The Disparate Effects of Partial Information in Bayesian Strategic Learning

Srikanth Avasarala[1], **Serena Wang**[2], **Juba Ziani**[1]

[1]Georgia Institute of Technology
[2]Harvard University
savasarala9@gatech.edu, serenalwang@g.harvard.edu, jziani3@gatech.edu

Abstract

We study how partial information about scoring rules affects fairness in strategic learning settings. In strategic learning, a learner deploys a scoring rule, and agents respond strategically by modifying their features—at some cost—to improve their outcomes. However, in our work, agents do not observe the scoring rule directly; instead, they receive a *noisy signal* of said rule. We consider two different agent models: (i) *naive* agents, who take the noisy signal at face value, and (ii) *Bayesian* agents, who update a prior belief based on the signal.

Our goal is to understand how disparities in outcomes arise between groups that differ in their costs of feature modification, and how these disparities vary with the *level of transparency* of the learner's rule. For naive agents, we show that utility disparities can grow unboundedly with noise, and that the group with lower costs can, perhaps counter-intuitively, be disproportionately harmed under limited transparency. In contrast, for Bayesian agents, disparities remain bounded. We provide a full characterization of disparities across groups as a function of the level of transparency and show that they can vary *non-monotonically* with noise; in particular, disparities are often minimized at *intermediate* levels of transparency. Finally, we extend our analysis to settings where groups differ not only in cost, but also in prior beliefs, and study how this asymmetry influences fairness.

1 Introduction

Machine learning systems are increasingly being deployed in high-stakes domains such as college admissions, lending, and hiring. A common assumption in the design of these models is that the data encountered at test time is drawn from the same distribution as the training data. However, this assumption breaks down when individuals *strategically* adapt their features in response to the deployed model. In real life, individuals may invest in test preparation, tune their resumes to match job descriptions and automated CV reviewing algorithms, or open dummy credit accounts to artificially improve metrics like credit usage—all in an effort to improve their predicted outcomes. This phenomenon, known as *strategic learning*, was first formalized by Hardt et al. (2016), and has become central to understanding robustness algorithmic decision-making.

While such strategic behavior may be rational from an individual perspective, it can have undesirable implications for society and in particular for fairness. When individuals differ in the cost of modifying their features—due to resource constraints, institutional barriers, or historical disadvantage—strategic adaptation can reinforce or amplify existing inequalities. Prior work has shown that disparities in the cost of feature manipulation can translate into disparities in both scoring and classification outcomes even when the underlying model treats all individuals equally (Milli et al. 2019; Hu, Immorlica, and Vaughan 2019).

A common, key assumption in much of the strategic classification literature is that individuals have *full knowledge* of the deployed model and can best respond accordingly. Yet this assumption is often unrealistic. In practice, scoring rules are rarely transparent. Banks and credit bureaus do not disclose the exact algorithms behind loan decisions or credit scores; recidivism prediction tools such as COMPAS operate as proprietary black boxes (Angwin et al. 2016); and AI-based hiring platforms use complex and often non-interpretable models to rank candidates. Therefore, addressing fairness in strategic learning environments also requires thinking about how individuals perceive deployed rules and algorithms, both when all agents have access to the same information about the deployed rule and when there are informational disparities across individuals.

These observations raise a critical question: how does *partial knowledge* of the model affect the fairness of strategic learning systems? More precisely, how do different levels of transparency—ranging from full disclosure to total opacity—influence disparities in outcomes across populations? And how do these effects depend on the assumptions we make about how individuals operate under uncertainty? In this work, we develop a framework to study fairness under partial information in strategic learning environments. We model agents who observe only a *noisy signal* of the deployed decision rule. We consider two types of agent responses to this signal: (i) *naive* agents, who treat the signal as ground truth and optimize against it directly (Jagadeesan, Mendler-Dünner, and Hardt 2021); and (ii) *Bayesian* agents, who combine the signal with a prior belief to form a posterior, and optimize based on this posterior belief on the deployed rule (Cohen et al. 2024). Our analysis focuses on how outcome disparities—measured both in terms of

model scores and individual utilities (that also take costs into account)——arise across groups that differ in their cost of feature modification. We examine how these disparities evolve as a function of the amount of information the learner releases about the model.

Summary of Contributions. Our paper makes the following contributions:

- In Section 4.1, we analyze the fairness of **naive agents** under partial transparency. We show that while *score* disparities remain constant, *utility* disparities vary monotonically with the level of noise in the model signal. Surprisingly, the group with lower costs of feature change can be disproportionately harmed by noisy information, leading to unbounded disparities in utility, due to "overspending" on ineffective modifications.
- In Section 4.2, we turn to **Bayesian agents** and fully characterize how score and utility disparities depend on the prior and noise level of the released signal. We characterize when disparities arise as a function of the parameter of the problem. Perhaps surprisingly, we show that disparities can be *non-monotone* in the level of information revealed by the learner and are minimized at intermediate transparency levels.
- In Section 5, we characterize disparities when groups differ not only in cost but also in their **prior beliefs**, expanding the model of Bechavod et al. (2022). We derive bounds on group disparities as a function of the *information overlap* of Bechavod et al. (2022).

Related work This work lies at the intersection of fairness and strategic behavior in algorithmic decision-making. A growing body of research has studied how individuals adapt their features to secure better outcomes from predictive models—known as *Strategic Classification* or *Strategic Learning*—, and how such behavior leads to disparate impacts across populations.

Strategic Learning was initially introduced by Hardt et al. (2016) and sparked a large area of research studying how agents respond to decision rules in learning systems (Braverman and Garg 2020; Dong et al. 2018; Zhang et al. 2022; Lechner, Urner, and Ben-David 2023; Chen, Liu, and Podimata 2020; Ahmadi et al. 2021; Sundaram et al. 2023). While early work primarily focused on agents "gaming" classifiers—in the context of loans, this could be seen as opening dummy credit cards to artificially lower credit utilization and inflate credit scores—, more recent papers have considered actual improvements—e.g., improving one's ability to repay loans on time—, as seen in the works of Kleinberg and Raghavan (2020); Shavit, Edelman, and Axelrod (2020); Harris, Heidari, and Wu (2021); Bechavod et al. (2022); Ebrahimi, Vaccaro, and Naghizadeh (2024).

A significantly smaller subset of this literature explicitly tackles fairness. Notably, Milli et al. (2019) and Hu, Immorlica, and Vaughan (2019) show that unequal ability to manipulate features (in the form in unequal feature manipulation costs) can lead to unfair outcomes. Other works examine disparities arising from population-level variation in feature distributions or strategic behavior itself (Jung et al. 2020; Liu et al. 2020; Estornell et al. 2023; Somerstep, Ritov, and Sun 2024). Most recently, Liu and Sun (2025) propose a dynamic pricing algorithm that ensures fairness by constraining strategic group misreporting.

Our work builds directly on recent advances in modeling strategic classification under partial information. The noisy response model introduced by Jagadeesan, Mendler-Dünner, and Hardt (2021) assumes agents do not see the exact classifier—rather, they see a noisy version of the deployed classifier, and "naively" best respond to this noisy signal. We build on the model of Jagadeesan, Mendler-Dünner, and Hardt (2021) by considering the disparate impacts of information revelation when such naive agents have different costs of modifying their features. Furthermore, we additionally consider a Bayesian model of agent behavior, where agents form and update beliefs about the decision rule based on observed information—rather than taking a noisy and potentially inaccurate signal at face value. This approach directly follows the recent work of Cohen et al. (2024), who first introduce Bayesian agents in the context of strategic classification. Our main departure from the work of Cohen et al. (2024) is that we study Bayesian agents strategic classification from a *fairness* perspective, and aim to understand how the information revealed to agents affects group-level disparities.

Our model is perhaps most closely related to Bechavod et al. (2022), who also consider agents without access to the true model and study resulting group outcomes. However, we note two key differences:

- *(i) Modeling of partial information:* they assume agents infer the model from peer samples, whereas we allow agents to observe noisy outputs of the model directly. Further, agents in Bechavod et al. (2022) are not Bayesian, and instead compute the deterministic "most reasonable" model based on the observed samples;
- *(ii) Role of the learner:* in our framework, the learner can control how much information is revealed via a tunable noise parameter, whereas in their work the learner has no direct influence over agent beliefs.

Finally, our approach connects with recent work on strategic classification with causal and informational structure (Ahmadi et al. 2022; Horowitz and Rosenfeld 2023; Efthymiou et al. 2025). While Efthymiou et al. (2025) study optimal effort allocation under uncertain causal graphs and classifiers, their focus is primarily on incentivizing agents to modify features in "desirable" ways. Our work differs by directly analyzing how information policies shape fairness under incomplete information.

2 Model and Preliminaries

We study a Stackelberg game between a *learner* (or principal) and a population of *agents* where each agent belongs to one of m sub-populations (or "groups"). We focus on the case of $m = 2$ in this work[1].

[1]Groups are statistically independent of each other, so all insights for $m = 2$ can be generalized to any m

Let the sub-populations (or groups) be denoted g_1 and g_2. Each agent i has a feature vector $x_i \in \mathbb{R}^d$, and is assigned a label $y_i \in \mathbb{R}$ by the learner. We assume[2] that the learner assigns scores *linearly*, i.e. that their score is given by

$$y_i = x_i^\top \theta_*,$$

where $\theta_* \in \mathbb{R}^d$ is the rule deployed by the learner.

However, each agent observes only partial information about the deployed model θ_*, in the form of a noisy signal about said model. Formally, and as in (Jagadeesan, Mendler-Dünner, and Hardt 2021), each agent observes a signal S of the model θ_* corrupted by an additive noise, given by:

$$S = \theta_* + \sigma Z. \tag{1}$$

where $\sigma \in \mathbb{R}_{\geq 0}$ is the parameter controlling the amount of variance in the noise, and Z is an independent sample from a standard Gaussian—i.e., we assume that $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. This signal can be interpreted as the information that the *learner* decided to reveal about their classifier; the lower the value of σ, the more information they reveal to agents. For example, credit scoring companies reveal *partial* information about their models (i.e., the weight they put on each feature, or which features they consider, but not how these features are computed). Our goal is to understand *how the level of information revelation by the learner affects disparities across populations* under varying i) costs and ii) initial information about the learner's model.

Principal-Agent interaction Agents first observe the signal S then strategically change their features to try to improve their score. For a given agent, we denote x their original features and x' their modified features. We also write $\Delta x \triangleq x' - x$ for simplicity of notation. To do so, they first build a posterior belief on θ_* based on their signal S, denoted by $\theta \sim \pi^S$. Given modified feature vector $x' \in \mathbb{R}^d$, an agent's utility function for moving from x to x' is given by

$$u(x, x'; g) = \text{score}(x'; g) - c(x, x'; g), \tag{2}$$

where

$$\text{score}(x'; g) = \mathbb{E}_{\theta \sim \pi_g^S}\left[x'^\top \theta\right]$$

is the expected score the agent gets under posterior π_g^S, and

$$c(x, x'; g) = \frac{1}{2}(x' - x)^\top A_g (x' - x)$$

is a Mahalanobis-distance-based cost function[3] for changing agent features from x to x'. The cost function is parametrized by $A_g \in \mathbb{R}^{d \times d}$, which is called the cost matrix for group g and is assumed to be positive definite (PD)[4].

We now consider two principled ways that agents interpret the signal S: (i) a *naive* agent that takes the signal S at face value, i.e., has the posterior belief $\hat{\theta} = S$, as in (Jagadeesan, Mendler-Dünner, and Hardt 2021); (ii) a *Bayesian* agent that computes their *posterior belief* on $\hat{\theta}$ based on both a prior belief π about the model and the signal S; this is the model of Cohen et al. (2024). Formally:

- (i) *Naive agent:* In the naive case, the agent solves the following optimization problem:

$$\begin{aligned}\hat{x}_g &= \arg\max_{x'} \quad u(x, x'; g) \\ &= \arg\max_{x'} \quad S^\top x' - \frac{1}{2}(x' - x)^\top A_g (x' - x).\end{aligned} \tag{3}$$

- (ii) *Bayesian agent:* In this case, an agent in group g has a prior distributional belief π on the learner's model, encoding their initial knowledge of the model. After observing realization s of signal S, the agent updates their prior in a Bayesian fashion to a new posterior belief given by $\pi_g^S(\theta) = \mathbb{P}_{\theta \sim \pi}\left[\hat{\theta} = \theta | S\right]$. This posterior is a distributional belief on the learner's model, to be interpreted as the perceived chance that the model is θ. The individual then changes their features from x to x' to maximize the utility conditioned on the observed signal in the following way

$$\begin{aligned}\hat{x}_g &= \arg\max_{x'} \quad \mathbb{E}_{\theta \sim \pi_g^S}\ u(x, x'; g | S) \\ &= \arg\max_{x'} \quad \mathbb{E}_{\theta \sim \pi_g^S}\left[\theta^\top x' | S\right] \\ &\quad - \frac{1}{2}(x' - x)^\top A_g (x' - x).\end{aligned} \tag{4}$$

  For the purpose of our analysis, we consider the same amount of information revealed across both groups. For a *Bayesian* agent, we assume the agent prior in a group to be $\pi_g \sim \mathcal{N}(\omega_g, \gamma^2 \mathbf{I}_d)$, where γ^2 is the variance of each component of the prior distribution vector.

Fairness metrics The main goal of the paper is to understand disparities across groups when it comes to strategically fitting the deployed decision rule. To do so, we now define two different measures of disparity that we use to quantify fairness between two groups in equilibrium, i.e. after agents in both the groups change their features.

1. *Score disparity:* The *score disparity* $\mathcal{F}_s$ is defined as

$$\begin{aligned}\mathcal{F}_s &= \mathbb{E}[\Delta \text{ score}_1] - \mathbb{E}[\Delta \text{ score}_2] \\ &= \mathbb{E}[{\theta_*}^\top \Delta x_1] - \mathbb{E}[{\theta_*}^\top \Delta x_2].\end{aligned} \tag{5}$$

2. *utility disparity:* The utility disparity $\mathcal{F}_u$ is defined as

$$\begin{aligned}\mathcal{F}_u &= \mathbb{E}[\Delta u_1] - \mathbb{E}[\Delta u_2] \\ &= \mathbb{E}[{\theta_*}^\top \Delta x_1 - \Delta x_1^\top A_1 \Delta x_1] \\ &\quad - \mathbb{E}[{\theta_*}^\top \Delta x_2 - \Delta x_2^\top A_2 \Delta x_2].\end{aligned} \tag{6}$$

The first definition, *score disparity*, measures the difference in average score improvement between the two groups, capturing how changes in features translate to perceived gains across populations (Bechavod et al. 2022). The second definition, *utility disparity*, extends this notion by incorporating the cost of change, reflecting the net benefit each group experiences after accounting for adjustment effort.

We now define regions of score and utility disparity based on their signs with respect to the parameterization $\mathcal{P}$.

[2] As is common in related work—see for example (Bechavod et al. 2022).

[3] Such ℓ_2 cost functions are common in the strategic learning literature, e.g. (Bechavod et al. 2022; Cohen et al. 2023)

[4] Similar assumptions are made in related work, e.g. Bechavod et al. (2022)

Definition 2.1. *We define the following three regions corresponding to $\mathcal{P}$ when using the metric $\mathcal{F}$*

1. Exploitation: *A parameterization corresponds to* Exploitation *if the metric $\mathcal{F} > 0$.*
2. Neutrality: *A parameterization corresponds to* Neutrality *if the metric $\mathcal{F} = 0$.*
3. Burden: *A parameterization corresponds to* Burden *if the metric $\mathcal{F} < 0$.*

The first definition, *Exploitation*, implies that the more advantaged group has higher improvement for the choice of parameters. *Neutrality* implies equal improvements shown by both groups. Finally, *Burden* indicates the region where the more advantaged group suffers a lower improvement from strategic change of feature.

Cost disparities The associated cost matrices for groups 1 and 2 are denoted A_1 and A_2 (respectively) and parameterized by $\mathcal{A} := (A_1, A_2)$. Importantly, in the rest of the paper, we assume that there are cost disparities between groups 1 and 2—similarly to the works of Hu, Immorlica, and Vaughan (2019) and Milli et al. (2019)—, and in particular, that group 1 is *advantaged* and incurs systematically *lower cost* compared to group 2. Formally:

Assumption 2.2. *$A_1 \prec A_2$, or equivalently, $A_2 - A_1$ is positive-definite.*

3 Preliminaries: Agent Best Responses

In this section, we compute the feature improvement vector Δx_g for an agent in a population g for both (i) *naive* agents and (ii) *Bayesian* agents by solving their optimization problems—respectively given by Eq 3 and Eq 4. Proofs of both lemmas are provided in the full version of the paper.

Lemma 3.1. *For a Naive agent in group g, the feature change Δx_g in equilibrium is given by:*

$$\Delta x_g = A_g^{-1} S_g. \tag{7}$$

Lemma 3.2. *For a Bayesian agent in group g, the feature change Δx_g in equilibrium is given by:*

$$\Delta x_g = A_g^{-1}[\omega_g + \beta(\sigma, \gamma)(S_g - \omega_g)] \tag{8}$$

where $\beta(\sigma, \gamma) := \frac{\gamma^2}{\gamma^2 + \sigma^2}$, and $S_g = \theta_* + \sigma Z_g$.

For the Naive agent, observe that the feature change does not depend on the noise level σ. Intuitively, this is because the Naive agent takes the signal at face value. In contrast, the Bayesian agent's response depends on a function of the noise of the signal, σ, and the noise of the prior, γ.

4 Fairness under Incomplete Information

Using the obtained analytical forms for improvements, we now perform fairness analyses using the defined metrics $\mathcal{F}_s$ and $\mathcal{F}_u$ for both a *naive* and *Bayesian* agent. Relevant proofs for this section can be found in the full version of the paper.

4.1 *Naive* agent

In this section, we characterize how disparities in outcomes arise as a function of the amount of information revealed by the learner for *naive* agents.

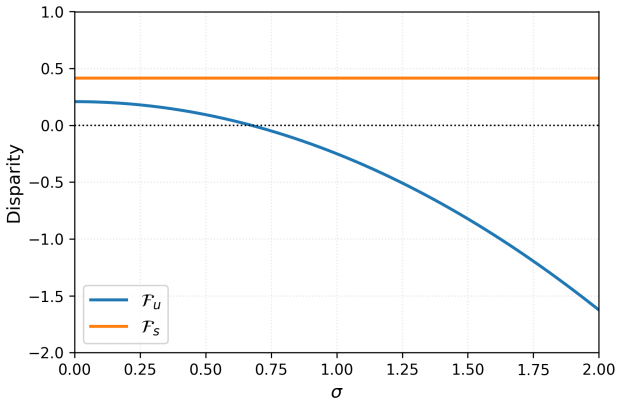

Figure 1: Comparison of *score-disparity* $\mathcal{F}_s$ and *utility-disparity* $\mathcal{F}_u$ as a function of σ for a *naive* agent, computed from Lemma. 4.1 and Lemma. 4.4 respectively. The dotted black line at zero of y-axis represents *Neutrality*. The parameters chosen are: $\theta_* = [1,\ 0.5]^\top, A_1 = diag(2, 1), A_2 = diag(4, 3)$

Score-disparity We start by providing a characterization of the *score-disparity*, and show that it is *invariant* to the amount of information σ released by the learner:

Theorem 4.1. *For a naive agent, in equilibrium:*

$$\mathcal{F}_s = {\theta_*}^\top (A_1^{-1} - A_2^{-1})\theta_*. \tag{9}$$

Corollary 4.2. *The* score-disparity *satisfies $\mathcal{F}_s > 0$ and is invariant w.r.t σ.*

Corollary 4.3. *$Var({\theta_*}^\top \Delta x_1 - {\theta_*}^\top \Delta x_2)$ is monotonically increasing in σ.*

For a *naive* agent, the *score-disparity* is only related to the cost matrix A_g and the true model θ_*. We note that this score is always positive, i.e. it always shows *Exploitation*. While score-disparity of a *naive* agent is agnostic to signal variance or information, it is important to note that the differences in variances of scores in both groups is monotonic in σ. This reflects variability in score improvement across individuals in a group, which gets nullified when computing the average score-improvement across the group.

Utility-disparity On the other hand, the *utility-disparity* shows a monotonic trend from favoring the advantaged group at higher information to unboundedly favoring the disadvantaged group at the limit of no information.

Theorem 4.4. *For a naive agent, in equilibrium the* utility-disparity *$\mathcal{F}_u$ is given by:*

$$\mathcal{F}_u = \frac{1}{2}\left[{\theta_*}^\top (A_1^{-1} - A_2^{-1})\theta_* - \sigma^2 (Tr(A_1^{-1}) - Tr(A_2^{-1}))\right], \tag{10}$$

where Tr(.) denotes matrix trace.

Lemma 4.5. *For a naive agent in equilibrium the utility-disparity $\mathcal{F}_u$ is monotonically decreasing as a function of σ. Neutrality occurs at $\sigma_r = \sqrt{\frac{{\theta_*}^\top (A_1^{-1} - A_2^{-1})\theta_*}{Tr(A_1^{-1}) - Tr(A_2^{-1})}}$. Exploitation*

occurs for $\sigma < \sigma_r$ and Burden occurs for $\sigma > \sigma_r$. Moreover, $\lim_{\sigma\to 0^+} \mathcal{F}_u(\sigma) = \frac{\mathcal{F}_s(\sigma)}{2}$ *and* $\lim_{\sigma\to\infty} \mathcal{F}_u = -\infty$.

The behavior of a *naive* agent is intuitive: as the noise level σ increases, agents observe the model at face value, without accounting for the underlying uncertainty. This leads them to rely on distorted signals, resulting in misallocated effort—often overspending on modifying features that have little impact on their score. Consequently, the lower-cost group, which can afford greater strategic feature changes, may in fact expand more cost to modify features due to overconfidence without increasing their score accordingly. Mathematically, as σ increases, the increased cost disparity given by the term $\sigma^2(Tr(A_1^{-1}) - Tr(A_2^{-1})) > 0$ in eq. 10 dominates the disparities in score, indicating that the lower cost group is, counter-intuitively, overspending compared to the high-cost group.

Relevant plots for both kinds of disparities for a *naive* agent are shown in Fig. 1. Interestingly, here, partial information revelation is optimal in terms of fairness: it is in the best interest of the learner interested in fairness to reveal σ_r amount of information, which achieves the minimum possible disparities in utilities while not affecting score disparities. This may however make it difficult for a policy maker in practice, over simpler mechanisms that just reveal all or no information about the deployed rule. This also implies a fairness-utility trade-off that complexifies a policy maker or learner's decision making: it is easy to see that utility for each of the group is maximized under full information, but fairness requires the learner to hold some information back.

4.2 Bayesian Agent

In this section, we compute the disparities between groups for Bayesian agents at equilibrium. We assume the same prior parameters for agents of both groups throughout and denote $\theta_0 \triangleq \omega_1 = \omega_2$ as the common prior mean, and γ the standard deviation of each component of the characteristic vector. We do so to isolate the effect of information parameter σ and the cost disparities on fairness, that arise *even when all agents have access to the same information*.

In this section, we show that the *score-disparity* is bounded and monotonic. We then show that *Neutrality* occurs at limited information if and only if the prior mean is 'poorly-aligned' in some sense with the true model. On the other hand, we show that the *utility-disparity* shows a transition from being monotonic to non-monotonic in the *level of information* σ revealed by the learner; this transition happens at a given, computable value of the *prior*'s standard deviation γ. We show conditions where *Neutrality* occurs at incomplete information, along with a natural condition where *Neutrality* occurs at several distinct points (in terms of signal variance σ^2) at incomplete information.

Score-disparity We start by providing a full characterization of score disparities. Before doing so, we remind the reader that $A_1 \prec A_2$, implying in particular that $A_1^{-1} - A_2^{-1}$ is a symmetric positive-definite matrix. For a given matrix A, we denote by $\sqrt{A}$ the unique positive-definite matrix Q such that $Q^\top Q = A$. For notational convenience, we define the following transformed variables:

$$k_{\theta_0} := \sqrt{A_1^{-1} - A_2^{-1}}\,\theta_0 \quad \text{and} \quad k_{\theta_*} := \sqrt{A_1^{-1} - A_2^{-1}}\,\theta_*,$$

Theorem 4.6. *For a Bayesian agent, in equilibrium:*

$$\begin{aligned}\mathcal{F}_s &= k_{\theta_0}^\top k_{\theta_*} + (k_{\theta_*}^\top k_{\theta_*} - k_{\theta_0}^\top k_{\theta_*})\beta_\gamma(\sigma) \\ &= (1-\beta_\gamma(\sigma))k_{\theta_0}^\top k_{\theta_*} + \beta_\gamma(\sigma)k_{\theta_*}^\top k_{\theta_*}, \end{aligned} \tag{11}$$

where $\beta_\gamma(\sigma) := \beta(\sigma,\gamma) = \frac{\gamma^2}{\gamma^2+\sigma^2}$.

From Eq 11, we observe that $\mathcal{F}_s$ is a weighted sum of two terms: $k_{\theta_0}^\top k_{\theta_*}$, weighted by a factor proportional to σ^2, and $k_{\theta_0}^\top k_{\theta_*}$, weighted by a factor proportional to γ^2. Both terms quantify alignment within the subspace spanned by the dominant eigenvectors of the matrix $A_1^{-1} - A_2^{-1}$. Specifically, the first term captures the alignment between the prior mean θ_0 and the ground truth θ_*, while the second term measures the concentration of the ground truth within this subspace. We now characterize all possible behaviors for $\mathcal{F}_s$:

Lemma 4.7. *In equilibrium, $\mathcal{F}_s$ is bounded, and is monotonic in σ. Specifically:*

1. *if $k_{\theta_0}^\top k_{\theta_*} < k_{\theta_*}^\top k_{\theta_*}$, then $\mathcal{F}_s$ is decreasing in σ.*
2. *if $k_{\theta_0}^\top k_{\theta_*} > k_{\theta_*}^\top k_{\theta_*}$ ($k_{\theta_0}^\top k_{\theta_*} \geq k_{\theta_*}^\top k_{\theta_*}$) , then $\mathcal{F}_s$ is increasing (non-decreasing) in σ.*
3. *At boundaries, $\mathcal{F}_s(0) = k_{\theta_*}^\top k_{\theta_*}$ and $\lim_{\sigma\to\infty}\mathcal{F}_s(\sigma) = k_{\theta_0}^\top k_{\theta_*}$.*

The monotonic behaviors of *score-disparity* as a function of σ are illustrated in Figure 2a (decreasing), Figure 2b (decreasing), and Figure 2c (increasing). Essentially, the conditions compare i) the alignment between the prior mean and the ground truth to ii) the concentration of the ground truth. Intuitively, when ii) is dominant, the initial prior does not matter, and information revelation is important to an agent. The more information is revealed (i.e. the smaller the σ), the more the advantaged group can leverage their cost edge over the disadvantaged group—$\mathcal{F}_s$ is bigger at smaller σ.

The preceding lemma outlines the general monotonic behavior of $\mathcal{F}_s$ based on the alignment between the prior and true models. In the Corollary below, we provide a condition where Neutrality exists for $\mathcal{F}_s$.

Lemma 4.8. *In equilibrium, $\mathcal{F}_s$ shows* Neutrality *if and only if $k_{\theta_0}^\top k_{\theta_*} < 0$, at $\sigma_r = \sqrt{-\frac{k_{\theta_*}^\top k_{\theta_*}}{k_{\theta_0}^\top k_{\theta_*}}}\gamma$.*

The optimal fairness-oriented revelation policy for the learner, with respect to *score-disparity* depends on the behavior of $\mathcal{F}_s$. For instance, if $\mathcal{F}_s$ is increasing, then the best policy is to reveal full information—since $\mathcal{F}_s$ attains its minimum at $\sigma = 0$, though this may be difficult to implement in practice. On the other hand, if $\mathcal{F}_s$ is decreasing, the optimal strategy is to reveal partial or no information, depending on whether *Neutrality* occurs (as shown in Lemma 4.8); in such cases, revealing partial information around σ_r may be ideal. Lemma 4.8 shows that one must also account for the extent of deterioration caused by the negative term $k_{\theta_0}^\top k_{\theta_*}$, which is weighted higher as σ increases. If *Neutrality* does not occur, then revealing no signal is the optimal policy.

We now state a natural condition that is particularly useful for the learner's analysis.

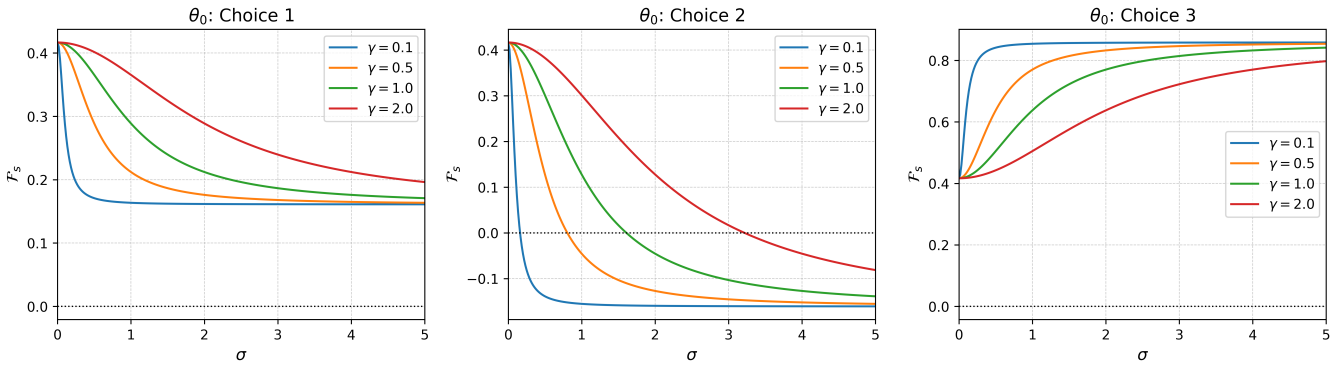

Figure 2: *Score-disparity* $\mathcal{F}_s$ as a function of σ for a *Bayesian* agent, across different values of γ. Each panel corresponds to a different prior mean θ_0: **(a)** $\theta_0 = [0.5,\ 2]^\top$, **(b)** $\theta_0 = -[0.5,\ 2]^\top$, **(c)** $\theta_0 = k[0.5,\ 2]^\top$ with $\|\theta_0\| = 2\|\theta^*\|$. The dotted black horizontal line indicates *Neutrality*. Parameters: $\theta^* = [1,\ 0.5]^\top$, $A_1 = \mathrm{diag}(2, 1)$, $A_2 = \mathrm{diag}(4, 3)$.

Corollary 4.9. *If* $\|k_{\theta_0}\| \leq \|k_{\theta_*}\|$, *then, in equilibrium:*

1. $\mathcal{F}_s$ *is monotonically non-increasing w.r.t* σ.
2. $\mathcal{F}_s$ *is bounded by* $(k_{\theta_0}^\top k_{\theta_*}, k_{\theta_*}^\top k_{\theta_*}]$.

Note that we use $\|.\|$ to denote the Euclidean norm for vectors and the spectral norm (i.e., largest singular value) for matrices. The assumption on the prior mean θ_0 in Corollary 4.9 ensures that the norm of θ_0 is comparable to that of the ground truth θ_*. More precisely, if $A_g = \alpha_g \mathbf{I}$, then the condition simplifies to $\|\theta_0\| \leq \|\theta_*\|$. This ensures that the prior does not assign disproportionately more "magnitude" or regularization than the true signal, maintaining consistency with the expected scale of the ground truth[5].

This also highlights that monotonic increase in $\mathcal{F}_s$ is expected to be less common, typically arising only in unconstrained settings. However, there do exist cases where such increasing behavior still occurs as shown in Figure 2c.

Utility-disparity We now characterize the *utility-disparity* for a Bayesian agent in equilibrium. We adopt the same notation as in the *score-disparity* analysis, specifically defining the following constants: $k_{\theta_0} = \sqrt{A_1^{-1} - A_2^{-1}}\ \theta_0$ and $k_{\theta_*} = \sqrt{A_1^{-1} - A_2^{-1}}\ \theta_*$.

Theorem 4.10. *Let* $m \triangleq (Tr(A_1^{-1}) - Tr(A_2^{-1}))$, *the* utility-disparity *for a Bayesian agent in equilibrium is given by:*

$$\mathcal{F}_u(\sigma) = -\frac{\beta_\gamma^2(\sigma)}{2}\left(\left(k_{\theta_*} - k_{\theta_0}\right)^2 + \sigma^2 m\right) + \beta_\gamma(\sigma)\left(k_{\theta_*} - k_{\theta_0}\right)^2 + \left(k_{\theta_0}^\top k_{\theta_*} - \frac{k_{\theta_0}^\top k_{\theta_0}}{2}\right), \quad (12)$$

where $\beta_\gamma(\sigma) := \beta(\sigma, \gamma) := \frac{\gamma^2}{\gamma^2 + \sigma^2}$.

The above expression leads to different cases in terms of both monotonicity and whether we observe *Neutrality*, *Exploitation*, or *Burden*. These cases are highlighted below:

[5]We believe this to be a relatively mild assumption. This occurs, for example, if a user perfectly knows the top k features, but ignores all other features, due to bounded rationality.

Lemma 4.11. *In equilibrium,* $\mathcal{F}_u$ *for a Bayesian agent, is bounded. Additionally, let* $\gamma_c := \sqrt{\frac{2}{m}}\ |k_{\theta_*} - k_{\theta_0}|$, *the following are possible behaviors for* $\mathcal{F}_u$*:*

Monotone case: *If* $\gamma \leq \gamma_c$, *then* $\mathcal{F}_u$ *is monotonically decreasing in* σ. *The following are sub-cases:*

1. *If* $k_{\theta_0}^\top k_{\theta_0} > 2k_{\theta_0}^\top k_{\theta_*}$, *then* $\mathcal{F}_u$ *attains Neutrality at a unique* σ.
2. *If* $k_{\theta_0}^\top k_{\theta_0} \leq 2k_{\theta_0}^\top k_{\theta_*}$, *then* $\mathcal{F}_u$ *shows Exploitation throughout for all* σ.

Non-monotone case: *If* $\gamma > \gamma_c$, *then* $\mathcal{F}_u$ *is non-monotone and attains a global minimum at* $\sigma_{\min}$ *given by*

$$\sigma_{min} = \sqrt{\frac{1}{1 - \frac{\gamma_c^2}{\gamma^2}}}\ \gamma. \quad (13)$$

$\mathcal{F}_u$ *is decreasing for* $\sigma \leq \sigma_{\min}$, *and increasing for* $\sigma > \sigma_{min}$. *The following are sub-cases:*

1. *If* $k_{\theta_0}^\top k_{\theta_0} > 2k_{\theta_0}^\top k_{\theta_*}$, *then* $\mathcal{F}_u$ *attains Neutrality at one point for some* $\sigma < \sigma_{\min}$.
2. *If* $k_{\theta_0}^\top k_{\theta_0} \leq 2k_{\theta_0}^\top k_{\theta_*}$, *then* $\mathcal{F}_u$ *has: no point of Neutrality if* $\mathcal{F}_u(\sigma_{\min}) > 0$, *one if* $\mathcal{F}_u(\sigma_{\min}) = 0$, *two if* $\mathcal{F}_u(\sigma_{\min}) < 0$.

Further, from Eq 12 it is easy to see the following values of $\mathcal{F}_u$ in boundary cases:

$$\mathcal{F}_u(0) = \frac{k_{\theta_*}^\top k_{\theta_*}}{2}, \quad \text{and} \quad \lim_{\sigma \to \infty} \mathcal{F}_u(\sigma) = k_{\theta_0}^\top k_{\theta_*} - \frac{k_{\theta_0}^\top k_{\theta_0}}{2}.$$

Figure 3 illustrates separate panels, each corresponding to one of the cases described in Lemma 4.11 for *utility-disparity*. Panel 3a shows monotonic-decreasing pattern with only *Exploitation*. In contrast, Panel 3b shows the same monotonic behavior with a single point of *Neutrality*. Panels 3c through 3e display non-monotonic behaviors: Panel 3c features one point of *Neutrality*, Panel 3d shows none, and Panel 3e presents two distinct *Neutrality* points.

We observe that for lower prior variance (i.e., $\gamma \leq \gamma_c$), $\mathcal{F}_u$ is monotonically decreasing. In contrast, for higher variance (i.e., $\gamma > \gamma_c$), $\mathcal{F}_u$ becomes non-monotonic, exhibiting

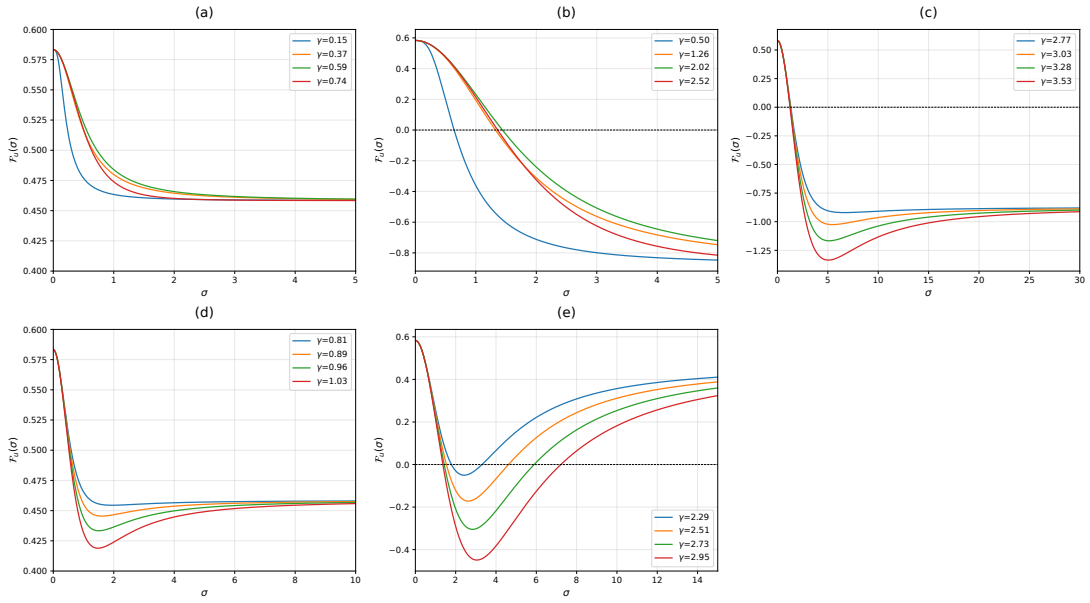

Figure 3: *Utility-disparity* $\mathcal{F}_u$ as a function of σ for a *Bayesian* agent, for different values of γ, where each panel—**(a)** through **(e)**—represents a chosen prior mean θ_0. The dotted black line on the y-axis denotes *Neutrality*.

a unique global minimum at $\sigma_{\min}$. In this regime, $\mathcal{F}_u$ decreases for $\sigma \leq \sigma_{\min}$ and increases for $\sigma > \sigma_{\min}$. In the monotonic case, the behavior—whether it is purely decreasing with no *Neutrality* or decreasing with a single *Neutrality* point—is closely aligned with the scenarios analyzed in score disparities (Case 1 of Lemma 4.7). Consequently, the learner's optimal revelation strategy for *utility-disparity* mirrors that for *score-disparity* in these cases—either revealing no information or revealing partial information around the point of *Neutrality*, depending on whether it occurs.

An interesting observation is that in both score and *utility-disparity*, the condition $k_{\theta_0}^\top k_{\theta_*} < 0$ guarantees similar behavior—i.e monotonic shift from *Exploitation*→ *Neutrality* → *Burden* as σ increases. This is intuitive, as the condition implies that the prior is misaligned with the deployed model, in which case information revelation provides an edge to the advantaged group (since the advantaged group can take advantage of knowing the model better than the disadvantaged group). I.e., $\sigma \to 0$ (more information) maximizes disparities between both groups. Under such conditions, the learner can adopt a consistent policy by revealing partial information, specifically around the region where *Neutrality* occurs for both disparities in scores and utilities.

In the non-monotonic cases where one point of *Neutrality* exists (cases 1 and, 2 with $\mathcal{F}_u(\sigma_{\min}) = 0$; as described in Lemma 4.11), the learner's optimal strategy closely resembles that of the monotonic decreasing case with a single *Neutrality* point—namely, to identify and operate in a neighborhood around the point of *Neutrality*. Even when no point of *Neutrality* exists in the non-monotonic regime (as seen in Case 2: with $\mathcal{F}_u(\sigma_{\min}) > 0$), the behavior around $\sigma_{\min}$ remains relevant, since it corresponds to the point of minimal *utility-disparity*, albeit characterized by *Exploitation*.

Finally, things become complex when two points of *Neutrality* emerge in the non-monotonic case—i.e., when $\mathcal{F}_u(\sigma_{\min}) < 0$—denoted by σ_{r_1} and σ_{r_2}, with $\sigma_{r_1} < \sigma_{r_2}$. The interval $(\sigma_{r_1}, \sigma_{r_2})$ defines a region of *Burden*, where the learner's intervention results in worsened *utility-disparity*. The presence of two *Neutrality* points, σ_{r_1} and σ_{r_2}, poses a significant challenge for the learner. The existence of an intermediate region where *utility-disparity* is negative—termed *Burden*—complicates the learner's signaling strategy, as it is unclear whether revealing more or less information will improve outcomes. In the Lemma below we show a region of selected parameters that corresponds to two distinct points of *Neutrality*.

Lemma 4.12. *In equilibrium, the following is a region* $\mathcal{R}$ *of relevant parameters that guarantee two points of* Neutrality *for* $\mathcal{F}_u$.

$$\mathcal{R} = \Bigg\{ k_{\theta_0}, k_{\theta_*}, m, \gamma : k_{\theta_0}^\top k_{\theta_0} < 2k_{\theta_0}^\top k_{\theta_*} \text{ and } \gamma > \max\Bigg\{ \sqrt{\frac{1}{m}\left(2k_{\theta_0}^\top k_{\theta_*} - k_{\theta_0}^\top k_{\theta_0} + 3k_{\theta_*}^\top k_{\theta_*}\right)}, \sqrt{\frac{2}{m}|k_{\theta_0} - k_{\theta_*}|} \Bigg\} \Bigg\} \tag{14}$$

Using the above feasibility region, we can infer that, given $k_{\theta_0}^\top k_{\theta_0} < 2k_{\theta_0}^\top k_{\theta_*}$, a sufficiently large prior variance γ^2 determined by the model parameters can lead to the existence of two distinct points of *Neutrality* for $\mathcal{F}_u$.

5 Fairness under Unequal Priors

In the previous section, we assumed that agents in both groups shared the same prior distribution. We now move away from this assumption and introduce an additional source of heterogeneity: namely, agents from different groups possess distinct prior means.

To do so, we adopt and expand on the model of Bechavod et al. (2022), where agents are assumed to only have access to a subspace of the overall feature space. In their work, if the true model is given by θ_*, group 1's ability to recover θ_* is limited to a subspace S_1 of $\mathbb{R}^d$, and group 2's is limited to a subspace S_2 of $\mathbb{R}^d$. Similarly, we will assume that for all $g \in \{1, 2\}$, group g's ability to reason about the learner's model is limited to subspace S_g; however, a significant departure from (Bechavod et al. 2022) is that we still consider *Bayesian* agents[6], who may still be uncertain about their knowledge of the model. Formally, we make the following assumption about the agents' priors:

Assumption 5.1. *For all groups $g \in \{1, 2\}$, let us denote Π_g be the projection operator to subspace S_g. We assume that in group g, an agent's prior is given by $\pi_g(\theta) \sim \mathcal{N}(\Pi_g\theta_*, \gamma^2\mathbf{I})$.*

Remark 5.2. *This projection to a subspace is especially useful in explaining discrepancies in groups on the basis of their feature characteristics. For instance, in bank loan approvals, suppose group A primarily includes applicants with formal employment—so their features emphasize steady income, tax returns, and employer verification. In contrast, group B may comprise individuals from gig or informal sectors, where key features pertain to inconsistent earnings, alternate credit scores, or cash flows from small businesses. Then each of the groups may only see a small and a different part of the space of features. This can also encode bounded rationality with disparate information about the model, where each group only focuses on the top k features, and may have different understanding of what these top-k features are. See (Bechavod et al. 2022) for more details.*

We note a few assumptions made here for simplicity of exposition. First, we assume that the prior is unbiased within S_g, i.e. centered around $\Pi_g\theta_*$. We do so to control for the effect of the alignment of the mean of the prior with the true model; i.e., our disparity results will hold *even in the best case when both groups have a credible prior*. Second, both groups have the same variance term γ^2–we isolate the effect of differences in prior means solely on disparities.

We also define the amount of overlap between both groups using the metric of Bechavod et al. (2022).

Definition 5.3 ((Bechavod et al. 2022))**.** *Given a true model $\theta_* \in \mathbb{R}^d$ and projections $\Pi_1, \Pi_2 \in \mathbb{R}^{d\times d}$, the information overlap-proxy between groups g_1 and g_2 is defined as*

$$r_{1,2}(\theta_*) := \|\Pi_1\theta_* - \Pi_2\theta_*\|.$$

We now analyze disparities in the context of the Bayesian model framework established above. Notably, our analysis is agnostic to the specific choice of the true model θ_* used by the learner; rather, it focuses on the interaction between group-specific feature subspaces (captured by Π_g) and group-dependent cost matrices (i.e A_g). This allows us to study group-level discrepancies that arise purely due to structural differences in feature representations.

5.1 Score-disparity

We first analyze fairness via *score-disparity* similar to that done in Section. 4.2. All the relevant proofs of the Section can be found in the full version of the paper.

Theorem 5.4. *For a Bayesian agent, the score disparity between both groups in equilibrium is given by:*

$$\mathcal{F}_s = {\theta_*}^\top(A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2)\theta_* + \big[{\theta_*}^\top(A_1^{-1} - A_2^{-1})\theta_* - {\theta_*}^\top(A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2)\theta_*\big]\beta_\gamma(\sigma). \quad (15)$$

We now characterize the behavior of $\mathcal{F}_s$ and outline the behavior in some natural cases.

When do Exploitation, Burden, and Neutrality occur?

Lemma 5.5. *In equilibrium, Exploitation with respect to $\mathcal{F}_s$ always occurs for all $\sigma \in \mathbb{R}^+$ and for all $\theta_* \neq \bar{0}$ if and only if $(A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2) + (\Pi_1A_1^{-1} - \Pi_2A_2^{-1})$ is positive semi-definite. Furthermore, if $\Pi_gA_g^{-1} = A_g^{-1}\Pi_g,\ g \in \{1, 2\}$, the condition simplifies to $A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2$ being positive semi-definite.*

Note: Commutativity of Π_g, A_g^{-1} is satisfied in many natural settings including for instance, when $A_g = \alpha_g\mathbf{I}$ for $\alpha_g > 0$ (see Appendix E. of Bechavod et al. (2022)).

Lemma 5.6. *If $A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2 \succeq 0$, then $\eta(\Pi_1) \subseteq \eta(\Pi_2)$, where $\eta(B)$ represents the null-space of a matrix $B \in \mathbb{R}^{d\times d}$. If the inequality is strict, then $\Pi_1 = \mathbf{I}$.*

Lemmas 5.5 and 5.6 imply that, under many conditions, when *Exploitation* occurs for all $\sigma \geq 0$ and for every true model θ_*, the feature subspace observed by the disadvantaged groups is contained is the feature subspace observed by the advantaged groups[7]. I.e., this corresponds to a situation where the advantaged group always has access to more information compared the disadvantaged group. Furthermore, *Exploitation* persists even in the limit of no information ($\sigma \to \infty$) for every θ_* if and only if the advantaged group's prior fully spans the feature subspace—this is however a corner case that may not arise in realistic scenarios.

We now state our condition for the existence of *Neutrality*:

Lemma 5.7. *In equilibrium, Neutrality with respect to $\mathcal{F}_s$ occurs for all $\theta_* \neq \bar{0}$ if and only if $(A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2) + (\Pi_1A_1^{-1} - \Pi_2A_2^{-1})$ is negative definite, at*

$$\sigma = \sigma_r \triangleq \sqrt{\frac{{\theta_*}^\top(A_1^{-1} - A_2^{-1})\theta_*}{{\theta_*}^\top(\Pi_2A_2^{-1} - \Pi_1A_1^{-1})\theta_*}}\gamma.$$

Furthermore, if $\Pi_gA_g^{-1} = A_g^{-1}\Pi_g,\ g \in \{1, 2\}$, then Neutrality occurs for all $\theta_* \neq \bar{0}$ if and only if $A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2$ is negative definite.

[6] Bechavod et al. (2022) assumes that θ_* is perfectly known and taken at face value within S_g for group g.

[7] Note that $\eta(\Pi_1) \subseteq \eta(\Pi_2)$ implies that the induced subspaces satisfy $S_2 \subset S_1$.

Lemma 5.8. *Given $A_2 \succ A_1$, if $(A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2) \prec 0$, then $\Pi_1 \neq \mathbf{I}$ and $\Pi_2 = \mathbf{I}$.*

Lemmas 5.7 and 5.8 imply that *Neutrality* could occur with respect to $\mathcal{F}_s$ for every true model parameter θ_* only when the disadvantaged group has full information of the true model θ_* while the advantaged group does not have full information on θ_*. This is expected to be a rare scenario, in fact showing that we do not expect neutrality to happen in practice with respect to scores for complex, black-box models. In practice, this means that a learner interested in fairness may have to aim for reducing *Exploitation* as much as possible but will never fully achieve *Neutrality*.

How does the level of information revelation affect fairness? We shift our attention towards monotonicity of $\mathcal{F}_s$ with respect to σ.

Definition 5.9. *For an orthogonal projection matrix Π, its complement $\Pi^\perp$ is defined as $\Pi^\perp = \mathbf{I} - \Pi$.*

We now move on to our main monotonicity result:

Lemma 5.10. *In equilibrium, $\mathcal{F}_s$ is monotonically non-increasing (respectively, non-decreasing) with σ for all θ_* if and only if $(A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp) + (\Pi_1^\perp A_1^{-1} - \Pi_2^\perp A_2^{-1})$ is positive semi-definite (respectively, negative semi-definite). The inequality is strict for all $\theta_* \neq \bar{0}$ — that is, $\mathcal{F}_s$ is strictly decreasing (increasing)—if the matrix is positive definite (negative definite). Furthermore, if $\Pi_g A_g^{-1} = A_g^{-1}\Pi_g$ for each $g \in \{1, 2\}$, the condition reduces to $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp$ being positive (negative) semi-definite, or positive (negative) definite for strict monotonicity.*

The above lemma relies on conditions: $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp$ being either positive/negative semi-definite or positive/negative definite. Below, we provide interpretations of these conditions:

Lemma 5.11. *Given $A_2 \succ A_1$, we have the following:*

1. *If $(A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp) \succeq 0$, then span$(\Pi_1) \subseteq$ span(Π_2). If the inequality is strict, then $\Pi_1 = \mathbf{0}$.*
2. *If $(A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp) \preceq 0$, then span$(\Pi_2) \subseteq$ span(Π_1). If the inequality is strict, then $\Pi_1 \neq \mathbf{0}$, and $\Pi_2 = \mathbf{0}$.*

Lemmas 5.10 and 5.11 together imply that when $\mathcal{F}_s$ is increasing in σ (for all θ_*), the prior feature subspace of the disadvantaged group is contained within that of the advantaged group. In this case, the disadvantaged group is disadvantaged in terms of *both* costs and information. As the learner reveals more information (lowers σ), he reduces the informational disparities across both groups, helping group 2; i.e. $\mathcal{F}_s$ becomes smaller at lower σ's. Similarly, when $\mathcal{F}_s$ is decreasing in σ, the prior feature subspace of the advantaged group is contained within that of the disadvantaged group. In this case, the learner revealing more information about the model strongly benefits the advantaged group, and disparities increase at low values of σ. Interestingly, and despite the advantaged group still having an edge in terms of cost of feature manipulations, $\mathcal{F}_s$ remains monotonic.

Overlapping, non-nested subspaces We now quantify the effect of overlap between prior subspaces of both groups on the disparity in scores at equilibrium. To do so, we let $A_1 = A_2 := A$ to isolate the effect of the information overlap itself, and upper bound $|\mathcal{F}_s|$ using the *information-overlap proxy* that we defined earlier.

Lemma 5.12. *Suppose $A_1 = A_2 = A$, then:*

$$|\mathcal{F}_s| \leq (1 - \beta_\gamma(\sigma))\|A\|^{-1}\|\theta_*\| \cdot r_{1,2}(\theta_*).$$

The above inequality relates *score disparity* between two groups to the information-overlap proxy $r_{1,2}(\theta_*)$ using an upper bound on $|\mathcal{F}_s|$. The bound decreases as the amount of information increases (i.e the RHS decreases, given all other parameters, as σ decreases) and equals 0 at full information ($\sigma = 0$) whereas at the limit of no information ($\sigma \longrightarrow \infty$), the bound goes to $\|A\|^{-1}\|\theta_*\|r_{1,2}(\theta_*)$. The bound suggests that more information overlap, perhaps unsurprisingly, leads to fewer disparities across groups.

5.2 Utility disparity

We now analyze the *utility disparity* for a Bayesian agent when both groups choose their prior means based on the subspace of feature information of model known to them. Throughout the section, we assume commutativity of A_g^{-1}, Π_g; consistent with the assumptions used in the analysis of the *score disparity*. We start by deriving the *utility disparity* for Bayesian agents:

Theorem 5.13. *Let $k_\pi := {\theta_*}^\top(A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2)\theta_*$, $k_A := {\theta_*}^\top(A_1^{-1} - A_2^{-1})\theta_*$, and $m := (Tr(A_1^{-1}) - Tr(A_2^{-1}))$. If $\Pi_g A_g^{-1} = A_g^{-1}\Pi_g$ for all $g \in \{1, 2\}$, the utility disparity between both groups in equilibrium is given by:*

$$\mathcal{F}_u(\sigma) = \frac{\beta_\gamma^2(\sigma)}{2}\big(k_\pi - k_A - \sigma^2 m\big) + \beta_\gamma(\sigma)\big(k_A - k_\pi\big) + \frac{k_\pi}{2}. \tag{16}$$

Note that for full and no information revelation, we have:

$$\mathcal{F}_u(0) = \frac{k_A}{2} > 0, \ \lim_{\sigma\to\infty} \mathcal{F}_u(\sigma) = \frac{k_\pi}{2}.$$

We now characterize all possible behaviors of $\mathcal{F}_u$:

Lemma 5.14. *In equilibrium, $\mathcal{F}_u$ is bounded. The following are possible behaviors of $\mathcal{F}_u$:*

1. ***Monotone case:** If $\gamma^2 \leq \frac{2}{m}(k_A - k_\pi)$, then $\mathcal{F}_u$ is monotonically decreasing in σ. The following are the sub-cases:*
   (a) *If $k_\pi < 0$, then $\mathcal{F}_u$ attains* Neutrality *at a unique σ.*
   (b) *If $k_\pi \geq 0$, then $\mathcal{F}_u$ does not attain* Neutrality *for all σ, and instead shows* Exploitation *for all σ.*
2. ***Non-monotone case:** If $\gamma^2 > \frac{2}{m}(k_A - k_\pi)$, then $\mathcal{F}_u$ attains a global minimum at a unique $\sigma_{\min}$ given by:*

$$\sigma_{min} = \sqrt{\frac{1}{1 - \frac{2}{m\gamma^2}(k_A - k_\pi)}}\,\gamma.$$

$\mathcal{F}_u$ is decreasing for $\sigma \leq \sigma_{\min}$, and increasing for $\sigma > \sigma_{\min}$. The following are sub-cases:
   (a) *If $k_\pi < 0$, then $\mathcal{F}_u$ attains* Neutrality *for some $\sigma < \sigma_{\min}$.*

(b) If $k_\pi \geq 0$, then $\mathcal{F}_u$ has no point of Neutrality *if $\mathcal{F}_u(\sigma_{\min}) > 0$, one if $\mathcal{F}_u(\sigma_{\min}) = 0$, two if $\mathcal{F}_u(\sigma_{\min}) < 0$.*

Similar to Lemma 4.11—the case of equal priors across groups, we observe both monotonic and non-monotonic behavior depending on the value of γ. In particular, when the value of γ is low enough, $\mathcal{F}_u$ is monotone in σ, and when γ is large enough, $\mathcal{F}_u$ exhibit non-monotonicities. However, we note here that the monotonic case may not arise when $k_A < k_\pi$ (unlike in Lemma 4.11 where there is always a value of γ sporting each regime). Another interesting observation is that: if $k_A \geq k_\pi$, then $\sigma_{\min} \geq \gamma$, otherwise if $k_A < k_\pi$, then $\sigma_{\min} < \gamma$. On the other hand, when both priors are the same, $\sigma_{\min}$ corresponding to the non-monotonic behavior always shows $\sigma_{\min} \geq \gamma$ (from eq. 13).

We now leverage the subspace-based matrix characterizations developed earlier to establish conditions under which $\mathcal{F}_u$ exhibits monotonic and non-monotonic behavior *agnostic* of θ_*, analogous to the approach taken for *score disparities* in the unequal prior setting:

Assumption 5.15. *In what follows, we additionally assume $A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2 \succeq 0$, and either $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp \succeq 0$ or $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp \preceq 0$.*

The constraint $A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2 \succeq 0$ is equivalent to '*no-burden*' for all θ_* at the limit of no information ($\sigma \to \infty$), where as the constraint on the positivity of $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp$ helps control the sign of $k_A - k_\pi$.

Lemma 5.16. *If $A_1^{-1}\Pi_1 - A_2^{-1}\Pi_2 \succeq 0$ and $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp \succeq 0$, then the following characterization holds for $\mathcal{F}_u$ for all θ_*:*

1. ***Monotone case:*** *If $\gamma^2\mathbf{I} \preceq \frac{2}{m}(A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp)$, then $\mathcal{F}_u$ is monotonically decreasing in σ, and does not attain* Neutrality *at any σ.*
2. ***Non-monotone case:*** *If $\gamma^2\mathbf{I} \succ \frac{2}{m}(A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp)$, then $\mathcal{F}_u$ attains a global minimum at $\sigma_{\min}$ given by:*

$$\sigma_{min} = \sqrt{\frac{1}{1 - \frac{2}{m\gamma^2}(k_A - k_\pi)}}\,\gamma.$$

Also, $\mathcal{F}_u$ attains no point of Neutrality if $\mathcal{F}_u(\sigma_{\min}) > 0$, one if $\mathcal{F}_u(\sigma_{\min}) = 0$, two if $\mathcal{F}_u(\sigma_{\min}) < 0$.

It is important to note that in Lemma 5.16, we expect the monotonic case to not happen in practice simultaneously for all θ_* as the $\gamma^2\mathbf{I} \preceq \frac{2}{m}(A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp)$ requirement implies strict positive definiteness of $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp$ which would here would imply $\Pi_1 = \mathbf{0}$ (as it follows from Lemma 5.6). Using similar arguments, we can show that if we assume $A_1^{-1}\Pi_1^\perp - A_2^{-1}\Pi_2^\perp \preceq 0$, then monotonic case is unrealistic and non-monotonic case holds for every $\gamma > 0$ and θ_*.The number σ where neutrality is attained depends on the sign of $\sigma_{\min}$ which further depends on γ.

We now isolate the effect of information overlap on $\mathcal{F}_u$ by assuming equal cost matrices for both groups:

Lemma 5.17. *Suppose $A_1 = A_2 = A$ and $\Pi_g A_g^{-1} = A_g^{-1}\Pi_g$ for $g \in \{1, 2\}$, then $|\mathcal{F}_u| \leq \frac{1}{2}(1 - \beta_\gamma(\sigma))^2\|A\|^{-1}\|\theta_*\|r_{1,2}(\theta_*)$.*

Similarly to the bound for *score disparity*, this bound decreases as the information increases, i.e., the right-hand side decreases with decreasing σ. It vanishes under full information ($\sigma = 0$) and approaches $\|A\|^{-1}\|\theta_*\|r_{1,2}(\theta_*)$ as $\sigma \to \infty$. However, the key difference is that this bound decays faster due to the presence of $\beta_\gamma^2(\sigma)$, in contrast to the $\beta_\gamma(\sigma)$ factor in the score disparity case.

6 Conclusion and Limitations

Our key insight is that the amount of information made available to agents, under cost asymmetries, has complex and often counter-intuitive implications for fairness.

For *naive agents*, transparency reduces utility disparities in a monotonic way, but can sometimes *harm* the *advantaged* group with lower costs by inducing over-investment in uninformative directions. Perhaps surprisingly, information relevation does not impact *expected score disparities*, though it does impact the level of randomness around said score disparities.

For *Bayesian agents*, disparities are *bounded*, and utility disparities are often minimized at intermediate levels of transparency. Counter-intuitively, more transparency is not always better; revealing information leads to a tension between i) helping the advantaged group who, with their lower costs, can take advantage of the additional information more efficiently than the disadvantaged group when it comes to the cost of changing features and ii) additional information helps both groups not invest sub-optimal feature modifications. Effect ii) mostly benefits the higher-cost, disadvantaged group when it comes to efficient feature modifications.

When groups differ in *informational priors*, score and utility disparities are shaped by the *alignment and overlap* of group-specific priors. This expands prior work (e.g., Bechavod et al. (2022)) by introducing Bayesian beliefs, where agents quantify uncertainty around their belief about the deployed classifier.

We characterize when *Neutrality*, *Exploitation*, and *Burden* occur, and identify settings where the learner can use the amount of information disclosure as a knob to reduce disparities. We believe that our work advances the understanding of fairness in strategic learning settings.

Limitations and Future Work. We focus on linear models and Gaussian noise, which allows us to derive useful initial insights on the level of information a learner should reveal. These models may not capture more complex models and uncertainty, though we believe they are providing useful first-order insights. Further, a learner may be interested not only in fairness, but also on deploying a model that is as accurate as possible; an interesting direction work future work is to characterize *accuracy-fairness* trade-offs as a function of how much information a learner reveals about their scoring or decision rule. We expected that this trade-off will provide further arguments for partial information revelation when agents may try to game the classifier, a phenomenon that is prevalent in practice.

Acknowledgements

Juba Ziani's research was supported by NSF CAREER Award IIS-2336236. Full version of the paper (including proofs) is available at https://arxiv.org/abs/2506.00627

References

Ahmadi, S.; Beyhaghi, H.; Blum, A.; and Naggita, K. 2021. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, 6–25.

Ahmadi, S.; Beyhaghi, H.; Blum, A.; and Naggita, K. 2022. On classification of strategic agents who can both game and improve. *arXiv preprint arXiv:2203.00124*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. ProPublica, May 23, 2016.

Bechavod, Y.; Podimata, C.; Wu, S.; and Ziani, J. 2022. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, 1691–1715. PMLR.

Braverman, M.; and Garg, S. 2020. The role of randomness and noise in strategic classification. *arXiv preprint arXiv:2005.08377*.

Chen, Y.; Liu, Y.; and Podimata, C. 2020. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33: 15265–15276.

Cohen, L.; Sharifi-Malvajerdi, S.; Stangl, K.; Vakilian, A.; and Ziani, J. 2023. Sequential strategic screening. In *International Conference on Machine Learning*, 6279–6295. PMLR.

Cohen, L.; Sharifi-Malvajerdi, S.; Stangl, K.; Vakilian, A.; and Ziani, J. 2024. Bayesian strategic classification. *arXiv preprint arXiv:2402.08758*.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 55–70.

Ebrahimi, R.; Vaccaro, K.; and Naghizadeh, P. 2024. The double-edged sword of behavioral responses in strategic classification: Theory and user studies. *arXiv preprint arXiv:2410.18066*.

Efthymiou, V.; Podimata, C.; Sen, D.; and Ziani, J. 2025. Incentivizing Desirable Effort Profiles in Strategic Classification: The Role of Causality and Uncertainty. *arXiv preprint arXiv:2502.06749*.

Estornell, A.; Das, S.; Liu, Y.; and Vorobeychik, Y. 2023. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 389–399.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Harris, K.; Heidari, H.; and Wu, S. Z. 2021. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34: 28728–28741.

Horowitz, G.; and Rosenfeld, N. 2023. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, 13233–13253. PMLR.

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

Jagadeesan, M.; Mendler-Dünner, C.; and Hardt, M. 2021. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, 4687–4697. PMLR.

Jung, C.; Kannan, S.; Lee, C.; Pai, M.; Roth, A.; and Vohra, R. 2020. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 677–678.

Kleinberg, J.; and Raghavan, M. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4): 1–23.

Lechner, T.; Urner, R.; and Ben-David, S. 2023. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*, 18714–18732. PMLR.

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

Liu, P.; and Sun, W. W. 2025. Fairness-aware Contextual Dynamic Pricing with Strategic Buyers. *arXiv preprint arXiv:2501.15338*.

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency*, 230–239.

Shavit, Y.; Edelman, B.; and Axelrod, B. 2020. Causal strategic linear regression. In *International Conference on Machine Learning*, 8676–8686. PMLR.

Somerstep, S.; Ritov, Y.; and Sun, Y. 2024. Algorithmic fairness in performative policy learning: Escaping the impossibility of group fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 616–630.

Sundaram, R.; Vullikanti, A.; Xu, H.; and Yao, F. 2023. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192): 1–38.

Zhang, X.; Khalili, M. M.; Jin, K.; Naghizadeh, P.; and Liu, M. 2022. Fairness interventions as (dis) incentives for strategic manipulation. In *International Conference on Machine Learning*, 26239–26264. PMLR.