UTMD Working Paper
The University of Tokyo
Market Design Center

UTMD-011

On Statistical Discrimination as a Failure of Social Learning: A Multi-Armed Bandit Approach

Junpei Komiyama
New York University

Shunya Noda
University of British Columbia

July 7, 2021

On Statistical Discrimination as a Failure of Social Learning: A Multi-Armed Bandit Approach

Junpei Komiyama Shunya Noda*

First Draft: October 2, 2020 Current Version: July 7, 2021

Abstract

We analyze statistical discrimination in hiring markets using a multi-armed bandit model. Myopic firms face workers arriving with heterogeneous observable characteristics. The association between the worker's skill and characteristics is unknown ex ante; thus, firms need to learn it. Laissez-faire causes perpetual underestimation: minority workers are rarely hired, and therefore, underestimation towards them tends to persist. Even a slight population-ratio imbalance frequently produces perpetual underestimation. We propose two policy solutions: a novel subsidy rule (the hybrid mechanism) and the Rooney Rule. Our results indicate that temporary affirmative actions effectively mitigate discrimination caused by insufficient data.

*Komiyama: Leonard N. Stern School of Business, New York University. E-mail: junpei.komiyama@gmail.com. Noda: Vancouver School of Economics, University of British Columbia. E-mail: shunya.noda@gmail.com. We are grateful to Itai Ashlagi, Tomohiro Hara, Yoko Okuyama, Masayuki Yagasaki, and the seminar participants at Happy Hour Seminar, the University of British Columbia, Tokyo Keizai University, the University of Tokyo, the AFCI Workshop in NeurIPS2020, the AI4SG Workshop in IJCAI 2020, the University of Texas at Austin, CyberAgent, Inc., WEAI International Conference 2021, JEA Spring Meeting 2021, CEA Annual Meeting 2021, and NASMES 2021 for their helpful comments. All remaining errors are our own.

1 Introduction

Statistical discrimination refers to discrimination against minority people, taken by fully rational and non-prejudiced agents. Previous studies have shown that, even in the absence of prejudice, discrimination can occur persistently because of various reasons, including the discouragement of human capital investment (Arrow, 1973; Foster and Vohra, 1992; Coate and Loury, 1993; Moro and Norman, 2004), information friction (Phelps, 1972; Cornell and Welch, 1996; Bardhi et al., 2020), and search friction (Mailath et al., 2000; Che et al., 2019). The literature has proposed various affirmative-action policies to solve statistical discrimination, with many having been implemented in practice.

This paper demonstrates that statistical discrimination may appear as a failure of social learning. We endogenize the evolution of biased beliefs and analyze its consequences. Our model assumes that (i) all firms (decision-makers) are fully rational and non-prejudiced (i.e., attempt to hire the most productive worker), and (ii) all workers are ex ante symmetric. We show that, even in such an environment, a biased belief could be generated and persist in the long run. We demonstrate that *temporary* affirmative actions effectively improve welfare and equality.

Although our model applies more broadly, we use the terminology of hiring markets to describe our model. We develop a *multi-armed bandit model* of social learning, in which many myopic and short-lived firms sequentially make hiring decisions. In each round, a firm hires one worker from a set of candidates. Each firm's utility is determined by the hired worker's *skill*, which cannot be observed directly until employment. However, as in the standard statistical discrimination model, each worker also has *observable characteristics* associated with their unobservable skills. At first, no firm knows precisely how to interpret a worker's characteristics to predict his skill. Therefore, firms need to learn the statistical association between characteristics and skills using data pertaining to past hiring cases (shared through, e.g., private communication, social media, and recommendation letters). Then, firms apply their statistical model to predict the skills of candidates.

Each worker belongs to a *group* that represents, for example, their gender, race, and ethnicity. We assume that the characteristics of workers who belong to different groups should be interpreted differently. This assumption is realistic. First, previous studies have revealed that underrepresented groups receive unfairly low evaluations.[1] When these eval-

[1]For example, Trix and Psenka (2003) study letters of recommendation for medical faculty and find that letters written for female applicants differ systematically from those written for male applicants. Hanna and Linden (2012) suggest that students who belong to lower caste (in India) tend receive unfairly lower exam scores. Conversely, as for teaching evaluation, MacNell et al. (2015) and Mitchell and Martin (2018) demonstrate that students rated the male identity significantly higher than the female identity. Hannák et al. (2017) study online freelance marketplaces and find that gender and race are significantly correlated

uations are used as the "observable characteristics," firms should be aware of the potential bias therein. Second, evaluations may reflect cultural differences and differences in living environments and social systems (Precht, 1998; Al-Ali, 2004). For example, firms must be familiar with the custom of writing recommendation letters to interpret letters correctly. Hence, the observable characteristics (curriculum vitae, exam score, grading report, recommendation letter, etc.) may provide very different implications even when their appearances are similar. If firms are impartial and aware of these potential biases, they should adjust the way of interpreting the characteristics by applying different statistical models to different groups.

When firms learn the statistical association from data, with some probability, the minority group is underestimated because of a large estimation error. Once the minority group is underestimated, it is difficult for a minority worker to appear to be the best candidate—even if he has the greatest skill among the candidates, the firm often dismisses this fact and tends to hire a majority worker. As long as firms only hire majority workers, society cannot learn about the minority group; thus, the imbalance persists even in the long run. We call this phenomenon *perpetual underestimation*.

We use a linear contextual bandit model to analyze the consequence of social learning. To evaluate the performance of policies, we use *regret*, one of the most popular criteria in multi-armed bandit literature. It measures the welfare loss in comparison with the first-best decision rule. When groups are ex ante identical, statistical discrimination is not only unfair but also socially inefficient: perpetual underestimation implies linear regret and vice versa. We focus on how regret grows as the total number of firms (denoted by N) increases. When regret is sublinear in N, firms make fair and efficient decisions in the long run.

We first analyze the equilibrium consequence of *laissez-faire* (no policy intervention). With a balanced population, no intervention is required to achieve no regret in the long run. That is, when the groups are ex ante symmetric and the population ratio is equal, laissez-faire results in $\tilde{O}(\sqrt{N})$ regret.[2] However, when the population ratio is unbalanced, this no longer holds, and expected regret is linear: $\tilde{\Omega}(N)$.

We study two policy interventions towards fair and efficient social learning. The first policy is a subsidy rule, based on the idea of *upper confidence bound* (UCB). UCB is an effective solution for balancing exploration and exploitation (Lai and Robbins, 1985; Auer et al., 2002). By incentivizing firms to take actions that are consistent with the recommendations of the UCB, social learning can promote no regret in the long run. The subsidy is

with worker evaluations.

[2] $\tilde{O}, \tilde{\Omega}$, and $\tilde{\Theta}$ are a Landau notations that ignore polylogarithmic factors. we often treat these factors as if it were constant because polylogarithmic factors grow very slowly ($o(N^\epsilon)$ for any exponent $\epsilon > 0$).

adjusted to the degree of information externality; thus, the total amount shrinks over time. Formally, we show that the UCB mechanism has the expected regret of $\tilde{O}(\sqrt{N})$. The subsidy required to implement the UCB mechanism is also $\tilde{O}(\sqrt{N})$.

Improving the UCB mechanism, this paper proposes a *hybrid* mechanism, which lifts affirmative actions upon the collection of a sufficiently rich data set. The hybrid mechanism takes advantage of spontaneous exploration: Once firms obtain a certain amount of data, the diversity of workers' characteristics naturally promotes learning about the minority group. The hybrid mechanism achieves $\tilde{O}(\sqrt{N})$ regret with $\tilde{O}(1)$ subsidy.

The second policy is the Rooney Rule, which requires each firm to interview at least one minority candidate as a finalist for each job opening. We analyze the effect of the Rooney Rule using a two-stage model in which firms observe additional signals of each finalist. The Rooney Rule enables minority workers to reveal the additional signal to the firm, which leaves a chance of breaking down the underestimation. Nevertheless, our evaluations of the Rooney Rule are mixed. The interviewing quota may unfairly deprive skilled majority workers of hiring opportunities, connoting reverse discrimination. This shortcoming is mitigated if the Rooney Rule is used temporarily.

2 Related Literature

Statistical discrimination Various studies have analyzed statistical discrimination both theoretically (Phelps, 1972; Arrow, 1973; Foster and Vohra, 1992; Coate and Loury, 1993; Cornell and Welch, 1996; Mailath et al., 2000) and experimentally (Neumark, 2018, is an excellent survey). We contribute to this literature by articulating a new channel of discrimination—the endogenous data imbalance and insufficiency. Similar to previous studies, we assume otherwise ex ante identical individuals from different groups to demonstrate how discrimination evolves and persists. Meanwhile, our results provide further indication that demographic minorities suffer from discrimination as an inevitable consequence of laissez-faire because the population imbalance crucially affects underestimation frequency.

More recently, several works (Bohren et al., 2019a,b; Monachou and Ashlagi, 2019) have demonstrated how misspecified beliefs about groups generate discrimination. Thus far, this literature has attributed belief misspecification to psychological biases and bounded rationality. In contrast, we demonstrate that misspecified beliefs may evolve and persist endogenously, even in the long run. Elsewhere, using a lab experiment, Dianat et al. (2020) demonstrated that the effects of affirmative action become short-lived if the action is lifted before people change their beliefs. Our hybrid mechanism provides a solution by algorithmically selecting the timing for lifting the subsidy program such that underestimation cannot persist.

The economics literature has extensively studied herding, information cascade, and social learning (e.g., Bikhchandani et al., 1992; Banerjee, 1992; Smith and Sørensen, 2000). Additionally, various papers have studied improvements to social welfare through subsidy for exploration (e.g., Frazier et al., 2014; Kannan et al., 2017) and selective information disclosure (e.g., Kremer et al., 2014; Papanastasiou et al., 2018; Immorlica et al., 2020; Mansour et al., 2020). We propose novel policy interventions to improve social learning in fairness and efficiency.

Multi-armed bandit learning A multi-armed bandit problem stems from the literature of statistics (Thompson, 1933; Robbins, 1952). This problem is driven by the question of how a single long-lived decision-maker can maximize his payoff by balancing exploration and exploitation. More recently, the machine learning community has proposed the contextual bandit framework, in which payoffs associated with "arms" (actions) depend not only on the hidden state but also additional information, referred to as "contexts" (Abe and Long, 1999; Langford and Zhang, 2008). We adopt the contextual bandit framework because context enables us to capture the diversity of worker characteristics.

Several previous studies have considered a linear contextual bandit problem and studied the performance of a "greedy" algorithm, which makes decisions myopically in accordance with the current information. Because firms take greedy actions under laissez-faire, their results are also relevant to our model. Bastani et al. (2020) and Kannan et al. (2018) have shown that a greedy algorithm leads to no regret in the long run, if the contexts are diverse enough.[3] Our contribution to this literature is a characterization of the relationship between the diversity of contexts and the rate of learning. Moreover, we show that the population ratio is crucial to the regret rate (Section 4.4), which differs strikingly from the previous results, which assumed that each arm belongs to a distinct group.

As an efficient intervention, Kannan et al. (2017) consider a contextually fair UCB-based subsidy rule. Although our subsidy policy also originates from the idea of UCB (Section 5), we establish a novel mechanism (the hybrid mechanism, Section 6) that reduces budget expenditure by utilizing spontaneous exploration.

The multi-armed bandit approach has recently been used to analyze labor markets. Bardhi et al. (2020) show that a small difference in prior beliefs about each worker's type could generate a significant difference in workers' payoffs. Elsewhere, Bergman et al. (2020) present a UCB method for screening job applicants before the interviewing stage. Their

[3] Kannan et al. (2018) indicate that the greedy algorithm achieves no regret if the decision-maker obtains a sufficient number of uniform samples in the beginning. Our simulation, included in Appendix 2, shows that our hybrid mechanism can be interpreted as a more efficient approach to collecting initial samples.

approach improves not only the quality but also the diversity of workers interviewed. This paper further demonstrates that the contextual bandit problem can usefully elucidate how discrimination occurs and persists in a hiring market.

Algorithmic fairness The literature on algorithmic fairness is growing in the field of machine learning. This literature has implicitly assumed exogenous asymmetry in worker skills and pursued the approaches to correct between-group inequality. To this end, "discrimination-aware" constraints such as demographic parity (Pedreschi et al., 2008; Calders and Verwer, 2010) and equal opportunity (Hardt et al., 2016) have been proposed, with several papers applying these constraints in the context of multi-armed bandit learning (Joseph et al., 2016) or more general sequential learning (Raghavan et al., 2018; Bechavod et al., 2019; Chen et al., 2020). In contrast, we mainly consider workers to be ex ante symmetric. When workers are symmetric, most such algorithmic fairness tools are not meaningful: For example, demographic parity, which requires firms to hire workers from each group in equal ratios, is (asymptotically) achieved by any sublinear regret decision rule even in the absence of fairness constraints.

Rooney Rule The Rooney Rule was originally introduced in the context of the hiring of National Football League senior staff[4] and later applied in many companies including Amazon[5] and Facebook.[6] However, theoretical analyses of the Rooney Rule are scarce. Kleinberg and Raghavan (2018) show that, when a recruiter is unconsciously biased against a group, the Rooney Rule not only improves the representation of that group but also leads to a higher payoff for the recruiter. To the best of our knowledge, this study (Section 7) constitutes the first attempt to demonstrate the advantage of the Rooney Rule by modeling unbiased agents.

3 Model

Basic Setting We develop a linear contextual bandit problem with myopic agents (firms). We consider a situation where N firms (indexed by $n = 1, \dots, N$) sequentially hire one

[4]The original version of the rule required league teams to interview ethnic-minority candidates for head coaching and senior football operation jobs. The rule is named after Dan Rooney, the former chairman of the league's diversity committee (Eddo-Lodge, 2017).

[5]`https://www.sec.gov/Archives/edgar/data/1018724/000119312518162552/d588714ddefa14a.htm`.

[6]Facebook COO Sheryl Sandberg stated that "*The company's 'diverse slate approach' is a sort of 'Rooney Rule,' the National Football League policy that requires teams to consider minority candidates.*" (O'Brien, 2018)

worker for each.[7] In each round n, a set of workers $I(n)$ (i.e., arms) arrives. Each worker $i \in I(n)$ takes no action, and firm n hires only one worker $\iota(n) \in I(n)$. Both firms and workers are short-lived. Upon round n ending, firm n's payoff is finalized, and all rejected workers leave the market.[8]

Each worker $i \in I$ belongs to a group $g \in G$. We assume that the population ratio is fixed: for every round n, the number of workers belonging to group g is $K_g \in \mathbb{N}$ and $K = \sum_{g \in G} K_g$. Slightly abusing the notation, we denote the group worker i belongs to by $g(i)$. Each worker i also has observable characteristics $\boldsymbol{x}_i \in \mathbb{R}^d$, with $d \in \mathbb{N}$ as their dimension. Finally, each worker i also has a skill $y_i \in \mathbb{R}$ that is not observable until worker i is hired. The characteristics and skills are random variables.

Because each firm's payoff is equal to the hired worker's skill y_i (plus the subsidy assigned to worker i as an affirmative action, if any), firms want to predict the skill y_i based on the characteristics $\boldsymbol{x}_i$. We assume that characteristics and skills are associated as $y_i = \boldsymbol{x}_i' \boldsymbol{\theta}_{g(i)} + \epsilon_i$, where $\boldsymbol{\theta}_g \in \mathbb{R}^d$ is a *coefficient parameter*, and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ i.i.d. is an unpredictable error term. We assume $||\boldsymbol{\theta}_g|| \leq S$ for some $S \in \mathbb{R}_+$, where $||\cdot||$ is the standard L2-norm. Since ϵ_i is unpredictable, $q_i \coloneqq \boldsymbol{x}_i' \boldsymbol{\theta}_{g(i)}$ is the best predictor of worker i's skill y_i.

The coefficient parameters $(\boldsymbol{\theta}_g)_{g \in G}$ are initially unknown. Hence, unless firms share information about past hires, firms are unable to predict each worker's skill y_i. We assume that all firms share information about hiring cases. Accordingly, when firm n makes a decision, in addition to the characteristics and groups of current workers $(\boldsymbol{x}_i, g(i))_{i \in I(n)}$, firm n can observe the characteristics and groups of all past candidates, $(\boldsymbol{x}_i, g(i))$ for all $i \in \bigcup_{n'=1}^{n-1} I(n')$, the past firms' decisions $(\iota(n'))_{n'=1}^{n-1}$, and the skills of previously hired workers $(y_{\iota(n')})_{n'=1}^{n-1}$.[9] We refer to all realizations of these variables as the *history* in round n, and denote it by $h(n)$. Formally, $h(n)$ is given by

$$h(n) = \left((\boldsymbol{x}_i, g(i))_{i \in I(n)}, (\boldsymbol{x}_i, g(i))_{i \in \bigcup_{n'=1}^{n-1} I(n')}, (\iota(n'))_{n'=1}^{n-1}, (y_{\iota(n')})_{n'=1}^{n-1} \right).$$

Note that, $h(n)$ does not include information about (i) the worker hired by firm n, or (ii) that worker's actual skill. This is because the notation $h(n)$ represents the information set

[7] While real-world firms are long-lived and hire multiple workers, the number of workers hired by one firm is typically much smaller than the total number of workers hired in a hiring market. Accordingly, even if we allowed firms to hire multiple (but a small number of) workers, the conclusion would not change qualitatively. Note also that various seminal papers within the social learning literature (such as Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000) have made the same assumption.

[8] This assumption is for the sake of simplicity. Since firms have no private information, the fact that a worker was previously rejected by another firm does not influence the worker's evaluation (given that the current firm can also observe the worker's characteristics).

[9] If firms only share the information about some workers, social learning would slow and underestimation would become even more severe.

firm n faces when it makes a hiring decision. We denote the set of all the possible histories in round n by $H(n)$. The firm's decision rule for hiring and the government's subsidy rule are defined as a function that maps a history to a hiring decision and the subsidy amount (described later). For notational convenience, we often omit $h(n)$.

Prediction We assume that firms are not Bayesian but *frequentists*. Hence, firms do not have a prior belief about the parameter $\boldsymbol{\theta}$ but estimate it only using the available data set. We expect that essentially the same results will be obtained with Bayesian firms (see Appendix C).

We assume that each firm predicts skill using *ridge regression* (L2-regularized least square).[10] Let $N_g(n)$ be the number of rounds at which group-g workers are hired before round n. Let $\boldsymbol{X}_g(n) \in \mathbb{R}^{N_g(n)\times d}$ be a matrix that lists the characteristics of group-g workers hired by round n: each row of $\boldsymbol{X}_g(n)$ corresponds to $\{\boldsymbol{x}_{\iota(n')} : \iota(n') = g\}_{n'=1}^{n-1}$. Likewise, let $Y_g(n) \in \mathbb{R}^{N_g(n)}$ be a vector that lists the skills of group-g workers hired by round n: each element of $Y_g(n)$ corresponds to $\{y_{\iota(n')} : \iota(n') = g\}_{n'=1}^{n-1}$. We define $\boldsymbol{V}_g(n) \coloneqq (\boldsymbol{X}_g(n))'\boldsymbol{X}_g(n)$. For a parameter $\lambda > 0$, we define $\bar{\boldsymbol{V}}_g(n) = \boldsymbol{V}_g(n) + \lambda \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. Firm n estimates the parameter as follows:

$$\hat{\boldsymbol{\theta}}_g(n) \coloneqq (\bar{\boldsymbol{V}}_g(n))^{-1}(\boldsymbol{X}_g(n))'Y_g(n).$$

Firm n predicts worker i's skill q_i, while substituting the true predicted skill $\boldsymbol{\theta}_g$ with *estimated skill* $\hat{\boldsymbol{\theta}}_g(n)$: $\hat{q}_i(n) \coloneqq \boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)}(n)$. Hence, $\hat{q}_i(n)$ and $\hat{\boldsymbol{\theta}}_g(n)$ depend on the history $h(n)$. The *ordinary least squares* (OLS) estimator corresponds to the ridge estimator with $\lambda = 0$. We use the ridge estimator instead of the OLS estimator to stabilize the small-sample inference. For example, for some history, $\boldsymbol{V}_g(n)$ may not have full rank, and the OLS estimator may not be well-defined. Even for such histories, the ridge estimator (with $\lambda > 0$) is always well-defined.

For analytical tractability, we assume that for the first $N^{(0)}$ rounds, each firm n must hire from a pre-specified group, g_n. We refer to the first $N^{(0)}$ rounds as the *initial sampling phase*. Let $N_g^{(0)} \coloneqq \sum_{n=1}^{N^{(0)}} \mathbf{1}[g_n = g]$ as the data size of initial sampling for group g, where $\mathbf{1}[\mathcal{A}] = 1$ if event $\mathcal{A}$ holds or 0 otherwise. The initial sampling phase is exogenous. That is, we ignore the incentives and payoffs of firms and assume that the characteristics $\boldsymbol{x}$ of the hired candidate constitute an i.i.d. sample of the corresponding group. We analyze mechanism, social welfare, and budget after round $n > N^{(0)}$.

[10] See Kennedy (2008), for example.

Mechanism In addition to worker skill, firms are also concerned about subsidies. We assume that firm preferences are risk-neutral and quasi-linear. Hence, if firm n hires worker i, its payoff (von-Neumann–Morgenstern utility) is given by $y_i + s_i$, where $s_i \in \mathbb{R}_+$ denotes the amount of the subsidy assigned to worker i.

In the beginning of the game, the government commits to a *subsidy rule* $s_i(n, \cdot) : H(n) \to \mathbb{R}_+$, which maps a history to a subsidy amount. Hence, once a history $h(n)$ is specified, firm n can identify the subsidy assigned to each worker $i \in I(n)$. Firm n attempts to maximize

$$\mathbb{E}\left[y_i + s_i(n; h(n)) \middle| h(n) \right] = \hat{q}_i(n; h(n)) + s_i(n; h(n)).$$

Firm n's *decision rule* $\iota(n, \cdot) : H(n) \to I(n)$ specifies the worker that firm n hires after history $h(n)$. We say that, a decision rule ι is *implemented* by a subsidy rule s_i if for all n and $h(n)$, we have

$$\iota(n; h(n)) = \arg\max_{i \in I(n)} \left\{ \hat{q}_i(n; h(n)) + s_i(n; h(n)) \right\}. \tag{1}$$

Throughout this paper, any ties are broken arbitrarily. We call a pair of a decision rule and subsidy rule a *mechanism*. We often drop $h(n)$ from the input of decision rule ι when it does not cause confusion.

Social Welfare We measure social welfare by the smallness of *regret*, which is the standard measure for evaluating the performance of algorithms in multi-armed bandit models:

$$\mathrm{Reg}(N) \coloneqq \sum_{n=N^{(0)}+1}^{N} \left\{ \max_{i \in I(n)} q_i - q_{\iota(n)} \right\}.$$

Since ϵ_i is unpredictable, it is natural to evaluate the performance of the algorithm (or the equilibrium consequence of the policy intervention) by comparing it with q_i. If the parameter $(\boldsymbol{\theta}_g)_{g \in G}$ were known, each firm could easily calculate q_i for each worker i and hire the best worker, $\iota(n) = \arg\max_{i \in I(n)} q_i$. In this case, regret would be zero. The goal of the policy design is to establish a mechanism that minimizes the expected regret $\mathbb{E}[\mathrm{Reg}(N)]$, where the expectation is taken on a random draw of workers. This aim is equivalent to maximizing the sum of the skill of workers hired.

Budget Some of the policies we study incentivize exploration through subsidies. The total budget required by a subsidy rule is also an important policy concern. The total amount of the subsidy is given by $\mathrm{Sub}(N) \coloneqq \sum_{n=N^{(0)}+1}^{N} s_{\iota(n)}(n)$.

4 Laissez-Faire

This section analyzes the equilibrium under *laissez-faire*; that is, the consequence of social learning in the absence of policy intervention.

Definition 1 (Laissez-Faire)**.** The *laissez-faire decision rule* always selects the worker who has the greatest estimated skill, i.e., $\iota(n) = \arg\max_{i \in I(n)} \hat{q}_i(n)$. This decision rule is implemented by the *laissez-faire subsidy rule*, which provide no subsidy $s_i(n) = 0$ after any history.

Laissez-faire makes no intervention. Each firm hires the worker with the greatest estimated skill, as predicted by the current data set. The multi-armed bandit literature refers to the laissez-faire decision rule as the *greedy algorithm*.

4.1 Symmetry and Diverse Characteristics

To articulate a failure of social learning, we make three assumptions. First, we focus on the two-group case.

Assumption 1 (Two Groups)**.** The population comprises two groups $G = \{1, 2\}$.

When we consider asymmetric equilibria, we refer to group 1 as the majority (dominant) group and group 2 as the minority (discriminated-against) group. The two-group assumption enables the elucidation of how the minority group is discriminated against by the majority group.

Second, we assume that groups are symmetric.

Assumption 2 (Symmetric Groups)**.** The characteristics of all groups are identical, and the coefficient parameters are the same across the groups. That is, a probability distribution F such that for all $i \in I$, $\boldsymbol{x}_i \sim F$, and there exists $\boldsymbol{\theta} \in \mathbb{R}^d$ such that, for all $g \in G$, $\boldsymbol{\theta}_g = \boldsymbol{\theta}$.

Note that although we assume that groups are symmetric, firms do not know the true parameters, and therefore, apply different statistical models to different groups. That is, even though the *true* coefficients are identical ($\boldsymbol{\theta}_g = \boldsymbol{\theta}'_g$ for all $g, g' \in G$), firms estimate them separately; thus, the values of the *estimated* coefficients are typically different ($\hat{\boldsymbol{\theta}}_g(n) \neq \hat{\boldsymbol{\theta}}_{g'}(n)$ for $g \neq g'$).

Although Assumption 2 is unrealistic (because the characteristics should evidently be interpreted differently), it is useful for elucidating how laissez-faire nourishes statistical discrimination. Under Assumption 2, agents are ex ante identical (as assumed in Arrow, 1973;

Foster and Vohra, 1992; Coate and Loury, 1993; Moro and Norman, 2004). Hence, the differences we observe in the equilibrium consequence are based entirely on social learning.

When workers are symmetric, discrimination is not only unfair but also inefficient. That is, given that there is no fundamental difference between groups, underestimation always produces larger regret (i.e., welfare loss). If we instead consider an asymmetric model, then fairness and efficiency would often conflict, and no regret would not indicate fair decision-making.

Third, we assume that characteristics are normally distributed, and therefore, the distribution is non-degenerate. This assumption captures the diversity of workers.

Assumption 3 (Normally Distributed Characteristics)**.** For every candidate i,

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{xg(i)}, \sigma_{xg(i)}^2 \boldsymbol{I}_d),$$

where $\boldsymbol{\mu}_{xg} \in \mathbb{R}^d$ and $\sigma_{xg} \in \mathbb{R}_{++}$ for every $g \in G$. We also denote $\boldsymbol{x}_i = \boldsymbol{\mu}_{xg(i)} + \boldsymbol{e}_{xi}$ to highlight the noise term $\boldsymbol{e}_{xi}$.

Note that when we have both Assumptions 2 and 3, then there exist $\boldsymbol{\mu}_x, \sigma_x$ such that $\boldsymbol{\mu}_{xg} = \boldsymbol{\mu}_x$ and $\sigma_{xg} = \sigma_x$ for all $g \in G$. Hence, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \sigma_x^2 \boldsymbol{I}_d)$ for all i.

4.2 Perpetual Underestimation

To determine whether or not social learning incurs linear expected regret, it is useful to check whether it results in *perpetual underestimation* with a significant probability.

Definition 2 (Perpetual Underestimation)**.** A group g_0 is *perpetually underestimated* if, for all $n > N^{(0)}$, we have $g(\iota(n)) \neq g_0$.

That is, when group g_0 is perpetually underestimated, no worker from group g_0 is hired after the initial sampling phase. If social learning generates perpetual underestimation with a significant probability, then linear expected regret often results. In particular, under Assumption 2, perpetual underestimation against any group $g \in G$ implies that firms fail to hire at least $(K_g/K)\left(N - N^{(0)}\right)$ best candidate, which is linear in N. Hence, the constant probability of perpetual underestimation (independent of N) precipitates linear expected regret.

Perpetual underestimation is not only inefficient but also unfair. Perpetual underestimation leads to a candidate belonging to an underestimated group not being hired, despite the groups being symmetric. This outcome happens because society cannot accurately predict the skills of minority workers due to insufficient data. Hence, in our model, perpetual discrimination can be regarded as a form of statistical discrimination.

4.3 Sublinear Regret with Balanced Population

This section analyzes the case of only one candidate arriving from each group during each period. The contextual variation implicitly urges firms to explore all the groups with some frequency. Consequently, laissez-faire has sublinear regret, implying that statistical discrimination is eventually resolved.

Theorem 1 (Sublinear Regret with a Balanced Population)**.** Suppose Assumptions 1, 2, and 3. Suppose also that $K_g = 1$ for $g = 1, 2$. Then, expected regret is bounded as

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{LF}}(N)] \leq C_{\mathrm{bal}}\sqrt{N}$$

where C_{bal} is a $\tilde{O}(1)$ factor that depends on model parameters.

Letting $\mu_x = ||\boldsymbol{\mu}_x||$, the factor C_{bal} is inverse proportional[11] to $\Phi^c(\mu_x/\sigma_x)$, which approximately scales as $\exp(-(\mu_x/\sigma_x)^2/2)$.

Proof. See Appendix 1. The explicit form of C_{bal} is also shown there. $\square$

To prove Theorem 1, we characterize the condition with which underestimation is spontaneously resolved. Let indices i_1 and i_2 denote the majority candidate and the minority candidate. With a constant (i.e., independent of N) probability, the minority group is underestimated (i.e, $\hat{\boldsymbol{\theta}}_2(n)$ is misestimated in such that $\boldsymbol{x}_{i_2}\hat{\boldsymbol{\theta}}_2(n) \ll \boldsymbol{x}_{i_2}\boldsymbol{\theta}_2$ often occurs) in early rounds due to a bad realization of the error term. Even in such a case, there is some probability of the minority candidate being hired. Since characteristics are diverse (i.e., $\sigma_x > 0$), with some probability, the majority candidate i_1 is not very good (i.e., $x_{i_1}\hat{\boldsymbol{\theta}}_1(n) \approx x_{i_1}\boldsymbol{\theta}_1$ is small). In such a round, $x_{i_1}\hat{\boldsymbol{\theta}}_1(n) < x_{i_2}\hat{\boldsymbol{\theta}}_2(n)$ holds despite group 2 being underestimated, and the minority candidate i_2 is hired. In such a case, firms update their belief about the minority, leading to a resolution of underestimation. Such events more frequently when workers have more diverse characteristics, i.e., μ_x/σ_x is large.

As anticipated by the theory of least squares, the standard deviation of $\hat{\boldsymbol{\theta}}_g(n)$ is proportional to $(\bar{\boldsymbol{V}}_g(n))^{-1/2}$, and we demonstrate that its diameter $(\lambda_{\min}(\bar{\boldsymbol{V}}_g(n)))^{-1/2}$ shrinks as $\tilde{O}(1/\sqrt{n})$, where $\lambda_{\min}$ is the minimum eigenvalue of a matrix. The regret per error is defined by this quantity, with the total regret being $\tilde{O}(\sum_{n \leq N}(1/\sqrt{n})) = \tilde{O}(\sqrt{N})$.

Theorem 1 indicates that statistical discrimination is resolved spontaneously when candidate variation is large. At a glance, this appears to contradict widely known results that state laissez-faire (greedy) may lead to suboptimal results in bandit problems due to underexploration. Selfish firms do not want to experiment with groups at their own risk. However,

[11]Here, $\Phi(x)$ is the cdf of the standard normal distribution and $\Phi^c(x) = 1 - \Phi(x)$.

the variation in characteristics naturally incentivizes selfish agents to explore the underestimated group, and therefore, with some additional conditions, we can bound the probability of perpetual underestimation.

Remark 1. Theorem 1 shares certain intuitions with the previous research (Kannan et al., 2018; Bastani et al., 2020) demonstrating that the variation in contexts (characteristics) improves the performance of the greedy algorithm (laissez-faire) in contextual multi-armed bandit problems. However, in contrast to Kannan et al. (2018), our theorem makes no assumptions regarding the length of the initial sampling phase. Theorem 1 in Bastani et al. (2020) corresponds to our paper's Theorem 1, and we further characterize the factor of the regret as a function of μ_x/σ_x rather than the diameter of the characteristics.

4.4 Large Regret with Unbalanced Population

While Theorem 1 implies that statistical discrimination is spontaneously resolved in the long run, it crucially relies on one unrealistic assumption—the balanced population ratio. In many real-world problems, the population ratio is unbalanced, and the discriminated group is often a demographic minority in the relevant market. We indeed find that the population ratio crucially impacts the equilibrium consequence under laissez-faire.

Theorem 2 (Substantial Regret with Unbalanced Populations)**.** Suppose Assumptions 1, 2, and 3. Suppose also that $K_2 = 1$ and $d = 1$. Let $K_1 > \log_2 N$. Then, under the laissez-faire decision rule, group 2 is perpetually underestimated with a probability of at least $C_{\mathrm{imb}} = \tilde{\Theta}(1)$. Accordingly, the expected regret associated with the laissez-faire decision rule is

$$\mathbb{E}\left[\mathrm{Reg}^{\mathrm{LF}}(N)\right] \geq \frac{C_{\mathrm{imb}}(N - N^{(0)})}{K} = \tilde{\Omega}(N).$$

Proof. See Appendix 2. The explicit form of C_{imb} is shown in Eq. (38). □

In the proof of Theorem 2, we evaluate the probability that the following two events occur: (i) $\hat{\theta}_2$ is underestimated ($\hat{\theta}_2/\theta = o(1)$), and (ii) The characteristics and skills of the hired majority workers are not very bad throughout rounds (i.e., $\max_{i:g(i)=1} x_i\hat{\theta}_1 \geq c\mu_x\theta$ for some constant $c > 0$). The probability of (i) is polylogarithmic to N (i.e., $\tilde{\Theta}(1)$) and the probability that (ii) consistently holds for all the rounds $n = N^{(0)} + 1, \ldots, N$ is polylogarithmic if $K_1 > \log_2 N$. When both (i) and (ii) occur, we always have $\max_{i \in I(n)\setminus\{i_2(n)\}} x_i\hat{\theta}_1 > x_{i_2(n)}\hat{\theta}_2$ (where $i_2(n)$ is the unique minority candidate of round n); thus, thus the minority worker is never hired.

Theorem 2 indicates that we should not be too optimistic about the consequence of laissez-faire. An unbalanced population ratio naturally favors the majority group. Once the minority group is underestimated and the majority candidate pool is reasonably large, then the minority group is afforded no hiring opportunity, perpetuating underestimation. We emphasize that only a small imbalance in the population ratio ($\log_2 N$ majority worker) suffices to induce linear regret. This insight applies to many real-world problems because unbalanced populations are commonplace.

Remark 2 (Endogenous Skills). In our framework, statistical distribution is attributed entirely to misinformation because skill distribution is assumed to be exogenous. If we additionally incorporate the choice of the human capital investments (Foster and Vohra, 1992; Coate and Loury, 1993), the misinformation naturally discourages minority workers from improving their skill. Therefore, if skill is endogenous, the welfare loss and inequality triggered by social learning would be even worse.

5 The Upper Confidence Bound Mechanism

Section 4 has discussed the equilibrium consequences of laissez-faire. We observed that an unbalanced population ratio leads to a substantial probability of underestimation being perpetuated. Policy intervention is demanded to improve social welfare and the fairness of the hiring market.

This section proposes a subsidy rule to resolve underestimation. We employ the idea of the *upper confidence bound* (UCB) algorithm (Lai and Robbins, 1985; Auer et al., 2002), which has widely been used in the literature on the bandit problem. The UCB algorithm balances exploration and exploitation by developing a confidence interval for the true reward and evaluating each arm's performance according to its upper confidence bound to achieve this balance. Firms are generally unwilling to follow the UCB's recommendations voluntarily; therefore, the government needs to provide a subsidy to incentivize firms to hire a candidate with the greatest UCB index. This section establishes a UCB-based subsidy rule and evaluates its performance.

The adaptive selection of candidates based on history can induce some bias, meaning the standard confidence bound no longer applies. To overcome this issue, we use martingale inequalities (Peña et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011). We here introduce the confidence interval for the true coefficient parameter, $(\boldsymbol{\theta}_g)_{g \in G}$.

Definition 3 (Confidence Interval)**.** Given the group g's collected data matrix $\bar{\boldsymbol{V}}_g(n)$, the *confidence interval* of group g's coefficient parameter $\boldsymbol{\theta}_g$ is given by

$$\mathcal{C}_g(n)$$

$$= \left\{ \bar{\boldsymbol{\theta}}_g \in \mathbb{R}^d : \left\| \bar{\boldsymbol{\theta}}_g - \hat{\boldsymbol{\theta}}_g(n) \right\|_{\bar{\boldsymbol{V}}_g(n)} \leq \sigma_\epsilon \sqrt{d \log \left(\frac{\det(\bar{\boldsymbol{V}}_g(n))^{1/2} \det(\lambda \boldsymbol{I}_d)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right\}$$

where $||\boldsymbol{v}||_{\boldsymbol{A}} = \sqrt{\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}}$ for a d-dimensional vector $\boldsymbol{v}$ and $d \times d$ matrix $\boldsymbol{A}$.

Abbasi-Yadkori et al. (2011) study the property of this confidence interval, and they prove that the true parameter $\boldsymbol{\theta}_g$ lies in $\mathcal{C}_g(n)$ with probability $1-\delta$ (Lemma 17). By choosing a sufficiently small δ,[12] it is "safe" to assess that worker i's skill is *at most*

$$\tilde{q}_i(n) \coloneqq \max_{\bar{\boldsymbol{\theta}}_{g(i)} \in \mathcal{C}_{g(i)}(n)} \boldsymbol{x}_i' \bar{\boldsymbol{\theta}}_{g(i)}.$$

We call $\tilde{q}_i(n)$ the *UCB index* of worker i's skill. Intuitively, $\tilde{q}_i(n)$ is worker i's skill in the most optimistic scenario. The confidence interval $\mathcal{C}_g(n)$ shrinks as we obtain more data about group g. Hence, the UCB index $\tilde{q}_i(n)$ converges to true predicted skill $q_i(n)$ as the size of the data set increases.

Definition 4 (UCB Decision Rule)**.** The UCB decision rule selects the worker with the greatest UCB index; i.e.,

$$\iota(n) = \arg\max_{i \in I(n)} \tilde{q}_i(n). \tag{2}$$

The UCB index $\tilde{q}_i(n)$ is close to the pointwise estimate $\hat{q}_i(n)$ when society has rich data about group $g(i)$, because $\mathcal{C}_{g(i)}(n)$ is small in such cases. However, when information about group $g(i)$ is insufficient, $\tilde{q}_i(n)$ is much larger than $\hat{q}_i(n)$, because the firm is unsure about the true skill of worker i and $\mathcal{C}_{g(i)}(n)$ is large. In this sense, the UCB decision rule offers affirmative actions towards underexplored groups.

The subsidy amount is proportional to the uncertainty surrounding the candidate's characteristics, which is represented by the confidence interval $\mathcal{C}_g(n)$ for $g = g(i)$. The magnitude of the confidence interval $\mathcal{C}_g(n)$ is inverse proportional to $\bar{\boldsymbol{V}}_g(n) = (\boldsymbol{X}_g(n))'\boldsymbol{X}_g(n) + \lambda \boldsymbol{I}_d$.[13] Hence, if the data $\boldsymbol{V}_g(n)$ do not vary substantially for a particular dimension of $\boldsymbol{x}_i$, then that dimension's prediction can be inaccurate. In such cases, the UCB decision rule recommends hiring a candidate that contributes to increasing that dimension's data. For example, when a candidate possesses skill previous hires do not, then the candidate's UCB index tends to become large.

[12] We typically choose $\delta = 1/N$ to make the confidence interval asymptotically correct in the limit of $N \to \infty$.

[13] The standard OLS has a confidence bound of the form $\boldsymbol{\theta}_g - \hat{\boldsymbol{\theta}}_g(n) \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{V}_g^{-1}(n))$ and thus $|\boldsymbol{\theta}_g - \hat{\boldsymbol{\theta}}_g(n)| \sim \sigma_\epsilon \boldsymbol{V}_g^{-1/2}(n)$. The price of adaptivity causes the martingale confidence bound $\mathcal{C}_g(n)$ to be larger than the OLS confidence bound for two factors: (i) $\sqrt{d}$ factor, and (ii) $\sqrt{\log(\det(\bar{\boldsymbol{V}}_g(n)))}$ factor. As discussed in Xu et al. (2018), the $\sqrt{d}$ factor unnecessarily overestimates the confidence bound in most cases.

The UCB decision rule efficiently balances exploration and experimentation. Accordingly, it produces sublinear regret in general environments.

Theorem 3 (Sublinear Regret of UCB)**.** Suppose Assumptions 3. Let $\mathrm{Reg}^{\mathrm{UCB}}$ be the regret from the UCB decision rule. Let $\lambda \geq \max(1, L^2)$. Then, by choosing $\delta = 1/N$, regret under the UCB decision rule is bounded as

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{UCB}}(N)] \leq C_{\mathrm{ucb}}\sqrt{N},$$

where C_{ucb} is $\tilde{O}(1)$ as a function of N.

Proof. See Appendix 3. The explicit form of C_{ucb} is shown in Eq. (42). □

$\tilde{O}(\sqrt{N})$ regret is the optimal rate for these sequential optimization problems under partial feedback (Chu et al., 2011). Hence, Theorem 3 states that the UCB decision rule effectively prevents perpetual underestimation and is asymptotically efficient. Note also that, differing from the case of laissez-faire, where the factor depends on the variation of the context (Theorem 1), Theorem 3 provides a reasonably small regret bound even when σ_x is very small.

To implement the UCB decision rule, we need to satisfy the firms' obedience condition (1) in conjuction with the UCB's decision rule (2). This paper focuses on the most straightforward subsidy rules.

Definition 5 (UCB Index Subsidy Rule)**.** The *UCB index subsidy rule* s subsidizes firm n to hire worker i who arrives by

$$s_i(n; h(n)) = \tilde{q}_i(n; h(n)) - \hat{q}_i(n; h(n)).$$

The UCB index subsidy rule aligns each firm's incentive with the maximization of the UCB index, thereby incentivizing firms to follow the UCB decision rule.

Theorem 4 (Sublinear Subsidy of the UCB Index Subsidy Rule)**.** Under the same assumptions as Theorem 3, the amount of the subsidy required by the UCB index subsidy rule is bounded as

$$\mathbb{E}[\mathrm{Sub}^{\mathrm{UCB\text{-}I}}(N)] \leq C_{\mathrm{ucb}}\sqrt{N}.$$

where C_{ucb} is an $\tilde{O}(1)$ factor that is the same as Theorem 3.

Proof. See Appendix 4. □

Remark 3. The UCB index subsidy rule is an *index policy* in the sense that the subsidy amount is independent of the information about rejected workers. The UCB index subsidy rule demands the smallest budget among all index policies implementing the UCB decision rule. We also consider non-index policies and compare the budget. See Appendix A for details.

6 The Hybrid Mechanism

The previous section demonstrated that the UCB mechanism effectively prevents perpetual underestimation and achieves sublinear regret in general environments. However, the UCB mechanism has one drawback: it continues subsidies in perpetuity. Even for a large n, there remains a gap between estimated skill $\hat{q}_i(n)$ and the UCB index $\tilde{q}_i(n)$. This is undesirable for several reasons. First, introducing a permanent policy is often more politically difficult than introducing a temporary policy. Second, a long-term distribution of subsidies tends to increase the required budget. Third, in addition to the subsidy itself, the permanent allocation of the subsidy features (unmodeled) administrative costs.

To overcome these limitations, we propose the *hybrid mechanism*, which initially uses the UCB mechanism but switches to laissez-faire by terminating the subsidy at some point. We abandon the UCB phase upon receiving sufficient minority-group data to induce spontaneous exploration. Similar to the UCB mechanism, our hybrid mechanism has $\tilde{O}(\sqrt{N})$ regret, and its expected total subsidy amount is $\tilde{O}(1)$ (while the UCB mechanism needs $\tilde{O}(\sqrt{N})$ subsidy).

The construction of the hybrid mechanism is as follows. Let $s_i^{\text{U-I}}(n) = \tilde{q}_i(n) - \hat{q}_i(n)$ be the size of the confidence bound. Note that, $s_i^{\text{U-I}}(n)$ corresponds to the amount of the subsidy allocated by the UCB index subsidy rule (Definition 5). The *hybrid index* $\tilde{q}_i^{\text{H}}$ is defined as

$$\tilde{q}_i^{\text{H}}(n; h(n)) := \begin{cases} \tilde{q}_i(n; h(n)) & \text{if } s_i^{\text{U-I}}(n; h(n)) > a\sigma_x ||\hat{\boldsymbol{\theta}}_{g(i)}(n; h(n))||, \\ \hat{q}_i(n; h(n)) & \text{otherwise,} \end{cases} \tag{3}$$

where $a \geq 0$ is the mechanism's parameter.

The hybrid index is literally a "hybrid" of estimated skill $\hat{q}_i(n)$ and the UCB index $\tilde{q}_i(n)$. If the difference between the UCB index and estimated skill surpasses the threshold (i.e., $s_i^{\text{U-I}}(n) > a\sigma_x ||\hat{\boldsymbol{\theta}}_{g(i)}(n)||$), the hybrid index is equal to the UCB index $\tilde{q}_i(n)$. The confidence bound $|\tilde{q}_i(n) - \hat{q}_i(n)|$ is large when society has insufficient knowledge about group $g(i)$, which is typically the case during early stages of the game. Once this gap falls below the threshold (i.e., $s_i^{\text{U-I}}(n) \leq a\sigma_x ||\hat{\boldsymbol{\theta}}_{g(i)}(n)||$), then the hybrid index switches to the estimated skill $\hat{q}_i(n)$. The hybrid decision rule is defined as the rule that hires the greatest hybrid index.

Definition 6 (Hybrid Decision Rule)**.** The *hybrid decision rule* selects the worker who has the greatest hybrid index; i.e.,

$$\iota^{\mathrm{H}}(n; h(n)) = \arg\max_{i \in I(n)} \tilde{q}_i^{\mathrm{H}}(n; h(n)).$$

Since the hybrid decision rule is a hybrid of the UCB decision rule and the laissez-faire decision rule, it can be implemented by mixing the laissez-faire subsidy rule and the UCB index subsidy rule.

Definition 7 (Hybrid Index Subsidy Rule)**.** Let $s_i^{\text{U-I}}$ be the UCB index subsidy rule. The *hybrid index subsidy rule* $s^{\text{H-I}}$ is defined by

$$s_i^{\text{H-I}}(n; h(n)) := \begin{cases} s_i^{\text{U-I}}(n; h(n)) & \text{if } s_i^{\text{U-I}}(n; h(n)) > a\sigma_x ||\hat{\boldsymbol{\theta}}_{g(i)}(n; h(n))||, \\ 0 & \text{otherwise.} \end{cases}$$

The following theorems characterize the regret and the total subsidies associated with the hybrid mechanism.

Theorem 5 (Performance of the Hybrid Mechanism)**.** Suppose Assumptions 1, 2, and 3. Then, by choosing $\delta = 1/N$, regret associated with the hybrid decision rule ι^{H} is bounded as

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{H}}(N)] \leq C_{\mathrm{hyb}} \sqrt{N}$$

where C_{hyb} is a factor that is $\tilde{O}(1)$ to N. Furthermore, for any $a > 0$, the total amount of the subsidy under the hybrid index subsidy rule ($\mathrm{Sub}^{\text{H-I}}$) is bounded as

$$\mathbb{E}[\mathrm{Sub}^{\text{H-I}}(N)] \leq C_{\text{hyb-sub}}.$$

where $C_{\text{hyb-sub}}$ is a factor that is $\tilde{O}(1)$ to N.

Proof. See Appendix 5. The explicit form of C_{hyb} is shown in Eq. (51). The explicit form of $C_{\text{hyb-sub}}$ is shown in Eq. (58). □

Theorem 5 states that the order of the regret under the hybrid decision rule is the same as the original UCB, and the subsidy amount is reduced to $\tilde{O}(1)$ (with respect to N). This is a substantial improvement from the UCB mechanism, which requires the $\tilde{O}(\sqrt{N})$ subsidy.

The threshold for switching from the UCB mechanism to laissez-faire is crucial for guaranteeing the performance of the hybrid mechanism. Our threshold, $a\sigma_x||\hat{\boldsymbol{\theta}}(n)||$, is determined such that the hybrid decision rule ι^{H} satisfies *proportionality*, a new concept that this paper establishes. We evaluate the performance of the hybrid decision rule by comparing it with the UCB decision rule. However, neither decision rule dominates the other because different decision rules generate different histories and data. To overcome this difficulty, we prove that the amount of exploration exerted by the hybrid decision rule is proportional to the UCB decision rule: there exists a constant $c > 0$ such that when the UCB rule ι^{U} hires worker i with a probability of p_i, then the hybrid rule ι^{H} hires worker i with a probability of at least cp_i given the same history. The proportionality guarantees that the hybrid rule resolves underestimation and secures the expected regret of $\tilde{O}(\sqrt{N})$. The formal statement of the proportionality appears in Lemma 26 in Appendix 5.

Remark 4 (Dependence on Parameter a)*.* There is a tradeoff between regret and subsidy. The constant on the top of regret is $\exp(a^2/4)$, which is increasing in a. By contrast, the constant on the top of subsidy is $\exp(3a^2/4)/a^2$, which goes to infinity as $a \to 0$. To balance this tradeoff, the government should select an appropriate value for a. Our simulations adopt $a = 1$ (see Section 8).

7 Interviews and the Rooney Rule

Although subsidy rules effectively resolve statistical discrimination, they are often difficult to implement in practice. This section articulates the advantages and disadvantages of the *Rooney Rule*, a regulation that requires each firm to invite at least one candidate from each group to an on-site interview. The Rooney Rule is easier to implement because it requires neither a subsidy nor meeting a hiring quota.

To incorporate the additional information firms acquire through the interview, we modify the model as follows. In the modified model, each round n comprises two stages. At the first stage, firm n observes the characteristics $\boldsymbol{x}_i$ of each arriving agent $i \in I(n)$. Based on $\boldsymbol{x}_i$, firm n selects a shortlist of *finalists* $I^F(n) \subseteq I(n)$, where $|I^F(n)| = K^F$ for some constant $K^F \in \mathbb{N}$. At the second stage, by interviewing finalists, firm n observes an additional signal η_i for each finalist i (as assumed in Kleinberg and Raghavan, 2018). Firm n predicts each finalist i's skill from the characteristics $\boldsymbol{x}_i$ and the additional signal η_i, and hires one worker from the set of finalists, $\iota(n) \in I^F(n)$. Firms are not allowed to hire a worker not selected as a finalist. After the firm's decision, the skill of the hired worker $y_{\iota(n)}$ is publicly disclosed.

We assume the following linear relationship between skill y_i and observable variables $\boldsymbol{x}_i$: $y_i = \boldsymbol{x}_i'\boldsymbol{\theta}_{g(i)} + \eta_i + \epsilon_i$ The "noise" term comprises two variables: η_i and ϵ_i. η_i is revealed as an additional signal when the firm chooses i as a finalist. However, ϵ_i remains unpredictable even after the interview.

For analytical tractability, we make the following two assumptions.

Assumption 4 (Two Finalists)**.** Each firm can invite only two finalists; i.e., $K^F = 2$.

Assumption 5 (Normal Additional Signals)**.** Each additional signal that a finalist reveals follows a normal distribution, $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$, i.i.d.

Remark 5. If $\sigma_\eta = 0$, then the two-stage model is equivalent to the one-stage model that we have considered in the previous sections.

7.1 Failure of Laissez-Faire in the Two-Stage Model

This subsection analyzes the performance of laissez-faire in this two-stage setting. The result is analogous to the one-stage case (Theorem 2) : laissez-faire often falls in perpetual underestimation, and therefore, has linear regret.

First, we define regret. As in the one-stage model, the benchmark is the first-best decision rule, which is the rule firms would apply if the coefficient parameter $\boldsymbol{\theta}$ were known. Clearly, the first-best decision rule would greedily invite top-K^F workers in terms of q_i to the final interview. We denote this set of finalists chosen by the first-best decision rule in round n by $\bar{I}^F(n)$. Formally, $\bar{I}^F(n)$ is obtained by solving the following problem:

$$\bar{I}^F(n) = \operatorname*{arg\,max}_{I' \subseteq I(n)} \sum_{i \in I'} q_i \quad \text{s.t. } |I'| = K^F.$$

After that, the first-best decision rule would observe the realization of η_i for $i \in \bar{I}^F(n)$, and then hires the worker i who has the greatest skill predictor: $q_i + \eta_i$. *Unconstrained two-stage regret* (U2S-Reg) is defined as the loss compared with this first-best decision rule. (This type of regret is named "unconstrained" because we later introduce an alternative definition.)

Definition 8 (Unconstrained Two-Stage Regret)**.** In the two-stage hiring model, the *unconstrained two-stage regret* U2S-Reg of decision rule ι is defined as follows:

$$\text{U2S-Reg}(N) = \sum_{n=1}^{N} \left\{ \max_{i \in \bar{I}^F(n)} (q_i + \eta_i) - \left(q_{\iota(n)} + \eta_{\iota(n)}\right) \right\}.$$

Under laissez-faire, the optimal strategy of firm n is to choose candidates greedily based on their estimated skills, i.e.,

$$I^F(n) = \operatorname*{arg\,max}_{I' \subseteq I(n)} \sum_{i \in I'} \hat{q}_i(n) \quad \text{s.t. } |I'| = K^F.$$

After observing the realization of the additional signals η_i, firm n selects the candidate who has the greatest estimated skill:

$$\iota(n) = \operatorname*{arg\,max}_{i \in I^F(n)} \{\hat{q}_i(n) + \eta_i\}.$$

Even in the two-stage model, laissez-faire has linear regret when the population ratio is unbalanced.

Theorem 6 (Failure of Laissez-Faire in the Two-Stage model)**.** Suppose Assumptions 1, 2, 3, 4, and 5. Suppose also that $K_2 = 1$ and $d = 1$. Let $K_1 - \log_2(K_1 + 1) > \log_2 N$. Then, under the laissez-faire decision rule, group 2 is perpetually underestimated with the probability $\tilde{\Omega}(1)$. Accordingly, the expected regret associated with the laissez-faire decision rule is

$$\mathbb{E}\left[\text{U2S-Reg}^{\text{LF}}(N)\right] = \tilde{\Omega}(N).$$

Proof. See Appendix 6. □

The proof idea of Theorem 6 is as follows. Under laissez-faire, each firm n interviews the two finalists with the greatest estimated skills, $\hat{q}_i(n)$. If both finalists belong to the majority group, then minority candidates are never hired, regardless of the η_i for each finalist. By evaluating the probability that both finalists are majority candidates, we derive the probability of perpetual underestimation. Thus, even in a two-stage setting, the laissez-faire decision has linear regret (when the population ratio is unbalanced).

7.2 The Rooney Rule and Exploration

Given laissez-faire does not mitigate perpetual underestimation, desirable policy intervention is necessary.

Definition 9 (Rooney Rule)**.** In the two-stage hiring model, the *Rooney Rule* requires each firm n to select at least one finalist from every group $g \in G$; i.e., for every n and every $g \in G$, $I^F(n)$ must satisfy

$$\left|\left\{i \in I^F(n) \mid g(i) = g\right\}\right| \geq 1. \tag{4}$$

Under Assumption 1 and 4, each firm interviews one majority candidate and one minority candidate. To analyze how the Rooney Rule resolves statistical discrimination, we introduce a weaker notion of regret, *constrained two-stage regret*.

Definition 10 (Constrained Two-Stage Regret)**.** In the two-stage hiring model, the *constrained two-stage regret* (C2S-Reg) of decision rule ι is defined as follows:

$$\text{C2S-Reg}(N) = \sum_{n=1}^{N} \left\{ \max_{i \in \breve{I}^F(n)} (q_i + \eta_i) - \left(q_{\iota(n)} + \eta_{\iota(n)}\right) \right\}.$$

where $\breve{I}^F(n)$ is given by

$$\breve{I}^F(n) = \arg\max_{I' \subseteq I(n)} \sum_{i \in I} q_i$$

$$\text{s.t. } |I'| = K^F, \tag{5}$$

$$\forall g \in G,\ |\{i \in I' \mid g(i) = g\}| \geq 1. \tag{6}$$

In plain words, $\breve{I}^F(n)$ is the best finalists who satisfy the constraint (4). If Eq. (4) is imposed as an "exogenous constraint" (rather than a policy), the first-best decision rule would interview $\breve{I}^F(n)$ to maximize social welfare. Constrained regret enables us to identify whether the Rooney Rule prevents perpetual underestimation: if perpetual underestimation occurs under the Rooney Rule, then the constrained regret is linear in N.

Under the Rooney Rule, myopic firm n greedily chooses candidates based on estimator $\hat{q}_i(n)$ subject to the following constraints:

$$I^F(n) = \arg\max_{I' \subseteq I(n)} \sum_{i \in I} \hat{q}_i(n) \quad \text{s.t. (5) and (6).}$$

and $\iota(n) = \arg\max_{i \in I^F(n)} \{\hat{q}_i(n) + \eta_i\}$.

The following theorem states that the Rooney Rule resolves underestimation.

Theorem 7 (Sublinear Constrained Regret under the Rooney Rule)**.** Suppose Assumptions 1, 2, 3, 4, and 5. Then, regret under the Rooney Rule is bounded as

$$\mathbb{E}\left[\text{C2S-Reg}^{\text{Rooney}}(N)\right] \leq C_{2\text{SR}}\sqrt{N}$$

where $C_{2\text{SR}}$ is $\tilde{O}(1)$ to N.

Proof. See Appendix 7. The explicit form of $C_{2\text{SR}}$ is shown in Eq. (64). □

The value $C_{2\text{SR}}$ is exponentially dependent[14] on signal variance σ_η, which implies that a sufficiently large σ_η is required to obtain a reasonable bound. That is, if the additional

[14] See definition of C_6 in the proof.

signal η_i tends to be informative (i.e., σ_η is large), the minority finalist frequently beats the majority finalist to be hired, despite its $\hat{q}_i$ being underestimated.

7.3 The Rooney Rule and Exploitation

Although the Rooney Rule prevents statistical discrimination (Theorem 7), it may worsen social welfare in terms of the original unconstrained regret. The intuition as follows. An unbalanced population ratio produces a significant probability that more than one majority candidate is highly skilled. In that case, the *true* predicted skill of the second-best majority candidate (q_i) is likely to be greater than that of the best minority candidate. This feature raises constant regret per round: when η_i is normally distributed, any finalist has a positive probability of being hired. Hence, the skill level of all candidates matters, and therefore, firms prefer to interview top-K^F candidates who have the greatest skill. The Rooney Rule prevents this outcome. This effect would present even when firms had perfect information about coefficients $\boldsymbol{\theta}$.

In summary, the loss from the constraint (4) is constant per round. Therefore, Rooney rule results in $\Omega(N)$ unconstrained regret for N rounds.

Theorem 8 (Linear Unconstrained Regret under the Rooney Rule)**.** Suppose Assumptions 1, 2, 3, 4, 5. Then, regret under the Rooney Rule is bounded as

$$\mathbb{E}\left[\text{U2S-Reg}^{\text{Rooney}}(N)\right] = \Omega(N).$$

The proof is straightforward from the argument above, and therefore, is omitted.

Although the laissez-faire and the Rooney Rule have linear unconstrained regret, these two results have different causes for the outcome in each case: laissez-faire produces linear regret due to underexploration, whereas the Rooney Rule produces linear regret due to underexploitation. One way to resolve this is by combining the two. By starting with the Rooney Rule and abolishing it after obtaining sufficiently rich data, we could mitigate the approach's disadvantage. Section 8 also testifies the performance of such a mechanism.

8 Simulation

This section reports the results of the simulations run to support our theoretical findings. Unless specified, model parameters are set as $d = 1, \mu_x = 3, \sigma_x = 2$, $\sigma_\epsilon = 2$, $\lambda = 1$, and $N = 1000$. Group sizes are set to be $(K_1, K_2) = (10, 2)$. The initial sample size is $N^{(0)} = K_1 + K_2$, and the sample size for each group is equal to its population ratio:

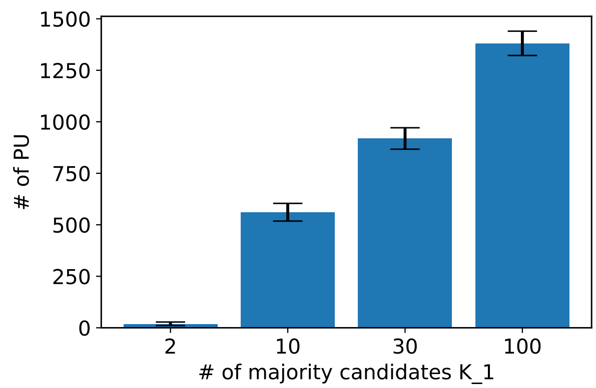

Figure 1: The frequency of perpetual underestimation across 4000 runs under laissez-faire.

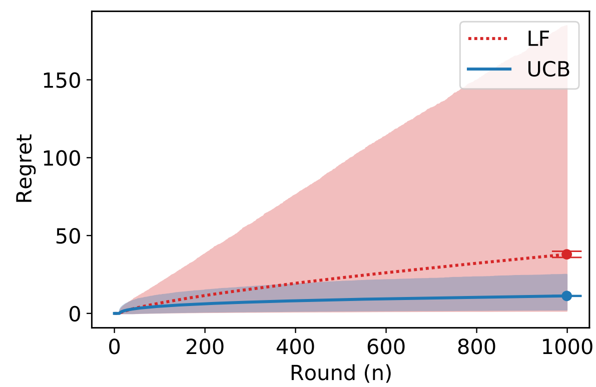

Figure 2: Comparison between the LF and UCB decision rules.

$N_1^{(0)} = K_1, N_2^{(0)} = K_2$. All results are averaged over 4000 runs. The value of δ in the confidence bound is set to 0.1.

For all bar graphs, the error bars represent the two-sigma binomial confidence intervals. For all line graphs, the lines are averages over sample paths, the areas cover between 5% and 95% percentiles of runs, and the error bars at $N = 1000$ are the two-sigma confidence intervals. The implementation of the simulation is available at https://github.com/jkomiyama/FairSocialLearning.

8.1 The Effects of Population Ratio

We testify how population ratio impacts the frequency of perpetual underestimation. The decision rule is fixed to laissez-faire (LF). We fix the number of minority candidates in each round to two (i.e., $K_2 = 2$) and vary the number of majority candidates ($K_1 = 2, 10, 30, 100$).

Figure 1 exhibits the simulation result. Consistent with our theoretical analyses, we observe that (i) as indicated by Theorem 1, laissez-faire rarely produces perpetual underestimation if the population is balanced (i.e., K_1 is close to $K_2 = 2$), and (ii) as indicated by Theorem 2, the larger the population of majority workers (i.e., K_1 increases), the more frequently perpetual underestimation occurs.

8.2 Laissez-Faire vs the UCB Mechanism

Figure 2 compares the regret associated with these two rules. As indicated by Theorem 2, our simulation shows that laissez-faire has a significant probability of underestimating the minority group. Consequently, laissez-faire sometimes causes perpetual underestimation, and regret grows linearly to n. Due to the possibility of perpetual underestimation, the

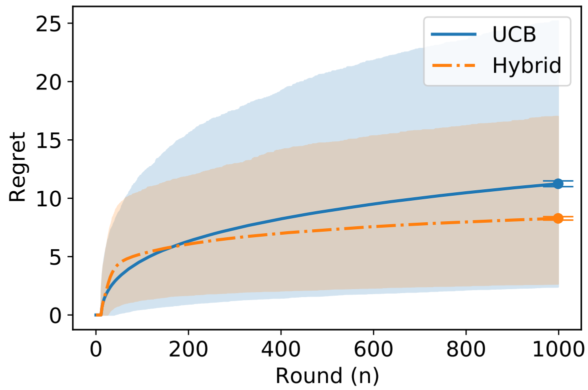

Figure 3: The regret associated with the UCB and hybrid decision rules.

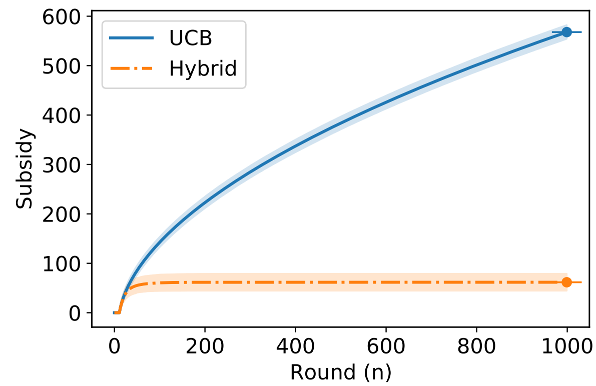

Figure 4: The budget required by the UCB and hybrid index subsidy rules.

confidence intervals of the sample paths (denoted by the red area) are very large, indicating the highly uncertain performance of laissez-faire. In contrast, consistent with Theorem 3, the UCB decision rule performs much more stably. Since the UCB rule avoids underexploration, it does not cause perpetual underestimation.

8.3 The UCB Mechanism vs the Hybrid Mechanism

Next, we compare the performance of the UCB and hybrid mechanisms. The parameter of the hybrid mechanism is set to be $a = 1$. Figure 3 shows the associated regret. As Theorems 3 and 5 anticipated, the regret associated with the two decision rules are similar (these two decision rules have the same order: $\tilde{O}(\sqrt{N})$). However, the hybrid decision rule outperforms the UCB decision rule in our simulation setting, likely because the hybrid decision rule suppresses overexploration of the minority group. Figure 4 compares the subsidy rules. As Theorems 4 and 5 predicted, the hybrid index subsidy rule requires a much smaller budget than the UCB index subsidy rule.

8.4 The Rooney Rule

This subsection compares the performance of the Rooney Rule with that of the laissez-faire decision rule. Figure 5 depicts the relationship between the frequency of perpetual underestimation and the informativeness of the signal obtained at the second stage (measured by σ_η^2, the variance of η_i) under both rules. When σ_η^2 is large, the Rooney Rule effectively resolves underestimation.

Figure 6 compares U2S-Reg associated with each rule. While both rules produce linear regret, the Rooney Rule suffers from more regret due to underexploitation. This shortcoming

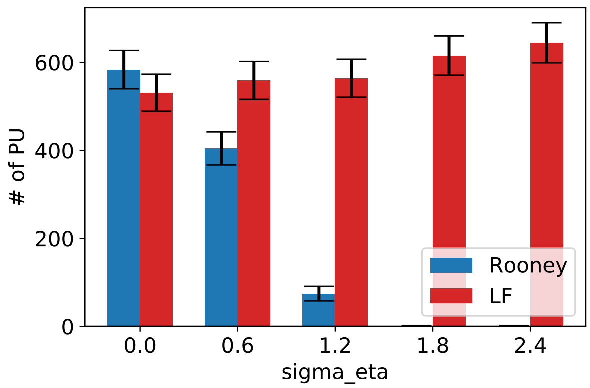

Figure 5: The frequency of perpetual underestimation in the two-stage model across 4000 runs.

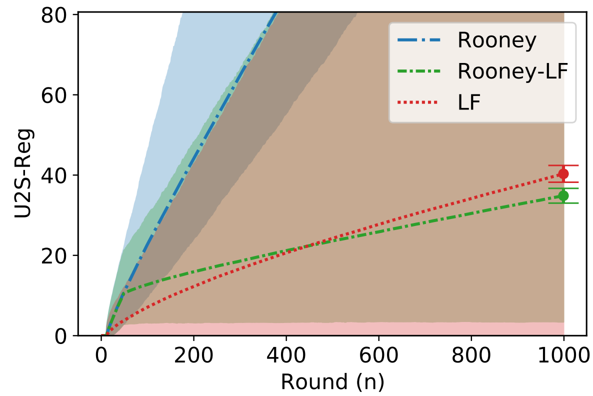

Figure 6: The U2S-Reg of laissez-faire and the Rooney Rule.

can be overcome by using the Rooney Rule as a temporary policy. the "Rooney-LF" decision rule begins with the Rooney Rule and shifts to laissez-faire after 100 rounds. This approach achieves both less regret and fairer hiring.

9 Conclusion

We have studied statistical discrimination using a contextual multi-armed bandit model. Our dynamic model articulates how a failure of social learning produces statistical discrimination. In our model, the insufficiency of data about minority groups is endogenously generated. This data shortage prevents firms from accurately estimating the skill of minority candidates. Consequently, firms tend to prefer hiring a majority worker, leading the data sufficiency to persist (perpetual underestimation). This form of statistical discrimination is not only unfair but also inefficient. We have demonstrated that an unbalanced population ratio leads laissez-faire to tend towards perpetual underestimation, an unfair and inefficient consequence.

We analyzed two possible policy interventions. One is subsidy rules that incentivize firms to hire minority candidates. Our hybrid mechanism achieves $\tilde{O}(\sqrt{N})$ regret with $\tilde{O}(1)$ subsidy. Another intervention is the Rooney Rule, which requires firms to interview at least one minority candidate. Our result indicates that terminating the Rooney Rule at an appropriate point would resolve statistical discrimination while maintaining the social welfare level. These results contrast with some of the previous studies (e.g., Foster and Vohra, 1992; Coate and Loury, 1993; Moro and Norman, 2004) demonstrating the possible counterproductivity of affirmative-action policies.

Our analyses of the two interventions provide a consistent policy implication: Affirmative actions effectively resolve statistical discrimination caused by data insufficiency, but such actions should be lifted upon acquiring sufficient information. Accordingly, a *temporary* affirmative action constitutes the best approach to resolving statistical discrimination as a social learning failure.

References

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári (2011): "Improved Algorithms for Linear Stochastic Bandits," in *Advances in Neural Information Processing Systems*, 2312–2320.

Abe, N. and P. M. Long (1999): "Associative Reinforcement Learning Using Linear Probabilistic Concepts," in *Proceedings of the Sixteenth International Conference on Machine Learning*, 3–11.

Al-Ali, M. N. (2004): "How to Get Yourself on the Door of a Job: A Cross-Cultural Contrastive Study of Arabic and English Job Application Letters," *Journal of Multilingual and Multicultural Development*, 25, 1–23.

Arrow, K. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton University Press, 3–33.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002): "Finite-Time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, 47, 235–256.

Banerjee, A. V. (1992): "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107, 797–817.

Bardhi, A., Y. Guo, and B. Strulovici (2020): "Early-Career Discrimination: Spiraling or Self-Correcting?" Working Paper.

Bastani, H., M. Bayati, and K. Khosravi (2020): "Mostly Exploration-Free Algorithms for Contextual Bandits," *Management Science*.

Bechavod, Y., K. Ligett, A. Roth, B. Waggoner, and S. Z. Wu (2019): "Equal Opportunity in Online Classification with Partial Feedback," in *Advances in Neural Information Processing Systems*, 8972–8982.

Bergman, P., D. Li, and L. Raymond (2020): "Hiring as Exploration," Working Paper 27736, National Bureau of Economic Research.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100, 992–1026.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2019a): "Inaccurate Statistical Discrimination," Working Paper.

Bohren, J. A., A. Imas, and M. Rosenberg (2019b): "The Dynamics of Discrimination: Theory and Evidence," *American Economic Review*, 109, 3395–3436.

Calders, T. and S. Verwer (2010): "Three naive Bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, 21, 277–292.

Che, Y.-K., K. Kim, and W. Zhong (2019): "Statistical Discrimination in Ratings-Guided Markets," Working Paper.

Chen, Y., A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis (2020): "The Fair Contextual Multi-Armed Bandit," in *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, 1810–1812.

Chu, W., L. Li, L. Reyzin, and R. Schapire (2011): "Contextual Bandits with Linear Payoff Functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

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

Cornell, B. and I. Welch (1996): "Culture, Information, and Screening Discrimination," *Journal of Political Economy*, 104, 542–571.

Dianat, A., F. Echenique, and L. Yariv (2020): "Statistical Discrimination and Affirmative Action in the Lab," Working Paper.

Ding, J., R. Eldan, and A. Zhai (2015): "On Multiple Peaks and Moderate Deviations for the Supremum of a Gaussian Field," *Annals of Probability*, 43, 3468–3493.

Eddo-Lodge, R. (2017): "Why I'm No Longer Talking to White People About Race," The Gurdian, https://www.theguardian.com/world/2017/may/30/why-im-no-longer-talking-to-white-people-about-race. Accessed on 08/20/2020.

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications.*, vol. 1 of *Third edition*, New York: John Wiley & Sons Inc.

Foster, D. and R. Vohra (1992): "An Economic Argument for Affirmative Action," *Rationality and Society*, 4, 176 – 188.

Frazier, P., D. Kempe, J. Kleinberg, and R. Kleinberg (2014): "Incentivizing Exploration," in *Proceedings of the fifteenth ACM conference on Economics and computation*, 5–22.

Gittins, J. C. (1979): "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 148–177.

Hanna, R. N. and L. L. Linden (2012): "Discrimination in Grading," *American Economic Journal: Economic Policy*, 4, 146–68.

Hannák, A., C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson (2017): "Bias in Online Freelance Marketplaces: Evidence from Taskrabbit and Fiverr," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1914–1933.

Hardt, M., E. Price, and N. Srebro (2016): "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems*, 3315–3323.

Immorlica, N., J. Mao, A. Slivkins, and Z. S. Wu (2020): "Incentivizing Exploration with Selective Data Disclosure," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 647–648.

Joseph, M., M. Kearns, J. H. Morgenstern, and A. Roth (2016): "Fairness in Learning: Classic and Contextual Bandits," in *Advances in Neural Information Processing Systems*, 325–333.

Kannan, S., M. Kearns, J. Morgenstern, M. Pai, A. Roth, R. Vohra, and Z. S. Wu (2017): "Fairness Incentives for Myopic Agents," in *Proceedings of the 2017 ACM Conference on Economics and Computation*, 369–386.

Kannan, S., J. H. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu (2018): "A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem," in *Advances in Neural Information Processing Systems*, 2227–2236.

Kaufmann, E. (2014): "Analyse de Stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources," Ph.D. thesis, Institut des sciences et technologies de Paris, thèse de doctorat dirigée par Cappé, Olivier et Garivier, Aurélien Signal et images Paris, ENST 2014.

Kennedy, P. (2008): *A Guide to Econometrics*, Hoboken, New Jersey: Wiley-Blackwell, chap. 12, 192–202, 6 ed.

Kleinberg, J. M. and M. Raghavan (2018): "Selection Problems in the Presence of Implicit Bias," in *9th Innovations in Theoretical Computer Science*, 33:1–33:17.

Kremer, I., Y. Mansour, and M. Perry (2014): "Implementing the 'Wisdom of the Crowd'," *Journal of Political Economy*, 122, 988–1012.

Lai, T. and H. Robbins (1985): "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, 6, 4 – 22.

Langford, J. and T. Zhang (2008): "The Epoch-Greedy Algorithm for Contextual Multi-Armed Bandits," in *Advances in Neural Information Processing Systems*, 817–824.

MacNell, L., A. Driscoll, and A. N. Hunt (2015): "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching," *Innovative Higher Education*, 40, 291–303.

Mailath, G. J., L. Samuelson, and A. Shaked (2000): "Endogenous Inequality in Integrated Labor Markets with Two-Sided Search," *American Economic Review*, 90, 46–72.

Mansour, Y., A. Slivkins, and V. Syrgkanis (2020): "Bayesian Incentive-Compatible Bandit Exploration," *Operations Research*, 68, 1132–1161.

Mitchell, K. M. and J. Martin (2018): "Gender Bias in Student Evaluations," *PS: Political Science and Politics*, 51, 648–652.

Monachou, F. G. and I. Ashlagi (2019): "Discrimination in Online Markets: Effects of Social Bias on Learning from Reviews and Policy Design," in *Advances in Neural Information Processing Systems*, 2145–2155.

Moro, A. and P. Norman (2004): "A General Equilibrium Model of Statistical Discrimination," *Journal of Economic Theory*, 114, 1–30.

Neumark, D. (2018): "Experimental Research on Labor Market Discrimination," *Journal of Economic Literature*, 56, 799–866.

O'Brien, S. A. (2018): "Facebook Commits to Seeking More Minority Directors," *CNN*, https://money.cnn.com/2018/05/31/technology/facebook-board-diversity/index.html. Accessed on 09/09/2020.

Papanastasiou, Y., K. Bimpikis, and N. Savva (2018): "Crowdsourcing Exploration," *Management Science*, 64, 1727–1746.

Pedreschi, D., S. Ruggieri, and F. Turini (2008): "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 560–568.

Peña, V. H., T. L. Lai, and Q.-M. Shao (2008): *Self-normalized processes: Limit theory and Statistical Applications*, Springer Science & Business Media.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659–661.

Precht, K. (1998): "A Cross-Cultural Comparison of Letters of Recommendation," *English for Specific Purposes*, 17, 241–265.

Raghavan, M., A. Slivkins, J. W. Vaughan, and Z. S. Wu (2018): "The Externalities of Exploration and How Data Diversity Helps Exploitation," in *Conference On Learning Theory*, PMLR, vol. 75, 1724–1738.

Rigollet, P. (2015): "High Dimensional Statistics," MIT OpenCourseWare, https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015/lecture-notes/. Accessed on 08/29/2020.

Robbins, H. (1952): "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 58, 527–535.

Rusmevichientong, P. and J. N. Tsitsiklis (2010): "Linearly Parameterized Bandits," *Mathematics of Operations Research*, 35, 395–411.

Smith, L. and P. Sørensen (2000): "Pathological Outcomes of Observational Learning," *Econometrica*, 68, 371–398.

Thompson, W. R. (1933): "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, 25, 285–294.

Trix, F. and C. Psenka (2003): "Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty," *Discourse and Society*, 14, 191–220.

Tropp, J. A. (2012): "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434.

Xu, L., J. Honda, and M. Sugiyama (2018): "A Fully Adaptive Algorithm for Pure Exploration in Linear Bandits," in *International Conference on Artificial Intelligence and Statistics*, 843–851.

Appendix

A The Design of Subsidy Rules

1 Pivot Subsidy Rules

This subsection provides a rationale for focusing on the UCB and hybrid index subsidy rules. First, we define an *index* and *index policy* as follows.

Definition 11 (Index)**.** A sequence of functions $Q = (Q_i)$ where $Q_i(n; \cdot) : H(n) \to \mathbb{R}$ is an *index* if for all n and $i \in I(n)$, $Q_i(n; \cdot)$ only depends on $\boldsymbol{X}_{g(i)}(n)$, $Y_{g(i)}(n)$, and $\boldsymbol{x}_i$. A subsidy rule s is an *index policy* if s is an index.

Our definition of the index and index policy is slightly weaker than the standard definition used by the multi-armed bandit literature (Gittins, 1979). A standard definition requires that the index of an arm depends only on the data generated by that arm. However, since we regard a set of arms as a group, it does not make sense to focus on the data generated by "one specific arm." Instead, our definition requires that the subsidy for worker i be independent of (i) the characteristics of the other agents in the same round (i.e., $\boldsymbol{x}_j$ for any $j \in I(n) \setminus \{i\}$) and (ii) the data about other groups (i.e., $\boldsymbol{X}_{g'}(n)$ for any $g' \neq g(i)$).

If a subsidy rule is an index policy, the government does not have to observe the characteristics of $I(n) \setminus \{i\}$ to determine the subsidy assigned to the employment of worker i. This is a practically desirable property: For many real-world problems, it is difficult for the government to observe the characteristics of candidate workers who are not hired, rendering a non-index policy hard to implement.

The estimated skill $\hat{q}$, the UCB index $\tilde{q}$, and the hybrid index $\tilde{q}^{\mathrm{H}}$ are indices. The UCB index subsidy rule and the hybrid index rule are index policies. Furthermore, they also belong to a class of *pivot subsidy rules* that are defined as follows:

Definition 12. Given that a decision rule ι that maximizes an index Q, i.e.,

$$\iota(n) = \arg\max_{i \in I(n)} Q_i(n),$$

a *pivot subsidy rule* s is specified by

$$s_i(n) = Q_i(n) - \hat{q}_i(n).$$

Thus, the UCB index subsidy rule is obtained by substituting $Q_i(n) = \tilde{q}_i(n)$ and the hybrid index subsidy rule is obtained by substituting $Q_i(n) = \tilde{q}_i^{\mathrm{H}}(n)$.

The following theorem states the optimality of the pivot subsidy rule. Among all index policies, the pivot subsidy rule requires the smallest subsidy amount under certain conditions.

Theorem 9 (Optimality of the Pivot Subsidy Rule)**.** Suppose that (i) a decision rule ι maximizes an index Q, (ii) $Q_i(n; h(n)) \geq \hat{q}_i(n; h(n))$ for all i, n, $h(n)$, and (iii) $\inf_{i,n,h(n)} Q_i(n; h(n)) = \hat{q}_i(n)$. Then, we have the following.

(a) A pivot subsidy rule implements ι.

(b) Let s be a pivot subsidy rule and s' be an arbitrary index policy that implements ι. Then, for all i, n and $h(n)$, we have

$$s_i(n; h(n)) \leq s'_i(n; h(n)).$$

Note that, the UCB index $\tilde{q}$ and the hybrid index $\tilde{q}^{\mathrm{H}}$ satisfy Conditions (ii) and (iii). Accordingly, among all index policies that implement the same decision rule, the UCB index subsidy rule and the hybrid index subsidy rule require the smallest subsidy.

Proof. (a) Since $\hat{q}_i(n) + s_i(n) = Q_i(n)$, firm n's payoff from hiring worker i is equal to $Q_i(n)$. Furthermore, since $Q_i(n) \geq \hat{q}_i(n)$, $s_i(n) \geq 0$ always holds. Accordingly, the pivot subsidy rule implements the targeted decision rule.

(b) For notational simplicity, we omit n, $\boldsymbol{X}_g$, Y_g from this proof. Define a correspondence $\mathcal{U}$ by

$$\mathcal{U}(Q_i; s') \coloneqq \{u_i \in \mathbb{R} \mid \exists i, \exists \boldsymbol{x}_i \text{ s.t. } \hat{q}_i(\boldsymbol{x}_i) + s'_i(\boldsymbol{x}_i) = u_i, Q_i = Q_i(\boldsymbol{x}_i)\}.$$

The set $\mathcal{U}(Q_i; s')$ represents the set all of firm n's possible payoffs from hiring a worker with index Q_i, given that the subsidy rule s' is used.

Clearly, subsidy rule s' implements the decision rule ι if and only if for all distinct i, j, $Q_j > Q_i$ implies

$$\min \mathcal{U}(Q_j; s') > \max \mathcal{U}(Q_i; s'). \tag{7}$$

Since $\min \mathcal{U}(\cdot; s')$ is an increasing function, it is continuous at all but countably many points. Thus, $\mathcal{U}(Q_i; s)$ is a singleton for almost all values of Q_i.

Now, suppose that $\mathcal{U}(Q_i^*; s')$ is not a singleton for some $Q_i^* \in \mathbb{R}_+$. Define Δ by

$$\Delta \coloneqq \max \mathcal{U}(Q_i^*; s') - \min \mathcal{U}(Q_i^*; s').$$

Define another subsidy rule s'' by setting

$$s_i''(\boldsymbol{x}_i) = \begin{cases} s_i'(\boldsymbol{x}_i) & \text{if } Q_i(\boldsymbol{x}_i) < Q_i^*, \\ \min \mathcal{U}(Q_i^*; s') - \hat{q}_i(\boldsymbol{x}_i) & \text{if } Q_i(\boldsymbol{x}_i) = Q_i^*, \\ s_i'(\boldsymbol{x}_i) - \Delta & \text{otherwise,} \end{cases}$$

for all i. Then, we have

$$\mathcal{U}(\tilde{q}_i; s'') = \begin{cases} \mathcal{U}(\tilde{q}_i; s') & \text{if } Q_i < Q_i^*, \\ \{\min \mathcal{U}(Q_i^*; s')\} & \text{if } Q_i = Q_i^*, \\ \mathcal{U}(Q_i; s') - \Delta & \text{otherwise,} \end{cases}$$

which implies that $\mathcal{U}(\cdot; s'')$ also satisfies (7), or equivalently, s'' also implements the decision rule ι. Furthermore, $s_i''(\boldsymbol{x}_i) \le s_i'(\boldsymbol{x}_i)$ for all $\boldsymbol{x}_i$, with a strict inequality for some $\boldsymbol{x}_i$. Accordingly, s'' needs a smaller budget than s'.

By the argument above, whenever $\mathcal{U}(\cdot; s')$ does not return a singleton for some Q_i, the subsidy amount can be improved by filling a gap. Now, we discuss the case that $\mathcal{U}(\cdot; s')$ returns a singleton for all Q_i; i.e., $\mathcal{U}$ reduces to a function. We use $u(Q_i; s')$ to represent the firm's utility when it hires a worker with the index Q_i. We have

$$s_i'(\boldsymbol{x}_i) = u(Q_i(\boldsymbol{x}_i); s') - \hat{q}_i(\boldsymbol{x}_i)$$

for all $\boldsymbol{x}_i$. Since we require that $s_i'(\boldsymbol{x}_i) \ge 0$ for all $\boldsymbol{x}_i$,

$$u(Q_i(\boldsymbol{x}_i); s') - \hat{q}_i(\boldsymbol{x}_i) \ge 0.$$

After some history, $\hat{q}_i$ may become arbitrarily close to Q_i. Accordingly, u must satisfy

$$u(Q_i; s) \ge Q_i \tag{8}$$

for all q. The pivot subsidy rule satisfies (8) with equalities for all q: The UCB index subsidy rule satisfies $s_i = Q_i - \hat{q}_i$, and therefore, $u(Q_i; s) = Q_i$ for all Q_i. Accordingly, the pivot subsidy rule demands the minimum possible budget.

$\square$

2 Cost-Saving Subsidy Rules

If a subsidy rule need not be an index policy, then a decision rule can be implemented with a smaller budget. The *cost-saving subsidy rule* provides a minimum subsidy to change the firm's hiring decision.

Definition 13 (Cost-Saving Subsidy Rule)**.** Given an arbitrary decision rule ι, a *cost-saving subsidy rule* s is specified by

$$s_i(n; h(n)) = \begin{cases} \max_{j \in I(n)} \hat{q}_j(n; h(n)) - \hat{q}_i(n; h(n)) & \text{if } i = \iota(n); \\ 0 & \text{otherwise.} \end{cases}$$

The UCB and hybrid cost-saving subsidy rules are obtained by applying Definition 13 to the UCB and hybrid decision rules.

A cost-saving subsidy rule subsidizes only the targeted worker, $\iota(n)$. Hence, for other workers $j \neq \iota(n)$, the payoff from the employment is $\hat{q}_j(n)$. The UCB cost-saving subsidy rule sets the subsidy amount $s_{\iota(n)}$ such that the payoff from hiring worker $\iota(n)$, which is $\hat{q}_{\iota(n)}(n) + s_{\iota(n)}(n)$, is equal to (or slightly larger than) the payoff from hiring the worker with the greatest estimated skill, $\max_{j \in I(n)} \hat{q}_j(n)$.

Clearly, the UCB cost-saving subsidy rule is the subsidy rule that requires the smallest budget to implement the UCB decision rule. Since fines (negative subsidies) are not allowed, the government cannot further discourage the employment of the other candidates, $j \in I(n) \setminus \{\iota(n)\}$. Hence, the UCB cost-saving subsidy rule requires the smallest budget among all subsidy rules that implements the decision rule (2).

Theorem 10 (Optimality of the Cost-Saving Subsidy Rule)**.**

(a) A cost-saving subsidy rule implements the decision rule with which the subsidy rule is associated.

(b) Let s be a cost-saving subsidy rule and s' be an arbitrary subsidy rule that implements the same decision rule. Then, for all i, n and $h(n)$, we have

$$s_i(n; h(n)) \leq s'_i(n; h(n)).$$

The proof is straightforward from the argument above.

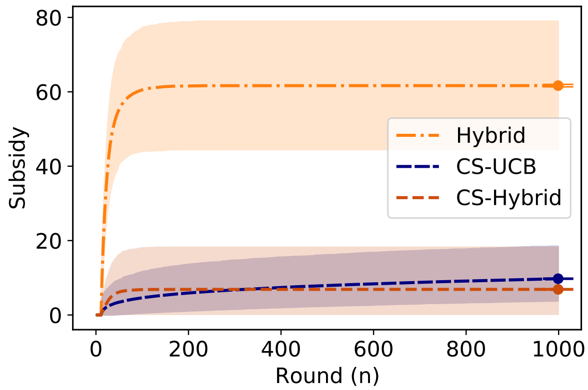

Figure 7: The UCB cost-saving subsidy rule vs the hybrid index subsidy rule and the hybrid cost-saving subsidy rule.

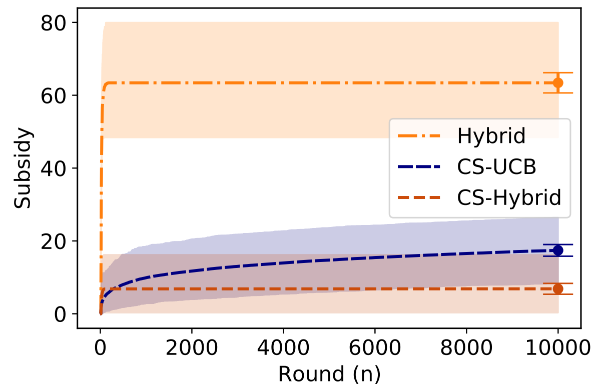

Figure 8: The UCB cost-saving subsidy rule vs the hybrid index subsidy rule and the hybrid cost-saving subsidy rule where $N = 10000$. The plot is an average over 50 runs.

Since the government hardly observes rejected candidates' characteristics, a cost-saving subsidy rule is difficult to implement. Nevertheless, since it provides the smallest subsidy for implementing a decision rule, its performance is a useful theoretical benchmark.

B Additional Simulation

1 The Pivot Subsidy Rule vs the Cost-Saving Subsidy Rule

Figure 7 compares the subsidy amount associated with the UCB cost-saving subsidy rule and the hybrid subsidy rules. The UCB index subsidy rule is excluded because it requires a much larger subsidy (see Figure 4). We observe that (i) two cost-saving subsidy rules require a similar subsidy amount (although the hybrid cost-saving subsidy rule performs slightly better), and (ii) the cost-saving method is very effective, even when compared to the hybrid index rule.

Although the subsidy amounts required by these two cost-saving rules appear similar in Figure 7, with more rounds, the hybrid cost-saving subsidy rule outperforms the UCB cost-saving subsidy rule. Figure 8 articulates this result. While the subsidy required by the hybrid cost-saving rule remains constant after a few (about 100) rounds, the subsidy required by the UCB cost-saving rule gradually grows. This result is also consistent with our theory: While the subsidy required by the hybrid rule is proven to be $\tilde{O}(1)$ (this is immediate from Theorems 5 and 10), there is no such guarantee for the UCB cost-saving subsidy rule. We conjecture that the subsidy required by the UCB cost-saving rule is $\tilde{O}(\sqrt{N})$.

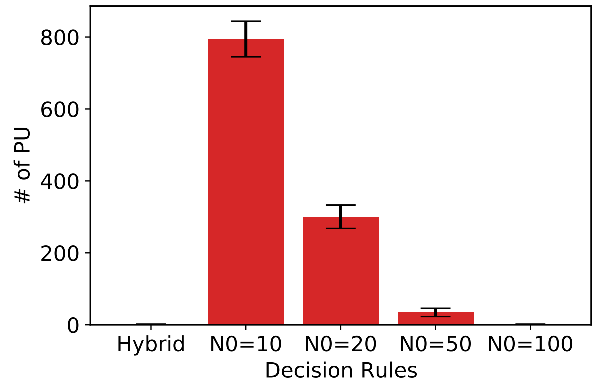

Figure 9: The frequency of perpetual underestimations across 4000 runs, each run is for $1,000$ rounds.

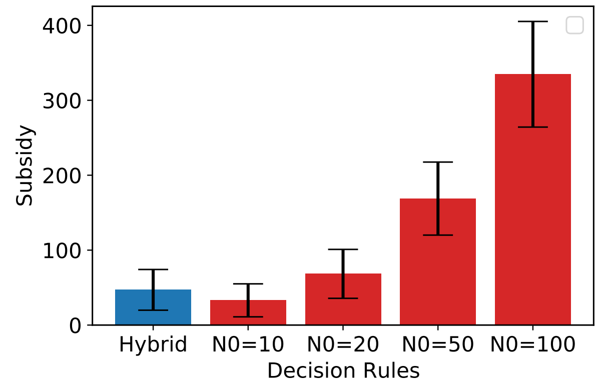

Figure 10: The total amount of subsidies ($1,000$ rounds). "Hybrid" denotes the hybrid cost-saving subsidy rule.

Note: The hybrid mechanism (with the cost-saving subsidy rule) vs laissez-faire with uniform sampling. N0 ($= N^{(0)}$) denotes the number of initial samples taken prior to laissez-faire being implemented.

2 The Hybrid Mechanism vs Uniform Sampling

Kannan et al. (2018) show that with sufficiently large initial samples (i.e., $N^{(0)}$ is large), the greedy algorithm (corresponding to laissez-faire in this paper) has sublinear regret.[15] Our analysis also indicates that the probability of perpetual underestimation is small when $N^{(0)}$ is large (see Lemma 22 for full details).

This "warm-start" version of laissez-faire might be presumed efficient. However, the warm-start approach carries several disadvantages. First, although we have thus far ignored the cost of acquiring initial samples for analytical tractability, we need to consider this cost if we want to *take* a sufficiently long warm-start period. Since uniform sampling ignores firms' incentives for hiring workers, practical implementation of it requires a large budget. Second, uniform sampling does not maximize any index. This precludes its implementation by any index policy. Third, uniform sampling is inefficient in terms of information acquisition because it is not adaptive to currently estimated parameters.

We argue that our hybrid mechanism (Section 6) is more efficient than laissez-faire with a warm start because it initially samples the data adaptively before switching to laissez-faire at an efficient time. Hence, we can naturally expect the hybrid mechanism outperforms laissez-faire with initial uniform sampling.

Figures 9 and 10 exhibit the simulation results comparing the hybrid mechanism with

[15] We also note that the number of initial samples required by the relevant theorem ($n_{\min}$ of Lemma 4.3) is very large and cannot be satisfied in our simulation setting: Letting $R = \sigma_x\sqrt{2\log(N)}$, we have $n_{\min} \geq 320R^2\log(R^2dK/\delta)/\lambda_0 \geq 10^3$.

laissez-faire for various initial samples. In this simulation, the number of initial samples for each group is proportional to the population ratio; i.e., $N_g^{(0)} = (K_g/K) \cdot N^{(0)}$.

Figure 9 measures the frequency of perpetual underestimations. As our theory indicated, the larger the initial sample, the less frequently perpetual underestimation occurs. Additionally, we observed no perpetual underestimation for the hybrid mechanism, because it solidly incentivizes hiring candidates from an underexplored group.

Figure 10 depicts the subsidy amount required by the cost-saving subsidy rules. Here, we can observe that the hybrid cost-saving subsidy rule outperforms laissez-faire with uniform sampling. Laissez-faire requires at least $N^{(0)} \geq 50$ samples to mitigate perpetual underestimation. However, when $N^{(0)} \geq 20$, the hybrid cost-saving subsidy rule requires a smaller budget than uniform sampling.

C Bayesian Approach

Thus far, we have assumed that firms and the regulator are frequentists, implying that they estimate the underlying parameter $(\boldsymbol{\theta}_g)_{g \in G}$ based solely on the data and without forming a "prior belief" about the parameter's distribution.

The frequentist approach is widely used in the multi-armed bandit literature because selecting a "prior belief" for implementing a Bayesian approach is difficult in practice. Since the frequentist approach is valid for *any* realization of the parameter $(\boldsymbol{\theta}_g)_{g \in G}$, it produces a more robust solution. Based on this trend, we have developed and analyzed a frequentist model.

Nevertheless, adopting a Bayesian setting leads to a similar conclusion, so long as all firms share a common prior belief. Specifically, when the common prior belief is endowed as a normal distribution, i.e.,

$$\boldsymbol{\theta}_g \sim \mathcal{N}(0, \kappa^2 \boldsymbol{I}_d),$$

then the posterior belief would also be a normal distribution, with its mean becoming $\boldsymbol{x}_i' \hat{\boldsymbol{\theta}}_{g(i)}(n)$.

Regarding the UCB mechanism, there exists a Bayesian version of confidence region[16] $\mathcal{C}_g^{\text{Bayes}}(n)$ such that

$$\Pr^{\text{Bayes}}\left(\bigcap_n \{\boldsymbol{\theta}_g \in \mathcal{C}_g^{\text{Bayes}}(n)\}\right) \geq 1 - \delta$$

[16]Here, $\Pr^{\text{Bayes}}$ denotes probability over the Bayes posterior. The bound here is derived from Eq. (4.8) in Kaufmann (2014).

by defining

$$
\begin{aligned}
&\mathcal{C}_g^{\text{Bayes}}(n) \\
&= \left\{ \bar{\boldsymbol{\theta}}_g \in \mathbb{R}^d : \left\| \bar{\boldsymbol{\theta}}_g - \hat{\boldsymbol{\theta}}_g(n) \right\|_{\bar{\boldsymbol{V}}_g(n)} \leq \sigma_\epsilon \sqrt{d + \log\left(\frac{\pi^2 N^2}{6\delta}\right) + 2\sqrt{d \log\left(\frac{\pi^2 N^2}{6\delta}\right)}} \right\}.
\end{aligned}
$$

Using $\mathcal{C}_g^{\text{Bayes}}(n)$, we can obtain a Bayesian version of the UCB mechanism.[17]

D Lemmas

This section describes the technical lemmas that are used for deriving the theorems.

The Hoeffding inequality, which is one of the most well-known versions of concentration inequality, provides a high-probability bound of the sum of bounded independent random variables.

Lemma 11 (Hoeffding Inequality)**.** Let $x_1, x_2, \ldots, x_n$ be i.i.d. random variables in $[0, 1]$. Let $\bar{x} = (1/n) \sum_{t=1}^n x_t$. Then,

$$
\begin{aligned}
\Pr\left[\bar{x} - \mathbb{E}[\bar{x}] \geq k\right] &\leq e^{-2nk^2} \\
\Pr\left[\bar{x} - \mathbb{E}[\bar{x}] \leq -k\right] &\leq e^{-2nk^2}
\end{aligned}
$$

and taking union bound yields

$$
\Pr\left[|\bar{x} - \mathbb{E}[\bar{x}]| \geq k\right] \leq 2e^{-2nk^2}.
$$

The following is a version of concentration inequality for a sum of squared normal variables.

Lemma 12 (Concentration Inequality for Chi-squared distribution)**.** Let $Z_1, Z_2, \ldots, Z_n$ be independent standard normal variables. Then,

$$
\Pr\left[\left|\frac{1}{n}\sum_{k=1}^n Z_k^2 - 1\right| \geq t\right] \leq 2e^{-nt^2/8}
$$

[17] Note that, to run the UCB mechanism in a model, the regulator needs to know the common prior belief of firms to calculate the confidence bound $\mathcal{C}_g^{\text{Bayes}}(n)$.

Lemma 13 (Normal Tail Bound (Feller, 1968))**.** Let $\phi(x) := \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ be the probability density function (pdf) of a standard normal random variable. Let $\Phi^c(x) = \int_x^\infty \phi(x')dx'$. Then,

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{e^{-x^2/2}}{\sqrt{2\pi}} \leq \Phi^c(x) \leq \frac{1}{x}\frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Lemma 14 (Largest Context, Theorem 1.14 in Rigollet (2015))**.** Let

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \sigma_x \boldsymbol{I}_d)$$

for each $i \in I(n)$. Let $\mu_x = ||\boldsymbol{\mu}_x||$ and

$$L_\delta := \mu_x + \sigma_x\sqrt{2d(2\log(KN) + \log(1/\delta))}$$

Then, with a probability at least $1-\delta$, we have

$$\forall i \in I(n), n \in [N],\ ||\boldsymbol{x}_i|| \leq L_\delta.$$

The following bounds the variance of a conditioned normal variable.

Lemma 15 (Conditioned Tail Deviation)**.** Let $x \sim \mathcal{N}(a,1)$ be a scalar normal random variable with its mean $a \in \mathbb{R}$ and unit variance. Then, for any $b \in \mathbb{R}$, the following two inequalities hold.

$$\mathrm{Var}(x|x \geq b) \geq \frac{1}{10}$$

Proof. Without loss of generality, we assume $b = 0$ (otherwise we can reparametrize $x' = x - b \sim \mathcal{N}(a-b,1)$). If $a \leq 0$, the pdf of conditioned variable $x|x \geq 0$ is $2\psi(x)$ for $x \geq 0$. Manual evaluation of this distribution[18] reveals that $\mathrm{Var}(x) \geq 1/10$. Otherwise ($a > 0$), the pdf of $x|x \geq b$ is $p(x) \geq \psi(x-a)$ for $x \geq a$, which implies $\mathrm{Var}(x|x \geq b) \geq \mathrm{Var}(z)$, where z be a "half-normal" random variable[19] with its cumulative distribution function

$$P(z) = \begin{cases} \Phi(z) & \text{if } z > 0 \\ 1/2 & \text{if } z = 0 \\ 0 & \text{otherwise} \end{cases}.$$

Manual evaluation of $\mathrm{Var}(z)$ also shows that $\mathrm{Var}(z) \geq 1/10$. □

[18] This distribution is called a folded normal distribution.

[19] Half of the mass lies in $z > 0$, the other half of mass is at $z = 0$.

The following diversity condition that simplifies the original definition of Kannan et al. (2018) is used to lower-bound the expected minimum eigenvalue of $\bar{\boldsymbol{V}}_g$.

Lemma 16 (Diversity of Multivariate Normal Distribution)**.** The context $\boldsymbol{x}$ is λ_0-diverse for $\lambda_0 > 0$ if for any $\hat{b} \in \mathbb{R}$, $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$

$$\lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}'|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right]\right) \geq \lambda_0.$$

Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \sigma_x \boldsymbol{I}_d)$. Then, the context $\boldsymbol{x}$ is λ_0-diverse with $\lambda_0 = \sigma_x^2/10$.

Proof.

$$\begin{aligned}\lambda_{\min}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}'|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right]\right) &= \min_{\boldsymbol{v}:||\boldsymbol{v}||=1} \mathbb{E}\left[(\boldsymbol{v}'\boldsymbol{x})^2|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right] \\ &\geq \min_{\boldsymbol{v}:||\boldsymbol{v}||=1} \mathrm{Var}\left[\boldsymbol{v}'\boldsymbol{x}|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right]\end{aligned}$$

Let $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ be the orthogonal bases. Without loss of generality, we assume $\hat{\boldsymbol{\theta}} = \theta_1 \boldsymbol{e}_1$ for some $\theta_1 \geq 0$ and $\mu_x = u_1\boldsymbol{e}_1 + u_2\boldsymbol{e}_2$ for some $u_1, u_2 \in \mathbb{N}$. Let

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \cdots + x_d\boldsymbol{e}_d.$$

Due to the property of the normal distribution, each coordinate x_l for $l \in [d]$ are independent of each other. We will show the variance of

$$\mathrm{Var}\left[x_l|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right] \geq \sigma_x^2/10, \tag{9}$$

which suffices to prove Lemma 16.

- For the first dimension, we have $x_1 \sim \mathcal{N}(u_1, \sigma_x^2)$ and

  $$\mathrm{Var}\left[x_1|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right] = \mathrm{Var}\left[x_1|x_1 \geq \hat{b}/\theta_1\right].$$

  Applying Lemma 15 with $x = \mathrm{sgn}(\hat{b}/\theta_1)/\sigma_x$, $a = \mu_x/\sigma_x$, and $b = |\hat{b}/\theta_1|$ yield

  $$\mathrm{Var}\left[x_1|x_1 \geq \hat{b}/\theta_1\right] \geq \sigma_x^2/10.$$

- For the second dimension, we have $x_2 \sim \mathcal{N}(u_2, \sigma_x^2)$ and

  $$\mathrm{Var}\left[x_2|\boldsymbol{x}'\hat{\boldsymbol{\theta}} \geq \hat{b}\right] = \mathrm{Var}\left[x_2\right] = \sigma_x^2 > \sigma_x^2/10.$$

- $(x_3, x_4, \ldots, x_d) \sim \mathcal{N}(0, \sigma_x^2 \boldsymbol{I}_{d-2})$. In other words, these characteristics are normally distributed and thus $\mathrm{Var}(x_l) = \sigma_x^2 > \sigma_x^2/10$.

In summary, we have Eq. (9), which concludes the proof. □

Lemma 17 (Abbasi-Yadkori et al. (2011))**.** Assume that $||\boldsymbol{\theta}_g|| \le S$. Let $\delta > 0$ be arbitrary. With a probability at least $1-\delta$, the true parameter $\boldsymbol{\theta}_g$ is bounded as

$$\forall n, \ \left\| \hat{\boldsymbol{\theta}}_g(n) - \boldsymbol{\theta}_g \right\|_{\bar{\boldsymbol{V}}_g(n)} \le \sigma_\epsilon \sqrt{2d \log \left(\frac{\det(\bar{\boldsymbol{V}}_g(n))^{1/2} \det(\lambda \boldsymbol{I})^{-1/2}}{\delta} \right)} + \lambda^{1/2} S. \tag{10}$$

Moreover, let $L = \max_{i,n} \|\boldsymbol{x}_i(n)\|_2$ and

$$\beta_n(L, \delta) = \sigma_\epsilon \sqrt{d \log \left(\frac{1 + nL^2/\lambda}{\delta} \right)} + \lambda^{1/2} S.$$

Then, with a probability at least $1-\delta$,

$$\forall n, \ \left\| \hat{\boldsymbol{\theta}}_g(n) - \boldsymbol{\theta}_g \right\|_{\bar{\boldsymbol{V}}_g(n)} \le \beta_n(L, \delta). \tag{11}$$

The following lemma is used in deriving a regret bound.

Lemma 18 (Abbasi-Yadkori et al. (2011))**.** Let $\lambda \ge 1$ and $L = \max_{n,i} \|\boldsymbol{x}_i(n)\|_2$. Then, the following inequality holds:

$$\sum_{n:\iota(n)=g} \left\| \boldsymbol{x}_{\iota(n)} \right\|^2_{(\bar{\boldsymbol{V}}_g(n))^{-1}} \le 2L^2 \log \left(\frac{\det(\bar{\boldsymbol{V}}_g(N))}{\det(\lambda \boldsymbol{I}_d)} \right)$$

for any group g.

The following inequality is used to bound the variation of the minimum eigenvalue of the sum of characteristics (contexts).

Lemma 19 (Matrix Azuma Inequality (Tropp, 2012))**.** Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be adaptive sequence of $d \times d$ symmetric matrices such that $\mathbb{E}_{k-1} \boldsymbol{X}_k = \boldsymbol{0}$ and $\boldsymbol{X}_k^2 \preceq \boldsymbol{A}_k^2$ almost surely, where $\boldsymbol{A} \succeq \boldsymbol{B}$ between two matrices denotes $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. Let

$$\sigma_A^2 := \left\| \frac{1}{n} \sum_k \boldsymbol{A}_k^2 \right\|$$

where the matrix norm is defined by the largest eigenvalue. Then, for all $t \geq 0$,

$$\Pr\left[\lambda_{\min}\left(\sum_k \boldsymbol{X}_k\right) \leq t\right] \leq d\exp(-t^2/(8n\sigma_A^2)).$$

Proof. The proof directly follows from Theorem 7.1 and Remark 3.10 in Tropp (2012). $\square$

The following lemma states that the selection bias makes its variance slightly ($O(1/\log K)$ times) smaller than the original variance.

Lemma 20 (Variance of Maximum, Theorem 1.8 in Ding et al. (2015))**.** Let $x_1, \ldots, x_K \in \mathbb{R}$ be i.i.d. samples from $\mathcal{N}(0,1)$. Let $I_{\max} = \arg\max_{i\in[K]} x_i$. Then, there exists a distribution-independent constant $C_{\text{varmax}} > 0$ such that

$$\mathrm{Var}[I_{\max}] \geq \frac{C_{\text{varmax}}}{\log(K)}.$$

E Proofs

0 Common Inequalities

In the proofs, we often ignore the events that happen with probability $O(1/N)$. Since the expected regret per round is at most $\max_i \boldsymbol{x}_i'\boldsymbol{\theta}_{g(i)} - \min_i \boldsymbol{x}_i'\boldsymbol{\theta}_{g(i)}$, which is $O(1)$ in expectation, the events that happen with probability $O(1/N)$ contributes to the regret by $O(N \times 1/N) = O(1)$, which are insignificant in our analysis.

Specifically, we regard all the contexts are bounded by $L_{1/N} = O(\sqrt{\log N}) = \tilde{O}(1)$ because

$$\Pr\left[\forall n \in [N], i \in I(n),\ ||\boldsymbol{x}_i(n)|| \leq L_{1/N}\right] \geq 1 - \frac{1}{N}. \quad \text{(by Lemma 14)} \tag{12}$$

Moreover, we also regard all the confidence bounds hold with

$$\beta_n\left(L_{1/N}, 1/N\right) \leq \beta_N\left(L_{1/N}, 1/N\right) = O(\sqrt{\log N}) = \tilde{O}(1)$$

because

$$\Pr\left[\forall n \in [N], g \in G,\ \left\|\hat{\boldsymbol{\theta}}_g(n) - \boldsymbol{\theta}_g\right\|_{\bar{\boldsymbol{V}}_g(n)} \leq \beta_n\left(L_{1/N}, \frac{1}{N}\right)\right] \geq 1 - \frac{|G|}{N}. \tag{13}$$

follows form Eq. (11) in Lemma 17,

Throughout the proof, we ignore the case these events do not hold. We also denote $L := L_{1/N}$ and $\beta_N = \beta_N(L, 1/N)$.

We next discuss the upper confidence bounds.

Remark 6 (Bound for $\tilde{\boldsymbol{\theta}}_i$). Let $\tilde{\boldsymbol{\theta}}_i = \arg\max_{\bar{\boldsymbol{\theta}}_{g(i)} \in \mathcal{C}_{g(i)}(n)} \boldsymbol{x}_i' \bar{\boldsymbol{\theta}}_{g(i)}$. By definition of $\tilde{\boldsymbol{\theta}}_i$, the following inequality always holds:

$$\forall n, \ \left\| \tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{g(i)}(n) \right\|_{\bar{\boldsymbol{V}}_g(n)} \leq \beta_N. \tag{14}$$

and Eq. (13) implies

$$\forall n, \ \boldsymbol{x}_i' (\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{g(i)}(n)) \geq 0. \tag{15}$$

Moreover, by using triangular inequality, we have

$$\left\| \tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_g(n) \right\|_{\bar{\boldsymbol{V}}_g(n)} \leq \left\| \tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_g(n) \right\|_{\bar{\boldsymbol{V}}_g(n)} + \left\| \hat{\boldsymbol{\theta}}_g(n) - \boldsymbol{\theta} \right\|_{\bar{\boldsymbol{V}}_g(n)}$$

and thus Eq. (13) implies

$$\forall n, \ \left\| \tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_g(n) \right\|_{\bar{\boldsymbol{V}}_g(n)} \leq 2\beta_N. \tag{16}$$

We use the calligraphic font to denote events. For two events $\mathcal{A}, \mathcal{B}$, let $\mathcal{A}^c$ be a complementary event and $\{\mathcal{A}, \mathcal{B}\} := \{\mathcal{A} \cap \mathcal{B}\}$. We also use prime to denote events that are close to the original event. For example, event $\mathcal{A}'$ is different from event $\mathcal{A}$ but these two events are deeply linked. Finally, we discuss the minimum eigenvalue. We denote $\boldsymbol{A} \succeq \boldsymbol{B}$ for two $d \times d$ matrices if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite: That is, $\lambda_{\min}(\boldsymbol{A} - \boldsymbol{B}) \geq 0$. Note that $\lambda_{\min}(\boldsymbol{A} + \boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B})$ and $\lambda_{\min}(\boldsymbol{A} + \boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A})$ if $\boldsymbol{B} \succeq \boldsymbol{0}$. We have $\boldsymbol{x}\boldsymbol{x}' \succeq \boldsymbol{0}$ for any vector $\boldsymbol{x} \in \mathbb{R}^d$.

1 Proof of Theorem 1

We first bound regret per round $\mathrm{reg}(n) := \mathrm{Reg}(n) - \mathrm{Reg}(n-1)$ in Lemma 21. Then, we prove Theorem 1.

Lemma 21 (Regret per Round)**.** Under the laissez-faire decision rule, the regret per round is bounded as:

$$\mathrm{reg}(n) \leq 2 \max_{i \in I(n)} \|\boldsymbol{x}_i\|_{\bar{\boldsymbol{V}}_g^{-1}} \left\| \boldsymbol{\theta}_{g(i)} - \hat{\boldsymbol{\theta}}_{g(i)} \right\|_{\bar{\boldsymbol{V}}_g}.$$

Proof. We denote the first-best decision rule by $i^*(n) := \arg\max_{i \in I(n)} \boldsymbol{x}_i' \boldsymbol{\theta}_{g(i)}$. Then,

$$\begin{aligned}
\mathrm{reg}(n) &= \boldsymbol{x}_{i^*}' \boldsymbol{\theta}_{g(i^*)} - \boldsymbol{x}_\iota' \boldsymbol{\theta}_{g(\iota)} \\
&\leq \boldsymbol{x}_{i^*}' \left(\hat{\boldsymbol{\theta}}_{g(i^*)} + \boldsymbol{\theta}_{g(i^*)} - \hat{\boldsymbol{\theta}}_{g(i^*)} \right) - \boldsymbol{x}_\iota' \left(\hat{\boldsymbol{\theta}}_{g(\iota)} + \boldsymbol{\theta}_{g(\iota)} - \hat{\boldsymbol{\theta}}_{g(\iota)} \right) \\
&\leq \boldsymbol{x}_{i^*}' \left(\boldsymbol{\theta}_{g(i^*)} - \hat{\boldsymbol{\theta}}_{g(i^*)} \right) - \boldsymbol{x}_\iota' \left(\boldsymbol{\theta}_{g(\iota)} - \hat{\boldsymbol{\theta}}_{g(\iota)} \right) \quad \text{(by the greedy choice of firm)}
\end{aligned}$$

$$\leq \|\boldsymbol{x}_{i^*}\|_{\bar{\boldsymbol{V}}_{g(i^*)}^{-1}} \left\|\boldsymbol{\theta}_{g(i^*)} - \hat{\boldsymbol{\theta}}_{g(i^*)}\right\|_{\bar{\boldsymbol{V}}_{g(i^*)}} + \|\boldsymbol{x}_{\iota}\|_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}} \left\|\boldsymbol{\theta}_{g(\iota)} - \hat{\boldsymbol{\theta}}_{g(\iota)}\right\|_{\bar{\boldsymbol{V}}_{g(\iota)}}$$
$$\text{(by the Cauchy–Schwarz inequality)}$$
$$\leq 2 \max_{i \in I(n)} \|\boldsymbol{x}_i\|_{\bar{\boldsymbol{V}}_{g(i)}^{-1}} \left\|\boldsymbol{\theta}_{g(i)} - \hat{\boldsymbol{\theta}}_{g(i)}\right\|_{\bar{\boldsymbol{V}}_{g(i)}}.$$

□

Now, we provide the proof of Theorem 1.

Proof. For ease of discussion, we assume $N^{(0)} = 0$. That is, there is no initial sampling phase. Extending our results to the case of $N^{(0)} > 0$ is trivial. We first show that regardless of estimated values $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, the candidate of group 2 is drawn with constant probability. Let $\mu_x = ||\boldsymbol{\mu}_x||$. Let

$$\mathcal{M}_1(n) = \left\{\boldsymbol{x}_1'(n)\hat{\boldsymbol{\theta}}_1 \leq 0\right\},$$
$$\mathcal{M}_2(n) = \left\{\boldsymbol{x}_2'(n)\hat{\boldsymbol{\theta}}_2 > 0\right\}.$$

The sign of $\boldsymbol{x}_1'\hat{\boldsymbol{\theta}}_1(n)$ is solely determined by the component of $\boldsymbol{x}_1(n)$ that is parallel to $\hat{\boldsymbol{\theta}}_1(n)$. This component is drawn from $\mathcal{N}(\mu_{x,\|}, \sigma_x)$ where $\mu_{x,\|}$ is the component of $\boldsymbol{\mu}_x$ that is parallel to $\hat{\boldsymbol{\theta}}_1(n)$. Therefore, for any $\hat{\boldsymbol{\theta}}_1$, we have[20]

$$\Pr[\mathcal{M}_1(n)] \geq \Phi^c(\mu_x/\sigma_x). \tag{17}$$

Likewise, for $\hat{\boldsymbol{\theta}}_2 \neq 0$, we have[21]

$$\Pr[\mathcal{M}_2(n)] \geq \Phi^c(\mu_x/\sigma_x) \tag{18}$$

Let $\mathcal{X}_2(n) = \{g(\iota(n)) = g\}$ for $g \in \{1, 2\}$. By using Eq. (17) and (18),

$$\begin{aligned}
\Pr[\mathcal{X}_2(n)] &= \Pr[x_1'(n)\hat{\theta}_1 < x_2'(n)\hat{\theta}_2] \\
&\geq \Pr[x_1'(n)\hat{\theta}_1 \leq 0 < x_2'(n)\hat{\theta}_2] \\
&= \Pr[\mathcal{M}_1(n), \mathcal{M}_2(n)] \\
&\geq \left(\Phi^c(\mu_x/\sigma_x)\right)^2. \qquad (19) \\
&\quad \text{(by Eq. (17), (18))}
\end{aligned}$$

[20] $\Pr[\mathcal{M}(n)] = \Phi^c(\mu_x/\sigma_x)$ when $\mu_{x,\|} = \mu_x$. That is, the direction of $\boldsymbol{\mu}_x$ is exactly the same as $\hat{\boldsymbol{\theta}}_1$.

[21] In the subsequent discussion, we do not care point mass $\hat{\boldsymbol{\theta}}_2 = 0$ of measure zero for $N_2(n) > 0$.

Let $N_2^{(\mathcal{M})}(n) = \sum_{n'=1}^{n} \mathbf{1}[\mathcal{M}_1(n'), \mathcal{X}_2(n')] \le N_2(n)$. Eq. (19) implies

$$\mathbb{E}[N_2^{(\mathcal{M})}(n)] \ge \left(\Phi^c(\mu_x/\sigma_x)\right)^2 n.$$

By using the Hoeffding inequality, with a probability at least $1 - 2/N^2$, we have

$$N_2^{(\mathcal{M})} \ge n\left((\Phi^c(\mu_x/\sigma_x))^2 - k\right) \tag{20}$$

for

$$k = \sqrt{\frac{\log(N)}{n}}.$$

Therefore, union bound over $n = 1, 2, \dots, N$ implies Eq. (20) holds with a probability at least $1 - \sum_n 2/N^2 = 1 - 2/N$.

In the following we bound the $\lambda_{\min}(\bar{\boldsymbol{V}}_g)$. Note that a hiring of a worker i_2 under events $\mathcal{M}_1(n), \mathcal{X}_2(n)$ satisfies a diversity condition (Lemma 16) with $\hat{b} = 0$, and we have

$$\lambda_{\min}(\mathbb{E}[\boldsymbol{x}_\iota \boldsymbol{x}_\iota' | \mathcal{M}_1(n), \mathcal{X}_2(n)]) \ge \lambda_0$$

with $\lambda_0 = \sigma_x^2/10$. Using the matrix Azuma inequality (Lemma 19) for subsequence $\{\boldsymbol{x}_\iota \boldsymbol{x}_\iota' : \mathcal{M}_1(n), \mathcal{X}_2(n)\}$ with $\boldsymbol{X} = \boldsymbol{x}_\iota \boldsymbol{x}_\iota' - \mathbb{E}[\boldsymbol{x}_\iota \boldsymbol{x}_\iota']$ and $\sigma_A = 2L^2$, for $t = \sqrt{32 N_2 \sigma_A^2} \log(dN)$, with probability $1 - 1/N$

$$\lambda_{\min}\left(\sum_{n:\iota(n)=2} \boldsymbol{x}_\iota \boldsymbol{x}_\iota'\right) \ge N_2^{(\mathcal{M})} \lambda_0 - t. \tag{21}$$

In summary, with probability $1 - 4/N$, Eq. (20) and (21) hold, and then, we have

$$\begin{aligned}
\lambda_{\min}(\bar{V}_2) &\ge N_2^{(\mathcal{M})} \lambda_0 - \sqrt{32 N_2 \sigma_A^2} \log(dN) \\
&\ge (n(\Phi^c(\mu_x/\sigma_x))^2 - k)\lambda_0 - \sqrt{32 N_2 \sigma_A^2} \log(dN) \\
&= n(\Phi^c(\mu_x/\sigma_x))^2 \lambda_0 - \tilde{O}(\sqrt{n}).
\end{aligned} \tag{22}$$

By using the symmetry of the two groups, exactly the same results as Eq. (22) holds for group 1.

In the following, we bound the regret as a function of $\min_g \lambda_{\min}(\bar{V}_g)$. Eq. (22) holds with probability $1 - O(1/N)$, and we ignore events of probability $O(1/N)$ that do not affect the analysis. The regret is bounded as

$$
\begin{aligned}
\mathrm{Reg}(N) &\le 2\sum_n \max_i \|\boldsymbol{x}_i\|_{\bar{\boldsymbol{V}}_{g(i)}^{-1}} \left\|\boldsymbol{\theta}_{g(i)} - \hat{\boldsymbol{\theta}}_{g(i)}\right\|_{\bar{\boldsymbol{V}}_{g(i)}} \quad \text{(by Lemma 21)} \\
&\le 2\sum_n \max_i \|\boldsymbol{x}_i\|_{\bar{\boldsymbol{V}}_{g(i)}^{-1}} \beta_N \quad \text{(by Eq. 13)} \\
&\le 2\sum_n \max_i \frac{||\boldsymbol{x}_i||}{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}\beta_N \quad \text{(by definition of eigenvalues)} \\
&\le 2\sum_n \max_i \frac{L}{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}\beta_N \quad \text{(by Eq. (12))} \\
&\le 2L\sum_n \max_i \min\left(\frac{1}{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}, \frac{1}{\lambda}\right)\beta_N \quad \text{(by } \lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)}) \ge \lambda) \\
&\le 2L\sum_n \min\left(\sqrt{\frac{1}{n(\Phi^c(\mu_x/\sigma_x))^2\lambda_0 - \tilde{O}(\sqrt{n})}}, \frac{1}{\lambda}\right)\beta_N \quad \text{(by Eq. (22))} \\
&\le 4L\sqrt{\frac{N}{(\Phi^c(\mu_x/\sigma_x))^2\lambda_0}}\beta_N + \tilde{O}(1) \\
&\qquad \left(\text{by} \sum_{n=C^2+1}^{N}\left\{\frac{1}{\sqrt{n - C\sqrt{n}}}\right\} = 2\sqrt{N} + \tilde{O}(1) \text{ for } C = \tilde{O}(1))\right)
\end{aligned}
\tag{23}
$$

which completes Proof of Theorem 1. □

2 Proof of Theorem 2

Proof. Since we consider $d = 1$ case in this theorem, we remove bold styles in scalar variables. In this proof, we assume $\mu_x\theta > 0$ and $\theta > 0$. The proof for the case of $\mu_x\theta < 0$ or $\theta < 0$ is similar. Let $\hat{\theta}_{g,t}$ be the value of $\hat{\theta}_g$ when group g candidate was chosen t times. With a slight abuse of notation, we use $i_2 = i_2(n)$ to denote the unique candidate of group 2 in each round n. We first define the several events that characterize the perpetual underestimation. That are,

$$
\begin{aligned}
\mathcal{P} &= \left\{\left|\hat{\theta}_{2,N_2^{(0)}}\right| < \frac{b}{2}\theta\right\} \\
\mathcal{P}'(n) &= \left\{x_{i_2(n)}\hat{\theta}_{2,N_2^{(0)}} < \frac{1}{2}\mu_x\theta\right\} \\
\mathcal{Q} &= \left\{\forall t \ge N_1^{(0)},\ \hat{\theta}_{1,t} \ge \frac{1}{2}\theta\right\} \\
\mathcal{Q}'(n) &= \left\{\exists i \text{ s.t. } g(i) = 1, x_i\hat{\theta}_{1,N_1(n)} \ge \frac{1}{2}\mu_x\theta\right\}
\end{aligned}
$$

where b is a small[22] constant that we specify later. $\mathcal{P}$ and $\mathcal{P}'$ are about the minority whereas $\mathcal{Q}$ and $\mathcal{Q}'$ are about the majority: Intuitively, Event $\mathcal{P}$ states that $\hat{\theta}_2$ is largely underestimated, and $\mathcal{P}'$ states that the minority candidate is undervalued. $\mathcal{Q}$ states that the majority parameter $\hat{\theta}_1$ is consistently lower-bounded, and $\mathcal{Q}'$ states the stability of the best candidate of the majority after n rounds. Under laissez-faire,

$$\bigcap_{n=N^{(0)}+1}^{N} (\mathcal{P}'(n) \cap \mathcal{Q}'(n))$$

implies the majority candidate is always chosen ($g(\iota) = 1$ for all n), which is exactly the perpetual underestimation of Definition 2. Therefore, proving

$$\Pr\left[\bigcap_{n=N^{(0)}+1}^{N} (\mathcal{P}'(n) \cap \mathcal{Q}'(n))\right] \geq \tilde{O}(1) \tag{24}$$

concludes the proof. We bound these events by the following lemmas and finally derives Eq. (24).

Lemma 22.

$$\Pr[\mathcal{P}] \geq C_1 b$$

for some constant C_1.

Proof. We denote $x_{i_2,t}$ for representing t-th sample of group 2 during the initial sampling phase, which is an i.i.d. sample from $\mathcal{N}(\mu_x, \sigma_x^2)$. Likewise, we also denote $y_{i_2,t} = x_{i_2,t}\theta + \epsilon_t$.

$$\begin{aligned}
\Pr[\mathcal{P}] &= \Pr\left[\left|\frac{\sum_{t=1}^{N_2^{(0)}} x_{i_2,t}(x_{i_2,t}\theta + \epsilon_t)}{\sum_{t=1}^{N_2^{(0)}} x_{i_2,t}^2 + \lambda}\right| \leq \frac{b}{2}\theta\right] \\
&= \Pr\left[\left|\sum_{t=1}^{N_2^{(0)}} x_{i_2,t}(x_{i_2,t}\theta + \epsilon_t)\right| \leq \frac{b}{2}\theta\left(\sum_{t=1}^{N_2^{(0)}} x_{i_2,t}^2 + \lambda\right)\right] \\
&= \Pr\left[-g(b) \leq \sum_{t=1}^{N_2^{(0)}} x_{i_2,t}(x_{i_2,t}\theta + \epsilon_t) \leq g(b)\right]
\end{aligned}$$

where

$$g(b) = \frac{b}{2}\theta\left(\sum_{t=1}^{N_2^{(0)}} x_{i_2,t}^2 + \lambda\right).$$

[22]We will specify $b = O(1/(\log N))$.

Let $x_{i_2,t} = \mu_x + e_t$. Define an event $\mathcal{R}$ as follows.

$$\mathcal{R} = \left\{ \sum_{t=1}^{N_2^{(0)}} e_t^2 \le 5\sigma_x^2 N_2^{(0)} \right\} \subseteq \left\{ \sum_{t=1}^{N_2^{(0)}} x_{i_2,t}^2 \le 2N_2^{(0)}(\mu_x^2 + 5\sigma_x^2) \right\}$$

where we used $x_{i_2,t}^2 = (\mu_x + e_t)^2 \le 2(\mu_x^2 + e_t^2)$ in the last transformation. By using Lemma 12, we have

$$\Pr[\mathcal{R}^c] \le 1 - 2e^{-2N_2^{(0)}} \le 1/4.$$

Moreover, let

$$\mathcal{S} = \left\{ \sum_{t=1}^{N_2^{(0)}} x_{i_2,t}^2 = \sum_{t=1}^{N_2^{(0)}} (\mu_x + e_t)^2 \ge N_2^{(0)} \mu_x^2 \right\}.$$

It is easy to confirm that $\Pr[\sum_n (\mu_x + e_t)^2 \ge N_2^{(0)} \mu_x^2] \ge 1/2$, and thus

$$\Pr[\mathcal{R} \cap \mathcal{S}] \ge 1 - 1/4 - 1/2 = 1/4. \tag{25}$$

Note that $\mathcal{S}$ implies

$$g(b) \ge \frac{b}{2} \theta N_2^{(0)} \mu_x + \lambda. \tag{26}$$

Conditioned on $x_{i_2,t}$, we have $x_{i_2,t}\epsilon_t \sim \mathcal{N}(0, x_{i_2,t}^2 \sigma_\epsilon^2)$. Moreover, by using the property on the sum of independent normal random variables,

$$\sum_t x_{i_2,t}\epsilon_t \sim \mathcal{N}(0, \sum_t x_{i_2,t}^2 \sigma_\epsilon^2). \tag{27}$$

Letting

$$\begin{aligned}
L_R &= \frac{-g(b) - \sum_t x_{i_2,t}^2 \theta}{\sigma_\epsilon \sqrt{\sum_t x_{i_2,t}^2}}, \\
U_R &= \frac{g(b) - \sum_t x_{i_2,t}^2 \theta}{\sigma_\epsilon \sqrt{\sum_t x_{i_2,t}^2}}, \\
M_R &= \frac{L_R + U_R}{2} = \frac{-\left(\sqrt{\sum_t x_{i_2,t}^2}\right)\theta}{\sigma_\epsilon},
\end{aligned}$$

we have

$$\Pr\left[-g(b) \le \sum_{t=1} (x_{i_2,t}^2 \theta + x_{i_2,t}\epsilon_n) \le g(b) \right]$$

$$
\begin{aligned}
&\geq \Pr\left[-g(b) \leq \sum_{t=1}(x_{i_2,t}^2\theta + x_{i_2,t}\epsilon_t) \leq g(b), \mathcal{R}, \mathcal{S}\right] \\
&\geq \Pr\left[-g(b) - \sum_{t=1} x_{i_2,t}^2\theta \leq \sum_{t=1} x_{i_2,t}\epsilon_t \leq g(b) - \sum_{t=1} x_{i_2,t}^2\theta, \mathcal{R}, \mathcal{S}\right] \\
&\geq \Pr[\mathcal{R}, \mathcal{S}] \min_{\{e_n:\mathcal{R},\mathcal{S}\}} \left[\int_{L_R}^{U_R} \phi(y)dy\right] \quad \text{(by Eq. (27))} \\
&\geq \frac{1}{4} \min_{\{e_n:\mathcal{R},\mathcal{S}\}} \left[\int_{L_R}^{U_R} \phi(y)dy\right]. \quad \text{(by Eq. (25))}
\end{aligned} \tag{28}
$$

The following bounds Eq. (28). The integral's bandwidth is

$$
U_R - L_R = \frac{2g(b)}{\sigma_\epsilon\sqrt{\sum_t x_{i_2,t}^2}} \geq \frac{2g(b)}{\sigma_\epsilon\sqrt{2N_2^{(0)}(\mu_x^2 + 5\sigma_x^2)}}. \quad \text{(by event } \mathcal{R})
$$

The value of $\phi(y)$ within $[M_R - 1, M_R + 1]$ is at least $\phi(M_R)/e^{1/2} \geq (1/2)\phi(M_R)$. Therefore,

$$
\int_{L_R}^{U_R} \phi(y)dy \geq \min\left(2, \frac{2g(b)}{\sigma_\epsilon\sqrt{2N_2^{(0)}(\mu_x^2 + 5\sigma_x^2)}}\right) \times \frac{\phi(M_R)}{2}. \tag{29}
$$

Moreover,

$$
\begin{aligned}
\phi(M_R) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(M_R)^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2 \sum_{t=1}^{N_2^{(0)}} x_{i_2,t}^2}{2\sigma_\epsilon^2}\right) \\
&\leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{2\theta^2 N_2^{(0)}(\mu_x^2 + 5\sigma_x^2)}{2\sigma_\epsilon^2}\right). \quad \text{(by event } \mathcal{R})
\end{aligned} \tag{30}
$$

By using these, we have

$$
\begin{aligned}
\int_{L_R}^{U_R} \phi(y)dy &\geq \min\left(2, \frac{2g(b)}{\sigma_\epsilon\sqrt{2(\mu_x^2 + 5\sigma_x^2)}}\right) \frac{\phi\left(M_R\right)}{2} \quad \text{(by Eq. (29))} \\
&= \min\left(1, \frac{g(b)}{\sigma_\epsilon\sqrt{2N_2^{(0)}(\mu_x^2 + 5\sigma_x^2)}}\right) \phi(M_R) \\
&= O\left(b\sqrt{N_2^{(0)}} \exp\left(-\frac{2\theta^2 N_2^{(0)}(\mu_x^2 + 5\sigma_x^2)}{2\sigma_\epsilon^2}\right)\right). \quad \text{(by Eq. (26), (30))}
\end{aligned}
$$

The exponent does not depend on b: Given all model parameters as constant, the probability of $\mathcal{P}$ is $O(b)$, which concludes the proof. $\square$

The following Lemma 23 on $\mathcal{Q}$ is about the stability of the mean estimator, which is widely used to prove lower bounds in multi-armed bandit problems. That is, for any $\Delta > 0$, a wide class of mean estimators $\hat{\theta}$ of θ satisfies

$$\Pr\left[\bigcup_{n=1}^{\infty}\left(\hat{\theta}(n) \geq \theta - \Delta\right)\right] \geq C \tag{31}$$

for some constant $C = C(\theta, \Delta) > 0$. Lemma 23 is a version Eq. (31) for our ridge estimator.

Lemma 23. There exists a constant n that is independent on N such that, with a warm-start of size $N_1^{(0)} \geq n$,

$$\Pr[\mathcal{Q}] \geq C_2$$

holds with $C_2 = 1/4$.

Proof. In this proof, we use $t \geq 0$ to denote the estimator where the t-th sample is drawn. For example, $\bar{V}_{g,t} := \bar{V}_g(n)$ of $n : N_1(n-1) = t$. Note that we consider $d = 1$ case and $\bar{V}_{1,t} = \sum_{t'=1}^{t} x_{1,t}^2 + \lambda$. By martingale bound (Eq. (10)), with probability $1-\delta$,

$$\forall t \geq 1, \quad |\hat{\theta}_{1,t} - \theta|\sqrt{\bar{V}_{1,t}} \leq \sigma_\epsilon \sqrt{\log\left(\frac{\bar{V}_{1,t}^{1/2}\lambda^{-1/2}}{\delta}\right)} + \lambda^{1/2}S. \tag{32}$$

Let $\delta = 1/2$. It follows from $\sqrt{\log x} \leq \sqrt{x}$ for any $x > 0$ that

$$\sqrt{\log\left(2\bar{V}_{1,t}^{1/2}\lambda^{-1/2}\right)} \leq \sqrt{2\bar{V}_{1,t}^{1/2}\lambda^{-1/2}}. \tag{33}$$

Therefore,

$$\begin{aligned}
|\hat{\theta}_{1,t} - \theta| &\leq \frac{\sigma_\epsilon \sqrt{\log\left(\frac{\bar{V}_{1,t}^{1/2}\lambda^{-1/2}}{\delta}\right)} + \lambda^{1/2}S}{\sqrt{\bar{V}_{1,t}}} \quad \text{(by Eq. (32))} \\
&\leq \frac{\sigma_\epsilon \sqrt{2\bar{V}_{1,t}^{1/2}\lambda^{-1/2}} + \lambda^{1/2}S}{\sqrt{\bar{V}_{1,t}}} \quad \text{(by (33))}
\end{aligned}$$

and thus

$$\forall t \geq N_1^{(0)},\ |\hat{\theta}_{1,t} - \theta| \leq \frac{1}{2}|\theta|$$

holds if

$$\sqrt{\bar{V}_{1,N_1^{(0)}}} \geq 2\theta \max\left(\sigma_\epsilon \sqrt{2\bar{V}_{1,N_1^{(0)}}^{1/2} \lambda^{-1/2}}, \lambda^{1/2} S\right)$$

whose sufficient condition for the initial sample size $N_1^{(0)}$ is

$$\bar{V}_{1,N_1^{(0)}} \geq \max\left[\frac{64}{\theta^4}(\sigma_\epsilon^4/\lambda^2), \frac{4}{\theta^2}\lambda S^2\right].$$

Note that $\Pr[\bar{V}_{1,N_1^{(0)}} \geq \mu_x^2 N_1^{(0)}] \geq 1/2$. Letting the observation noise σ_ϵ and regularizer λ be constants, constant size of warm-start is enough to assure this bound with probability $C_2 = 1/2 \times 1/2 = 1/4$. □

The following lemma states that, when $\hat{\theta}_2$ is very small, the estimated quality $x_{i_2}\hat{\theta}_2$ of the minority group is likely to be small.

Lemma 24. There exists a constant C_3, C_4 that is independent of N such that

$$\Pr[\mathcal{P}'(n)|\mathcal{P}] \geq 1 - C_3 \exp\left(-C_4/b\right) \tag{34}$$

holds.

Proof.

$$\begin{aligned}
\Pr[\mathcal{P}'(n)|\mathcal{P}] &\geq 1 - \Pr\left[x_{i_2}(n) \geq \frac{2}{b}\right] \\
&\geq 1 - \Phi^c\left(\frac{1}{\sigma_x}\left(\frac{2}{b} - \mu_x\right)\right) \\
&\geq 1 - \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{\sigma_x}\left(\frac{2}{b} - \mu_x\right)\right), \quad \text{(by Lemma 13)}
\end{aligned}$$

where we have assumed $\left(\frac{2}{b} - \mu\right)/\sigma_x \geq 1$ in the last transformation (which holds for sufficiently small b). Eq. (34) holds for $C_3 = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\mu/\sigma_x}$ and $C_4 = 2/\sigma_x$. □

Lemma 25.

$$\Pr[\mathcal{Q}'(n) \mid \mathcal{Q}] \geq 1 - (1/2)^{K_1}.$$

Event $\mathcal{Q}'(n)$ states that all the candidates' estimated quality $x_i\hat{\theta}$ is not below mean. Lemma 25 states that the probability of $\mathcal{Q}'(n)$ is exponentially small to the number of candidates. The proof of Lemma 25 directly follows from the symmetry of normal distribution and independence of characteristics $\boldsymbol{x}_i$.

Proof of Theorem 2, continued. By using Lemmas 22–25, we have

$$\Pr[\mathcal{P}] \geq C_1 b \tag{35}$$
$$\Pr[\mathcal{Q}] \geq C_2 \tag{36}$$
$$\Pr[\mathcal{P}'(n)|\mathcal{P}] \geq 1 - C_3 \exp(-C_4 b) \tag{37}$$
$$\Pr[\mathcal{Q}'(n)|\mathcal{Q}] \geq 1 - (1/2)^{K_1}.$$

From these equations, the probability of perpetual underestimation is bounded as:

$$\begin{aligned}
&\Pr\left[\bigcup_n \{\iota(n) = 1\}\right] \\
&\geq \Pr\left[\bigcup_n \{\mathcal{P}'(n), \mathcal{Q}'(n)\}, \mathcal{P}, \mathcal{Q}\right] \\
&\geq \Pr[\mathcal{P}]\Pr[\mathcal{Q}]\Pr\left[\bigcup_n \{\mathcal{P}'(n), \mathcal{Q}'(n)\} \mid \mathcal{P}, \mathcal{Q}\right] \quad \text{(by the independence of } \mathcal{P} \text{ and } \mathcal{Q}) \\
&\geq C_1 b \times C_2 \times (1 - N C_3 \exp(-C_4 b)) \times \left(1 - N\left(\frac{1}{2}\right)^{K_1}\right) \quad \text{(by the union bound)} \qquad (38)
\end{aligned}$$

which, by letting $b = O(1/\log(N))$ and $K_1 > \log_2(N)$, is $\tilde{O}(1)$. □

3 Proof of Theorem 3

Proof. Let $\mathrm{reg}(n) = \mathrm{Reg}(n) - \mathrm{Reg}(n-1)$. Notice that under the UCB decision rule,

$$\iota(n) = \max_{i \in I(n)} (\boldsymbol{x}_i' \tilde{\boldsymbol{\theta}}_i(n)). \tag{39}$$

By Lemma 17, with a probability at least $1 - \delta$, the true parameter of group g lies in $\mathcal{C}_g$, and thus

$$\boldsymbol{x}_i' \tilde{\boldsymbol{\theta}}_i(n) \geq \boldsymbol{x}_i' \boldsymbol{\theta}_g \tag{40}$$

for each $i \in I(n)$.

Let $i^* = i^*(n) := \arg\max_{i \in I(n)} \boldsymbol{x}_i' \boldsymbol{\theta}_{g(i)}$ be the first-best worker, and $g^* = g(i^*)$ be the group i^* belongs to. The regret in round n is bounded as

$$\mathrm{reg}(n) = \boldsymbol{x}_{i^*}' \boldsymbol{\theta}_{g^*} - \boldsymbol{x}_\iota' \boldsymbol{\theta}_{g(\iota)}$$

$$
\begin{aligned}
&\le \boldsymbol{x}_{i^*}'\tilde{\boldsymbol{\theta}}_{i^*} - \boldsymbol{x}_\iota'\boldsymbol{\theta}_{g(\iota)} \quad \text{(by Eq. (40))}\\
&\le \boldsymbol{x}_\iota'\tilde{\boldsymbol{\theta}}_\iota - \boldsymbol{x}_\iota'\boldsymbol{\theta}_{g(\iota)} \quad \text{(by Eq. (39))}\\
&\le ||\boldsymbol{x}_\iota'||_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}} \left\|\boldsymbol{\theta}_{g(\iota)} - \tilde{\boldsymbol{\theta}}_\iota\right\|_{\bar{\boldsymbol{V}}_{g(\iota)}} \quad \text{(by the Cauchy–Schwarz inequality)}\\
&\le ||\boldsymbol{x}_\iota'||_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}}\beta_N. \quad \text{(by Eq. (13))}
\end{aligned} \tag{41}
$$

The total regret is bounded as:

$$
\begin{aligned}
\mathrm{Reg}(N) = \sum_n \mathrm{reg}(n) &\le \sqrt{N\sum_n \mathrm{reg}(n)^2} \quad \text{(by the Cauchy–Schwarz inequality)}\\
&\le 2\beta_N\sqrt{N\sum_n ||\boldsymbol{x}_\iota'||^2_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}}(n)}\\
&\le 2\beta_N\sqrt{2NL^2\sum_{g\in G}\log(\det(\bar{\boldsymbol{V}}_g(N)))} \quad \text{(by Lemma 18)}\\
&\le \tilde{O}(\sqrt{N|G|})
\end{aligned} \tag{42}
$$

where we have used the fact that $\log(\det(\bar{\boldsymbol{V}}_g)) = O(\log(N)) = \tilde{O}(1)$. □

4 Proof of Theorem 4

Proof. We bound the amount of total subsidy $\mathrm{Sub}(N)$.

$$
\begin{aligned}
\mathrm{Sub}(N) &:= \sum_n \boldsymbol{x}_{\iota(n)}'(\tilde{\boldsymbol{\theta}}_\iota - \hat{\boldsymbol{\theta}}_{g(\iota)})\\
&\le ||\boldsymbol{x}_\iota'||_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}}\beta_N, \quad \text{(by Eq. (14))}
\end{aligned}
$$

which is the same as Eq. (41) and thus the same bound as regret applies. □

5 Proof of Theorem 5

Proof. We adopt "slot" notation for each group. Group g is allocated K_g slots and at each round n, one candidate arrives for each slot. We use index $i \in [K]$ to denote each slot: Although $\boldsymbol{x}_i$ at two different rounds n,n' $(= \boldsymbol{x}_i(n), \boldsymbol{x}_i(n'))$ represent different candidates, they are from the identical group $g = g(i)$. In summary, we use index i to represent the i-th slot and with a slight abuse of argument. We also call candidate i to represent the candidate of slot i. Note that this does not change any part of the model, and the slot notation here is for the sake of analysis.

Under the hybrid decision rule, a firm at each round hires the candidate of the largest index. That is,

$$\iota(n) = \arg\max_{i \in I(n)} \tilde{q}_i^{\mathrm{H}}(n)$$

where $\tilde{q}_i^{\mathrm{H}}$ is defined at Eq. (3). We also denote $\tilde{\iota}(n) = \arg\max_{i \in I(n)} \boldsymbol{x}_i' \tilde{\boldsymbol{\theta}}_i$. That is, $\tilde{\iota}$ indicates the candidate who would have been hired if we have used the standard UCB decision rule (Eq. (2))

The following bounds the regret into estimation errors of $\tilde{\iota}$ and ι.

$$\begin{aligned}
\mathrm{reg}(n) &= \boldsymbol{x}_{i^*}' \boldsymbol{\theta}_{g^*} - \boldsymbol{x}_{\iota}' \boldsymbol{\theta}_{g(\iota)} \\
&\le \boldsymbol{x}_{i^*}' \tilde{\boldsymbol{\theta}}_{i^*} - \boldsymbol{x}_{\iota}' \boldsymbol{\theta}_{g(\iota)} \quad \text{(by Eq. (15))} \\
&\le \boldsymbol{x}_{\tilde{\iota}}' \tilde{\boldsymbol{\theta}}_{\tilde{\iota}} - \boldsymbol{x}_{\iota}' \boldsymbol{\theta}_{g(\iota)} \quad \text{(by definition of } \tilde{\iota}\text{)} \\
&= \boldsymbol{x}_{\tilde{\iota}}' \tilde{\boldsymbol{\theta}}_{\tilde{\iota}} - \boldsymbol{x}_{\iota}' \tilde{\boldsymbol{\theta}}_{\iota} + \boldsymbol{x}_{\iota}' (\tilde{\boldsymbol{\theta}}_{\iota} - \boldsymbol{\theta}_{g(\iota)}) \\
&\le \boldsymbol{x}_{\tilde{\iota}}' (\tilde{\boldsymbol{\theta}}_{\tilde{\iota}} - \hat{\boldsymbol{\theta}}_{g(\tilde{\iota})}) + \boldsymbol{x}_{\iota}' (\tilde{\boldsymbol{\theta}}_{\iota} - \boldsymbol{\theta}_{g(\iota)}). \quad \text{(by definition of } \iota\text{)}
\end{aligned} \tag{43}$$

Here,

$$\begin{aligned}
\boldsymbol{x}_{\tilde{\iota}}' (\tilde{\boldsymbol{\theta}}_{\tilde{\iota}} - \hat{\boldsymbol{\theta}}_{g(\tilde{\iota})}) &\le ||\boldsymbol{x}_{\tilde{\iota}}'||_{\bar{\boldsymbol{V}}_{g(\tilde{\iota})}^{-1}} \left\| \tilde{\boldsymbol{\theta}}_{\tilde{\iota}} - \hat{\boldsymbol{\theta}}_{g(\tilde{\iota})} \right\|_{\bar{\boldsymbol{V}}_{g(\tilde{\iota})}} \quad \text{(by the Cauchy–Schwarz inequality)} \\
&\le ||\boldsymbol{x}_{\tilde{\iota}}'||_{\bar{\boldsymbol{V}}_{g(\tilde{\iota})}^{-1}} \beta_N. \quad \text{(by Eq. (14))} \\
&\le \frac{||\boldsymbol{x}_{\tilde{\iota}}'||}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(\tilde{\iota})})}} \beta_N \quad \text{(by definition of eigenvalues)} \\
&\le \frac{L}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(\tilde{\iota})})}} \beta_N. \quad \text{(by Eq. (12))}
\end{aligned} \tag{44}$$

Moreover, the estimation error of candidate ι is bounded as

$$\begin{aligned}
\boldsymbol{x}_{\iota}' (\tilde{\boldsymbol{\theta}}_{\iota} - \boldsymbol{\theta}_{g(\iota)}) &\le ||\boldsymbol{x}_{\iota}'||_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}} \left\| \tilde{\boldsymbol{\theta}}_{\iota} - \boldsymbol{\theta}_{g(\iota)} \right\|_{\bar{\boldsymbol{V}}_{g(\iota)}} \quad \text{(by the Cauchy–Schwarz inequality)} \\
&\le 2||\boldsymbol{x}_{\iota}'||_{\bar{\boldsymbol{V}}_{g(\iota)}^{-1}} \beta_N \quad \text{(by Eq. (16))} \\
&\le \frac{2||\boldsymbol{x}_{\iota}'||}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(\iota)})}} \beta_N \quad \text{(by definition of eigenvalues)} \\
&\le \frac{2L}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(\iota)})}} \beta_N. \quad \text{(by Eq. (12))}
\end{aligned} \tag{45}$$

Based on the above bounds, the regret is bounded as follows.

$$
\begin{aligned}
\mathrm{Reg}(N) &= \sum_{n=1}^{N} \mathrm{reg}(n) \\
&\le \sum_{n=1}^{N} \left(\frac{2}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(\iota)})}} + \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(\tilde{\iota})})}} \right) L\beta_N \\
&\quad \text{(by Eq.(43), (44), (45))} \\
&\le 2L\beta_N \sum_{i\in[K]} \sum_{n=1}^{N} \mathbf{1}[\iota = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} \\
&\quad + L\beta_N \sum_{i\in[K]} \sum_{n=1}^{N} \mathbf{1}[\tilde{\iota} = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}}.
\end{aligned} \tag{46}
$$

Eq. (46) consisted of two components. The first component is the estimation error of the hired candidate ι. The second component is the estimation error of $\tilde{\iota}$, the candidate who would have hired if we had posed the UCB decision rule. The Hybrid decision rule ι can be different from the UCB decision rule $\tilde{\iota}$, which is the main challenge of deriving regret bound in the hybrid decision rule.

We first define the following events

$$
\begin{aligned}
\mathcal{V}_i(n) &:= \left\{ \boldsymbol{x}_i(n)'(\tilde{\boldsymbol{\theta}}_i(n) - \hat{\boldsymbol{\theta}}_{g(i)}(n)) \le a\sigma_x \left\| \hat{\boldsymbol{\theta}}(n) \right\| \right\}, \\
\mathcal{W}_i(n) &:= \{\tilde{\iota}(n) = i\}, \\
\mathcal{X}_i(n) &:= \{\iota(n) = i\}, \\
\mathcal{X}'_i(n) &:= \left\{ \boldsymbol{x}_i(n)'\hat{\boldsymbol{\theta}}(n) \ge \arg\max_{j\neq i} \tilde{q}_j^{\mathrm{H}} \right\} \subseteq \mathcal{X}_i.
\end{aligned}
$$

Event $\mathcal{V}_i$ states that the candidate i is not subsidized. Event $\mathcal{W}_i$ states that i would have been hired if it was subsidized in the UCB decision rule. Event $\mathcal{X}_i$ states that i is hired and $\mathcal{X}'_i$ states that i is hired regardless of the subsidy.

The following lemma is the crux of bounding the components in Eq. (46).

Lemma 26 (Proportionality)**.** The following two inequalities hold.

$$
\Pr[\mathcal{X}'_i] \ge \exp(-a^2/2)\Pr[\mathcal{W}_i], \tag{47}
$$

$$
\Pr[\mathcal{X}'_i] \ge \exp(-a^2/2)\Pr[\mathcal{X}_i]. \tag{48}
$$

Proof. We first prove, for any $c \in \mathbb{R}$, $d > 0$,

$$\Pr\left[\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)} \geq c\right] \geq \exp(-d^2/2) \Pr\left[\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)} \geq c - d\left(\sigma_x \left\|\hat{\boldsymbol{\theta}}_{g(i)}\right\|\right)^2\right]. \tag{49}$$

Let $x_{\|} := (\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)})/||\hat{\boldsymbol{\theta}}_{g(i)}||$ be the projection of $\boldsymbol{x}_i$ into the direction of $\hat{\boldsymbol{\theta}}_{g(i)}$. Then, $\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)} = x_{\|}||\hat{\boldsymbol{\theta}}_{g(i)}||$. From the symmetry of a normal distribution, $x_{\|}||\hat{\boldsymbol{\theta}}_{g(i)}||$ is drawn from a normal distribution with its variance $(\sigma_x||\hat{\boldsymbol{\theta}}_{g(i)}||)^2$, from which Eq. (49) follows.

Eq. (47) follows by letting $c = \max_{j\neq i} \tilde{q}_j^{\mathrm{H}}$, $d = a$ because

$$\begin{aligned} \mathcal{W}_i &\subseteq \left\{\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)} \geq c - d\left(\sigma_x \left\|\hat{\boldsymbol{\theta}}_{g(i)}\right\|\right)^2\right\} \\ \mathcal{X}_i' &\supseteq \left\{\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)} \geq c\right\} \end{aligned}$$

Eq. (48) also follows by letting $c = \max_{j\neq i} \tilde{q}_j^{\mathrm{H}}$ and $d = a$

$$\begin{aligned} \mathcal{X}_i &\subseteq \left\{\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}}_i \geq c\right\} \\ \mathcal{X}_i' &\supseteq \left\{\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}}_i \geq c + d\left(\sigma_x \left\|\hat{\boldsymbol{\theta}}_{g(i)}\right\|\right)^2\right\} \end{aligned}$$

and exactly the same discussion as Eq. (49) applies for[23]

$$\Pr\left[\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}}_i \geq c + d\left(\sigma_x \left\|\hat{\boldsymbol{\theta}}_{g(i)}\right\|\right)^2\right] \geq \exp(-d^2/2) \Pr\left[\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}}_i \geq c\right]. \tag{50}$$

□

Lemma 26 is intuitively understood as follows. Assume that candidate i would have been hired under the UCB rule. The candidate may not be hired under the hybrid rule because it can cut subsidies for that candidate. However, there is a constant probability such that a slightly better ("$a\sigma$-good") candidate appears on slot i, and such a candidate is hired under the hybrid rule.

The following two lemmas, which utilizes Lemma 26, bounds the two terms of Eq. (46).

Lemma 27.

$$\mathbb{E}\left[\sum_{n=1}^{N} \mathbf{1}[\iota = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}}\right] \leq \frac{2e^{a^2/4}}{\lambda_0}\sqrt{N} + O(1).$$

[23] Note that $\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)}$ in Eq. (49) is replaced by $\boldsymbol{x}_i'\tilde{\boldsymbol{\theta}}_i$ in Eq. (50), which does not change the subsequent derivations at all.

Lemma 28.

$$\mathbb{E}\left[\sum_{n=1}^{N}\mathbf{1}[\tilde{\iota}=i]\frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}}\right] \leq \frac{2e^{a^2/4}}{\lambda_0}\sqrt{N}+O(1).$$

With Lemmas 27 and 28, the regret is bounded as

$$\begin{aligned}
\mathrm{Reg}(N) &\leq 2L\beta_n(L,1/N)\sum_{i\in[K]}\sum_{n=1}^{N}\mathbf{1}[\iota=i]\frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}}\\
&\quad + L\beta_n(L,1/N)\sum_{i\in[K]}\sum_{n=1}^{N}\mathbf{1}[\tilde{\iota}=i]\frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} \quad \text{(by Eq. (46))}\\
&\leq 6L\beta_n(L,1/N)K\frac{e^{a^2/4}\sqrt{N}}{\lambda_0}+\tilde{O}(1) \quad \text{(by Lemma 27 and 28)}
\end{aligned} \tag{51}$$

which completes the proof of Theorem 5. □

The following is the proof of Lemma 27.

Proof. Let $N_i(n)$ be the number of the rounds before n such that the worker of slot i is selected. Let τ_t be the first round such that $N_i(n)$ reaches t and $N_{i,t}=\sum_{n\leq\tau_t}\mathbf{1}[\mathcal{X}_i'(n)]$. Lemma 26 implies $\mathbb{E}[N_{i,t}]\geq e^{-a^2/2}t$ and applying the Hoeffding inequality on binary random variables $(\mathbf{1}[\mathcal{X}_i'(\tau_1)],\mathbf{1}[\mathcal{X}_i'(\tau_2)],\ldots,...,\mathbf{1}[\mathcal{X}_i'(\tau_t)])$ yields

$$\Pr\left[N_{i,t}<\left(e^{-a^2/2}t-\sqrt{(\log N)t}\right)\right]\leq\frac{2}{N^2}. \tag{52}$$

By using this, we have

$$\begin{aligned}
&\Pr\left[\bigcap_{t=1}^{N}\left\{N_{i,t}<\left(e^{-a^2/2}t-\sqrt{(\log N)t}\right)\right\}\right]\\
&\leq\sum_t\Pr\left[N_{i,t}<\left(e^{-a^2/2}t-\sqrt{(\log N)t}\right)\right] \quad \text{(by union bound)}\\
&\leq\sum_t\frac{2}{N^2} \quad \text{(by Eq. (52))}\\
&\leq\frac{2}{N}.
\end{aligned}$$

In the following, we focus on the case

$$N_{i,t}\geq e^{-a^2/2}t-\sqrt{(\log N)t}, \tag{53}$$

which occurs with a probability at least $1 - 2/N$.

Let $\bar{\boldsymbol{V}}_i(n) := \sum_{n' \le n} \mathbf{1}[\iota = i]\boldsymbol{x}_i\boldsymbol{x}_i' \preceq \bar{\boldsymbol{V}}_{g(i)}(n)$. The context $\boldsymbol{x}_i$ conditioned on event $\mathcal{X}_i'$ satisfies assumptions in Lemma 16 with $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_i$ and $\hat{b} = \max_{j \ne i} \tilde{q}_j^{\mathrm{H}}$. We have,

$$
\begin{aligned}
\sum_{n=1}^{N} \mathbf{1}[\iota = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} &\le \sum_{n=1}^{N} \mathbf{1}[\iota = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i)}} \quad (\text{by } \bar{\boldsymbol{V}}_{g(i)} \succeq \bar{\boldsymbol{V}}_i) \\
&\le \sum_{n=1}^{N} \sum_{t=1}^{N} \mathbf{1}[\iota = i, N_i(n) = t] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i)}} \\
&\quad (\text{by } N_i(N) \le N) \\
&\le \sum_{t=1}^{N} \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))}}. \\
&\quad (\text{by } \mathbf{1}[\iota = i, N_i(n) = t] \text{ occurs at most once})
\end{aligned}
$$

In other words, lower-bounding $\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))$ suffices the regret bound, which we demonstrate in the following.

We have

$$
\begin{aligned}
\mathbb{E}\left[\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))\right] &\ge \lambda_{\min}\Big(\sum_n \mathbb{E}[\mathbf{1}[\mathcal{X}_i'(n)]\boldsymbol{x}_i\boldsymbol{x}_i']\Big) \\
&\ge \lambda_0 N_{i,t}. \quad (\text{by Lemma 16})
\end{aligned}
$$

By using the matrix Azuma inequality (Lemma 19), with probability of at least $1 - 1/N$

$$
\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t)) \ge \left(\lambda_0 N_{i,t} - \sqrt{32 N_{i,t} \sigma_A^2 \log(dN)}\right) \tag{54}
$$

where $\sigma_A = 2L^2$. By using Eq. (53), (54), we have

$$
\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t)) \ge \lambda_0 e^{-a^2/2} t - O(\sqrt{t})
$$

and thus

$$
\begin{aligned}
\sum_{t=1}^{N} \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))}} &\le \sum_{t=1}^{N} \frac{1}{\sqrt{\lambda_0 e^{-a^2/2} t - O(\sqrt{t})}} \\
&\le \frac{2e^{a^2/4}}{\lambda_0} \sqrt{N} + O(1).
\end{aligned}
$$

□

The following is the proof of Lemma 28.

Proof. Let $N_i^{\mathcal{W}_i}(n) = \sum_{n' \le n} \mathbf{1}[\mathcal{W}_i]$ and let τ_t be the first round such that $N_i^{\mathcal{W}_i}(n)$ reaches t and $N_{i,t} = \sum_{n \le \tau_t} \mathbf{1}[\mathcal{X}_i'(n)]$. The following discussions are very similar to the one of Lemma 27, which we write for completeness. Then, we have

$$
\begin{aligned}
&\Pr\left[\bigcap_{t=1}^{N} \left\{ N_{i,t} < \left(e^{-a^2/2}t - \sqrt{(\log N)t}\right)\right\}\right] \\
&\le \sum_t \Pr\left[N_{i,t} < \left(e^{-a^2/2}t - \sqrt{(\log N)t}\right)\right] \quad \text{(by union bound)} \\
&\le \sum_t \frac{2}{N^2} \quad \text{(by Lemma 26 and the Hoeffding inequality)} \\
&\le \frac{2}{N}.
\end{aligned}
$$

In the following, we focus on the case

$$
N_{i,t} \ge e^{-a^2/2}t - \sqrt{(\log N)t} \tag{55}
$$

that occurs with a probability at least $1 - 2/N$.

We have,

$$
\begin{aligned}
\sum_{n=1}^{N} \mathbf{1}[\tilde{\imath} = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} &\le \sum_{n=1}^{N} \mathbf{1}[\tilde{\imath} = i] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i)}} \quad \text{(by } \bar{\boldsymbol{V}}_{g(i)} \succeq \bar{\boldsymbol{V}}_i\text{)} \\
&\le \sum_{n=1}^{N} \sum_{t=1}^{N} \mathbf{1}[\tilde{\imath} = i, N_i^{\mathcal{W}_i}(n) = t] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i)}} \\
&\le \sum_{t=1}^{N} \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))}}. \\
&\quad \text{(by } \{\tilde{\imath} = i\} \text{ increments } N_i^{\mathcal{W}_i}\text{)}
\end{aligned}
$$

The following lower-bounds $\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))$.

We have

$$
\begin{aligned}
\mathbb{E}\left[\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))\right] &\ge \lambda_{\min}\left(\sum_n \mathbb{E}[\mathbf{1}\left[\mathcal{X}_i'(n)\right] \boldsymbol{x}_i \boldsymbol{x}_i']\right) \\
&\ge \lambda_0 N_{i,t}. \quad \text{(by Lemma 16)}
\end{aligned}
$$

By using the matrix Azuma inequality (Lemma 19), at least $1-1/N$

$$\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t)) \geq \left(\lambda_0 N_{i,t} - \sqrt{32 N_{i,t}\sigma_A^2}\log(dN)\right) \tag{56}$$

where $\sigma_A = 2(L(1/N))^2$. By using Eq. (55), (56), we have

$$\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t)) \geq \lambda_0 e^{-a^2/2} t - O(\sqrt{t})$$

and thus

$$\begin{aligned}\sum_{t=1}^{N} \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_i(\tau_t))}} &\leq \sum_{t=1}^{N} \frac{1}{\sqrt{\lambda_0 e^{-a^2/2} t - O(\sqrt{t})}} \\ &\leq \frac{2e^{a^2/4}}{\lambda_0}\sqrt{N} + O(1).\end{aligned}$$

We here bound the amount of the subsidy. Eq. (44), (45) imply

$$\begin{aligned}\boldsymbol{x}_i'\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{g(i)}\right) &\leq \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} L\beta_N \\ \left|\boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_{g(i)} - \boldsymbol{\theta}_{g(i)}\right| &\leq 2\frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} L\beta_N\end{aligned}$$

and thus the subsidy $s_i^{\text{H-I}}(n) = 0$ for

$$\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)}) \geq \left(\frac{2L\beta_N}{\|\boldsymbol{\theta}\|}\right)^2 \max\left(1, \frac{1}{a^2\sigma_x^2}\right) =: C_s = \tilde{O}(1). \tag{57}$$

Hence, it follows that

$$\begin{aligned}\text{Sub}(N) &= \sum_n s_\iota^{\text{H-I}}(n) \\ &\leq \sum_n \sum_i \mathbf{1}[\mathcal{X}_i] s_\iota^{\text{H-I}}(n) \\ &\leq L\beta_N \sum_i \sum_n \mathbf{1}[\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)}) \leq C_s] \frac{1}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_{g(i)})}} \quad \text{(by Eq. (57))} \\ &\leq L\beta_N \sum_i \sum_t \mathbf{1}[\lambda_0 e^{-a^2/2} t - O(\sqrt{t}) \leq C_s] \frac{1}{\sqrt{\lambda_0 e^{-a^2/2} t - O(\sqrt{t})}} \\ &\quad \text{(by the same discussion as Lemma 27)}\end{aligned}$$

$$\leq L\beta_N K \sum_t \mathbf{1}[\lambda_0 e^{-a^2/2} t \leq C_s] \frac{1}{\sqrt{\lambda_0 e^{-a^2/2} t}} + \tilde{O}(1)$$

$$\leq L\beta_N K \frac{2e^{a^4/2}}{\lambda_0} \sqrt{\frac{C_s e^{a^2/2}}{\lambda_0}} + \tilde{O}(1) = \tilde{O}(1). \tag{58}$$

□

Note that C_s diverges as $a \to +0$. The bound of the subsidy is meaningful for $a > 0$. If $a = 0$, the hybrid mechanism is reduced to the UCB mechanism, and thus Theorem 4 for UCB applies.

6 Proof of Theorem 6

We modify the proof of Theorem 2. Accordingly, we use the same notation as the proof of Theorem 2 unless we explicitly mention.

We define

$$\mathcal{Q}''(n) = \left\{ \exists i^A, i^B \text{ s.t. } g(i^A) = g(i^B) = 1, i^A \neq i^B, \text{ and } x_i \hat{\theta}_{1,N_1(n)} \geq \frac{1}{2}\mu_x \theta \text{ for } i = i^A, i^B \right\}.$$

When the event $\mathcal{Q}''(n)$ occur, there are two majority workers whose estimated skill $\hat{q}_i(n)$ is larger than its mean.

Lemma 29.

$$\Pr[\mathcal{Q}''(n)|\mathcal{Q}] \geq 1 - (K_1 + 1)\left(\frac{1}{2}\right)^{K_1}. \tag{59}$$

Event $\mathcal{Q}''(n)$ states that the second order statistics of $\{\hat{q}_i\}_{i:g(i)=1}$ is below mean. Lemma 29 states that this event is exponentially unlikely to K_1. By the symmetry of normal distribution and independence of characteristics $\boldsymbol{x}_i$, each candidate is likely to be below mean with probability $1/2$, and the proof of Lemma 29 directly follows by counting the combinations such that at most one of the worker(s) are above mean.

When we have $\mathcal{P}'(n)$ and $\mathcal{Q}''(n)$ for all n, then for every round n, the top-2 workers in terms of quality $\hat{q}_i(n)$ are from the majority. In this case, the minority worker is not hired regardless of additional signal η_i. Accordingly, this is a sufficient condition for a perpetual underestimation.

The following is the proof of Theorem 6.

Proof. By using Lemmas 22, 23, 24, and 29, we have (35), (36), (37), and (59). From these equations, the probability of perpetual underestimation is bounded as:

$$\begin{aligned}
&\Pr\left[\bigcup_n \{\iota(n) = 1\}\right] \\
&\geq \Pr\left[\bigcup_n \{\mathcal{P}'(n), \mathcal{Q}''(n)\}, \mathcal{P}, \mathcal{Q}\right] \\
&\geq \Pr\left[\mathcal{P}\right] \Pr\left[\mathcal{Q}\right] \Pr\left[\bigcup_n \{\mathcal{P}'(n), \mathcal{Q}''(n)\} \mid \mathcal{P}, \mathcal{Q}\right] \quad \text{(by the independence of } \mathcal{P} \text{ and } \mathcal{Q}) \\
&\geq C_1 b \times C_2 \times (1 - N C_3 \exp(-C_4 b)) \times \left(1 - N(K_1 + 1)\left(\frac{1}{2}\right)^{K_1}\right) \quad \text{(by the union bound)},
\end{aligned}$$

which, by letting $b = O(1/\log(N))$ and $K_1 + \log_2(K_1 + 1) \geq \log_2 N$, is $\tilde{O}(1)$. □

7 Proof of Theorem 7

Proof. We have

$$\begin{aligned}
|\boldsymbol{x}_i'(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g)| &\leq ||\boldsymbol{x}_i||_{\bar{\boldsymbol{V}}_g^{-1}} \left\|\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\right\|_{\bar{\boldsymbol{V}}_g} \\
&\leq \frac{L}{\lambda_{\min}(\bar{\boldsymbol{V}}_g)} \beta_n \quad \text{(by Eq. (12) and (13))} \\
&\leq \frac{L}{\lambda} \beta_N \quad (\text{by } \bar{\boldsymbol{V}}_g \succeq \lambda \boldsymbol{I}_d) \\
&=: C_5 = \tilde{O}(1).
\end{aligned} \tag{60}$$

Let i_1 and i_2 be the finalists chosen from group 1 and 2, respectively. Define the following event:

$$\mathcal{J}(n) = \{\eta_{i_1}(n) - \eta_{i_2}(n) > 2C_5\}.$$

Under $\mathcal{J}$, the finalist of group 1 is chosen because Eq. (60) implies that $|\boldsymbol{x}_{i_1}'\hat{\boldsymbol{\theta}}_{g_1} - \boldsymbol{x}_{i_2}'\hat{\boldsymbol{\theta}}_{g_2}| \leq 2C_5$ and thus $\boldsymbol{x}_{i_1}'\hat{\boldsymbol{\theta}}_{g_1} + \eta_{i_1} - \boldsymbol{x}_{i_2}'\hat{\boldsymbol{\theta}}_{g_2} + \eta_{i_2} > 0$. Note that $\eta_{i_1} - \eta_{i_2}$ is drawn from $\mathcal{N}(0, 2\sigma_\eta^2)$. Let $C_6 = \Phi^c(\sqrt{2}C_5/\sigma_\eta)$. Then,

$$\Pr[\mathcal{J}(n)] = C_6.$$

Let $N_1^{\mathcal{J}} = \sum_{n'=1}^{n-1} \mathbf{1}[g(\iota) = 1, \mathcal{J}] \leq N_1(n)$ be the number of hiring of group 1 under event

$\mathcal{J}$. By using the Hoeffding inequality, with probability $1 - 1/N^2$ we have

$$N_1^{\mathcal{J}} \geq nC_6 - \sqrt{n \log(N)}. \tag{61}$$

By taking union bound, Eq. (61) holds for all n with probability $1 - \sum_n 1/N^2 \geq 1 - 1/N$. From now, we evaluate $\lambda_{\min}\left(\bar{\boldsymbol{V}}_1(n)\right)$. It is easy to see that

$$\begin{aligned}
\bar{\boldsymbol{V}}_1 &:= \sum_{n'=1:\iota(n')=g}^{n} \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}' + \lambda I \\
&\geq \sum_{n'=1:\iota(n')=g}^{n} \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}' \\
&\geq \sum_{n'=1:\mathcal{J}}^{n} \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}'.
\end{aligned}$$

In the following, we lower-bound the quantity

$$\lambda_{\min}(\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i'|\mathcal{J}]) \geq \min_{\boldsymbol{v}:||\boldsymbol{v}||=1} \lambda_{\min}(\mathrm{Var}[\boldsymbol{v}'\boldsymbol{x}_i|\mathcal{J}]).$$

Note that $i_1 = \arg\max_{i:g(i)=1} \boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_1$ is biased towards the direction of $\hat{\boldsymbol{\theta}}_1$, and we cannot use the diversity condition (Lemma 16). Let $\boldsymbol{v}_\parallel$ and $\boldsymbol{v}_\perp$ be the component of $\boldsymbol{v}$ that is parallel to and perpendicular to $\hat{\boldsymbol{\theta}}_1$.[24] It is easy to confirm that $\mathrm{Var}[\boldsymbol{v}_\perp'\boldsymbol{x}_i] = ||\boldsymbol{v}_\perp||^2\sigma_x^2$ because selection of $\arg\max_i \boldsymbol{x}_i'\hat{\boldsymbol{\theta}}_g$ does not yield any bias in perpendicular direction. Regarding $\boldsymbol{v}_\parallel$, Lemma 20 characterize the variance, which is slightly[25] smaller than the original variance due to biased selection. That is,

$$\min_{\boldsymbol{v}:||\boldsymbol{v}||=1} \lambda_{\min}(\mathrm{Var}[\boldsymbol{v}'\boldsymbol{x}_i|\mathcal{J}]) \geq \sigma_x\left(\frac{C_{\mathrm{varmax}}}{\log(K)}||\boldsymbol{v}_\parallel||^2 + ||\boldsymbol{v}_\perp||^2\right) \geq \sigma_x\frac{C_{\mathrm{varmax}}}{\log(K)}.$$

By using the matrix Azuma inequality (Lemma 19) with $\sigma_A = 2L^2$, for $t = \sqrt{32N_1^{\mathcal{J}}\sigma_A^2\log(dN)}$, with probability $1 - 1/N$

$$\lambda_{\min}(\bar{\boldsymbol{V}}_1) \geq \sigma_x\frac{C_{\mathrm{varmax}}}{\log(K)}N_g^{\mathcal{J}} - t. \tag{62}$$

Combining Eq. (61) and (62), with a probability at least $1 - 2/N$, we have

$$\lambda_{\min}(\bar{\boldsymbol{V}}_1(n)) \geq \sigma_x\frac{C_p}{\log(K)}n - \tilde{O}(\sqrt{n}) \tag{63}$$

[24] $||\boldsymbol{v}_\parallel||^2 + ||\boldsymbol{v}_\perp||^2 = 1$.
[25] $O(1/\log K)$

where $C_p = C_6 C_{\mathrm{varmax}} = \tilde{O}(1)$. By symmetry, exactly the same bound as Eq. (63) holds for group 2. Finally, by using similar transformations as Eq. (23), the regret is bounded as

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(N)] &\le 2\sum_{n=1}^{N} \max_{i\in[K]} \left| \boldsymbol{x}_i'(n)(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g) \right| \\
&\le 2\sum_{n=1}^{N} \frac{L}{\sqrt{\lambda_{\min}(\bar{\boldsymbol{V}}_g)}} \beta_N \quad \text{(by Eq. (12), (13))} \\
&\le 2L\beta_N \sum_{n=1}^{N} \sqrt{\frac{\log(K)}{\sigma_x C_p n - \tilde{O}(\sqrt{n})}} \quad \text{(by Eq. (63))} \\
&\le 4L\beta_N \sqrt{\frac{N\log(K)}{\sigma_x C_p}} + \tilde{O}(1) = \tilde{O}(\sqrt{N})
\end{aligned} \tag{64}$$

which concludes the proof. □