Meritocratic Fairness for Cross-Population Selection

Michael Kearns * Aaron Roth † Zhiwei Steven Wu‡

Abstract

We consider the problem of selecting a pool of individuals from several populations with incomparable skills (e.g. soccer players, mathematicians, and singers) in a *fair* manner. The quality of an individual is defined to be their relative rank (by cumulative distribution value) within their own population, which permits cross-population comparisons. We study algorithms which attempt to select the highest quality subset despite the fact that true CDF values are not known, and can only be estimated from the finite pool of candidates. Specifically, we quantify the regret in quality imposed by "meritocratic" notions of fairness, which require that individuals are selected with probability that is monotonically increasing in their true quality. We give algorithms with provable fairness and regret guarantees, as well as lower bounds, and provide empirical results which suggest that our algorithms perform better than the theory suggests. We extend our results to a sequential batch setting, in which an algorithm must repeatedly select subsets of individuals from new pools of applicants, but has the benefit of being able to compare them to the accumulated data from previous rounds.

1 Introduction

Consider the following common academic (or similar) hiring scenario: The dean has promised your department 3 faculty slots, in any areas. Your goal is to hire the best candidates possible — but how should you identify them? An immediate problem is that candidates are incomparable across subfields, because, among other things, standards of publication, citation counts, and letter-writing styles can vary considerably across subfields. An attractive way to rank candidates is according to how strong they are relative to others working in the same field, to whom they are directly comparable. If we model each subfield as corresponding to a different distribution over metrics that are monotonically increasing in candidate quality, this is the value we get when we evaluate the CDF function of the distribution on a candidate's realized value. But because the number of candidates each year is small, simply comparing each candidate to their direct competitors this year — i.e. taking their *empirical* CDF values as truth — would lead to a noisy ranking: it could be that due to chance, the best candidate this year in subfield A would be a mediocre candidate in a typical year, and the top two candidates in subfield B would each be the top candidate in a typical year. We would prefer to evaluate our success by considering the unknown *true* CDF value of each candidate.[1] Similar situations, in which we must select a high

*University of Pennsylvania

†University of Pennsylvania

‡University of Minnesota. Work done when ZSW was a PhD student at the University of Pennsylvania.

[1] Letters of recommendation often seek to communicate this information, with statements like "This candidate is among the top 5 students I have seen in my 16 years as a professor."

quality set of candidates from multiple, mutually incomparable groups, arise frequently. In college admissions, distinguished talent in mathematics, football, and trombone playing can all be enough to merit admission — but how do you compare a math olympiad competitor to a quarterback? In track, men's and women's race times come from different distributions — but it is exactly within-group CDF value that wins races. Some affirmative action policies are premised on the assertion that SAT scores and other measures may not be directly comparable across different groups (e.g. due to only advantaged groups having the financial resources for test preparation courses and multiple retakes).

For various reasons, in these settings we may also be concerned with the *fairness* of our choices.[2] But what should fairness mean? In this paper, we take inspiration from Dwork et al. (2012) who propose that fairness should mean that "similar individuals are treated similarly", where "similarity" is measured with respect to some task specific metric. In our setting, the natural task-specific metric is the true within-group CDF value for each individual. On its own, this is compatible with the goal of selecting the best candidates, but in our work, the main obstacle is that we do not know the true CDF value of each individual, and can only approximate this from data. We study the degree to which fairness and optimality are compatible with one another in this setting.

1.1 Our Results

We study a setting in which we wish to select k individuals out of a pool of n for some task. The individuals are drawn from d populations, each represented by a different distribution over real numbers.[3] The number of draws from each distribution may differ. The "quality" of an individual is defined to be their (true) CDF value, as evaluated on the distribution from which they were drawn. An algorithm is evaluated based on the (expected) quality of the k individuals it selects.

The meritocratic fairness definition we propose informally asks that lower quality individuals are never (probabilistically) favored over higher quality individuals. When formulating this definition, we have a choice as to how to incorporate randomness. The strongest formulation possible (*ex-post* fairness) does not involve randomness, and simply requires that every individual actually selected has quality at least that of every individual not selected. The weakest formulation (*ex-ante* fairness) incorporates the randomness of the selection of the population from the underlying distribution, and informally requires that for any pair of individuals, the higher quality individual is selected with weakly higher probability than the lower quality individual, where the randomness is over the realization of the population from the underlying distributions, as well as any internal randomness of the mechanism. An intermediate formulation (*ex-interim* fairness) requires informally that higher quality individuals be selected with weakly higher probability than lower quality individuals, where the probability is computed over the randomness of the mechanism, but *not* over the selection of the population. Roughly speaking, these choices correspond to what an individual may know and still be satisfied by a promise of "fairness". Individuals should be satisfied with ex-post fairness even after the choices of the mechanism are made, with full knowl-

[2] With respect to men's and women's sports, equal opportunity is legislated in Title IX. With respect to faculty hiring, fairness concerns can arise because the proportion of women can vary substantially across subfields. For example, as reported in Cohoon et al. (2011), the percentage of female authors varies from 10% to 44% across ACM conferences, when averaged over the 10 year period from 1998-2008.

[3] We study the simple setting in which each individual is represented by a 1-dimensional "score" — e.g. a credit score, a time in the 100m dash, etc. — which itself may encapsulate or summarize many features into a single value. Generalizing this work to richer representations is an interesting direction for future work.

	Exact Fairness	Approximate Fairness
Ex-Ante	Regret $O(1/n)^\dagger$ (Lemma 2.8)	Regret $O(1/n)^\dagger$ (Lemma 2.8)
Ex-Interim	Regret $\Omega(1)$ (Theorem 5.1).	Regret $\tilde{O}(\sqrt{k}/n)$ (Theorem 3.7)
Ex-Post	Impossible	Regret $\tilde{O}(\sqrt{k}/n)^*$ (Theorem 4.2)

Table 1: An informal summary of results. The bounds are stated in the case when the populations have sizes within a constant factor of one another – see the theorem statements for the precise bounds. † When the population sizes are the same. *Exact ex-post fairness within each population, approximate ex-interim fairness between populations, and selects *approximately* k individuals.

edge of the applicant pool — that is, they should be satisfied with the actual outcome, regardless of the algorithm used to reach it. In contrast, individuals with full knowledge of the applicant pool should still be satisfied with ex-interim fairness *before* the mechanism makes its decisions — that is, they should feel satisfied that the *algorithm* used is fair. An individual should only be satisfied by ex-ante fairness if she has no knowledge of the applicant pool (and so can consider it a random variable) before the choices are made.

Given such a spectrum of fairness constraints, we observe that the strongest ex-post fairness is impossible to achieve, whereas the weakest ex-ante fairness is sometimes easy to achieve: when the population sizes are the same, it is satisfied by the mechanism that simply selects the k individuals with highest empirical CDF values.[4] Our main results therefore concern the cost (in terms of the expected quality of the selected applicants) of asking for the stronger notion of ex-interim fairness. We show that satisfying an exact variant of this constraint requires the selection algorithm to select uniformly at random amongst all individuals, and hence obtain only trivial utility guarantees, but that subject to an approximate relaxation of this constraint, it is possible to recover asymptotically optimal utility bounds. We show that when we further relax the problem, to allow the algorithm to select *approximately* k individuals (rather than exactly k), it is possible to recover asymptotically optimal utility bounds while satisfying *ex-post* fairness guarantees within each sub-population, and approximate ex-interim fairness guarantees across populations. We summarize our results in Table 1. We complement our theoretical results with empirical simulations which emphasize that both the utility and fairness guarantees of our algorithms are better in practice than our theorems promise.

Finally, we remark on an interesting property of our upper bounds: they are *oblivious*, in the sense that they do not make use of the raw scores associated with each individual — only their empirical CDF ranking. As such, our upper bounds can be viewed as universal distributions over permutations (of empirical CDF rankings) that satisfy a fairness guarantee, rather than algorithms. Our lower bounds apply not just to oblivious algorithms, but to any algorithm, even those that can make use of raw scores (or indeed, even knowledge of the family of distributions from which populations are drawn).

1.2 Related Work

This paper fits into a rapidly growing line of work studying "fairness" in learning settings that is now too large to summarize fully, and so we discuss only the most closely related work. Our

[4] However, for the cases in which the populations are not the same size, we do not know of better utility guarantees for ex-ante fairness than those we derive for the stronger notion of ex-interim fairness.

definition of fairness is in the spirit of Dwork et al. (2012), who propose that individual fairness should mean that "similar individuals are treated similarly" with respect to some underlying task-specific metric. As with the work of Joseph et al. (2016) and Jabbari et al. (2016), we define the metric to be a measure of quality already present in the model (in our case, the CDF values of individuals) but unknown to the algorithm, except through samples. It is this necessity to learn the underlying metric that poses the tension between the fairness constraint and the accuracy goal. Although in this line of work, we adopt a definition that merely requires "better individuals be treated better" according to the true *unknown* metric, this necessarily requires that "similar individuals be treated similarly" with respect to empirical estimates of the metric.

Technically, our work includes adaptations of techniques in differential privacy (Dwork et al., 2006). Specifically, we adopt variants of the "report noisy max" algorithm (Dwork and Roth, 2014), and Raskhodnikova and Smith's "exponential mechanism for scores with varying sensitivities" (Raskhodnikova and Smith, 2016), which is itself a variant of the exponential mechanism (McSherry and Talwar, 2007).

2 Model and Preliminaries

There are d different populations, indexed by j. For each population j, there is a pool of candidates with their raw scores (and henceforth observations) drawn i.i.d. from some unknown continuous distribution $\mathcal{F}_j$ over $\mathbb{R}$. Let $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_d$ denote the product distribution. We will slightly abuse notation and write x_{ij} to denote both the individual i in the population j and her associated observation, and write X to denote the set of all candidates. Let m_j be the size of the candidate pool from population j, $n = \sum_j m_j$ be the size of the total population, and $m = \min_j m_j$ be the smallest population size. Each individual x_{ij} is associated with the following values.

- A *cumulative distribution function (CDF)* value $\mathcal{F}_j(x_{ij}) = \Pr_{\mathcal{F}_j}[x < x_{ij}]$,[5] and an empirical CDF value $\widehat{\mathcal{F}_j}(x_{ij}) = \frac{1}{m_j}\sum_{i'=1}^{m_j} \mathbf{1}[x < x_{i'j}]$.
- A *complementary cumulative distribution function (CCDF)* value: $p_{ij} = 1 - \mathcal{F}_j(x_{ij})$ and an empirical CCDF value $\hat{p}_{ij} = \frac{1}{m_j}\sum_{i'=1}^{m_j} \mathbf{1}[x \geq x_{i'j}]$.

A selection algorithm π takes all the n observations X drawn from different distributions as input, and (randomly) selects k individuals as outputs. We will write $\pi(X, x_{ij})$ (or π_{ij} for simplicity) to denote the selection probability over the individual x_{ij}. The *utility* for selecting an individual x_{ij} is her true CDF value $\mathcal{F}_j(x_{ij})$. Equivalently, the *loss* for selecting an individual x_{ij} is the true CCDF value p_{ij}. The *loss* for an algorithm π on input X is then defined as

$$\mathcal{L}(\pi, X) = \frac{1}{k}\sum_{x_{ij} \in X} \pi(X, x_{ij})(1 - \mathcal{F}_j(x_{ij}))$$

and the *expected loss* of the algorithm is $\mathbb{E}_{X \sim \mathcal{F}}[\mathcal{L}(\pi, X)]$.

[5] We adopt a slightly different definition from the standard one: $\mathcal{F}_j(x_{ij}) = \Pr_{\mathcal{F}_j}[x \leq x_{ij}]$.

2.1 Fairness Formulation

Our goal is design selection algorithms subject to a *meritocratic fairness* notion that requires that less qualified candidates (in terms of CDF values) are never preferred over more qualified ones. We will present three different formulations of such notion based on the different forms of randomness we are considering.

First, the weakest formulation is the following *ex-ante fairness*, which guarantees fairness over the randomness of both the random draws of the candidates and the coin flips of the algorithm.

Definition 2.1 (Ex-Ante Fairness)**.** An algorithm π satisfies *ex-ante fairness* if for any pair of candidates $x_{ij}, x_{i'j'}$ with CDF values $\mathcal{F}_j(x_{ij}) > \mathcal{F}_{j'}(x_{i'j'})$, their selection probabilities (when they are in the pool) satisfy

$$\mathbb{E}\Big[\pi(X, x_{ij})\Big] \geq \mathbb{E}\Big[\pi(X, x_{i'j'})\Big]$$

where the expectations are taken over the $(n-2)$ random draws of all the other candidates.

An intermediate formulation of fairness is the following *ex-interim fairness*, which guarantees fairness over the randomness of the algorithms (but not the realizations of X) on almost all of inputs drawn from the distribution.

Definition 2.2 (Exact Ex-Interim Fairness)**.** Let $\delta \in (0,1)$. An algorithm π satisfies *δ-exact ex-interim fairness* if with probability at least $1-\delta$ over the realized observations X, for any pair of individuals $x_{ij}, x_{i'j'} \in X$,

$$\pi(X, x_{ij}) > \pi(X, x_{i'j'}) \quad \text{only if} \quad \mathcal{F}_j(x_{ij}) > \mathcal{F}_{j'}(x_{i'j'})$$

We also consider the following relaxation:

Definition 2.3 (Approximate Ex-Interim Fairness)**.** An algorithm π satisfies *(ε, δ)-approximate ex-interim fairness* if with probability at least $1-\delta$ over the realized observations X, for any pair of individuals $x_{ij}, x_{i'j'} \in X$,

$$\pi(X, x_{ij}) > e^{\varepsilon}\, \pi(X, x_{i'j'}) \quad \text{only if} \quad \mathcal{F}_j(x_{ij}) > \mathcal{F}_{j'}(x_{i'j'})$$

Remark 2.4. *We note that this relaxation of ex-interim fairness bears a similarity to the definition of differential privacy (Dwork et al., 2006), and indeed, techniques from the differential privacy literature will prove useful in designing algorithms to satisfy it.*

Perhaps the strongest formulation is the following *ex-post fairness* condition, which requires that an individual is selected only if a more qualified individual is also selected.

Definition 2.5 (Ex-post Fairness)**.** An algorithm π satisfies *ex-post fairness* if any pair of individuals x_{ij} and $x_{i'j'}$ such that $\mathcal{F}_j(x_{ij}) > \mathcal{F}_{j'}(x_{i'j'})$, the individual $x_{i'j'}$ is admitted only if x_{ij} is also selected.

Note that any algorithm that satisfies ex-post fairness must admit a prefix of individuals from each population, which is also sufficient to guarantee *within population* ex-post fairness, but that this is not sufficient to satisfy the constraint between populations.

It is not hard to see that satisfying ex-post fairness in the generality that we have defined it is impossible, since it requires perfectly selecting the k true best CDF values from only sample data. Thus, the primary focus of our paper is on ex-interim fairness. Unless we specify differently, the term "fair" and "fairness"refer to ex-interim fairness.

2.2 Oblivious Algorithms

A special class of selection algorithms is the class of *oblivious* algorithms, which select candidates with probabilities that only depend on their empirical CDF values, not on their observations.

Definition 2.6 (Oblivious Algorithms)**.** An algorithm π is *oblivious* if for any pair of input observations X and X' that induce the same empirical CDF values over the candidates, $\pi(X) = \pi(X')$.

All of our algorithms presented in this paper are oblivious. As a result, we need to make no assumption on the underlying distributions to achieve both fairness and utility guarantees. Moreover, the utility guarantee of an oblivious algorithm can be characterized as follows.

Lemma 2.7. *The expected loss achieved by any oblivious algorithm π is the expected average empirical CCDF values among the selected candidates.*

A very simple example of an oblivious algorithm is Greedy which selects the k individuals with the highest empirical CDF values (breaking ties uniformly at random).

Lemma 2.8. *Suppose that the populations sizes are the same, that is, $m_j = m$ for each j. The algorithm* Greedy *satisfies ex-ante fairness and has an expected loss at most $\frac{k}{2n} + \frac{1}{m}$.*

To simplify our bounds on the expected loss, we will use $k/2n$ as our benchmark and define the *regret* of an algorithm π to be $\mathcal{R}(\pi) = \mathbb{E}_{X\sim\mathcal{F}}[\mathcal{L}(\pi, X)] - \frac{k}{2n}$.

3 An Approximately Fair Algorithm

In this section, we provide an algorithm that satisfies approximate fairness in the sense of Definition 2.3. We will present our solution in three steps.

1. First, we provide confidence intervals for the candidates' CCDF values p_{ij} based on their empirical CCDF values $\hat{p}_{ij}$. As we show, our bound has a tighter dependence on p_{ij}, which gives better utility guarantee than using the standard DKW inequality of Dvoretzky et al. (1956).

2. Next, we give a simple subroutine NoisyTop that randomly selects k individuals out of n based on their "scores". We show that individuals with similar scores will have close selection probabilities under this subroutine. This subroutine is similar to the "Report Noisy Max" algorithm (Dwork and Roth, 2014).

3. Then, we will use the deviation bound in the first step to assign scores to the candidates. We show that running NoisyTop based on these scores give approximate fairness and low regret guarantees. These scores are computed in a way similar to the generalized exponential mechanism of Raskhodnikova and Smith (2016).

3.1 Confidence Intervals for CCDF Values

We will first give the following concentration inequality specialized for the uniform distribution over $(0, 1)$.

Lemma 3.1. *Fix any $n \in \mathbb{N}$. Let $x_1, x_2, \ldots, x_n$ be i.i.d. draws from the uniform distribution over $(0,1)$. Then with probability at least $1-\delta$, for any $p \in (0,1)$,*

$$|p - \hat{p}| \leq \sqrt{\ln(2n/\delta)}\left(\sqrt{\frac{3p}{n}} + \frac{2}{n}\right)$$

where $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[x_i < p]$.

To translate this result into a deviation bound on the CCDF values, first note that CCDF values for any distribution $\mathcal{F}_j$ are drawn from the uniform distribution over $(0,1)$, so the bound applies immediately to the CCDF values. By a standard calculation, we can also get a bound in terms of the empirical CCDF value $\hat{p}_{ij}$ as shown below.

Lemma 3.2. *For each $j \in [d]$, draw m_j points $X_j = \{x_{ij}\}_{i=1}^{m_j}$ i.i.d. from $\mathcal{F}_j$. For each point x_{ij}, let p_{ij} be its true CCDF value and $\hat{p}_{ij}$ be its empirical CCDF value in $\mathcal{F}_j$. Then with probability at least $1-\delta$ over the n random draws,*

$$|p_{ij} - \hat{p}_{ij}| \leq 9\sqrt{\frac{\hat{p}_{ij}}{m}\ln(2n/\delta)}$$

where $m = \min_j m_j$ and $n = \sum_{j=1}^{d} m_j$.

Remark 3.3. *The standard DKW inequality gives a bound of $\tilde{O}(\sqrt{1/m})$. Our bound gives a tighter dependence for small empirical CCDF value $\hat{p}_{ij}$. For example, when $\hat{p}_{ij} = 1/m$, we obtain a bound of $\tilde{O}(1/m)$.* [6]

3.2 The NoisyTop Subroutine

Given a set of individuals with scores $Y = \{y_1, \ldots, y_n\}$, the subroutine NoisyTop will first perturb each score by adding independent noise drawn from the Laplace distribution,[7] and output the k individuals with the minimum noisy scores (ties broken arbitrarily). We will now show that NoisyTop has the following desirable "Lipschitz" property—individuals with similar scores are chosen with similar probabilities. This is crucial for obtaining approximate fairness.

Algorithm 1 NoisyTop$(\{y_1, y_2, \ldots, y_n\}, \alpha, k)$

Input: n numbers $\{y_1, y_2, \ldots, y_n\}$ and parameter α
For each $i \in [n]$: let $\tilde{y}_i = y_i + \mathrm{Lap}(\alpha)$
Output: the k indices with the smallest $\tilde{y}_i$

Lemma 3.4. *Let $i, j \in [n]$ be such that $\Delta = y_i - y_j \geq 0$. Let P_i and P_j denote the probabilities that the two indices i and j are output by* NoisyTop$(\{y_1, y_2, \ldots, y_n\}, \alpha)$ *respectively. Then $P_i \leq P_j \leq P_i \exp(2\Delta/\alpha)$.*

Proof. Let $\tilde{y}_i$ and $\tilde{y}_j$ be the noisy scores for i or j. We will introduce a new random variable Q to denote the value of the $(k-1)$-st lowest noisy value, not counting $\tilde{y}_i$ and $\tilde{y}_j$. We will slightly abuse

[6] As shown in Corollary 3.8, this also gives an improvement over regret when k is small ($\tilde{O}(\sqrt{k}/n)$ versus $\tilde{O}(\sqrt{1/n})$).
[7] The Laplace distribution $\mathrm{Lap}(b)$ has density function $f(x) = \exp(-|x|/b)$.

notation and write $\Pr[R = r]$ as a shorthand for the pdf of any random variable R evaluated at r. The ratio $\frac{P_i}{P_j}$ can then be written as

$$\frac{\int_{q\in\mathbb{R}} \Pr[Q=q]\Big(\int_{t\in\mathbb{R}} \Pr[\tilde{y}_j = t]\Pr[\tilde{y}_i < \min\{t,q\}]dt\Big)dq}{\int_{q\in\mathbb{R}} \Pr[Q=q]\Big(\int_{t\in\mathbb{R}} \Pr[\tilde{y}_i = t]\Pr[\tilde{y}_j < \min\{t,q\}]dt\Big)dq} \tag{1}$$

For any fixed value $r \in \mathbb{R}$, we also have the following based on the Laplace distribution,

$$\begin{aligned}\frac{\Pr[\tilde{y}_i = r]}{\Pr[\tilde{y}_j = r]} &= \frac{\frac{1}{2\alpha}\exp\left(-\frac{|r-y_i|}{\alpha}\right)}{\frac{1}{2\alpha}\exp\left(-\frac{|r-y_j|}{\alpha}\right)} \\ &= \exp\left(\frac{|r-y_j|}{\alpha} - \frac{|r-y_i|}{\alpha}\right)\end{aligned}$$

By the triangle inequality we know that $|r-y_j| - |r-y_i| \le \Delta$. It follows that for any t and q,

$$\exp(-\Delta/\alpha) \le \frac{\Pr[\tilde{y}_i = t]}{\Pr[\tilde{y}_j = t]} \le \exp(\Delta/\alpha) \quad \text{and,}$$

$$\begin{aligned}\frac{\Pr[\tilde{y}_i < \min\{q,t\}]}{\Pr[\tilde{y}_j < \min\{q,t\}]} &= \frac{\int_{r<\min\{q,t\}} \Pr[\tilde{y}_i = r]\,dr}{\int_{r<\min\{q,t\}} \Pr[\tilde{y}_j = r]\,dr} \\ &\le \exp(\Delta/\alpha)\end{aligned}$$

Plugging these bounds into Equation (1), we get $\frac{P_i}{P_j} \le \exp(2\Delta/\alpha)$. The inequality that $P_i/P_j \le 1$ follows directly from $y_i \ge y_j$. □

3.3 Wrapping Up

We will present our algorithm FairTop by combining the methods in the previous two sections. In the light of Lemma 3.2, we will define the following confidence interval width function on the empirical CCDF values

$$c(\hat{p}) = 9\ln(2n/\delta)\sqrt{\hat{p}/m}$$

and a normalized score function $s(\hat{p}) = \hat{p}/c(\hat{p})$. We have that any candidate is guaranteed a score not much lower than a less qualified one.

Lemma 3.5. *Let $x, y \in [0,1]$ be the (true) CCDF values for two individuals such that $x \le y$. Let $\hat{x}, \hat{y}$ be the empirical CCDF values respectively. Suppose that $|x-\hat{x}| \le c(\hat{x})$ and $|y-\hat{y}| \le c(\hat{y})$, then $s(\hat{x}) - s(\hat{y}) \le 1$.*

Our algorithm FairTop (presented in Algorithm 4) proceeds by first computing the normalized score of every candidates based on their empirical CCDF values, and then calling NoisyTop to output k individuals. We will first establish the approximate fairness guarantee.

Theorem 3.6. *The algorithm* FairTop *instantiated with parameters ε and δ satisfies (ε, δ)-approximate fairness.*

Algorithm 2 FairTop($X = \{x_{ij}\}, \varepsilon, \delta, k, m$)

Input: candidates' observations X, fairness parameters ε, δ, number of selected individuals k, and smallest population size m

For each individual $x_{ij} \in X$

Compute the empirical CCDF value $\hat{p}_{ij}$ and the associated score $s(\hat{p}_{ij})$

Run NoisyTop($\{s(\hat{p}_{ij}\}, 2/\varepsilon, k$)

Proof sketch. By Lemma 3.2, we know that with probability $1-\delta$, for every candidate x_{ij}, the true and empirical CCDF values satisfy $|p_{ij} - \hat{p}_{ij}| \le c(\hat{p}_{ij})$. This means that for any pair of individuals a and a' with CCDF values $p_a < p_{a'}$ (that is, a is more qualified than a'), we also have $s(\hat{p}_a) \le s(\hat{p}_{a'}) + 1$ by Lemma 3.5. Finally, by the result of Lemma 3.4 and the instantiation of NoisyTop, we guarantee that a' will not be selected with substantially higher probability: $\pi_a \exp(\varepsilon) \ge \pi_{a'}$, which recovers the approximate fairness guarantee. □

Our algorithm also has a diminishing regret guarantee:

Theorem 3.7. *Fix any $\beta \in (0,1)$. Then with probability at least $1-\beta$, the algorithm* FairTop *instantiated with fairness parameters ε and δ has regret bounded by*

$$\left(\frac{1}{\varepsilon}\sqrt{\left(\frac{k}{n}+\frac{1}{m}\right)\frac{1}{m}}+\frac{1}{m\varepsilon^2}\right)\cdot \text{polylog}(n, 1/\beta, 1/\delta)$$

Thus for example, as the smallest sampled population size m grows (fixing k and ε), our regret rapidly approaches 0. To understand the utility guarantee better, we will state the regret bound for the following natural scaling, which is also examined in the simulations of Section 7:

Corollary 3.8. *Consider an instance with two population of sizes m_1 and m_2 such that $m_1 = \alpha m_2$ for some constant $\alpha \ge 1$. Suppose we instantiate* FairTop *with parameter $\varepsilon = \Theta(1)$, then the regret is at most $\tilde{O}\left(\frac{\sqrt{k}}{m}\right)$.*

4 Within Population Ex-Post Fairness

In this section, we provide a variant of the FairTop algorithm that satisfies approximate ex-interim fairness across different populations, but also *ex-post fairness* within each population. The key idea here is that since we know the ranking of the candidates true qualities within each population, we can guarantee ex-post fairness within populations as long as we select a prefix of candidates in each population. This will however come at a cost — our algorithm will no longer select *exactly* k individuals, but only *approximately* k individuals.

Similar to FairTop, the algorithm AboveThre (presented in Algorithm 3) also computes the normalized scores for each candidate. Instead of perturbing the scores, AboveThre computes a noisy threshold T_j for each population by adding Laplace noise to $s(k/n)$. The algorithm then selects all candidates with scores above the noisy threshold. Because the algorithm selects a prefix of the raw scores within each population, within population ex-post fairness is immediate. We also show that AboveThre also achieves approximate ex-interim fairness.

Algorithm 3 AboveThre($X = \{x_{ij}\}, \varepsilon, \delta, k, m$)

Input: observations X, fairness parameters ε, δ, target number of selected individuals k, smallest population size m

For each individual x_{ij}

Compute her empirical CCDF value $\hat{p}_{ij}$ and the associated score $s(\hat{p}_{ij})$

For each population j

Compute a noisy threshold $T_j = s(k/n) + \nu_j$ where

ν_j is drawn from $\mathrm{Lap}(1/\varepsilon)$

Select candidates x_{ij} with scores $s(\hat{p}_{ij})$ above T_j

Theorem 4.1. *The algorithm* AboveThre *instantiated with fairness parameters ε and δ satisfies both (ε, δ)-approximate ex-interim fairness and* ex-post fairness *within each population.*

Note that were the algorithm to take all the individuals with scores above $s(k/n)$, it would select a (k/n) fraction from each population and therefore select k people in total. Due to the noisy thresholds, the algorithm will only select approximately k individuals. We will now establish the utility guarantee of AboveThre and show that the number of selected individuals is roughly $k \pm \tilde{O}(\sqrt{k})$ when $m = \Theta(n)$.

Theorem 4.2. *Fix any $\beta \in (0,1)$. With probability at least $1-\beta$, the algorithm* AboveThre *instantiated with fairness parameters ε and δ has regret bounded by*

$$\left(\frac{1}{m\varepsilon^2} + \frac{\sqrt{k}}{\varepsilon\sqrt{mn}}\right) \cdot \mathrm{polylog}(n, d, 1/\delta, 1/\beta),$$

and selects a total number of $\hat{k}$ individuals with

$$|k - \hat{k}| \leq d + \left(\frac{n}{m\varepsilon^2} + \frac{\sqrt{nk}}{\varepsilon\sqrt{m}}\right) \cdot \mathrm{polylog}(n, d, 1/\delta, 1/\beta)$$

5 Lower Bound for Exact Fairness

We will show that it is impossible to achieve exact ex-interim fairness with non-trivial regret guarantees.

Theorem 5.1. *Fix any $\delta < 0.0002$ and any δ-fair algorithm π. There exist two distributions $\mathcal{F}_1$ and $\mathcal{F}_2$ over the two populations such that if algorithm π takes m observations drawn from each distribution as input, and must select at least $k = \Omega(m^{1/2+\alpha})$ individuals for any $\alpha > 0$, π incurs a regret of $\Omega(1)$.*

The main idea is to show that there exist distributions $\mathcal{F}_1$ and $\mathcal{F}_2$ such that any fair algorithm will essentially have to select uniformly at random across $\Omega(m)$ individuals, which incurs regret $\Omega(1)$. We will proceed via Bayesian reasoning. Suppose that the observations from the two populations are drawn from two different unit-variance Gaussian distributions $\mathcal{N}(\mu_1, 1)$ and $\mathcal{N}(\mu_2, 1)$, and both means μ_1 and μ_2 are themselves drawn from the prior $\mathcal{N}(0, 1)$. The following lemma characterizes the posterior distribution on the mean given a collection of observations.

Lemma 5.2 (Murphy (2007))**.** *Suppose that a mean parameter μ is drawn from a prior distribution $\mathcal{N}(0,1)$. Let $D = (x_1, x_2, \ldots, x_m)$ be m i.i.d. draws from the distribution $\mathcal{N}(\mu, 1)$. Then the posterior distribution of μ conditioned on D is the Gaussian distribution $\mathcal{N}(\hat{\mu}, \sigma^2)$, where $\hat{\mu} = \frac{\sum_i x_i}{m+1}$ and $\sigma^2 = \frac{1}{m+1}$.*

The result above shows that conditioned on any m draws from the Gaussian distribution, there is constant probability that the true mean will be bounded away from the posterior maximum likelihood estimate by at least $\Omega(1/\sqrt{m})$. With this observation, we will partition the real line into the following intervals: Given any posterior mean $\hat{\mu}$ any integer $r \geq 1$, let the two intervals $I_r^+(\hat{\mu})$ and $I_r^-(\hat{\mu})$ be

$$I_r^+(\hat{\mu}) = [\hat{\mu} + (r-1)/\sqrt{m}, \hat{\mu} + r/\sqrt{m}] \quad \text{and } I_r^-(\hat{\mu}) = [\hat{\mu} - r/\sqrt{m}, \hat{\mu} - (r-1)/\sqrt{m}]$$

The intervals capture the uncertainty we have regarding the CDF values of the observations x_{ij}. Let $X_j = (x_{1j}, x_{2j}, \ldots, x_{mj}\}$ denote the m draws from each distribution $\mathcal{F}_j$, $\hat{\mu}_j = \frac{\sum_i x_{ij}}{m+1}$ be the posterior mean for μ_j conditioned on the draws. Consider two individuals x_{i1} and $x_{i'2}$ such that $x_{i1} \in I_r^+(\hat{\mu}_1)$ and $x_{i'2} \in I_{r+1}^+(\hat{\mu}_2)$. Even though $(x_{i'2} - \hat{\mu}_2) > (x_{i1} - \hat{\mu}_1)$, there is a constant probability that their CDF values satisfy $\mathcal{F}_1(x_{i1}) > \mathcal{F}_2(x_{i'2})$. Any fair algorithm therefore must play these two individuals in these "neighboring" intervals with equal probabilities.

Next, we show that with high probability over the realizations of the true mean μ and the m draws X, all of the $O(m \log m)$ intervals around the posterior mean will be "hit" by points in X.

Lemma 5.3. *Fix any $c < 1$ and $\beta \in (0,1)$. Let mean μ be drawn from $\mathcal{N}(0,1)$, $\hat{r} = \sqrt{cm \log m} + 1$ and $X = (x_1, x_2, \ldots, x_m)$ be m i.i.d. draws from $\mathcal{N}(\mu, 1)$. Let $\hat{\mu} = \frac{\sum_i x_i}{m+1}$. Then except with probability $2\hat{r} \exp\left(-\frac{m^{(1/2-c/2)}}{\sqrt{2\pi}}\right) + 3\beta$ over the joint realizations of μ and X, the following holds*

- *for all $r \leq \hat{r} - 2\sqrt{2\ln(2/\beta)}$, there exist two draws $x_r^+, x_r^- \in X$ such that $x_r^+ \in I_r^+(\hat{\mu})$ and $x_r^- \in I_r^-(\hat{\mu})$;*
- *the number of points that are bigger than $\hat{\mu} + (\hat{r} - 2\sqrt{2\ln(2/\beta)})/\sqrt{m}$ is no more than*

$$m^{1-c/2} + m^{1/2-c/4}\sqrt{3\ln(1/\beta)}$$

We show that the event that all of the consecutive intervals are occupied for both populations will force a fair algorithm to play all the individuals in these intervals with equal probability. More formally, fix any c and sufficiently small constant β, let $\hat{r} = \sqrt{cn \log n} + 1$ and let $Y = \{x_{ij} \mid x_{ij} \in I_r^+(\hat{\mu}_j) \vee x_{ij} \in I_r^-(\hat{\mu}_j)$ for some $r \leq \hat{r} - 2\sqrt{2\ln(2/\beta)}\}$. Consider the following events:

- FullChain(X_1, X_2): for all $r \leq \hat{r} - 2\sqrt{2\ln(2/\beta)}$ and $j \in \{1,2\}$, both the intervals $I_r^+(\hat{\mu}_j)$, $I_r^-(\hat{\mu}_j)$ contain at least one point in X_j,
- UARChain(π, X_1, X_2): the points in Y are selected by the algorithm π with equal probabilities.

Lemma 5.4. *Fix any δ-fair algorithm π for some $\delta < 0.0002$. With probability at least $1/2$ over the realizations of μ_1, μ_2, X_1 and X_2, the event* FullChain(X_1, X_2) *implies* UARChain(π, X_1, X_2).

Proof sketch for Theorem 5.1. The combination of Lemmas 5.3 and 5.4 shows that with constant probability over μ_1, μ_2 and X, π will need to select $\Omega(m)$ individuals with equal probabilities, which leads to an expected regret of $\Omega(1)$ over the draws of μ_1, μ_2 This means there exist distributions $\mathcal{F}_1 = \mathcal{N}(\mu_1^*, 1)$ and $\mathcal{F}_2 = \mathcal{N}(\mu_2^*, 1)$ under which π incurs $\Omega(1)$ regret. □

6 Sequential Batch Setting

We now study an extension to the sequential batch setting, in which the algorithm selects individuals in T rounds. In each round t, for each population j, there are m_j new candidates with their observations drawn i.i.d. from the distribution $\mathcal{F}_j$. At each round t, the algorithm needs to select k individuals from this pool X^t. Let X_j^t be the set of observations from population j accumulated after the first t rounds. In particular, for any observation x and population j, let the *historical CCDF value* be

$$\hat{q}_j^t(x) = \frac{1}{|m_j t|} \sum_{x' \in X_j^t} \mathbf{1}[x' \geq x]$$

As t grows large, the empirical CCDF values become better estimates for the true CCDF values. We give a variant of the FairTop algorithm, called SeqBatchthat achieves (ε, δ)-approximate fairness in every round, and incurs average regret over time diminishing as $\tilde{O}\left(\frac{1}{\varepsilon\sqrt{mT}}\right)$.

Algorithm 4 SeqBatch$(X^t, \varepsilon, \delta, k, m)$

Input: candidates' scores $\{X^t\}_{t=1}^T$ over T rounds, fairness parameters ε, δ, number of selected individuals k, and smallest population size mt

For each round t

For each individual $x_{ij}^t \in X^t$

Compute her historical empirical CCDF value $\hat{p}_{ij}^t = \hat{q}_j^t(x_{ij}^t)$ and the associated score

$$s(\hat{p}_{ij}^t) = \frac{\hat{p}_{ij}^t}{c(\hat{p}_{ij}^t)} \text{ with } c(x) = \sqrt{\ln(2/\delta)/(2mt)} \tag{2}$$

Run NoisyTop$(\{s(\hat{p}_{ij}^t\}, 2/\varepsilon, k)$ to select k individuals

Theorem 6.1. *Fix any round t and $\beta \in (0,1)$. Then with probability at least $1-\beta$, the instantiation of* SeqBatch *at round t with fairness parameters ε and δ satisfies (ε, δ)-approximate fairness and has regret bounded by*

$$\left(\frac{1}{\varepsilon\sqrt{mt}}\right)\text{polylog}(n, 1/\delta, 1/\beta).$$

The following result follows directly from the Dvoretzky-Kiefer-Wolfowith (DKW) inequality.

Theorem 6.2 (DKW Inequality (Dvoretzky et al., 1956))**.** *Let $A = \{x_1, \ldots, x_m\}$ be i.i.d. draws from some distribution $\mathcal{F}$ over $\mathbb{R}$. For each $x \in \mathbb{R}$, let $\mathcal{F}(x)$ denote the CDF value of $\mathcal{F}$ evaluated at x and let $\mathcal{F}_n(x) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}[x_i < x]$. Then with probability $1-\delta$, for any $x \in \mathbb{R}$,*

$$|\mathcal{F}(x) - \mathcal{F}_n(x)| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

The following is just a variant of Lemma 3.5 tailored to the new score function in Equation (2) and historical CCDF values.

Claim 6.3. *Let $c(\cdot)$ and $s(\cdot)$ be functions defined in Equation (2). Let $x, y \in [0,1]$ be the (true) CCDF values for two individuals such that $x \le y$. Let $\hat{x}, \hat{y}$ be the historical CCDF values respectively. Then $|x - \hat{x}| \le c(\hat{x})$, $|y - \hat{y}| \le c(\hat{y})$, and $s(\hat{x}) - s(\hat{y}) \le 1$.*

Proof of Theorem 6.1. Then the (ε, δ)-approximate fairness guarantee at any single round follows with the exact same reasoning as the proof for Theorem 3.6.

In the instantiation of the algorithm SeqBatch at each round t, the subroutine NoisyTop draws n random variables $\{\nu_{ij}\}$ from the distribution $\mathrm{Lap}(2/\varepsilon)$. For each pair (i, j), we know from the property of Laplace distribution that

$$\Pr\left[|\nu_{ij}| \ge t\left(\frac{2}{\varepsilon}\right)\right] \le \exp(-t).$$

Applying union bound, for any $\beta \in (0,1)$, we have for all pairs of (i, j) that

$$|\nu_{ij}| \le \frac{2\ln(n/\beta)}{\varepsilon} \equiv \gamma.$$

We will condition on this event for the remainder of the proof. For any value x_{ij}, let s_{ij} and $\hat{s}_{ij}$ be the score and noisy score associated with individual x_{ij}^t. By our bound on $|\nu_{ij}|$, we get $|s_{ij} - \hat{s}_{ij}| \le \gamma$.

Fix any integer $r \le k$. Let s^r be the r-th smallest values among the scores s_{ij},let $\hat{s}^r$ be the associated noisy score, and let $\hat{p}_r$ be the associated historical CCDF value. Therefore, for some candidate x_{ij} to have the r-th smallest noisy scores, her score must satisfy

$$s_{ij} \le s_r + 2\gamma,$$

which implies that

$$\hat{p}_{ij} \le \hat{p}_r + 2\gamma\sqrt{\frac{\ln(2/\delta)}{2mt}} = \hat{p}_r + \frac{4\ln(n/\beta)}{\varepsilon}\sqrt{\frac{\ln(2/\delta)}{2mt}}$$

Therefore, the historical empirical CCDF value in each of the selected candidate increases by at most $\frac{4\ln(n/\beta)}{\varepsilon}\sqrt{\frac{\ln(2/\delta)}{2mt}}$, which corresponds to the regret due to Lemma 2.7. □

7 Simulations

We conclude by discussing some illustrative simulation results for FairTop, along with comparisons to simpler algorithms without fairness guarantees. The simulations were conducted on data in which the raw scores for each population $i = 1, 2$ were drawn from $\mathcal{N}(\mu_i, 1)$ respectively, and the μ_i themselves were chosen randomly from $\mathcal{N}(0, 1)$. Thus befitting the motivation for our model, the raw scores are not directly comparable between populations. While we varied the population sizes, they were held in the fixed ratio $m_1/m_2 = 2$ and $k = \lceil 0.1(m_1 + m_2) \rceil$.

For such a simulation with population sizes $m_1 = 100$ and $m_2 = 50$, Figure 1 shows the underlying scores computed by FairTop (which depend only on the *empirical* CDF values) for each member of both populations, but sorted according to their *true* CDF values so that the transpositions that occur between emprical and true CDFs are apparent; the red points are for the larger population and green for the smaller. Overlaid on this arc of underlying scores is a black plot illustrating sample post-noise scores when $\varepsilon = 10$. As we can see, re-sorting the points by their noisy scores will result in a significant amount of additional reshuffling.

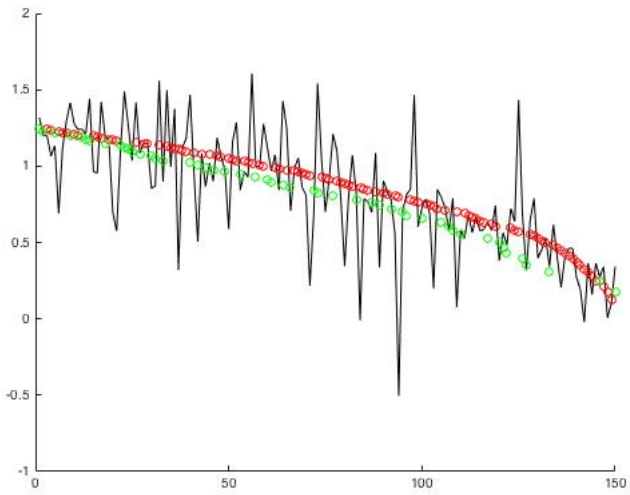

Figure 1: Sample underlying and noisy scores as a function of true CDF rank for $\varepsilon = 10$.

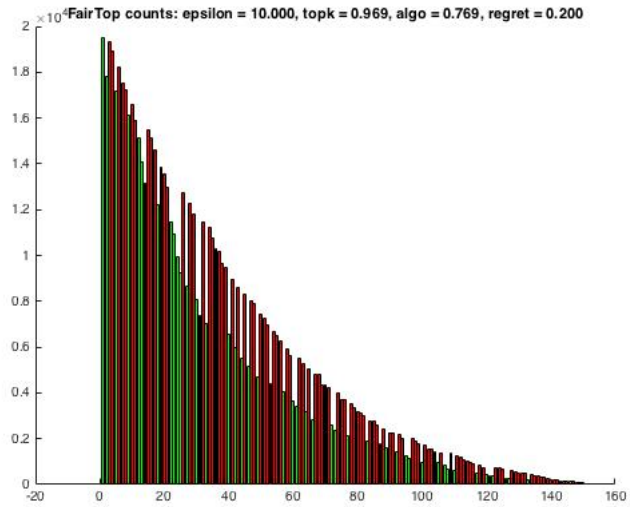

Figure 2: Empirical distribution of selection counts as a function of true CDF rank for $\varepsilon = 10$.

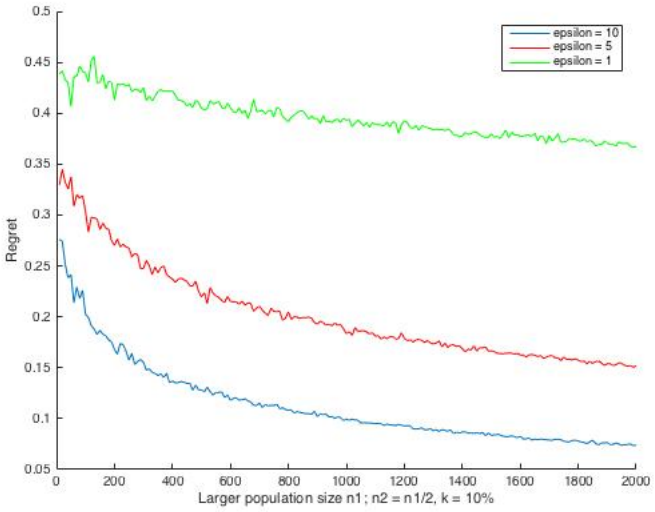

Figure 3: Regret as a function of population sizes, varying ε.

Figure 2 illustrates the induced distribution over chosen individuals; here we show the results of resampling the Laplace noise (again at $\varepsilon = 10$) for 100,000 trials, and choosing the top k post-noise scores across populations. The ordering is again by true CDF values and the same color coding is used. At this value of ε the distribution is biased towards better true CDF values but still enjoys strong fairness properties. For example, the "unfairness ratio" (maximum ratio of the number of times a worse CDF value is chosen to a better CDF value is chosen) is only 1.56 (note that this is *substantially* stronger than the bound of e^{10} guaranteed by our theorem). It is also visually clear that FairTop is treating similar CDF values similarly, both within and between populations.

Nevertheless, the regret of FairTop for these population sizes and ε is nontrivial (roughly 0.20 regret compared to the best k true CDF values). Of course, as per Theorem 3.7 by increasing ε we can reduce regret to any desired level at the expense of weakened fairness guarantees. However, as per Corollary 3.8 even for fixed ε (and therefore fixed fairness properties), regret diminishes rapidly in the natural scaling where the population sizes grow, but in a fixed ratio. This is illustrated empirically in Figure 3, where for varying choices of ε we plot regret as $m_1, m_2 \to \infty$ with $m_1/m_2 = 2$.

We now briefly compare the properties of FairTop to simpler approaches that generally enjoy lower regret but have no fairness properties. Perhaps the simplest is to pick the k highest ranked individuals by *empirical* CDF rank. This method will in general have very low regret, but since it is deterministic, any trial in which it doesn't select the top k *true* CDF values has no fairness guarantee (i.e. the unfairness ratio will be infinite), and this happens in approximately approximately 87% of trials under the simulation parameters above (and approaches 100% as populations grow in fixed ratio).

Perhaps the most natural "learning" approach is to use the raw scores to obtain estimated population means $\hat{\mu}_i$ (or more generally to estimate the unknown parameters of some known or assumed parametric form) and then use the CDFs of $\mathcal{N}(\hat{\mu_1}, 1)$ and $\mathcal{N}(\hat{\mu_2}, 1)$ to select the k best

individuals across the two populations. This again has generally lower regret than FairTop, but is deterministic and without fairness guarantees, with approximately 53% of trials resulting in unbounded unfairness ratio (approaching 100% as populations grow in fixed ratio).

But the main drawback of such a learning approach in comparison to the data-oblivious FairTop is its need for realizability. For instance, if we change the population 2 scores to be drawn from the uniform distribution over a wide range, but the learning approach continues to assume normality in each population, it will virtually always choose only members of population 2, a clear and dramatic violation of any intuitive notion of fairness. This is of course due the fact that the highest scores in population 2 appear to have extraordinarily high CDF values when (incorrectly) assumed to have been drawn from a normal distribution. In contrast FairTop, since it doesn't even consider the actual scores but only generic properties of the relationship between empirical and true CDF values, will behave exactly the same, in both fairness and regret, regardless of how the underlying scores are generated.

References

Cohoon, J. M., Nigai, S., and Kaye, J. J. 2011. Gender and computing conference papers. *Communications of the ACM 54*, 8, 72–80.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist. 27*, 3 (09), 642–669.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. Springer, 265–284.

Dwork, C. and Roth, A. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*, 3–4, 211–407.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. 2016. Fair learning in Markovian environments. *arXiv preprint arXiv:1611.03071*.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

McSherry, F. and Talwar, K. 2007. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE, 94–103.

Murphy, K. P. 2007. Conjugate Bayesian analysis of the Gaussian distribution.

Raskhodnikova, S. and Smith, A. D. 2016. Lipschitz extensions for node-private graph statistics and the generalized exponential mechanism. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, I. Dinur, Ed. IEEE Computer Society, 495–504.

A Basic Probability Tools

Theorem A.1 (Multiplicative Chernoff Bound)**.** *Fix any distribution $\mathcal{P}$ over $[0,1]$. Let $x_1, x_2, \ldots, x_n$ be independent random variables drawn from $\mathcal{P}$. Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\mu = \mathbb{E}_{x\sim\mathcal{P}}[x]$. Then*

$$\Pr[\overline{x} > (1+\varepsilon)\mu] < \exp(-\varepsilon^2\mu n/3) \qquad \textit{and} \qquad \Pr[\overline{x} < (1-\varepsilon)\mu] < \exp(-\varepsilon^2\mu n/2)$$

We will also use the following Gaussian tail bound.

Fact A.2. *Suppose that X is drawn from $\mathcal{N}(\mu, \sigma^2)$. Then*

$$\Pr[X \geq \mu + t] \leq \exp\left(\frac{-t^2}{2\sigma^2}\right) \qquad \textit{for all } t \geq 0.$$

B Missing Proofs for Section 2

Proof of Lemma 2.7. Without loss of generality, we assume that x_{ij} is the individual of rank i in the population j. By the definition of oblivious algorithms, we know that the selection probability π_{ij} is fixed and is independent of the observations (raw scores) in X. Then the expected loss of the algorithm is

$$\mathop{\mathbb{E}}_{X\sim\mathcal{F}}[\mathcal{L}(\pi, X)] = \sum_{x_{ij}\in X} \pi_{ij} \mathop{\mathbb{E}}_{X\sim\mathcal{F}}\Big[[1 - \mathcal{F}_j(x_{ij})]\Big]$$

The expected average empirical CCDF among the selected individuals is

$$\sum_{x_{ij}\in X} \pi_{ij}(1 - \hat{\mathcal{F}}_j(x_{ij})).$$

Note that

$$\mathop{\mathbb{E}}_{X\sim\mathcal{F}}\Big[(1 - \hat{\mathcal{F}}_j(x_{ij}))\Big] = (1 - \mathcal{F}_j(x_{ij}))$$

This implies that

$$\mathop{\mathbb{E}}_{X\sim\mathcal{F}}\left[\sum_{x_{ij}\in X} \pi_{ij}(1 - \hat{\mathcal{F}}_j(x_{ij}))\right] = \sum_{x_{ij}\in X} \pi_{ij} \mathop{\mathbb{E}}_{X\sim\mathcal{F}}\Big[(1 - \hat{\mathcal{F}}_j(x_{ij}))\Big] = \sum_{x_{ij}\in X} \pi_{ij} \mathop{\mathbb{E}}_{X\sim\mathcal{F}}\Big[(1 - \mathcal{F}_j(x_{ij}))\Big] = \mathop{\mathbb{E}}_{X\sim\mathcal{F}}[\mathcal{L}(\pi, X)]$$

which recovers our claim. □

Proof of Lemma 2.8. The utility guarantee follows directly from the fact that the algorithm is selecting from the top $\lceil km/n \rceil$ candidates in each population.

To show the ex-ante fairness guarantee, consider two individuals x_{ij} and $x_{i'j'}$ with CDF values such that $\mathcal{F}_j(x_{ij}) > \mathcal{F}_{j'}(x_{i'j'})$. Then let Y_{ij} and $Y_{i'j'}$ be their ranks within their populations respectively. Then over the randomness of the other $(n-2)$ draws, $Y_{i'j'}$ stochastically dominate Y_{ij}. Since the selection probability is monotically decreasing in the candidate's rank, we have

$$\mathbb{E}\Big[\pi(X, x_{ij})\Big] \geq \mathbb{E}\Big[\pi(X, x_{i'j'})\Big].$$

Thus, the algorithm Greedy satisfies ex-ante fairness. □

C Missing Proofs for Section 3

Proof of Lemma 3.1. For each $j' \in \{0, 1, \ldots, n\}$, let $a_{j'} = j'/n$ and $\hat{a}_{j'} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[x_i < a_{j'}]$. There exists some j such that $a_j \leq p \leq a_{j+1}$ such that

$$a_{j+1} - \hat{a}_{j+1} - 1/n \leq p - \hat{p} \leq a_j - \hat{a}_j + 1/n.$$

To see this, observe that there exists some j such that $a_j \leq p \leq a_{j+1}$, $\hat{a}_j \leq \hat{p} \leq \hat{a}_{j+1}$ and $a_{j+1} - a_j = 1/n$. Furthermore, $\hat{a}_j \leq \hat{p} \leq \hat{a}_{j+1}$. It follows that

$$\begin{aligned} p - \hat{p} &\leq (a_j + 1/n) - \hat{p} \\ &\leq a_j - \hat{a}_j + 1/n \end{aligned}$$

The other side of the inequality follows similarly.

By applying both the Multiplicative Chernoff Bound and the union bound, we have with probability at least $1-\delta$ that for all j',

$$|a_{j'} - \hat{a}_{j'}| \leq \sqrt{\frac{3a_{j'}\ln(2n/\delta)}{n}}$$

We will condition on this event for the remainder of the proof. Then it follows that for any $p \in [0,1]$, we have

$$\begin{aligned}
|p - \hat{p}| &\leq \sqrt{\frac{3a_{j+1}\ln(2n/\delta)}{n}} + \frac{1}{n} \\
&\leq \sqrt{\frac{3(p+1/n)\ln(2n/\delta)}{n}} + \frac{1}{n} \\
&\leq \sqrt{\frac{3p\ln(2n/\delta)}{n}} + \frac{2\sqrt{\ln(2n/\delta)}}{n}
\end{aligned}$$

which completes our bound. □

By a similar analysis, we can get the following confidence interval bound across multiple populations, which gives the following result of Lemma C.1.

Lemma C.1. *For each $j \in [d]$, draw m_j points $\{a_{ij}\}$ independently from the uniform distribution over $(0,1)$. Then with probability at least $1-\delta$ over the n random draws, for all a_{ij}*

$$|a_{ij} - \hat{a}_{ij}| \leq \sqrt{\ln(2n/\delta)}\left(\sqrt{\frac{3p}{m_j}} + \frac{2}{m_j}\right)$$

where $\hat{a}_{ij} = \frac{1}{m_j}\sum_{i'=1}^{m_j} \mathbf{1}[a_{ij} < a_{ij}]$ and $n = \sum_{j=1}^{d} m_j$.

We can then translate this result into the following Lemma C.2, which is an intermediate step for showing Lemma 3.2. The key idea is that that both the CDF and CCDF values for points drawn from any distribution are distributed according to the uniform distribution over $(0,1)$.

Lemma C.2. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_d$ be probability distributions over $\mathbb{R}$. For each $j \in [d]$, draw m_j points $X_j = \{x_{ij}\}_{i=1}^{m_j}$ i.i.d. from $\mathcal{F}_j$. For each point x_{ij}, let p_{ij} be its true CCDF value and $\hat{p}_{ij}$ be its empirical CCDF value in $\mathcal{F}_j$. Then with probability at least $1-\delta$ over the n random draws, for all points x_{ij}*

$$|p_{ij} - \hat{p}_{ij}| \leq \sqrt{\ln(2n/\delta)}\left(\sqrt{\frac{3p_{ij}}{m_j}} + \frac{2}{m_j}\right)$$

where $n = \sum_{j=1}^{l} m_j$.

We are now ready to prove Lemma 3.2.

Proof of Lemma 3.2. Since there are n draws from the distribution, we know that $\hat{p}_{ij} \geq 1/n$ for all i. By the result of Lemma C.2, we know that with probability $1-\delta$, for all $i \in [n]$,

$$|p_{ij} - \hat{p}_{ij}| \leq \sqrt{\ln(2n/\delta)}\left(\sqrt{\frac{3p_{ij}}{m}} + \frac{2}{m}\right)$$

Let $\Delta_{ij} = |p_{ij} - \hat{p}_{ij}|$. Then we have $p_{ij} \leq \hat{p}_{ij} + \Delta_{ij}$. This means

$$\Delta_{ij} \leq \sqrt{\ln(2n/\delta)}\left(\sqrt{\frac{3(\hat{p}_{ij} + \Delta_{ij})}{m}} + \frac{2}{m}\right)$$

Solving for Δ_{ij}, we get

$$\Delta_{ij} \leq \frac{3a^2 + 4a}{2m} + \frac{1}{2}\sqrt{\frac{9a^4 + 24a^3 + 12a^2 m\hat{p}_{ij}}{m^2}}$$

where $a = \sqrt{\ln(2n/\delta)} > 1$. Since $\hat{p}_{ij} \geq 1/n$, we can simplify the RHS of the inequality we get

$$\Delta_{ij} \leq \ln(2n/\delta)\left(\frac{13}{2n} + \sqrt{\frac{3\hat{p}_{ij}}{n}}\right) \leq 9\ln(2n/\delta)\sqrt{\frac{\hat{p}_{ij}}{n}}$$

which recovers the stated bound. □

Proof of Lemma 3.5. We can derive the following:

$$\begin{aligned}
s(\hat{x}) - s(\hat{y}) &\leq \frac{\sqrt{\hat{x}m}}{9\ln(2n/\delta)} - \frac{\sqrt{\hat{y}m}}{9\ln(2n/\delta)} \\
&\leq \frac{\sqrt{m}}{9\ln(2n/\delta)}\left(\sqrt{\hat{x}} - \sqrt{\hat{y}}\right) \\
&= \frac{\sqrt{m}}{9\ln(2n/\delta)}\left(\frac{\hat{x} - \hat{y}}{\sqrt{\hat{x}} + \sqrt{\hat{y}}}\right) \\
&\leq \frac{\sqrt{m}}{9\ln(2n/\delta)}\left(\frac{x - y + c(\hat{x}) + c(\hat{y})}{\sqrt{\hat{x}} + \sqrt{\hat{y}}}\right) \\
\text{(because } x \leq y\text{)} \qquad &\leq \frac{\sqrt{m}}{9\ln(2n/\delta)}\left(\frac{c(\hat{x}) + c(\hat{y})}{\sqrt{\hat{x}} + \sqrt{\hat{y}}}\right) \\
&\leq \left(\frac{\sqrt{m}}{9\ln(2n/\delta)}\right)\left(\frac{9\ln(2n/\delta)}{\sqrt{m}}\right)\left(\frac{\sqrt{\hat{x}} + \sqrt{\hat{y}}}{\sqrt{\hat{x}} + \sqrt{\hat{y}}}\right) = 1
\end{aligned}$$

which recovers the stated bound. □

Proof of Theorem 3.6. By the result of Lemma 3.2, we know that with probability at least $1-\delta$, for each point x_{ij} in the draw,

$$|p_{ij} - \hat{p}_{ij}| \leq 9\sqrt{\frac{\hat{p}_{ij}}{m}}\ln(2n/\delta) = c(\hat{p}_{ij})$$

where let p_{ij} be its true CCDF value and $\hat{p}_{ij}$ be the empirical CCDF value of x_{ij} in $\mathcal{F}_j$. We will condition on this accuracy guarantee on the empirical CCDF values.

Let i_1 and i_2 be two individuals with true CCDF values of f_{i_1} and f_{i_2} and empirical CCDF values $\hat{f}_{i_1}$ and $\hat{f}_{i_2}$ in their populations such that $f_{i_1} - f_{i_2} \geq 0$. It follows that

$$|f_{i_1} - \hat{f}_{i_1}| \leq c(\hat{f}_{i_1}) \qquad \text{and,} \qquad |f_{i_2} - \hat{f}_{i_2}| \leq c(\hat{f}_{i_2})$$

By the result of Lemma 3.5, we know that $s(\hat{f}_{i_1}) - s(\hat{f}_{i_2}) \geq -1$. It follows from Lemma 3.4 that $P_{i_1} \leq \exp(\varepsilon) P_{i_2}$, where P_{i_1} and P_{i_2} denote the probabilities that the two individuals will be selected in the instantiation of FairTop with fairness parameter ε and δ. □

Proof of Theorem 3.7. In our instantiation of the algorithm FairTop, the subroutine NoisyTop draws n random variables $\{v_{ij}\}$ from the distribution $\text{Lap}(2/\varepsilon)$. For each pair (i, j), we know from the property of Laplace distribution that

$$\Pr\left[|v_{ij}| \geq t\left(\frac{2}{\varepsilon}\right)\right] \leq \exp(-t).$$

Applying union bound, for any $\beta \in (0, 1)$, we have for all pairs of (i, j) that

$$|v_{ij}| \leq \frac{2\ln(n/\beta)}{\varepsilon} \equiv \gamma.$$

We will condition on this event for the remainder of the proof. For any value x_{ij}, let s_{ij} and $\hat{s}_{ij}$ be the score and noisy score associated individual. By our bound on $|v_{ij}|$, we get $|s_{ij} - \hat{s}_{ij}| \leq \gamma$.

Fix any integer $r \leq k$. Let s^r be the r-th smallest values among the scores s_{ij} and let $\hat{s}^r$ be the associated noisy score. Since s is a monotonically increasing function, we know that

$$s(r/n - 1/m) \leq s^r \leq s(r/n + 1/m)$$

Suppose that an individual with empirical CCDF value $\hat{p}^r$ and score $s(\hat{p}^r)$ has the r-th lowest noisy score $\hat{s}(\hat{p}^r)$. Since we know there are at least r noisy scores below $s(r/n + 1/m) + \gamma$, we must have

$$\hat{s}(\hat{p}^r) \leq s(r/n + 1/m) + \gamma$$

Since $|\hat{s}(\hat{p}^r) - s(\hat{p}^r)| \leq \gamma$, we also have

$$s(\hat{p}^r) \leq s(r/n + 1/m) + 2\gamma.$$

Plugging in our definition of the score function s,

$$\frac{\sqrt{\hat{p}m}}{9\ln(2n/\delta)} \leq \frac{\sqrt{(r/n + 1/m)m}}{9\ln(2n/\delta)} + 2\gamma$$

Solving for $\hat{p}$, we get the following bound

$$\hat{p} \leq \left(\frac{r}{n} + \frac{1}{m}\right) + \frac{2A}{\varepsilon}\sqrt{\left(\frac{r}{n} + \frac{1}{m}\right)\frac{1}{m}} + \frac{A^2}{m\varepsilon^2} \leq \left(\frac{r}{n} + \frac{1}{m}\right) + \frac{2A}{\varepsilon}\sqrt{\left(\frac{k}{n} + \frac{1}{m}\right)\frac{1}{m}} + \frac{A^2}{m\varepsilon^2}$$

where $A = 18\gamma\ln(2n/\delta) = 36\ln(n/\beta)\ln(2n/\delta)$. Since the algorithm will output k individuals in the end, the average empirical CCDF values is therefore no more than

$$\left(\frac{k}{2n} + \frac{1}{m}\right) + \frac{2A}{\varepsilon}\sqrt{\left(\frac{k}{n} + \frac{1}{m}\right)\frac{1}{m}} + \frac{A^2}{m\varepsilon^2}.$$

Our regret bound follows by applying Lemma 2.7. □

D Missing Proofs in Section 4

Proof of Theorem 4.1. Ex-post fairness within each population directly follows from how we select the individuals. To show (ε, δ)-approximate fairness across populations, consider a pair of individuals i_1 and i_2 in populations 1 and 2 (without loss of generality) such that their true CCDF values satisfy $p_{i_1} \geq p_{i_2}$. Let $\hat{p}_{i_1}$ and $\hat{p}_{i_2}$ be their associated empirical CCDF values respectively, and let P_{i_1} and P_{i_2} denote the probabilities that the two individuals are selected by the algorithm. In particular, the ratio between the two probabilities can be written as

$$\begin{aligned}\frac{P_{i_1}}{P_{i_2}} &= \frac{\Pr[T_1 \geq s(\hat{p}_{i_1})]}{\Pr[T_2 \geq s(\hat{p}_{i_2})]} \\ &= \frac{\Pr[\nu_1 \geq s(\hat{p}_{i_1}) - s(k/n)]}{\Pr[\nu_2 \geq s(\hat{p}_{i_2}) - s(k/n)]}\end{aligned}$$

To bound the ratio in the last expression above, we will consider two cases. Suppose that $s(\hat{p}_1) \geq s(\hat{p}_2)$. Since both random variables ν_1 and ν_2 are drawn from the Laplace distribution $\mathrm{Lap}(1/\varepsilon)$, we know that

$$\frac{\Pr[\nu_1 \geq s(\hat{p}_{i_1}) - s(k/n)]}{\Pr[\nu_2 \geq s(\hat{p}_{i_2}) - s(k/n)]} \leq 1$$

Now suppose that $s(\hat{p}_1) < s(\hat{p}_2)$. By Lemma 3.2, we know that with probability at least $1 - \delta$, for all individual x_{ij}, the empirical and true CCDF values satisfy $|p_{ij} - \hat{p}_{ij}| \leq c(\hat{p}_{ij})$. We will condition on this event for the remainder of the proof. By Lemma 3.5, we have $|s(\hat{p}_{i1}) - s(\hat{p}_{i2})| \leq 1$. Thus,

$$\begin{aligned}&\frac{\Pr[\nu_1 \geq s(\hat{p}_{i_1}) - s(k/n)]}{\Pr[\nu_2 \geq s(\hat{p}_{i_2}) - s(k/n)]} \\ &= \frac{\int_{t \geq s(\hat{p}_{i_1}) - s(k/n)} \exp\left(-\varepsilon \left|t - s(\hat{p}_{i_1}) + s(k/n)\right|\right)}{\int_{t \geq s(\hat{p}_{i_2}) - s(k/n)} \exp\left(-\varepsilon \left|t - s(\hat{p}_{i_2}) + s(k/n)\right|\right)} \\ &\leq \exp(\varepsilon |s(\hat{p}_{i_1}) - s(\hat{p}_{i_2})|) \\ &\leq \exp(\varepsilon).\end{aligned}$$

This concludes the proof for (ε, δ)-approximate fairness. □

Proof of Theorem 4.2. Since each noise random variable ν_j is drawn from the Laplace distribution $\mathrm{Lap}(1/\varepsilon)$, we have

$$\Pr[|\nu_j| \geq t/\varepsilon] = \exp(-t)$$

By applying union bound, we know that with probability at least $1 - \beta$,

$$\max_j |\nu_j| \leq \frac{\ln(d/\beta)}{\varepsilon} \equiv E.$$

We will condition on this event for the remainder of the proof, which is the case except with probability β. By our definition of the noisy thresholds, we have for each $j \in [d]$, $T_j \in [s(k/n) - E, s(k/n) + E]$. For each $j \in [d]$, let $\hat{p}^j$ be highest empirical CCDF value among the admitted individuals in population j. It follows that

$$s(\hat{p}^j) \leq s(k/n) + E \qquad \text{and} \qquad s(\hat{p}^j + 1/m_j) \geq s(k/n) - E$$

Plugging in the definition of the score function s and solving for $\hat{p}^j$, we get

$$\frac{k}{n} - \frac{1}{m_j} + \frac{A^2}{m\varepsilon^2} - \frac{2A\sqrt{k}}{\varepsilon\sqrt{mn}} \le \hat{p}^j \le \frac{k}{n} + \frac{A^2}{m\varepsilon^2} + \frac{2A\sqrt{k}}{\varepsilon\sqrt{mn}} \tag{3}$$

where $A = 9\ln(2n/\delta)\ln(d/\beta)$. Note that the average loss within each population is at most $\hat{p}^j/2$, so the average loss in population j is at most

$$\frac{\hat{p}^j}{2} \le \frac{k}{2n} + \frac{A^2}{2m\varepsilon^2} + \frac{A\sqrt{k}}{\varepsilon\sqrt{mn}}$$

Therefore, the average empirical CCDF values among all selected individuals is no more than $\frac{k}{2n} + \frac{A^2}{2m\varepsilon^2} + \frac{A\sqrt{k}}{\varepsilon\sqrt{mn}}$. Our regret bound follows by applying Lemma 2.7.

Let k_j be the number of individuals we admit in population j. For all $j \in [d]$, we then have

$$\frac{km_j}{n} - 1 + \frac{m_j A^2}{m\varepsilon^2} - \frac{2m_j A\sqrt{k}}{\varepsilon\sqrt{mn}} \le k_j \le \frac{km_j}{n} + \frac{m_j A^2}{m\varepsilon^2} + \frac{2m_j A\sqrt{k}}{\varepsilon\sqrt{mn}}$$

Summing the inequalities above over all j and using the fact that $\sum_j m_j = n$, we know that the total number of admitted individuals satisfies

$$k - d + \frac{A^2 n}{m\varepsilon^2} - \frac{2A\sqrt{nk}}{\varepsilon\sqrt{m}} \le \sum_j k_j \le k + \frac{A^2 n}{m\varepsilon^2} + \frac{2A\sqrt{nk}}{\varepsilon\sqrt{m}}$$

which recovers the stated bound. □

E Missing Proofs for Section 5

Lemma E.1. *Fix any $0 < c < 1$ and mean μ. Let $\hat{r} = \sqrt{cm\log m} + 1$ and $X = (x_1, x_2, \ldots, x_m)$ be m i.i.d. draws from $\mathcal{N}(\mu, 1)$. Then for all $r \le \hat{r}$, both intervals $I_r^+(\mu)$ and $I_r^-(\mu)$ contain at least one point in X except with probability*

$$2\hat{r}\exp\left(-\frac{n^{(1/2-c/2)}}{\sqrt{2\pi}}\right)$$

over the realizations of X.

Proof. We will make use of the following claim about the unit-variance Gaussian distribution that follows directly from the density function of the standard Gaussian distribution.

Claim E.2. *Let X be a random number drawn from $\mathcal{N}(\mu, 1)$. Then for any integer $r \ge 1$,*

$$\Pr[X \in I_r^+(\mu)] = \Pr[X \in I_r^-(\mu)] \le \frac{1}{\sqrt{2m\pi}}\exp\left(-\frac{r^2}{2m}\right)$$

The probability that any X_i falls outside of $I_r^+(\mu)$ for any $r \le \hat{r}$ is at most

$$1 - \frac{1}{\sqrt{2m\pi}}\exp\left(-\frac{(r-1)^2}{2m}\right) \le 1 - \frac{1}{\sqrt{2m\pi}}\exp\left(-\frac{c\log m}{2}\right) = 1 - \frac{1}{\sqrt{2\pi}}\frac{1}{m^{1/2+c/2}}$$

It follows that the probability that all the random draws X fall outside of I_r^+ for any $r \le \hat{r}$ is at most

$$\left(1 - \frac{1}{\sqrt{2\pi}} \frac{1}{m^{1/2+c/2}}\right)^m = \left[\left(1 - \frac{1}{\sqrt{2\pi}} \frac{1}{m^{(1/2+c/2)}}\right)^{\sqrt{2\pi} m^{(1/2+c/2)}}\right]^{m^{(1/2-c/2)}/\sqrt{2\pi}}$$
$$\le \exp\left(-\frac{m^{(1/2-c/2)}}{\sqrt{2\pi}}\right)$$

where the inequality follows from the fact that $(1 - 1/x)^x \le 1/e$ for any $x > 0$. The same upper bound also applies to I_r^- since Gaussian is a symmetrical distribution. Thus, for any $r \le \hat{r}$,

$$\Pr[\forall i, X_i \notin I_r^+] = \Pr[\forall i, X_i \notin I_r^-] \le \exp\left(-\frac{m^{(1/2-c/2)}}{\sqrt{2\pi}}\right).$$

Then the stated bound follows from a direct application of union bound for the $2\hat{r}$ intervals. □

Proof of Lemma 5.3. By the Gaussian tail above we know that, for any random variable x drawn from the distribution $\mathcal{N}(\mu, 1)$, we have

$$\Pr\left[x \ge \mu + \frac{\hat{r}}{\sqrt{m}}\right] \le \exp\left(\frac{-\hat{r}^2}{2m}\right) \le \exp\left(\frac{-cm \log m}{2m}\right) = m^{-c/2}$$

By applying the multiplicative Chernoff bound in Theorem A.1, we can show that the number of points in X that are larger than $\mu + \hat{r}/\sqrt{m}$ is at most

$$m^{1-c/2} + m^{1/2-c/4}\sqrt{3\ln(1/\beta)} \equiv A$$

with probability at least $1 - \beta$ over the draws of X. We will condition on this event for the remainder of the argument.

Let $\overline{x} = \frac{\sum_i x_i}{m}$ denote the empirical mean of the draws in X. We know that $\overline{x}$ is distributed according to $\mathcal{N}(\mu, 1/m)$. By the Gaussian tail bound, we get

$$\Pr[|\overline{x} - \mu| > t] \le 2\exp\left(-\frac{mt^2}{2}\right)$$

This means with probability at least $1 - \beta$ over the random draws of X that

$$|\overline{x} - \mu| \le \sqrt{\frac{2\ln(2/\beta)}{m}}.$$

Since $\mu \sim \mathcal{N}(0, 1)$, we know with probability at least $1 - \beta$ over the realization of μ that $|\mu| \le \sqrt{2\ln(2/\beta)}$. It follows that with probability at least $1 - 2\beta$ the joint realizations of μ and X,

$$\begin{aligned}
|\hat{\mu} - \mu| &= \left|\left(\frac{m}{m+1}\right)\overline{x} - \mu\right| \\
&\le \left(\frac{m}{m+1}\right)|\overline{x} - \mu| + \frac{|\mu|}{m+1} \\
&\le \sqrt{\frac{2\ln(2/\beta)}{m}} + \frac{\sqrt{2\ln(2/\beta)}}{m} \\
&\le \frac{2\sqrt{2\ln(2/\beta)}}{\sqrt{m}}
\end{aligned}$$

We will condition on this event for the rest of the proof. It follows that the number of points in X that are bigger than $\hat{\mu} + (\hat{r} - 2\sqrt{2\ln(2/\beta)})/\sqrt{n}$ is no more than A.

From Lemma E.1, we know that except with probability $2\hat{r}\exp\left(-\frac{n^{(1/2-c/2)}}{\sqrt{2\pi}}\right)$ over the realization of X that all intervals $I_r^+(\mu)$ and $I_r^-(\mu)$ with $r \le \hat{r}$ contain at least one point in X. We will also condition on this event. Since μ and $\hat{\mu}$ differ by at most the length of $2\sqrt{2\ln(2/\beta)}$ intervals, we know that all intervals $I_r^+(\hat{\mu})$ and $I_r^-(\hat{\mu})$ with $r \le \hat{r} - 2\sqrt{2\ln(2/\beta)}$ contain at least one point in X. □

The following useful lemma follows directly from Bayes law.

Lemma E.3. *Consider the following two experiments. In the first, let $\mu \sim \mathcal{N}(0,1)$ and $X = (x_1, x_2, \ldots, x_n)$ be i.i.d. draws $\mathcal{N}(\mu, 1)$, and W denote the joint distribution over (μ, X). In the second, let $\mu \sim \mathcal{N}(0,1)$ and $X = (x_1, x_2, \ldots, x_n)$ be i.i.d. draws from $\mathcal{N}(\mu, 1)$, and then re-draw the mean $\mu' \sim \mathcal{N}(\hat{\mu}, \sigma_n^2)$ from its posterior distribution with $\hat{\mu} = \frac{\sum_i X_i}{n+1}$ and $\sigma_n^2 = \frac{1}{n+1}$. Let $(\mu', X) \sim W'$. Then W and W' are identical distributions.*

Proof of Lemma 5.4. In the following, we will write μ to denote the means (μ_1, μ_2), write $X \sim \mu$ to denote that the draws X_1 and X_2 are drawn from $\mathcal{N}(\mu_1, 1)$ and $\mathcal{N}(\mu_2, 1)$, P to denote the Gaussian priors over the means, and write $P|X$ to denote the posterior distributions over μ conditioned on the draws X. Define the set $\textsc{Unfair}(\pi, \mu)$ to be the set of realized observations X_1 and X_2 that cause the algorithm π to violate exact fairness.

Suppose that the realized observations X satisfy $\textsc{FullChain}(X_1, X_2)$, but $\textsc{UARChain}(\pi, X_1, X_2, c)$ does not hold. This means there exists two individuals with observations x_{ij} and $x_{i'j'}$ such that their selection probabilities $\pi_{ij} \neq \pi_{ij}$ and

$$|(x_{i1} - \hat{\mu}_1) - (x_{i'2} - \hat{\mu}_2)| \le \frac{2}{\sqrt{n}}$$

Note that the posterior on each μ_j conditioned on the observations X_j is distributed according to the Gaussian distribution $\mathcal{N}(\hat{\mu}_j, \frac{1}{n+1})$. It follows that for sufficiently large n, we have

$$\Pr_{\mu \sim P|X}[\mathcal{F}_1(x_{i1}) > \mathcal{F}_2(x_{i'2})], \Pr_{\mu \sim P|X}[\mathcal{F}_1(x_{i1}) < \mathcal{F}_2(x_{i'2})] > 0.0004$$

It follows that for any such set of realized observations X,

$$\Pr_{\mu' \sim P|X}[X \in \textsc{Unfair}(\pi, \mu')] > 2\delta \tag{4}$$

Next, since the algorithm π is δ-fair, for any fixed means μ_1 and μ_2,

$$\Pr_{X \sim \mu}[X \in \textsc{Unfair}(\pi, \mu)] \le \delta.$$

Furthermore, given our prior distribution P over the means μ, we also have

$$\Pr_{\mu \sim P, X \sim \mu}[X \in \textsc{Unfair}(\pi, \mu)] \le \delta$$

By the result of Lemma E.3, we also have

$$\Pr_{\mu \sim P, X \sim \mu, \mu' \sim P|X}[X \in \textsc{Unfair}(\pi, \mu')] \le \delta.$$

By Markov inequality, we get

$$\Pr_{\mu\sim P, X\sim\mu}\left[\Pr_{\mu'\sim P|X}[X \in \textsc{Unfair}(\pi,\mu')] \geq 2\delta\right] \leq 1/2.$$

This means with probability at least 1/2 over the joint distribution over the means μ and draws X, the following holds

$$\Pr_{\mu'\sim P|X}[X \in \textsc{Unfair}(\pi,\mu')] \leq 2\delta$$

Note that inequality (4) shows that the above does not hold if X satisfies $\textsc{FullChain}(X_1, X_2)$, but $\textsc{UARChain}(\pi, X_1, X_2, c)$ does not hold. Thus, with probability at least 1/2 over the means and observations, $\textsc{FullChain}(X_1, X_2)$ implies that $\textsc{UARChain}(\pi, X_1, X_2, c)$. □

Proof for Theorem 5.1. Fix any $\alpha \in (0,1)$ and let $c = 1 - 2\alpha$. Let $\hat{r} = \sqrt{cn\log n} + 1$, β to be some sufficiently small constant, and

$$N = n^{1-c/2} + n^{1/2-c/4}\sqrt{3\ln(1/\beta)} \qquad \text{and} \qquad \delta_1 = 2\hat{r}\exp\left(-\frac{n^{(1/2-c/2)}}{\sqrt{2\pi}}\right) + 3\beta$$

We know from Lemma 5.4 that

$$\Pr_{\mu\sim P, X\sim\mu}[\textsc{FullChain}(X_1, X_2) \Rightarrow \textsc{UARChain}(\pi, X_1, X_2, c)] \geq 1/2$$

Also, based on Lemma 5.3, we know that

$$\Pr_{\mu\sim P, X\sim\mu}[\textsc{FullChain}(X_1, X_2)] \geq 1 - \delta_1$$

Together, we have

$$\Pr_{\mu\sim P, X\sim\mu}[\textsc{UARChain}(\pi, X_1, X_2, c)] = \Pr_{\mu\sim P}\left[\Pr_{X\sim\mu}[\textsc{UARChain}(\pi, X_1, X_2, c)]\right] \geq 1/2 - \delta_1 > 1/4$$

where the last inequality holds for sufficiently large n and sufficiently small β. This implies that there exists some means μ_1^* and μ_2^* such that

$$\Pr_{X\sim\mu}[\textsc{UARChain}(\pi, X_1, X_2, c)] > 1/4.$$

That is if the observations are drawn from $\mathcal{N}(\mu_1^*, 1)$ and $\mathcal{N}(\mu_2^*, 1)$, then with probability more than 1/4, the algorithm will need to choose the points in Y with equal probabilities.

Suppose that the algorithm π needs to select $k = 4N$ individuals from the two populations. This means that with probability at least $1 - \beta$, the algorithm must select $k/2$ individuals from the set Y. We know that Y contains at least $n/2$ points from each population (for sufficiently large n), we can guarantee that with probability at least 3/4 the average true CCDF values for the observations in Y is $\Omega(1)$. In other words, the average loss incurred by π is $\Omega(1)$ if the algorithm needs to select the points in Y with equal probability. □