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 decreasing 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.

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. Similar situations, in which we must select a high quality set of candidates from multiple, mutually incomparable groups, arise frequently. 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. 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. 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 knowledge 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.

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.

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 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; 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 fairness than those we derive for the stronger notion of ex-interim fairness.
differential privacy (Dwork et al., 2006). Specifically, we adopt variants of the “report noisy max” algorithm (Dwork & Roth, 2014), and Raskhodnikova and Smith’s “exponential mechanism for scores with varying sensitivities” (Raskhodnikova & Smith, 2016), which is itself a variant of the exponential mechanism (McSherry & 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 F_j over \mathbb{R}. Let \mathcal{F} = F_1 \times \cdots \times 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 F_j(x_ij) = \Pr_{F_j}[x < x_ij], and an empirical CDF value \hat{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 - 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 \mathcal{A} takes all the n observations X drawn from different distributions as input, and (randomly) selects k individuals as outputs. We will write \mathcal{A}(X, x_{ij}) (or \mathcal{A}_{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 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 \mathcal{A} on input X is then defined as

L(\mathcal{A}, X) = \frac{1}{k} \sum_{x_{ij} \in X} \mathcal{A}(X, x_{ij})(1 - F_j(x_{ij}))

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.

<table>
<thead>
<tr>
<th></th>
<th>Exact Fairness</th>
<th>Approximate Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex-Ante</td>
<td>Regret O(1/n) * (Lemma 2.8)</td>
<td>Regret O(1/n) * (Lemma 2.8)</td>
</tr>
<tr>
<td>Ex-Interim</td>
<td>Regret O(1) (Theorem 5.1)</td>
<td>Regret O(\sqrt{k/n}) (Theorem 3.7)</td>
</tr>
<tr>
<td>Ex-Post</td>
<td>Impossible</td>
<td>Regret O(\sqrt{k/n}) * (Theorem 4.2)</td>
</tr>
</tbody>
</table>

3We adopt a slightly different definition from the standard one: F_j(x_{ij}) = \Pr_{F_j}[x \leq x_{ij}].

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

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 \mathcal{A} satisfies ex-ante fairness if for any pair of candidates x_{ij}, x_{i'j'} with CDF values F_j(x_{ij}) > F_j(x_{i'j'}), their selection probabilities (when they are in the pool) satisfy

\[ \mathbb{E}[\mathcal{A}(X, x_{ij})] \geq \mathbb{E}[\mathcal{A}(X, x_{i'j'})] \]

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 \mathcal{A} satisfies \delta-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,

\[ \mathcal{A}(X, x_{ij}) > \mathcal{A}(X, x_{i'j'}) \quad \text{only if} \quad F_j(x_{ij}) > F_j(x_{i'j'}) \]

We also consider the following relaxation:

Definition 2.3 (Approximate Ex-Interim Fairness). An algorithm \mathcal{A} satisfies (\varepsilon, \delta)-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,

\[ \mathcal{A}(X, x_{ij}) > \varepsilon \mathcal{A}(X, x_{i'j'}) \quad \text{only if} \quad F_j(x_{ij}) > F_j(x_{i'j'}) \]

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 \( A \) satisfies ex-post fairness if any pair of individuals \( x_{ij} \) and \( x_{ij'} \) such that \( F_j(x_{ij}) > F_j(x_{ij'}) \), the individual \( x_{ij'} \) 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 \( A \) is oblivious if for any pair of input observations \( X \) and \( X' \) that induce the same empirical CDF values over the candidates, \( A(X) = A(X') \).

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 \( A \) is the expected average empirical CCDF values among the selected candidates.

A very simple example of an oblivious algorithm is \textsc{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 \textsc{Greedy} satisfies ex-ante fairness and has an expected loss at most \( \frac{k}{2m} + \frac{1}{m} \).

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

### 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 \textsc{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 & Roth, 2014).

3. Then, we will use the deviation bound in the first step to assign scores to the candidates. We show that running \textsc{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 & 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 \( \hat{p} = \frac{1}{n} \sum_{i=1}^{n} 1[x_i < p] \).

To translate this result into a deviation bound on the CCDF values, first note that CCDF values for any distribution \( F_j \) 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 \( F_j \). For each point \( x_{ij} \), let \( p_{ij} \) be its true CCDF value and \( \hat{p}_{ij} \) be its empirical CCDF value in \( F_j \). Then with probability at least \( 1 - \delta \) over the \( n \) random draws,

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

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 \(O(\sqrt{1/m})\). Our bound gives a tighter dependence for small empirical CCDF value \(\hat{p}_{ij}\). For example, when \(p_{ij} = 1/m\), we obtain a bound of \(\hat{O}(1/m)\). \(^6\)

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\((\{y_1, \ldots, y_n\}, \alpha, k)\)

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

**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\((\{y_1, \ldots, y_n\}, \alpha)\) 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 notation and write \(\Pr[R=r] = \int_{q \in \mathbb{R}} \Pr[Q=q] \left( \Pr[\tilde{y}_i = t] \Pr[\tilde{y}_j < \min\{t, q\}] dt \right) dq\)

\[\int_{q \in \mathbb{R}} \Pr[Q=q] \left( \Pr[\tilde{y}_i = t] \Pr[\tilde{y}_j < \min\{t, q\}] dt \right) dq\]

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

\[\Pr[\tilde{y}_i = r] = \frac{1}{2\alpha} \exp \left( -\frac{|r-y_i|}{\alpha} \right)\]

\[\Pr[\tilde{y}_j = r] = \frac{1}{2\alpha} \exp \left( -\frac{|r-y_j|}{\alpha} \right)\]

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

\[\exp(-\Delta/\alpha) \leq \frac{\Pr[\tilde{y}_i = t]}{\Pr[\tilde{y}_j = t]} \leq \exp(\Delta/\alpha)\]

Plugging these bounds into Equation (1), we get \(Pr[\tilde{y}_i < \min\{q, t\}] \leq \frac{\int_{r<\min\{q,t\}} \Pr[\tilde{y}_i = r] dr}{\int_{r<\min\{q,t\}} \Pr[\tilde{y}_j = r] dr} \leq \exp(\Delta/\alpha)\).

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 \(x \leq y\). Let \(\hat{x}, \hat{y}\) be the empirical CCDF values respectively. Suppose that \(|x - \hat{x}| \leq c(\hat{x})\) and \(|y - \hat{y}| \leq c(\hat{y})\), then \(s(\hat{x}) - s(\hat{y}) \leq 1\).

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

**Input:** candidates’ observations \(X\), fairness parameters \(\varepsilon, \delta\), 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\((s(\hat{p}_{ij}), 2/\varepsilon, k)\)

Our algorithm FairTop (presented in Algorithm 2) proceeds by first computing the normalized score of every candidate 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 \(\varepsilon, \delta\) satisfies \((\varepsilon, \delta)\)-approximate fairness.

**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 \(|p_{ij} - \hat{p}_{ij}| \leq c(\hat{p}_{ij})\). 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 \(s(\hat{p}_a) \leq s(\hat{p}_{a'}) + 1\) 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: \(A_a \exp(\varepsilon) \geq A_{a'}\), 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 $\varepsilon$ and $\delta$ has regret bounded by

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

Thus for example, as the smallest sampled population size $m$ grows (fixing $k$ and $\varepsilon$), 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 $\alpha \geq 1$. Suppose we instantiate FAIRTOP with parameter $\varepsilon = \Theta(1)$, then the regret is at most $\hat{O} \left( \frac{\sqrt{k}}{m} \right)$.

### 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 ABOVEHRE (presented in Algorithm 3) also computes the normalized scores for each candidate. Instead of perturbing the scores, ABOVEHRE 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 ABOVEHRE also achieves approximate ex-interim fairness.

**Theorem 4.1.** The algorithm ABOVEHRE instantiated with fairness parameters $\varepsilon$ and $\delta$ satisfies both $(\varepsilon, \delta)$-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 ABOVEHRE and show that the number of selected individuals is roughly $k \pm \hat{O}(\sqrt{k})$ when $m = \Theta(n)$.

**Algorithm 3 ABOVEHRE**

**Input:** observations $X$, fairness parameters $\varepsilon, \delta$, 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 $\nu_j$ is drawn from Lap$(1/\varepsilon)$

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

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

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

and selects a total number of $\hat{k}$ individuals with $|k - \hat{k}| \leq d + \left( \frac{n}{m\varepsilon^2} + \frac{\sqrt{nk}}{\varepsilon \sqrt{m}} \right) \cdot \text{polylog}(n, d, 1/\delta, 1/\beta)$

### 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 $\delta$-fair algorithm $A$. There exist two distributions $F_1$ and $F_2$ over the two populations such that if algorithm $A$ takes $m$ observations drawn from each distribution as input, and must select at least $k = \Omega(m^{1/2 + \alpha})$ individuals for any $\alpha > 0$, $A$ incurs a regret of $\Omega(1)$.

The main idea is to show that there exist distributions $F_1$ and $F_2$ 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 $N(\mu_1, 1)$ and $N(\mu_2, 1)$, and both means $\mu_1$ and $\mu_2$ are themselves drawn from the prior $N(0, 1)$. 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 $\mu$ is drawn from a prior distribution $N(0, 1)$. Let $D = (x_1, x_2, \ldots, x_m)$ be $m$ i.i.d. draws from the distribution $N(\mu, 1)$. Then the posterior distribution of $\mu$ conditioned on $D$ is the Gaussian distribution $N(\hat{\mu}, \sigma^2)$, where $\hat{\mu} = \frac{\sum x_i}{m+1}$ and $\sigma^2 = \frac{1}{m+1}$. 

---

*Meritocratic Fairness for Cross-Population Selection*

---
Next, we show that with high probability over the realiza-
tions of observations \(x_{ij}\). Let \(X_j = \{x_{1j}, x_{2j}, \ldots, x_{mj}\}\) denote the \(m\) draws from each distribution \(X_j, \mu_j = \frac{\sum x_{ij}}{m+1}\) be the posterior mean for \(\mu_j\) conditioned on the draws. Consider two individuals \(x_{1i}\) and \(x_{1j}\) such that \(x_{1i} \in I^+ (\hat{\mu}_i)\) and \(x_{1j} \in I^+ (\hat{\mu}_j)\). Even though \((x_{1j} - x_{1i}) > 0\), there is a constant probability that their CDF values satisfy \(F_1(x_{1i}) > F_2(x_{1j})\). 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 \(\mu\) and the \(m\) draws, the following holds

\[
\begin{align*}
I^+ (\hat{\mu}) &= [\hat{\mu} + (r - 1)/\sqrt{m}, \hat{\mu} + r/\sqrt{m}] \\
I^- (\hat{\mu}) &= [\hat{\mu} - r/\sqrt{m}, \hat{\mu} - (r - 1)/\sqrt{m}].
\end{align*}
\]

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 \(X_j, \mu_j = \frac{\sum x_{ij}}{m+1}\) be the posterior mean for \(\mu_j\) conditioned on the draws. Consider two individuals \(x_{1i}\) and \(x_{1j}\) such that \(x_{1i} \in I^+ (\hat{\mu}_i)\) and \(x_{1j} \in I^+ (\hat{\mu}_j)\). Even though \((x_{1j} - x_{1i}) > 0\), there is a constant probability that their CDF values satisfy \(F_1(x_{1i}) > F_2(x_{1j})\). Any fair algorithm therefore must play these two individuals in these “neighboring” intervals with equal probabilities.

We conclude by discussing some illustrative simulation re-
sults for FAIRTOP, along with comparisons to simpler algo-
rithms without fairness guarantees. The simulations were conducted on data in which the raw scores for each pop-
ulation \(i = 1, 2\) were drawn from \(\mathcal{N}(\mu_i, 1)\) respectively, and the \(\mu_i\) 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 \(k = [0.1(m_1 + m_2)]\).

For such a simulation with population sizes \(m_1 = 100\) and \(m_2 = 50\), Figure 1(a) 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 empirical 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(b) 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 $\varepsilon$ 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 $\varepsilon$ is nontrivial (true CDF values and the same color coding is used). At this value of $\varepsilon$ 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.

Figure 1(c) illustrates the regret as a function of population sizes, for $\varepsilon = 1$ (green), 5 (red), and 10 (blue).

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 $N(\hat{\mu}_1, 1)$ and $N(\hat{\mu}_2, 1)$ 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


