Algorithmic Fairness in Performative Policy Learning: Escaping the Impossibility of Group Fairness

Seamus Somerstep
smrstep@umich.edu
University of Michigan
Ann Arbor, MI, USA

Ya'acov Ritov
yritov@umich.edu
University of Michigan
Ann Arbor, MI, USA

Yuekai Sun
yuekai@umich.edu
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

In many prediction problems, the predictive model affects the distribution of the prediction target. This phenomenon is known as performativity and is often caused by the behavior of individuals with vested interests in the outcome of the predictive model. Although performativity is generally problematic because it manifests as distribution shifts, we develop algorithmic fairness practices that leverage performativity to achieve stronger group fairness guarantees in social classification problems (compared to what is achievable in non-performative settings). In particular, we leverage the policymaker's ability to steer the population to remedy inequities in the long term. A crucial benefit of this approach is that it is possible to resolve the incompatibilities between conflicting group fairness definitions.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Group Fairness; Impossibility Theorems; Performative Prediction; Long Term Fairness

ACM Reference Format:
Seamus Somerstep, Ya'acov Ritov, and Yuekai Sun. 2024. Algorithmic Fairness in Performative Policy Learning: Escaping the Impossibility of Group Fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3630106.3658929

1 INTRODUCTION

Automated decision-making and support systems that rely on predictive models are often used to make consequential decisions in criminal justice [7, 59], lending [54], and healthcare [16], but their long-term impacts on the population are poorly understood. Most prior work on algorithmic fairness assumes a *static* population and focuses on *allocative equality* [12]. For example, consider the common fairness definition equalized odds [28]. It requires a predictive model to incur false positives and false negatives at equal rates across demographic groups; *i.e.* it requires the model to allocate errors equally between groups. It does not consider long-term impacts of errors across demographic groups: for example, it may be easier for members of an advantaged group to overturn an error (as compared to members of a disadvantaged group), so errors are more consequential for the disadvantaged group. In the long term, this can exacerbate inequalities in the population in ways that simply enforcing equalized odds cannot address.

Motivated by concerns about the long-term impacts of predictive models, there is a line of work that embeds predictive models as *policies* in dynamic models of populations and studies how predictive models *steer* the population [13, 29, 34, 63]. Following this line of work, we study the enforcement of group fairness in the long term. Our contributions are as follows.

(1) We formulate a new long-term group fairness constraint inspired by algorithmic reform/reparation [15, 24], as well as long-term versions of the three traditional group fairness constraints.
(2) We show that as long as the policymaker has enough flexibility in the way they remedy historical inequities, it is possible for them to steer populations towards a reformed state while maintaining group fairness. As a consequence, this shows that it is possible, in the long term, to simultaneously satisfy traditionally conflicting group fairness constraints.
(3) We provide a reduction framework for computationally enforcing long-term group fairness. The convergence rates and generalization guarantees of this framework are provided.

As mentioned, a key consequence of our results is that in the long term, it is possible to *simultaneously* remedy historical disparities and satisfy multiple group fairness definitions that are traditionally incompatible. This is in contrast to previous research on compatible group fairness. The common theme of previous work is that of *avoidance*, *i.e.* practitioners should alter policies so that the different notions of fairness are fully satisfied at separate times or relaxed versions of group fairness are satisfied simultaneously. Instead, our focus is on *resolution* of incompatibility of group fairness, *i.e.* practitioners should implement policies that eliminate the inequity that leads to the impossibility of enforcing multiple forms of group fairness.

1.1 Related Work

We first cover the prior work on resolving group fairness incompatibilities, this is a non-exhaustive list and a thorough survey is given in [57]. The study of incompatibilities in group fairness was initiated simultaneously by the impossibility results in the works [9, 36]. In the years since then, researchers have sought to extend these results and modify fairness practices to partially resolve them. The

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658929

work [3] covers impossibility results for several group fairness metrics derived from confusion matrices in the context of recidivism. Beyond just group fairness, trade-offs between optimal accuracy, fairness, and resource allocation for different groups have also been explored [17, 56]. The line of work [6, 37, 43] studies algorithmic fairness in ML systems with human elements, although ultimately the aggregate decision making process (human plus machine) is still burdened by impossibility theorems. Another avenue of approach is to relax the sufficiency and separation requirements to compatibility [8, 25, 41, 51, 56], although direct trade-offs between fairness notions remain, and often such relaxations do not guarantee a desired level of fairness. Philosophically, our work aligns with the recent idea of *substantive fairness*, which argues that policies should aim to eliminate historical inequities, treating causes of disparity rather than symptoms of disparity [15, 24]. One of the main contributions of this work is to formalize *substantive fairness* into an algorithmic framework and study the feasibility of erasing historical disparities with ML systems.

Concern amongst the fairness community regarding the long-term impacts of predictive models has grown, originating with a line of work that models fair policies and populations as a dynamical system[13, 29, 34]. The long-term fairness framework that we study is based on the recently developed idea of performative prediction, a line of work that studies model-induced distribution shift [5, 32, 45, 46, 53]. Closely related lines of work on the enforcement of fairness in performative settings include [18, 58, 63, 64]. The first designs a Markov chain oriented framework of performative fairness, the second solves the problem of estimating down stream fairness impacts of policies, the third studies fairness in strategic environments though not with a goal of compatible group fairness. In general, these generally consider a *stateful* performative prediction setting and cast the task of steering the underlying dynamical system as an optimal control or reinforcement learning problem. In contrast, our problem setting prioritizes the discrimination at the steady state and is not concerned with intermediate time steps.

Although it predates performative prediction, strategic classification [27] is a common example of performative prediction. As such, attention has been paid to the study of fairness in strategic settings. The authors of [19] show that traditional fairness interventions in strategic settings can actually exacerbate discrimination. A similar perspective on traditional fairness constraints in strategic settings is given in [39]. In [65] the effect of fairness interventions on the incentive for strategic manipulation is studied. The work [31] shows that *cost discrimination* in strategic classification will cause violations of group fairness constraints in the long term. The authors of [40] also study cost discrimination in strategic classification, showing that subsidies for the disadvantaged group can alleviate discrimination concerns. The common goal for each of these works (and others in this area) is to study how strategic interventions make discrimination worse if either no fairness intervention is used or if a *traditional* fairness intervention (that does not account for strategic agents) is used. The key difference between this line of work and our work is that this line of work considers strategic behavior as a problem to be overcome, while we leverage strategic behavior for greater fairness achievement than is possible in non-strategic environments.

The running example of performative prediction/ a strategic setting in this paper is labor market models; the study of such models has a rich history in economics and long precedes the idea of performative prediction. A comprehensive survey of this field is given in [20]. The works [2, 55] presented initial labor market models and analyzed the equilibrium of these models at the worker level. In [55], discrimination in labor markets due to exogenous groups is studied, while [2] surprisingly shows that even markets with endogenous groups will have discriminatory equilibrium. Later, the authors [10] formulated these ideas in the celebrated *Coate and Loury* model of labor markets. Extensions on this model are numerous; the authors of [50] and [49] develop a model where wages are set by inter-employer competition and groups of workers actually benefit from discrimination of others. The line of work [11, 21–23] studies the efficacy of color-blind policies in preventing discrimination. We will primarily work with our own models of labor markets, inspired by [60, 62].

Our primary goal is to study the feasibility of (fairly) equating response distributions between two groups in the long term. This is closely related to the goal of incentivizing agents to improve some desirable quality. The authors of [47] show that this involves causal modeling, and our work is no exception: various labor market models will play the role of the causal model in our study. The works [48, 62] focus on improving the agents' overall welfare; they show that there is often a trade-off between the learner and the agents' utilities. *Performative power* [26] measures the ability of the learner to steer the agents in performative prediction. As we shall see, the firm in the aforementioned labor market models possess enough performative power to equate ex post response distributions despite ex ante disparities between the two groups. Finally, the causal strategic labor market model that we develop is an example of an *outcome performativity* problem. *outcome performativity* is introduced in [35] along with efficient omniprediction algorithms for outcome performativity problems. In general, the novelty of our work is our focus on a) driving improvement with the goal of equating disparate groups and b) doing so without discriminating against the advantaged group.

2 GROUP FAIRNESS IN PERFORMATIVE POLICY LEARNING

Consider a binary classification problem in which samples correspond to a population of individuals invested in learned policies (often referred to as strategic agents) characterized by $Z = (X, Y) \in \mathcal{X} \times \{0, 1\}$ and a protected attribute denoted G. This setting is characterized as a problem in performative prediction. Performative prediction is a distribution shift setting, where the implementation of a policy $f : \mathcal{X} \times G \to \{0, 1\}$ triggers responses from the individuals invested in the policy, leading to a new distribution of data $\mathcal{D}(f, G) \in \Delta(Z \times G)$. Throughout, we will assume that group proportions remain constant (say $\mathbb{P}(G = g) = \lambda_g$) for any policy f. Under this assumption, we can write

$$\mathcal{D}(f, G) = \sum \lambda_g \mathcal{D}(f, G)|G = g \triangleq \sum \lambda_g \mathcal{D}(f, g) \qquad (2.1)$$

In performative prediction, The policy maker's goal is to make a policy that maximizes their expected reward, *taking into account*

the strategic response of the individuals to their decisions:

$$\max_{f\in\mathcal{F}} \mathrm{EPR}(f) \triangleq \sum \lambda_g \mathbb{E}_{Z'\sim\mathcal{D}(f,g)}\big[r(f(\cdot,g);Z')\big], \qquad (2.2)$$

where $\mathcal{F}$ is a policy class, $\mathcal{D}(f,g)$ is the distribution map that encodes the long-term impacts of the policy on the subset of the population with sensitive trait value $G=g$, $r(f(\cdot,g);z)$ is the policy maker's reward function that measures their reward from applying a policy f to an agent z (this agent also has group membership g), and λ_g is the proportion of the population with sensitive trait value $G=g$. The objective function in (2.2) is often called the performative (expected) utility and measures the *ex post* reward of policies. Throughout this paper, we will mark random variables drawn from an *ex post* distribution as $Z'=(X',Y')$.

A recurring instance of performative policy learning in this work is the hiring firm's problem in Coate-Loury-type models of labor markets [20].

Example 2.1 (Continuous labor market example [61]). *Consider an employer that wants to hire skilled workers who reside in one of two identifiable groups $G\in\{A,D\}; G\sim Ber[\lambda]$. The workers are represented as (S,X,Y,G) quintuples. $S\in\mathbb{R}$ is a worker's (latent) base skill level, $Y\in\{0,1\}$ a workers productivity and $X\in\mathbb{R}$ be a noisy productivity assessment (e.g. the outcome of an interview). Throughout, it is assumed that conditioned on S, productivity is independent of G and that*

$$Y|S \overset{d}{=} Ber[\sigma(S)].$$

The productivity assessment X is independent of G given Y and specified by a conditional CDF

$$I(X\mid y) \triangleq \mathbb{P}\{X\le x \mid Y=y\}$$

that decreases in y (at a fixed x). We note that this also specifies the generation of X as

$$\Phi(X\mid s) \triangleq \sigma(s)I(X|1) + (1-\sigma(s))I(X|0).$$

Intuitively, the assumption that $I(X\mid y)$ decreases in y at a fixed x, requires the productivity assessment to be "unbiased" (in the sense of a statistical test). Under this unbiased assumption, the optimal hiring policy for the firm is of the form $f(x,\theta,g)=\mathbf{1}_{x\ge\theta_g}$. Let $u(f,y)$ be the firm's received utility from hiring ($f(x)=1$) or not hiring ($f(x)=0$) a worker with productivity y. The firm's expected utility for policy $f:\mathbb{R}\times\{A,D\}\to\{0,1\}$ is $\mathbb{E}\big[u(f(X,g),Y)\big]$, so the firm's utility maximization problem is

$$\max_f \sum \lambda_g \mathbb{E}\big[u(f(X,g),Y)\big].$$

As in [10], we allow the workers to improve their skills (at a cost) in response to the firm's policy. Let $w>0$ be the wage paid to hired workers and $c_g(s,s')>0$ be the sensitive trait dependent cost to workers of improving their skills from s to s'. We assume c is non-increasing in s and non-decreasing in s'. The expected utility a worker with group membership $G=g$ receives from increasing their skill level from s to s' is

$$u_w(f,s,s',g) \triangleq \int_{\mathbb{R}} w f(x,g) d\Phi(x\mid s') - c_g(s,s'),$$

so a strategic worker changes their skill level to maximize their expected utility. We encode the ex-post *workers' skill level, skill level assessment, and productivity as*

$$\begin{aligned} S' &\triangleq \arg\max_{s'} u_w(f,s,s',g),\\ X'\mid S' &\sim \Phi(x\mid S'),\\ Y'\mid S' &\sim Ber[\sigma(S')] \end{aligned}$$

Note that the (conditional) distribution of a worker's skill level assessment, given their skill level, remains the same before and after the worker changes their skill level. To account for the strategic behavior of the workers, the firm solves the performative *policy learning problem:*

$$\max_f \sum \lambda_g \mathbb{E}_{Z'\sim\mathcal{D}(f,g)}\big[u(f(X',g),Y')\big]$$

The workers do not respond instantly to the employer's hiring policy; it takes them a while. Thus, we interpret $\mathcal{D}(f,G)$ as the *long term* distribution of the workers' skill levels and assessments in response to the employer's hiring policy. More concretely, imagine a labor market in which the workers slowly turn over: new workers enter the workforce and old workers retire constantly. As workers enter the workforce, they make their human capital investment decisions in response to the employer's (contemporaneous) hiring policy. Over a long period, the labor force population will converge to $\mathcal{D}(f,G)$.

2.1 Standard fairness constraints are insufficient in performative prediction

The standard way to enforce fairness in policy learning problems is to equalize certain fairness metrics between demographic groups (indicated by a demographic attribute $G\in\mathcal{G}$). This is often done by imposing fairness constraints on the policy learning problem. In general, fairness constraints fall into one of three types:

$$\begin{aligned} &\text{(demographic parity DP)} && f(X)\perp\!\!\!\perp G,\\ &\text{(separation)} && f(X)\perp\!\!\!\perp G\mid Y,\\ &\text{(sufficiency)} && Y\perp\!\!\!\perp G\mid f(X). \end{aligned}$$

To see each of the traditional group fairness constraints in action, consider example 2.1. Enforcing separation requires identical *ex-ante* hiring rates between workers from the advantaged and disadvantaged groups with the same skill level, while enforcing sufficiency requires the *ex-ante* (distribution of) skill levels of the hired workers from the majority and minority groups to be the same. Finally, enforcing demographic parity simply requires that the *ex-ante* hiring rites for individuals be the same across the groups.

In the long-term setting, there are two main issues with such group fairness constraints. First, they only focus on the policy and not its long-term impacts on the population: the policy f appears in all three constraints, but the distribution map $\mathcal{D}(f,G)$ that encodes the long-term impacts of f is absent. In other words, group fairness constraints enforce equal treatment, but ignore the long-term impacts of equal treatment on the population. Consider example 2.1, only *ex-ante* quantities of the workers are considered. This leads us to consider constraints that focus on the long-term impacts of the policy on the population. Instead of enforcing *ex-ante* equal treatment, we seek *ex-post* equality of certain fairness metrics.

Second, traditional group fairness constraints are also plagued by incompatibilities. Despite the intuitive nature of DP, separation and

sufficiency, [9, 36] prove that it is generally impossible for a policy to simultaneously satisfy two of DP, separation and sufficiency.

Theorem 2.2 (Chouldechova [9], Kleinberg et al. [36]). *It is impossible to find a joint distribution on $(f(X), Y, G)$ that satisfies two of DP, separation, and sufficiency, unless one of the following hold:*

(1) $\mathbb{P}(f(X) = Y) = 1$
(2) $Y \perp\!\!\!\perp \mathcal{G}$

This result is generally interpreted as an impossibility result: it is impossible to simultaneously satisfy separation and sufficiency. Although there are two cases in which separation and sufficiency are compatible, they are considered pathological. The first case, perfect prediction, is generally unachievable because the Bayes error rate in most practical prediction problems is non-zero. The second case, independent responses, is also considered pathological because the policymaker has no control over the distribution of responses. However, in performative settings, the policymaker can steer the population so that response distributions are equal, suggesting that it is possible to resolve the incompatibilities between separation and sufficiency in performative settings by enforcing group-independent responses *ex-post*. In the next section, we build on this observation to resolve the incompatibility between separation, sufficiency, and demographic parity in performative settings.

2.2 Equality of Outcomes

The preceding developments suggest that the goal of a fairness-conscious policymaker should be to eliminate disparities between demographic groups in the long term (instead of myopically enforcing group fairness regardless of the long term impacts). Because we are interested in equating the different demographic groups *ex-post* (where an individual from each group ends up) rather than equating the different demographic groups *ex-ante* (where an individual from each group comes from), we refer to this constraint as equality of outcomes.

Definition 2.3 (Equality of outcomes). *A policy f satisfies equality of outcomes with respect to metric $m(f, g)$ if and only if $m(f, g)$ is constant over each possible value of the sensitive trait g.*

We emphasize that fairness metrics $m(\cdot, \cdot)$ are not metrics in the distance sense, but rather a quantity that measures a long-term outcome of interest for strategic individuals. Since we are interested in *ex-post* fairness, each metric will measure quantities associated with the ex-post distribution ($\mathcal{D}(f, G)$). Finally, for simplicity of presentation, we will present each metric for the case of binary classification (*i.e.* both $f(X', g)$ are in $Y' \in \{0, 1\}$). The extension to the multiclass or continuous case is straightforward.

The goal of algorithmic reform in example 2.1 is to equalize the disparities in human capital investment among minority and majority workers. Choosing the fairness metric in definition 2.3 to be worker productivity, we can encode this goal as an instance of definition 2.3.

Definition 2.4 (equality of responses). *A policy f satisfies equality of responses if it satisfies equality of outcomes with respect to the metric $m_{res}(f, g) = \mathbb{E}_{\mathcal{D}(f,g)}[Y']$.*

Equal responses is a concept of fairness unique to the long-term setting (and is our interpretation of the aim of algorithmic reform/reparation); we note that it does not imply the long-term analog of group fairness defined next. We opt to refer to equality of outcomes with respect to any metric containing the policy f as equality of treatment.

Definition 2.5 (Equality of treatment). *A policy f satisfies equality of treatment if f meets all the following criteria.*

(1) *The policy f satisfies equality of outcomes with respect to metric $m_{par}(f, g) = \mathbb{E}_{\mathcal{D}(f,g)}[f(X', g)]$.*
(2) *The policy f satisfies equality of outcomes with respect to both metrics*

$$m_{FPR}(f, g) = \frac{\mathbb{E}_{\mathcal{D}(f,g)}[\mathbf{1}\{Y' = 0\}\mathbf{1}\{f(X', g) = 1\}]}{\mathbb{E}_{\mathcal{D}(f,g)}[1 - Y']}$$

$$m_{FNR}(f, g) = \frac{\mathbb{E}_{\mathcal{D}(f,g)}[\mathbf{1}\{Y' = 1\}\mathbf{1}\{f(X', g) = 0\}]}{\mathbb{E}_{\mathcal{D}(f,g)}[Y']}$$

(3) *The policy f satisfies equality of outcomes with respect to both metrics*

$$m_{PPV}(f, g) = \frac{\mathbb{E}_{\mathcal{D}(f,g)}[\mathbf{1}\{Y' = 1\}\mathbf{1}\{f(X', g) = 1\}]}{\mathbb{E}_{\mathcal{D}(f,g)}[f(X', g)]}$$

$$m_{NPV}(f, g) = \frac{\mathbb{E}_{\mathcal{D}(f,g)}[\mathbf{1}\{Y' = 0\}\mathbf{1}\{f(X', g) = 0\}]}{\mathbb{E}_{\mathcal{D}(f,g)}[1 - f(X', g)]}$$

We we wish to emphasize that requirements one, two and three are simply the long term analogs of demographic parity, separation and sufficiency respectively. Thus, the enforcement of equality of treatment implies the long-term enforcement of multiple fairness constraints that are incompatible in the short term.

Of course, our ultimate goal is for policymakers to implement policies that satisfy equality of responses and equality of treatment *simultaneously*. We point out that these goals are not necessarily disjoint. As previous impossibility theorems have shown, in most cases satisfying equality of responses is a prerequisite to satisfying equality of treatment. We show in the following proposition that satisfying equality of treatment and equality of responses is equivalent to enforcing the independence of the joint distribution $(Y', f(X', G))$ and G.

Proposition 2.6. *A policy f satisfies equality of treatment and equality of responses if and only if the joint distribution $(Y', f(X', G))$ is independent of G.*

As mentioned, equality of responses is a mathematical formalization of the line of work on algorithmic reform/reparation [15, 24]. This line of work "escapes" from incompatibilities between group fairness definitions by questioning the goal of satisfying those definitions. It argues that the underlying goal of enforcing fairness is to remedy injustices. From this perspective, traditional algorithmic fairness definitions are merely a flawed indicator of the true goal, so it is inconsequential if they are incompatible. We encode the goal of reform or reparation mathematically as closing or eliminating disparities in the responses of interest among demographic groups. Furthermore, we study the enforcement of equality of responses and equality of treatment simultaneously. In our formulation, this implies the attainment of algorithmic reform/reparation *without*

discrimination. In [15, 24], it is often argued that reform can only be achieved by discriminating against the advantaged group. In contrast to this, our analysis will show that it is often possible to achieve reform (equality of responses) fairly (equality of treatment).

3 FEASIBILITY OF EQUALITY OF OUTCOMES

A crucial and non-trivial question remains. Is it feasible to enforce equality of treatment and equality of responses simultaneously? Note that this requires the same policy to equate *ex-post* responses and to satisfy equality of outcomes with respect to each long-term group fairness metric. If we deploy one policy to steer the population so that the response distributions are equal and another policy to satisfy the equality of treatment constraints, the second policy may steer the population away from equated responses, and we end up in a cycle that achieves neither goal. Unfortunately, the goal of equal responses and equal treatment may not be possible, even in performative settings; *i.e.* there may not be a policy that achieves both the treatment and the response goals. This is because the distribution map $\mathcal{D}(\theta, G)$ depends on the sensitive attributes (*e.g.* because it encodes inequities in the *ex-ante* distribution or inequities in the agent response map). Thus, equating the responses *ex-post* requires disparate treatment of the group (*i.e.* the policy-maker cannot simply implement the same policy for each group). On the other hand, equality of treatment forbids a disparate allocation of *ex-post* errors between groups. In this section, we study the feasibility of enforcing equality of treatment (and thus equality of responses) in labor market models.

3.1 Impossibility Results

We start by establishing impossibility results to elucidate problem structures that preclude equality of treatment and equality of responses in labor market models. We consider two types of disparities: human capital investment cost disparities and *ex-ante* skill disparities. In order to introduce *ex-ante* skill disparities, we define the notion of stochastic dominance.

Definition 3.1 (stochastic dominance). *Consider two real valued random variables A and B. Then A stochastically dominates B if for all* $x \in \mathbb{R}$, $\mathbb{P}(A \leq x) \leq \mathbb{P}(B \leq x)$.

We return to the continuous labor market model 2.1 in which an employer hires workers from two demographic groups. In order to keep the example as equitable as possible (so that it is as easy as possible for the employers to achieve equality of treatment), recall that we assume that the skill assessment process is fair ($X \perp\!\!\!\perp G \mid Y$) and the same wage is paid to hired workers from both groups. Thus, the only *ex-ante* differences allowed between workers in the two groups are the *ex-ante* distribution of their skill levels and the cost of human capital investment.

Theorem 3.2. *Assume the following:*

(1) *The worker cost function is of the form* $c(s', s, g) = \frac{c_g}{2}(s' - s)_+^2$, *and* $min_g(c_g)$ *is large enough to enforce strong convexity of the agents optimization problem*
(2) *Exactly one of two forms of market discrimination is present:*
   (a) *There is a difference in* ex-ante *skill levels (specifically* $S|G = A$ *stochastically dominates* $S|G = D$*).*
   (b) *The is a difference in the cost of human capital investment of the form* $c_A < c_D$.

Then, under either form of discrimination, group-blind policies that ignore the demographic attribute of the workers G cannot achieve equality of responses. Furthermore, hiring policies that satisfy equality of outcomes with respect to $m_{FPR}(f, g)$ *and* $m_{FNR}(f, g)$ *are necessarily group-blind.*

3.2 Alternative performative models

We present the preceding negative results to emphasize that simultaneously achieving equality of responses and equality of treatment with respect to various fairness metrics is not trivial. To overcome the difference *ex-ante* between workers in the two groups, employers must offer additional incentives to the *ex-ante* least skilled group to close the skill gap and achieve equal responses. Unfortunately, this prevents them from treating workers from the two groups equally because employers only have a single degree of freedom (the hiring threshold θ). This suggests that it may be possible for employers with more degrees of freedom to equalize *ex-post* response distributions with an (*ex-post*) fair hiring policy.

In this section, we introduce two models, the first is a generic model of causal strategic classification inspired by the work [60]. The second is a modification of the causal strategic classification model to better model labor markets. The causal strategic classification model provides a simplified framework for theoretical questions on feasibility and generalization, while also demonstrating that our fairness framework is applicable to a wide variety of learning settings.

Example 3.3 (causal strategic classification [60]). *Consider a learning setting, in which samples correspond to strategic agents that posses features and a sensitive trait* $(X, G) \in \mathbb{R}^d \times \{A, D\}$. *Conditioned on sensitive trait membership G, features X are generated from the* ex-ante *distribution*

$$X|G \stackrel{d}{=} P_G.$$

Conditioned on the features X, the agent responses $Y \in \{0, 1\}$ *are generated via the Bernoulli variable*

$$Y|X \stackrel{d}{=} Ber[\sigma(\beta^T X)].$$

Note that this implies $Y \perp\!\!\!\perp G|X$. *The learner wishes to accurately classify the agents using the features and sensitive trait. As such, the learner deploys predictions* $f(X, G) \in \{0, 1\}$ *generated with the model*

$$f(X, G)|X, G \stackrel{d}{=} Ber[\sigma(\theta_G^T X)].$$

In response to model choice θ_g, *an agent (with trait g and* ex-ante *features x) is allowed to take some action a to improve their standing with the learner (at cost* $C_g(a)$*). The agents act rationally, optimizing their utility:*

$$a(\theta_g) = \arg\max_{a' \in \mathbb{R}^k} [\theta_g]^T [x + M_{d \times k} a'] - C_g(a')$$

Upon selecting action $a(\theta_g)$ *the agent's features are* ex-post $x' = x + Ma(\theta_g)$. *The matrix M is an* effort conversion *matrix, encoding the improvement of the feature* x_i *from the action* a_j *for each* $i, j \in [d] \times [k]$.

At the population level, the ex-post feature distribution X' for the group g is given by

$$P'_g \overset{d}{=} T_{\#}(P_g; \theta_g, M, C_g); \text{ where } T(x; \theta_g, M, C_g) = x + Ma(\theta_g).$$

Conditioned on G and X ex-post *responses are generated by*

$$Y'|X' \overset{d}{=} Ber[\sigma(\beta^T X')],$$

and ex-post predictions are generated by

$$f(X', G)|X', G \overset{d}{=} Ber[\sigma(\theta_G^T X')].$$

To prevent arbitrary inflation of all agent outcomes, the learner is subject to regularization penalty $||\theta||_2^2$. From the above, the learners ex-post risk is

$$R(\theta) = \sum_{g \in \mathcal{G}} \lambda_g [\mathbb{P}(f(X', G) \neq Y'|G = g) + ||\theta_g||_2^2].$$

To better model a labor market, we modify example 3.3. The learning setting will correspond to a labor market and the strategic agents will correspond to workers. The key difference is that, rather than features, workers now have a profile of latent skills, and each skill contributes to the productivity of a worker. Consequently, firms now view a noisy measurement of each skill based on a factor model. This provides workers with multiple ways of investing in their human capital and firms with flexibility in their hiring policies. As we shall see, this additional flexibility is crucial for the enforcement of equality of treatment.

Example 3.4 (Modified Labor Market Model). *Consider the causal strategic classification set up (example 3.3). In the context of a labor market, the learner corresponds to a hiring firm, and each strategic agent is a worker. Workers are encoded by pairs $(S, G) \in \mathbb{R}^d \times \{A, D\}$, with S corresponding to a **latent** skill profile; as before $S \not\perp\!\!\!\perp G$ and we say $S|G = g \overset{d}{=} P_g$. The productivity of the workers in the group g is generated by $Y|G = g \sim Ber[\sigma(\beta^T S)]; S \sim P_g$. Rather than observing latent skill profiles, the hiring firm views interview outcomes generated by $X = \Lambda S + \epsilon$, with $\Lambda \in \mathbb{R}^{p \times d}$ a matrix of factor loadings. This skill assessment model is motivated by item response theory (IRT) models of test outcomes [42]. The results of the interviews are used to make hiring decisions through policy $f(X, G)|X, G \sim Ber[\sigma(\theta_G^T X)]$.*

A worker in group g with an initial skill profile s can take action (possibly training, studying, or additional education) in response to θ_g through the causal strategic classification response mechanism

$$a(\theta_g) = \arg\max_{a' \in \mathbb{R}^k} [\Lambda^T \theta_g]^T [s + M_{d \times k} a'] - C(a', g).$$

The ex-post skill profiles are $S' = S + Ma(\theta_g), S \sim P_g$, which propagates to the ex-post interviews X', productivity Y' and hiring decisions $f(X', g)$. The firm seeks to maximize some (regularized) ex-post reward/profit

$$R(\theta) = \sum_{g \in G} \lambda_g \mathbb{E}[r(S', X', \theta_g)|G = g] - ||\theta_g||^2$$

3.3 Feasibility of Equality of Treatment in Alternative Models

In contrast to the models in example 2.1, in the causal strategic classification setting (and the alternative labor market mode)l, the set of policies that enforces equality of treatments (and thus equality of responses by 2.6) is non-empty. In fact, it contains a stratified manifold of dimension $O(d)$, so the set is quite large in some sense. We will study this set in two scenarios: correcting *ex-ante* feature/skill disparities and correcting cost-of-improvement disparities. For the purposes of a theoretical analysis, we operate under some simplifying assumptions.

Assumption 3.5.

(1) *The agent (worker) cost is quadratic: $C(a, g) = \frac{c_g}{2}||a||_2^2$, the effort matrix M is of the form $M = diag[B]; B \in \{0, 1\}^d$.*
(2) *The ex ante features (latent skill profiles) and interview outcome distributions are Gaussian. In Example 3.3 $X|G \overset{d}{=} \mathcal{N}(\mu_G, I)$ while in example 3.4 $S|G \overset{d}{=} \mathcal{N}(\mu_G, I)$ and $\epsilon \overset{d}{=} \mathcal{N}(0, I)$.*

The quadratic cost assumption is standard in the strategic learning literature, [32, 33, 60]. The effort matrix is of the form $\text{diag}[B]; B \in \{0, 1\}^d$, if each skill is improved by a distinct action and only *some* skills can be improved. Each assumption (including normality) is primarily for mathematical convenience; we expect that feasibility will hold under a wide class of choices for $C(a), \Lambda, M$ and measures on S, ϵ.

Under assumption 3.5, the feasibility of equality of treatment can be studied by analyzing two disjoint constraint sets, one that pertains to parameters that correspond to "manipulable" features and another that pertains to "nonmanipulable" features, which we now define.

Definition 3.6 (Manipulable features). *Given an effort matrix $M = diag[B]; B \in \{0, 1\}^d$, and a general feature vector $v \in \mathbb{R}^p$, we let $v_m = \{v_i \in v; \mathbf{1}_{\{M_i = 1\}}\}$ and $v_u = \{v_i \in v; \mathbf{1}_{\{M_i = 0\}}\}$ be the manipulable and nonmanipulable features, respectively.*

Additionally, we will assign $2d_m$, $2d_u$ as the dimensions of the parameter spaces $(\theta_{m,A}, \theta_{m,D}), (\theta_{u,A}, \theta_{u,D})$.

Theorem 3.7. *Consider the learning setting of example 3.3 with the minor assumption that the vectors $\{(\mu_{A,m}, -\mu_{D,m}), (\beta_m, -\beta_m)\}$ are not co-linear, and the vectors $\{(\mu_{A,u}, -\mu_{D,u}), (\beta_m, -\beta_m)\}$ are not co-linear. Suppose that one of the two following forms of discrimination is present:*

(1) *Ex-ante distribution discrimination: $c_A = c_D = 1$, but $\mu_A^T \beta > \mu_D^T \beta$,*
(2) *Cost of improvement discrimination: $\mu_A = \mu_D = 0$ but $c_A < c_D$.*

Then under (1) or (2) there exist stratified manifolds $\mathcal{M}_m$, $\mathcal{M}_u$ such that $dim[\mathcal{M}_u] = d_u - 2$, and $dim[\mathcal{M}_m] = d_m - 2$ and any learner decision $\theta = (\theta_A, \theta_D)$ that satisfies $(\theta_{A,m}, \theta_{D,m}) \in \mathcal{M}_m$ and $(\theta_{A,u}, \theta_{D,u}) \in \mathcal{M}_u$ also satisfies equality of treatment and equality of responses.

Corollary 3.8. *Consider the modified labor market model (example 3.4). Assume that worker discrimination of the form (1) or (2) is present. Then if $d_m = d_u \approx d/2$ there exists a stratified manifold $\mathcal{M}$ of dimension $O(d)$ such that any $\theta = (\theta_A, \theta_D) \in \mathcal{M}$ satisfies equality of treatment.*

Theorem 3.7 and corollary 3.8 state that in the causal strategic classification model/modified labor market model, the set of policies

that enforce long-term fairness and correct differences in worker skills or costs contains a manifold of dimension $O(d)$. Here, d is interpreted as the "number of skills" a worker can possess. This result clarifies the importance of flexibility in human capital investment. If d is not large enough, then the feasible subsets of Theorem 3.7 will be either too small to provide policy makers flexibility or even empty in extreme cases. The assumptions of theorem 3.7 also provide an important form of *policy maker* flexibility; the existence of skills that are immune to performative effects implies the existence of policy parameters that can be adjusted to change error rates between groups *without* impacting downstream responses.

Open questions on the feasibility of equality of treatment in labor markets with multiple forms of discrimination or continuous outcomes remain. The assumption that only one form of discrimination is present (cost of education discrimination *or ex-ante* skill discrimination) is necessary for our analysis but not necessary for feasibility (see Figure 1 for a numerical example). The discrete nature of worker productivity and firm hiring decisions is also not strictly necessary; for example. The argument of theorem 3.7 immediately implies feasibility (assuming Gaussian features) in the Causal Strategic Least Squares model posited in [60]. This model is both the inspiration for our labor market model and can also be interpreted in the context of a labor market with continuous worker productivity.

4 A REDUCTION ALGORITHM FOR EQUALITY OF OUTCOMES

In practice, a policy maker often does not have complete knowledge of the *ex-ante* and *ex-post* distributions but instead only observes some samples from the *ex-ante* distribution and has some model for how individuals respond to their policy. We turn to the question of implementing equality of outcomes under such conditions, providing a reduction algorithm (inspired by Agarwal et al. [1]) adapted to the performative setting. By proposition 2.6 a policy maker can implement equality of treatment and equality of responses by enforcing the independence of the joint distribution $(Y', f(X', G))$ and G. We propose that the policymaker enforce this through a series of moment inequality constraints:

$$M\mu(f) \leq c,$$

$$\mu(f)_{ij} = \mathbb{E}[h_i(f(X'), Y')|G = g_j].$$

Throughout this section, we will work with the causal strategic classification setting (example 3.3) in the two group setting. In this model, equal responses and equal treatment can be enforced through a series of 6 moment constraints.

Example 4.1. *Consider the causal strategic classification example 3.3 with two possible groups $\{A, D\}$. In this setting, the condition $(Y', f(X', G)) \perp\!\!\!\perp G$ is equivalent to the constraint:*

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mathbb{E}[Y'|A] \\ \mathbb{E}[Y'|D] \\ \mathbb{E}[f(X', G)|A] \\ \mathbb{E}[f(X', G)|D] \\ \mathbb{E}[f(X', G)Y'|A] \\ \mathbb{E}[f(X', G)Y'|D] \end{pmatrix} \leq 0_6,$$

while the ex-post *risk is given by*

$$EPR(\theta) = \sum_{g\in\{A,D\}} \lambda_g [\mathbb{P}_{\mathcal{D}(f,g)}(f(X', g) \neq Y') + ||\theta_g||_2^2].$$

The case of multi-class classification or multiple sensitive traits is relatively similar. For multiclass classification, enforcement of the independence requirement will require the addition of higher-order moment constraints, and moment constraints may simply be repeated for each possible sensitive trait combination in the case of multiple sensitive traits.

Recall that the policy maker is vested in minimizing some *ex-post* risk, therefore, in aggregate, the policymaker must solve a constrained optimization problem in f. The primal problem is

$$\mathcal{L}(f; \lambda) = \min_{f\in\mathcal{F}} \max_{\lambda\geq 0} \mathrm{EPR}(f) + \lambda^T (M\mu(f) - c). \quad (4.1)$$

In practice, the policymaker only observes samples $\{Z_i\}_{i=1}^n$ from the *ex-ante* distribution. We will assume that the policymaker has access to some correctly specified model of $\mathcal{D}(f, g)$ at the sample level. For example, a hiring firm in a labor market would need to be aware of each worker's cost-adjusted utility optimization problem. Using *ex-ante* samples $\{Z_i\}_{i=1}^n$ and knowledge of $\mathcal{D}(f)$, the policymaker can obtain the natural empirical estimates of $\mathrm{EPR}(f)$ and $\mu(f)$ (denoted $\widehat{\mathrm{EPR}}(f)$ and $\hat{\mu}(f)$). As an example of obtaining estimates for $\mathrm{EPR}(f)$ and $\mu(f)$ from $\mathcal{D}(f)$ we return to the modified labor market model.

Example 4.2 (Equality of outcomes estimates in causal strategic classification). *Recall the setting of Example 3.3. Consider an agent with sensitive trait $G = g$ and ex-ante features x, upon viewing policy θ_g, the agent invests in their own features via $x' = x + Ma(\theta_g, g)$. Given an* ex-ante *sample of skill features from group g, $\{x_i\}_{i=1}^n$ a reasonable choice of estimates for $EPR(f)$ and $\mu(f)$ are*

$$\mathbb{E}[\widehat{Y'|G = g}] = \frac{1}{n}\sum_{i=1}^n \sigma(\beta^T(x_i + Ma(\theta_g, g))),$$

$$\mathbb{E}[f(\widehat{X', G})|G = g] = \frac{1}{n}\sum_{i=1}^n \sigma(\theta_g^T(x_i + Ma(\theta_g, g))),$$

$$\mathbb{E}[f(\widehat{X', G)Y'}|G = g] = \frac{1}{n}\sum_{i=1}^n \sigma(\beta^T(x_i + Ma(\theta_g, g)))\sigma(\theta_g^T(x_i + Ma(\theta_g, g))),$$

$$\widehat{EPR}(\theta_g) = \frac{1}{n}\sum_{i=1}^n \sigma(\theta_g^T(x_i + M_g a(\theta_g, g)))(1 - \sigma(\beta^T(x_i + M_g a(\theta_g, g))))$$

$$+(1 - \sigma(\theta_g^T(x_i + M_g a(\theta_g, g))))(\sigma(\beta^T(x_i + M_g a(\theta_g, g)))).$$

After attaining estimates $\hat{\mu}(f)$ and $\widehat{\mathrm{EPR}}(f)$, equation 4.1 can be replaced with said estimates. Furthermore, for convergence reasons, a L_1 norm constraint is placed on the dual variable λ. Finally, due to statistical error, a relaxation $\hat{c} = c + \nu$ is allowed on the moment constraint. If the learner instead opts to solve the dual problem (the justification for this is expanded upon in Appendix A), the final result is

$$\mathcal{L}(f; \lambda) = \max_{\lambda\geq 0; ||\lambda||_1 \leq B} \min_{f\in\mathcal{F}} \widehat{\mathrm{EPR}}(f) + \lambda^T (M\hat{\mu}(f) - \hat{c}). \quad (4.2)$$

From here, the iterates of the dual variables are obtained using mirror ascent on the dual variable with the potential function $\phi(\lambda) = -\lambda ln(\lambda)$, which algorithm 1 lays out explicitly.

Algorithm 1 Reduction for Equality of Outcomes

Require: Error tolerance ϵ, step size η, samples $\{Z_i\}_{i=1}^n$, fairness constraints $M, \hat{c}, \mu$, initial iterate v_0

1: **for** $t = 0, 1, \ldots$ **do**
2: Scale dual: $\lambda_{k,t} \leftarrow B\frac{e^{v_{k,t}}}{1+\sum_k e^{v_{k,t}}}$
3: Get policymakers best decision: $f_t \leftarrow \arg\min_{f\in\mathcal{F}} \widehat{\mathrm{EPR}}(f) + \lambda_t^T(M\hat{\mu}(f) - \hat{c})$
4: **if** (λ_t, f_t) is an ϵ-saddle point **then**
5: **return** (λ_t, f_t)
6: **else**
7: Update iterates: $v_{t+1} \leftarrow v_t + \eta_t(M\hat{\mu}(f_t) - \hat{c})$
8: **end if**
9: **end for**

Two elements of the algorithm 1, are nontrivial: (i) the ϵ-saddle point stopping criteria; (ii) the attainment of the best decision of the policy makers f_t.

The ϵ-approximate saddle point stopping criteria: A primal dual pair (f_t, λ_t) is a ϵ-saddle point if the following hold:

$$\mathcal{L}(f_t, \lambda_t) \leq \min_{f\in\mathcal{F}} \mathcal{L}(f, \lambda_t) + \epsilon$$

$$\mathcal{L}(f_t, \lambda_t) \geq \max_{\lambda\geq 0; ||\lambda||_1 \leq B} \mathcal{L}(f_t, \lambda) - \epsilon$$

Checking the first criteria reduces to a problem in attainment of the policy makers best decision f_t[1]. The second requirement requires solving a linear program with an L1 inequality constraint, a well-studied problem [4].

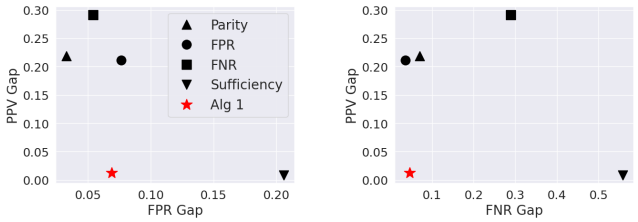
Obtaining of the policymakers best decision f_t: Attaining the best long-term policy for risk function $\widehat{\mathrm{EPR}}(f) + \lambda_t^T(M\hat{\mu}(f) - \hat{c})$ is generally a nontrivial problem. Previous works have established methods for obtaining the best policy f under the assumption that the policy maker knows the map $\mathcal{D}(f)$ [30, 38, 62]. Such methods are generally specific to a particular $\mathcal{D}$, and since our focus is on fairness, we will assume that the policymaker has access to some oracle which produces such an f. This is a strong assumption, and our methodology is limited to performative maps that allow for such an oracle.

4.1 Algorithm 1 in the Modified Labor Market Model

As an application of Algorithm 1, we study the problem of enforcing equality of outcomes and equality of responses in the modified labor market (example 3.4) when both *ex-ante* distributions **and** cost of education are different between each group. We assume that $\Lambda = I$, so that the firm has an unbiased estimate of each skill. Figure 1 demonstrates the performance of algorithm 1 on a held-out test set of workers. Prior work on long-term fairness studied the impact of enforcing standard fairness constraints in the long term, and, as such, we utilize this as a base line. Algorithm 1 is compared to policies that equalize one *ex-ante* fairness metric, including false positive rate, false negative rate, demographic parity, and sufficiency.

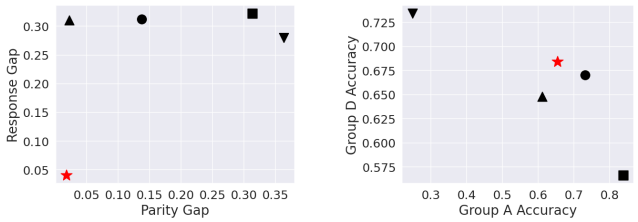
Figure 1 **(a)** and **(b)** demonstrate that in the long term, the policy deployed by Algorithm 1 enforces both sufficiency and separation. Figure 2 **(a)** shows that this same policy satisfies equality of responses and (long-term) parity. Finally, Figure 2 **(b)** also

Figure 1



(a) **(b)**

Figure 2



(a) **(b)**

demonstrates that enforcing equality of treatment and equality of responses will equate the accuracy of a policy between the two groups. Due to statistical error (exacerbated by the opaque nature of worker skills), perfect fairness is not achieved on the test set. We emphasize that this is not due to an issue of *feasibility*, figure 3 demonstrates that algorithm 1 can attain nearly zero fairness violation on the training set.

5 SUMMARY AND DISCUSSION

In this paper, we studied fairness in performative settings in which the policymaker has the ability to steer the population. We showed that it is possible for the policymaker to remedy existing inequities in the population. In particular, we showed that by equating the distribution of responses Y between groups in classification problems, it is possible for the policymaker to simultaneously satisfy multiple notions of group fairness that are generally incompatible in non-performative settings. However, we also showed that this is not always possible: if the policymaker does not have enough flexibility in how they can equalize base rates, then it is unfortunately impossible, even in performative settings, to resolve the longstanding incompatibilities between group fairness definitions. Another limitation of our approach is that the policymaker must be aware of the long-term impacts of their policies on the population. Although this requirement is necessary, it limits the applicability of the approach. One possible direction for future work is to develop methods that help the policymaker estimate the effects of their policies on the population. Such methods can be combined with our approach to steer sociotechnical systems towards more equitable states.

Our work is also aligned with the goals of algorithmic reform/reparations. By considering reform/reparations mathematically, we show that it is somewhat possible to achieve the goals

of reform without unequal treatment. This is especially desirable in application domains in which unequal treatment is illegal or impractical. For example, consider the problem of underrepresentation of women in the tech sector, especially in technical roles [14]. The authors of Davis et al. [15] suggest that employers in the tech sector should adopt a "reparative" approach to equalize the representation of men and women, even if this entails explicitly discriminating against men. They justify explicit discrimination by appealing to the historical injustices that led to the dearth of women in the tech sector and the need to remedy such injustices. Although a discriminatory approach is likely to be limited by various labor laws, our results suggest that it may be possible to equalize the representation of men and women while treating men and women fairly. We hope that our results lead to more serious consideration of "reparative" approaches in algorithmic decision making.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation (NSF) under grants no. 2027737 and 2113373.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 60–69.

[2] Kenneth Arrow. 1971. The Theory of Discrimination. *Labor Economics vol 4* (1971).

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv:1703.09207 [stat]* (March 2017). arXiv:1703.09207 [stat]

[4] Stephen P. Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK ; New York.

[5] Gavin Brown, Shlomi Hod, and Iden Kalemaj. 2022. Performative Prediction in a Stateful World. In *International Conference on Artificial Intelligence and Statistics*.

[6] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. 2019. From Soft Classifiers to Hard Decisions: How fair can we be? arXiv:1810.02003 [cs.LG]

[7] Ben Casselman and Dana Goldstein. 2015. The New Science of Sentencing. https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

[8] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 319–328. https://doi.org/10.1145/3287560.3287586

[9] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163. https://doi.org/10.1089/big.2016.0047

[10] Stephen Coate and Glenn C. Loury. 1993. Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review* 83, 5 (1993), 1220–1240. jstor:2117558

[11] Ashley C Craig and Jr Fryer, Roland G. 2017. *Complementary Bias: A Model of Two-Sided Statistical Discrimination*. Working Paper 23811. National Bureau of Economic Research.

[12] Kate Crawford. 2017. The Trouble with Bias.

[13] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 525–534. https://doi.org/10.1145/3351095.3372878

[14] Jeffrey Dastin. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters* (Oct. 2018).

[15] J. L. Davis, A. Williams, and M. W. Yang. 2021. Algorithmic reparation. *Big Data and Society* (2021).

[16] Rahul C. Deo. 2015. Machine Learning in Medicine. *Circulation* 132, 20 (Nov. 2015), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[17] Kate Donahue and Jon Kleinberg. 2020. Fairness and utilization in allocating resources with uncertain demand. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM. https://doi.org/10.1145/3351095.3372847

[18] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. 2023. Group-Fair Classification with Strategic Agents *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 389–399. https://doi.org/10.1145/3593013.3594006

[19] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. 2023. Group-Fair Classification with Strategic Agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 389–399. https://doi.org/10.1145/3593013.3594006

[20] Hanming Fang and Andrea Moro. 2011. Chapter 5 - Theories of Statistical Discrimination and Affirmative Action: A Survey. Handbook of Social Economics, Vol. 1. North-Holland, 133–200. https://doi.org/10.1016/B978-0-444-53187-2.00005-X

[21] Roland G. Fryar and Glenn C. Loury. 2005. Affirmative Action in Winner-Take-All Markets. *The Journal of Economic Inequality* (2005).

[22] Roland G. Fryar and Glenn C. Loury. 2013. Valuing Diversity. *Journal of Political Economy* (2013).

[23] Roland G. Fryar, Glenn C. Loury, and Tolga Yuret. 2008. An Economic Analysis of Color-Blind Affirmative Action. *The Journal of Law, Economics, and Organization.* (2008).

[24] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy and Technology* 35, 4 (Oct. 2022). https://doi.org/10.1007/s13347-022-00584-6

[25] Limor Gultchin, Vincent Cohen-Addad, Sophie Giffard-Roisin, Varun Kanade, and Frederik Mallmann-Trenn. 2022. Beyond Impossibility: Balancing Sufficiency, Separation and Accuracy. arXiv:2205.12327 [cs.LG]

[26] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. 2022. Performative Power. *arXiv:2203.17232 [cs, econ]* (March 2022). arXiv:2203.17232 [cs, econ]

[27] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS '16)*. Association for Computing Machinery, New York, NY, USA, 111–122. https://doi.org/10.1145/2840728.2840730

[28] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[29] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2692–2701.

[30] Guy Horowitz and Nir Rosenfeld. 2023. A Tale of Two Shifts: Causal Strategic Classification. *https://arxiv.org/pdf/2302.06280.pdf* (2023).

[31] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/3287560.3287597

[32] Zachary Izzo, Lexing Ying, and James Zou. 2021. How to Learn When Data Reacts to Your Model: Performative Gradient Descent. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 4641–4650.

[33] Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. 2023. Supply-Side Equilibria in Recommender Systems. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

[34] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 240–248. https://doi.org/10.1145/3287560.3287578

[35] Michael P. Kim and Juan C. Perdomo. 2023. Making Decisions under Outcome Performativity. arXiv:2210.01745 [cs, stat]

[36] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)*. Berkeley, CA. arXiv:1609.05807

[37] Claire Lazar Reich and Suhas Vijaykumar. 2021. A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPICS.FORC.2021.4

[38] Sagi Levanon and Nir Rosenfeld. 2021. Strategic Classification Made Practical. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 6243–6253.

[39] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. *arXiv:1803.04383 [cs, stat]* (March 2018). arXiv:1803.04383 [cs, stat]

[40] Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The Disparate Equilibria of Algorithmic Decision Making when Individuals Invest Rationally. In *ACM Conference on Fairness, Accountability, and Transparency in Machine Learning*.

[41] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. 2020. Too Relaxed to Be Fair. In *Proceedings of the 37th International Conference on Machine Learning*

(Proceedings of Machine Learning Research, Vol. 119), Hal Daumé III and Aarti Singh (Eds.). PMLR, 6360–6369. https://proceedings.mlr.press/v119/lohaus20a.html
[42] F. M. Lord. 1980. *Applications of Item Response Theory To Practical Testing Problems.* Routledge, New York. https://doi.org/10.4324/9780203056615
[43] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. *arXiv:1802.06309 [cs, stat]* (Feb. 2018). arXiv:1802.06309 [cs, stat]
[44] Andreas Maurer. 2016. A vector-contraction inequality for Rademacher complexities. arXiv:1605.00251 [cs.LG]
[45] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. 2022. Anticipating Performativity by Predicting from Predictions. In *Advances in Neural Information Processing Systems*.
[46] Celestine Mendler-Dünner, Juan C. Perdomo, Tijana Zrnic, and Moritz Hardt. 2020. Stochastic Optimization for Performative Prediction. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, 4929–4939.
[47] John Miller, Smitha Milli, and Moritz Hardt. 2020. Strategic Classification Is Causal Modeling in Disguise. *arXiv:1910.10362 [cs, stat]* (Feb. 2020). arXiv:1910.10362 [cs, stat]
[48] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The Social Cost of Strategic Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 230–239. https://doi.org/10.1145/3287560.3287576
[49] Andrea Moro and Peter Norman. 2003. Affirmative Action in a Competitive Economy. *Journal of Public Economics* 87, 3-4 (March 2003), 567–594. https://doi.org/10.1016/S0047-2727(01)00121-9
[50] Andrea Moro and Peter Norman. 2004. A General Equilibrium Model of Statistical Discrimination. *Journal of Economic Theory* 114, 1 (Jan. 2004), 1–30. https://doi.org/10.1016/S0022-0531(03)00165-0
[51] Kirtan Padh, Diego Antognini, Emma Lejal Glaude, Boi Faltings, and Claudiu Musat. 2021. Addressing Fairness in Classification with a Model-Agnostic Multi-Objective Algorithm. arXiv:2009.04441 [cs.LG]
[52] Randall D. Penfield and Gregory Camilli. 2006. 5 Differential Item Functioning and Item Bias. In *Handbook of Statistics*, C. R. Rao and S. Sinharay (Eds.). Psychometrics, Vol. 26. Elsevier, 125–167. https://doi.org/10.1016/S0169-7161(06)26005-X
[53] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 7599–7609.
[54] Case-Kevin Petrasic, Benjamin Saul, James Greig, and Katherine Lamberth. 2017. Algorithms and Bias: What Lenders Need to Know. White & Case LLP.
[55] Edmund Phelps. 1972. The Statistical Theory of Racism and Sexism. *The American Economic Review* (1972).
[56] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf
[57] Manish Raghavan. 2023. What Should We Do when Our Ideas of Fairness Conflict? *Commun. ACM* 67, 1 (dec 2023), 88–97. https://doi.org/10.1145/3587930
[58] Miriam Rateike, Isabel Valera, and Patrick Forré. 2023. Designing Long-term Group Fair Policies in Dynamical Systems. arXiv:2311.12447 [cs.AI]
[59] Cynthia Rudin. 2013. Predictive Policing: Using Machine Learning to Detect Patterns of Crime. *Wired* (Aug. 2013).
[60] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. 2020. Causal Strategic Linear Regression. In *International Conference on Machine Learning*.
[61] Seamus Somerstep, Yuekai Sun, and Ya'acov Ritov. 2023. Learning in Reverse Causal Strategic Environments with Ramifications on Two Sided Markets. In *NeurIPS 2023 Workshop on Algorithmic Fairness through the Lens of Time (AFT2023)*.
[62] Seamus Somerstep, Yuekai Sun, and Yaacov Ritov. 2024. Learning in reverse causal strategic environments with ramifications on two sided markets. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=vEfmVS5ywF
[63] Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. 2023. Long-Term Fairness with Unknown Dynamics. In *Thirty-Seventh Conference on Neural Information Processing Systems*.
[64] Sebastian Zezulka and Konstantin Genin. 2023. Performativity and Prospective Fairness. arXiv:2310.08349 [cs.CY]
[65] Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, and Mingyan Liu. 2022. Fairness Interventions as (Dis)Incentives for Strategic Manipulation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 26239–26264. https://proceedings.mlr.press/v162/zhang22l.html

A THEORETICAL PROPERTIES OF ALGORITHM 1

In our development of the ex-post reduction algorithm (1), we opted to solve the Dual problem rather than the primal problem, to achieve a fair policy. This is only justifiable if strong duality of problem 4.1 holds. Unfortunately, moment constraints of the form 4.1 are often not convex and thus strong duality may not hold. To fix this issue (as is often done in other reduction methods), we can slightly generalize algorithm 1 to allow for randomized policies $Q \in \Delta(\mathcal{F})$, that first select a policy f at random (with $\mathbb{P}(Q = f) = Q(f)$), then make a prediction.

As long as group membership is independent of the policy selected, *i.e.* for any $Q \in \Delta(\mathcal{F})$ the events $1\{G = g\}$ and $1\{Q = f\}$ are independent, one can show that $\mu(Q)$ and EPR are linear in Q.

Proposition A.1. *Suppose that all $Q \in \Delta(\mathcal{F})$ satisfy $Q \perp\!\!\!\perp G$. Then the following holds for all $Q \in \Delta(\mathcal{F})$:*

(1) $\mu(Q) = \sum_{f\in\mathcal{F}} Q(f)\mu(f)$.
(2) $EPR(Q) = \sum_{f\in\mathcal{F}} Q(f)EPR(f)$

The implication of proposition A.1 is that both the *ex-post* risk and fairness constraints are linear in Q; this in turn will give us strong duality. In terms of Q the policymaker's primal problem is

$$\mathcal{L}(Q;\lambda) = \min_{Q\in\Delta}\max_{\lambda\geq 0} \mathrm{EPR}(Q) + \lambda^T(M\mu(Q) - c). \quad \text{(A.1)}$$

Because this problem is linear in Q and λ, the domains of Q and λ are convex, and the equality of treatments constraint is feasible (theorem 3.7) the solution to A.1 will be the unique saddle point (Q^*, λ^*), which algorithm 1 (appropriately modified to include randomization) will converge to. Specifically, if $\{f_t\}_{t=1}^T$ and ${\lambda_t}_{t=1}^T$ are the iterates of algorithm 1, then the empirical measure $Q_T = \frac{1}{T}\sum_{t=1}^T f_t$ and the mean $\bar{\lambda}_T = \frac{1}{T}\sum_{t=1}^T \lambda_t$ will eventually converge to an appropriate saddle point.

Proposition A.2 (Agarwal et al. [1]Theorem 1). *Let $Q_T = \frac{1}{T}\sum_{t=1}^T f_t$, $\bar{\lambda}_T = \frac{1}{T}\sum_{t=1}^T \lambda_t$ be the empirical distribution (resp. average) of the primal (resp. dual) iterates of algorithm 1. Let $\rho = \sup_f ||M\mu(f) - c||_\infty$, K be the total number of moment constraints, and $\eta_t = \sqrt{log(K)+1}/\rho\sqrt{t}$. Then $(\bar{\lambda}_T, Q_T)$ is an ϵ_T saddle point with $\epsilon_T = 2\rho B\sqrt{(log(K)+1)/T}$*

Besides loss of precision due to optimization error, questions on the statistical error of policies produced by algorithm 1 remain. Unlike the case of optimization error, the statistical error analysis is not identical to the analysis in [1]. This is due to the presence of performativity, which can affect the uniform convergence of any estimator. For concreteness, we consider the causal strategic classification example, recovering the classic parametric rate.

Assumption A.3.

(1) *The learning setting is example 3.3 with $M\mu(\theta) - c$ of the form in example 4.1, and $\widehat{EPR}(\theta)$, $\hat{\mu}(\theta)$ of the form in example 4.2.*
(2) *The parameter space $\Theta \subset \mathbb{R}^d$ is compact and $||X||_\infty$ is bounded above with probability one. Furthermore, the response $Ma(\theta)$ is bounded.*

Theorem A.4. *Let n_A, n_D denote the number of samples observed by the policymaker from each group. Suppose $(\hat{Q}, \hat{\lambda})$ is an ϵ-saddle*

point of 4.2 with $\nu = O(min(n_A, n_D)^{-1/2})$ *and* $c = 0_6$. *Let* Q^* *minimize* $EPR(Q)$ *subject to* $M\mu(Q) \leq c$. *Then with probability at least* $1 - 7\delta$ *the distribution* $\hat{Q}$ *satisfies*

$$EPR(\hat{Q}) \leq EPR(Q^*) + 2\epsilon + \tilde{O}(min(n_A, n_D)^{-1/2})$$

$$||M\mu(\hat{Q})||_\infty \leq \frac{1+2\epsilon}{B} + \tilde{O}(min(n_A, n_D)^{-1/2})$$

where $\tilde{O}$ *hides square root dependence on* $ln(1/\delta)$.

In practice, randomization is often an unnecessary complication, and for simplicity, experiments are performed utilizing the non-randomized version of algorithm 1.

B COATE AND LOURY RESULT

In this section we cover a similar impossibility theorem to 2.2 for the Coate and Loury model.

Example B.1 (Coate-Loury labor market model [10]). *Consider an employer that wishes to hire skilled workers which reside in one of two identifiable groups* $G \in \{A, D\}; G \sim Ber[\lambda]$. *The workers are represented as* (X, Y, G) *tuples. Here* $Y|G = g \in \{0, 1\}$ *drawn from a Bernoulli*(π_g) *distribution represents the qualification/skill of a worker, and* $X \in [0, 1]$ *some noisy signal (possible the outcome of an skill assessment) drawn from CDF* $\Phi(X|Y)$. *We will assume that* $\Phi(X|Y = 1)$ *stochastically dominates* $\Phi(X|Y = 0)$, *and additionally that* $X \perp\!\!\!\perp G|Y$. *As such the hiring firm opts to deploy hiring policy* $f(x, \theta, g) = 1_{x \geq \theta_g}$. *The performative aspect of the model is that post deployment of any hiring policy* θ_g, *the workers select their qualification level in response to the employer's policy. If* $w > 0$ *is the wage paid to a worker and* C_g *is the (random and drawn from CDF* E_g*) cost of attaining qualification, then for a worker from group* g *with observed cost* c_g *the utility of each selecting each option of skill* y *is*

$$u_w(\theta_g, y, c_g) \triangleq \begin{cases} \int_{[\theta_g,1]} wd\Phi(x \mid 1) - c_g & \text{worker selects } y = 1, \\ \int_{[\theta_g,1]} wd\Phi(x \mid 0) & \text{if the worker remains unskilled,} \end{cases}$$

Each worker acts rationally and selects the y *that maximizes their utility, at the sample level the performative map for a worker with sensitive trait* g *is*

$$Y' \leftarrow \arg\max_{y' \in \{0,1\}} u_w(f, Y, y'),$$
$$X' \mid Y' \sim \Phi(x \mid Y').$$

In aggregate, the proportion of qualified workers (in a given group) is updated via

$$\pi_g(\theta_g) = E_g(w(P(X > \theta_g|Y = 1) - P(X > \theta_g|Y = 0))).$$

The workers' do not respond instantly to the employer's hiring policy; it takes them a while. Thus we interpret $\mathcal{D}(f)$ *as the* long term *distribution of the workers' skill levels and assessments in response to the employer's hiring policy. More concretely, imagine a labor market in which the workers slowly turn over: new workers enter the workforce and old workers retire constantly. As workers enter the workforce, they make their human capital investment decisions in response to the employer's (contemporaneous) hiring policy. Over a long period, the labor force will converge to* $\mathcal{D}(f)$. *To account for the long-term effects of their hiring policy on the labor market, the employer solves the* performative *policy learning problem:*

$$R(\theta) = \sum_{g \in G} \lambda_g [p_+ \mathbb{P}(X > \theta_g \mid y = 1)$$
$$\pi_g(\theta_g) - p_- \mathbb{P}(X > \theta_g \mid y = 0)(1 - \pi_g(\theta_g))],$$

where p_+ *and* p_- *are the firm's utilities for hiring a qualified and unqualified worker respectively.*

Proposition B.2. *Consider the discrete labor market example B.1, recall the mechanism of discrimination:*

1. *Wages are independent of group membership.*
2. *There is no differential item functioning (DIF). [52] in the skill assessment* $(X \perp\!\!\!\perp G \mid Y)$
3. *Cost of education is discriminatory, i.e.* $E_A(c) \overset{d}{\neq} E_D(c)$.

Thus, the only *source of discrimination is the cost of education. Suppose that this discrimination is of the following form: the cost random variables* C_A, C_D *are unbounded and* C_D *stochastically dominates* C_A, *so that group* D *is strictly disadvantaged through the cost of education. Then, for any hiring policy* $\theta = (\theta_A, \theta_D)$ *that satisfies equality of treatments (and thus equality of outcomes), it holds that* $\pi_A(\theta_A) = \pi_D(\theta_D) = 0$.

Proof. Let TPR, FPR denote the true positive and false positive rate of a classifier. Note that $\pi_g(\theta_g) = G_g(w(\text{TPR}(\theta_g) - \text{FPR}(\theta_g)))$. Note that ex-post separation would require that $\text{TPR}(\theta_A) = \text{TPR}(\theta_D)$ and $\text{FPR}(\theta_A) = \text{FPR}(\theta_D)$. However by the assumption that $c > 0 \implies G_A(c) > G_D(c)$, any policy (θ_A, θ_D) that satisfies ex post separation and $w(\text{TPR}(\theta_A) - \text{FPR}(\theta_A)) > 0$ satisfies $G_A(w(\text{TPR}(\theta_A) - \text{FPR}(\theta_A))) > G_D(w(\text{TPR}(\theta_D) - \text{FPR}(\theta_D)))$ thus any policy that satisfies ex post equality must satisfy $\pi(\theta_A) = \pi(\theta_D) = 0$. □

C SECTION 2 AND SECTION 3 PROOFS

C.1 section 2 proofs

Proposition C.1. *A policy* f *satisfies equality of treatment and equality of responses if and only if the joint distribution* $(Y', f(X', G))$ *is independent of* G.

Proof of proposition 2.6: Suppose a policy f satisfies fairness definitions 2.4 and 2.5. We have

$$\phi(f(X'), Y'|G) = p(f(X')|Y', G)p(Y'|G)$$
$$\overset{\text{defs 2.4+2.5}}{(f;z)} = p(f(X')|Y')p(Y') = \phi(f(X'), Y')$$

On the other hand suppose $\phi(f(X'), Y'|G = g) = \phi(f(X'), Y')$ for $g \in |G|$. Trivially we must have $f(X') \perp\!\!\!\perp G$ and $Y' \perp\!\!\!\perp G$. For separation note that $p(f(X')|Y', G) = \phi(f(X'), Y'|G)/(p(Y'|G)) = \phi(f(X'), Y')/(p(Y')) = p(f(X')|Y')$. Sufficiency will follow from an essentially identical argument. □

C.2 Proof of Theorem 3.7

Lemma C.2. *Under assumption 3.5 an agent in group* g *selects action* $a(\theta_g, g) = \frac{1}{c_g} M\theta_g$.

Proof. Each agent solves

$$a(\theta_g) = \arg\max_{a' \in \mathbb{R}^d} \theta_g^T [x_g + Ma'] - \frac{c_g}{2}||a'||_2^2.$$

We can check the first order optimality condition to see that $a(\theta_g)$ must solve

$$M^T\theta_g - c_g a(\theta_g) = 0$$

By assumption M is diagonal and thus symmetric, so $M^T = M$ and $\frac{1}{c_g} M\theta_g = a(\theta_g)$. □

Lemma C.3. *Any policy $\theta = (\theta_A, \theta_D)$ that satisfies $(\theta_g^T X_g', \beta^T X_g') \perp\!\!\!\perp G$ also satisfies equality of treatment and equality of outcomes.*

Proof. This can be seen by applying, the tower property of conditional expectation, conditioning on X, then use the assumption that $(\theta_g^T X_g', \beta^T X_g') \perp\!\!\!\perp G$. For example, to see that equality of responses holds note that we have:

$$\mathbb{E}[Y'|G=A] = \mathbb{E}'_X \mathbb{E}[Y'|G=A, X'] = \mathbb{E}'_X[\sigma(\beta^T X')|G=A]$$

$$= \mathbb{E}'_X[\sigma(\beta^T X')|G=D] = \mathbb{E}[Y'|G=D]$$

Equality of treatment will follow from the exact same argument applied to the quantities $\hat{Y}'$ and $Y'\hat{Y}'$ (see also example 4.1). □

We now provide the explicit structure of the stratified manifolds in theorem 3.7, and corresponding proofs. For notational convenience, let $\mathrm{Aff}(k)$ (resp. $\mathbb{S}(k)$) be the set of k-dimensional affine subspaces (resp. k-dimensional hyperspheres) of a context-dependent ambient space, and $\mathbb{O}(k)$ the group of $k \times k$ orthogonal matrices.

Theorem C.4 (Theorem 3.7 ex-ante skill discrimination). *Consider the learning setting of Example 3.3. Suppose $c_A = c_D = 1$, the vectors $\{(\mu_{A,m}, -\mu_{D,m}), (\beta_m, -\beta_m)\}$ are not co-linear and the vectors $\{(\mu_{A,u}, -\mu_{D,u}), (\beta_m, -\beta_m)\}$ are not co-linear. Then there exist sets $\mathcal{Z}_m \subset R^{2d_m}$ and $\mathcal{Z}_u \in R^{2d_u}$ of the form:*

$$\mathcal{Z}_m = \cup_{U \in \mathcal{U}_m} Z_m(U);\ Z_m(U) \in \mathit{Aff}(d_m - 2)$$

$$\mathcal{U}_m = \{U \in \mathbb{O}(d_m) : \begin{pmatrix} \beta_m^T, -\beta_m^T \\ \mu_{A,m}^T, -\mu_{D,m}^T \end{pmatrix} \not\perp \mathit{Null}[I_{d_m}, -U^T]\}$$

$$\mathcal{Z}_u = \cup_{U \in \mathcal{U}_u} Z_u(U); Z_u(U) \in \mathit{Aff}(d_u - 2)$$

$$\mathcal{U}_u = \{U \in \mathbb{O}(d_u) : \begin{pmatrix} \beta_u^T, -\beta_u^T \\ -\mu_{A,u}^T, \mu_{D,u}^T \end{pmatrix} \not\perp \mathit{Null}[I_{d_u}, -U^T]\}$$

such that for any learner decision $\theta = (\theta_A, \theta_D)$ which satisfies $(\theta_{A,m}, \theta_{D,m}) \in \mathcal{Z}_m$ and $(\theta_{A,u}, \theta_{D,u}) \in \mathcal{Z}_u$ also satisfies equality of treatment.

Proof. We have that any policy $\theta = (\theta_A, \theta_D)$ that satisfies or $(\theta_g^T X_g', \beta^T X_g') \perp\!\!\!\perp G$ also satisfies equality of treatment. By Lemma C.2 $(\theta_g^T X_g', \beta^T X_g')$ are generated in the manner $\beta^T X_g' = \beta^T(X_g + M\theta_g)$ and $\theta_g^T X_g' = \theta_g^T(X_g + M\theta_g)$. Under the normality assumption in 3.5, the ex-post joint distribution of the (pre-discretized) responses and predictions conditioned on sensitive trait is

$$(\theta_g^T X_g', \beta^T X_g') \sim \mathcal{N}(\tilde{\mu}_g, \tilde{\Sigma}_g)$$

$$\tilde{\mu}_g = (\theta_g^T(\mu_g + M\theta_g), \beta^T(\mu_g + M\theta_g))$$

$$\tilde{\Sigma}_g = \begin{pmatrix} ||\theta_g||^2 & \theta_g^T\beta \\ \theta_g^T\beta & ||\beta||^2 \end{pmatrix}$$

As mentioned, equality of responses plus equality of treatment will be upheld by any policy that satisfies $(\theta_A^T X_A', \beta^T X_A') \overset{d}{=} (\theta_D^T X_D', \beta^T X_D')$. Thus, the equality of treatment constraint set can be studied by setting $\tilde{\mu}_A = \tilde{\mu}_D$ and $\tilde{\Sigma}_A = \tilde{\Sigma}_D$, the resulting constraints are:

$$\begin{aligned} \theta_A^T M\theta_A &= \theta_D^T M\theta_D \\ \theta_A^T M\beta + \beta^T\mu_A &= \theta_D^T M\beta + \beta^T\mu_D \\ \theta_A^T\mu_A &= \theta_D^T\mu_D \\ ||\theta_A||^2 &= ||\theta_D||^2 \\ \theta_A^T\beta &= \theta_D^T\beta \end{aligned} \tag{C.1}$$

The next observation is that we can decompose this feasibility requirement into two, one which pertains to parameters that correspond to "manipulable" features, and another which pertains to "non-manipulable" features. For a general vector v let $v_{m(i)} = v_i 1_{\{M_i=1\}}$ and $v_{u(i)} = v_i 1_{\{M_i=0\}}$. The constraint C.2 will be satisfied by any (θ_A, θ_D) that satisfies both the following:

$$\begin{aligned} ||\theta_{m,A}||^2 &= ||\theta_{m,D}||^2 \\ \theta_{m,A}^T\beta_m - \theta_{m,D}^T\beta_m &= \beta^T\mu_D - \beta^T\mu_A \\ \theta_{m,A}^T\mu_{m,A} &= \theta_{m,D}^T\mu_{m,D} \end{aligned} \tag{C.2}$$

$$\begin{aligned} ||\theta_{u,A}||^2 &= ||\theta_{u,D}||^2 \\ \theta_{u,A}^T\beta_u - \theta_{u,D}^T\beta_u &= -\beta^T\mu_D + \beta^T\mu_A \\ \theta_{u,A}^T\mu_{u,A} &= \theta_{u,D}^T\mu_{u,D} \end{aligned} \tag{C.3}$$

This can be seen by making note of the following identities: $||\theta||_2 = ||\theta_m||^2 + ||\theta_u||^2$, $\theta^T v = \theta_m^T v_m + \theta_u^T v_u$, $\theta^T M\theta = ||\theta_m||^2$, and $\theta^T M v = \theta_m^T v_m$. The forms of constraints C.3, C.4 are nearly identical so we only provide an analysis of the "manipulable" constraints (constraint C.3). On-wards let $2d_m$ be the dimension of the parameter space $(\theta_{m,A}, \theta_{m,D})$

We begin with the quadratic constraint $||\theta_{m,A}||^2 = ||\theta_{m,D}||^2$. The key observation is that this is satisfied if and only if there exists $U \in \mathbb{O}(d_m)$ such that $\theta_{m,A} = U\theta_{m,D}$. Thus, for a fixed $U \in \mathbb{O}(d_m)$ we can write the constraint set as

$$\mathcal{Z}(U) : (\theta_{m,A}, \theta_{m,D}) \text{ s.t. } : \begin{pmatrix} \beta_m & -\beta_m \\ \mu_{m,A} & -\mu_{m,D} \\ I & -U^T \end{pmatrix} \begin{pmatrix} \theta_{m,A} \\ \theta_{m,D} \end{pmatrix} = \begin{pmatrix} b_0 \\ 0 \\ 0_{d_m} \end{pmatrix}.$$

Any choice of matrix U is satisfactory. However, in order to exclude any U that results in $\dim[Z(U)] = 0$, we only select U such that $(\beta_m, -\beta_m)^T \not\perp \mathrm{Null}[I, -U^T]$ and $(\mu_{m,A}, -\mu_{m,D})^T \not\perp \mathrm{Null}[I, -U^T]$ Together, the constraint set is a union of $d_m - 2$ dimensional subspaces:

$$\mathcal{Z} = \cup_{U \in \mathcal{U}} \mathcal{Z}(U)$$

$$\mathcal{U} = \{U \in \mathbb{O}(d_m) : \begin{pmatrix} \beta_m, -\beta_m \\ \mu_{A,m}, -\mu_{D,m} \end{pmatrix} \not\perp \mathrm{Null}[I, -U^T]\}$$

□

Proof of corollary 3.8 (ex-ante skill discrimination). Note that now the joint distribution of $(\theta_g^T X_g', \beta^T S_g')$ satisfies

$$(\theta_g^T X_g', \beta^T S_g') \sim \mathcal{N}(\tilde{\mu}_g, \tilde{\Sigma}_g)$$

$$\tilde{\mu}_g = (\theta_g^T\Lambda(\mu_g + M\Lambda^T\theta_g), \beta^T(\mu_g + M\Lambda^T\theta_g))$$

$$\tilde{\Sigma}_g = \begin{pmatrix} \theta_g^T\Lambda\Lambda^T\theta_g + ||\theta_g||^2 & \beta^T\Lambda^T\theta_g \\ \beta^T\Lambda^T\theta_g & ||\beta||^2 + 1 \end{pmatrix}$$

The constraint $||\theta_A||_2^2 = ||\theta_D||_2^2$ is satisfied for any $U \in \mathbb{O}(q)$ and pair θ such that $\theta_A = U\theta_D$, thus for every U there is a q-dimensional plane that satisfies the constraint $||\theta_A||_2^2 = ||\theta_D||_2^2$. We begin with the initial stratified manifold $\mathcal{Z}' = \cup_{U\in\mathbb{O}_q}\text{Null}[I - U^T]$. By the above, any $\theta \in \mathcal{Z}'$ satisfies $||\theta_A||^2 = ||\theta_D||^2$. Let $V'_U \in \mathcal{Z}$ satisfy $\text{span}\{\begin{pmatrix}\Lambda^T & 0\\ 0 & \Lambda^T\end{pmatrix} v; v \in V'_U\}\} \cong \mathbb{R}^{2d}\}$. Under the transformation $T : \theta \to \tilde{\theta} \in \mathbb{R}^{2d}$; $\tilde{\theta} = (\Lambda^T\theta_A, \Lambda^T\theta_D)$, the remaining constraints (in terms of $\tilde{\theta}$) are identical to those in the proof of 3.7. Thus, by theorem 3.7 there is a manifold $\tilde{\mathcal{M}} \subset \mathbb{R}^{2d}$ such that any $\tilde{\theta} \in \tilde{\mathcal{M}}$ satisfies all remaining constraints necessary for equality of treatment. Thus, any $\theta \in \mathcal{M}(U) \triangleq T^{-1}(\tilde{M}) \cap V'_U$ satisfies equality of treatment, and by the full rank property of Λ restricted to V'_U, $\mathcal{M}(U)$ is a manifold of dimension at least $O(d)$. □

Theorem C.5 (Theorem 3.7 cost discrimination). *Consider the causal strategic classification setting (example 3.3). Suppose $\mu_A = \mu_D = 0$ and $c_A \neq c_D$. There exist sets $\mathcal{Z}_{m,A}$, $\mathcal{Z}_{m,D}$, $\mathcal{Z}_{u,A}$, $\mathcal{Z}_{u,D}$ of the form:*

$$\mathcal{Z}_{m,A} = \cup_{(k_1,k_2)\in\mathcal{K}} \mathcal{S}^{A,m}_{(k_1,k_2)}; \mathcal{S}^{A,m}_{(k_1,k_2)} \in \mathbb{S}(d_m - 2)$$
$$\mathcal{Z}_{D,m} = \cup_{(k_1,k_2)\in\mathcal{K}} \mathcal{S}^{D,m}_{(k_1,k_2)}; \mathcal{S}^{U,m}_{(k_1,k_2)} \in \mathbb{S}(d_m - 2)$$
$$\mathcal{Z}_{A,u} = \cup_{(k_1,k_2)\in\mathcal{K}} \mathcal{S}^{A,u}_{(k_1,k_2)}; \mathcal{S}^{A,u}_{(k_1,k_2)} \in \mathbb{S}(d_u - 2)$$
$$\mathcal{Z}_{D,u} = \cup_{(k_1,k_2)\in\mathcal{K}} \mathcal{S}^{D,u}_{(k_1,k_2)}; \mathcal{S}^{D,u}_{(k_1,k_2)} \in \mathbb{S}(d_u - 2)$$

$$\mathcal{K} = \{(k_1,k_2) \in \mathbb{R}^+ \times \mathbb{R} : k_1 > \frac{\sqrt{\frac{c_A}{c_D}}|k_2|}{min(||\beta_u||, ||\beta_m||)}\}$$

such that for any policy $\theta = (\theta_{A,m}, \theta_{A,u}, \theta_{D,m}, \theta_{D,u})$ which satisfies $\theta \in \mathcal{Z}_{m,A} \times \mathcal{Z}_{m,D} \times \mathcal{Z}_{u,A} \times \mathcal{Z}_{u,D}$ also satisfies ex-post *equality.*

Proof. By lemma C.2 $(\theta_g^T X'_g, \beta^T X'_g)$ are generated in the manner $\beta^T X'_g = \beta^T(X_g + \frac{1}{c_g} M\theta_g)$ and $\theta_g^T X'_g = \theta_g^T(X_g + \frac{1}{c_g} M\theta_g + \epsilon)$.

Under the normality assumption in 3.5, the ex-post joint distribution of the (pre-discretized) outcomes and predictions conditioned on sensitive trait is

$$(\theta_g^T X'_g, \beta^T X'_g) \sim \mathcal{N}(\tilde{\mu}_g, \tilde{\Sigma}_g)$$
$$\tilde{\mu}_g = (\frac{1}{c_g}\theta_g^T M\theta_g, \frac{1}{c_g}\beta^T M\theta_g)$$
$$\tilde{\Sigma}_g = \begin{pmatrix} ||\theta_g||^2 & \theta_g^T\beta \\ \theta_g^T\beta & ||\beta||^2 \end{pmatrix}$$

By the above lemma equality of treatments will be satisfied by any policy that satisfies $(\theta_A^T X'_A, \beta^T X'_A) \overset{d}{=} (\theta_D^T X'_D, \beta^T X'_D)$. Thus, using a nearly identical argument as the proof for the first part of theorem C.5 the equality of treatments constraint set can be studied by setting $\tilde{\mu}_A = \tilde{\mu}_D$ and $\tilde{\Sigma}_A = \tilde{\Sigma}_D$, the resulting constraints are:

$$\begin{aligned} \frac{1}{c_A}\theta_A^T M\theta_A &= \frac{1}{c_D}\theta_D^T M\theta_D \\ \frac{1}{c_A}\theta_A^T M\beta &= \frac{1}{c_D}\theta_D^T M\beta \\ ||\theta_A||^2 &= ||\theta_D||^2 \\ \theta_A^T\beta &= \theta_D^T\beta \end{aligned} \tag{C.4}$$

Again, we can decompose this feasibility requirement into four constraints, each which pertains to parameters that correspond to "manipulable"/"non-manipulable" and a sensitive trait value . The constraint C.5 will be satisfied by any (θ_A, θ_D) that satisfies all four the following for any pair (k_1, k_2):

$$\begin{aligned} ||\theta_{m,A}||^2 &= \frac{c_A}{c_D}k_1^2 \\ \theta_{m,A}^T\beta_m &= \frac{c_A}{c_D}k_2 \end{aligned} \tag{C.5}$$

$$\begin{aligned} ||\theta_{m,D}||^2 &= k_1^2 \\ \theta_{m,D}^T\beta_m &= k_2 \end{aligned} \tag{C.6}$$

$$\begin{aligned} ||\theta_{u,A}||^2 &= k_1^2 \\ \theta_{u,A}^T\beta_u &= k_2 \end{aligned} \tag{C.7}$$

$$\begin{aligned} ||\theta_{u,D}||^2 &= \frac{c_A}{c_D}k_1^2 \\ \theta_{u,D}^T\beta_u &= \frac{c_A}{c_D}k_2 \end{aligned} \tag{C.8}$$

We have again used each of the following identities $||\theta||_2 = ||\theta_m||^2 + ||\theta_u||^2$, $\theta^T v = \theta_m^T v_m + \theta_u^T v_u$, $\theta^T M\theta = ||\theta_m||^2$, and $\theta^T Mv = \theta_m^T v_m$. Consider the first constraint set C.5; the constraint $||\theta_{m,A}||^2 = \frac{c_A}{c_D}k_1^2$ is a hypersphere (of dimension $d_{m,A} - 1$) with radius $\frac{c_A}{c_D}k_1^2$. The constraint $\theta_{m,A}^T\beta_m = \frac{c_A}{c_D}k_2 \iff \theta_{m,A}^T\frac{\beta_m}{||\beta_m||} = \frac{c_A}{c_D}\frac{k_2}{||\beta_m||}$ is a hyperplane. It is easy to see that the intersection between these two geometric objects is either empty, (if $|\frac{c_A}{c_D}\frac{k_2}{||\beta_m||}| > \sqrt{\frac{c_A}{c_D}k_1^2}$) or is a hypersphere of dimension $d_{m,A} - 2$ (if $|\frac{c_A}{c_D}\frac{k_2}{||\beta_m||}| < \sqrt{\frac{c_A}{c_D}k_1^2}$). Note that an identical argument can be applied to each of the other constraint sets, C.6, C.7, C.8. Each of these sets will be non-empy so long as $|k_1| > \frac{\sqrt{\frac{c_A}{c_D}}k_2}{\min(||\beta_u||, ||\beta_m||)}$. □

Proof of corollary 3.8 (cost discrimination). Note that now the joint distribution of $(\theta_g^T X'_g, \beta^T S'_g)$ satisfies

$$(\theta_g^T X'_g, \beta^T S'_g) \sim \mathcal{N}(\tilde{\mu}_g, \tilde{\Sigma}_g)$$
$$\tilde{\mu}_g = (\theta_g^T\Lambda(\frac{1}{c_g}M\Lambda^T\theta_g), \beta^T(\frac{1}{c_g}M\Lambda^T\theta_g))$$
$$\tilde{\Sigma}_g = \begin{pmatrix} \theta_g^T\Lambda\Lambda^T\theta_g + ||\theta_g||^2 & \beta^T\Lambda^T\theta_g \\ \beta^T\Lambda^T\theta_g & ||\beta||^2 + 1 \end{pmatrix}$$

The constraint $||\theta_A||_2^2 = ||\theta_D||_2^2$ is satisfied for any $U \in \mathbb{O}(q)$ and pair θ such that $\theta_A = U\theta_D$, thus for every U there is a q-dimensional plane that satisfies the constraint $||\theta_A||_2^2 = ||\theta_D||_2^2$. We begin with the initial stratified manifold $\mathcal{Z}' = \cup_{U\in\mathbb{O}_q}\text{Null}[I - U^T]$, using the above, any $\theta \in \mathcal{Z}'$ satisfies $||\theta_A||^2 = ||\theta_D||^2$. Let $V'_U \in \mathcal{Z}$ satisfy $\text{span}\{\begin{pmatrix}\Lambda^T & 0\\ 0 & \Lambda^T\end{pmatrix} v; v \in V'_U\}\} \cong \mathbb{R}^{2d}\}$. Under the transformation $T : \theta \to \tilde{\theta} \in \mathbb{R}^{2d}$; $\tilde{\theta} = (\Lambda^T\theta_A, \Lambda^T\theta_D)$, the remaining constraints (in terms of $\tilde{\theta}$) are identical to those in the proof of C.5. Thus, by theorem $C.5$ there is a manifold $\tilde{\mathcal{M}} \subset \mathbb{R}^{2d}$ such that any $\tilde{\theta} \in \tilde{\mathcal{M}}$ satisfies all remaining constraints necessary for equality of treatment. Thus, any $\theta \in \mathcal{M}(U) \triangleq T^{-1}(\tilde{M}) \cap V'_U$ satisfies equality

of treatment, and by the full rank property of Λ restricted to V'_U, $\mathcal{M}(U)$ is a manifold of dimension at least $O(d)$. □

C.3 Proof of impossibility results

Proof of theorem 3.2: Under the unbiasedness assumption on test signals, the optimal hiring policy for the firm will be of the form $f(x,g) = \mathbf{1}\{x \geq \theta_g\}$. Consider a firm that uses a group blind threshold policies of the form $f(x,g) = \mathbf{1}\{x \geq \theta\}$ to hire workers, where $\theta \in \mathbb{R}$ is the threshold value for both groups. The workers' expected utility for increasing skills from s to s^* is

$$u_w(f, s, s^*, g) = w\bar{F}(\theta \mid s^*) - c(s, s^*), \tag{C.9}$$

where $\bar{F}(\theta \mid s^*) \triangleq \mathbb{P}\{X > \theta \mid S = s^*\}$ is the survival function of the skill level assessment.

In the labor market example, the worker (in group g) response

$$s'(s, fg) \triangleq \arg\max_{s^*} u_w(f, s, s^*, g)$$

is generally non-decreasing. For example, suppose $c_A = c_D = c$ and $c(s, s^*) = \frac{c}{2}[s^* - s]_+^2$, where $[\,\cdot\,]_+ \triangleq \max\{0, \cdot\}$ is the ReLU function, then the derivative of the worker response is

$$\frac{\partial s'(s,f,g)}{\partial s} = -\frac{\partial_s \partial_{s^*} c(y, s^*)|_{s^*=s'(s,f,g)}}{\partial^2_{s^*} u_w(f, y, s^*)|_{s^*=s'(s,f,g)}} = \frac{c - w\partial^2_y \bar{F}(t \mid s'(s,f,g))}{c},$$

so $\frac{\partial s'(s,f,g)}{\partial s} > 0$ as long as c is large enough. In general, We are unconcerned with settings in which $s'(s,f,g)$ is not non-decreasing in s this would be both unintuitive and unrealistic. Under the assumption that $S|A$ stochastically dominates $S|D$, the non-decreasingness of $s'(s,f,g)$ in s implies $S'|A$ is stochastically dominates $S'|D$, which precludes equal *ex post* responses (in particular $\mathbb{E}[Y'|A] > \mathbb{E}[Y'|D]$).

On the other hand, suppose that $S|A \overset{d}{=} S|D$ but $c_A < c_D$, keeping the assumption that policies are group blind. Note that

$$\frac{\partial}{\partial c} s'(s,f,c) = \frac{w\partial^2_{s^*}\bar{F}(t|s^*)|_{s^*=s'(s,f,c)} - c}{[s'(s,f,c) - s]_+}$$

This implies that if c_A, c_D are large enough with $c_A < c_D$ two workers from each group with $s_A = s_D$ will have $s'(s,f,c_A) > s'(s,f,c_D)$ implying $S'|A$ stochastically dominates $S'|D$, which precludes equal *ex post* responses.

Finally, the fact that only group blind policies will satisfy equality of outcomes with respect to $m_{\text{FPR}}(f,g)$ and $m_{\text{FNR}}(f,g)$ follows immediately from the market assumption that X' is independent of G given Y'. □

D APPENDIX A PROOFS

D.1 Proof of Proposition A.1

Proof. We prove the statement on $\mu(Q)$, the proof for $\text{EPR}(Q)$ is identical. We have:

$$\mu_{ij}(Q) = \mathbb{E}_{(X',Y'),Q,\mathcal{G}}[h_i(Q(X'), Y')|\mathcal{G} = g_j]$$

By the law of total (conditional) expectation this is equivalent to

$$\mu_{ij}(Q) = \sum_{f\in\mathcal{F}} P(Q=f)\mathbb{E}_{(X',Y'),Q}[h_i(Q(X'), Y')|G = g_j, Q = f]P(Q = f|G = g_j)$$

By assumption $P(Q = f|G = g_j) = P(Q = f)$. Note also that $\mathbb{E}_{(X',Y')\sim\mathcal{D}(Q),Q}[h_i(Q(X'), Y')|G = g_j, Q = f] = \mathbb{E}_{(X',Y')\sim\mathcal{D}(f)}[h_i(f(X'), Y')|\mathcal{G} = g_j] = \mu_{ij}(f)$. Thus

$$\mu_{ij}(Q) = \sum_{f\in\mathcal{F}} P(Q=f)\mu_{ij}(f) = \sum_{f\in\mathcal{F}} P(Q=f)\mu_{ij}(f) = \sum_{f\in\mathcal{F}} Q(f)\mu_{ij}(f)$$

□

D.2 Proof of Generalization Error

Proposition D.1 (Agarwal et al. [1] Lemma 3). *Suppose that the constraint $M\hat{\mu}(Q) \leq \hat{c}$ is feasible. Then if $\hat{Q}$ is a ϵ-saddle point of equation 4.2 the following holds:*

$$||M\hat{\mu}(\hat{Q}) - \hat{c}||_\infty \leq \frac{1+2\epsilon}{B}$$

Proposition D.2 (Agarwal et al. [1] Lemma 2). *If $\hat{Q}$ is an ϵ-saddle point, then for all Q such that $M\hat{\mu}(Q) \leq c$ the distribution $\hat{Q}$ satisfies*

$$\widehat{EPR}(\hat{Q}) \leq \widehat{EPR}(Q) + 2\epsilon.$$

The technical tool we use to study the generalization properties of algorithm 1 is the Rademacher complexity, which we now define.

Definition D.3. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [0,1]$ and ϵ_i be i.i.d. Rademacher random variables. the Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathcal{R}_n(\mathcal{F}) \triangleq \sup_{x_1,\ldots,x_n\in\mathcal{X}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)|$$

The primary obstacle to proving a generalization bound will be to establish bounds on the Rademacher complexity of the function classes $\mu_i(\theta), \hat{\text{EPR}}(\theta)$. Beyond this the analysis is a standard application of the arguments in [1].

Lemma D.4. *Let $\mathcal{X} \triangleq \{x \in \mathbb{R}^d; ||x|| \leq C\}$, let $\mathcal{H}_\theta = \{\theta^T x; x \in \mathcal{X}, \theta \in \Theta\}$.*

$$\hat{\mu}_1(\theta) = \sum_{i=1}^n \sigma(\theta^T(x_i + Ma(\theta))),$$

$$\hat{\mu}_2 = \sum_{i=1}^n \sigma(\beta^T(x_i + Ma(\theta))),$$

$$\hat{\mu}_3 = \sum_{i=1}^n \sigma(\theta^T(x_i + Ma(\theta)))\sigma(\beta^T(x_i + Ma(\theta))).$$

Then the following hold:

(1) *With probability at least $1-\delta$ for all $\theta \in \Theta$,*

$$|\hat{\mu_1}(\theta) - \mu_1(\theta)| \leq 2R_n(\mathcal{H}_\theta) + 2\sup_{\theta\in\Theta}|\theta^T Ma(\theta)|/(\sqrt{n}) + \sqrt{\frac{ln(2/\delta)}{2n}} \sim \tilde{O}(n^{-1/2})$$

(2) *With probability at least $1-\delta$ for all $\theta \in \Theta$*

$$|\hat{\mu_2}(\theta) - \mu_2(\theta)| \leq 2||\beta||C/\sqrt{n} + 2\sup_{\theta\in\Theta}|\beta^T Ma(\theta)|/\sqrt{n} + \sqrt{\frac{ln(2/\delta)}{2n}} \sim \tilde{O}(n^{-1/2})$$

(3) *With probability at least $1-\delta$ for all $\theta \in \Theta$*

$$|\hat{\mu_3}(\theta) - \mu_3(\theta)| \leq$$
$$2\sqrt{2}||\beta||C/\sqrt{n} + 2\sqrt{2}\sup_{\theta\in\Theta}|\beta^T Ma(\theta)|/\sqrt{n} + 2\sqrt{2}R_n(\mathcal{H}_\theta)$$

$$+2\sqrt{2}\sup_{\theta\in\Theta}|\theta^T Ma(\theta)|/(\sqrt{n})+\sqrt{\frac{ln(2/\delta)}{2n}}\sim\tilde{O}(n^{-1/2})$$

Proof. (1) Let $\mathcal{F}_\theta$ be the class of functions $f_\theta : s \to \sigma(\theta^T(x+Ma(\theta))) \in [0,1]$. By a standard concentration inequality, with probability at least $1-\delta$ for all $\theta\in\Theta$,

$$|\hat{\mu_1}(\theta)-\mu_1(\theta)|\le 2R_n(\mathcal{F}_\theta)+\sqrt{\frac{ln(2/\delta)}{2n}}$$

Thus it remains to find an appropriate bound for $R_n(\mathcal{F}_\theta)$. Note that

$$R_n(\mathcal{F}_\theta)=\sup_{x_1,\dots,x_n\in\mathcal{X}}\mathbb{E}_{\epsilon_i}[\sup_{\theta\in\Theta}|\frac{1}{n}\sum_{i=1}^n\epsilon_i\sigma(\theta^T(x_i+Ma(\theta)))|]$$

$$\overset{\text{Talegrands}}{\le}\sup_{x_1,\dots,x_n\in\mathcal{X}}\mathbb{E}_{\epsilon_i}[\sup_{\theta\in\Theta}|\frac{1}{n}\sum_{i=1}^n\epsilon_i\theta^T x_i+\epsilon_i\theta^T Ma(\theta)|]$$

$$\le\sup_{x_1,\dots,x_n\in\mathcal{X}}\mathbb{E}_{\epsilon_i}[\sup_{\theta\in\Theta}|\frac{1}{n}\sum_{i=1}^n\epsilon_i\theta^T x_i|]+\mathbb{E}_{\epsilon_i}[\sup_{\theta\in\Theta}|\frac{1}{n}\epsilon_i\theta^T Ma(\theta)|]$$

$$\le R_n(\mathcal{H}_\theta)+\sup_{\theta\in\Theta}|\theta^T Ma(\theta)|/\sqrt{n}$$

(2) By an identical argument to the case of μ_1 we have that with probability at least $1-\delta$ for all $\theta\in\Theta$ we have

$$|\hat{\mu_2}(\theta)-\mu_2(\theta)|\le\sup_{x_1,\dots,x_n\in\mathcal{X}}\mathbb{E}_\epsilon|\frac{1}{n}\sum_{i=1}^n\epsilon_i\beta^T x_i|+\sup_{\theta\in\Theta}|\beta^T Ma(\theta)|/\sqrt{n}$$

The first term is trivially bounded above by the empirical Rademacher complexity of the function class $\{w^T x; x\in\mathcal{X}, ||w||_2\le||\beta||_2\}$ which is bounded above by $||\beta||C/\sqrt{n}$

(3) Let $\mathcal{B}_\theta$ be the function class $b_\theta : x\to\sigma(\beta^T(x+Ma(\theta)))\in[0,1]$, and let Π_θ be the function class $\pi_\theta : x\to\sigma(\beta^T(x+Ma(\theta)))\sigma(\theta^T(x+Ma(\theta)))\in[0,1]$. Note that this second function class can be thought of as the composition of the (1-Lipschitz on $[0,1]\times[0,1]$) function $\psi(x,y)=xy$ and the vector valued function $v(x)=(\sigma(\beta^T(x+Ma(\theta))),\sigma(\theta^T(x+Ma(\theta))))$. Thus by corollary 4 of [44] we have

$$\mathcal{R}_n(\Pi_\theta)\le\sqrt{2}[\mathcal{R}_n(\mathcal{B}_\theta)+\mathcal{R}_n(\mathcal{F}_\theta)]$$

From here we simply plug in the upper bounds for $\mathcal{R}_n(\mathcal{B}_\theta)$ and $\mathcal{R}_n(\mathcal{F}_\theta)$ attained in part 1 and 2 and the standard concentration inequality used throughout.

The fact that each quantity (1,2,3) is $\sim\tilde{O}(n^{-1/2})$ follows from the well-known result that $\mathcal{R}_n(\mathcal{H}_\theta)\sim O(n^{-1/2})$ if x and θ are bounded. □

Proof of theorem A.4.

Note that $||M\mu(\hat{Q})||_\infty\le||M(\hat{\mu}(\hat{Q})-\mu(\hat{Q}))||_\infty+||M\hat{\mu}(\hat{Q})||_\infty$. By proposition D.1 (and choice of ν) it holds that $||M\hat{\mu}(\hat{Q})||_\infty\le\frac{1+2\epsilon}{B}$. By the form of M, $||M(\hat{\mu}(\hat{Q})-\mu(\hat{Q}))||_\infty\le 2||\hat{\mu}(\hat{Q})-\mu(\hat{Q})||_\infty$. Then by lemma D.4 and a union bound, with probability at least $1-6\delta$, it holds that $||\hat{\mu}(\hat{Q})-\mu(\hat{Q})||_\infty\le O(\min(n_A,n_D)^{-1/2})+8\sqrt{\frac{ln(2/\delta)}{2\min(n_A,n_D)}}$

Additionally, by our choice of ν, $||M(\hat{\mu}(Q^*))||\le\hat{c}$, so by proposition D.2 $\widehat{\text{EPR}}(\hat{Q})\le\widehat{\text{EPR}}(Q^*)+2\epsilon$. By the argument of part 3 of lemma D.4 (and the additive nature of Rademacher complexity), with probability at least $1-\delta$

$$|\widehat{\text{EPR}}(\hat{Q})-\text{EPR}(\hat{Q})|\le O(\min(n_A,n_D)^{-1/2})+\sqrt{\frac{ln(2/\delta)}{2\min(n_A,n_D)}}$$

$$|\widehat{\text{EPR}}(Q^*)-\text{EPR}(Q^*)|\le O(\min(n_A,n_D)^{-1/2})+\sqrt{\frac{ln(2/\delta)}{2\min(n_A,n_D)}}$$

Thus with probability at least $1-\delta$ it holds that $\text{EPR}(\hat{Q})\le\text{EPR}(Q^*)+2\epsilon+O(\min(n_A,n_D)^{-1/2})+2\sqrt{\frac{ln(2/\delta)}{2\min(n_A,n_D)}}$. A final union bound completes the proof. □

E EXPERIMENTS

E.1 Experimental Details

All experiments were done on Google Colab using only a CPU. The chosen parameters for data generation are $\beta=1_{10}$, $\mu_A=0.5*1_{10}$, $\mu_D=0.1*1_{10}$, $c_A=\frac{4}{2}||a||_2^2$, $c_D=\frac{10}{2}||a||_2^2$, $\Sigma=I_{10\times10}$, $\Lambda=I_{10\times10}$ with *ex-ante* skills generated from $\mathcal{N}(\mu_g,\Sigma)$ distributions. Each base line was also implemented using a reduction method. Step size η_t was selected according to the convergence theory, B was selected to achieve at least 10^{-5} fairness violation on the training set. Training and test sizes of 500 samples were used, with a random seed of 0 for the training set and a random seed of 1 for the test set.

E.2 Additional Experiments

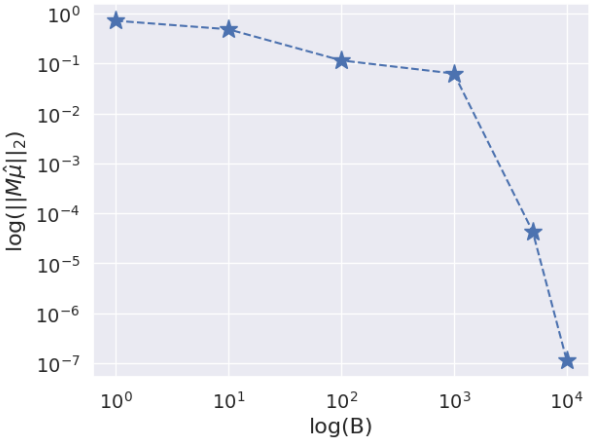


Figure 3: Equality of treatment + Equality of responses on training set