Centralized Selection with Preferences in the Presence of Biases

L. Elisa Celis [1] **Amit Kumar** [2] **Nisheeth K. Vishnoi** [1] **Andrew Xu** [1]

Abstract

This paper considers the scenario in which there are multiple institutions, each with a limited capacity for candidates, and candidates, each with preferences over the institutions. A central entity evaluates the utility of each candidate to the institutions, and the goal is to select candidates for each institution in a way that maximizes utility while also considering the candidates' preferences. The paper focuses on the setting in which candidates are divided into multiple groups and the observed utilities of candidates in some groups are biased–systematically lower than their true utilities. The first result is that, in these biased settings, prior algorithms can lead to selections with sub-optimal true utility and significant discrepancies in the fraction of candidates from each group that get their preferred choices. Subsequently, an algorithm is presented along with proof that it produces selections that achieve near-optimal group fairness with respect to preferences while also nearly maximizing the true utility under distributional assumptions. Further, extensive empirical validation of these results in real-world and synthetic settings, in which the distributional assumptions may not hold, are presented.

1. Introduction

This paper studies a centralized selection problem that occurs in high-stakes contexts such as education and employment. In this setting there are p institutions, each with capacities $k_1, k_2, \ldots, k_p$, and n candidates, and $\sum_i k_i$ is much less than n. A central entity evaluates the utility of each candidate to the institutions (same for each institution) and the candidates declare preferences for the institutions. The high-level goal of this entity is to then select a subset of at most k_ℓ candidates for ℓ-th institution while trying to maximize total utility and ensuring the preferences of the candidates are taken into account. Examples include the Joint Seat Allocation Authority (JOSAA) in India, which places over 50,000 students in top engineering colleges each year (Baswana et al., 2019; JOSAA, 2022), and comparable centralized allocation mechanisms across China (Wikipedia contributors, 2024b). Other examples include public schools in NYC and Boston, which have used centralized admissions systems (Roth & Sotomayor, 1992; Roth & Peranson, 1999), and online labor markets (such as TaskRabbit) that use centralized systems to match workers to tasks based on their skills and preferences for working hours, location, and type-of-task (Hannák et al., 2017; TaskRabbit, 2014).

Several works in Game Theory (Roth, 1989; Aziz et al., 2016), Machine Learning (Das & Kamenica, 2005; Muthukrishnan, 2009; Liu et al., 2020; 2021; Min et al., 2022), and Economics (Roth, 1985; Roth & Peranson, 1999; Liu et al., 2014; Rastegari et al., 2013) examine this centralized selection problem and its various forms. The variant of interest here is where candidates' utilities $u_1, u_2, \ldots, u_n \geq 0$ are cardinal and their preferences are ordinal. The utilities can denote candidates' test scores, performance in interviews, or suitability for hiring. Candidate i's preferences are given by a ranking σ_i of the p institutions, ordered from most to least preferred. The Gale-Shapley algorithm (Gale & Shapley, 1962) for this problem orders candidates in decreasing order of utility and assigns them to their most preferred institution which still has empty slots. This algorithm can be shown to maximize utility and ensure a type of stability–there are no two candidates who both prefer each other's assigned institution to the ones they have been assigned.

However, the true or latent utilities of candidates are often estimated through evaluation processes like tests, interviews, or user ratings, which are known to yield biased estimates against underprivileged groups (Rooth, 2010; Kite & Whitley, 2016; Régner et al., 2019; Elsesser, 2019; Moss-Racusin et al., 2012; Wennerås & Wold, 1997; Lyness & Heilman, 2006; Celis et al., 2023). Standardized exams like the SAT, GRE, and ACT, which are often pivotal in college admissions, disproportionately affect candidates from minority and low-income demographics (Elsesser, 2019). Further examples of biased evaluations include biased peer-review of fellowship applications against women (Wennerås & Wold, 1997), and stricter promotion standards for women in managerial positions (Lyness & Heilman, 2006).

[1]Yale University [2]IIT Delhi. Correspondence to: Nisheeth K. Vishnoi <nisheeth.vishnoi@yale.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Selections based on such biases in estimated utilities may adversely affect the opportunities of candidates from disadvantaged groups in multiple ways. For one, if an algorithm makes selections in an attempt to maximize estimated utilities, the representation of candidates from disadvantaged groups could be significantly low. Moreover, candidates from disadvantaged groups might be assigned to lower preferences compared to other groups. This latter inequality of opportunity may, in turn, influence candidates' future educational, employment, and economic opportunities (Lecher & Varner, 2021; Laverde, 2022; Whitehurst, 2017; Hannák et al., 2017; Bogen & Rieke, 2018), and is the focus of this paper. For instance, a study by (Laverde, 2022) showed that within the Boston Public Schools' assignment system, white pre-kindergarten students are more frequently placed in higher-performing schools compared to minority students, contradicting the system's intended purpose of ensuring fair access to quality education; see also (Lecher & Varner, 2021; Whitehurst, 2017). And, besides disadvantaging candidates from certain groups, biases in utilities also reduce the utility for institutions, who may be assigned candidates with a low true utility but a high estimated utility.

To reduce the adverse effects of such biases, constraints that increase the representation of underprivileged candidates, anti-bias training that aims to reduce bias in evaluations, and anonymized evaluations that aim to blind the evaluator to the socially salient attributes of candidates have been proposed (Goldin & Rouse, 2000; Reskin & McBrier, 2000; Sowell, 2008; Gawande, 2010; Bohnet, 2016; Zestcott et al., 2016; Agarwal, 2020). A line of work in the algorithmic fairness literature (Kleinberg & Raghavan, 2018; Celis et al., 2020; Emelianov et al., 2020; Celis et al., 2021) studies the effectiveness of *group-wise* constraints that require at least a certain number of candidates from disadvantaged groups to be selected, in the special case of the problem with only *one* institution (and, hence, candidates' preferences are vacuous). For instance, (Kleinberg & Raghavan, 2018) considers a model of bias where candidates are divided into two groups: the advantaged group G_1 and the disadvantaged group G_2. For an *unknown* $0 < \beta \leq 1$, the estimated utility of each disadvantaged candidate i is $\widehat{u}_i = \beta \cdot u_i$, and that of each advantaged candidate j is $\widehat{u}_j = u_j$. (Kleinberg & Raghavan, 2018; Celis et al., 2020) show that when the utilities are i.i.d., then requiring a proportional number of candidates from both advantaged and disadvantaged groups to be selected improves the latent utility of the assignment. However, these results do not generalize to the case when $p > 1$. Simple examples show that just ensuring group representation in assignments doesn't guarantee that candidates from each group will be selected for their top choice institution. Specifically, disadvantaged groups may see lower selection rates for their preferred institutions due to biases. We discuss additional related works in detail in Appendix A.

Our contributions. We study the question: To what extent can we achieve simultaneous goals of preference fairness, representational fairness, and utility maximization in the multi-institution selection problem with preferences in the presence of biases? We assume that the true utilities and preferences are i.i.d. from distributions $\mathcal{D}$ and $\mathcal{L}$ respectively; see Section 2. We consider preference-based fairness metric $\mathscr{P}$ which is the expectation of the ratio of the fraction of candidates from the disadvantaged group (G_2) that get assigned to their most preferred institution to the analogous fraction for G_1 (Equation (3)). To measure representational fairness, we consider the metric $\mathscr{R}$ which is the expectation of the ratio of the fraction of candidates in G_2 that are selected to that of G_1 (Equation (2)). To measure utility, we consider a utility ratio $\mathscr{U}$, which is defined as the expected value of the ratio of the true utility achieved in the presence of bias to that achieved without any bias (Equation (1)).

Our first result studies the performance of the Gale-Shapley algorithm and provides upper bounds for $\mathscr{U}$ and $\mathscr{P}$ as a function of β; see Theorem 3.1 and Appendix C.2 for extensions. When the distribution $\mathcal{D}$ is uniform on $[0, 1]$ and β approaches 0, both preference-based and representational fairness decrease to 0, while the utility ratio approaches $\frac{2}{3}$.

We consider a type of representational constraint: institution-wise constraints that require a proportional number of candidates from each group to be matched to *each* institution. The algorithm $\mathcal{A}_{\text{inst-wise}}$ (Algorithm 3) then outputs an assignment that is group-wise stable subject to satisfying these institution-wise constraints. Our second result is that $\mathcal{A}_{\text{inst-wise}}$ has $\mathscr{P} \approx 1$, $\mathscr{R} \approx 1$, $\mathscr{U} \approx 1$ (Theorem 3.2). This result holds for any distributions $\mathcal{D}$ and $\mathcal{L}$. The proofs of both our results rely on a novel Lipschitz property of the Gale-Shapley algorithm and non-trivial concentration and scaling results for the total number of candidates who are assigned their top choice by it; see Section C.1 for details.

We provide empirical validation of our theoretical results on synthetic and real-world datasets in Section 4 and Appendices G to I to demonstrate their robustness under deviations from assumptions. In Section 4, we study the performance of $\mathcal{A}_{\text{inst-wise}}$ on real-world data from the 2009 IIT-JEE. In Section 4, we evaluate $\mathcal{A}_{\text{inst-wise}}$ on a synthetic dataset when the preferences of the two groups are drawn from different distributions. In Appendix H, we evaluate the performance of $\mathcal{A}_{\text{inst-wise}}$ under different models of bias. In all simulations, $\mathcal{A}_{\text{inst-wise}}$ maintains a high preference-based fairness and utility ratio compared to the baselines. For instance, for the IIT-JEE dataset, we observe that $\mathcal{A}_{\text{inst-wise}}$ achieves significantly higher preference-based fairness ($\geq$0.9) than group-wise constraints ($\leq$0.50) (Figure 1). In Appendix G, we also study the performance of institution-wise constraints under relaxed bounds. Finally, in Appendix I, we show that Theorem 3.1 is robust for different utility and bias models.

2. Model

Here we present the centralized selection problem, the model of bias, the model for utilities and preferences, and the fairness/utility metrics. Each component follows prior works in fairness and/or centralized assignment literature.

Centralized selection with preferences. Given n candidates and p institutions with capacities $\boldsymbol{k} = (k_1, k_2, \ldots, k_p)$ the goal is to select $K := \sum_{\ell=1}^{p} k_\ell$ candidates and assign them to the institutions subject to the capacity constraint. We consider the centralized setting where each candidate i has a true of *latent* utility $u_i \geq 0$ that captures the value generated if candidate i is selected–it does not depend on the institution. Further, each candidate also has a preference list σ_i over the set $[p] := \{1, 2, \ldots, p\}$ of institutions. A selection M is a partial function from the set of candidates $[n]$ to the set of institutions $[p]$. For any M and index $\ell \in [p]$, let $M^{-1}(\ell)$ denote the set of candidates assigned to the ℓ-th institution. We extend this notation to a subset of $S \subseteq [p]$ of institutions – $M^{-1}(S)$ denotes the set of candidates assigned to an institution in S. One of the goals of the selection problem is to find an M that maximizes the latent utility $\sum_{i \in M^{-1}([p])} u_i$ of the candidates assigned to an institution. another desirable property is stability: consider two candidates i and j with i having higher utility, then $M(i)$ is preferred over $M(j)$ in σ_i. The Gale-Shapley algorithm is widely used for finding such an M (Gale & Shapley, 1962; Roth & Peranson, 1999) as they have many desirable properties, including latent-utility maximization and stability (Roth, 1985; Roth & Sotomayor, 1992).

Remark 2.1. *This model of centralized selection with preferences mirrors several real-world admission systems, where each student is represented by a single utility value and schools rank these values to manage the complexities of large-scale admission processes. For example, the Joint Entrance Examination (JEE Main), overseen by India's National Testing Agency, is among the most prestigious annual college entrance exams globally. In 2023, it drew over 1.17 million candidates. Candidates are evaluated in Mathematics, Physics, and Chemistry, with their scores combined into a single ranking for all participating institutions. For additional information, refer to the JEE Main FAQ (National Testing Agency, 2024). The* Civil Services Examination in India *is also an annual test for entering around 15 civil services. Attracting over 500,000 applicants annually, the exam has ten sections. Scores from these sections are merged into one overall score. For more details, see (Wikipedia contributors, 2024a). China's Gaokao, an annual standardized test for undergraduate admissions, attracted over 12.9 million candidates in 2023. Students are assessed in six subjects, with their total score calculated as a weighted sum of these scores. Applicants select universities based on this score and their regional preferences. More details are available in (Chen & Kesten, 2017).*

Model of bias in utilities. We study the model introduced by (Kleinberg & Raghavan, 2018). In this model, the candidates are divided into two groups: the advantaged group G_1 and the disadvantaged group G_2. The estimated utilities in this model are parameterized by an *unknown* bias parameter $0 < \beta \leq 1$: given β, the estimated utility $\widehat{u}_i$ for a disadvantaged candidate is β-times their latent utility, i.e., $\widehat{u}_i = \beta \cdot u_i$, and that of an advantaged candidate j is the same as their latent utility, i.e., $\widehat{u}_i = u_i$.

The motivation to consider $\beta < 1$ and, hence, $\hat{u}_i \leq u_i$, comes from contexts (discussed in the introduction) where the estimated utility $\hat{u}_i$ of a candidate in the disadvantaged group systematically underestimates their true utility u_i. That said, even in this bias model, one can allow for $\hat{u}_i \geq u_i$ by setting $\beta \geq 1$. Our results continue to hold when $\beta \geq 1$. Moreover, in our empirical results, we consider the setting where β may have some noise; see H.1. The motivation to assume β is unknown comes from the fact that the analysis is focused on the "one-round" setting of the selection problems. Finally, the bias model assumes that the group identities are known and are reported truthfully. This model is easily generalized to multiple disjoint groups by introducing a bias parameter for each group (Celis et al., 2020).

Generative model of utilities and preferences. Following (Kleinberg & Raghavan, 2018; Celis et al., 2020; Emelianov et al., 2020), we let the latent utility u_i of each candidate i be drawn some distribution $\mathcal{D}$ independent of all other candidates. Further, we assume that the preference list σ_i of each candidate i is drawn independently from some distribution $\mathcal{L}$ of the set of all preference lists of p institutions.

I.I.D. utilities encode the fact that there are no systematic differences in utilities across groups. Different choices of $\mathcal{D}$ arise in different contexts: measures of popularity and success are observed to have power-law distributions (Clauset et al., 2009), the percentile of candidates in a population has a uniform distribution on $[0, 100]$, and normal distributions model utilities in standardized tests (Dorans, 2002).

I.I.D. preference lists encode the assumption that candidates in either group have the same distribution of preferences over institutions. This is motivated by contexts where preferences are largely determined by the "quality" of the institution: For instance, most candidates applying to the JOSAA in India have similar preferences over universities (Livemint, 2022; Verma, 2022), in Boston and New York Public Schools, most parents prefer high-performing schools over others (Laverde, 2022), and workers naturally prefer tasks with higher pay per time over tasks with a lower pay per time (Hannák et al., 2017). Note that the i.i.d. assumption allows candidates to have different preferences. For example, if $\mathcal{L}$ assigns probability 0.5 to each of two preference lists σ_1 and σ_2 over $[p]$, then roughly half the candidates have preference σ_1 (and similarly for σ_2).

Utility ratio. For a fixed utility vector $\boldsymbol{v}$ and a preference vector $\boldsymbol{\sigma}$, given a selection $M_{\boldsymbol{v},\boldsymbol{\sigma}} : [n] \to [p]$, we define the utility of $M_{\boldsymbol{v},\boldsymbol{\sigma}}$ as the sum of utilities of all candidates that are assigned to any institution, i.e., $\mathsf{U}(M_{\boldsymbol{v},\boldsymbol{\sigma}}) := \sum_{i \in M^{-1}([p])} v_i$. Since $\hat{\boldsymbol{u}}$ is a scaled-down version of $\boldsymbol{u}$, note that the maximum value $\mathsf{U}(\cdot)$ can take is the sum of top K entries in the true utility vector $\boldsymbol{u}$; we denote this by $\mathsf{U}^\star(\boldsymbol{u})$. We measure the performance of an assignment with respect to $\mathsf{U}^\star$. Consider an algorithm $\mathcal{A}$ that, given preferences $\boldsymbol{\sigma}$ and utilities $\widehat{\boldsymbol{u}}$, outputs a selection $M_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}$. The utility-ratio of $\mathcal{A}$ is defined as:

$$\mathscr{U}_{\mathcal{D},\mathcal{L}}(\mathcal{A}) := \mathbb{E}_{\boldsymbol{u}\sim\mathcal{D},\boldsymbol{\sigma}\sim\mathcal{L}} \frac{\mathsf{U}(M_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}})}{\mathsf{U}^\star(\boldsymbol{u})}. \tag{1}$$

It is important to note that the algorithm's assignment relies on the *estimated* utilities $\widehat{\boldsymbol{u}}$ of the candidates. However, when measuring the utility ratio, we use the *true* utilities of the selected candidates.

Fairness metrics. There is a wide body of work on metrics for fairness (Barocas et al., 2019). The two fairness metrics considered here are grounded in the idea of "proportional representation", a common and useful notion of fairness in many contexts. *Representational fairness* disregards candidate preferences and only considers how many candidates from each group are assigned to at least one institution. For group $j \in \{1, 2\}$, let ρ_j denote the fraction of candidates in G_j that are selected by $M_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}$. Representational fairness measures the disparity in the values of ρ_1 and ρ_2. We consider a multiplicative notion of comparing the performance across groups. For an algorithm $\mathcal{A}$, its representational fairness is defined as:

$$\mathscr{R}_{\mathcal{D},\mathcal{L}}(\mathcal{A}) := \mathbb{E}_{\boldsymbol{u}\sim\mathcal{D},\boldsymbol{\sigma}\sim\mathcal{L}} \frac{\min_{j\in\{1,2\}} \rho_j}{\max_{j'\in\{1,2\}} \rho_{j'}}. \tag{2}$$

This definition is similar in spirit to the fairness metrics used by past works (Celis et al., 2019): they require the value of some desirable quality to be similar for different groups. By definition, $\mathscr{R}_{\mathcal{D},\mathcal{L}}(\mathcal{A})$ is a value between 0 and 1. $\mathscr{R}_{\mathcal{D},\mathcal{L}}(\mathcal{A})$ is close to 1 if the $\mathcal{A}$ ensures that a proportional number of candidates from each group are assigned to at least one institution. The larger the value of $\mathscr{R}_{\mathcal{D},\mathcal{L}}(\mathcal{A})$ is the more "fair" $\mathcal{A}$ is.

Preference-based fairness captures the disparity in the fraction of candidates in each group that are assigned to their "top" preferences. The definition of top preference is context-dependent: if there are a small number of institutions, as in school admissions, the first preference of students is important (Laverde, 2022) (see also (Singh & Joachims, 2018; Celis et al., 2018; Manning et al., 2010)), and if there are a large number of institutions, as in the online labor market, then the relevant notion may be the first ℓ preferences for some $1 \le \ell \le p$. Let π_j denote the fraction of candidates in group G_j that get their first preference. For an algorithm $\mathcal{A}$, its preference-based fairness is defined as:

$$\mathscr{P}_{\mathcal{D},\mathcal{L}}(\mathcal{A}) := \mathbb{E}_{\boldsymbol{u}\sim\mathcal{D},\boldsymbol{\sigma}\sim\mathcal{L}} \frac{\min_{j\in\{1,2\}} \pi_j}{\max_{j'\in\{1,2\}} \pi_{j'}}. \tag{3}$$

$\mathscr{P}_{\mathcal{D},\mathcal{L}}(\mathcal{A})$ is a value between 0 and 1 and large values are preferable. These three metrics extend to the setting of multiple disjoint groups by considering the minimum and maximum over all groups. Similarly, for $\ell \in [p]$, one can consider the performance $\pi_j^{(\ell)}$, which considers the fraction of candidates in group G_j that get an assignment from among their top-ℓ choices; and similarly define $\mathscr{P}^{(\ell)}_{\mathcal{D},\mathcal{L}}(\mathcal{A})$. When $\mathcal{D}$ and $\mathcal{L}$ are clear, we drop the subscripts from $\mathscr{U}_{\mathcal{D},\mathcal{L}}, \mathscr{R}_{\mathcal{D},\mathcal{L}}$, and $\mathscr{P}_{\mathcal{D},\mathcal{L}}$ and use $\mathscr{U}$, $\mathscr{R}$, and $\mathscr{P}$, $\mathscr{P}^{(\ell)}$ etc.

3. Theoretical Results

In this section, we first present theoretical results on the effect of the bias parameter on the fairness and utility metrics of a stable assignment algorithm. Subsequently, we develop an algorithm for enhancing these metrics by adding institutional representation constraints. Finally, we present an overview of the ideas involved in proving these results.

Performance of a stable assignment in the presence of bias. We consider how stable allocation algorithms that take into account the preferences of participating members perform in the presence of bias. In our special setting, where the preferences of institutions are derived from a centralized evaluation of candidates, there is a unique stable assignment of candidates that can be obtained by the following algorithm: consider the candidates in decreasing order of observed utilities. When considering a candidate, assign it to its most preferred institution that still has an available spot (see Definition C.3 for the definition of a stable assignment, and Proposition C.4 for the uniqueness of such an assignment). This algorithm, denoted $\mathcal{A}_{\text{st}}$, is formally described in Algorithm 1. When there is no bias, i.e. $\beta = 1$, it can be shown that $\mathcal{A}_{\text{st}}$ has near-optimal representational and preference-bases fairness, and near-optimal utility (see Corollary C.27). Our first result shows that the performance of $\mathcal{A}_{\text{st}}$ deteriorates as β decreases:

Theorem 3.1. *Consider an instance where the utilities of the candidates are drawn from the uniform distribution on* $[0, 1]$ *and the distribution over preferences is arbitrary. Assume* $n_1 = n_2 = K$. *Then,* $\mathscr{P}(\mathcal{A}_{\text{st}}) \le \beta + O\left(\frac{p\sqrt{\log n}}{\sqrt{n}}\right)$, $\mathscr{R}(\mathcal{A}_{\text{st}}) = \beta \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right)$, *and* $\mathscr{U}(\mathcal{A}_{\text{st}}) = \frac{2}{3} + \frac{4\beta}{3(\beta+1)^2} \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right)$.

The above result shows that when the bias parameter β is close to zero, then the fairness metrics $\mathscr{R}(\mathcal{A}_{\text{st}})$ and $\mathscr{U}(\mathcal{A}_{\text{st}})$ deteriorate substantially, and there is a significant drop in the utility of the selected candidates as well. It is also worth noting that the error terms in the theorem decay as $p/\sqrt{n}$ and are negligible for reasonable values of n and p. The

proof of this result appears in Section C.2 and develops novel results on Lipschitz and concentration properties of a stable assignment in our setting. It is also worth noting that the proof first shows a high probability bound on the utility and fairness metrics and then uses this to bound the expectation of the respective metrics. We present an outline of the proof at the end of this section.

The theorem extends to arbitrary values of n_1, n_2 and K (see Theorem C.12 for details). For example, when $n_1 = n_2$ and $K = \alpha n_1$ for some parameter $\alpha \geq 1 - \beta$, then both $\mathscr{R}(\mathcal{A}_{\text{st}})$ and $\mathscr{P}(\mathcal{A}_{\text{st}})$ are at most $\frac{\alpha-(1-\beta)}{\alpha\beta+(1-\beta)}$. This result generalizes significantly to arbitrary log-concave distributions of utilities that include (truncated) Gaussian and Pareto (see Section C.2.4 and also empirical results in Section I.1). It is also worth noting that such a result does not hold for arbitrary distributions of utilities (see Example C.30). We also empirically validate the robustness of Theorem 3.1 on other models of bias (see Sections I.2, I.3).

Enhancing preference-based fairness via institution-wise representational constraints. Our second result shows an algorithm that can enhance preference-based fairness while ensuring near-optimal utility.

Theorem 3.2. *Let $\eta_1, \eta_2, \eta_3 > 0$ be parameters such that $|G_j| \geq \eta_1 n$ for each $j \in \{1, 2\}$, $K \geq \eta_2 n$, and $k_\ell \geq \eta_3 K$ for each $\ell \in [p]$. There is an allocation algorithm $\mathcal{A}_{\text{inst-wise}}$ (stated formally in Algorithm 3) such that, for any distribution of utilities and preference lists, and bias parameter β, $\mathscr{P}(\mathcal{A}_{\text{inst-wise}}) \geq 1 - O\left(\frac{p\sqrt{\log K}}{\eta_1\eta_3\sqrt{K}}\right)$, $\mathscr{U}(\mathcal{A}_{\text{inst-wise}}) \geq 1 - O\left(\frac{\sqrt{\log n}}{\sqrt{\eta_2 n}}\right)$ and $\mathscr{R}(\mathcal{A}_{\text{inst-wise}}) = 1$.*

The error bounds in the theorem are negligible in typical scenarios. Consider for example a setting where $|G_1| = |G_2| = n/2, K = cn$ for some constant c, and all p institutions have the same capacity. Then $\eta_1 = 1/2, \eta_2 = c$ and $\eta_3 = 1/p$. Thus the error bounds for $\mathscr{P}(\mathcal{A}_{\text{inst-wise}})$ and $\mathscr{U}(\mathcal{A}_{\text{inst-wise}})$ decay as $p^2/\sqrt{K}$ and $1/\sqrt{n}$ respectively. The proof of this result appears in Section C.3. As in the case of Theorem 3.1, we show a stronger statement that the fairness and the utility metrics of $\mathcal{A}_{\text{inst-wise}}$ are close to 1 with high probability. The algorithm $\mathcal{A}_{\text{inst-wise}}$ circumvents the strong lower bounds mentioned in Theorem 3.1 arising due to the bias parameter β. It is worth emphasizing that the results in Theorem 3.2 hold for arbitrary distributions $\mathcal{D}$ and $\mathcal{L}$ over utilities and preferences, and the stated bounds in the theorem do not depend on the bias parameter. The theorem does require that the distribution of the utilities and the preference lists of candidates are identical and independent. Removing this i.i.d. assumption seems challenging and we leave it as an interesting avenue for future work. We now motivate the ideas in the algorithm $\mathcal{A}_{\text{inst-wise}}$ by arguing that several natural algorithms do not satisfy the conclusions of Theorem 3.2:

(i) Algorithm $\mathcal{A}_{\text{inst-wise}}$: As Theorem 3.1 shows, running the algorithm $\mathcal{A}_{\text{st}}$ directly on the original instance could lead to a significant loss in utility and preference-based fairness. A natural fix would be to first learn (and remove the effect of) the bias parameter β and then apply $\mathcal{A}_{\text{st}}$ on the "corrected" input. We give a detailed discussion in Section B on why such an approach is highly impractical.

(ii) Algorithm $\mathcal{A}_{\text{inst-wise}}$ with proportional group-wise representational constraints: This approach is rooted in in prior works (Kleinberg & Raghavan, 2018; Celis et al., 2020) on the single institution setting. For instance, we could run $\mathcal{A}_{\text{st}}$ on the set of candidates selected from the top $|G_j|/n$ fraction from each group G_j (see Algorithm 2 for a formal description). While this algorithm can be shown to enhance representational fairness while obtaining near-optimal utility, it may have low preference-based fairness (see Section B).

(iii) Algorithms that directly enforce preference-based fairness: An approach for generalizing the known results for the single institution setting to multiple institutions while directly enforcing high preference-based fairness would be the following (assume for the sake of simplicity that $|G_1| = |G_2|$) – each institution ℓ considers the candidates that prefer ℓ as their first choice. We assign the highest utility candidates from this subset to institution ℓ while ensuring that no more than half the capacity at this institution is allocated to a single group (in case some slots are vacant at the end, we can fill them with the highest utility unassigned candidates). Such an algorithm $\mathcal{A}$ would indeed achieve high preference-based fairness, but may not obtain near-optimal utility even when there is no bias – see Section B for an example.

We consider a different family of representational constraints: institution-wise constraints. These constraints require an adequate number of candidates from each group to be assigned to *each* institution. Formally, we require that for each institution $\ell \in [p]$, $|G_j|/n \cdot k_\ell$ candidates from each group G_j are assigned to it, where k_ℓ is the capacity of the ℓ-th institution. The algorithm $\mathcal{A}_{\text{inst-wise}}$, which is a strict generalization of the approaches for single institution setting in (Kleinberg & Raghavan, 2018; Celis et al., 2020), satisfies the proportional representational constraints as follows: We invoke two independent instantiations of the algorithm $\mathcal{A}_{\text{st}}$ described at the beginning of Section 3. The first instance is run with the capacity of each institution ℓ set to $k_\ell \cdot n_1/n$ and restricting the candidates to G_1 only. Similarly, the second instance is run with the capacity of each institution ℓ set to $k_\ell \cdot n_2/n$ and restricting the candidates to those from G_2 only. A formal description of $\mathcal{A}_{\text{inst-wise}}$ appears in Algorithm 3. It is worth noting that this algorithm does not explicitly enforce preference-based constraints and does not require the knowledge of β. It is easy to show that $\mathcal{A}_{\text{inst-wise}}$ is group-wise stable, Pareto-efficient,

and strategy-proof (see Section B). The algorithm $\mathcal{A}_{\text{inst-wise}}$ and the results in Theorem 3.2 can also be extended to settings with multiple groups and preference-based fairness metric involving a fraction of candidates that are assigned an institution among their top-ℓ choices for a parameter ℓ.

We empirically validate the conclusions of Theorem 3.2 on the following robust settings: (i) real-world data from India's centralist IIT admission system, where the i.i.d. assumption on the utilities and the preference lists of candidates may not hold and compare $\mathcal{A}_{\text{inst-wise}}$ with algorithms based on group-wise constraints only (see Section 4), (ii) relaxed versions of the institutional constraints where less than $|G_j|/n_j \cdot k_\ell$ slots may be reserved in the ℓ-the institution for each group G_j (see Algorithm 4 for the formal description of such an algorithm and Section G for related empirical results), (iii) execute $\mathcal{A}_{\text{inst-wise}}$ on data generated using other models of bias (see Sections H.1, H.2), and (iv) allow two different distributions over the preferences for candidates from the two respective groups (see Section E.2). Note that our results cannot be extended directly to settings where institutions have varying utilities for candidates. Indeed, in such settings, the algorithm $\mathcal{A}_{\text{st}}$ (restricted to candidates from the same group) is not well defined. One could replace $\mathcal{A}_{\text{st}}$ with another stable matching algorithm, but such an algorithm may not have both near-optimal utility and preference-based fairness.

Outline of the proofs of Theorems 3.1 and 3.2. We now describe the main ideas in the proofs of Theorem 3.1 and Theorem 3.2. It is well known that the algorithm $\mathcal{A}_{\text{st}}$, also known as a *serial dictatorship mechanism*, is stable (see e.g. (Svensson, 1999)). However, proving either of these results requires us to delve much deeper into the properties of the algorithm $\mathcal{A}_{\text{st}}$. In particular, we study the Lipschitz properties of the assignment given by $\mathcal{A}_{\text{st}}$ and the concentration properties of this assignment when preference lists are drawn from a distribution. We begin by stating the following concentration bound.

Lemma 3.3 (Informal; see Lemma C.9). *Let G be a subset of candidates belonging to the same group. Let S be the subset of selected candidates and $\pi(S)$ denote the candidates in S that receive their top choice. Then,* $\Pr\left[||S \cap G| - \mathbb{E}[|S \cap G|]| \geq t\right] \leq 2e^{-2t^2/|G|}$, *and* $\Pr\left[||\pi(S) \cap G| - \mathbb{E}[|\pi(S) \cap G|]| \geq t\right] \leq 2e^{-2t^2/(p^2K)}$.

We now give an outline of the proof of this result. We can express $|S \cap G|$ as $\sum_{i \in G} X_i$, where X_i are negatively correlated random variables. Therefore, we can apply Chernoff-type concentration bounds here. However, the quantity $|\pi(S) \cap G|$ is much trickier to handle. Clearly, $|\pi(S) \cap G|$ can be again be written as $\sum_{i \in G} Y_i$, where Y_i is an indicator variable denoting whether the i-th candidate in G received the most preferred choice. Unlike the case of X_i's, the variables Y_i's may not be independent or negatively correlated. One possibility is to use concentration bounds relying on the Lipschitz property of the assignment (see e.g., Theorem C.2), i.e., changing the preference list of only one candidate changes the assignment slightly. However, one can construct examples for general stable matching instances (see Example C.5), where such a change can dramatically alter the stable assignment. Our problem has the additional feature that all the institutions agree on a central ranking of the candidates. For such settings, we show that there is a unique stable assignment (Proposition C.4), and is given by algorithm $\mathcal{A}_{\text{st}}$ (formally described in Algorithm 1). We can compare two runs of $\mathcal{A}_{\text{st}}$ on instances that differ on only one candidate. Let S and S' be the set of candidates selected by $\mathcal{A}_{\text{st}}$ in these two runs. We show that the size of the set difference between $\pi(S)$ and $\pi(S')$, i.e., $|(\pi(S) \setminus \pi(S')) \cup (\pi(S') \setminus \pi(S))|$ is at most $p + 1$ (Lemma C.6). This Lipschitz property of $\mathcal{A}_{\text{st}}$ allows us to show the second concentration bound mentioned in Lemma 3.3.

We now give an overview of the proof of Theorem 3.1 (and the more general result Theorem C.12). Assume for the sake of simplicity that $|G_1| = |G_2|$. We first consider $\mathscr{R}(\mathcal{A}_{\text{st}})$. Let S be the set of candidates selected by $\mathcal{A}_{\text{st}}$. The metric $\mathscr{R}(\mathcal{A})$ considers the expectation of the ratio $\frac{|S \cap G_2|}{|S \cap G_1|}$. Using Lemma 3.3, we can closely approximate this by $\frac{\mathbb{E}[|S \cap G_2|]}{\mathbb{E}[|S \cap G_1|]}$. We now find the threshold $\Delta \in [0, 1]$ such that S is closely approximated by candidates with utility at least Δ. This allows us to estimate both of the desired expectations. The utility metric $\mathscr{U}(\mathcal{A}_{\text{st}})$ can be estimated by evaluating the expected utility of top $\mathbb{E}[|S \cap G_j|]$ candidates from each of the groups G_j. We now consider $\mathscr{P}(\mathcal{A}_{\text{st}})$ which is defined as $\mathbb{E}\left[\frac{\pi(S \cap G_2)}{\pi(S \cap G_1)}\right]$ (recall that $\pi(X)$ denotes the number of candidates in a set X that are assigned their top choice by $\mathcal{A}_{\text{st}}$). Again, using the concentration bound in Lemma 3.3, it suffices to bound $\frac{\mathbb{E}[\pi(S \cap G_2)]}{\mathbb{E}[\pi(S \cap G_1)]}$. However, we cannot simply write down a closed form expression for $\mathbb{E}[\pi(S \cap G_j)]$. Instead, we show (here Δ is as defined above):

Lemma 3.4 (Informal; see Proposition C.26). *Let A and B be the set of candidates with observed utility in the range $[\Delta, 1]$ and $[0, \Delta]$ respectively. Conditioning on $g_1 := |A|, g_1' := |B \cap G_1|, g_2' := |B \cap G_2|$, $\pi(G_1)$ and $\pi(G_2)$ are closely approximated by $\pi(A) + \frac{g_1'}{g_1' + g_2'} \cdot \pi(B)$ and $\frac{g_2'}{g_1' + g_2'} \cdot \pi(B)$ respectively.*

Using the above result along with the fact that candidates in A are more likely to get their first choice than in B, and hence $\pi(A)/g_1 \geq \pi(B)/(g_1' + g_2')$ (Proposition C.25), we can bound $\mathscr{U}(\mathcal{A}_{\text{st}})$. Thus, we show all the three desired bounds in Theorem 3.1.

We now give an overview of the proof of Theorem 3.2. The metric $\mathscr{R}(\mathcal{A}_{\text{inst-wise}}) = 1$ by the definition of the algorithm

$\mathcal{A}_{\text{inst-wise}}$. Indeed, we ensure that the number of selected candidates from each group G_j is proportional to $|G_j|$. The utility ratio $\mathscr{U}(\mathcal{A}_{\text{st}})$ is also close to 1 for a similar reason: the institution-wise representational constraints imply that the total utility of the top K candidates is close to the sum over each group G_j of the total utility of the top $\frac{|G_j|}{n} \cdot K$ candidates from G_j. We now consider $\mathscr{P}(\mathcal{A}_{\text{inst-wise}})$, which is defined as $\mathbb{E}\left[\frac{|\pi(S \cap G_2)|/|G_2|}{|\pi(S \cap G_1)|/|G_1|}\right]$. Lemma 3.3 allows us to approximate this by the ratio of the corresponding expectations $\mathbb{E}[|\pi(S \cap G_j)|/|G_j|]$. In light of the representational constraints, one would expect $\mathbb{E}[|\pi(S \cap G_j)|]/|G_j|]$ to be the same for both groups. Somewhat surprisingly, this fact turns out to be false (see Example C.10). In particular, consider the following two instances (consisting of just one group): in the instance $\mathcal{I}$ there are n candidates and capacity k_j at each institution $j \in [p]$; whereas in the "scaled" instance $\mathcal{I}_a$, where $a \geq 1$, there are n/a candidates with k_j/a capacity at each institution $j \in [p]$. Let M and M_a be the corresponding assignments given by $\mathcal{A}_{\text{st}}$, and $\pi(M)$ and $\pi(M_a)$ be the number of candidates who receive their top choices in these two assignments. Then $\mathbb{E}[\pi(M)]$ may not equal $a\,\mathbb{E}[\pi(M_a)]$. However, we can bound their gap:

Lemma 3.5. $\left|\frac{\mathbb{E}[\pi(M)]}{a} - \mathbb{E}[\pi(M_a)]\right| = O(p\sqrt{K \log K})$.

The proof of this result relies on yet another Lipschitz property of the algorithm $\mathcal{A}_{\text{st}}$: changing the capacity of an institution by one unit changes the allocation for at most p candidates (Lemma C.7). Lemma 3.5 allows us to show that $\mathbb{E}\left[\frac{|\pi(S \cap G_2)|/|G_2|}{|\pi(S \cap G_1)|/|G_1|}\right]$ is close 1. Thus, we can show that all three metrics are near optimal for $\mathcal{A}_{\text{inst-wise}}$.

4. Empirical Results

We empirically examine the robustness of Theorem 3.2 under non-ideal conditions. We evaluate Theorem 3.2 using data from India's centralized IIT admissions system, and evaluate a setting in which preferences are not drawn from a single distribution. The code and data can be found here[1].

Baselines and metrics. We consider three algorithms in each simulation: the unconstrained algorithm ($\mathcal{A}_{\text{st}}$), the algorithm subject to group-wise constraints ($\mathcal{A}_{\text{group}}$), and the algorithm that satisfies institution-wide constraints ($\mathcal{A}_{\text{inst-wise}}$). Pseudocodes are provided in Appendix D. We consider two groups, a general and a disadvantaged group, and measure preference-based fairness using $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(3)}$ as defined in Section 2. We generate preferences via the well-studied Mallows model, which mirrors many real-world scenarios (Mallows, 1957). Preferences are distributed according to a central ranking ρ and the Kendall-Tau distance (d_{KT}) between rankings, defined as the number of pairwise disagreements between two rankings. A dispersion parameter, $0 < \phi \leq 1$, defines the distribution such that for any ranking σ, $\Pr(\sigma) \propto \phi^{d_{\text{KT}}(\sigma,\rho)}$.

Simulation with real-world data. We use admissions data from India's centralized admissions system for the Indian Institutes of Technology (IITs) to evaluate institution-wise constraints in a real-world, non-ideal setting.

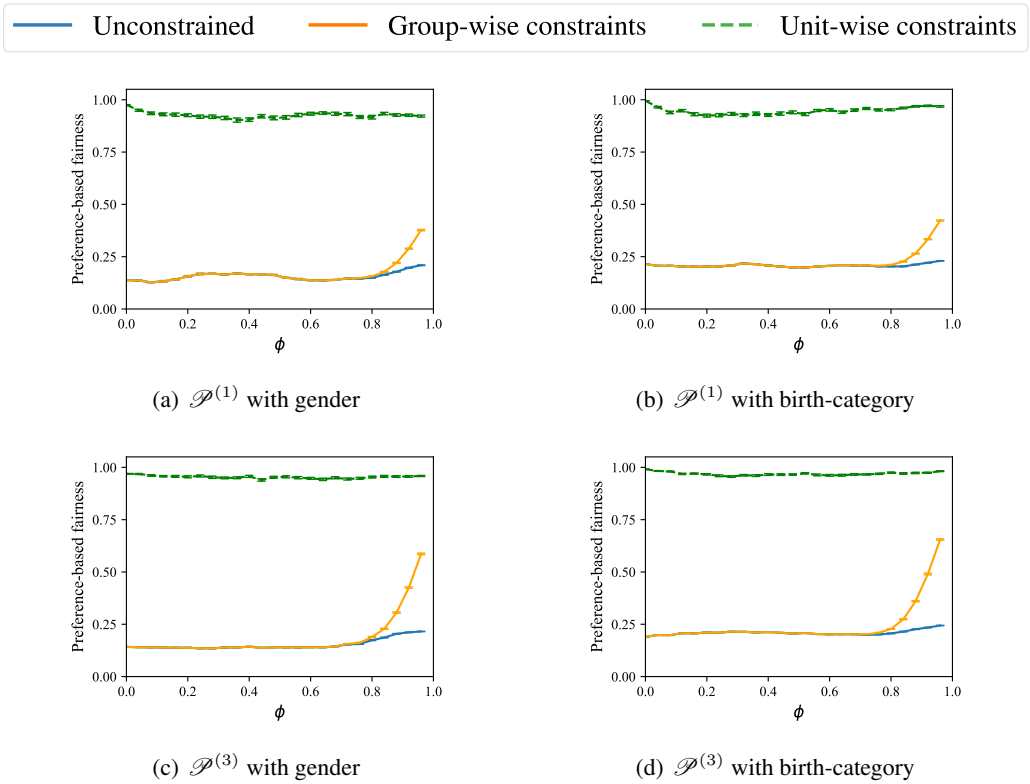
Data. Admission to IITs is determined by a centralized system that considers an applicant's score in the Joint Entrance Exam (JEE) as their utility and their preferences over different major-institute pairs (Baswana et al., 2019). The data contains the test scores of 384,977 students from IIT-JEE 2009, self-reported gender and official birth category, opening and closing ranks, and capacities of major-institute pairs (JOSAA, 2009). Individual preferences were not reported, so we model preferences over institutions, defined as major-institute pairs, using a Mallows distribution with ρ defined by increasing order of $C_{(M,I)}$, the closing All India Rank of candidates admitted into the major M at institute I in 2009. Further discussion about the dataset, the notion of latent utilities, and preference generation is in Appendix E.

Setup. We run two simulations, one considering gender and the other considering birth category as the protected attribute. We consider the $p = 33$ institutions corresponding to the top major-institute pairs in ρ. The last index such that the p-th pair (M, I) in ρ has $C_{(M,I)} \leq 1000$ is $p = 33$. We fix k_ℓ to be the 2009 capacity of the ℓ-th major-institute pair in ρ. For $\phi \in [0, 1]$, we report the average of $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(3)}$ over 50 iterations, where we regenerate student preferences according to the Mallows distribution in each iteration. Implementation details appear in Appendix E.

Results and observations. Results appear in Figure 1. Our main observation is that institution-wise constraints achieve near-optimal preference-based fairness across all preference distributions, whereas group-wise constraints do not. For all ϕ, $\mathscr{P}^{(1)}(\mathcal{A}_{\text{inst-wise}}) \geq 0.9$ and $\mathscr{P}^{(3)}(\mathcal{A}_{\text{inst-wise}}) \geq 0.95$. On the other hand, group-wise constraints and no constraints result in low preference-based fairness: $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}}) \leq 0.5$, $\mathscr{P}^{(3)}(\mathcal{A}_{\text{group}}) \leq 0.7$, and $\mathscr{P}^{(3)}(\mathcal{A}_{\text{st}}) \leq 0.25$. When $\phi \leq 0.75$, $\mathscr{P}^{(3)}(\mathcal{A}_{\text{group}}) \leq 0.25$. Under all conditions, institution-wide constraints achieve near-optimal preference-based fairness, higher than both other algorithms. Given $\phi \leq 0.75$, which likely models the admissions system for IITs, there is no distinguishable difference between $\mathcal{A}_{\text{st}}$ and $\mathcal{A}_{\text{group}}$ in terms of preference-based fairness.

Simulation with distinct preference distributions. We also evaluate the efficacy of institution-wise constraints when groups have different preference distributions. This tests Theorem 3.2 when the assumption that preferences are generated from a single distribution is not fulfilled. This situation can occur in many real-life scenarios: notably, His-

[1] https://github.com/sandrewxu/CentralizedSelectionwithPreferenceBias



(a) $\mathscr{P}^{(1)}$ with gender

(b) $\mathscr{P}^{(1)}$ with birth-category

(c) $\mathscr{P}^{(3)}$ with gender

(d) $\mathscr{P}^{(3)}$ with birth-category

Figure 1: Preference-based fairness as measured by $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(3)}$ using either gender or birth-category as the protected attribute with data from the 2009 JEE test under centralized admission (see Section 4 for details). The x-axis denotes ϕ, the dispersion parameter of the Mallows distribution. Error bars denote the standard error of the mean over 50 iterations. Institution-wise constraints achieve near-optimal preference-based fairness.

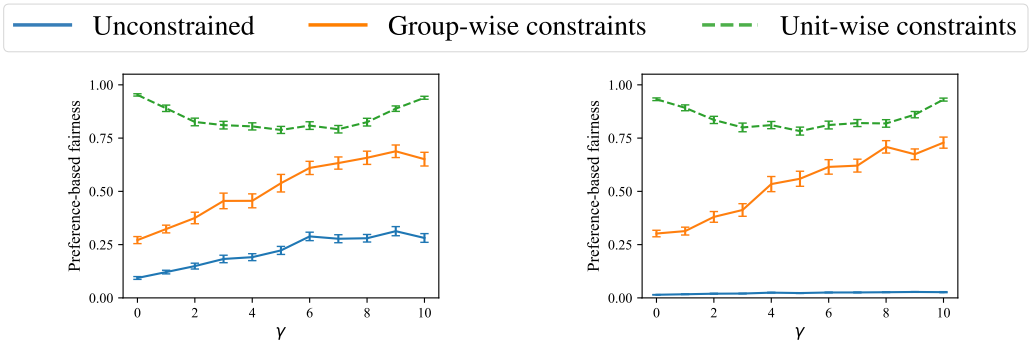
torically Black Colleges and Universities (HBCUs) in the United States are highly prestigious and a coveted choice for black students, but are rarely a top choice for students of other racial groups.

Data (Latent utilities and preferences). We consider two different distributions for latent utility: $\mathcal{D}_{\text{Gauss}}$, the truncated Gaussian distribution bounded below by 0, and $\mathcal{D}_{\text{Pareto}}$, the Pareto distribution with shape parameter $\alpha = 3$. The Gaussian distribution is used because it often models the outcomes of standardized tests (Dorans, 2002). The fat-tailed Pareto distribution with shape parameter $2 \leq \alpha \leq 3$ models many real-life phenomena such as citations of a paper or success in creative professions (Clauset et al., 2009). When generating preferences, we use the Mallows distribution with $\phi = 0.25$ centered around ρ_1, ρ_2 for groups G_1 and G_2. For distance $0 \leq \gamma \leq \gamma_{\max} := \binom{p}{2}$, ρ_1 and ρ_2 are randomly generated under the constraint $d_{\text{KT}}(\rho_A, \rho_B) = \gamma$.

Setup. We set $n = 1000, p = 5$, and $k_i = 100$ for $i \in [p]$. For $\mathcal{D} \in \{\mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$ and $\beta \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$, we vary $\gamma \in [0, \gamma_{max}]$ and calculate $\mathscr{P}^{(1)}$ and $\mathscr{U}$ over 50 iterations.

Results and observations. Figure 2 shows the results using the $\mathscr{P}^{(1)}$ metric with $\beta = \frac{1}{4}$. Results with other β and a discussion of $\mathscr{U}$ appear in Appendix F. We observe that implementing institution-wide constraints retains a high degree of preference-based fairness, while group-wise constraints do not. For all γ, $\mathscr{P}^{(1)}(\mathcal{A}_{\text{inst-wise}}) \geq 0.75$ while $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}}) \leq 0.7$. When $\gamma = 0$, central ranking is consistent across groups and $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}}) \leq 0.3$ and $\mathscr{P}^{(1)}(\mathcal{A}_{\text{inst-wise}}) \geq 0.9$. As γ increases, $\mathscr{P}(\mathcal{A}_{\text{inst-wise}})$ slightly decreases, then increases again. This is because an increase in γ decreases the ability of $\mathcal{A}_{\text{inst-wise}}$ to guarantee high $\mathscr{P}$, but as γ approaches γ_{max}, rankings tend to become more opposite and more people from each group match with their top choice institutions, resulting in high $\mathscr{P}$. The preference-based fairness obtained through group-wise constraints and no constraints generally increases with γ. Even when groups follow different preference distributions, institutional constraints outperform group-wise constraints, especially when $\gamma \leq 3$.

To better model the complexities of socio-economic biases, we also explore other models of bias. In Appendix H.1, we utilize the noisy β-bias model, where β is sampled from a truncated Gaussian distribution with a standard deviation of 0.1 within the range $[0, 1]$. Our findings reveal that for $0.15 \leq \beta \leq 0.85$, the fairness achieved in this model is within 0.05 of that predicted by the β- bias model. We also consider the implicit variance model in Appendix H.2,



(a) $\mathcal{D}_{\text{Gauss}}$ with $\beta=\frac{1}{4}$

(b) $\mathcal{D}_{\text{Pareto}}$ with $\beta=\frac{1}{4}$

Figure 2: Preference-based fairness as measured by $\mathscr{P}^{(1)}$, with synthetic data where non-i.i.d. preferences are generated from Mallows distributions. The x-axis denotes γ, the Kendall-Tau distance between the central rankings, and the error bars denote the standard error of the mean over 50 iterations. We observe that institution-wise constraints achieve higher preference-based fairness than group-wise and unconstrained settings.

which introduces Gaussian noise directly to the true utility of each candidate. The mean of the noise is centered at 0, and the variance is higher for the disadvantaged group. We vary the standard deviation of the Gaussian noise within the interval $[0, 2]$ and observe that our algorithm maintains a preference-based fairness above 0.9, while the introduced noise reduces the fairness of baseline algorithms.

Recognizing that a rigorous implementation of Theorem 3.2 may seem challenging in some contexts, we also consider an algorithm that relaxes the strict capacity reservation constraint of institutions in Appendix G. We introduce the parameter α, which ranges from 0 to 1, to allow a proportionate fraction of an institution's capacity to be reserved for various groups. By adjusting α to values less than 1, we offer institutions the flexibility to proportionally reserve a smaller portion of their capacity, making the algorithm adaptable to diverse policy requirements and institutional constraints. We find that this can provide an effective method for improving $\mathscr{P}$, with a natural tradeoff between preference-based fairness and the strength of the constraints on the institutions.

Finally, we empirically evaluate Theorem 3.1 under different utility distributions, as well under the noisy β-bias model and the implicit variance model in Appendix I to see that under these settings, the predictions of Theorem 3.1 still generally hold.

5. Conclusion, Limitations, and Future Work

This paper formalizes and studies the impact of biases against socio-economic groups for a centralized selection problem. Our first result shows that biases in utilities can result in assignments that not only lower utility and decrease representation but also fail to conform to preference-based notions of fairness (Theorem C.12). We present a family of institution-wise representational constraints and show that the algorithm $\mathcal{A}_{\text{inst-wise}}$ achieves both preference-based fairness and near-optimal utility (Theorem 3.2). We extensively evaluate the performance of $\mathcal{A}_{\text{inst-wise}}$ on real-world and synthetic datasets, along with several deviations from the assumptions of Theorem 3.2. We observe that it achieves high utility and near-optimal preference-based fairness compared to baselines. Theorem 3.2 relies on the assumption that the utilities of individuals are independently and identically distributed. This i.i.d. assumption is not specific to our work and also applies to the results of the prior works (Kleinberg & Raghavan, 2018; Celis et al., 2020) for the special case of subset selection (where there is only one institution and hence, the preferences of the candidates are vacuous). Extending these prior works in the non-i.i.d. setting is already an interesting direction. The proof of Theorem 3.2 can be generalized to more than two groups by reserving capacity at each institution for each of the groups in proportion to their sizes and running $\mathcal{A}_{\text{st}}$ on several independent instances, one for each group.

Future research could explore settings where different institutions value candidates differently, such as the hospital-resident matching problem, particularly in the presence of biases. If institutions' utilities are uncorrelated, adapting our results might need completely new approaches. This could involve modeling how institutions' rankings of candidates correlate. A starting point might be the model proposed by (Castera et al., 2022), which assumes that different institutions' utilities for a candidate are correlated.

Finally, we emphasize that discrimination is a systemic problem and our work only studies a part of it. To holistically counter bias and avoid unintended negative impacts, it is also important to study the long-term consequences of interventions, e.g., investments in skill development (Heidari & Kleinberg, 2021) and the impact on biases in evaluations (Celis et al., 2021), and evaluate the efficacy of interventions in specific contexts both pre- and post-deployment.

Acknowledgments

This work was funded in part by grants from Tata Sons Private Limited, Tata Consultancy Services Limited, and Titan, and NSF Awards CCF-2112665 and IIS-2045951. We acknowledge Anay Mehrotra for his help in executing some components of the empirical work and proofs at an earlier stage of the project. We acknowledge Vineet Goyal and Bhaskar Ray Chaudhari for initial discussions regarding theoretical results.

Impact statement

The goal of this paper is to present an algorithm that can enhance fairness in centralized selection problems despite the systemic biases present due to socioeconomic contexts and/or machine learning algorithms. This algorithm has the potential to positively impact areas such as hiring and admissions. It can be used as one of a multipronged approach to counter systemic social biases. As with any intervention intended to improve equity, we note that the same algorithm could be used instead to discriminate by selecting parameters that oppose rather than improve fairness. The implementation of the algorithm should be determined collaboratively with stakeholders to ensure it works towards equity both explicitly and implicitly. We encourage the continued responsible and ethical application of these advancements, including an evaluation of the long-term effects of such interventions.

References

Agarwal, P. *Sway: Unravelling Unconscious Bias*. Bloomsbury Publishing, 2020. ISBN 9781472971371.

Ahsanullah, M., Nevzorov, V. B., and Shakil, M. *An Introduction to Order Statistics*, volume 8. Springer, 2013.

Alderman, H. and King, E. M. Gender differences in parental investment in education. *Structural Change and Economic Dynamics*, 9(4):453–468, 1998.

Aziz, H., Biró, P., Gaspers, S., de Haan, R., Mattei, N., and Rastegari, B. Stable matching with uncertain linear preferences. In Gairing, M. and Savani, R. (eds.), *Algorithmic Game Theory*, pp. 195–206, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53354-3.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019.

Baswana, S., Chakrabarti, P. P., Kamakoti, V., Kanoria, Y., Kumar, A., Patange, U., and Chandran, S. Joint seat allocation: An algorithmic perspective, 2015.

Baswana, S., Chakrabarti, P. P., Chandran, S., Kanoria, Y., and Patange, U. Centralized Admissions for Engineering Colleges in India. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pp. 323–324, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367929.

Biswas, A., Patro, G. K., Ganguly, N., Gummadi, K. P., and Chakraborty, A. Toward fair recommendation in two-sided platforms. *ACM Trans. Web*, 16(2), dec 2021. ISSN 1559-1131.

Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? In *FORC*, volume 156 of *LIPIcs*, pp. 3:1–3:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

Boehmer, N., Celis, L. E., Huang, L., Mehrotra, A., and Vishnoi, N. K. Subset selection based on multiple rankings in the presence of bias: Effectiveness of fairness constraints for multiwinner voting score functions. In *ICML*, 2023.

Bogen, M. and Rieke, A. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias, December 2018. https://www.upturn.org/reports/2018/hiring-algorithms/.

Bohnet, I. *What Works: Gender Equality by Design*. Harvard University Press, 2016. ISBN 9780674545991.

Castera, R., Loiseau, P., and Pradelski, B. S. Statistical discrimination in stable matchings. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pp. 373–374, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504.

Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with Fairness Constraints. In *ICALP*, volume 107 of *LIPIcs*, pp. 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2018.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *FAT*, pp. 319–328. ACM, 2019.

Celis, L. E., Mehrotra, A., and Vishnoi, N. K. Interventions for ranking in the presence of implicit bias. In *FAT**, pp. 369–380. ACM, 2020.

Celis, L. E., Hays, C., Mehrotra, A., and Vishnoi, N. K. The Effect of the Rooney Rule on Implicit Bias in the Long Term. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 678–689, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Celis, L. E., Kumar, A., Mehrotra, A., and Vishnoi, N. K. Bias in evaluation processes: An optimization-based

model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

CensusInfo India 2011: Final Population Totals. CensusInfo India 2011: Final Population Totals, 2011. http://censusindia.gov.in/2011census/censusinfodashboard/index.html.

Chen, Y. and Kesten, O. Chinese college admissions and school choice reforms: A theoretical analysis. *Journal of Political Economy*, 125(1):99–139, 2017.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Matroids, matchings, and fairness. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2212–2220. PMLR, 2019.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4): 661–703, 2009.

Das, S. and Kamenica, E. Two-sided bandits and the dating market. In *IJCAI*, volume 5, pp. 19. Citeseer, 2005.

Dorans, N. J. The recentering of SAT® scales and its effects on score distributions and score interpretations. *ETS Research Report Series*, 2002(1):i–21, 2002.

Dubhashi, D. P. and Panconesi, A. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009. ISBN 978-0-521-88427-3.

Elsesser, K. Lawsuit Claims SAT And ACT Are Biased—Here's What Research Says, 2019. URL https://www.forbes.com/sites/kimelsesser/2019/12/11/lawsuit-claims-sat-and-act-are-biased-heres-what-research-says/?sh=177187663c42.

Emelianov, V., Gast, N., Gummadi, K. P., and Loiseau, P. On fair selection in the presence of implicit variance. In *EC*, pp. 649–675. ACM, 2020.

Freeman, R., Micha, E., and Shah, N. Two-sided matching meets fair division. *arXiv preprint arXiv:2107.07404*, 2021.

Gale, D. and Shapley, L. S. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Garg, N., Li, H., and Monachou, F. Standardized tests and affirmative action: The role of bias and variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 261, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Gawande, A. *The Checklist Manifesto: How to Get Things Right*. Henry Holt and Company, 2010. ISBN 9781429953382.

Goldin, C. and Rouse, C. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, 90(4):715–741, September 2000.

Goto, M., Kojima, F., Kurata, R., Tamura, A., and Yokoo, M. Designing matching mechanisms under general distributional constraints. *American Economic Journal: Microeconomics*, 9(2):226–62, May 2017.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. 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*, CSCW '17, pp. 1914–1933, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350.

Heidari, H. and Kleinberg, J. Allocating Opportunities in a Dynamic Model of Intergenerational Mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 15–25, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Irving, R. W. An efficient algorithm for the "stable roommates"' problem. *Journal of Algorithms*, 6(4):577–595, 1985. ISSN 0196-6774.

JOSAA. IIT-JEE Brochure. https://mba.ind.in/forum/attachments/f2/2103d1348557572-iit-jee-brochure-downloads-iit-jee-brochure.pdf, 2009.

JOSAA. Seat Information for IITs, NITs, IIEST, IIITs and Other-GFTIs for the Academic Year 2022-23, 2022. https://josaa.admissions.nic.in/applicant/seatmatrix/seatmatrixinfo.aspx.

Kamada, Y. and Kojima, F. Efficient matching under distributional constraints: Theory and applications. *American Economic Review*, 105(1):67–99, January 2015.

Kamada, Y. and Kojima, F. Stability concepts in matching under distributional constraints. *Journal of Economic Theory*, 168:107–142, 2017. ISSN 0022-0531.

Karni, G., Rothblum, G. N., and Yona, G. On Fairness and Stability in Two-Sided Matchings. In Braverman,

M. (ed.), *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 92:1–92:17, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.92.

Kite, M. and Whitley, B. *Psychology of Prejudice and Discrimination: 3rd Edition*. Taylor & Francis, 2016. ISBN 9781317227229.

Kleinberg, J. M. and Raghavan, M. Selection problems in the presence of implicit bias. In *ITCS*, volume 94 of *LIPIcs*, pp. 33:1–33:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

Kumar, M. R. RTI Complaint, 2009. Decision No. CIC/SG/C/2009/001088/5392, Complaint No. CIC/SG/C/2009/001088.

Kumar, R. SC, ST, OBC representation in Indian education is dismal, upper-caste nexus persists. *The Print*, March 2021. https://theprint.in/campus-voice/sc-st-obc-representation-in-indian-education-is-dismal-upper-caste-nexus-persists/627217/.

Laverde, M. Distance to schools and equal access in school choice systems. Technical report, Boston College, 2022. https://www.bc.edu/content/dam/bc1/schools/mcas/economics/pdf/working-papers/1000/wp1046.pdf.

Lecher, C. and Varner, M. NYC's School Algorithms Cement Segregation. This Data Shows How. *The Markup*, May 2021. https://themarkup.org/machine-learning/2021/05/26/nycs-school-algorithms-cement-segregation-this-data-shows-how.

Liu, L. T., Mania, H., and Jordan, M. Competing bandits in matching markets. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1618–1628. PMLR, 26–28 Aug 2020.

Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Liu, Q., Mailath, G. J., Postlewaite, A., and Samuelson, L. Stable matching with incomplete information. *Econometrica*, 82(2):541–587, 2014.

Livemint. IIT seat allotment: IIT Bombay still the top choice for JEE rank holders, September 2022. https://www.livemint.com/news/india/iit-seat-allotment-iit-bombay-still-the-top-choice-for-jee-rank-holders-11663932899833.html.

Lyness, K. S. and Heilman, M. E. When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology*, 91(4):777, 2006.

Mallows, C. L. Non-null ranking models. i. *Biometrika*, 44 (1/2):114–130, 1957.

Manlove, D. *Algorithmics Of Matching Under Preferences*. Series On Theoretical Computer Science. World Scientific Publishing Company, 2013. ISBN 9789814425261.

Manning, C., Raghavan, P., and Schütze, H. Introduction to Information Retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

Mehrotra, A. and Vishnoi, N. K. Maximizing submodular functions for recommendation in the presence of biases. In Ding, Y., Tang, J., Sequeda, J. F., Aroyo, L., Castillo, C., and Houben, G. (eds.), *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pp. 3625–3636. ACM, 2023a.

Mehrotra, A. and Vishnoi, N. K. Maximizing submodular functions for recommendation in the presence of biases. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 3625–3636, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9781450394161.

Mehrotra, A., Pradelski, B. S. R., and Vishnoi, N. K. Selection in the presence of implicit bias: The advantage of intersectional constraints. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 599–609, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522.

Min, Y., Wang, T., Xu, R., Wang, Z., Jordan, M., and Yang, Z. Learn to match with no regret: Reinforcement learning in markov matching markets. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19956–19970. Curran Associates, Inc., 2022.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., and Handelsman, J. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012.

Muthukrishnan, S. Ad exchanges: Research issues. In *International Workshop on Internet and Network Economics*, pp. 1–12. Springer, 2009.

National Testing Agency. Frequently Asked Questions - Joint Entrance Examination (Main), 2024. URL https://jeemain.nta.ac.in/images/Jeemain-FAQ.pdf.

Passariello, C. Tech Firms Borrow Football Play to Increase Hiring of Women, September 2016. URL https://www.wsj.com/articles/tech-firms-borrow-football-play-to-increase-hiring-of-women-1474963562.

Rastegari, B., Condon, A., Immorlica, N., and Leyton-Brown, K. Two-sided matching with partial information. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pp. 733–750, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450319621.

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., and Huguet, P. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature human behaviour*, 3(11):1171–1179, 2019.

Reskin, B. F. and McBrier, D. B. Why Not Ascription? Organizations' Employment of Male and Female Managers. *American Sociological Review*, 65(2):210–233, 2000. ISSN 00031224.

Rooth, D.-O. Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3): 523–534, 2010.

Roth, A. E. Common and conflicting interests in two-sided matching markets. *European Economic Review*, 27(1): 75–96, 1985. ISSN 0014-2921. Special Issue on Market Competition, Conflict and Collusion.

Roth, A. E. Two-sided matching with incomplete information about others' preferences. *Games and Economic Behavior*, 1(2):191–209, 1989.

Roth, A. E. and Peranson, E. The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *American Economic Review*, 89(4):748–780, September 1999.

Roth, A. E. and Sotomayor, M. Chapter 16: Two-sided matching. In *Handbook of Game Theory with Economic Applications*, volume 1, pp. 485–541. Elsevier, 1992.

Rothblum, U. G. Characterization of stable matchings as extreme points of a polytope. *Mathematical Programming*, 54:57–67, 1992.

Salem, J. and Gupta, S. Closing the gap: Mitigating bias in online résumé-filtering. In *WINE*, pp. 471, 2020.

Singh, A. and Joachims, T. Fairness of Exposure in Rankings. In *KDD*, pp. 2219–2228. ACM, 2018.

Sowell, T. *Affirmative Action Around the World: An Empirical Study*. Yale University Press, 2008. ISBN 9780300128352.

Svensson, L.-G. Strategy-proof allocation of indivisible goods. *Social Choice and Welfare*, 16(4):557–567, 1999.

TaskRabbit. Unveiling the New TaskRabbit, 2014. https://www.taskrabbit.com/blog/unveiling-the-new-taskrabbit.

Vande Vate, J. H. Linear programming brings marital bliss. *Operations Research Letters*, 8(3):147–153, 1989. ISSN 0167-6377.

Verma, T. JoSAA Counselling 2022: IIT Bombay, IIT Delhi and IIT Kanpur are most-preferred choices of top 1000 rankers in the past. *The Indian Express*, September 2022. https://indianexpress.com/article/education/jee-advanced-results-2022-date-iit-bombay-most-preferred-institute-among-top-1000-rankers-followed-by-iit-delhi-iit-kanpur-8140297/.

Wennerås, C. and Wold, A. Nepotism and Sexism in Peer-Review. *Nature*, 387(6631):341–343, May 1997. ISSN 1476-4687.

Whitehurst, G. J. R. New evidence on school choice and racially segregated schools. *Brookings*, December 2017. https://www.brookings.edu/research/new-evidence-on-school-choice-and-racially-segregated-schools/.

Wikipedia contributors. Civil services examination — wikipedia, the free encyclopedia, 2024a. URL https://en.wikipedia.org/w/index.php?title=Civil_Services_Examination&oldid=1133321967.

Wikipedia contributors. Gaokao — wikipedia, the free encyclopedia, 2024b. URL https://en.wikipedia.org/w/index.php?title=Gaokao&oldid=1125543653.

Williams, R. Why girls in India are still missing out on the education they need, March 2013. https://www.theguardian.com/education/2013/mar/11/indian-children-education-opportunities.

Zestcott, C. A., Blair, I. V., and Stone, J. Examining the presence, consequences, and reduction of implicit bias in health care: a narrative review. *Group Processes & Intergroup Relations*, 19(4):528–542, 2016.

Contents

1 Introduction **1**

2 Model **3**

3 Theoretical Results **4**

4 Empirical Results **7**

5 Conclusion, Limitations, and Future Work **9**

A Additional Related Work **16**

B Additional Discussion on Theorem 3.2 **17**

C Proofs **19**

C.1 Properties of a stable assignment of candidates to institutions . . . 19

C.2 Proof of Theorem 3.1 and extensions . . . 24

C.2.1 Step 1: Bounding $\mathscr{R}$ . . . 24

C.2.2 Step 2: Bounding $\mathscr{U}$ . . . 27

C.2.3 Step 3: Bounding $\mathscr{P}$ . . . 28

C.2.4 Extension of Theorem 3.1 to log-concave densities . . . 31

C.3 Proof of Theorem 3.2 . . . 33

C.3.1 Step 1: Bounding $\mathscr{R}$ . . . 34

C.3.2 Step 2: Bounding $\mathscr{U}$ . . . 34

C.3.3 Step 3: Bounding $\mathscr{P}$ . . . 34

D Pseudocode for Algorithms with Institution-Wise, Group-Wise and No Constraints **35**

E Additional Discussion and Figures for the Real-World Simulation With JEE 2009 Scores **37**

E.1 Additional figures and implementation details . . . 37

E.2 Simulation that varies the dispersion parameter of the preference distribution using synthetic data . . . 38

F Additional Discussion and Plots for Simulation with Synthetic Data from Section 4 **40**

F.1 Discussion of utility ratio . . . 40

F.2 Additional plots . . . 40

G Empirically Studying Institution-Wise Constraints Under Relaxed Bounds **41**

H Empirical Results Considering Noise in Estimated Utilities **42**

H.1 Simulation considering noise in the bias parameter . . . 42

H.2 Simulation with noise in estimated utilities under the implicit variance model 43

I Empirical Validation of Theorem 3.1 **44**

I.1 Pareto and truncated Gaussian distributions of utilities 45

I.2 Noise in the bias parameter 46

I.3 Implicit variance model 46

A. Additional Related Work

The game-theoretic analysis of assignment and selection problems began with the foundational work by Gale & Shapley (1962). Since then, a substantial body of literature, including (Roth, 1985; 1989; Roth & Peranson, 1999; Das & Kamenica, 2005; Rastegari et al., 2013; Liu et al., 2014; Aziz et al., 2016; Liu et al., 2020; 2021; Min et al., 2022), has expanded on this topic, as detailed in overviews like (Roth & Sotomayor, 1992; Manlove, 2013). This research examines various aspects of assignment problems. It explores game-theoretic properties such as stability and truthfulness (Roth, 1985; Roth & Peranson, 1999), analyzes these properties under conditions of incomplete or uncertain information (Roth, 1989; Rastegari et al., 2013; Liu et al., 2014; Aziz et al., 2016), and uses machine learning tools to impute preferences and utilities (Das & Kamenica, 2005; Liu et al., 2020; 2021; Min et al., 2022).

Several works empirically study biases in algorithmic assignment/selection systems used in the real world, from centralized admissions systems to online labor markets (Whitehurst, 2017; Hannák et al., 2017; Lecher & Varner, 2021; Laverde, 2022). (Castera et al., 2022) complement these empirical studies with theoretical analysis: They consider a model of bias utilities of candidates in the disadvantaged group has a higher noise than the utilities of candidates in the advantaged group and show that, somewhat surprisingly, increasing the noise for the disadvantaged candidates worsens the outcomes for both groups by reducing the probability that a candidate is assigned their first preference. In comparison, we consider a different model of bias (Section 2) and bound the effect of bias on representational and preference-based fairness $\mathscr{P}$. Our result implies a bound on the ratio of the probability that a disadvantaged candidate gets their first preference compared to an advantaged candidate. However, unlike (Castera et al., 2022), we also demonstrate that institution-wise representational constraints can mitigate the adverse effect of bias (Theorem 3.2 and Section 4). Motivated by these, several works design algorithms for finding assignments subject to different types of fairness constraints (Chierichetti et al., 2019; Biswas et al., 2021; Freeman et al., 2021; Karni et al., 2022). Notably, the algorithm of (Baswana et al., 2019) has been implemented in an Indian context. Unlike these works, we focus on mitigating the adverse effects of bias in the centralized selection problem (Theorem 3.2).

The effect of representational constraints. Several works study fundamental algorithmic tasks when the inputs may be biased. Examples include, subset selection (Kleinberg & Raghavan, 2018; Emelianov et al., 2020; Salem & Gupta, 2020; Celis et al., 2021; Garg et al., 2021; Mehrotra et al., 2022; Mehrotra & Vishnoi, 2023b; Boehmer et al., 2023; Mehrotra & Vishnoi, 2023a), ranking (Celis et al., 2020), and classification (Blum & Stangl, 2020). Subset selection and ranking are particularly relevant to our study.

(Kleinberg & Raghavan, 2018) proposes a mathematical framework for understanding bias and demonstrates that implementing the Rooney Rule—a type of group-wise constraint (Passariello, 2016)—enhances the latent utility of selected subsets. (Emelianov et al., 2020) explore a bias model where estimated utilities are noisier for disadvantaged candidates, showing positive impacts of group-wise constraints on latent utility. Additionally, (Boehmer et al., 2023) analyzes the benefits of group-wise constraints in multi-winner elections—a form of subset selection with submodular objectives based on multiple preference lists.

These studies explore various dimensions of group-wise constraints' effectiveness. For instance, (Salem & Gupta, 2020) focus on online selection scenarios, (Celis et al., 2021) investigate the long-term effects of biases on evaluators, (Mehrotra et al., 2022) broaden (Kleinberg & Raghavan, 2018)'s model to intersecting protected groups, and (Mehrotra & Vishnoi, 2023b) adapt it to submodular objectives in recommendation systems. Our work builds on (Kleinberg & Raghavan, 2018)'s bias model but diverges by considering multiple institutions and focusing on ensuring fairness relative to candidates' preferences, where group-wise constraints alone are insufficient.

Ranking, which requires ordering candidates where earlier positions are more desired, can be regarded as a specialized form of assignment. Each rank position acts as an institution with a single capacity, and all candidates share identical preferences for earlier ranks over later ones.

(Celis et al., 2020) extend (Kleinberg & Raghavan, 2018)'s insights to ranking by showing that maintaining proportional group representation at each rank ensures near-optimal latent utility. However, the institution-specific constraints we propose differ significantly from the prefix-based constraints of (Celis et al., 2020), which assume uniform candidate preferences and are not applicable in contexts like ours where institutions vary in capacity. In addition, our study incorporates preference-based fairness, a crucial element in the settings we aim to address (Theorem 3.2).

Matchings under distributional constraints. Several works study matchings under distributional constraints; see (Goto et al., 2017; Kamada & Kojima, 2017; 2015) and the references therein. Both–the works on distributional constraints in matching and this work–consider the task of computing a centralized assignment between two sets of agents (denoting, e.g., individuals, institutions, or items). However, there are several differences: Works on distributional constraints in matching consider the formulation where agents on both sides have preferences, whereas we consider the problem where agents on one side (denoting candidates) have individual preferences and the agents on the other side (denoting institutions) have certain (common) utilities for the individuals and their goal is utility maximization. Our motivation comes from centralized admission contexts such as the admissions administered by the Joint Seat Allocation Authority in India where institutions share a common utility, e.g., scores in a standardized exam, and the goal of a centralized body is to match individuals to institutions to maximize the (total) utility of the institutions while respecting the preferences of the individuals as far as possible. Moreover, the constraints considered by works in matching and mechanism design are similar to the ones considered in this work: the constraints are lower bounds or upper bounds on the number of individuals that can be matched to one or a set of institutions. The key difference is that the constraints in these works apply to all individuals, whereas we consider different constraints for candidates in different socially salient groups (to counter bias against certain groups). To the best of our knowledge, unlike this work, works on distributional constraints in matchings do not consider biases in the preferences and/or utilities.

Other generalizations of stable assignments. There has been much work on generalizations of stable matchings, including weighted stable matchings and stable matchings in non-bipartite settings (see e.g., (Vande Vate, 1989; Irving, 1985; Rothblum, 1992)). In the context of bias, (Castera et al., 2022) considered a stable assignment setting with two groups of candidates and two institutions where both sides have preferences. In this model, the correlation between the two utility values assigned to a candidate by the institutions depends on the group to which the candidate belongs. However, the distribution of the utility of a candidate for a specific institution remains the same for both groups. In contrast, the utilities of a candidate for different institutions remain the same in our model, and hence, are always perfectly correlated. Instead, the distribution of the utility of a candidate is group-specific – for one of the groups, the utility values get scaled down by the bias parameter β.

B. Additional Discussion on Theorem 3.2

In this section, we give a detailed discussion on issues with potential approaches for deriving an algorithm with the guarantees mentioned in Theorem 3.2. We also discuss properties of the algorithm $\mathcal{A}_{\text{inst-wise}}$.

Properties of the algorithm $\mathcal{A}_{\text{inst-wise}}$. We now state some additional properties of the algorithm $\mathcal{A}_{\text{inst-wise}}$:

(a) **Group-wise stability**: The algorithm $\mathcal{A}_{\text{inst-wise}}$ has the property that the assignment it produces is stable when restricted to candidates from the same group. Indeed, by definition, $\mathcal{A}_{\text{inst-wise}}$ executes $\mathcal{A}_{\text{st}}$ on each group (with certain institutional capacities). Since $\mathcal{A}_{\text{st}}$ outputs a stable assignment, it follows that $\mathcal{A}_{\text{inst-wise}}$ outputs a group-wise stable assignment.

(b) **Strategy proof property**: Assuming candidates' socially salient groups are pre-determined, $\mathcal{A}_{\text{inst-wise}}$ is strategy proof for the candidates with respect to their respective groups. Moreover, candidates have no incentive to misreport their preferences. Indeed, if a candidate that prefers institution i as a top choice, reports another institution i' as its top choice, then the likelihood that $\mathcal{A}_{\text{inst-wise}}$ assigns it i only decreases.

(c) **Pareto efficiency**: The algorithm $\mathcal{A}_{\text{inst-wise}}$ is Pareto-efficient. Consider two candidates i and i' with identical preference lists and $\widehat{u}_i > \widehat{u}_{i'}$. Then the institution assigned to i is preferable (to i and i') to the institution assigned to i'.

Algorithms that learn the bias parameter β. One potential approach for proving Theorem 3.2 would be learn the parameter β from the given data. Once β is known, the utilities of the disadvantaged group can be corrected. In some contexts, if the scenario consisted of repeated interactions, and we had access to historical data, one may learn the most likely value of β. Exploring various interventions in situations involving repeated centralized matching is a crucial direction for future research. However, we follow prior works (see e.g. (Kleinberg & Raghavan, 2018; Celis et al., 2020; Emelianov et al., 2020; Mehrotra et al., 2022)) that focus on one round-settings. In such scenarios, obtaining an estimate of β can be difficult. We also note that algorithms dependent on estimates of β may degrade due to inaccuracies in these estimates and are not resilient to explicit biases, such as those from human evaluators. For instance, consider a setting where an interviewer

estimates the utilities of the candidates. One potential approach could be to re-scale the estimated utilities reported by the interviewer. However, if the interviewer is explicitly biased, they can assign disproportionately low scores to a particular group of candidates; this would make any attempt to learn β and re-scale utilities insufficient. In contrast, the algorithms that are typically used in the real world settings do not rely on estimates of β and hence, are more resilient to explicit biases.

Algorithms with group-wise representational constraints. We now consider algorithms that add proportional representation constraints to $\mathcal{A}_{\text{st}}$. More concretely, we select the top $|G_j|/n$ fraction of candidates from each group G_j. This ensures that the representational fairness of such an algorithm is 1. To maintain group-wise stability property, we now run $\mathcal{A}_{\text{st}}$ on these candidates. See Algorithm 2 for a formal description of this algorithm $\mathcal{A}_{\text{group}}$. While this algorithm can be shown to enhance representational fairness while obtaining near-optimal utility, preference-based fairness may be low. Indeed, consider an example consisting of two institutions, each with capacity $K/2$, and $|G_1| = |G_2| = K/2$. Assume that each candidate prefers the first institution to the second one. Since $n = K$, the algorithm selects all the candidates. Assuming β is close to 0, the observed utilities of candidates from G_2 would be higher than those from G_1. Therefore, the algorithm $\mathcal{A}_{\text{group}}$ would assign the first institutions to candidates from the first group only. Thus, the preference-based fairness metric of $\mathcal{A}_{\text{group}}$ will be close to 0. We validate this fact empirically as well (see Figure 2), where $\mathcal{A}_{\text{inst-wise}}$ performs much better than $\mathcal{A}_{\text{group}}$ in the preference-based fairness metric.

Algorithms that explicitly enforce preference-based fairness. We consider algorithms that explicitly enforce preference-based fairness and maximize utility subject to such a constraint. More concretely, consider the following algorithm (assume for the sake of simplicity that $|G_1| = |G_2|$): each institution ℓ considers the candidates that prefer ℓ as their first choice. We assign the highest utility candidates from this subset to institution ℓ while ensuring that no more than half the capacity at this institution is allocated to a single group (in case some slots are vacant at the end, we can fill them with the highest utility unassigned candidates). Such an algorithm $\mathcal{A}$ would indeed achieve high preference-based fairness, but may not obtain near-optimal utility even when there is no bias. The following example illustrates the issue: suppose there are two institutions A and B, each with capacity $K/2$, and K candidates in each of the two groups. The distribution over preferences is such that $0.75K$ candidates from each of the groups prefer institution A to B. Each candidate's utility is drawn uniformly from $[0, 1]$.

We now show that the above-mentioned algorithm has sub-optimal utility in this example. We recall the following fact (see Proposition C.22): suppose we choose m values independently and uniformly at random from the interval $[0, 1]$, and let $Y_{k,m}$ denote the sum of the highest k values among these m values. Then $\mathbb{E}[Y_{k,m}] = k - k(k+1)/2(m+1)$.

Now, the expectation of optimal utility in the above-mentioned example, i.e., the expected utility of the top K candidates is the same as $\mathbb{E}[Y_{K,2K}] = 1 - \frac{K}{2K+1} \approx 0.75K$. Now, there are $0.75K$ candidates in each group preferring institution A. The algorithm selects the top $0.25K$ such candidates from each group. Therefore, the expected utility of the candidates assigned to institution A is: $2\,\mathbb{E}[Y_{0.25K,0.75K}] \approx 0.5K(1 - 1/6) \approx 0.42K$. Now, there are $0.25K$ candidates from each group preferring institution B, and the algorithm selects all of them. Therefore, the expected utility of the candidates assigned to institution B is $2\,\mathbb{E}[Y_{0.25K,0.25K}] = 0.25K$. Thus, the total expected utility of the candidates selected by this algorithm = $0.125K + 0.42K = 0.67K$, which is below the optimal value of $0.75K$. Intuitively, this gap occurs because the set of candidates with optimal utility is expected to have $0.75K$ candidates that prefer A; whereas the algorithm selects only $0.5K$ such candidates.

This example can be extended to create an instance such that the expected optimal utility is almost twice the expected utility of candidates selected by the algorithm. Consider a setting with n candidates, p institutions, and total capacity K with each institution having K/p capacity. Assume that $n \gg K$ and the utility of each agent is drawn uniformly from $[0, 1]$. Now, suppose the distribution over preference lists is such that $n - K$ candidates prefer the first institution as their top choice, and for every other institution ℓ, K/p candidates prefer ℓ as their top choice. Now, the expected optimal utility of top K candidates is about $K(1 - K/2n) \approx K$ since $n \gg K$. For an institution ℓ other than the first institution, let J_ℓ denote the K/p candidates that prefer ℓ as their first choice. The algorithm would assign J_ℓ to institution ℓ. The total expected utility of candidates in J_ℓ is $\frac{K}{2p}$. The total utility of the candidates assigned to the first institution can be at most $\frac{K}{p}$. Summing over all institutions, we see that the total expected utility of selected candidates is at most $\frac{Kp}{2p} + K \approx \frac{K}{2}$, which is roughly half of the expected optimal utility.

C. Proofs

In this section, we formally prove Theorem 3.1 (and its generalization Theorem C.12) and Theorem 3.2. In Section C.1, we prove crucial properties of a stable assignment in our setting. These properties are needed in proving Theorem 3.1 and Theorem 3.2. In particular, we show that there is a unique stable assignment of candidates to institutions (Proposition C.4). Further, this stable assignment has the following Lipschitz properties: changing the preference list of one candidate or changing the capacity of an institution by one unit affects the assignment of at most $p+1$ candidates (Lemma C.6, Lemma C.7). Using this Lipschitz property, we show that given any instance and a subset S of candidates, the number of selected candidates in S and the number of candidates in S receiving their top choice are concentrated around their respective means (C.9). Finally, we show that when we consider an instance $\mathcal{I}$ (on a single group of candidates only), and *scale* it by a factor $a \geq 1$, i.e., scale down the number of candidates and the capacities by factor a, the expected fraction of selected candidates that get their top choice in the two instances remain close (Lemma 3.5).

In Section C.2, we prove Theorem C.12 which is a generalization of Theorem 3.1 to arbitrary values of $|G_1|, |G_2|$ and K. In Sections C.2.1, C.2.2 and C.2.3, we bound the metrics $\mathscr{R}(\mathcal{A}_{\text{st}}), \mathscr{U}(\mathcal{A}_{\text{st}})$ and $\mathscr{P}(\mathcal{A}_{\text{st}})$ respectively. We show the properties of the algorithm $\mathcal{A}_{\text{inst-wise}}$ by proving Theorem 3.2 in Section C.3. We show that $\mathscr{R}(\mathcal{A}_{\text{inst-wise}}), \mathscr{U}(\mathcal{A}_{\text{inst-wise}})$ and $\mathscr{P}(\mathcal{A}_{\text{inst-wise}})$ are near-optimal in Sections C.3.1, C.3.2 and C.3.3 respectively. We now mention the concentration bounds that will be useful in the subsequent proofs.

Preliminary concentration bounds. We state the concentration inequalities that will be useful in the analysis.

Theorem C.1 (Hoeffding's Bound). *Let $X_1, \ldots, X_m$ be independent random variables such that $X_i \in [0,1]$ for each $i = 1, \ldots, m$. Let X denote $X_1 + \ldots + X_m$ and $\mu := \mathbb{E}[X]$. Then*

$$\Pr[X - \mu(X) \geq t] \leq e^{-2t^2/m},$$

and

$$\Pr[X - \mu(X) \leq t] \leq e^{-2t^2/m}.$$

The above bound also holds when the random variables $X_1, \ldots, X_m$ are negatively correlated (Dubhashi & Panconesi, 2009). For more general functions than the sum of the random variables, similar concentration bounds hold provided the function is Lipschitz.

Theorem C.2 ((Dubhashi & Panconesi, 2009))**.** *Let $T : \Omega^n \to \Re$ be a function of n variables, each lying in a domain Ω, such that for any two vectors $\boldsymbol{a}$ and $\boldsymbol{a'}$ differing in only one coordinate,*

$$|T(\boldsymbol{a}) - T(\boldsymbol{a'})| \leq c,$$

for some parameter c. Let $X_1, \ldots, X_n$ be n independent random variables, each taking values in Ω. Then,

$$\Pr\left[|T(X_1, \ldots, X_n) - \mathbb{E}[T]| \geq t\right] \leq 2e^{-2t^2/(c^2 n)}.$$

Here $\mathbb{E}[T]$ denotes the expectation of $T(Y_1, \ldots, Y_n)$, where the random variables Y_j are independent, and each $Y_j \sim X_j$.

C.1. Properties of a stable assignment of candidates to institutions

We fix an input $\mathcal{I}$ consisting of n candidates and p institutions, with institution ℓ having capacity k_ℓ. The utilities of the candidates are drawn independently from a distribution $\mathcal{D}$ over $[0,1]$ and the preferences over the institutions are drawn independently for a distribution $\mathcal{L}$. This input is instantiated by a pair of tuples $(\widehat{\boldsymbol{u}}, \boldsymbol{\sigma})$ sampled from $\mathcal{D} \times \mathcal{L}$, where $\widehat{\boldsymbol{u}}$ is the sequence of observed utilities of the n candidates and $\boldsymbol{\sigma}$ is the sequence of their preferences. We begin by defining a stable assignment.

Definition C.3. *Given $\widehat{\boldsymbol{u}}, \boldsymbol{\sigma}$, consider an assignment M of candidates to institutions. We say that M is* stable *if for every pair of candidates i and institution $j \neq M(i)$, the following condition must not hold: candidate i prefers j to its assignment $M(i)$, and institution j prefers i to all the candidates assigned to it by M.*

Observe that a stable assignment cannot leave a slot unassigned if $n \geq K$. Given arbitrary rankings of candidates by each institution and vice versa, stable assignments always exist; and they may not be unique. However, in the special setting where the institutions agree on a central ranking of the candidates, there is a unique stable assignment.

Proposition C.4. *Given $\widehat{\boldsymbol{u}}, \boldsymbol{\sigma}$, there is a unique stable assignment of candidates to institutions.*

Proof. We prove this by induction on the number of candidates. As a base case, when there are no candidates, the statement is vacuously true.

Suppose the claim is true for $n-1$ candidates, and consider an instance $\mathcal{I} := (\widehat{\boldsymbol{u}}, \boldsymbol{\sigma})$ with n candidates. Let M be a stable assignment. Let $i^\star$ be the candidate with the highest utility $\widehat{u}_{i^\star}$. We first claim M must assign $\widehat{u}_{i^\star}$ to its most preferred institution, say $j^\star$. Suppose not. Consider the candidate $i^\star$ and the institution $j^\star$. The candidate $i^\star$ prefers $j^\star$ to $M(i)$. Further, $j^\star$ prefers $i^\star$ to all the candidates assigned to it by M. But this contradicts the fact that M is stable.

Now we define an instance $\mathcal{I}'$ obtained by restricting $\mathcal{I}$ to the candidates in $[n] \setminus \{i^\star\}$, and institution $j^\star$ having $k_{j^\star} - 1$ capacity (all other institutions have the same capacity as in $\mathcal{I}$). The assignment M' given by the restriction of M to $[n] \setminus \{i^\star\}$ is also a stable assignment for the instance $\mathcal{I}'$. By induction hypothesis, M' is uniquely determined by the instance $\mathcal{I}'$. It follows that M is also uniquely determined. □

We shall use $\mathcal{A}_{\mathrm{st}}$ to denote the unique stable assignment algorithm. The proof of Proposition C.4 shows that $\mathcal{A}_{\mathrm{st}}$ can be specified as follows: sort the candidates in decreasing order of estimated utility. Now consider the candidates in this order, and when considering candidate i, assign it to the highest preferred institution with a vacant slot. It is also worth noting that $\mathcal{A}_{\mathrm{st}}$ does not depend on the actual values of the estimated utilities, but only on the relative ordering of the candidates based on these values. A formal description of this algorithm is given in Algorithm 1.

Lipschitz properties of the stable assignment $\mathcal{A}_{\mathrm{st}}$. We now show that changing the preference of only one candidate in an instance changes the resulting assignment in a few coordinates. This result shall be useful in deriving concentration bounds for the number of candidates that get their most preferred choice of institution. We first give an example showing that such Lipschitz properties may not hold in general stable matching instances.

Example C.5. *Consider an instance consisting of two institutions A and B, each having capacity $K/2$. There are K candidates that are partitioned into two types: type-I candidates prefer institution A to B and type-II candidates prefer B to A. There are $K/2$ candidates of both types. Now, institution A prefers candidates of type-II to those of type-I (the preference ordering for candidates of the same type can be arbitrary), whereas institution B prefers candidates of type-I to those of type-II.*

Consider the following assignment M: assign all type-II candidates to A and all type-I candidates to B. It is easy to check that this is a stable assignment. Now, suppose we remove a candidate of type-I. We argue that any stable assignment M' must assign all type-I candidates to A. Since there are only $K-1$ candidates, one slot shall remain unassigned. Suppose this slot is from institution B. Since there $K/2$ candidates of type-II, M' must have assigned a candidate j of type-II to A. But this causes an instability because j prefers B to A. Therefore, this vacant slot must be from institution A. Now, if M' assigns a type-I candidate i to B, it is again an instability because i prefers A and A has a vacant slot. Thus, M' must assign all type-I candidates to A (and all type-II candidates to B).

We see that M and M' do not agree on the assignment of any candidate. Similar extreme examples can be shown to exist if we change the preference list of one candidate or change the capacity of an institution by one unit.

The above example does not apply to our setting because all institutions agree on a common ranking of the candidates. We show the following Lipschitz property of the assignment given by $\mathcal{A}_{\mathrm{st}}$:

Lemma C.6. *Consider instances $\mathcal{I}$ and $\mathcal{I}'$ on a set of n candidates with the utility vector given by $\widehat{\boldsymbol{u}}$. Assume that the corresponding preference orderings $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma'}$ differ in at most one coordinate (i.e., there is at most one candidate $i^\star$ such that $\sigma_{i^\star} \neq \sigma'_{i^\star}$). Let M and M' be the assignments given by $\mathcal{A}_{\mathrm{st}}$ on the instances $\mathcal{I}$ and $\mathcal{I}'$ respectively. Then, M and M' differ in at most $p+1$ candidates. Further, for any institution $\ell \in [p]$, the set of candidates assigned to ℓ in the two assignments, i.e., $M^{-1}(\ell)$ and $M'^{-1}(\ell)$, differ in at most 1 candidate (other than the candidate $i^\star$).*

Proof. Order the candidates in decreasing order of $\widehat{u}_i$ values. Let $i^\star$ be the candidate for which $\sigma_{i^\star} \neq \sigma'_{i^\star}$. We run the algorithm $\mathcal{A}_{\mathrm{st}}$ on these two instances. Recall that the algorithm $\mathcal{A}_{\mathrm{st}}$ considers the candidates in decreasing order of utility, and when considering a candidate i, it assigns i to the most preferred institution with a vacant slot.

Let M_i and M'_i be the resulting assignments after $\mathcal{A}_{\text{st}}$ has considered the first $(i-1)$ candidates in the instances $\mathcal{I}$ and $\mathcal{I}'$ respectively. Clearly, $M_{i^\star} = M'_{i^\star}$. We claim that the following invariant holds for each $i > i^\star$: let J_i be the set of institutions for which M_i occupies more slots than M'_i. Define J'_i similarly. Then $|J_i|, |J'_i| \leq 1$. Further, if $j \in J_i$, then M_i occupies exactly one more slot on institution j than that in M'_i (and similarly for J'_i).

We prove the above statement by induction on i. When $i = i^\star + 1$, this follows from the fact that $M_{i^\star} = M'_{i^\star}$ and candidate $i^\star$ may get assigned to two different institutions in the respective instances. Now suppose $i > i^\star + 1$ and assume that the induction hypothesis holds for M_i and M'_i. In the next step, the candidate i needs to be assigned. Suppose M and M' assign candidate i to j and j' respectively. If $j = j'$, then it is easy to see that the invariant conditions hold for M_{i+1} and M'_{i+1} as well. Hence, assume $j \neq j'$. Assume w.l.o.g. that candidate i prefers j to j'. Then institution j must be full in the assignment M'_i (but it is not full in M_i). Therefore $j \in J'_i$ (in fact, by the induction hypothesis, $J'_i = \{j\}$). By induction hypothesis, j had only one vacant slot in M_i, and this becomes full in M_{i+1}. Thus, institution j is fully occupied in both M_{i+1} and M'_{i+1}, and hence does not belong to J'_{i+1}. Now i is assigned to j' by M'_{i+1}. If $j' \in J_i$, then J_{i+1} becomes empty. Otherwise, j' gets added (as the only institution) to J'_{i+1}. In either case, the invariant condition continues to hold after candidate i has been assigned in both instances. This completes the proof of the induction hypothesis.

Let M and M' be the resulting assignments in the two instances. Let D be the set of candidates i (other than $i^\star$) such that M and M' assign i to different institutions. For each $i \in D$, we associate an institution $j(i)$ as follows: it is the institution among $M(i)$ and $M'(i)$ which is preferred by i.

We claim that for any two distinct $i_1, i_2 \in D$, $j(i_1) \neq j(i_2)$. Indeed, suppose $\widehat{u}_{i_1} > \widehat{u}_{i_2}$. Then $\mathcal{A}_{\text{st}}$ considers i_1 before i_2. Assume w.l.o.g. that M assigns i_1 to $j(i_1)$. Consider the assignments M and M' just before i_1 is considered. It must be the case that $j(i_1)$ is already full in M'. By the invariant condition, $j(i_1)$ would have only one vacant slot in the assignment M at this time. This vacant slot would get occupied by i_1. Thus $j(i_1)$ gets full in both the assignments. Henceforth, no candidate can be assigned to $j(i_1)$. Thus, $j(i_2) \neq j(i_1)$.

The above claim shows that $|D| \leq p$ and this proves the first statement in the lemma. The second statement follows from the fact that D contains at most one candidate that is assigned to a particular institution ℓ by either M or M'. □

The above result shows that if two instances differ in the preference ordering of only one candidate, the resulting assignments can differ for at most $(p+1)$ candidates. Hence, the number of candidates that get their first choice in the two instances may also differ by at most $(p+1)$. A similar result holds if two instances differ slightly in the available capacity at each institution.

Lemma C.7. *Consider instances $\mathcal{I}$ and $\mathcal{I}'$ on a set of n candidates and p institutions with the utility vector given by $\widehat{\boldsymbol{u}}$ and preference lists given by $\boldsymbol{\sigma}$. In both the instances, each institution $\ell \in [p]$ has capacity k_ℓ, except for one particular institution $\ell^\star$ where $\mathcal{I}$ has capacity $k_{\ell^\star}$, but $\mathcal{I}'$ has capacity $k_{\ell^\star} + 1$. Let M and M' be the assignments given by $\mathcal{A}_{\text{st}}$ on the instances $\mathcal{I}$ and $\mathcal{I}'$ respectively. Then, M and M' differ in at most p candidates. Further, for any institution $\ell \in [p]$, the set of candidates assigned to ℓ in the two assignments, i.e., $M^{-1}(\ell)$ and $M'^{-1}(\ell)$, differ in at most 1 candidate.*

Proof. We use Lemma C.6 to prove this result. Create a new instance $\mathcal{I}_1$ which has $n+1$ candidates and $p+1$ institutions. The first candidate, call it candidate 0, is not present in the instances $\mathcal{I}$ and $\mathcal{I}'$; but candidate $i \in [n]$ in $\mathcal{I}_1$ corresponds to candidate i in $\mathcal{I}$ (or $\mathcal{I}'$). Similarly, label the first institution in $\mathcal{I}_1$ as 0, and the remaining institutions are in 1-1 correspondence with those in $\mathcal{I}$ or $\mathcal{I}'$.

Candidate 0 has the highest utility and has institution $\ell^\star$ as its most preferred institution. For the rest of the candidates $i \in [n]$, their preference list in $\mathcal{I}_1$ is given by σ_i along with institution 0 as the last choice. It is easy to check that when we run $\mathcal{A}_{\text{st}}$ on $\mathcal{I}_1$ to get an assignment M_1, candidate 0 is assigned to $\ell^\star$. The next $K := \sum_\ell k_\ell$ candidates are assigned to the same institutions as in M. Thus, among the first $K+1$ candidates, M and M_1 agree on all candidates except for candidate 0.

Now, create another instance $\mathcal{I}_2$ which is the same as $\mathcal{I}_1$ except that candidate 0 has institution 0 as the first choice. It is now easy to verify that the resulting assignment M_2 agrees with M' on all the first $K+1$ candidates, except for candidate 0. Applying Lemma C.6 to $\mathcal{I}_1$ and $\mathcal{I}_2$, we see that for each institution $\ell \in [p]$, $M_1^{-1}(\ell)$ and $M_2^{-1}(\ell)$ differ in at most 1 candidate. The lemma now follows from the fact that M and M_1 agree on all candidates in $[K]$, and similarly for M' and M_2. □

Applying the above lemma repeatedly, we get the following result.

Corollary C.8. *Consider two instances $\mathcal{I}$ and $\mathcal{I}'$ on n candidates and p institutions with the same utility and preference lists of the candidates. For each $\ell \in [p]$, let k_ℓ and k'_ℓ denote the capacity of institution ℓ in the instances $\mathcal{I}$ and $\mathcal{I}'$ respectively. Let D denote $\sum_\ell |k_\ell - k'_\ell|$. Then the stable assignments in the two instances differ on at most $p \cdot D$ candidates.*

Concentration bounds for assignments generated by the stable assignment. We now apply the Lipschitz property proved in Lemma C.6 to show that several random variables associated with the assignment on a random instance are concentrated around their respective means.

Lemma C.9. *Consider an instance $\mathcal{I}$ on n candidates and p institutions with total capacity K, where the utilities and the preference lists of candidates are drawn independently from distributions $\mathcal{D}$ and $\mathcal{L}$ respectively. Let G be a subset of candidates belonging to the same group and S denote the set of candidates selected by executing $\mathcal{A}_{\text{st}}$ on $\mathcal{I}$. Then.*

$$\Pr\left[\Big||S \cap G| - \mathbb{E}[|S \cap G|]\Big| \geq t\right] \leq 2e^{-2t^2/|G|}.$$

Condition on the observed utilities of the candidates. Let $\pi(S)$ be the set of candidates that are assigned their first choice by $\mathcal{A}_{\text{st}}$ and G be any subset of candidates. Then,

$$\Pr\left[\Big||\pi(S) \cap G| - \mathbb{E}[|\pi(S) \cap G|]\Big| \geq t\right] \leq 2e^{-2t^2/(p^2K)},$$

where the probability is over the random choice of preference lists of candidates.

Proof. For each candidate i, let X_i be the indicator variable for the event that candidate i is selected by $\mathcal{A}_{\text{st}}$. Note that X_i is selected only if $\widehat{u}_i$ appears among the top-K observed utilities. Now observe that if i, i' are two candidates belonging to the same group, then X_i and $X_{i'}$ are negatively correlated. Indeed, suppose both belong to group G_j for some $j \in \{1, 2\}$. Condition on the number of candidates selected by $\mathcal{A}_{\text{st}}$ from G_j: say this number is K_j. Let E be the event that K_j candidates from G_j are selected. Then i is selected if and only if $\widehat{u}_i$ occurs among the top K_j candidates from G_j. Thus, $\Pr[X_i = 1|E] = \frac{1}{|G_j|}$. Now, $\Pr[X_i = 1|X_{i'} = 1, E] = \frac{1}{|G_j|-1} < \Pr[X_i = 1|E]$. Removing the conditioning (since this holds for all choices of number of selected candidates from G_j), we see that $\Pr[X_i = 1|X_{i'} = 1] < \Pr[X_i = 1]$. Thus, the random variables $X_i, i \in G$, are negatively correlated.

Let S be the set of selected candidates. Then $|S \cap G| = \sum_{i \in G} X_i$. The first inequality in the lemma now follows from applying Theorem C.1 for negatively correlated random variables.

We now prove the second inequality in the lemma. We have conditioned on the observed utilities of the candidates. Order the candidates in decreasing order of observed utility. Now, let Z_i be the preference list chosen by candidate i. Observe that the random variables Z_i are independent. Also observe that the assignment generated by $\mathcal{A}_{\text{st}}$ is a function of $Z_1, \ldots, Z_K$ (since we have fixed the utilities of all the candidates). Let $T(Z_1, \ldots, Z_K)$ be the number of candidates in G that are assigned their top choice. Lemma C.6 shows that if $(Z'_1, \ldots, Z'_K)$ are another sequence of preference lists of the K candidates respectively with $Z_i = Z'_i$ for all but one candidate, then $|T(Z_1, \ldots, Z_k) - T(Z'_1, \ldots, Z'_K)| \leq p$. The desired inequality in the lemma now follows by applying Theorem C.2 to the function $T(\cdot)$. □

Approximate additivity properties of the stable assignment. Consider a setting given by p institutions, n candidates, and capacity k_ℓ for each institution $\ell \in [p]$. Assume that the true utilities are drawn from a density $\mathcal{L}$ and the preference lists are drawn from a distribution $\mathcal{D}$. Let $\mathcal{I}$ denote this instance. For a real $a \geq 1$, define the *scaled* instance $\mathcal{I}_a$ as follows: the set of institutions is the same as in $\mathcal{I}$, and the utilities of candidates and the preference lists are also drawn from $\mathcal{L}$ and $\mathcal{D}$ respectively. But there are n/a candidates, and each institution ℓ has k_ℓ/a slots (assume that all of these quantities are integers). Given utilities $\widehat{\boldsymbol{u}}$ and preference lists $\boldsymbol{\sigma}$ of the n/a candidates, let M_a denote the assignment obtained by running $\mathcal{A}_{\text{st}}$ on $\widehat{\boldsymbol{u}}, \boldsymbol{\sigma}$. Let $\pi(M_a)$ denote the number of candidates that are assigned their first choice in the assignment M_a. We shall use M to denote the assignment M_1 for the original instance $\mathcal{I}$ (which is the same as $\mathcal{I}_1$).

It turns out that for any fixed integer a, $\mathbb{E}[\pi(M_a)]$ may not be equal to $\mathbb{E}[\pi(M)/a]$.

Example C.10. *Consider a setting with two institutions, say a and b, with each institution having capacity 1. Further, assume that there are 2 candidates. The utility of each candidate is drawn from the uniform distribution on $[0, 1]$. The*

preference list of each candidate is drawn uniformly at random from the two orderings $\{(a,b),(b,a)\}$. We claim that the expected number of candidates that are assigned their top choice is $3/2$. Indeed, the first candidate (i.e., the one with the higher utility) always gets the first choice. The second candidate gets the first choice if and only if its preference list is different from that of the first candidate. This happens with probability $1/2$. Hence, the expectation of the number of candidates getting their top choice is 1.5.

However now consider the same instance as above, but now each institution has capacity 2 and there are 4 candidates. In other words, we are scaling the above instance by doubling the capacity and the number of candidates. A routine calculation shows that the expected number of candidates receiving their top choice is strictly larger than 3. Indeed, let the candidates (ordered by decreasing utility be 1,2,3,4). Now two possibilities can happen:

- *Candidates 1 and 2 have the same preference list: Assume w.l.o.g. that the two candidates have preference list (a,b). Then both the candidates are assigned institution a. Now it is easy to check that the remaining two candidates get their top choice with probability $1/2$ each. Hence, the expected number of candidates receiving their first choice is 3.*
- *Candidates 1 and 2 have distinct preference lists. In this case, candidate 3 shall always get its first choice, and candidate 4 receives its first choice with probability $1/2$. Hence, the desired expectation is 3.5.*

Since both the above cases can occur with equal probability, the expected number of candidates receiving their first choice is 3.25, which is strictly larger than 3.

On a positive note, we now show that the gap between the quantities $\mathbb{E}[\pi(M_a)]$ and $\mathbb{E}[\pi(M_1)/a]$ can be bounded by $O(p\sqrt{K\log K})$.

Lemma 3.5. $\left|\frac{\mathbb{E}[\pi(M)]}{a} - \mathbb{E}[\pi(M_a)]\right| = O(p\sqrt{K\log K})$.

Proof. We condition on the utilities and the preference list of the n candidates in $\mathcal{I}$. Let M be the assignment obtained by running $\mathcal{A}_{\text{st}}$ on this instance. Observe that M is a deterministic assignment since we have fixed the utilities and the preference lists of the candidates.

Now, we generate a (random) instance of $\mathcal{I}_a$ as follows: we choose n/a out of these n candidates in $\mathcal{I}$ uniformly at random. Call this random subset X. We consider the corresponding instance $\mathcal{I}_a$ where the candidates are given by X, and for each $i \in X$, its utility is given by $\widehat{u}_i$ and its preference list is given by σ_i. Let M_X denote the resulting assignment (which is a random assignment because X is a random subset of $[n]$). We now relate $\mathbb{E}[\pi(M_X)]$ and $\pi(M)$ (recall that M is a deterministic assignment).

Proposition C.11.

$$\left|\mathop{\mathbb{E}}_{X}[\pi(M_X)] - \pi(M)/a\right| \le O(p\sqrt{K\log K}).$$

Proof. For an institution $\ell \in [p]$, let J_ℓ be the set of candidates assigned to ℓ by M, i.e., $J_\ell := M^{-1}(\ell)$. Define $X_\ell := X \cap J_\ell$. Define $\mathcal{E}_\ell$ as the following event:

$$||X_\ell| - |J_\ell|/a| \le 8\cdot\sqrt{K\log K}.$$

Since $|J_\ell| = k_\ell$, we claim that $\mathcal{E}_\ell$ happens with probability at least $1 - 1/K^4$. Indeed, for each $i \in J_\ell$, define the indicator variable I_i which is 1 iff $i \in X_\ell$. The variables I_i are negatively correlated and $X_\ell = \sum_{i\in J_\ell} I_i$. The desired concentration bound now follows from Theorem C.1.

Let $X_0 := X \cap \pi(M)$ be the number of candidates in X that are assigned their first choice by the assignment M. Define $\mathcal{E}_0$ as the event that

$$||X_0| - \pi(M)/a| \le 8\cdot\sqrt{K\log K}.$$

We can again argue as above that $\mathcal{E}_0$ occurs with probability at least $1 - 1/K^4$. Thus, all the events $\mathcal{E}_\ell$, $\ell = 0,\ldots,p$, occur with probability at least $1 - 1/K^3$.

Assume that $\mathcal{E}_\ell$ occurs for all $\ell \in \{0\}\cup[p]$, and now, we condition on the cardinalities of the sets X_ℓ. Define $x_\ell := |X_\ell|$. Define a (random) instance $\mathcal{I}'$ where the set of candidates is given by X and each institution ℓ has x_ℓ seats. Then the assignment M restricted to X is a stable assignment for this instance. Let $\pi(\mathcal{I}')$ be the number of candidates in $\mathcal{I}'$ that are

assigned their first preference by the stable assignment. Then, $\pi(\mathcal{I}') = |X_0|$. For an institution ℓ, the fact that $\mathcal{E}_\ell$ occurs implies that $|k_\ell/a - x_\ell| \le 8\sqrt{K \log K}$.

Recall that the instance $\mathcal{I}_a$ is defined on X as the set of candidates and institution ℓ has k_ℓ/a seats. Comparing $\mathcal{I}_a$ and $\mathcal{I}'$, and applying Corollary C.8, we see that

$$|\pi(M_X) - \pi(\mathcal{I}')| = O(p\sqrt{K \log K}).$$

Since $\pi(\mathcal{I}') = |X_0|$, using the definition of the event $\mathcal{E}_0$, we see that

$$|\pi(M_X) - \pi(M)/a| = O(p\sqrt{K \log K}).$$

Since the above inequality holds for any choice of $x_\ell, \ell \in \{0\} \cup [p]$, it also holds when we remove conditioning on the sizes of the sets X_ℓ. Now if any of the events $\mathcal{E}_\ell, \ell \in \{0\} \cup [p]$, do not occur, we upper bound $\pi(M_x)$ by K. Since the probability of such an event is at most $1/K^3$,

$$\left|\mathop{\mathbb{E}}_{X}[\pi(M_X)] - \pi(M)/a\right| \le O(p\sqrt{K \log K}) + \frac{K}{K^3} = O(p\sqrt{K \log K}).$$

This proves the desired claim. □

Now consider an input to the instance $\mathcal{I}$ with n candidates where $(\widehat{\boldsymbol{u}}, \boldsymbol{\sigma})$ is drawn from the distribution $\mathcal{D}^n \times \mathcal{L}^n$. Now if we choose a random subset X of n/a candidates, the resulting distribution on $(\widehat{\boldsymbol{u}}, \boldsymbol{\sigma})$ restricted to X is drawn from $\mathcal{D}^{n/a} \times \mathcal{L}^{n/a}$. Thus, $\mathbb{E}_X(\pi(M_X))$ is same as $\mathbb{E}(\pi(I_a))$ and hence, the lemma follows from Proposition C.11. □

C.2. Proof of Theorem 3.1 and extensions

In this section, we prove the following result that characterizes the fairness and utility metrics of the stable assignment algorithm $\mathcal{A}_{\text{st}}$ as a function of the parameter β. Theorem 3.1 follows from the following result specialized to the setting where $n_1 = n_2 = K$.

Theorem C.12. *Consider an instance where the utilities of the candidates are drawn from the uniform distribution on* $[0, 1]$ *and the preference lists are drawn from an arbitrary distribution. Assume there are constants* η_1, η_2 *such that* $n_j \ge \eta_1 n$ *for each* $j \in \{1, 2\}$, $K \ge \eta_2 n$. *For any algorithm* $\mathcal{A}$ *that, given* $\widehat{\boldsymbol{u}}$ *and* $\boldsymbol{\sigma}$, *outputs an assignment maximizing the estimated utility,*

$$\mathscr{R}(\mathcal{A}) = \max\left\{\frac{K - n_1(1-\beta)}{K\beta + n_2(1-\beta)}, 0\right\} \pm O\left(\tfrac{\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right)$$
$$\mathscr{U}(\mathcal{A}) = \frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K, n)} \pm O\left(\tfrac{\sqrt{\log n}}{\eta_2\sqrt{n}}\right)$$

Here $f(x, y) := x - \frac{x(x+1)}{2(y+1)}$, $\alpha_2 := K - \alpha_1$, $\alpha_1 := K - \frac{n_2}{\beta n_1 + n_2} \cdot \max\{K - (1-\beta)n_1, 0\}$. *Assuming* $\mathcal{A}$ *is the stable assignment algorithm* $\mathcal{A}_{\text{st}}$, $\mathscr{P}(\mathcal{A})$ *is at most*

$$\max\left\{\frac{K - n_1(1-\beta)}{K\beta + n_2(1-\beta)}, 0\right\} + O\left(\tfrac{p\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right)$$

We divide the proof into three parts. In the three parts, we bound $\mathscr{R}$, $\mathscr{U}$, and $\mathscr{P}$ respectively. In the first two parts, the algorithm $\mathcal{A}$ is any utility maximizing algorithm, but in the third part, $\mathcal{A}$ is the same as $\mathcal{A}_{\text{st}}$ (observe that $\mathcal{A}_{\text{st}}$ is also utility maximizing and hence, the first two parts apply to $\mathcal{A}_{\text{st}}$ as well).

C.2.1. Step 1: Bounding $\mathscr{R}$

For any $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_n)$, let $M_{\widehat{\boldsymbol{u}}, \boldsymbol{\sigma}}$ be the assignment output by $\mathcal{A}$ given $\widehat{\boldsymbol{u}}$ and $\boldsymbol{\sigma}$. Recall that $\mathcal{A}$ outputs an assignment maximizing the estimated utilities and hence, selects the top K candidates according to their

$\widehat{u}_i$ values. Let $S_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}$ be the subset of candidates that are assigned in $M_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}$, i.e., $S_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}} = M^{-1}_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}([p])$. When $\widehat{\boldsymbol{u}}, \boldsymbol{\sigma}$ is clear from the context, we drop the subscripts in $M_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}$ and $S_{\widehat{\boldsymbol{u}},\boldsymbol{\sigma}}$.

The main step in bounding $\mathscr{R}$ (and $\mathscr{U}$) is to compute bounds on $|S \cap G_j|$, $j \in \{1, 2\}$, that hold with high probability. These bounds are stated in the following lemma. Recall that n_j denotes $|G_j|$.

Lemma C.13. *The following events happen with probability at least $1 - O(1/n^2)$: for each $j \in \{1, 2\}$,*

$$|S \cap G_j| \in \alpha_j \pm O(\sqrt{n \log n}),$$

where $\alpha_2 := K - \alpha_1$ and α_1 is defined as follows

$$\alpha_1 := K - \frac{n_2}{\beta n_1 + n_2} \cdot \max\{K - (1-\beta)n_1, 0\}. \tag{4}$$

Proof of Lemma C.13. With probability 1, S is unique as the observed utilities $(\widehat{u}_i : i \in G_j), j \in \{1, 2\}$, are drawn i.i.d. from continuous distributions (namely, the uniform distributions on $[0, 1]$ for G_1 and on $[0, \beta]$ for G_2). Since all candidates in G_2 have observed utility at most β, any candidate in G_1 with an observed utility larger than β would get preference over the candidates in G_2. Motivated by this, we give the following definition:

Definition C.14. *Given values $v_1, v_2 \in [0, 1]$, and $j \in \{1, 2\}$, let $G_j(v_1, v_2)$ denote the subset $\{i \in G_j : v_1 \le \widehat{u}_j \le v_2\}$ of G_j.*

Now, we define several desirable events and show that each of them holds with high probability.

Definition C.15 (Event $\mathcal{E}_1$). *Define $\mathcal{E}_1$ as the event that*

$$|G_1(\beta, 1)| \in [(1-\beta)n_1 - 8\sqrt{n \log n}, (1-\beta)n_1 + 8\sqrt{n \log n}].$$

Proposition C.16. *The event $\mathcal{E}_1$ occurs with probability at least $1 - 2/n^4$.*

Proof. We apply Theorem C.1 with $X_i = 1$ iff candidate $i \in G_1$ belongs to $A_{\beta,1}$. Note that $\widehat{u}_i = u_i$ for each $i \in G_1$. Clearly, $\mathbb{E}[X_i] = (1 - \beta)$. Hence, it follows from Theorem C.1 that

$$\Pr\left[\Big||G_1(\beta, 1)| - (1-\beta)n_1\Big| \ge 8\sqrt{n \log n}\right] \le 2/n^4.$$

This proves the desired result. □

The above result suffices to estimate $|S \cap G_j|$ when $K \le (1 - \beta)n_1$:

Proposition C.17. *Suppose $K \le (1-\beta)n_1$. Then $|S \cap G_1| \ge K - 8\sqrt{n \log n}$ with probability at least $1 - 1/n^4$. Therefore, $|S \cap G_2| \le 8\sqrt{n \log n}$ with probability at least $1 - 1/n^4$.*

Proof. Assume that the event $\mathcal{E}_1$ occurs. Then

$$|G_1(\beta, 1)| \ge n_1(1-\beta) - 8\sqrt{n \log n} \ge K - 8\sqrt{n \log n}.$$

Since any candidate in $G_1(\beta, 1)$ gets preference over any candidate in G_2, the claim follows from Proposition C.16. □

When $K \le (1 - \beta)n_1$, the quantity $\alpha_1 = K$ and $\alpha_2 = 0$. Thus, Proposition C.20 follows from Proposition C.17 in this special case. Therefore, assume that $K > (1 - \beta)n_1$ for the rest of the proof. Define $K' := K - (1 - \beta)n_1$ and let Δ denote $\frac{K'}{n_1 + n_2/\beta}$. The parameter Δ has been chosen such that the expected number of candidates whose observed utility lies in the range $[\beta - \Delta, 1]$ is exactly K. Indeed, $\mathbb{E}[|G_1(\beta - \Delta, 1)|]$ is $(1 - \beta + \Delta)n_1$ and $\mathbb{E}[|G_2(\beta - \Delta, 1)|]$ is $\Delta n_2/\beta$ (note that the observed utility of a candidate in G_2 is uniformly distributed in the range $[0, \beta]$). Therefore, the expected number of candidates whose estimated utility lies in the range $[\beta - \Delta, 1]$ is

$$(1 - \beta + \Delta)n_1 + \Delta n_2/\beta = (1-\beta)n_1 + \Delta\,(n_1 + n_2/\beta) = K. \tag{5}$$

The rest of the proof shows that the above-mentioned events occur with high probability.

Definition C.18 (Events $\mathcal{E}_2$ and $\mathcal{E}_3$). *Define $\mathcal{E}_2$ as the event that*

$$|G_1(\beta - \Delta, 1)| \in n_1(1 - \beta + \Delta) \pm 8\sqrt{n \log n},$$

and $\mathcal{E}_3$ as the event that

$$|G_2(\beta - \Delta, 1)| \in n_2\Delta/\beta \pm 8\sqrt{n \log n}.$$

The following result follows in the same manner as Proposition C.16.

Proposition C.19. *Each of the events $\mathcal{E}_2$ and $\mathcal{E}_3$ occur with probability at least $1 - 2/n^4$.*

Using union bound, Proposition C.19 shows that both the events $\mathcal{E}_2, \mathcal{E}_3$ occur with probability at least $1 - 4/n^4$.

Proposition C.20. *Assume that the events $\mathcal{E}_2, \mathcal{E}_3$ occur. Then the number of candidates from G_1 and from G_2 that get selected by the assignment M lie in the range $n_1(1 - \beta + \Delta) \pm 8\sqrt{n \log n}$ and $\Delta n_2/\beta \pm 8\sqrt{n \log n}$ respectively.*

Proof. Suppose, for the sake of contradiction, that the assignment M selects more than $n_1(1-\beta+\Delta)+8\sqrt{n \log n}$ candidates from G_1. Since $K = n_1(1 - \beta + \Delta) + \Delta n_2/\beta$ (using (5)), the assignment M selects less than $\Delta n_2/\beta - 8\sqrt{n \log n}$ candidates from G_2. Since the event $\mathcal{E}_2$ occurs, the assignment M selects a candidate $i \in G_1$ with $\widehat{u}_i < \Delta$. Since the event $\mathcal{E}_3$ occurs, there must be a candidate $i \in G_2$ with $\widehat{u}_i > \Delta$ such that i is not assigned by M. But this is a contradiction. Thus, M selects at most $n_1(1 - \beta + \Delta) + 8\sqrt{n \log n}$ candidates from G_1.

The argument that the assignment M selects more than $n_1(1-\beta+\Delta)-8\sqrt{n \log n}$ candidates from G_1 follows similarly. □

It is easy to check that the quantities α_1 and α_2 as defined in the statement of Lemma C.13 are equal to $n_1(1 - \beta + \Delta)$ and $\Delta n_2/\beta$ respectively. Thus, Lemma C.13 follows from Proposition C.19 and Proposition C.20. □

Using Lemma C.13, we can easily bound $\mathcal{R}(\mathcal{A})$. First, consider the case when $K \leq (1 - \beta)n_1$. In this case the quantities α_1 and α_2 are K and 0 respectively. Lemma C.13 shows that with probability at least $1 - 1/n^2$ (recall that q_j denotes the fraction of candidates from G_j that get assigned by M),

$$\frac{q_2}{q_1} \leq \frac{n_1}{n_2} \cdot \frac{8\sqrt{n \log n}}{K - 8\sqrt{n \log n}} \leq \frac{8\sqrt{n \log n}}{\eta_1\eta_2 n/2} = O\left(\frac{\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right),$$

where the second last inequality follows from the fact that $K \geq \eta_2 n \geq 16\sqrt{n \log n}$ (for large enough n) and $n_2 \geq \eta_1 n \geq \eta_1 n_1$. Thus, with probability at least $1 - 1/n^2$, $\frac{\min(q_1,q_2)}{\max(q_1,q_2)} = O\left(\frac{\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right)$. Since the l.h.s. above can never exceed 1, we see that

$$\mathcal{R}(\mathcal{A}) = \mathbb{E}\left[\frac{\min(q_1, q_2)}{\max(q_1, q_2)}\right] \leq \left(1 - 1/n^2\right) O\left(\frac{\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right) + \frac{1}{n^2} = O\left(\frac{\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right).$$

This proves the desired bound on $\mathcal{R}(\mathcal{A})$ when $K \leq (1 - \beta)n_1$.

Now consider the case when $K > (1 - \beta)n_1$. Proposition C.20 implies that, with probability at least $1 - 1/n^2$,

$$\frac{q_2}{q_1} \leq \frac{n_1}{n_2} \cdot \frac{\Delta n_2/\beta + 8\sqrt{n \log n}}{n_1(1 - \beta + \Delta) - 8\sqrt{n \log n}} = \frac{\Delta/\beta}{(1 - \beta + \Delta)} \cdot \left(1 - \frac{8\sqrt{n \log n}}{n_1(1 - \beta + \Delta)}\right)^{-1} \cdot \left(1 + \frac{8\sqrt{n \log n}}{\Delta n_2/\beta}\right). \tag{6}$$

Recall that $\Delta = \frac{K'}{n_1+n_2/\beta}$, where $K' = K - (1 - \beta)n_1$. Using this, we see that

$$1 - \beta + \Delta \geq \frac{K}{n_1 + n_2} = \frac{K}{n} \geq \eta_2. \tag{7}$$

Combining the above observations with the fact that $n_1, n_2 \geq \eta_1 n$, the r.h.s. of (6) can be expressed as

$$\frac{q_1}{q_2} \leq \frac{\Delta/\beta}{1 - \beta + \Delta} + O\left(\frac{\sqrt{\log n}}{\eta_1\eta_2\sqrt{n}}\right). \tag{8}$$

A similar argument shows that with probability at least $1 - 1/n^2$,

$$\frac{q_1}{q_2} \geq \frac{\Delta/\beta}{1-\beta+\Delta} - O\left(\frac{\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right). \tag{9}$$

A direct calculation shows that $1 - \beta + \Delta \geq \Delta/\beta$. Indeed, this is the same as showing $\Delta \leq \beta$, which holds when $n \geq K$. Thus, assuming (8) holds, $\frac{\min(q_1,q_2)}{\max(q_1,q_2)} = \frac{q_1}{q_2} - \left(\frac{\sqrt{\log n}}{\eta_1\sqrt{n}}\right)$. Therefore,

$$\mathcal{R}(\mathcal{A}) = \mathbb{E}\left[\frac{\min(q_1,q_2)}{\max(q_1,q_2)}\right] \overset{(8),(9)}{=} \frac{\Delta/\beta}{1-\beta+\Delta} \pm O\left(\frac{\sqrt{\log n}}{\eta_1\sqrt{n}}\right).$$

Using the definition of Δ, one can check that

$$\frac{\Delta/\beta}{1-\beta+\Delta} = \frac{K-(1-\beta)n_1}{\beta K + (1-\beta)n_2}.$$

This proves the desired bound on $\mathcal{R}(\mathcal{A})$ as stated in Theorem C.12.

C.2.2. Step 2: Bounding $\mathscr{U}$

We now bound $\mathscr{U}(\mathcal{A})$; we use a similar strategy as the one used for bounding $\mathcal{R}(\mathcal{A})$. We first derive concentration bounds on the sum of random values drawn from a certain range. We begin with a standard result on k-th order statistic for n independent draws from the uniform distribution on $[0,1]$:

Fact C.21 ((Ahsanullah et al., 2013))**.** *Let $X_1, \ldots, X_n$ be n independent random variables, each drawn from the uniform distribution on $[0,1]$. Let Y_k denote the k-th largest value in $\{X_1, \ldots, X_n\}$. For all $k \in [n]$, $\mathbb{E}[Y_k] = \frac{n+1-k}{n+1}$.*

The above result implies the following concentration bound for the sum of top-k among n values drawn uniformly at random from $[0,1]$:

Proposition C.22. *Let $X_1, \ldots, X_n$ be n i.i.d. random variables, each drawn uniformly at random from $[0,1]$. Given a parameter $k \in [n]$, define the function $T_k : [0,1]^n \to \Re$ as the sum of the top k values among $x_1, \ldots, x_n$. Then*

$$\Pr\left[|T_k(X_1, \ldots, X_n) - f(k,n)| \geq t\right] \leq 2e^{-2t^2/n}.$$

Here $f(k,n) := k - \frac{k(k+1)}{2(n+1)}$.

Proof. Fact C.21 implies that $\mathbb{E}[T_k] = \sum_{i=1}^{k} \frac{n+1-i}{n+1} = k - \frac{k(k+1)}{2(n+1)} = f(k,n)$. It is easy to check that the function T_k satisfies the conditions in Theorem C.2, and hence the desired result follows from Theorem C.2. □

Now we define several events:

Definition C.23. *Define $\mathcal{E}_4$ as the following event: for each $j \in \{1,2\}$, the top-α_j utilities among candidates in the group G_j lies in the range $f(\alpha_j, n_j) \pm 8\sqrt{n \log n}$. Define $\mathcal{E}_5$ as the event that the top-K utilities among all the candidates $[n]$ lies in the range $f(K,n) \pm 8\sqrt{n \log n}$. Finally, define $\mathcal{E}_6$ as the event that for each $j \in \{1,2\}$, $|S \cap G_j|$ lies in the range $\alpha_j \pm 8\sqrt{n \log n}$.*

Recall that $S = M^{-1}([p])$ is the set of candidates that are assigned an institution by M. Proposition C.22 shows that each of the events $\mathcal{E}_4$ and $\mathcal{E}_5$ happen with probability at least $1 - 2/n^4$, and Lemma C.13 shows that $\mathcal{E}_6$ happens with probability at least $1 - O(1/n^2)$. Applying union bound, all these three events happen with probability at least $1 - O(1/n^2)$.

Proposition C.24. *Assume that the events $\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6$ happen. Then*

$$\frac{\mathsf{U}(M_{\hat{\boldsymbol{u}},\boldsymbol{\sigma}})}{\mathsf{U}^\star(\boldsymbol{u})} = \frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K,n)} \pm O\left(\frac{\sqrt{\log n}}{\eta_2\sqrt{n}}\right).$$

Proof. Assume that the events $\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6$ occur. Since M assigns at most $\alpha_j + 8\sqrt{n \log n}$ candidates from group G_j (event $\mathcal{E}_6$), and the top-α_j candidates from G_j have total utility at most $f(\alpha_j, n_j) + 8\sqrt{n \log n}$, the total utility of the candidates in $S \cap G_j$ is at most

$$f(\alpha_j, n_j) + 16\sqrt{n \log n},$$

because the utility of any particular candidate can be at most 1. Thus, the total utility of the candidates selected by M is at most

$$\sum_{j=1}^{2} f(\alpha_j, n_j) + 32\sqrt{n \log n}.$$

Event $\mathcal{E}_5$ implies that $\mathsf{U}^\star(\boldsymbol{u}) \geq f(K, n) - 8\sqrt{n \log n}$. Thus,

$$\frac{\mathsf{U}(M_{\hat{\boldsymbol{u}},\boldsymbol{\sigma}})}{\mathsf{U}^\star(\boldsymbol{u})} \leq \frac{\sum_{j=1}^{2} f(\alpha_j, n_j) + 32\sqrt{n \log n}}{f(K, n) - 8\sqrt{n \log n}}.$$

Now, $f(K, n) = K - \frac{K(K+1)}{2(n+1)} \geq K/2 \geq \eta_2 n/2$. Similarly $f(\alpha_1, n_1) + f(\alpha_2, n_2) \geq \alpha_1/2 + \alpha_2/2 = K/2 \geq \eta_2 n/2$. Therefore, the r.h.s. above is at most

$$\frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K, n)} \left(1 + O\left(\frac{\sqrt{\log n}}{\eta_2 \sqrt{n}}\right)\right) \leq \frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K, n)} + O\left(\frac{\sqrt{\log n}}{\eta_2 \sqrt{n}}\right),$$

where the last inequality follows from the fact that $\sum_{j=1}^{2} f(\alpha_j, n_j) \leq f(K, n)$. A similar argument shows that

$$\frac{\mathsf{U}(M_{\hat{\boldsymbol{u}},\boldsymbol{\sigma}})}{\mathsf{U}^\star(\boldsymbol{u})} \geq \frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K, n)} - O\left(\frac{\sqrt{\log n}}{\eta_2 \sqrt{n}}\right).$$

This proves the desired result. □

Since the ratio $\frac{\mathsf{U}(M_{\hat{\boldsymbol{u}},\boldsymbol{\sigma}})}{\mathsf{U}^\star(\boldsymbol{u})}$ always lies in the range $[0, 1]$, and the event $\mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6$ happens with probability at least $1 - O(1/n^2)$, it follows that

$$\mathscr{U}(\mathcal{A}) = \mathbb{E}\left[\frac{\mathsf{U}(M_{\hat{\boldsymbol{u}},\boldsymbol{\sigma}})}{\mathsf{U}^\star(mbu}\right] = \frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K, n)} \pm O\left(\frac{\sqrt{\log n}}{\eta_2 \sqrt{n}}\right) + O(1/n^2) = \frac{\sum_{j=1}^{2} f(\alpha_j, n_j)}{f(K, n)} \pm O\left(\frac{\sqrt{\log n}}{\eta_2 \sqrt{n}}\right).$$

This proves the desired bound on $\mathscr{U}(\mathcal{A})$ as stated in Theorem C.12.

C.2.3. Step 3: Bounding $\mathscr{P}$

In this section, we bound $\mathscr{P}(\mathcal{A})$. We use the notation developed in Appendix C.2.1. Recall that $G_j(v_1, v_2)$ denotes the set of candidates $i \in G_j$ for which $\widehat{u}_i \in [v_1, v_2]$. We assume that $K \leq (1 - \beta)n_1$ (the other case is only easier). We also assume that the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ occur (as defined in Definition C.15 and Definition C.18). More specifically, we condition on the sizes of the sets $G_1(\beta, 1)$, $G_1(\beta - \Delta, \beta)$ and $G_2(\beta - \Delta, \beta)$. Call these values g_1, g_1', g_2' respectively. Note that the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ require that

$$g_1 \in (1 - \beta)n_1 \pm 8\sqrt{n \log n},\ g_1 + g_1' \in n_1(1 - \beta + \Delta) \pm 8\sqrt{n \log n},\ g_2' \in n_2 \Delta/\beta \pm 8\sqrt{n \log n}. \tag{10}$$

One manner in which we can condition on the above events is as follows:

(i) Select g_1 values uniformly at random in the range $[\beta, 1]$ and assign these (observed) utility values to the first g_1 candidates.

(ii) Select $g_1' + g_2'$ values uniformly at random in the range $[\beta - \Delta, \beta]$ and assign these estimated utility values to the next $g_1' + g_2'$ candidates.

(iii) Select $n - (g_1 + g_1' + g_2')$ values uniformly at random in the range $[0, \beta - \Delta]$ and assign these estimated utility values to the remaining candidates.

(iv) Assign the first g_1 candidates (who were assigned utility values in step (i)) to G_1. Consider the next $g_1' + g_2'$ candidates who were assigned values in step (ii) above. Select g_1' candidates (without replacement) uniformly at random from this subset of candidates and assign them to G_1, and the remaining g_2' candidates to G_2. Finally, randomly partition the remaining $n - (g_1 + g_1' + g_2')$ candidates (used in step (iii) above) between G_1 and G_2 in a similar manner.

Finally, we assign each candidate i a random ordering σ_i chosen independently from the distribution $\mathcal{D}$. Now we run $\mathcal{A}_{\text{st}}$ on these candidates. Note that $\mathcal{A}_{\text{st}}$ does not depend on the group to which a candidate belongs. Hence, $\mathcal{A}_{\text{st}}$ only depends on the first three steps mentioned above and does not depend on step (iv).

For the sake of brevity of notation, let A denote the first g_1 candidates in $[n]$ (i.e., the candidates that are assigned values in step (i) above) and B denote the next $g_1' + g_2'$ candidates (i.e, the ones assigned values in step (ii) above). Let $\pi(A)$ and $\pi(B)$ denote the set of candidates in A and B respectively that are assigned their most preferred institutions by the algorithm $\mathcal{A}_{\text{st}}$. Using Lemma C.9, we see that

$$\Pr[|\pi(A) - \mathbb{E}[\pi(A)]| \geq p\sqrt{8n \log n}] \leq 1/n^2, \tag{11}$$

and

$$\Pr[|\pi(B) - \mathbb{E}[\pi(B)]| \geq p\sqrt{8n \log n}] \leq 2/n^2. \tag{12}$$

We now relate $\mathbb{E}[\pi(A)]$ and $\mathbb{E}[\pi(B)]$.

Proposition C.25. $\mathbb{E}[\pi(A)]/g_1 \geq \mathbb{E}[\pi(B)]/(g_1' + g_2')$.

Proof. For a candidate i, let D_i be the event that it is assigned the top preferred institution. We now argue that for any $i \in [n]$, $\Pr[D_i] \geq \Pr[D_{i+1}]$. We condition on the assignment of institutions to the first $(i-1)$ candidates. After this assignment, let S be the set of institutions that have a vacant slot. Now let $\sigma \sim \mathcal{D}$ be the preference list for candidate i and $j^\star$ be the first institution in this list. Then $\Pr[D_i]$ is equal to the probability that $j^\star \in S$. Now after assigning an institution to i, the set of institutions with vacant slots will be a subset of S. Therefore, $\Pr[D_{i+1}]$is at most the probability that $j^\star \in S$.

Thus, we see that $\Pr[D_i] \geq \Pr[D_{i+1}]$ for all i. Now,

$$\mathbb{E}[\pi(A)] = \sum_{i=1}^{g_1} \Pr[D_i] \geq g_1 \Pr[D_{g_1}],$$

and

$$\mathbb{E}[\pi(B)] = \sum_{i=g_1+1}^{g_1+g_1'+g_2'} \Pr[D_i] \leq (g_1' + g_2') \Pr[D_{g_1}].$$

The desired result now follows from the above two inequalities. □

We can now estimate the number of candidates from each of the groups G_j that are assigned their most preferred institution.

Proposition C.26. *With probability at least $1 - O(1/n^2)$, the number of candidates from G_1 and G_2 that are assigned their most preferred institution choices are $\mathbb{E}[\pi(A)] + \frac{g_1'}{g_1'+g_2'} \cdot \mathbb{E}[\pi(B)] \pm O(p\sqrt{n \log n})$ and $\frac{g_2'}{g_1'+g_2'} \cdot \mathbb{E}[\pi(B)] \pm O(p\sqrt{n \log n})$ respectively.*

Proof. All the candidates in A are assigned to G_1. By (11), the number of candidates in A that are assigned their top choices lies in the range

$$\mathbb{E}[\pi(A)] \pm p\sqrt{8n \log n} \tag{13}$$

with probability at least $1 - 1/n^2$.

Now, we condition on the subset $\pi(B)$ of candidates in B that get their first choice – let this be B'. Recall that we randomly select g_1' candidates from B and assign them to G_1. For a candidate $i \in B'$, let X_i be the probability that i is assigned to G_1 and $i \in B'$. It is easy to see that $\mathbb{E}[X_i] = \frac{g_1'}{g_1'+g_2'}$. The random variables X_i are not independent but are negatively correlated. Therefore, applying Hoeffding's bound to $\sum_i X_i$, we see that with probability at least $1 - 1/n^2$, the number of candidates in $B' \cap G_1$ is

$$\frac{g_1'|B'|}{g_1' + g_2'} \pm p\sqrt{8n \log n}.$$

Further, (12) shows that $|B'| \in \mathbb{E}[\pi(B)] \pm p\sqrt{8n \log n}$ with probability at least $1 - 2/n^2$. Combining the above two observations, we see that with probability at least $1 - 3/n^2$, the number of candidates from $G_1 \cap B$ that are assigned their top preferences lies in the range

$$\frac{g_1' \, \mathbb{E}[\pi(B)]}{g_1' + g_2'} \pm p\sqrt{16n \log n}. \tag{14}$$

Combining (13) and (14), we see that the number of candidates in $G_1 \cap (A \cup B)$ that are assigned their top choices lies in the range

$$\mathbb{E}[\pi(A)] + \frac{g_1' \, \mathbb{E}[\pi(B)]}{g_1' + g_2'} \pm O(p\sqrt{n \log n}).$$

with probability at least $1 - 4/n^2$. Proposition C.20 shows that all but $O(\sqrt{n \log n})$ selected candidates lie in the set $A \cup B$. This proves the desired bound on the number of candidates in G_1 that are assigned their first choice. The corresponding bound for G_2 follows similarly. □

Recall that π_j denotes the *fraction* of candidates in G_j that are assigned their top choice. Proposition C.26 and Proposition C.25 show that

$$\frac{\pi_2}{\pi_1} \leq \frac{n_1}{n_2} \cdot \frac{g_2' \pm O(p\sqrt{n \log n})}{g_1 + g_1' \pm O(p\sqrt{n \log n})} = \frac{n_1}{n_2} \frac{g_2'}{g_1 + g_1'} + O\left(\frac{p\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right), \tag{15}$$

where the last inequality follows from the fact that $n_1/n_2 \geq 1/\eta_1$ and $g_1 + g_1'$ is $\Omega(\eta_2 n)$ (using (7) and (10)). Using (10) again, we get

$$\frac{\pi_2}{\pi_1} \leq \frac{\Delta/\beta}{(1-\beta) + \Delta} + O\left(\frac{p\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right) = \frac{K - n_1(1-\beta)}{\beta K + (1-\beta) n_2} + O\left(\frac{p\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right),$$

where the last equality follows from the definition of Δ. The above expression does not involve g_1, g_1', g_2' and hence, holds even after removing the conditioning on these parameters. In other words, the above bound holds if the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ occur.

Since $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ occur with probability at least $1 - O(1/n^4)$, and the ratio $\frac{\min(\pi_1,\pi_2)}{\max(\pi_1,\pi_2)} \leq 1$, we see that

$$\mathbb{E}\left[\frac{\min(\pi_1,\pi_2)}{\max(\pi_1,\pi_2)}\right] \leq \frac{K - n_1(1-\beta)}{\beta K + (1-\beta) n_2} + O\left(\frac{p\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right) + O(1/n^4) = \frac{K - n_1(1-\beta)}{\beta K + (1-\beta) n_2} + O\left(\frac{p\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right).$$

This proves the desired bound on $\mathscr{P}(\mathcal{A})$. Thus we completed the proof of Theorem C.12. The special setting of Theorem 3.1 now follows immediately. We restate this result below:

Theorem 3.1. *Consider an instance where the utilities of the candidates are drawn from the uniform distribution on* $[0,1]$ *and the distribution over preferences is arbitrary. Assume* $n_1 = n_2 = K$. *Then,* $\mathscr{P}(\mathcal{A}_{\text{st}}) \leq \beta + O\left(\frac{p\sqrt{\log n}}{\sqrt{n}}\right)$, $\mathscr{R}(\mathcal{A}_{\text{st}}) = \beta \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right)$, *and* $\mathscr{U}(\mathcal{A}_{\text{st}}) = \frac{2}{3} + \frac{4\beta}{3(\beta+1)^2} \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right)$.

Proof. The use notation in the statement of Theorem C.12. The parameters η_1 and η_2 can be set to 1. Since $n_1 = n_2 = K$, $K - n_1(1-\beta) = n_1\beta$ and $K\beta + n_2(1-\beta) = n_1$. Therefore, $\mathscr{R}(\mathcal{A}_{\text{st}}) = \beta \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right)$, and $\mathscr{P}(\mathcal{A}_{\text{st}}) \leq \beta + O\left(p\frac{\sqrt{\log n}}{\sqrt{n}}\right)$. We now compute $\mathscr{U}(\mathcal{A}_{\text{st}})$. Observe that $\alpha_1 = K - \frac{K}{\beta+1} = \frac{\beta \cdot K}{\beta+1}$ and $\alpha_2 = \frac{K}{\beta+1}$. Using these, we see that (up to an error of $O(1/n)$)

$$f(\alpha_1, n_1) = \frac{2\beta + \beta^2}{2(\beta+1)^2}, \; f(\alpha_2, n_2) = \frac{2\beta + 1}{2(\beta+1)^2}, \; f(K, n) = \frac{3}{8}.$$

Substituting these values in Theorem C.12 yields the desired bound on $\mathscr{U}(\mathcal{A}_{\text{st}})$. □

The following result shows that when $\beta = 1$, the algorithm has near-optimal fairness and utility also follows directly from the proof of Theorem C.12.

Corollary C.27. *Consider an instance where the utilities of the candidates are drawn from the uniform distribution on* $[0, 1]$ *and the preference lists are drawn from an arbitrary distribution. Assume there are constants* η_1, η_2 *such that* $n_j \geq \eta_1 n$ *for each* $j \in \{1, 2\}$, $K \geq \eta_2 n$. *When the bias parameter* $\beta = 1$,

$$\mathscr{R}(\mathcal{A}_{\text{st}}) \geq 1 - O\left(\frac{\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right)$$
$$\mathscr{U}(\mathcal{A}_{\text{st}}) \geq 1 - O\left(\frac{\sqrt{\log n}}{\eta_2 \sqrt{n}}\right)$$
$$\mathscr{P}(\mathcal{A}_{\text{st}}) \geq 1 - O\left(\frac{p\sqrt{\log n}}{\eta_1 \eta_2 \sqrt{n}}\right).$$

Proof. The bounds of $\mathcal{R}(\mathcal{A}_{\text{st}})$ and $\mathscr{U}(\mathcal{A}_{\text{st}})$ follow directly from Theorem C.12 by substituting $\beta = 1$. The bound on $\mathscr{P}(\mathcal{A}_{\text{st}})$ requires an explanation because Theorem C.12 only gives a lower bound on $\mathscr{P}(\mathcal{A}_{\text{st}})$. The main observation is that the set A defined in Proposition C.26 is empty. Hence, as in (15), $\mathscr{P}(\mathcal{A}_{\text{st}})$ can be expressed as

$$\frac{n_1}{n_2} \cdot \frac{g_2' \pm O(p\sqrt{n \log n})}{g_1' \pm O(p\sqrt{n \log n})}.$$

The rest of the argument now follows as in the proof of Theorem C.12. □

C.2.4. Extension of Theorem 3.1 to log-concave densities

In this section, we show that the Theorem 3.1 extends to the setting where the utilities of candidates are drawn from an arbitrary log-concave density. We show that both the representational and the preference-based fairness metrics deteriorate with β. Finally, we give an example (Example C.30) showing that such deterioration may not happen for an arbitrary distribution over utilities.

Theorem C.28. *Let* $\mathcal{D}$ *be a distribution on* $[0, 1]$ *with log-concave density* f *and* $n_1 = n_2 = K$. *Then, for any utility maximizing algorithm* $\mathcal{A}$, $\mathscr{R}(\mathcal{A}) \leq 2\beta \ln(1/\beta) \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right)$. *and* $\mathscr{P}(\mathcal{A}_{\text{st}}) \leq 2\beta \ln(1/\beta) + O\left(\frac{p\sqrt{\log n}}{\sqrt{n}}\right)$. *Further, as* β *approaches 0,* $\mathscr{U}(\mathcal{A}_{\text{st}})$ *approaches a value that is at most* $\mathbb{E}[\mathscr{U}(\mathcal{A}_{\text{st}})]$ *is at most* $\frac{U(K,K)+U(O(\sqrt{\log K},K)}{U(K,2K)}$. *Here* $U(k, m)$ *denotes the expected sum of the top* k*-values out of* m *values chosen independently from the distribution given by* f.

Remark C.29. *(i) As in the statement of Theorem 3.1, we can also derive an upper bound on the utility metric* $\mathscr{U}(\mathcal{A})$. *But this bound shall depend on the specific density; (ii) We can also generalize the above statement to the more general setting of Theorem C.12.*

Proof. We give a sketch of the main ideas in the proof. These can be formalized as in the proof of Theorem C.12. We assume that $\beta \ln(1/\beta) \leq 0.5$, otherwise the desired bounds hold trivially.

Bounding $\mathscr{R}(\mathcal{A}_{\text{st}})$. The proof proceeds in two steps. In the first step, we reduce the problem of bounding $\mathscr{R}(\mathcal{A}_{\text{st}})$ to the special case when $f(x) \propto e^{-\lambda x}$ for some parameter λ. In the second step, we carry out explicit calculations for this special case.

We now give details of the first step of the proof. Given a value $v \in [0, 1]$, let X_v denote the number of candidates whose estimated utility lies in the range $[v, 1]$. The key idea, as in the proof of Theorem C.12, is to find the value Δ such that $\mathbb{E}[X_\Delta] = K$ (recall that the total available capacity is K). One can then show using concentration bounds (as in Theorem C.12) that the number of selected candidates is close (up to an error bound of $O(\sqrt{K})$) to $\mathbb{E}[X_\Delta]$. Now let $F(x)$ denote the c.d.f. of f, i.e., for a value $x \in [0, 1]$, $F(x) = \int_0^x f(x)dx$. Then $\mathbb{E}[X_\Delta \cap G_1] = K(1 - F(\Delta)$. Now, $X_\Delta \cap G_2$ consists of those candidates in G_2 whose true utility is at least Δ/β. Therefore, $\mathbb{E}[X_\Delta \cap G_2] = K(1 - F(\Delta/\beta))$ (recall that $n_1 = K$). Thus, the parameter Δ satisfies the condition that

$$\begin{aligned} K(1 - F(\Delta)) + K(1 - F(\Delta/\beta)) &= K \\ i.e., \quad 1 - F(\Delta/\beta) &= F(\Delta). \end{aligned} \tag{16}$$

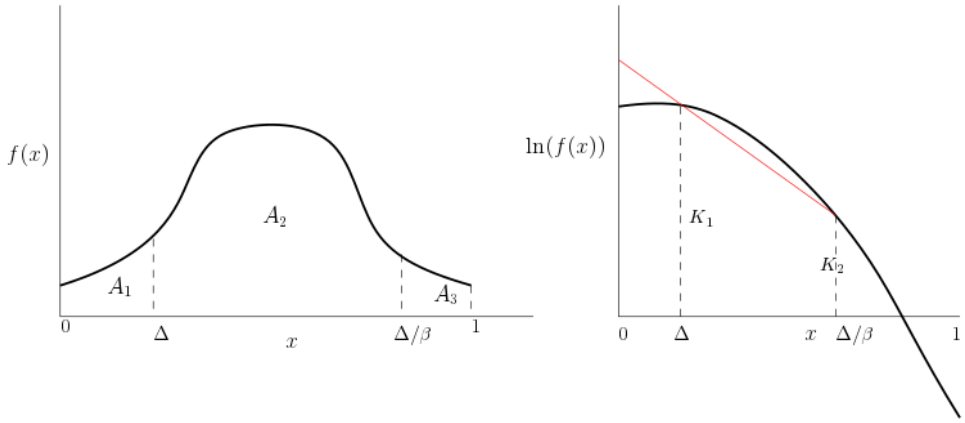


Figure 3: Illustration of the proof of Theorem C.28: The figure on the left plots the density $f(x)$. The quantities A_1, A_2, A_3 are the areas under the curve for the intervals $[0, \Delta], [\Delta, \Delta/\beta], [\Delta/\beta, 1]$. The figure on the right plots $g(x) := \ln(f(x))$, which is a concave function. The line $L(x)$ is shown in red.

This condition can be visualized as follows. We give some notation first. Given a non-negative function $h : [0, 1] \to \Re$, values $v_1 \leq v_2$ lying in $[0, 1]$, let $\mathsf{Area}_h(v_1, v_2)$ denote the area under h from v_1 to v_2, i.e., $\int_{v_1}^{v_2} h(x)dx$. Let A_1, A_2 and A_3 denote $\mathsf{Area}_f(0, \Delta), \mathsf{Area}_f(\Delta, \Delta/\beta)$ and $\mathsf{Area}_f(\Delta/\beta, 1)$ respectively (see Figure 3). Then the condition (16) is the same as $A_3 = A_1$. Further, $\mathbb{E}[X_\Delta \cap G_1] = K(1 - F(\Delta)) = K(A_2 + A_3)$ and $\mathbb{E}[X_\Delta \cap G_2] = K(1 - F(\Delta/\beta)) = K \cdot A_3$. Since we are interested in the ratio of these two expectations, we define

$$\gamma := \frac{A_3}{A_2 + A_3} = \frac{A_1}{A_1 + A_2},$$

where the equality follows from the fact that $A_1 = A_3$. Now, let $g(x)$ denote $\ln(f(x))$. Since f is log-concave, g is a concave function. Let $L(x)$ be the line that passes through the points $(\Delta, g(\Delta)$ and $(\Delta/\beta, g(\Delta/\beta))$ (see Figure 3). We assume that $L(x)$ is of the form $-\lambda x + b$, for a non-negative parameter λ (the case when $\lambda > 0$ is similar). Let $\widehat{f}(x)$ denote $e^{L(x)} = Ce^{-\lambda x}$ for some parameter C. Since g is concave, $L(x) \geq g(x)$ for $x \in [0, \Delta] \cup [\Delta/\beta, 1]$, and $L(x) \leq g(x)$ for $x \in [\Delta, \Delta/\beta]$. Therefore, if A'_1, A'_2, A'_3 denote $\mathsf{Area}_{\widehat{f}}(0, \Delta), \mathsf{Area}_{\widehat{f}}(\Delta, \Delta/\beta)$ and $\mathsf{Area}_{\widehat{f}}(\Delta/\beta, 1)$, then $A'_1 \geq A_1, A'_2 \leq A_2, A'_3 \geq A_3$. Since the ratio $\frac{a}{a+b}$ is an increasing function of a and decreasing function of b for any non-negative real a, b, it follows that

$$\gamma \leq \min\left(\frac{A'_1}{A'_1 + A'_2}, \frac{A'_3}{A'_2 + A'_3}\right). \tag{17}$$

Now, the above expression only involves the function $\widehat{f}$. Now we carry out the second step in the proof. There are two cases:

- $\lambda \leq \frac{\beta}{\Delta} \ln(1/\beta)$: We know that $\widehat{f}(x)$ is of the form $Ce^{-\lambda x}$. Hence,

$$\frac{A'_1}{A'_1 + A'_2} = \frac{\mathsf{Area}_{\widehat{f}}(0, \Delta)}{\mathsf{Area}_{\widehat{f}}(0, \Delta/\beta)} \leq 2\lambda\Delta \leq 2\beta \ln(1/\beta),$$

  where the second last inequality follows from the fact that $\mathsf{Area}_{\widehat{f}}(0, \Delta) \leq \max_{x \in [0,1]} \widehat{f}(X) \cdot \Delta = C\Delta$ and $\mathsf{Area}_{\widehat{f}}(0, \Delta/\beta) = \frac{C - Ce^{-\lambda\Delta/\beta}}{\lambda} \geq \frac{C - C\beta}{\lambda} \geq \frac{C}{2\lambda}$, where we have used the fact that $\beta \leq 1/2$ (since $\beta \ln(1/\beta)$ has been assumed to be at most 0.5).

- $\lambda > \frac{\beta}{\Delta} \ln(1/\beta)$: In this case, we see that

$$\frac{A'_3}{A'_2 + A'_3} = \frac{e^{-\lambda\Delta/\beta} - e^{-\lambda}}{e^{-\lambda\Delta} - e^{-\lambda}} \leq \frac{e^{-\lambda\Delta/\beta}}{e^{-\lambda\Delta}},$$

where the last hand inequality follows from the fact that if $0 \le a \le b$, then $\frac{a}{b} \ge \frac{a-t}{b-t}$ for any $t \ge 0$. Since the r.h.s. above is a decreasing function of λ (because $\beta < 1$), we see that it is at most $\frac{\beta}{e^{-\beta \ln(1/\beta)}} \le 2\beta$, because $\beta \ln(1/\beta)$ has been assumed to be at most 0.5.

Thus, we see that in either of the two cases, $\gamma \le 2\beta \ln(1/\beta)$. Now, let S be the set of candidates selected by $\mathcal{A}_{\text{st}}$. Arguing as in the proof of Theorem C.12, we can show that

$$\mathbb{E}\left[\frac{|S \cap G_2|}{|S \cap G_2|}\right] = \frac{A_3 \pm O(\sqrt{K \log k})}{A_2 + A_3 \pm O(\sqrt{K \log K}} + O\left(\frac{1}{n}\right).$$

Since $A_2 + A_3 \ge 0.5$ (as $A_1 + A_2 + A_3 = 1$ and $A_1 = A_3$), the above can be written as

$$\gamma \pm O\left(\frac{\sqrt{\log n}}{\sqrt{n}}\right).$$

Since $\gamma \le 2\beta \ln(1/\beta)$, this completes the required bound on $\mathscr{R}(\mathcal{A}_{\text{st}})$.

Bounding $\mathscr{U}(\mathcal{A}_{\text{st}})$. Given positive integers k, m, let $U(k, m)$ be the expectation of the top k values from m values chosen independently from the density f. As β goes to 0, we have shown that the number of selected agents from G_1 goes to $O(\sqrt{\log K})$ with high probability. Therefore, as β goes to 0, $\mathbb{E}[\mathscr{U}(\mathcal{A}_{\text{st}})]$ is at most

$$\frac{U(K, K) + U(O(\sqrt{\log K}, K)}{U(K, 2K)},$$

where the denominator gives the optimal utility of the top-K candidates from the set of $2K$ available candidates, and the numerator gives an upper bound on the utilities of top $O(\sqrt{K \log K})$ agents from G_2 and the agents from G_1. This completes the desired bound on $\mathscr{U}(\mathcal{A}_{\text{st}})$.

Bounding $\mathscr{P}(\mathcal{A}_{\text{st}})$. Finally, we bound $\mathscr{U}(\mathcal{A}_{\text{st}})$. The proof proceeds as in the proof in Theorem C.12 with the following changes: let Δ be defined as above by (16). As in Section C.2.1, we define the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ as follows: $\mathcal{E}_1$ is the event that $G_1(\beta, 1) \in (1 - F(\beta))n_1 \pm 8\sqrt{n \log n}$, $\mathcal{E}_2$ is the event that $G_1(\beta - \Delta, 1) \in (1 - F(\beta - \Delta))n_1 \pm 8\sqrt{n \log n}$ and $\mathcal{E}_3$ is the event that $G_2(\beta - \Delta, 1) \in (1 - F(1 - \Delta/\beta))n_2 \pm 8\sqrt{n \log n}$. The rest of the arguments proceed as in Section C.2.2. This completes the proof of the desired result. □

We now show an example illustrating that the result in Theorem C.28 does not extend to arbitrary distributions.

Example C.30. *Consider a distribution that places 0.5 probability mass at 0 and 1 each (although this is a discrete distribution, the example can be easily extended to continuous distributions as well). Assume $\beta \in (0, 1)$, $|G_1| = |G_2| = K$ and $p = 1$. Among the K candidates from each group G_1, roughly $K/2$ candidates have a utility of 1, and the rest have 0 utility. Similarly, about $K/2$ candidates from G_2 have utility β, and the remaining have utility 0. Thus, among the top K candidates, both groups have roughly $K/2$ candidates each. Thus, the representational fairness remains close to 1 even when β goes to 0. Contrast this with the case when the distribution f is log-concave. As Theorem C.28 shows, in such a setting the representational fairness goes to 0 as β goes to 0.*

C.3. Proof of Theorem 3.2

In this section, we prove Theorem 3.2. For ease of readability, we restate Theorem 3.2 below.

Theorem 3.2. *Let $\eta_1, \eta_2, \eta_3 > 0$ be parameters such that $|G_j| \ge \eta_1 n$ for each $j \in \{1, 2\}$, $K \ge \eta_2 n$, and $k_\ell \ge \eta_3 K$ for each $\ell \in [p]$. There is an allocation algorithm $\mathcal{A}_{\text{inst-wise}}$ (stated formally in Algorithm 3) such that, for any distribution of utilities and preference lists, and bias parameter β, $\mathscr{P}(\mathcal{A}_{\text{inst-wise}}) \ge 1 - O\left(\frac{p\sqrt{\log K}}{\eta_1 \eta_3 \sqrt{K}}\right)$, $\mathscr{U}(\mathcal{A}_{\text{inst-wise}}) \ge 1 - O\left(\frac{\sqrt{\log n}}{\sqrt{\eta_2 n}}\right)$ and $\mathscr{R}(\mathcal{A}_{\text{inst-wise}}) = 1$.*

The algorithm $\mathcal{A}_{\text{inst-wise}}$ proceeds by running $\mathcal{A}_{\text{st}}$ on two independent instances, $\mathcal{I}_j, j \in \{1, 2\}$. In $\mathcal{I}_j$, there are n_j candidates belonging to group G_j and institution $\ell \in [p]$ has $k_{\ell,j} := k_\ell \cdot n_j/n$ seats. Let M_j be the assignment obtained by running $\mathcal{A}_{\text{st}}$ on $\mathcal{I}_j$; and M denote the resulting assignment for $G_1 \cup G_2$, i.e., all the candidates in $[n]$. The proof is divided

into three steps. In the first step, we prove the representational fairness guarantee of M, i.e., $\mathscr{R}(\mathcal{A}_{\text{inst-wise}}) = 1$, and discuss how it implies the bound on the utility ratio $\mathscr{U}(\mathcal{A}_{\text{inst-wise}})$ in the second step. In the third step, we analyze $\mathcal{A}_{\text{inst-wise}}$ to conclude the proof. The bounds on representational fairness and utility follow from the institution-wise constraints. The bulk of the technical non-triviality in showing the desired bound on $\mathscr{U}(\mathcal{A}_{\text{inst-wise}})$ lies in proving Lemma 3.5.

C.3.1. Step 1: Bounding $\mathscr{R}$

Let S be the set of candidates matched in M, i.e., $S := M^{-1}([p])$. Let $S_j := S \cap G_j$ be the candidates in G_j that are assigned by M_j. Let K_j denote $K \cdot n_j/n$. Observe that K_j is the total available capacity in the instance $\mathcal{I}_j$. Since $n \geq K$, $n_j \geq K_j$ and hence $|S_j| = K_j$. Recall that q_j denotes the fraction of the candidates from G_j that are assigned by M. Therefore,

$$\frac{q_1}{q_2} = \frac{K_1/n_1}{K_2/n_2} = \frac{K/n}{K/n} = 1.$$

Similarly, $\frac{q_2}{q_1} = 1$. This shows that $\mathscr{R}(\mathcal{A}_{\text{inst-wise}}) = 1$.

C.3.2. Step 2: Bounding $\mathscr{U}$

We now bound $\mathscr{U}(\mathcal{A}_{\text{inst-wise}})$. A random input for the instance can be generated as follows: (i) sample utilities and preferences for the n candidates independently from the distribution $\mathcal{D}$ and $\mathcal{L}$ respectively; (ii) select a subset of n_1 candidates uniformly at random from the set of n candidates, and assign them to G_1, and assign the remaining candidates to G_2.

Now, we condition on the utility values of the n candidates obtained in step (i) above. We show that conditioned on these values, the expected utility ratio is close to 1, and hence, $\mathscr{U}(\mathcal{A}_{\text{inst-wise}})$ is close to 1 after removing this conditioning as well.

Let S be the set of candidates with utility among the top K values. We use Theorem C.1 for negatively correlated random variables as follows. For a candidate $i \in S$, let X_i be the indicator random variable that is 1 if and only if i is assigned to group G_1. Then the variables X_i's are negatively correlated, and $\mathbb{E}[X_i] = \frac{n_1}{n}$. Now observe that $|S \cap G_1| = \sum_{i=1}^{K} X_i$, and hence, $\mathbb{E}[|S \cap G_1| = \frac{n_1 K}{n}$. Applying Theorem C.1, we see that

$$\Pr\left[|S \cap G_1| \leq \frac{n_1 K}{n} - 2\sqrt{K \log K}\right] \leq \frac{1}{2K^2}.$$

Similarly,

$$\Pr\left[|S \cap G_2| \leq \frac{n_2 K}{n} - 2\sqrt{K \log K}\right] \leq \frac{1}{2K^2}.$$

The above two inequalities show that with probability at least $1 - \frac{1}{K^2}$, $\mathcal{A}_{\text{inst-wise}}$ selects the top $K - 4\sqrt{K \log K}$ candidates from S. Let U denote the total utility of the candidates in S. Then total utility of the top $K - 4\sqrt{K \log K}$ candidates in S is at least $U\left(1 - \frac{4\sqrt{K \log K}}{K}\right)$. Thus, $\mathbb{E}[sU(\mathcal{A}_{\text{inst-wise}})] \geq \left(1 - \frac{1}{K^2}\right) \cdot \left(1 - \frac{4\sqrt{K \log K}}{K}\right) \geq 1 - O\left(\frac{\sqrt{\log K}}{\sqrt{K}}\right) = 1 - O\left(\frac{\sqrt{\log n}}{\sqrt{\eta_2 n}}\right)$. This gives the desired bound on $\mathscr{U}(\mathcal{A}_{\text{inst-wise}})$.

C.3.3. Step 3: Bounding $\mathscr{P}$

Recall that $\mathcal{I}_j$ denotes the instance corresponding to candidates in the group G_j, and M_j denotes the assignment obtained by running $\mathcal{A}_{\text{st}}$ on G_j. Let $\pi(M_j)$ denote the number of candidates that are assigned the most preferred choice by M_j. Using Lemma C.9, we see for each $j \in \{1, 2\}$:

$$\Pr[|\pi(M_j) - \mathbb{E}[\pi(M_j)]| \geq 8p\sqrt{K \log K}] \leq 1/K^2.$$

Applying Lemma 3.5 with $a = n/n_j$, we get

$$\left|\frac{n_j}{n} \cdot \mathbb{E}[\pi(M)] - \mathbb{E}[\pi(M_j)]\right| = O(p\sqrt{K \log K}),$$

where $\pi(M)$ denotes the expected number of candidates that are assigned their first choice in the assignment M obtained when we run $\mathcal{A}_{\text{st}}$ on the entire input $\mathcal{I}$. Combining the above two inequalities, we see that with probability at least $1 - 1/K^2$,

the following event happens for each $j \in \{1, 2\}$:

$$\pi(M_j) = \frac{n_j}{n} \cdot \mathbb{E}[\pi(M)] \pm O(p\sqrt{K \log K}). \tag{18}$$

Recall that π_j denotes $\pi(M_j)/n_j$. Hence,

$$\frac{\pi_2}{\pi_1} = \frac{\mathbb{E}[\pi(M)] \pm O(np\sqrt{K \log K})/n_2}{\mathbb{E}[\pi(M)] \pm O(np\sqrt{K \log K})/n_1}.$$

Now, $\mathbb{E}[\pi(M)] \geq \min_\ell k_\ell = \eta_3 K$ and $n_1, n_2 \geq \eta_1 n$. Therefore, we can bound the above expression as follows

$$\frac{\pi_2}{\pi_1} \geq 1 - O\left(\frac{p\sqrt{\log K}}{\eta_1 \eta_3 \sqrt{K}}\right).$$

The same bound applies for $\frac{\pi_1}{\pi_2}$. Since the above-mentioned events happen with probability at least $1 - 1/K^2$, we see that

$$\mathbb{E}\left[\frac{\min(\pi_1, \pi_2)}{\max(\pi_1, \pi_2)}\right] \geq (1 - 1/K^2) \cdot \left(1 - O\left(\frac{p\sqrt{\log K}}{\eta_1 \eta_3 \sqrt{K}}\right)\right) = 1 - O\left(\frac{p\sqrt{\log K}}{\eta_1 \eta_3 \sqrt{K}}\right).$$

This proves the desired bound on $\mathscr{P}(\mathcal{A}_{\text{inst-wise}})$, and completes the proof of Theorem 3.2.

D. Pseudocode for Algorithms with Institution-Wise, Group-Wise and No Constraints

In this section, we provide the pseudocodes for the three algorithms we compare in Section 4, as well as extensions for relaxed bounds in Appendix G. These include a special case of the Gale-Shapley algorithm ($\mathcal{A}_{\text{st}}$; Algorithm 1), the algorithm to satisfy group-wise proportional representation constraints ($\mathcal{A}_{\text{group}}$; Algorithm 2), the algorithm to satisfy institution-wise constraints ($\mathcal{A}_{\text{inst-wise}}$; Algorithm 3), and modifications of Algorithm 2 and Algorithm 3 under relaxed bounds. We implement algorithms and run all empirical work in `Python3`. The code implementations of algorithms, fairness metrics, as well as simulation procedures and data, can be found here.[2]

Unconstrained algorithm $\mathcal{A}_{\text{st}}$. This is a special case of the Gale-Shapley algorithm for stable matching (Gale & Shapley, 1962) where all the institutions agree on a common central ranking of the candidates. The algorithm considers the candidates in decreasing order of observed utility. When considering a candidate, it assigns it to the highest preferred institution that still has an available slot.

Algorithm 1 Unconstrained algorithm ($\mathcal{A}_{\text{st}}$)

Input: S (set of candidates); $\boldsymbol{k} = (k_1, k_2, ..., k_p)$ (institutional capacities); $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ (preferences for $i \in S$); $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$ (estimated utilities for $i \in S$)

```
r_unmatched ← candidates in S sorted in order of decreasing utility
M ← assignment filled with -1 (no institution)
for candidate a in r_unmatched do
    M_a ← top choice institution n such that k_n > 0
    k_n ← k_n − 1
    until k_n = 0 for all n ∈ k
end for
return M
```

Algorithm with group-wise constraints. The algorithm first selects the top (according to observed utilities) $\frac{|G_j|}{n} \cdot K$ candidates from each group G_j. It then runs the algorithm $\mathcal{A}_{\text{st}}$ to assign these K candidates to institutions. This guarantees representation-based fairness for both groups.

Algorithm with institution-wise constraints. This algorithm creates two independent instances, one for each group. In the instance for group G_j, the capacity at an institution i is set to $\frac{|G_j|}{n} \cdot k_i$. Then $\mathcal{A}_{\text{st}}$ is independently executed on each of these two instances.

[2] https://github.com/sandrewxu/CentralizedSelectionwithPreferenceBias

Algorithm 2 Group-wise algorithm ($\mathcal{A}_{\text{group}}$)

Input: G_1 (set of candidates in group 1); G_2 (set of candidates in group 2); $\boldsymbol{k} = (k_1, k_2, \ldots, k_p)$ (institutional capacities); $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ (preferences for each $i \in S$); $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$ (estimated utilities for each $i \in S$)

$m_1 \leftarrow (\sum_\ell k_\ell) \cdot |G_1|/n$
$S_1 \leftarrow$ first m_1 candidates in G_1 sorted in order of decreasing utility
$m_2 \leftarrow (\sum_\ell k_\ell) \cdot |G_2|/n$
$S_2 \leftarrow$ first m_2 candidates in G_2 sorted in order of decreasing utility
$S_0 \leftarrow S_1 \cup S_2$
return $\mathcal{A}_{\text{st}}(S_0, \boldsymbol{k}, \sigma, \hat{u})$

Algorithm 3 Institution-wise algorithm ($\mathcal{A}_{\text{inst-wise}}$)

Input: G_1 (set of candidates in group 1); G_2 (set of candidates in group 2); $\boldsymbol{k} = (k_1, k_2, \ldots, k_p)$ (institutional capacities); $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ (preferences for each $i \in S$); $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$ (estimated utilities for each $i \in S$)

$A \leftarrow \{k_i \cdot |G_1|/n : i \in [p]\}$ */* Define capacity vector A for the p institutions for candidates in G_1 */*
$B \leftarrow \{k_i \cdot |G_2|/n : i \in [p]\}$ */* Define capacity vector B for the p institutions for candidates in G_2*/*
$M_1 \leftarrow \mathcal{A}_{\text{st}}(G_1, A, \{\sigma_i : i \in A\}, \{\hat{u}_i : i \in G_1\})$
$M_2 \leftarrow \mathcal{A}_{\text{st}}(G_2, B, \{\sigma_i : i \in B\}, \{\hat{u}_i : i \in G_2\})$
return $M_1 \cup M_2$

Algorithm with relaxed group-wise constraints. This algorithm satisfies relaxed group-wise constraints (Appendix G) specified by a parameter α, where $0 \leq \alpha \leq 1$. For each group G_j, it first selects the top $\alpha K \cdot \frac{|G_j|}{n}$ candidates. From the remaining pool of $n - \alpha K$ candidates, it selects the top $K - \alpha K$ candidates. Finally, the algorithm $\mathcal{A}_{\text{st}}$ is executed on these K selected candidates. Observe that when $\alpha = 0$, this algorithm is same as the unconstrained algorithm $\mathcal{A}_{\text{st}}$; and when $\alpha = 1$, we recover the algorithm $\mathcal{A}_{\text{group}}$.

Algorithm 4 Relaxed group-wise constraints algorithm

Input: G_1 (set of candidates in group 1); G_2 (set of candidates in group 2); $\boldsymbol{k} = (k_1, k_2, \ldots, k_p)$ (institutional capacities); $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ (preferences for each $i \in S$); $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$ (estimated utilities for each $i \in S$); α (proportion of strictness)

$m_1 \leftarrow \alpha \cdot \sum k_\ell \cdot |G_1|/n$ */* Minimum representation for candidates in G_1 is proportional to $|G_1|$ and α */*
$m_2 \leftarrow \alpha \cdot \sum k_\ell \cdot |G_2|/n$ */* Minimum representation for candidates in G_2 is proportional to $|G_2|$ and α */*
$m_{excess} \leftarrow \sum k_\ell - m_1 - m_2$ */* Capacities not allocated to a specific group when $\alpha < 1$ */*
$r_{unmatched} \leftarrow$ candidates in S sorted in order of decreasing utility
$M \leftarrow$ assignment filled with -1 (no institution)
for candidate a in $r_{unmatched}$ **do**
  $M_a \leftarrow$ top choice institution n such that $k_n > 0$ and $m_{group} > 0$ or $m_{excess} > 0$
  $k_n \leftarrow k_n - 1$
  if $m_{group} > 0$ **then**
    $m_{group} \leftarrow m_{group} - 1$ */* m_{group} is m_1 if the candidate is in G_1 and m_2 if the candidate is in G_2 */*
  else
    $m_{excess} \leftarrow m_{excess} - 1$
  end if
  until $k_n = 0$ for all $n \in k$
end for
return M

Algorithm with relaxed institution-wise constraints. This algorithm satisfies relaxed institution-wise constraints (Appendix G) specified by a parameter α, where $0 \leq \alpha \leq 1$. For each institution i and group G_j, it reserves $\alpha k_i \cdot |G_j|/n$ slots for candidates from group G_j, and the remaining $(1 - \alpha)k_i$ "excess" slots are available for candidates from either group. The algorithm $\mathcal{A}_{\text{st}}$ is now modified as follows: it again considers candidates in decreasing order of observed utility values. When considering a candidate a, say from group G_j, it assigns it to the highest ranked institution i that has an available slot

in either the ones reserved for G_j or in the unreserved pool. Observe that if $\alpha = 1$, we recover $\mathcal{A}_{\text{inst-wise}}$, and when $\alpha = 1$, we execute the unconstrained algorithm $\mathcal{A}_{\text{st}}$.

Algorithm 5 Relaxed institution-wide constraints algorithm

Input: G_1 (set of candidates in group 1); G_2 (set of candidates in group 2); $\boldsymbol{k} = (k_1, k_2, \ldots, k_p)$ (institutional capacities); $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ (preferences for each $i \in S$); $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$ (estimated utilities for each $i \in S$); α (proportion of strictness)

$k_{G_1} \leftarrow \{k_i \cdot \alpha \cdot |G_1|/n : i \in [p]\}$ */∗ Minimum institutional representation for candidates in G_1, proportional to $|G_1|$ and α ∗/*
$k_{G_2} \leftarrow \{k_i \cdot \alpha \cdot |G_2|/n : i \in [p]\}$ */∗ Minimum institutional representation for candidates in G_2, proportional to $|G_2|$ and α ∗/*
$k_{excess} \leftarrow \{k_i \cdot (1 - \alpha \cdot (|G_1| + |G_2|)/n) : i \in [p]\}$ */∗ Capacities not allocated to a specific group when $\alpha < 1$ ∗/*
$r_{unmatched} \leftarrow$ candidates in S sorted in order of decreasing utility
$M \leftarrow$ assignment filled with -1 (no institution)
for candidate a in $r_{unmatched}$ **do**
  $M_a \leftarrow$ top choice institution n such that $k_{group}[n] > 0$, or $k_{excess}[n] > 0$
  if $k_{group}[n] > 0$ **then**
    $k_{group}[n] \leftarrow k_{group}[n] - 1$ */∗ k_{group} is k_{G_1} if the candidate is in G_1 and k_{G_2} if the candidate is in G_2 ∗/*
  else
    $k_{excess}[n] \leftarrow k_{excess}[n] - 1$
  end if
  until $k_n = 0$ for all $n \in k$
end for
return M

E. Additional Discussion and Figures for the Real-World Simulation With JEE 2009 Scores

E.1. Additional figures and implementation details

Distribution of scores. The 23 IITs are among the most prestigious engineering schools in India. Each year, admissions into the IITs are decided through a centralized matching system, where applicants are ranked based on their scores in the Joint Entrance Exam (JEE). Their score is their estimated utility, which is then matched with preference across different major-institute pairs (Baswana et al., 2019). In 2009, the JEE test scores were released in response to a Right to Information application (Kumar, 2009). The preferences of candidates were not released. The data contains the test scores of 384,977 students who took the IIT-JEE 2009. In addition to scores, the dataset has each student's self-reported binary gender and birth category. Birth category is officially assigned as a socioeconomic label, and the most privileged group is the general category (Baswana et al., 2015; Sowell, 2008). A previous analysis of the dataset shows the discrepancies between student birth category or gender as well as score distributions (Celis et al., 2023). Under the β-bias model, this would equate to a bias parameter of $\beta = 0.69$ when the protected attribute is gender and $\beta = 0.52$ when the protected attribute is the birth category.

Utilities. The notion of latent utility is not available through our real-world data, so we consider JEE scores as estimated utility. This is because if we consider candidates of equal "latent utility", the ones from disadvantaged groups perform poorer on standardized tests, our measure of estimated utility (Elsesser, 2019). In India, this could be because fewer girls attend primary school (Alderman & King, 1998; CensusInfo India 2011: Final Population Totals), and many girls are forced to drop out of school to get married or help with housework (Williams, 2013). Because of these societal differences, we expect a female student to perform worse on the JEE than a male student of comparable "latent utility". Students in birth categories other than the general category face similar disadvantages (Kumar, 2021).

Preferences. The dataset does not include individual preferences, so we aggregated information from the opening and closing ranks of major-institution pairs, (M, I). We ordered pairs by their closing ranks $C_{(M,I)}$ in the general category and considered the range $[O_{(M,I)}, C_{(M,I)}]$ of the ranks of students admitted to (M, I) (Table 1). We observed that if $C_{(M_1,I_1)}$ is substantially less than $C_{(M_2,I_2)}$, then most applicants will strictly prefer the earlier ranked program, which is consistent with those of journalists (Livemint, 2022; Verma, 2022). We model preferences over programs using the Mallows distribution (specified by d_{KT}) defined by a dispersion parameter $0 \leq \phi \leq 1$ and the central ranking ρ of program-institute pairs by their closing ranks $C_{(M,I)}$. The capacities of each program were publicly available online. We considered the closing All India

Major-Institution Pair	Opening Rank	Closing Rank
CSE (4Yr), IIT Bombay	3	**86**
EE (4Yr), IIT Bombay	8	**109**
CSE (4Yr), IIT Delhi	1	**154**
CSE (4Yr), IIT Kanpur	2	**181**
CSE (4Yr), IIT Madras	5	**215**
EE (4Yr), IIT Delhi	108	**241**
EE w/ Micro (5Yr), IIT Bombay	117	**245**
CSE (5Yr), IIT Delhi	187	**278**
EE (4Yr), IIT Madras	32	**310**
EE w/ Info. & Comm. Tech. (5Yr), IIT Delhi	284	**369**
EE w/ Comm. & SP (5Yr), IIT Bombay	266	**379**
EE (4Yr), IIT Kanpur	39	**416**
CSE (5Yr), IIT Kanpur	216	**422**
ME (4Yr), IIT Bombay	72	**494**
CSE (5Yr), IIT Madras	333	**502**
CSE (4Yr), IIT Kharag.	276	**527**
EE (5Yr), IIT Kanpur	423	**608**
ME (4Yr), IIT Delhi	237	**634**
ME w/ CAD & Auto. (5Yr), IIT Bombay	419	**637**
EE w/ Micro & VLSI (5Yr), IIT Madras	339	**716**
ME w/ CIM (5Yr), IIT Bombay	556	**757**
EE w/ Power (4Yr), IIT Delhi	477	**758**
EEC (4Yr), IIT Kharag.	133	**762**
EE w/ Comm. Eng. (5Yr), IIT Madras	458	**764**
Math. & Comp. (5Yr), IIT Delhi	348	**789**
ME (4Yr), IIT Kanpur	497	**806**
ME (4Yr), IIT Madras	275	**820**
CSE (5Yr), IIT Kharag.	431	**877**
EE (4Yr), IIT Kharag.	596	**920**
Chem. Eng. (4Yr), IIT Bombay	244	**928**
EE w/ Power Sys. & Elec. (5Yr), IIT Madras	773	**937**
CSE (4Yr), IIT Roorkee	471	**984**
ME (5Yr), IIT Kanpur	808	**992**

Table 1: Top 33 major-institution pairs sorted by closing ranks in the 2009 IIT-JEE. See Appendix E for a discussion of the dataset.

Rank for the general category as opposed to across all categories. This is because the JOSSA has already instituted quotas for students from more disadvantaged groups to be admitted into certain programs. As such, a difference in individual preferences in the underprivileged groups would affect the overall closing ranks significantly and the general ranks give a better picture.

Algorithm implementation. While implementing our algorithm, we implemented limits on the number of candidates that were considered for the sake of lowering runtime. When considering all 384,977 candidates, there was a fixed bound: the candidate in the disadvantaged group that had the lowest JEE score possible to still be matched under group- or institution-wise constraints but not under the unconstrained algorithm. Any candidate, regardless of group, with a lower JEE score would never be considered to be matched. For the sake of runtime, we found the lowest AIR applicable with both birth and gender as the protected attribute and stopped generating preferences after that rank (5000 and 7000, respectively).

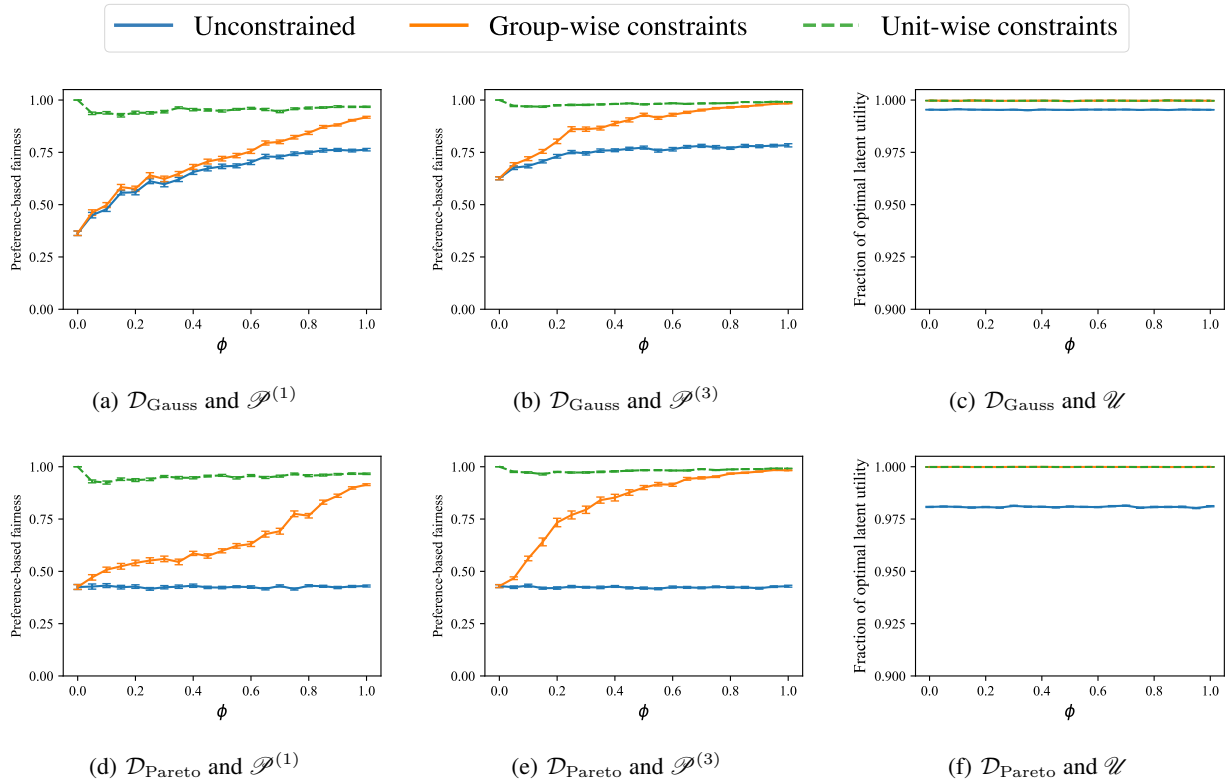
E.2. Simulation that varies the dispersion parameter of the preference distribution using synthetic data

This simulation considers the effect of dispersion of preferences on the three algorithms we present by altering the dispersion parameter, ϕ, of the Mallows distribution. A small ϕ indicates that preferences are more closely aligned to the central ranking,

and $\phi \to 1$ indicates that preferences are drawn more randomly than based on the central ranking. Since Theorem 3.2 does not depend on the shape of the preference distribution, we expect the preference-based fairness of $\mathcal{A}_{\text{inst-wise}}$ to be near optimal regardless of ϕ, but we would like to see the effect of ϕ on $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$.

Setup. We fix $n = 1000$, $p = 5$, and $k_i = 100$ for $i \in [p]$, For $\mathcal{D} \in \{\mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$, we set $\beta = \frac{3}{4}$ and vary $\phi \in [0, 1]$. We report the average and standard error of $\mathscr{P}^{(1)}$, $\mathscr{P}^{(3)}$, and $\mathscr{U}$ over 50 iterations.

Results and observations. The results are reported in Figure 4. Our main observations are that ϕ does not seem to have a large impact on preference-based fairness for $\mathcal{A}_{\text{inst-wise}}$, $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$ generally increase with ϕ, and ϕ does not affect the utility ratio. As expected by Theorem 3.2, $\mathcal{A}_{\text{inst-wise}}$ holds a high preference-based fairness ($\mathscr{P}^{(1)} \geq 0.9$, $\mathscr{P}^{(3)} \geq 0.95$). As ϕ increases, preference-based fairness for $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$ generally increase. The exception is with $\mathcal{A}_{\text{st}}$ when utilities are sampled from $\mathcal{D}_{\text{Pareto}}$, where the preference-based fairness of the unconstrained algorithm does not change with ϕ. To closely model real-world scenarios, we choose for $\phi \leq 0.5$ because the central ranking is often important. We also choose $\phi \geq 0.1$ to be able to differentiate between $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$. When utilities are drawn from $\mathcal{D}_{\text{Gauss}}$, the difference between $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}})$ and $\mathscr{P}^{(1)}(\mathcal{A}_{\text{st}})$ is within 0.05 when $\phi \leq 0.10$. Based on these constraints, we set $\phi = 0.25$ as the default in our simulations. The utility ratio does not change with ϕ because it does not depend on preference distribution. This is further discussed in Appendix F. Additionally, we observe that when varying ϕ between 0 and 1, the group-wise algorithm increases at a much earlier ϕ in synthetic data (Figure 4) compared to real-world data (Table 1). This is likely because the real-world dataset had 33 institutions compared to the 5 institutions used in synthetic simulations. With that many institutions, it would be much more difficult to increase the preference-based fairness, especially measured by the $\mathscr{P}^{(1)}$ metric.



(a) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{P}^{(1)}$ (b) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{P}^{(3)}$ (c) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{U}$

(d) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(1)}$ (e) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(3)}$ (f) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{U}$

Figure 4: $\mathscr{P}^{(1)}$, $\mathscr{P}^{(3)}$, and $\mathscr{U}$ measured for synthetic data when the dispersion parameter for the Mallows distribution is varied over $\phi \in [0, 1]$. In (a), we see the preference-based fairness measured by $\mathscr{P}^{(1)}$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. In (b), we measure $\mathscr{P}^{(3)}$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (c) shows the utility ratio when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (d), (e), and (f) show the results of (a), (b), and (c), respectively, when utilities are generated from $\mathcal{D}_{\text{Pareto}}$. Our main observation is that ϕ does not have a large impact on preference-based fairness for $\mathcal{A}_{\text{inst-wise}}$, while $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$ generally increase with ϕ. See Appendix E.2 for details and discussion. The x-axis denotes α, the y-axis denotes $\mathscr{P}^{(1)}$, $\mathscr{P}^{(3)}$, or $\mathscr{U}$, and the error bars denote the standard error of the mean over 50 iterations.

F. Additional Discussion and Plots for Simulation with Synthetic Data from Section 4

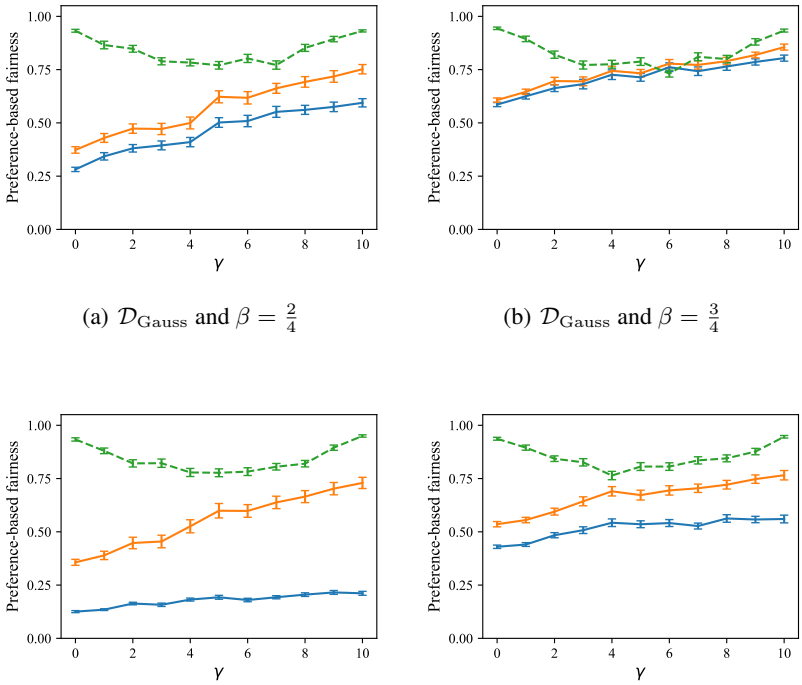
In this section, we report the preference-based fairness, $\mathscr{P}^{(1)}$, for additional values of β in the simulation from Section 4 (Figure 5). We also discuss the utility ratio in the simulation.

F.1. Discussion of utility ratio

We do not include results about utility ratio for synthetic data because they do not vary with respect to preferences for any of our algorithms. Specifically, $\mathcal{A}_{\text{st}}$ always selects the K candidates with the highest observed utilities, where K is the total capacity of all institutions. Likewise, under the assumption that $|G_1| = |G_2|$, $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{inst-wise}}$ both assign the top $K/2$ candidates in terms of utility from each group to institutions. This guarantees that $\mathscr{U}(\mathcal{A}_{\text{group}}) = \mathscr{U}(\mathcal{A}_{\text{inst-wise}}) \approx 1$ for all γ. Intuitively, $\mathscr{U}(\mathcal{A}_{\text{st}})$ increases with respect to β: this is detailed in Theorem 3.1 and visualized in Figure 11(a).

F.2. Additional plots

Results and observations. The results of $\mathscr{P}^{(1)}$ for additional values of β support the discussion made in Section 4. One additional observation is that as β increases, $\mathscr{P}^{(1)}(\mathcal{A}_{\text{inst-wise}})$ is not affected but $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}})$ and $\mathscr{P}^{(1)}(\mathcal{A}_{\text{st}})$ increase at comparable levels of β. This is because the estimated utilities of the disadvantaged candidates increase as β increases.



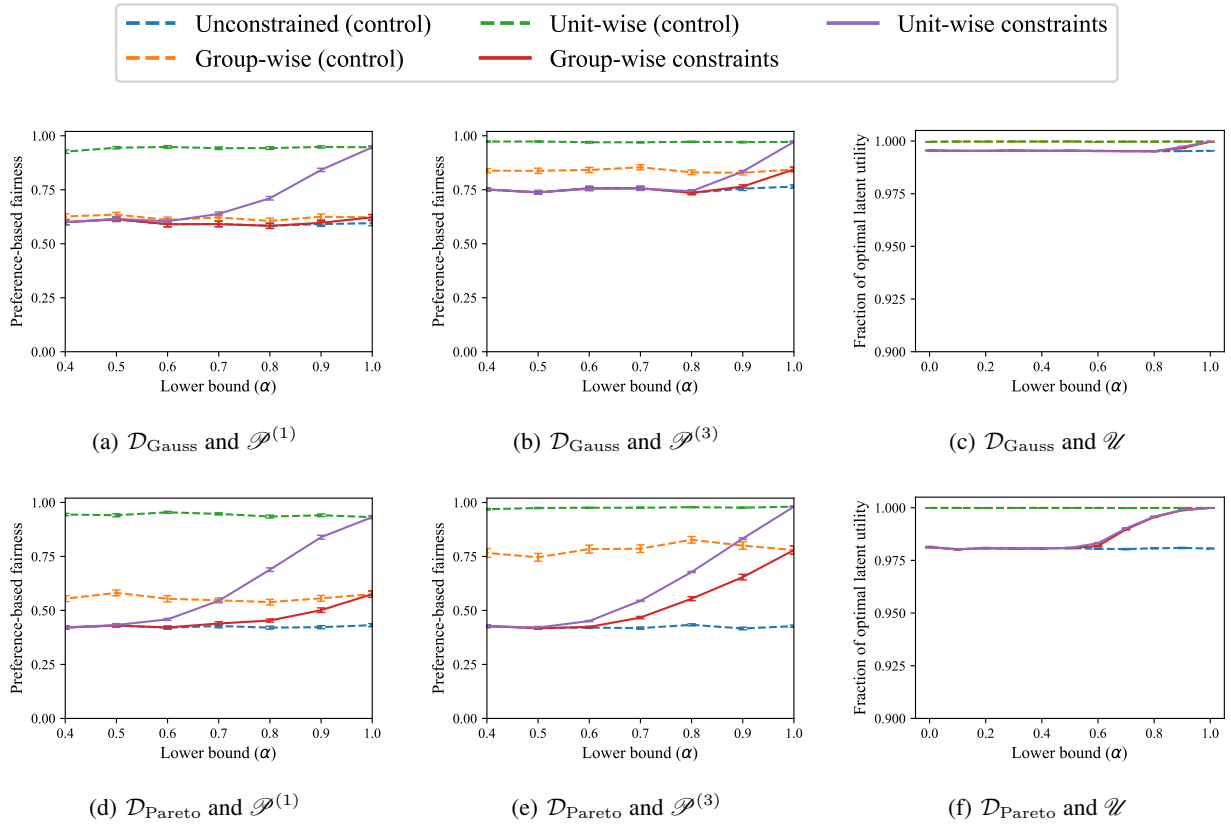
(a) $\mathcal{D}_{\text{Gauss}}$ and $\beta = \frac{2}{4}$

(b) $\mathcal{D}_{\text{Gauss}}$ and $\beta = \frac{3}{4}$

(c) $\mathcal{D}_{\text{Pareto}}$ and $\beta = \frac{2}{4}$

(d) $\mathcal{D}_{\text{Pareto}}$ and $\beta = \frac{3}{4}$

Figure 5: Preference-based fairness as measured by the top-1 metric, $\mathscr{P}^{(1)}$, with synthetic data where non-i.i.d. preferences are generated from Mallows distributions. The x-axis denotes γ, the Kendall-Tau distance between the central rankings, and the error bars denote the standard error of the mean over 50 iterations. (a) shows $\mathscr{P}^{(1)}$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$ with $\beta = \frac{2}{4}$. (b) modifies β to $\beta = \frac{3}{4}$. (c) and (d) are equivalent to (a) and (b), respectively, when utilities are drawn from $\mathcal{D}_{\text{Pareto}}$. See Section 4 and Appendix F for details and discussion. We observe that institution-wise constraints achieve higher preference-based fairness than group-wise and unconstrained settings.



(a) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{P}^{(1)}$ (b) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{P}^{(3)}$ (c) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{U}$

(d) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(1)}$ (e) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(3)}$ (f) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{U}$

Figure 6: Preference-based fairness measured by $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(3)}$, and the utility ratio measured by $\mathscr{U}$ for relaxed institution-wise and group-wise bounds parameterized by strictness ratio $0 \leq \alpha \leq 1$. We compare results with relaxed bounds (solid lines) and results with strict bounds (dotted lines). In (a), we measure $\mathscr{P}^{(1)}$ when utilities are drawn from $\mathcal{D}_{\text{Gauss}}$ In (b), we measure $\mathscr{P}^{(3)}$ with utility distribution $\mathcal{D}_{\text{Gauss}}$. In (c), we measure the utility ratio for $0 \leq \alpha \leq 1$ with utilities from $\mathcal{D}_{\text{Gauss}}$. (d), (e), and (f) are equivalent to (a), (b), and (c), respectively, except that utilities are drawn from $\mathcal{D}_{\text{Pareto}}$ instead of $\mathcal{D}_{\text{Gauss}}$. Our main observation is that institution-wise constraints and group-wise constraints only begin to increase preference-based fairness compared to no constraints when the lower bound is high. See Appendix G for details and discussion. The x-axis denotes α, and the error bars denote the standard error of the mean over 50 iterations.

G. Empirically Studying Institution-Wise Constraints Under Relaxed Bounds

This section considers the effectiveness of $\mathcal{A}_{\text{inst-wise}}$ under relaxed bounds. Specifically, instead of instituting a strict proportional quota for each institution, we consider the effect of a relaxed bound parameterized by $0 \leq \alpha \leq 1$. If $\alpha = 1$, the bound is strict and if $\alpha = 0$, there is no bound. In general, the quotas required by $\mathcal{A}_{\text{inst-wise}}$ are α times the available capacity at each institution, with the remaining capacity available for candidates from either group. This is similar when considering relaxed group-wise constraints, except on an institution-wise basis rather than on a representational basis. Pseudocode for the algorithms are provided in Appendix D. The purpose of this simulation is to analyze the practicality of more relaxed real-world constraints beyond the theoretical work done in Theorem 3.2. To do this, we measure the preference-based fairness and utility ratio of relaxed constraints compared to stringent ones (as required by $\mathcal{A}_{\text{inst-wise}}$).

Setup. We fix $n = 1000$, $p = 5$, and $k_i = 100$ for $i \in [p]$. For $\mathcal{D} \in \{\mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$, we set $\beta = \frac{3}{4}$ because β is often high in real-world situations. We generate preferences from a single central ranking using a Mallows distribution with $\phi = 0.25$ ($\gamma = 0$). We vary the relaxed bound parameter, α, from 0 (no bounds) to 1 (strict bounds) with two groups of the same size, i.e., $|G_1| = |G_2|$. We report the average and standard error of $\mathscr{P}^{(1)}$, $\mathscr{P}^{(3)}$, and $\mathscr{U}$ over 50 iterations.

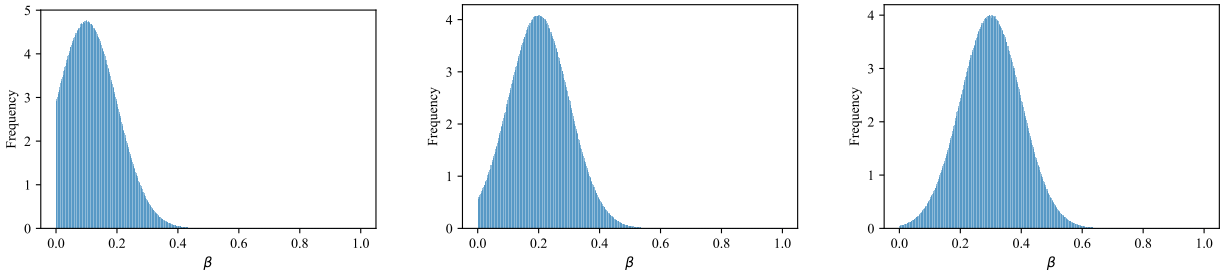
Results and observations. The results are presented in Figure 6. We know that the bounds instituted by $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{A}_{\text{group}}$ will not affect the preference-based fairness until the bound exceeds the proportion that $\mathcal{A}_{\text{st}}$ achieves. Based on

the empirical results for $\beta = \frac{3}{4}$, we started the bounds at $\alpha = 0.4$ to better illustrate the changes. Our main observation is that preference-based fairness begins to increase for $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{A}_{\text{group}}$ as $\alpha > 0.5$, but the utility ratio does not increase until between $\alpha = 0.6$ ($\mathcal{D}_{\text{Pareto}}$) or $\alpha = 0.8$ ($\mathcal{D}_{\text{Gauss}}$). Under a Pareto distribution, $\mathscr{P}^{(1)}$ for $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{A}_{\text{group}}$ begin to increase at $\alpha = 0.5$ from 0.45 to 0.90 ($\mathcal{A}_{\text{inst-wise}}$) and 0.55 ($\mathcal{A}_{\text{group}}$) as $\alpha \to 1$. Under a truncated Gaussian distribution, $\mathscr{P}^{(1)}$ for $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{A}_{\text{group}}$ begin to increase at $\alpha = 0.6$ from 0.55 to 0.90 ($\mathcal{A}_{\text{inst-wise}}$) and 0.60 ($\mathcal{A}_{\text{group}}$) as the bound becomes 50%. We see that institution-wise constraints are only effective when strictly enforced (for $\beta = \frac{3}{4}$, $\alpha > 0.5$). For situations in which β is higher, the bound would have to be even higher. We also see that preference-based fairness for the relaxed version of $\mathcal{A}_{\text{inst-wise}}$ begins to increase at a lower α than the corresponding utility ratio does. Lower bounds may not be effective at all, particularly if the unconstrained algorithm already guarantees the low ratio.

H. Empirical Results Considering Noise in Estimated Utilities

H.1. Simulation considering noise in the bias parameter

In this section, we consider noise in the bias parameter of the β-bias model, which we refer to as the noisy β-bias model. Instead of a fixed bias parameter, we sample β from a Gaussian distribution truncated to the interval $[0, 1]$ with a standard deviation of 0.1. Examples of the noisy β-bias model centered around $\beta = 0.1, 0.2, 0.3$ are shown in Figure 7. This standard deviation ensures that for non-extreme values of β ($0.1 \leq \beta \leq 0.9$), the expected value closely adheres to the center of the distribution (within 0.02). Under these non-ideal conditions, we observe how closely empirical data adheres to the theoretical results of Theorem 3.2, as well as the effect of noise on $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$.

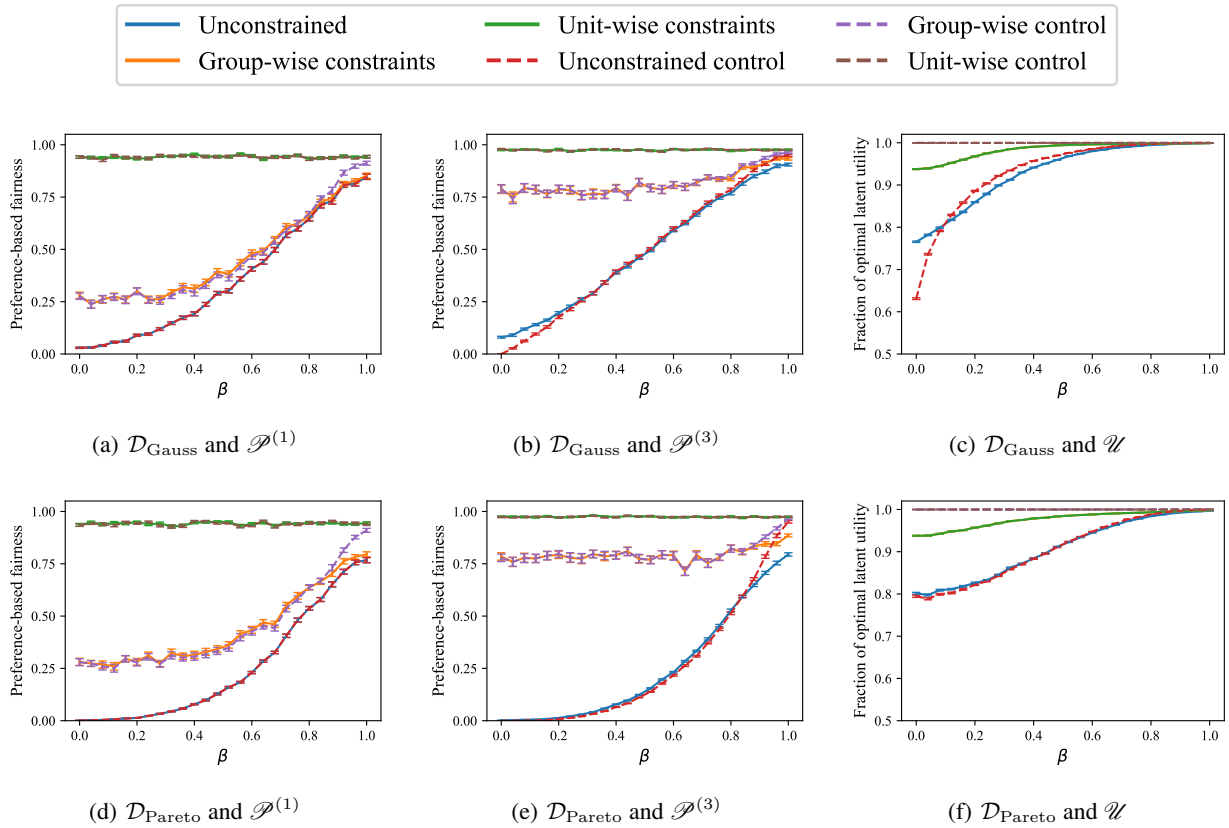


(a) $\beta = 0.1$ (b) $\beta = 0.2$ (c) $\beta = 0.3$

Figure 7: Distributions of β under the noisy β-bias model with standard deviation 0.1. (a) shows the distribution when $\beta = 0.1$ with mean 0.129. (b) shows the distribution when $\beta = 0.2$ with mean 0.205. (c) shows the distribution when $\beta = 0.3$ with mean 0.300. The noisy β-bias model is detailed in Appendix H.1.

Setup. We fix the parameters $n = 1000$, $p = 5$, and $k_\ell = 100$ for each $\ell \in [5]$. We vary $\beta \in [0, 1]$ and sample a bias parameter for each person in the disadvantaged group from the Gaussian distribution centered at β, truncated to the interval $[0, 1]$ with standard deviation 0.1. We draw latent utilities from $\mathcal{D} \in \{\mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$, and draw preferences i.i.d. from a Mallows distribution with $\phi = 0.25$. We plot preference-based fairness ($\mathscr{P}^{(1)}, \mathscr{P}^{(3)}$) and utility ratio ($\mathscr{U}$) over 50 iterations.

Results and observations. Results are plotted in Figure 8. Our main observation is that even with noise in β, the preference-based fairness of $\mathcal{A}_{\text{inst-wise}}$ is not affected (within 0.01) and is still high. When looking at $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{st}}$, noise can skew preference-based fairness for extreme β ($\beta \leq 0.15, \beta \geq 0.85$). Specifically, when β approaches 1, preference-based fairness in the presence of noise is less than preference-based fairness without noise. Under the condition of $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(3)}$, the difference in $\mathscr{P}^{(3)}$ was 0.1 for $\mathcal{A}_{\text{group}}$ and 0.2 for $\mathcal{A}_{\text{st}}$ at $\beta = 1$. Noise also affected the utility ratio. At low β, there was a difference between the utility ratio of the simulations with noise and without noise, but as β approached 1, the utility ratios converged. Significantly, for $\beta \geq 0.6$, the difference between the utility ratio with and without noise was within 0.02. Overall, noise had minimal impact on preference-based fairness for institution-wise constraints, but had a notable impact on group-wise and unconstrained algorithms for $\beta \geq 0.85$ or $\beta \leq 0.15$. The noise did not have a significant effect on the utility ratio for large β.



(a) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{P}^{(1)}$ (b) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{P}^{(3)}$ (c) $\mathcal{D}_{\text{Gauss}}$ and $\mathscr{U}$

(d) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(1)}$ (e) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{P}^{(3)}$ (f) $\mathcal{D}_{\text{Pareto}}$ and $\mathscr{U}$

Figure 8: Preference-based fairness is measured by $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(3)}$, and utility ratio is measured under the noisy β-bias model. The control lines with no noise (dotted) are compared to the experimental lines with noise (solid). (a) measures $\mathscr{P}^{(1)}$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (b) measures $\mathscr{P}^{(3)}$ under the same utility distribution. (c) shows the utility ratio when utilities are sampled from $\mathcal{D}_{\text{Gauss}}$. Note that the utility ratio lines for $\mathcal{A}_{\text{group}}$ and $\mathcal{A}_{\text{inst-wise}}$ overlap completely. (d), (e), and (f) are equivalent to (a), (b), and (c), respectively, except that utilities are drawn from $\mathcal{D}_{\text{Pareto}}$ instead of $\mathcal{D}_{\text{Gauss}}$. 5 Our main observation is that noise decreases the preference-based fairness at high β and decreases the utility ratio of $\mathcal{A}_{\text{inst-wise}}$ at low β. See Appendix H.1 for details and discussion. The x-axis denotes β, the y-axis denotes $\mathscr{P}^{(1)}$, $\mathscr{P}^{(3)}$, or $\mathscr{U}$, and the error bars denote the standard error of the mean over 50 iterations.

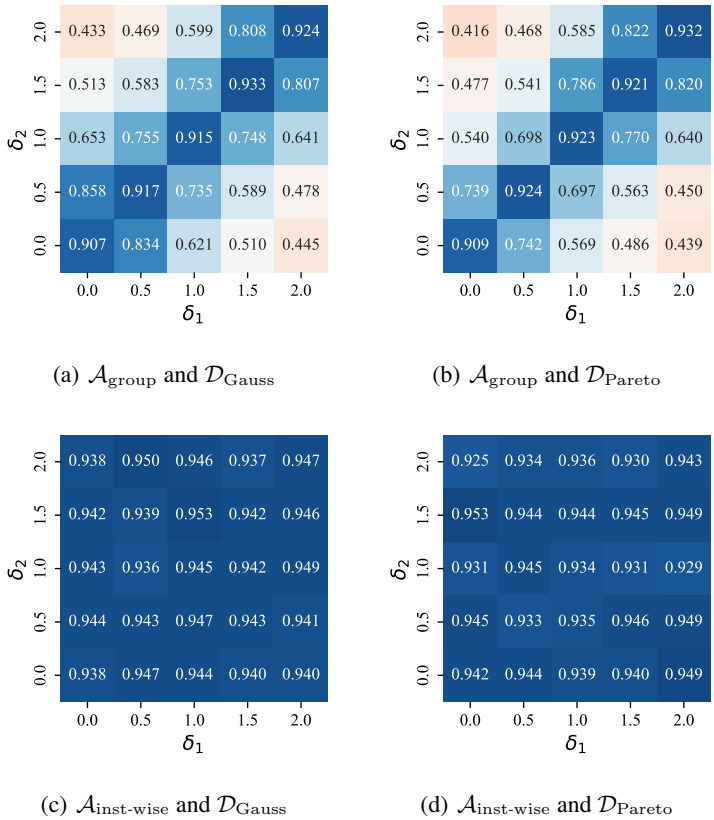
H.2. Simulation with noise in estimated utilities under the implicit variance model

In this section, we consider noise in the estimated utilities under the implicit variance model, which is an instance in which Theorem 3.2 may not hold (Emelianov et al., 2020). The main purpose of this simulation is to test whether or not the results of Theorem 3.2 hold under a different bias model. Since noise minimally changes the distribution as a whole, we expect institution-wise constraints to be relatively unaffected and would like to see how well group-wise constraints hold as a comparison. Given parameters $\delta_1 \geq 0$ and $\delta_2 \geq 0$ controlling the extent of the noise, for each item i, the estimated utility of item i is

$$\widehat{u}_{i,\sigma} := \begin{cases} u_i + \delta_1 \cdot \zeta_i & \text{if } i \in G_1, \\ u_i + \delta_2 \cdot \zeta_i & \text{if } i \in G_2, \end{cases} \quad \text{where} \quad \zeta_i \sim \mathcal{N}(0,1).$$

Setup. We fix the parameters $n = 1000$, $p = 5$, and $k_\ell = 100$ for each $\ell \in [5]$. For each choice of $\mathcal{D} \in \{\mathcal{D}_{\text{Pareto}}, \mathcal{D}_{\text{Gauss}}\}$, we vary both δ_1 and δ_2 over $[0, 2]$ and report the average preference-based fairness (both $\mathscr{P}^{(1)}$ and $\mathscr{P}^{(3)}$) over 50 iterations using both group-wise constraints ($\mathcal{A}_{\text{group}}$) and institution-wise constraints ($\mathcal{A}_{\text{inst-wise}}$). For this simulation, we let $\phi = 0.25$. As such, we are assuming that there is a consistent central ranking with variances in individual preferences.

Results and observations. We report results for $\mathscr{P}^{(1)}$ in Figure 9 and for $\mathscr{P}^{(3)}$ in Figure 10. Our main observation is that as expected, noise had minimal effect on $\mathcal{A}_{\text{inst-wise}}$. We also see that when $\delta_1 = \delta_2$, $\mathcal{A}_{\text{group}}$ does nearly as well as $\mathcal{A}_{\text{inst-wise}}$, but the noise has a much greater effect on $\mathcal{A}_{\text{group}}$. For all values of δ_1 and δ_2 tested, $\mathcal{A}_{\text{inst-wise}}$ retained a high preference-based fairness. $\mathscr{P}^{(1)} \geq 0.93$ and $\mathscr{P}^{(3)} \geq 0.96$ for both $\mathcal{D}_{\text{Gauss}}$ and $\mathcal{D}_{\text{Pareto}}$. $\mathcal{A}_{\text{group}}$ also retained high preference-based fairness when $\delta_1 = \delta_2$ ($\mathscr{P}^{(1)} \geq 0.90$, $\mathscr{P}^{(3)} \geq 0.95$). However, when $\delta_1 \neq \delta_2$, there are cases where $\mathscr{P}^{(1)} \leq 0.50$ and $\mathscr{P}^{(3)} \leq 0.7$. The difference between the preference-based fairness resulting from $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{A}_{\text{group}}$ can approach 0.50 ($\mathscr{P}^{(1)}$) and 0.27 ($\mathscr{P}^{(3)}$) such as when $\delta_1 = 2$ and $\delta_2 = 0$. Institution-wise constraints maintain near-optimal preference-based fairness under noise in estimated utilities, while group-wise constraints only do so when the noise parameters are equal.



(a) $\mathcal{A}_{\text{group}}$ and $\mathcal{D}_{\text{Gauss}}$

(b) $\mathcal{A}_{\text{group}}$ and $\mathcal{D}_{\text{Pareto}}$

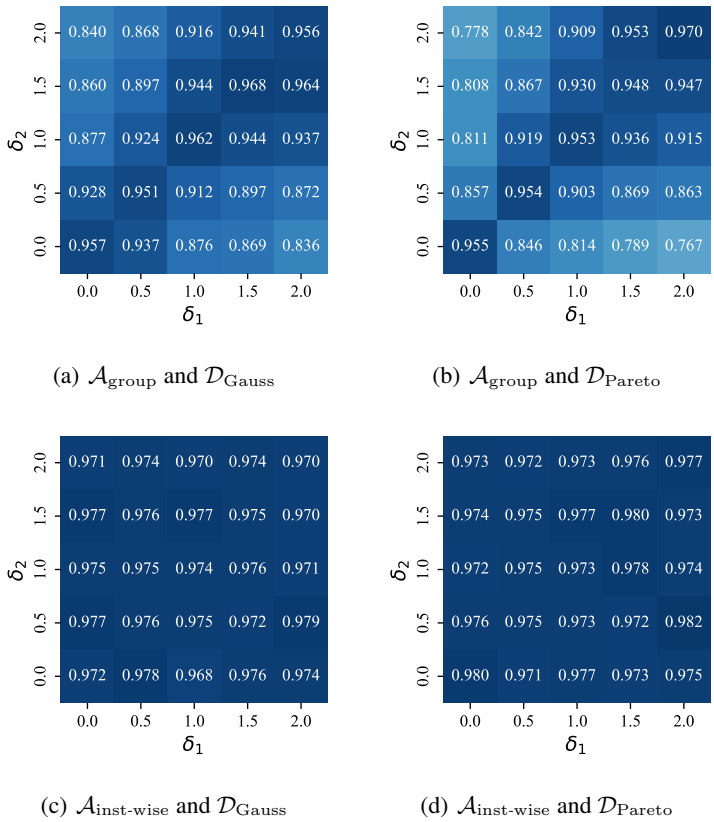
(c) $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{D}_{\text{Gauss}}$

(d) $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{D}_{\text{Pareto}}$

Figure 9: Preference-based fairness as measured by $\mathscr{P}^{(1)}$ under the implicit variance model, where the heatmap scales from 0 (red) to 1 (blue). (a) shows $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}})$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (b) shows $\mathscr{P}^{(1)}(\mathcal{A}_{\text{group}})$ when utilities are generated from $\mathcal{D}_{\text{Pareto}}$. (c) shows $\mathscr{P}^{(1)}(\mathcal{A}_{\text{inst-wise}})$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (d) shows $\mathscr{P}^{(1)}(\mathcal{A}_{\text{inst-wise}})$ when utilities are generated from $\mathcal{D}_{\text{Pareto}}$. Our main observation is that even with noise, institution-wise constraints maintain high preference-based fairness, while group-wise constraints may not. See Appendix H.2 for details and discussion. The x-axis denotes δ_1, the y-axis denotes δ_2, and the values are the average result over 50 iterations.

I. Empirical Validation of Theorem 3.1

In this section, we empirically show that Theorem 3.1 extends to more robust settings than those required by the theorem. In particular, Theorem 3.1 assumes that the latent utilities of candidates are drawn from the uniform distribution on $[0, 1]$ and the β-bias model is used to generate estimated utilities. An extension of the theorem, Theorem C.28, proves a similar result for all log-concave densities, where upper bounds on $\mathscr{R}$ and $\mathscr{U}$ are given. Here, we conduct an empirical study using synthetic data to show that the implications of Theorem 3.1 and Theorem C.28 continue to hold when the utilities are drawn from the truncated Gaussian distribution $\mathcal{D}_{\text{Gauss}}$ and the Pareto distribution $\mathcal{D}_{\text{Pareto}}$ (Appendix I.1); when the bias parameter β is stochastic (Appendix I.2); and when the utilities are biased using the implicit variance model (Appendix I.3).



(a) $\mathcal{A}_{\text{group}}$ and $\mathcal{D}_{\text{Gauss}}$

(b) $\mathcal{A}_{\text{group}}$ and $\mathcal{D}_{\text{Pareto}}$

(c) $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{D}_{\text{Gauss}}$

(d) $\mathcal{A}_{\text{inst-wise}}$ and $\mathcal{D}_{\text{Pareto}}$

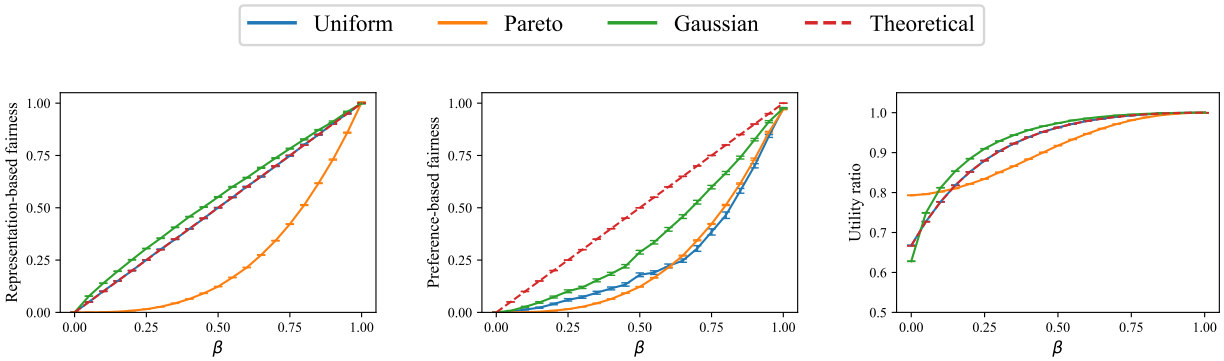
Figure 10: Preference-based fairness as measured by $\mathscr{P}^{(3)}$ under the implicit variance model, where the heatmap scales from 0 (red) to 1 (blue). (a) shows $\mathscr{P}^{(3)}(\mathcal{A}_{\text{group}})$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (b) shows $\mathscr{P}^{(3)}(\mathcal{A}_{\text{group}})$ when utilities are generated from $\mathcal{D}_{\text{Pareto}}$. (c) shows $\mathscr{P}^{(3)}(\mathcal{A}_{\text{inst-wise}})$ when utilities are generated from $\mathcal{D}_{\text{Gauss}}$. (d) shows $\mathscr{P}^{(3)}(\mathcal{A}_{\text{inst-wise}})$ when utilities are generated from $\mathcal{D}_{\text{Pareto}}$. Our main observation is that even with noise, institution-wise constraints maintain high preference-based fairness, while group-wise constraints may not. See Appendix H.2 for details and discussion. The x-axis denotes δ_1, the y-axis denotes δ_2, and the values are the average result over 50 iterations.

I.1. Pareto and truncated Gaussian distributions of utilities

We validate the results of Theorem 3.1 under the uniform distribution. Furthermore, we also validate Theorem C.28, which extends the results of Theorem 3.1 to all log-concave densities, by also testing $\mathcal{D}_{\text{Gauss}}$ and $\mathcal{D}_{\text{Pareto}}$.

Setup. We generate latent utilities u, where $u \sim \mathcal{D} \in \{\mathcal{D}_{\text{Unif}}, \mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$. In setting $n = 10000$, $p = 5$, and $k_\ell = 1000$ for each $\ell \in [5]$, preferences are generated from a Mallows distribution with $\phi = 0.25$ and $\mathcal{A}_{\text{st}}$ is run on the resulting data. We set $n = 10000$ to minimize the error term. We vary β in the range $[0, 1]$. We calculate $\mathscr{R}$, $\mathscr{P}^{(1)}$, and $\mathscr{U}$ for over 50 iterations.

Results and observations. Results are shown in Figure 11. Our main observation is that for the uniform distribution, $\mathscr{R}$ and $\mathscr{U}$ closely follow the theoretical results states in Theorem 3.1. This means the error term is likely small and negligible. In this simulation, $\mathscr{R}$ for the uniform distribution closely models theoretical results (± 0.01) and $\mathscr{U}$ closely models theoretical predictions for all β. Since $\mathscr{P}$ has an upper bound consistent with $\mathscr{R}$, that statement also holds. We calculate $\mathscr{P}^{(1)}$ to visualize the lower bound of $\mathscr{P}$. Under the premise of no noise for $\mathcal{D}_{\text{Gauss}}$ or $\mathcal{D}_{\text{Pareto}}$, we find that the results for $\mathcal{D}_{\text{Gauss}}$ closely follow the results for the uniform distribution, with differences ≤ 0.1 for both representational fairness and utility ratio. The Pareto distribution did not closely match the predicted results and had a lower representational fairness for all β. The Pareto distribution resulted in a higher utility ratio for $\beta \leq 0.15$ and a lower one for higher β. In general, the results for all three distributions match the results of Theorem 3.1 for $\mathscr{R}$ and $\mathscr{U}$, with the occasional exception from $\mathcal{D}_{\text{Pareto}}$.



(a) $\mathscr{R}$ (b) $\mathscr{P}^{(1)}$ (c) $\mathscr{U}$

Figure 11: Representation-based fairness is measured by $\mathscr{R}$, preference-based fairness is measured by $\mathscr{P}^{(1)}$, and utility ratio is measured under the $beta$-bias model with different utility distributions. (a) measures $\mathscr{R}$ when utilities are generated from $\mathcal{D}_{\text{Unif}}$, $\mathcal{D}_{\text{Gauss}}$, and $\mathcal{D}_{\text{Pareto}}$, as well as the theoretical results for $\mathcal{D}_{\text{Unif}}$. (b) measures $\mathscr{P}^{(1)}$, the lower bound for preference-based fairness. (c) measures the utility ratios of the three utility distributions. Our main observation is that for the uniform distribution, $\mathscr{R}$ and $\mathscr{U}$ closely follow the theoretical results stated in Theorem 3.1. See Appendix I for details and discussion. The x-axis denotes β, the y-axis denotes $\mathscr{R}$, $\mathscr{P}^{(1)}$, or $\mathscr{U}$, and the error bars denote the standard error of the mean over 50 iterations.

I.2. Noise in the bias parameter

We further test the robustness of Theorem 3.1 by adding noise to the bias parameter, with the procedure detailed in Appendix H.1.

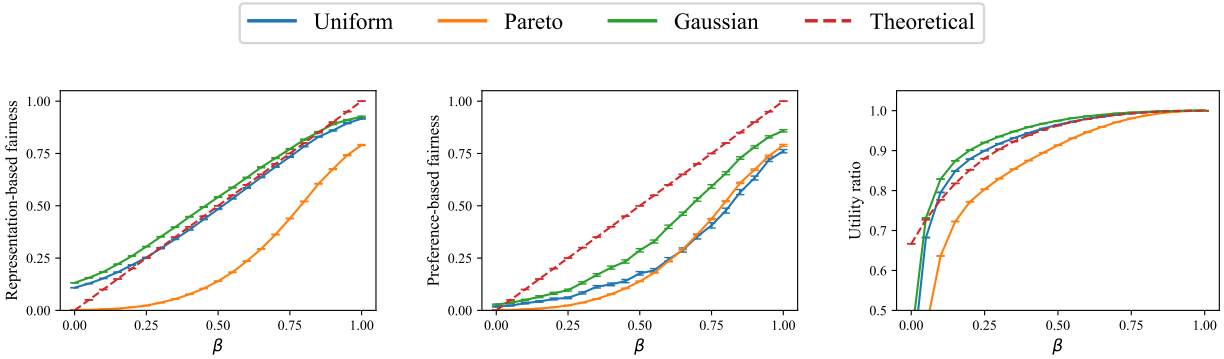
Setup. We generate latent utilities u, where $u \sim \mathcal{D} \in \{\mathcal{D}_{\text{Unif}}, \mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$. In setting $n = 10000$, $p = 5$, and $k_\ell = 1000$ for each $\ell \in [5]$, preferences are generated from a Mallows distribution with $\phi = 0.25$ and $\mathcal{A}_{\text{st}}$ is run on the resulting data. We set $n = 10000$ to minimize the error term. We vary $\beta \in [0, 1]$ with noise, where β is sampled over a Gaussian distribution truncated to $[0, 1]$ with a standard deviation of 0.1 (detailed in Appendix H.1). We calculate $\mathscr{R}$, $\mathscr{P}^{(1)}$, and $\mathscr{U}$ for over 50 iterations.

Results and observations. Results are shown in Figure 12. Our main observation is that even under noise in the bias parameter, for the uniform distribution, $\mathscr{R}$ and $\mathscr{U}$ closely follow the theoretical results stated in Theorem 3.1. The representational fairness and utility ratio of all three distributions resembled the theoretical results, although $\mathcal{D}_{\text{Pareto}}$ deviated the most. We observe that the data is skewed for extreme values of β ($\beta \leq 0.25, \beta \geq 0.75$). When considering the graph measuring $\mathscr{R}$ with noise, we see that $\mathscr{R}$ for the uniform distribution and $\mathcal{D}_{\text{Gauss}}$ is higher than without noise at low beta (At $\beta = 0$, within $0.10 - 0.15$). At $\beta = 1$, $\mathscr{R}$ for all 3 distributions with noise is lower than without noise (within $0.1 - 0.25$). When looking at $\mathscr{P}^{(1)}$, the most noticeable difference is when $\beta \to 1$. At $\beta = 1$, the difference lies between 0.15 ($\mathcal{D}_{\text{Unif}}$) and 0.25 ($\mathcal{D}_{\text{Gauss}}$). When looking at $\mathscr{U}$, there is almost no difference for $\beta \geq 0.25$, but for low β, noise decreases $\mathscr{U}$. Even under noise in the bias parameter, all three distributions of utilities follow the trends indicated in the theoretical results of Theorem 3.1 for $\mathscr{R}$ and $\mathscr{U}$, with occasional exceptions in the case of $\mathcal{D}_{\text{Pareto}}$.

I.3. Implicit variance model

We also test Theorem 3.1 by adding noise to estimated utilities under a simplified version of the implicit variance model detailed in Appendix H.2. While considering the implicit variance model, we fix the variance parameter δ_1 for group G_1 to 0 and only vary the parameter δ_2 for group G_2 between 0 and 2.

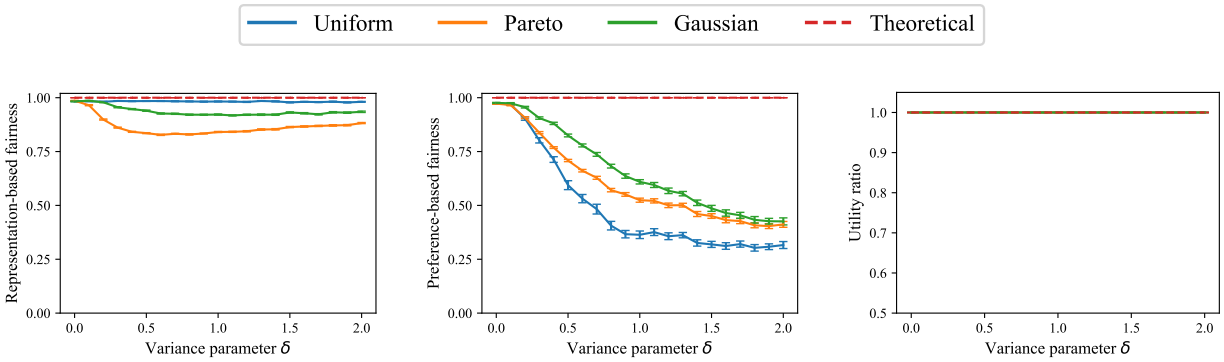
Setup. We generate latent utilities u, where $u \sim \mathcal{D} \in \{\mathcal{D}_{\text{Unif}}, \mathcal{D}_{\text{Gauss}}, \mathcal{D}_{\text{Pareto}}\}$. We add noise to half of the utilities under the implicit variance model. Here $n = 10000$, $p = 5$, and $k_\ell = 1000$ for each $\ell \in [5]$, preferences are generated from a Mallows distribution with $\phi = 0.25$ and $\mathcal{A}_{\text{st}}$ is run on the resulting data. We set $n = 10000$ to minimize the error term. We fix $\beta = 1$ and vary $\delta = \delta_2$, the noise parameter, from 0 to 2. We calculate $\mathscr{R}$, $\mathscr{P}^{(1)}$, and $\mathscr{U}$ for over 50 iterations.



(a) $\mathscr{R}$ with noise in β (b) $\mathscr{P}^{(1)}$ with noise in β (c) $\mathscr{U}$ with noise in β

Figure 12: Representation-based fairness is measured by $\mathscr{R}$, preference-based fairness is measured by $\mathscr{P}^{(1)}$, and utility ratio is measured under the noisy β-bias model with different utility distributions. (a) measures $\mathscr{R}$ when utilities are generated from $\mathcal{D}_{\text{Unif}}$, $\mathcal{D}_{\text{Gauss}}$, and $\mathcal{D}_{\text{Pareto}}$, as well as the theoretical results for $\mathcal{D}_{\text{Unif}}$. (b) measures $\mathscr{P}^{(1)}$, the lower bound for preference-based fairness. (c) measures the utility ratios of the three utility distributions. Our main observation is that noise skews the results for extreme β, but they generally still adhere to Theorem 3.1. See Appendix I for details and discussion. The x-axis denotes β, the y-axis denotes $\mathscr{R}$, $\mathscr{P}^{(1)}$, or $\mathscr{U}$, and the error bars denote the standard error of the mean over 50 iterations.

Results and observations. Results are shown in Figure 13. Our main observation is that the representational fairness and utility ratio of all three distributions resembled the theoretical results, while preference-based fairness for all three distributions decreased as δ increased. In this simulation, $\mathscr{R}$ for the uniform distribution closely models theoretical results (± 0.05) and $\mathscr{U}$ almost exactly models theoretical predictions for all β. Since $\mathscr{P}$ has an upper bound consistent with $\mathscr{R}$, that statement also holds. We calculate $\mathscr{P}^{(1)}$ to visualize the lower bound of $\mathscr{P}$. Under the implicit variance model, where noise was added to the estimated utilities, we see that it does not affect $\mathscr{U}$ at all ($\mathscr{U} = 1$). In terms of representational fairness, we observe that the fairness resulting from $\mathcal{D}_{\text{Gauss}}$ and $\mathcal{D}_{\text{Pareto}}$ decrease slightly due to noise. They drop to 0.90 and 0.80, respectively, and then plateau after $\delta \geq 0.5$. When looking at preference-based fairness in Figure 13(b), we see that as noise increases, the lower bound on preference-based fairness ($\mathscr{P}^{(1)}$) decreases from 1 to as low as 0.45 for $\mathcal{D}_{\text{Pareto}}$ and $\mathcal{D}_{\text{Gaussian}}$ and 0.30 for $\mathcal{D}_{\text{Unif}}$ at $\delta = 2$. We do see that although representational fairness and utility ratio stay consistent, noise can decrease the preference-based fairness guaranteed by these three distributions of preferences, specifically when there is noise in the estimated utilities under the implicit variance model.



(a) $\mathscr{R}$ with noise in utilities (b) $\mathscr{P}^{(1)}$ with noise in utilities (c) $\mathscr{U}$ with noise in utilities

Figure 13: Representation-based fairness is measured by $\mathscr{R}$, preference-based fairness is measured by $\mathscr{P}^{(1)}$, and utility ratio is measured under the implicit variance model with different utility distributions. (a) measures $\mathscr{R}$ when utilities are generated from $\mathcal{D}_{\text{Unif}}$, $\mathcal{D}_{\text{Gauss}}$, and $\mathcal{D}_{\text{Pareto}}$, as well as the theoretical results for $\mathcal{D}_{\text{Unif}}$. (b) measures $\mathscr{P}^{(1)}$, the lower bound for preference-based fairness. (c) measures the utility ratios of the three utility distributions. Our main observation is that $\mathscr{R}$ and $\mathscr{U}$ closely follow the theoretical results stated in Theorem 3.1, but preference-based fairness decreases as noise increases. See Appendix I for details and discussion. The x-axis denotes δ, the y-axis denotes $\mathscr{R}$, $\mathscr{P}^{(1)}$, or $\mathscr{U}$, and the error bars denote the standard error of the mean over 50 iterations.