, M. I.; and Srebro, N. 2019. From Fair Decision Making to Social Equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 359–368. ACM.

Narang, A.; Faulkner, E.; Drusvyatskiy, D.; Fazel, M.; and Ratliff, L. J. 2022. Multiplayer Performative Prediction: Learning in Decision-Dependent Games. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (arXiv:2201.03398)*.

Newton, D. 2021. From Admissions to Teaching to Grading, AI is Infiltrating Higher Education. *The Hechinger Report*.

Nocedal, J.; and Wright, S. J. 1999. *Numerical optimization*. Springer.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Perdomo, J. C.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2021. Performative Prediction. *arXiv:2002.06673 [cs, stat]*. ArXiv: 2002.06673.

Raab, R.; and Liu, Y. 2021. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34.

Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P.; et al. 2022. Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, 1270(10.6028).

Wen, M.; Bastani, O.; and Topcu, U. 2019. Fairness with Dynamics. *arXiv preprint arXiv:1901.08568*.

Yin, T.; Raab, R.; Liu, M.; and Liu, Y. 2023. Long-Term Fairness with Unknown Dynamics. *arXiv preprint arXiv:2304.09362*.

Zhang, X.; Khalili, M. M.; Tekin, C.; and Liu, M. 2019. Group Retention when Using Machine Learning in Sequential Decision Making: The Interplay between User Dynamics and Fairness. In *Advances in Neural Information Processing Systems*, 15243–15252.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469.