Fair Algorithms for Infinite and Contextual Bandits

Matthew Joseph [*1] Michael Kearns [†1] Jamie Morgenstern [‡1] Seth Neel [§ 2] Aaron Roth [¶ 1]

[1]Computer and Information Science, University of Pennsylvania
[2]Statistics Department, The Wharton School, University of Pennsylvania

Abstract

We study fairness in linear bandit problems. Starting from the notion of meritocratic fairness introduced in Joseph et al. [2016], we carry out a more refined analysis of a more general problem, achieving better performance guarantees with fewer modelling assumptions on the number and structure of available choices as well as the number selected. We also analyze the previously-unstudied question of fairness in infinite linear bandit problems, obtaining instance-dependent regret upper bounds as well as lower bounds demonstrating that this instance-dependence is necessary. The result is a framework for meritocratic fairness in an online linear setting that is substantially more powerful, general, and realistic than the current state of the art.

1 Introduction

The problem of repeatedly making choices and learning from choice feedback arises in a variety of settings, including granting loans, serving ads, and hiring. Encoding these problems in a *bandit* setting enables one to take advantage of a rich body of existing bandit algorithms. UCB-style algorithms, for example, are guaranteed to yield no-regret policies for these problems.

Joseph et al. [2016], however, raises the concern that these no-regret policies may be *unfair*: in some rounds, they will choose options with lower expected rewards over options with higher expected rewards, for example choosing less qualified job applicants over more qualified ones. Consider a UCB-like algorithm aiming to hire all qualified applicants in every round. As time goes on, any no-regret algorithm must behave unfairly for a vanishing fraction of rounds, but the total number of *mistreated* people – in hiring, people who saw a less qualified job applicant hired in a round in which they themselves were not hired – can be large (see Figure 1).

Joseph et al. [2016] then design no-regret algorithms which minimize mistreatment and are fair in the following sense: their algorithms (with high probability) never at any round place higher selection probability on a less qualified applicant than on a more qualified applicant. However, their analysis assumes that there are k well-defined groups, each with its own mapping from features to expected rewards; at each round exactly one individual from each group arrives; and exactly one individual is chosen in each round. In the hiring setting, this equates to assuming that a company receives one job applicant from each group and must

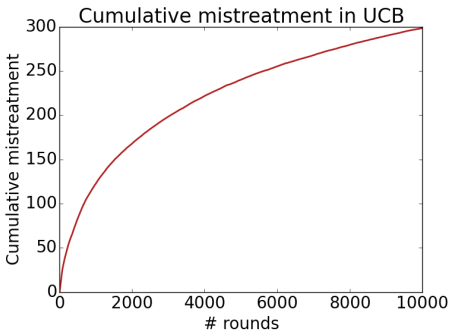

Figure 1: Cumulative mistreatments for UCB. See Section 7.7 in supplement for details and additional experimental evaluation of the structure of mistreatment.

[*]majos@cis.upenn.edu
[†]mkearns@cis.upenn.edu
[‡]jamiemor@cis.upenn.edu
[§]sethneel@wharton.upenn.edu
[¶]aaroth@cis.upenn.edu

hire exactly one (rather than m or all qualified applicants) introducing an unrealistic element of competition and unfairness both between applicants and between groups.

The aforementioned assumptions are unrealistic in many practical settings; our work shows they are also *unnecessary*. Meritocratic fairness can be defined without reference to groups, and algorithms can satisfy the strictest form of meritocratic fairness without any knowledge of group membership. Even without this knowledge, we design algorithms which will be fair with respect to *any* possible group structure over individuals. In Section 2, we present this general definition of fairness. The definition further allows for the number of individuals arriving in any round to vary, and is sufficiently flexible to apply to settings where algorithms can select $m \in [k]$ individuals in each round. By virtue of the definition making no reference to groups, the model makes no assumptions about how many individuals arriving at time t belong to any group. A company can then consider a large pool of applicants, not necessarily stratified by race or gender, with an arbitrary number of candidates from any one of these populations, and hire one or m or even every qualified applicant.

We then present a framework for designing meritocratically fair online linear contextual bandit algorithms. In Section 3, we show how to design fair algorithms when at most some finite number k of individuals arrives in any round (the linear contextual bandits problem [Abe et al., 2003, Auer, 2002]), as well as when m individuals may be chosen in each round (the "multiple play" introduced and studied absent fairness in Anantharam et al. [1987]). We therefore study a much more general model than [Joseph et al., 2016] and, in Section 3, substantially improve upon their black-box regret guarantees for linear bandit problems using a technical analysis specific to the linear setting.

However, these regret bounds still scale (polynomially) with k, the maximum number of individuals seen in any given round. This may be undesirable for large k, thus motivating the investigation of fair algorithms for the *infinite* bandit setting (the online linear optimization with bandit feedback problem Flaxman et al. [2005]).[1] In Section 4 we provide such an algorithm via an adaptation of our general confidence interval-based framework that takes advantage of the fact that optimal solutions to linear programs must be *extreme points* of the feasible region. We then prove, subject to certain assumptions, a regret upper bound that depends on Δ_{gap}, an instance-dependent parameter based on the distance between the best and second-best extreme points in a given choice set.

In Section 5 we show that this instance dependence is almost tight by exhibiting an infinite choice set satisfying our assumptions for which *any* fair algorithm must incur regret dependent polynomially on Δ_{gap}, separating this setting from the online linear optimization setting absent a fairness constraint. Finally, we justify our assumptions on the choice set by in Section 6 exhibiting a choice set that both violates our assumptions and admits *no* fair algorithm with nontrivial regret guarantees. A condensed presentation of our methods and results appears in Figure 2.

Finally, we note that our algorithms share an overarching logic for reasoning about fairness. These algorithms all satisfy fairness by *certifying optimality*, never giving preferential treatment to x over y unless the algorithm is *certain* that x has higher reward than y. The algorithms accomplish this by computing confidence intervals around the estimated rewards for individuals. If two individuals have overlapping confidence intervals, we say they are *linked*; if x can be reached from y using a sequence of linked individuals, we say they are *chained*.

1.1 Related Work and Discussion of Our Fairness Definition

Fairness in machine learning has seen substantial recent growth as a subject of study, and many different definitions of fairness exist. We provide a brief overview here; see e.g. Berk et al. [2017] and Corbett-Davies et al. [2017] for detailed descriptions and comparisons of these definitions.

Many extant fairness notions are predicated on the existence of *groups*, and aim to guarantee that certain groups are not unequally favored or mistreated. In this vein, Hardt et al. [2016] introduced the notion of *equality of opportunity*, which requires that a classifier's predicted outcome should be independent of a

[1] We note that both the finite and infinite settings have infinite numbers of potential candidates: the difference arises in how many choices an algorithm has in a given round.

# selected each round	# options each round	Technique	Notes	Regret
Exactly $j \leq k$	$\leq k$	Play all of chains in descending order, randomizing over last chain as necessary to pick exactly j	Requires randomness	$\tilde{O}\left(dkj\sqrt{T}\right)$
Unconstrained	$\leq k$	Select all in every chain with highest UCB > 0	Deterministic	$\tilde{O}\left(dk^2\sqrt{T}\right)$
Exactly 1	∞ bounded convex set $\Delta_{\text{gap}} > 0$	Play uniquely best point or UAR from entire set	Requires randomness	$\tilde{O}\left(c \cdot \log(T)/\Delta_{\text{gap}}^2\right)$ $\tilde{\Omega}(1/\Delta_{\text{gap}})$ $\Omega(T)$ for $\Delta_{\text{gap}} = 0$

Figure 2: A description of various settings in which our framework provides fair algorithms. In all cases, fairness can be imposed only across pairs for any partitioning of the input space; the bounds here assume they bind across all pairs, and are therefore worst-case upper bounds. See Section 4 for a complete explanation of the distribution-dependent constant c in the regret bound for the infinite case.

protected attribute (such as race) conditioned on the true outcome, and they and Woodworth et al. [2017] have studied the feasibility and possible relaxations thereof. Similarly, Zafar et al. [2017] analyzed an equivalent concurrent notion of (un)fairness they call *disparate mistreatment*. Separately, Kleinberg et al. [2017] and Chouldechova [2017] showed that different notions of group fairness may (and sometimes must) conflict with one another.

This paper, like Joseph et al. [2016], departs from the work above in a number of ways. We attempt to capture a particular notion of *individual* and *weakly meritocratic* fairness that holds *throughout the learning process*. This was inspired by Dwork et al. [2012], who suggest fair treatment equates to treating "similar" people similarly, where similarity is defined with respect to an assumed pre-specified task-specific metric. Taking the fairness formulation of Joseph et al. [2016] as our starting point, our definition of fairness does not promise to correct for past inequities or inaccurate or biased data. Instead, it assumes the existence of an accurate mapping from features to true quality for the task at hand[2] and promises fairness while learning and using this mapping in the following sense: any *individual* who is currently more qualified (for a job, loan, or college acceptance) than another individual will always have at least as good a chance of selection as the less qualified individual.

The one-sided nature of this guarantee, as well as its formulation in terms of quality, leads to the name *weakly meritocratic* fairness. Weakly meritocratic fairness may then be interpreted as a minimal guarantee of fairness: an algorithm satisfying our fairness definition cannot favor a worse option but is not required to favor a better option. In this sense our fairness requirement encodes a necessary variant of fairness rather than a completely sufficient one. This makes our upper bounds (Sections 3 and 4) relatively weaker and our lower bounds (Sections 5 and 6) relatively stronger.

We additionally note that our fairness guarantees require fairness *at every step of the learning process*. We view this as an important point, especially for algorithms whose learning processes may be long (or even continuous). Furthermore, while it may seem reasonable to relax this requirement to allow a small fraction of unfair steps, it is unclear how to do so without enabling discrimination against a correspondingly small population.

Finally, while our fairness definition draws from Joseph et al. [2016], we work in what we believe to be a

[2] Friedler et al. [2016] provide evidence that providing fairness from bias-corrupted data is quite difficult.

significantly more general and realistic setting. In the finite case we allow for a variable number of individuals in each round from a variable number of groups and also allow selection of a variable number of individuals in each round, thus dropping several assumptions from Joseph et al. [2016]. We also analyze the previously unstudied topic of fairness with infinitely many choices.

2 Model

Fix some $\beta \in [-1,1]^d$, the underlying linear coefficients of our learning problem, and T the number of rounds. For each $t \in [T]$, let $C_t \subseteq D = [-1,1]^d$ denote the set of available choices in round t. We will consider both the "finite" action case, where $|C_t| \leq k$, and the infinite action case. An algorithm $\mathcal{A}$, facing choices C_t, picks a subset $P_t \subseteq C_t$, and for each $x_t \in P_t$, $\mathcal{A}$ observes reward $y_t \in [-1,1]$ such that $\mathbb{E}[y_t] = \langle \beta, x_t \rangle$, and the distribution of the noise $\eta_t = y_t - \langle \beta, x_t \rangle$ is sub-Gaussian (see Section 7.1 for a definition of sub-Gaussian).

Refer to all observations in round t as $Y_t \in [-1,1]^{|P_t|}$ where $Y_{t,i} = y_{t,i}$ for each $x_{t,i} \in P_t$. Finally, let $\mathbf{X}_t = [X_1; \ldots; X_t], \mathbf{Y}_t = [Y_1; \ldots; Y_t]$ refer to the design and observation matrices at round t.

We are interested in settings where an algorithm may face size constraints on P_t. We consider three cases: the standard linear bandits problem ($|P_t| = 1$), the multiple choice linear bandits problem ($|P_t| = m$), and the heretofore unstudied (to the best of the authors' knowledge) case in which the size of P_t is unconstrained. For short, we refer to these as 1-bandit, m-bandit, and k-bandit.

Regret The notion of regret we will consider is that of pseudo-regret. Facing a sequence of choice sets $C_1, \ldots, C_T$, suppose $\mathcal{A}$ chooses sets $P_1, \ldots, P_T$.[3] Then, the expected reward of $\mathcal{A}$ on this sequence is $\text{Rew}(\mathcal{A}) = \mathbb{E}\left[\sum_{t \in [T]} \left[\sum_{x_t \in P_t} y_t\right]\right]$.

Refer to the sequence of feasible choices[4] which maximizes expected reward as $P_{*,1} \subseteq C_1, \ldots, P_{*,T} \subseteq C_T$, defined with full knowledge of β.

Then, the **pseudo-regret** of $\mathcal{A}$ on any sequence is defined as

$$\text{Rew}(P_{*,1}, \ldots, P_{*,T}) - \text{Rew}(\mathcal{A}) = R(T).$$

The **pseudo-regret** of $\mathcal{A}$ refers to the maximum pseudo-regret $\mathcal{A}$ incurs on any sequence of choice sets and any $\beta \in [-1,1]^d$. If $R(T) = o(T)$, then $\mathcal{A}$ is said to be **no-regret**. If, for any input parameter $\delta > 0$, $R(T)$ upper-bounds the expectation of the rewards of the sequence chosen by $\mathcal{A}$ with probability $1 - \delta$, then we call this a *high-probability* regret bound for $\mathcal{A}$.

Fairness Consider an algorithm $\mathcal{A}$, which chooses a sequence of *probability distributions* $\pi_1, \pi_2, \ldots, \pi_T$ over feasible sets to pick, $\pi_t \in \Delta(2^{C_t})$. Note that distribution π_t depends upon $C_1, \ldots, C_t$, the choices $P_1, \ldots, P_{t-1}$, and $Y_1, \ldots, Y_{t-1}$.

We now give a formal definition of fairness of an algorithm for the 1-bandit, m-bandit, and k-bandit problems. We adapt our fairness definition from Joseph et al. [2016], generalizing from discrete distributions over finite action sets to mixture distributions over possibly infinite action sets. We slightly abuse notation and refer to the probability density and mass functions of an element $x \in C_t$: this refers to the marginal distribution of x being chosen (namely, the probability that x belongs to the set picked according to the distribution π_t).

Definition 1 (Weakly Meritocratic Fairness)**.** We say that an algorithm $\mathcal{A}$ is *weakly meritocratic* if, for any input $\delta \in (0,1]$ and any β, with probability at least $1 - \delta$, at every round t, for every $x, x' \in C_t$ such that $\langle \beta, x \rangle \geq \langle \beta, x' \rangle$:

[3] If these are randomized choices, the randomness of $\mathcal{A}$ is incorporated into the expected value calculations.

[4] We assume these have the appropriate size for each problem we consider: singletons in the 1-bandit problem, size at most m in the m-bandit problem, and arbitrarily large in the k-bandit problem.

- If π_t is a discrete distribution: For $g_t(x) = \pi_t(x)$ (the probability mass function)

$$g_t(x) \geq g_t(x').$$

- If π_t is a continuous distribution: For $g_t(x) = f_t(x)$ (the probability density function)

$$g_t(x) \geq g_t(x').$$

- If π_t can be written as a mixture distribution: $\sum_i \alpha_i \pi_{ti}, \sum_i \alpha_i = 1$, such that each constituent distribution $\pi_{ti} \in \Delta(2^{C_t})$ is either discrete or continuous and satisfies one of the above two conditions.

For brevity, as consider only this fairness notion in this paper, we will refer to weakly meritocratic fairness as "fairness". We say $\mathcal{A}$ is **round-fair** at time t if π_t satisfies the above conditions.

This definition can be easily generalized over any partition $\mathcal{G}$ of D, by requiring this weak monotonicity hold *only for pairs x, x' belonging to different elements of the partition G, G'*. The special case above of the singleton partition is the most stringent choice of partition. We focus our analysis on the singleton partition as a minimal worst-case framework, but this model easily relaxes to apply only across groups, as well as to only requiring "one-sided" monotonicity, where monotonicity is required only for pairs where the more qualified member belongs to group G rather than G'.

Remark 1. In the k-bandit setting, Definition 1 can be simplified to require, with probability $1 - \delta$ over its observations, an algorithm *never* select a less-qualified individual over more-qualified one in any round, and can be satisfied by deterministic algorithms.

3 Finite Action Spaces: Fair Ridge Regression

In this section, we introduce a family of fair algorithms for linear 1-bandit, m-bandit, and the (unconstrained) k-bandit problems. Here, an algorithm sees a slate of at most k distinct individuals each round and selects some subset of them for reward and observation. This allows us to encode settings where an algorithm repeatedly observes a new pool of k individuals, each represented by a vector of d features, then decides to give some of those individuals loans based upon those vectors, observes the quality of the individuals to whom they gave loans, and updates the selection rule for loan allocation. The regret of these algorithms will scale polynomially in k and d as the algorithm gets tighter estimates of β.

All of the algorithms are based upon the following template. They maintain an estimate $\hat{\beta}_t$ of β from observations, along with confidence intervals around the estimate. They use $\hat{\beta}_t$ to estimate the rewards for the individuals on day t and the confidence interval around $\hat{\beta}_t$ to create a confidence interval around each of these estimated rewards. Any two individuals whose intervals overlap on day t will picked with the same probability by the algorithm. Call any two individuals whose intervals overlap on day t *linked*, and any two individuals belonging to the transitive closure of the linked relation *chained*. Since any two linked individuals will chosen with the same probability, any two chained individuals will also be chosen with the same probability.

An algorithm constrained to pick exactly $m \in [k]$ individuals each round will pick them in the following way. Order the chains by their highest upper confidence bound. In that order, select all individuals from each chain (with probability 1 while that results in taking fewer than m individuals. When the algorithm arrives at the first chain for which it does not have capacity to accept every individual in the chain, it selects to fill its capacity uniformly at random from that chain's individuals. If the algorithm can pick any number of individuals, it will pick all individuals chained to any individual with positive upper confidence bound.

We now present the regret guarantees for fair 1-bandit, m-bandit, and k-bandit using this framework.

Theorem 1. *Suppose, for all t, η_t is 1-sub-Gaussian, $C_t \subseteq [-1, 1]^d$, and $||x_t||_2 \leq 1$ for all $x_t \in C_t$, and $||\beta|| \leq 1$. Then,* RidgeFair$_1$, RidgeFair$_m$, *and* RidgeFair$_{\leq k}$ *are fair algorithms for the 1-bandit, m-bandit, and k-bandit problems, respectively. With probability $1 - \delta$, for $j \in \{1, m, k\}$, the regret of* RidgeFair$_j$ *is*

$$R(T) = O\left(dkj\sqrt{T}\log\left(\frac{T}{\delta}\right)\right) = \tilde{O}(dkj\sqrt{T}).$$

We pause to compare our bound for 1-bandit to that found in Joseph et al. [2016]. Their work supposes that each of k groups has an independent d-dimensional linear function governing its reward and provides a fair algorithm regret upper bound of $\tilde{O}\left(T^{\frac{4}{5}}k^{\frac{6}{5}}d^{\frac{3}{5}}, k^3\right)$. To directly encode this setting in ours, one would need to use a single dk-dimensional linear function, yielding a regret bound of $\tilde{O}\left(dk^2\sqrt{T}\right)$. This is an improvement on their upper bound for all values of T for which the bounds are non-trivial (recalling that the bound from Joseph et al. [2016] becomes nontrivial for $T > d^3k^6$, while the bound here becomes nontrivial for $T > d^2k^4$). We also briefly observe that RidgeFair$_{\leq k}$ satisfies an additional "fairness" property: with high probability, it always selects *every* available individual with positive expected reward.

Each of these algorithms will use ℓ_2-regularized least-squares regressor to estimate β. Given a design matrix $\mathbf{X}$, response vector $\mathbf{Y}$, and regularization parameter $\gamma \geq 1$ this is of the form $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \gamma I)^{-1}\mathbf{X}^T\mathbf{Y}$. Valid confidence intervals (that contain β with high probability) are nontrivial to derive for this estimator (which might be biased); to construct them, we rely on martingale matrix concentration results [Abbasi-Yadkori et al., 2011].

We now sketch how the proof of Theorem 1 proceeds, deferring a full proof (of this and all other results in this paper) and pseudocode to the supplementary materials. We first establish that, with probability $1 - \delta$, for all rounds t, for all $x_{t,i} \in C_t$, that $y_{t,i} \in [\ell_{t,i}, u_{t,i}]$ (i.e. that the confidence intervals being used are valid). Using this fact, we establish that the algorithm is fair. The algorithm plays any two actions which are linked with equal probability in each round, and any action with a confidence interval above another action's confidence interval with weakly higher probability. Thus, if the payoffs for the actions lie anywhere within their confidence intervals, RidgeFair is fair, which holds as the confidence intervals are valid.

Proving a bound on the regret of RidgeFair requires some non-standard analysis, primarily because the widths of the confidence intervals used by the algorithm do not shrink uniformly. The sum of the widths of the intervals of our *selected* (and therefore observed) actions grows sublinearly in t. UCB variants, by virtue of playing an action a with highest upper confidence bound, have regret in round t bounded by a's confidence interval width. RidgeFair, conversely, suffers regret equal to the *sum* of the confidence widths of the chained set, while only receiving feedback for the action it actually takes. We overcome this obstacle by relating the sum of the confidence interval widths of the linked set to the sum of the widths of the selected actions.

4 Fair algorithms for convex action sets

In this section we analyze linear bandits with infinite choice sets in the familiar 1-bandit setting.[5] We provide a fair algorithm with an instance-dependent sublinear regret bound for infinite choice sets – specifically convex bodies – below. In Section 5 we match this with lower bounds showing that instance dependence is an unavoidable cost for fair algorithms in an infinite setting.

A naive adaptation of RidgeFair to an infinite setting requires maintenance of infinitely many confidence intervals and is therefore impractical. We instead assume that our choice sets are convex bodies and exploit the resulting geometry: since our underlying function is linear, it is maximized at an *extremal* point. This simplifies the problem, since we need only reason about the relative quality of extremal points. The relevant quantity is Δ_{gap}, a notion adapted from Dani et al. [2008] that denotes the difference in reward between the best and second-best extremal points in the choice set. When Δ_{gap} is large it is easier to confidently identify the optimal choice and select it deterministically without violating fairness. When Δ_{gap} is small, it is more difficult to determine which of the top two points is best – and since deterministically selecting the wrong one violates fairness for any points infinitesimally close to the true best point, we are forced to play randomly from the entire choice set.

Our resulting fair algorithm, FairGap, proceeds as follows: in each round it uses its current estimate of β to construct confidence intervals around the two choices with highest estimated reward and selects the higher one if these intervals do not overlap; otherwise, it selects uniformly at random from the entire convex

[5]Note that no-regret guarantees are in general impossible for infinite choice sets in m-bandit and k-bandit settings, since the continuity of the infinite choice sets we consider makes selecting multiple choices while satisfying fairness impossible without choosing uniformly at random from the entire set.

body. We prove fairness and bound regret by analyzing the rate at which random exploration shrinks our confidence intervals and relating it to the frequency of exploitation, a function of Δ_{gap}. We begin by formally defining Δ_{gap} below.

Definition 2 (Gap, adapted from Dani et al. [2008])**.** Given sequence of action sets $C = (C_1, \ldots, C_T)$, define Ω_t to be the set of extremal points of C_t, i.e. the points in C_t that cannot be expressed as a proper convex combination of other points in C_t, and let $x_t^* = \max_{x \in C_t} \langle \beta, x \rangle$. The *gap* of C_t is

$$\Delta_{\text{gap}} = \min_{1 \le t \le T} \left(\inf_{x_t \in \Omega_t, x_t \neq x_t^*} \langle \beta, x_t^* - x_t \rangle \right)$$

the minimum of <beta, xt_star - xt> when xt is chosen from the set of optimal points but does not equal to xt_star (xt is the second best action)

Δ_{gap} is a lower bound on difference in payoff between the optimal action and any other extremal action in any C_t. When $\Delta_{\text{gap}} > 0$, this implies the existence of a unique optimal action in each C_t. Our algorithm (implicitly) and our analysis (explicitly) exploits this quantity: a larger gap enables us to confidently identify the optimal action more quickly.

We now present the regret and fairness guarantees for FairGap.

Theorem 2. *Given sequence of action sets $C = (C_1, \ldots, C_T)$ where each C_t has nonzero Lebesgue measure and is contained in a ball of radius r and feedback with R-sub-Gaussian noise,* FairGap *is fair and achieves*

$$\textsc{Regret}(T) = O\left(\frac{r^6 R^2 \ln(2T/\delta)}{\kappa^2 \lambda^2 \Delta_{gap}^2} \right)$$

where $\kappa = 1 - r\sqrt{\frac{2\ln\left(\frac{2dT}{\delta}\right)}{T\lambda}}$ *and* $\lambda = \min_{1 \le t \le T} \left[\lambda_{\min}(\mathbb{E}_{x_t \sim UAR C_t}[{x_t}^T x_t]) \right]$

A full proof of FairGap's fairness and regret bound, as well as pseudocode, appears in the supplement. We sketch the proof here: our proof of fairness proceeds by bounding the influence of noise on the confidence intervals we construct (via matrix Chernoff bounds) and proving that, with high probability, FairGap constructs correct confidence intervals. This requires reasoning about the spectrum of the covariance matrix of each choice set, which is governed by λ, a quantity which, informally, measures how quickly we learn from uniformly random actions. [6]. With correct confidence intervals in hand, fairness follows almost immediately, and to bound regret we analyze the rate at which these confidence intervals shrink.

The analysis above implies identical regret and fairness guarantees when each C_t is finite. For comparison, the results of Section 3 guarantee $\textsc{Regret}(T) = O(dk\sqrt{T})$. This result, in comparison, enjoys a regret independent of k which may prove especially useful for cases involving large k.

Finally, our analysis so far has elided any computational efficiency issues arising from sampling randomly from C. We note that it is possible to circumvent this issue by relaxing our definition of fairness to *approximate fairness* and obtain similar regret bounds for an efficient implementation. We achieve this using results from the broad literature on sampling and estimating volume in convex bodies, as well as recent work on finding "2nd best" extremal solutions to linear programs. Full details appear in Section 7.4 of the Supplement.

5 Instance-dependent Lower Bound for Fair Algorithms

We now present a lower bound instance for which any fair algorithm *must* suffer gap-dependent regret. More formally, we show that when each choice set is a square, i.e. $C_t = [0, 1]^2$ for all t, for any fair algorithm $\textsc{Regret}(T) = \tilde{\Omega}(1/\Delta_{\text{gap}})$ with probability at least $1 - \delta$. This also implies the weaker result that no fair algorithm enjoys an instance-independent sub-linear regret bound $o(T)$ holding uniformly over all β. We therefore obtain a clear separation between fair learning and the unconstrained case Dani et al. [2008], and show that an instance-dependent upper bound like the one in Section 4 is unavoidable. Our arguments establish fundamental constraints on fair learning with large choice sets and quantify through the Δ_{gap}

[6] λ can be computed directly for finite C_t or approximated by any positive lower bound for infinite C_t and substituted directly into our results.

parameter how choice set geometry can affect the performance of fair algorithms. The lower bound employs a Bayesian argument resembling that in Joseph et al. [2016] but with a novel "chaining" argument suited to infinite action sets. We present the result for $d = 2$ for simplicity; the proof technique holds in any dimension $d \geq 2$.

Theorem 3. *For all t let $C_t = [-1, 1]^d$, $\beta \in [-1, 1]^d$, and $y_t = \langle x_t, \beta \rangle + \eta_t$, where $\eta_t \sim U[-1, 1]$. Let $\mathcal{A}$ be any fair algorithm. Then for every gap Δ_{gap}, there is a distribution over instances with gap $\Omega(\Delta_{gap})$ such that any fair algorithm has regret $\textsc{Regret}(T) = \tilde{\Omega}(1/\Delta_{gap})$ with probability $1 - \delta$.*

We a sketch of the central ideas in the proof, relegating a full proof to the Supplement. We start with the fact that any fair algorithm $\mathcal{A}$ is required to be fair for any value β of the linear parameter. Thus if we draw $\beta \sim \tau$, $\mathcal{A}$ must be round-fair for all $t \geq 1$ with probability at least $1 - \delta$, where now the probability includes the random draw $\beta \sim \tau$. Then Bayes' rule implies that the procedure that draws $\beta \sim \tau$ and then plays according to $\mathcal{A}$ is identical to the procedure which at each step t re-draws β from its posterior distribution given the past $\tau|_{h_t}$.

Next, given the prior τ, $\mathcal{A}$'s round fairness at step t requires that (with high probability) if $\mathcal{A}$ plays action x with higher probability than action y, we must have

$$\mathbb{P}_{\beta \sim \tau|h_t} \left[\langle \beta, x \rangle > \langle \beta, y \rangle \right] > \frac{3}{4}. \tag{1}$$

This enables us to reason about the fairness and regret of the algorithm via a specific analysis of the posterior distribution $\tau|_{h_t}$. We formalize this argument in Lemmas 7 and 8. This Bayesian trick, first applied in Joseph et al. [2016], is a general technique useful for proving fairness lower bounds.

We then show that for a choice of prior specific to our choice set C, that two things hold: (i) whenever $\tau|_{h_t} = \tau$, Equation 1 forces $\mathcal{A}$ to play uniformly from C, and (ii) with high probability $\tau = \tau|_{h_t}$ until $t > \tilde{\Omega}(1/\epsilon)$, where ϵ is a parameter of the prior that acts as a proxy for Δ_{gap}. Playing an action uniformly from C incurs $\Omega(1)$ regret per round, so these two facts combine to show that with high probability $\textsc{Regret}(T) = \tilde{\Omega}(1/\epsilon)$.

Finally we consider $\textsc{Regret}(T)$ conditional on the event that $\Delta_{\text{gap}}(\beta) > \delta \cdot \epsilon$, which by our construction of τ happens with probability $1 - \delta$. Let τ_{gap} be the conditional distribution of β given that $\Delta_{\text{gap}}(\beta) > \delta \cdot \epsilon$. Then

$$\mathbb{P}_{\beta \sim \tau} \left[\textsc{Regret}(T) \geq \Omega\left(\frac{1}{\epsilon}\right) \right] \leq \mathbb{P}_{\beta \sim \tau_{gap}} \left[\textsc{Regret}(T) \geq \Omega\left(\frac{1}{\epsilon}\right) \right] (1 - \delta) + \delta$$

which implies

$$\mathbb{P}_{\beta \sim \tau_{gap}} \left[\textsc{Regret}(T) \geq \Omega\left(\frac{1}{\epsilon}\right) \right] \geq \frac{1 - 2\delta}{1 - \delta}.$$

Note that $\frac{1-2\delta}{1-\delta} \to 1$ as $\delta \to 0$, and so this is a high-probability bound. Since for every β in the support of τ_{gap}, we have that $\Delta_{\text{gap}}(\beta) \geq \delta \cdot \epsilon$, we've exhibited a distribution τ_{gap} such that when $\beta \sim \tau_{gap}$, with high probability, $\textsc{Regret}(T) = \tilde{\Omega}(1/\epsilon) = \tilde{\Omega}(1/\Delta_{\text{gap}})$, as desired.

The proof uses the fact that when $\tau = \tau|_{h_t}$, Equation 1 forces $\mathcal{A}$ to play uniformly at random. This happens by transitivity: if Equation 1 forces $\mathcal{A}$ to play x equiprobably with y and y equiprobably with z, then x must be played equiprobably with z. The fact that any two actions in C can be connected via such a (finite) transitive chain is illustrated in Figure 7.5 and formalized in Lemma 10.

Remark 2. We note that this impossibility result only holds for $d \geq 2$. When $d = 1$, the choice set reduces to $[-1, 1]$, and similarly $\beta \in [-1, 1]$. Thus, the optimal action is $\text{sign}(\beta)$). It takes $O(1/\beta^2)$ observations to determine the sign of β, so a simple fair algorithm may play randomly from $[-1, 1]$ until it has determined $\text{sign}(\beta)$, and then play $\text{sign}(\beta)$ for every following round. Because the maximum per-round regret of any action is $O(\beta)$, and because the maximum cumulative regret obtained by the algorithm is with high probability $O(\beta \cdot 1/\beta^2) = O(1/\beta)$, the regret of this simple algorithm over T rounds is $O(\min(\beta \cdot T, 1/\beta^2))$. Taking the worst case over β, we see that this quantity is bounded uniformly by $O(\sqrt{T})$, a sublinear parameter independent regret bound.

6 Zero Gap: Impossibility Result

Section 4 presents an algorithm for which the sublinear regret bound has dependence $1/\Delta_{\text{gap}}^2$ on the instance gap. Section 5 exhibits an choice set C with a $\tilde{\Omega}(1/\Delta_{\text{gap}})$ dependence on the gap parameter. We now exhibit a choice set C for which $\Delta_{\text{gap}} = 0$ for every β, and for which no fair algorithm can obtain non-trivial regret for any value of β. This precludes even instance-dependent fair regret bounds on this action space, in sharp contrast with the unconstrained bandit setting.

Theorem 4. *For all t let $C_t = S^1$, the unit circle, and $\eta_t \sim$ Unif$(-1, 1)$. Then for any fair algorithm $\mathcal{A}$, $\forall \beta \in S^1, \forall T \geq 1$, we have*

$$\mathbb{E}_\beta[\textsc{Regret}\,(T)] = \Omega(T).$$

S^1 makes fair learning difficult for the following reasons: since S^1 has no extremal points, there is no finite set of points which for any β contains the uniquely optimal action, and for any point in S^1, and any finite set of observations, there is another point in S^1 for which the algorithm cannot confidently determine relative reward. Since this property holds for *every* point, the fairness constraint transitively requires that the algorithm play every point uniformly at random, at every round.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays – part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, AC-32(Nov):968–976, 1987.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov): 397–422, 2002.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230*, 2017.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of ITCS 2012*, pages 214–226. ACM, 2012.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. In *arXiv*, volume abs/1609.07236, 2016. URL `http://arxiv.org/abs/1609.07236`.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, volume abs/1610.02413, 2016. URL `http://arxiv.org/abs/1610.02413`.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, Jan 2017.

Erik M Lindgren, Alexandros G Dimakis, and Adam Klivans. Facet guessing for finding the m-best integral solutions of a linear program. In *NIPS Workshop on Optimization for Machine Learning*, 2016.

László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8 (1-2):1–230, 2015.

Santosh Vempala. Geometric random walks: a survey. *Combinatorial and computational geometry*, 52(573-612):2, 2005.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of World Wide Web Conference*, 2017.

7 Appendix

7.1 Sub-Gaussian Definition

Sub-Gaussian random variables have moment generating functions bounded by the Gaussian moment generating function, and hence can be controlled via Chernoff bounds.

Definition 3. A random variable X with $\mu = \mathbb{E}[X]$ is $R > 0$ *sub-Gaussian* if, for all $t \in \mathbb{R}$, $\mathbb{E}\left[e^{t(X-\mu)}\right] \leq e^{Rt^2/2}$.

7.2 Proofs from Section 3

We start with full pseudocode for RidgeFair$_m$.

Proof of Theorem 1. We first claim that confidence intervals are valid: that with probability $1-\delta$, for all $t \in [T]$ and all $x_{t,i} \in C_t$, $y_{t,i} \in [\ell_{t,i}, u_{t,i}]$. Assuming this claim, we prove that RidgeFair$_m$ is fair. With probability $1-\delta$, for all rounds t and all individuals $x_{t,i}$, $y_{t,i} \in [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$. So, for any pair of individuals $x_{t,i}, x_{t,j} \in C_t$, if $y_{t,i} > y_{t,j}$, then $\hat{y}_{t,i} + w_{t,i} \geq \hat{y}_{t,j} - w_{t,j}$. So, if j belongs to some chain from which arms are selected, either i belongs to a higher chain or the same chain as j. Every individual belonging to a higher chain is played with weakly higher probability to any individual belonging to a lower chain, and every two individuals belonging to the same chain are played with equal probability, so i is played with weakly higher probability than j. Thus, at all rounds and for all pairs of individuals, the fairness constraint is satisfied by this distribution over P_t, and so RidgeFair$_m$ is fair.

We now prove the confidence intervals are valid: that with probability $1-\delta$, for all $t \in [T]$ and all $x_{t,i} \in C_t$, $y_{t,i} \in [\ell_{t,i}, u_{t,i}]$. We adopt the notation in Abbasi-Yadkori et al. [2011]: let $\bar{V}_t = \mathbf{X}_t{}^T\mathbf{X}_t + \gamma I$, where $\mathbf{X}_t$ is the design matrix at time t. Let $\hat{\beta}_t = (\bar{V}_t)^{-1}\mathbf{X}_t{}^T\mathbf{Y}_t$ be the regularized least squares estimator at time t.

Consider a feature vector $x_{t,i}$ at time t. For a d-dimensional vector z and a $d \times d$ positive definite matrix A, let $\langle z_1, z_2 \rangle_A$ denote $z_1^t A z_2$. Let η_t be the noise sequence prior to round t. Then, we have $\hat{\beta}_t = (\bar{V}_t)^{-1}\mathbf{X}_t{}^T(\mathbf{X}_t\beta + \eta_t)$. Then some matrix algebra in the proof of Theorem 2 of Abbasi-Yadkori et al. [2011] shows

$$x_{t,i} \cdot (\hat{\beta}_t - \beta) = x_{t,i}{}^T(\bar{V}_t)^{-1}\mathbf{X}_t{}^T\eta_t - \gamma x_{t,i}{}^T(\bar{V}_t)^{-1}\beta,$$

which using the above notation gives

$$x_{t,i} \cdot (\hat{\beta}_t - \beta) = \langle x_{t,i}, \mathbf{X}_t{}^T\eta_t \rangle_{(\bar{V}_t)^{-1}} - \gamma \langle x_{t,i}, \beta \rangle_{(\bar{V}_t)^{-1}}$$

Applying Cauchy-Schwarz,

$$|x_{t,i} \cdot \left(\hat{\beta}_t - \beta\right)| \leq ||x_{t,i}||_{(\bar{V}_t)^{-1}}(||\mathbf{X}_t{}^T\eta_t||_{(\bar{V}_t)^{-1}} + \sqrt{\gamma})$$

which follows from the fact that $||\beta||_{(\bar{V}_t)^{-1}} \leq \frac{1}{\sqrt{\gamma}}$ (a basic corollary of the Rayleigh quotient, and the fact that by assumption $||\beta|| \leq 1$.

We now present a result derived from [Abbasi-Yadkori et al., 2011] that will help us upper bound this quantity. The upper bound on $||\mathbf{X}^T\eta||$ at the bottom of page 13 of Abbasi-Yadkori et al. [2011] and the upper bound on $\log(\det(\bar{V}_t))$ at the top of page 15, combined with our assumption that our noise is 1-sub-Gaussian, implies that

$$
\begin{aligned}
||\mathbf{X}_t{}^T \eta_t||_{(\bar{V}_t)^{-1}} &\leq \sqrt{d\log\left(1+t/d\gamma\right) - 2\log\delta} \\
&= \sqrt{d\log\left(1+t/\gamma\right) + 2\log\frac{1}{\delta}} \\
&\leq \sqrt{2d\log\left(1+t/\gamma\right) + 2d\log\frac{1}{\delta}} \\
&= \sqrt{2d\log\left(\frac{1}{\delta}\left(1+t/d\gamma\right)\right)} \\
&\leq \sqrt{2d\log\left(\frac{1}{\delta}\left(1+t/\gamma\right)\right)} \\
&= \sqrt{2d\log\left(\frac{1+t/\gamma}{\delta}\right)}.
\end{aligned}
$$

Using this result and combining the inequalities we get that over all rounds $t \geq 0$ with probability $1-\delta$

$$
|x_{t,i} \cdot \left(\hat{\beta}_t - \beta\right)| \leq ||x_{t,i}||_{(\bar{V}_t)^{-1}} \left(\sqrt{2d\log\left(\frac{1+t/\gamma}{\delta}\right)} + \sqrt{\gamma}\right) \tag{2}
$$

and therefore the claim that the confidence intervals are valid holds.

Regret bound for RidgeFair$_1$ We now proceed with upper-bounding the regret of RidgeFair$_1$. With probability $1-\delta$, the confidence intervals are valid. We will condition on that event for the analysis of the regret of the algorithm, since this regret bound will hold with high probability (namely, with probability $1-\delta$).

We start with a bound that will be useful for analyzing the width of our confidence intervals. The top of page 15 of of [Abbasi-Yadkori et al., 2011] notes that $\log\det(\bar{V}_t) \leq d\log(\gamma + t/d)$, and we combine this with the fact that $\sum_{t=1}^{T} ||x_{t,i}||^2_{\bar{V}_t} \leq 2\log\det(\bar{V}^T) - 2\log\det(V)$ (proven as part of Lemma 11 in [Abbasi-Yadkori et al., 2011]) to get that

$$
\sum_{t=1}^{T} ||x_{t,i}||^2_{(\bar{V}_t)^{-1}} \leq 2d\log\left(1 + \frac{T}{d\gamma}\right). \tag{3}
$$

We now have all the tools needed to analyze the algorithm's regret. First note that the choice of the algorithm is a singleton, i.e. that $P_t = \{\hat{i}_t\}$, for some $\hat{i}_t \in S_t$, the active chained set. Further, since the confidence intervals are valid, $i_{*,t} \in S_t$ for the best action $i_{*,t} \in C_t$. By the definition of S_t, the instantaneous regret $r_{t,i}$ for any $i \in S_t$ is at most $r_{t,i} \leq \sum_{j \in S_t} w_{t,j}$ (as any $i \in S_t$ is chained to some other arm in S_t). So, we have that

$$
\begin{aligned}
R(T) &\leq \sum_t r_{t,\hat{i}_t} \\
&\leq \sum_t \sum_{j \in S_t} 2w_{t,j} && \text{Conditioning on w.p. } 1-\delta \text{ valid confidence intervals} \\
&= 2\sum_t |S_t| \cdot \mathbb{E}\left[w_{t,\hat{i}_t}\right] && \text{When uniformly selecting } \hat{i}_t \in S_t\text{; note this holds w.p. 1 conditioned on valid CI} \\
&\leq 2k \cdot \sum_t \mathbb{E}\left[w_{t,\hat{i}_t}\right] \\
&= 2k \cdot \mathbb{E}\left[\sum_t w_{t,\hat{i}_t}\right] && \text{By linearity of expectation} \\
&= 2k \cdot \mathbb{E}\left[\sum_t ||x_{t,\hat{i}_t}||_{(\bar{V}_t)^{-1}} \left(\sqrt{2d\log\left(\frac{1+t/\gamma}{\delta}\right)} + \sqrt{\gamma}\right)\right] && \text{By definition of } w_{t,i} \\
&= 2k \cdot \mathbb{E}\left[\sum_t ||x_{t,\hat{i}_t}||_{(\bar{V}_t)^{-1}} \left(\sqrt{2d\log\left(\frac{1+t/\gamma}{\delta}\right)}\right) + \sum_t ||x_{t,\hat{i}_t}||_{(\bar{V}_t)^{-1}} \left(\sqrt{\gamma}\right)\right] \\
&\leq 2k\mathbb{E}\left[\sqrt{\sum_t ||x_{t,\hat{i}_t}||^2_{(\bar{V}_t)^{-1}}} \cdot \left(\sqrt{\sum_t 2d\log\left(\frac{1+t/\gamma}{\delta}\right)} + \sqrt{\sum_t \gamma}\right)\right] && \text{By Cauchy-Schwartz} \\
&\leq 2k\sqrt{2d\log\left(1+\frac{T}{d\gamma}\right)} \cdot \left(\sqrt{\sum_t 2d\log\left(\frac{1+t/\gamma}{\delta}\right)} + \sqrt{\sum_t \gamma}\right) && \text{By Equation 3} \\
&\leq 2k\sqrt{2d\log\left(1+\frac{T}{d\gamma}\right)} \cdot \left(\sqrt{2dT\log\left(\frac{1+T/\gamma}{\delta}\right)} + \sqrt{T\gamma}\right)
\end{aligned}
$$

or $R(T) = O\left(dk\sqrt{T}\log\left(\frac{T}{\delta}\right)\right) = \tilde{O}(dk\sqrt{T})$ for $\gamma = 1$, as desired.

Regret bound for RidgeFair$_m$ This regret bound relies on a similar analysis to RidgeFair$_1$, with the following changes. The algorithm now selects m individuals, not all from the top chain, but instead from several chains. For each of the top m choices $x \in P_{*,t}$, we relate reward of that choice to the reward of one of our choices in the following way. For each d, consider the dth top chain. We claim that if the dth top chain contains n_d of the top m choices, our algorithm selects n_d individuals from the dth top chain. We prove this claim by induction. First, however, we notice that every individual in the dth top chain has strictly higher reward than every individual in any lower chain. For the first top chain, P_t contains either the entire chain or m from the top chain. As every individuals in the top chain has strictly higher reward than any other individuals, in the former case, every individual in the first top chain belongs to $P_{*,t}$; in the latter case, $P_{*,t}$ is entirely contained in the top chain. Thus, P_t and $P_{*,t}$ contain either all individuals in the top chain or exactly m of them. Then, assuming the claim for the first $d-1$ top chains, both P_t and $P_{*,t}$ have the same "capacity" for individuals in the dth top chain (and therefore either take all of the dth top chain or fewer but the same number from it). This proves the claim.

Then, we relate the reward of an $i \in P_t$ with some action in $P_{*,t}$ belonging to same chain. Following the previous claim, we can form a matching between $P_{*,t}$ and P_t for which all matches belong to the same chains. Then, the analysis of RidgeFair$_1$ bounds the difference between the reward of any individual in the dth top chain to any other individual in the dth top chain. Summing up over all m choices, the total regret for all of P_t is at most m times the loss suffered in 1-bandit.

Regret bound for RidgeFair$_{\leq k}$ The regret bound for this case reduces to lower-bounding the amount of reward incurred by playing arms with negative reward. Any individual selected by RidgeFair$_{\leq k}$ is within the sum of the widths of the confidence intervals in its chain, one of which has UCB which is positive. So, the reward of any action chosen is at least $-\sum_{i \in S_t^d} w_{t,t} i$ for S_t^d the dth top chain, or at most the sum of all k interval widths. Thus, summing up over all individuals selected, one gets regret which is at most k times worse than that for RidgeFair$_1$. □

7.3 Proofs from Section 4

We begin with the full pseudocode for FairGap.

We start our proof of Theorem 4 with a lemma bounding the contribution of noise to our confidence intervals.

Lemma 1. *Let $\eta_1, \ldots, \eta_T$ be T i.i.d draws of R-sub-Gaussian noise. Then*

$$\mathbb{P}\left[\left|\sum_{i=1}^{T} \eta_i\right| \geq R\sqrt{2T\ln(2T/\delta)}\right] \leq \delta/2T.$$

Proof of Lemma 1. A Hoeffding bound, in the general case for unbounded variables, implies that

$$\mathbb{P}\left[\left|\sum_{i=1}^{T} \eta_i\right| \geq c\right] \leq 2\exp(-c^2/2R^2)$$

so taking $c = R\sqrt{2T\ln(2T/\delta)}$ yields the desired result. □

Next, since the regret bound we will prove depends on $\lambda = \min_{1\leq t\leq T}\left[\lambda_{\min}(\mathbb{E}_{x_t\sim \text{UAR} C_t}[{x_t}^T x_t])\right]$, the minimum smallest eigenvalue of the expected outer product of a vector x_t drawn uniformly at random from each C_t we will need $\lambda > 0$ in order for this bound to make sense. We prove this in another lemma.

Lemma 2. *Given sequence of action sets $C = (C_1, \ldots, C_T)$ where each C_t has nonzero Lebesgue measure and is contained in a ball of radius r, $\lambda = \min_{1\leq t\leq T}\left[\lambda_{\min}(\mathbb{E}_{x_t\sim UAR C_t}[{x_t}^T x_t])\right] > 0$.*

Proof of Lemma 2. It suffices to prove that $\lambda_{\min}(\mathbb{E}_{x_t\sim \text{UAR} Ct}[x_t^T x_t]) > 0$ for each $1 \leq t \leq T$. $x^T x$ is positive semidefinite, so it is immediate that $\lambda \geq 0$. Assume $\lambda = 0$. Then there exists nonzero $z \in \mathbb{R}^d$ such that $z\mathbb{E}_{x\sim \text{UAR} C}[x^T x]z^T = 0$, so by linearity of expectation

$$\mathbb{E}_{x\sim \text{UAR} C}[||xz^T||^2] = 0.$$

However, $||xz^T||^2$ is a non-negative random-variable with expectation 0 and must therefore be 0 with probability 1. It follows that $x \in z^\perp$, so

$$\mathbb{P}_{x\sim \text{UAR} C}\left[x \in z^\perp\right] = 1,$$

$z^\perp$ is a $d-1$ dimensional subspace of $\mathbb{R}^d$, and thus has measure 0. We can decompose $C = (C\cap z^\perp)\bigcup(C\cap\left(z^\perp\right)^c)$, and since $\mathbb{P}_{x\sim \text{UAR} C}\left[x \in z^\perp\right] = 1$, this forces $\mathbb{P}_{x\sim \text{UAR} C}\left[x \in (C \cap {z^\perp}^c)\right] = 0$. By definition of the uniform distribution

$$\mathbb{P}_{x\sim \text{UAR} C}\left[x \in (C \cap {z^\perp}^c)\right] = \frac{\mu(C \cap {z^\perp}^c)}{\mu(D)} \implies \mu(C \cap {z^\perp}^c) = 0.$$

But $\mu(D) = \mu(C \cap z^\perp) + \mu(C \cap {z^\perp}^c) = \mu(C \cap z^\perp) + 0 \leq \mu(z^\perp) = 0$, where the second to last line follows since $C \cap z^\perp \subset z^\perp$. This contradicts our assumption that $\mu(C) > 0$, so $\lambda > 0$. □

Finally, since FairGap relies on constructed confidence intervals to guide its choice of actions, its correctness (both in terms of its regret guarantee and its fairness) relies on the correctness of those confidence intervals, stated in the following lemma. Its proof relies on a natural argument using matrix Chernoff bounds to bound the contribution of noise to FairGap's estimation of $\hat{\beta}_t$ and, consequently, the accuracy of its confidence intervals.

Lemma 3. *Given sequence of action sets $C_{=}(C_1, \ldots, C_T)$ where each C_t has nonzero Lebesgue measure and is contained in a ball of radius r, with probability at least $1-\delta$, in every round t every confidence interval $[\langle\hat{\beta}_t, x\rangle - w_t, \langle\hat{\beta}_t, x\rangle + w_t]$ constructed by* FairGap *contains its true mean $\langle\beta, x\rangle$.*

Proof of Lemma 3. Note first that FairGap has two kinds of rounds: in round t, it either plays uniformly at random from
Ct or deterministically plays $\hat{x_t}^*$, its estimate of the optimal extremal point in
Ct. In any round t with uniform random play FairGap immediately cannot violate fairness, as $\pi_t(x) = 1/\mu($
$Ct)$ for all $x \in C_t$. As a result, to prove fairness it suffices to show that for any t-step execution of FairGap,

$$\mathbb{P}_{C_1,\ldots,Ct}\left[\text{deterministically play } \hat{x}_i^* \neq x_i^* \text{ in any round } i\right] \leq \delta$$

x_i^* is the true optimal point in C_i, and $t > 4dr^4/\delta\lambda^2$ (since for smaller i FairGap just plays uniformly at random).

In round $t+1$ after observing $x_1 \sim_{\text{UAR}} C_1, \ldots, x_t \sim_{\text{UAR}}$
Ct, for every $x \in \Omega$ we have

$$\begin{aligned}
|\langle x, \hat{\beta} - \beta\rangle| &= |\langle x, (X^TX)^{-1}X^T(X\beta + \eta) - \beta\rangle| \\
&= |x^T\beta + x^T(X^TX)^{-1}X^T\eta - x^T\beta| \\
&= |x^T(X^TX)^{-1}X^T\eta|
\end{aligned}$$

where $X \in \mathbb{R}^{t\times d}$ is the design matrix of $x_1, \ldots, x_i$ and $\eta \in \mathbb{R}^t$ is its noise vector. We can then decompose $X^T\eta$ by round as

$$\begin{aligned}
|x^T(X^TX)^{-1}X^T\eta| &= \left|x^T(X^TX)^{-1}\sum_{i=1}^{t} x_i\eta_i\right| \\
&= \left|\sum_{i=1}^{t} x^T(X^TX)^{-1}x_i\eta_i\right| \\
&\leq \sum_{i=1}^{t}\left[||x^T(X^TX)^{-1}x_i|| \cdot |\eta_i|\right] \\
&\leq \sum_{i=1}^{t}\sqrt{xx^T \cdot x_ix_i^T} \cdot \lambda_{\max}((X^TX)^{-1}) \cdot |\eta_i| \\
&\leq r^2 \cdot \lambda_{\max}((X^TX)^{-1}) \cdot \left|\sum_{i=1}^{t}\eta_i\right| \\
&= \frac{r^2}{\lambda_{\min}(X^TX)} \cdot \left|\sum_{i=1}^{t}\eta_i\right|
\end{aligned}$$

where the second inequality follows from the fact that

$$||(X^TX)^{-1}|| = \sqrt{\lambda_{\max}([X^TX]^{-1}[(X^TX)^{-1}]^T)} = \sqrt{\lambda_{\max}([(X^TX)^{-1}]^2)} = \lambda_{\max}((X^TX)^{-1}),$$

the third inequality follows from the assumed bound on each C_i, and the final equality follows from $\lambda_{\max}(A^{-1}) = \frac{1}{\lambda_{\min}(A)}$. To upper bound this quantity, we now lower bound $\lambda_{\min}(X^TX)$.

To do so, we first note that for any $1 \leq i \leq t$ and any $x \in C_i$ we have $\lambda_{\max}(x^T x) \leq r^2$ by the Gershgorin circle theorem, which states that a square matrix has maximum eigenvalue bounded by its largest absolute row or column sum. Next, by linearity of expectation

$$\lambda_{\min}(\mathbb{E}_{x_1 \sim_{\text{UAR}} C_1, \ldots, x_t \sim_{\text{UAR}} C t}[X^T X]) = \lambda_{\min}\left(\sum_{i=1}^{t} \mathbb{E}_{x_i \sim_{\text{UAR}} C_i}[x_i^T x_i]\right) \geq t\lambda$$

for $\lambda = \min_{1 \leq i \leq t}\left[\lambda_{\min}(\mathbb{E}_{x_i \sim_{\text{UAR}} C_i}[x_i^T x_i])\right]$.Taking this together with a matrix Chernoff bound (see e.g. Tropp et al. [2015]) yields

$$\mathbb{P}\left[\lambda_{\min}(X^T X) \leq \kappa t \lambda\right] \leq d e^{(-(1-\kappa)^2 t\lambda / 2r^2)}$$

for any $\kappa \in [0, 1)$. Setting

$$\kappa = 1 - \sqrt{\frac{2r^2 \ln\left(\frac{2dt}{\delta}\right)}{t\lambda}}$$

this implies

$$\mathbb{P}\left[\lambda_{\min}(X^T X) \leq \kappa t \lambda\right] < \frac{\delta}{2t}$$

where $\kappa \in [0, 1)$ since $t > 2r^2 \ln(2dt/\delta)/\lambda$. Combining this with Lemma 1 and a union bound, we get that with probability $\geq 1 - \delta/t$

$$\begin{aligned}\frac{r^2}{\lambda_{\min}(X^T X)} \cdot \left|\sum_{i=1}^{t} \eta_i\right| &\leq \frac{r^2}{\kappa t \lambda} \cdot R\sqrt{2t \ln(2t/\delta)} \\ &= \frac{r^2 R\sqrt{2\ln(2t/\delta)}}{\kappa\lambda\sqrt{t}}.\end{aligned}$$

Taking a union bound over t rounds, it follows that with probability at least $1 - \delta$ through t rounds every constructed confidence interval around $\langle \hat{\beta}, x \rangle$ contains $\langle \beta, x \rangle$. Since FairGap only plays $\hat{x}^*$ deterministically when the confidence intervals around $\hat{x}^*$ and other extremal points do not overlap, this means that with probability at least $1 - \delta$ FairGap correctly identifies x^*. FairGap is therefore fair. □

Taken together, these lemmas let us prove Theorem 4.

Proof. Proof of Theorem 4 We begin by proving fairness. By Lemma 3, with probability at least $1 - \delta$ every confidence interval constructed by FairGap contains its true mean. Conditioning on this correctness of confidence intervals, since FairGap only chooses an action x_1 non-uniformly when $U_1 \cap U_2 = \emptyset$, it follows that any action chosen non-uniformly by FairGap is optimal. Thus, with probability at least $1 - \delta$ FairGap never chooses a suboptimal action x with higher mixture density $\pi_t(x)$ than a superior action x', and FairGap is fair.

While FairGap plays at random from C (for some number of rounds at least $4dr^4/\delta\lambda^2$), it incurs at most $2r$ regret per round. The algorithm incurs 0 regret once the confidence intervals around the top two extremal points no longer intersect. A sufficient condition is therefore

$$\frac{r^2 \cdot R \cdot \sqrt{2\ln(2T/\delta)}}{\kappa\lambda\sqrt{T}} < \frac{\Delta_{\text{gap}}}{2}$$

which we rearrange into

$$\frac{8r^4 R^2 \ln(2T/\delta)}{\kappa^2 \lambda^2 \Delta_{\text{gap}}^2} < T.$$

After this many rounds, with probability $\geq 1 - \delta$, FairGap identifies the optimal arm in *every* C_t and incurs no further regret.

Thus, the regret in total is at most

$$\sum_{t=1}^{L} 2r^2 + \delta T \leq \frac{16r^6R^2\ln(2T/\delta)}{\kappa^2\lambda^2\Delta_{\text{gap}}^2} + \delta T$$

where $L = \frac{8r^4R^2\ln(2t/\delta)}{\kappa^2\lambda^2\Delta_{\text{gap}}^2}$ and $\delta \leq 1/(T^{1+c})$ then implies the claim. □

7.4 Efficient Approximate Version of Section 4

In this section we describe an efficient implementation of FairGap using approximate fairness.

Recall that FairGap requires some method of sampling uniformly at random from a given convex body C_t, a problem that has attracted extensive attention over the past few decades (see Vempala [2005] for a survey of results). For our purposes, the primary contribution of this literature is that one cannot do better than *approximately* uniform random sampling from a convex set C_t under polynomial time constraints.

Since our current definition of fairness fails without a perfectly uniform distribution over actions, efficiency necessitates a relaxation of our definition to approximate fairness for infinite action spaces. Intuitively, approximate fairness will require that an algorithm (with high probability) always uses a distribution that is at least "almost" fair.

Definition 4 (ϵ-Approximate Fairness)**.** Given sequence of action sets $C = (C_1, \ldots, C_T)$, we say that algorithm $\mathcal{A}$ is *ϵ-approximately fair* if, for any inputs $\delta \in (0, 1], \epsilon > 0$ and for all β, with probability at least $1 - \delta$ at every round t there exists a fair distribution ${\pi_t}^f$ such that

$$||\pi_t - {\pi_t}^f|| < \epsilon$$

where π_t is $\mathcal{A}$'s choice distribution over C_t in round t and $\|\cdot\|$ denotes total variation distance.

We call this ϵ-approximate fairness to highlight that a single ϵ is input to the algorithm $\mathcal{A}$ in question, but will often shorthand this as *approximate fairness*.

Below we provide an approximately fair algorithm that, subject to additional assumptions on choice set structure, obtains similar regret guarantees as FairGap efficiently. We modify FairGap as follows: first, we replace each call to a random sample with a hit-and-run random walk scheme Lovász and Vempala [2006] to efficiently sample approximately uniformly at random.

We use the following lemma from Lovász and Vempala [2006] to upper-bound the mixing time hit-and-run requires to approach a near-uniform distribution in its walk over C_t.

Lemma 4. *[Corollary 1.2 in Lovász and Vempala [2006]] Let S be a convex set that contains a ball of radius r' and is contained in a ball of radius r. Then, starting from a point $x \in S$ at a distance α from the boundary, after*

$$c > 10^{11}d^3\left(\frac{r}{r'}\right)^2 \ln\left(\frac{r}{\alpha\epsilon}\right)$$

steps of a hit-and-run random walk the random walk induces a probability distribution P over points in S such that P is ϵ-close to uniform in total variation distance.

Next, we show that, with an additional assumption on the structure of C_t, FairGap's subroutine TopTwo can be implemented efficiently via the following known result [Lindgren et al., 2016].

Lemma 5 (Lindgren et al. [2016])**.** *Let C_t be defined by m intersecting half-planes. Then there exists an algorithm running in time polynomial in m and d which computes the two vertices which maximize $\hat{\beta}_t$ over C_t.*

This algorithm enables us to compute $\text{TopTwo}(C_t, \hat{\beta}_t)$ efficiently.

The following lemma guarantees that the distributions over histories generated by FairGap and Approx-FairGap are "close" during exploration.

Lemma 6. *Let $C = (C_1, \ldots, C_T)$ be a sequence of action sets where each action set, in addition to satisfying the assumptions of Theorem 4, is an intersection of polynomially many halfspaces and contains a ball of radius r'. Then through t rounds of exploration*

$$||P_{\pi_1,\ldots,\pi_t \sim \textsc{FairGap}} - P_{\pi_1,\ldots,\pi_t \sim \textsc{ApproxFairGap}(\epsilon/t)}||_{TV} < \epsilon$$

where each P represents distributions over possible exploration histories generated by FairGap *and* ApproxFairGap(ϵ/t) *respectively.*

Proof of Lemma 6. By construction, during exploration each π_i output by ApproxFairGap(ϵ) has a distribution within ϵ/t of a uniform distribution in total variation distance. Since these samples are independent, each distribution over $\pi_1, \ldots, \pi_t$ forms a product distribution, and the additivity of total variation distance over product distributions implies the claim. □

Combining the results above lets us prove that ApproxFairGap is fair, efficient, and obtains a similar regret bound as FairGap.

Theorem 5. *Consider an action set C that, in addition to satisfying the assumptions of Theorem 4, is an intersection of polynomially many halfspaces and contains a ball of radius r'. Then through T steps given inputs $\delta' = \delta/2$ and $\epsilon' = \min(\epsilon/T, \delta/2T^2)$,* ApproxFairGap$(\epsilon')$ *is efficient, ϵ-approximately fair, and achieves regret*

$$\textsc{Regret}\,(T) = O\left(\frac{r^6 R^2 \ln(4T/\delta)}{\kappa^2 \lambda^2 \Delta_{gap}^2}\right)$$

where $\kappa = 1 - r\sqrt{\frac{2\ln\left(\frac{2dT}{\delta}\right)}{T\lambda}}$.

Proof. Proof of Theorem 5 In each round t, FairGap performs (at most) three computation-intensive operations. First, it computes a least squares estimator $\hat{\beta}_t$, which may be maintained online and updated in poly(d) time. Next, it calls subroutine TopTwo $(C_t, \hat{\beta}_t)$ to compute (x_1, x_2) in poly(d, m) time via Lemma 5. Finally, it may choose an action (approximately) uniformly at random from C_t, which also takes polynomial time via Lemma 4. It follows that each round t of FairGap takes polynomial time, so FairGap is efficient.

To prove that ApproxFairGap is approximately fair, as in the exact case we analyze ApproxFairGap's split between exploration and exploitation. In exploration rounds, by Lemma 4 we know that each random sample is ϵ/T-close to a true uniform distribution and therefore satisfies ϵ-approximate fairness immediately.

We now bound the probability of violating fairness during exploitation. This can only happen if in some round t ApproxFairGap misidentifies the optimal extremal point x_t^* to exploit and instead deterministically plays $\hat{x}_t^* \neq x_t^*$. Since ApproxFairGap only uses exploration rounds to construct its design matrix, the identified $\hat{x}_t^*$ is a deterministic function of the $k \leq t-1$ exploration rounds $h_1, \ldots, h_k$ seen before round t. Lemma 6 implies that FairGap and ApproxFairGap have distributions over $h_1, \ldots, h_t$ within ϵ of each other. We then combine two facts. First, here FairGap has at most $\delta/2$ probability of constructing incorrect confidence intervals assuming perfect uniform random sampling. Second, ApproxFairGap has probability at most $\epsilon \leq \delta/2T^2$ of identifying a $\hat{x}_t^*$ different from that of FairGap by the above argument. A union bound then implies that ApproxFairGap has probability at most $\delta/2$ of identifying a different $\hat{x}_t^*$ than FairGap. Combining the probability of FairGap failing and ApproxFairGap failing to approximate FairGap, we get that ApproxFairGap has probability at most $\delta/2 + \delta/2 = \delta$ of misidentifying x^*. Thus ApproxFairGap is ϵ-approximately fair.

To analyze ApproxFairGap's regret, note that in the case where ApproxFairGap correctly identifies x_t^*, ApproxFairGap's use of $\delta/2$ rather than δ adds a factor of 2 inside the log in the original regret statement of FairGap. Next, ApproxFairGap incorrectly identifies x_t^* with probability at most δ by the logic above, so taking $\delta \leq 1/T^{1+c}$ as in the proof of Theorem 4 implies the claim. □

7.5 Proofs from Section 5

Proof of Theorem 3. Let $\mathcal{A}$ be a fair algorithm. For any input δ, $\mathcal{A}$ is round-fair for all $t \geq 1$ with probability $1-\delta$. Since this holds for any β with probability at least $1-\delta$, then it necessarily holds with probability at least $1-\delta$ over any prior τ on β with support contained in the unit rectangle. Our first lemma gives an alternative way to view the framework which draws $\beta \sim \tau$ and then plays according to $\mathcal{A}$.

Let x_t denote the action chosen by $\mathcal{A}$ at time step t, and let y_t denote the observed reward. Let the joint distribution of $((x_1, y_1), \dots (x_t, y_t), \beta)$ be denoted by W_t. Lemma 7 is similar in content to Lemma 4 in Joseph et al. [2016]; its proof follows from Bayes' Rule.

Lemma 7. *Let β' at time t be drawn from $\tau|h_t$, its posterior distribution given the observed sequence of choices and rewards $h_t = ((x_1, y_1), \dots (x_{t-1}, y_{t-1})) \in (C \times \mathbb{R})^{t-1}$. Then let W_t' be the joint distribution of $(h_t, (x_t, y_t), \beta')$. W_t and W_t' are identical distributions.*

Lemma 7 states that whether the instance draws $\beta \sim \tau$ once and then plays according to $\mathcal{A}$, or re-draws β from its posterior at each time-step, the joint distribution on instances and observations is unchanged at each step. We can thus assume without loss of generality that, given a prior τ, at each time step t we redraw $\beta \sim \tau|_{h_t}$. Taking this posterior viewpoint, we have the following lemma.

Lemma 8. *Given a fixed prior τ, let $\mathcal{A}$ be fair and let $\beta \sim \tau$. Let π_t be the distribution on actions of $\mathcal{A}$ at time t, and let f_t be the pdf of π_t. Then with probability at least $1-4\delta$, at each time t, if $\mathbb{P}_{\tau|h_t}\left[\langle \beta, y\rangle > \langle \beta, x\rangle\right] > \frac{1}{4}$, then $f_t(y) \geq f_t(x)$.*

This means that with probability at least $1-4\delta$, whenever the posterior distribution at time t tells us that point y has a higher reward than point x with probability at least $\frac{1}{4}$ over the posterior distribution of β, we must play y with at least the same probability as x.

We will use this lemma, in combination with results about a specific posterior, to constrain the possible actions any fair algorithm can take.

We now introduce the specific prior τ. Let β have prior distribution $\tau \sim \{1\} \times U[-\epsilon, -\epsilon]$. We first analyze the posterior distribution of β. We then show that with probability at least $1-4\delta$, until the posterior distribution differs from the prior, Lemma 8 forces $\mathcal{A}$ to play uniformly from C_t.

Suppose that we have observed $(x_1, y_1) \dots (x_{t-1}, y_{t-1})$. Since the prior in the second coordinate is $U[-\epsilon, \epsilon]$, and the noise η_i is also uniform, the posterior in the second coordinate is uniform over all β_2 consistent with the observed data in the following sense: since the noise $\eta_{t'}$ is bounded, each pair $(x_{t'}, y_{t'})$ gives a bound on β_2. Combining $y_{t'} = x_{t',1} + \beta_2 x_{t',2} + \eta_{t'}$ and $\eta_{t'} \in [-1, 1]$ we get

$$\beta_2 \in [l_{t'}, u_{t'}] = \left[\min\left(\frac{y_{t'} - x_{t',1} - 1}{x_{t',2}}, \frac{y_{t'} - x_{t',1} + 1}{x_{t',2}}\right), \max\left(\frac{y_{t'} - x_{t',1} - 1}{x_{t',2}}, \frac{y_{t'} - x_{t',1} + 1}{x_{t',2}}\right)\right]. \quad (4)$$

Since by the prior we know $\beta_2 \in [-\epsilon, \epsilon]$, we say that β_2 is consistent with h_t if $\beta_2 \in [-\epsilon, \epsilon]$ and $\beta_2 \in [\sup_{t'} l_{t'}, \inf_{t'} u_{t'}]$. This is the content of the following lemma.

Lemma 9. *Let $y_{t'} = \langle \beta, x_{t',}\rangle + \eta_{t'}, \eta_{t'} \sim U[-1, 1]$, and $\beta \sim \{1\} \times U[-\epsilon, \epsilon]$. Then $\tau(\beta_2|h_t)$ is uniform on the set of β_2 consistent with h_t.*

We now define and analyze S, the number of rounds required before the posterior distribution of β_2 becomes non-uniform. Each $(x_{t'}, y_{t'})$ gives the constraint on β given in Equation 4. This only changes the posterior from the prior if $l_{t'} > -\epsilon$ or $u_{t'} < \epsilon$. Assume first that $x_{t',2} > 0$ (by symmetry, a similar argument holds for $x_{t',2} < 0$). Then $u_{t'} = \dfrac{y_{t'} - x_{t',1} + 1}{x_{t',2}}$ and we can calculate

$$\begin{aligned}
\mathbb{P}\left[\frac{y_{t'} + 1 - x_{t',1}}{x_{t',2}} < \epsilon\right] &= \mathbb{P}\left[\frac{\eta_{t'} + x_{t',2}\beta_2 + 1}{x_{t',2}} < \epsilon\right] \\
&= \mathbb{P}\left[\eta_{t'} + 1 < x_{t',2}(\epsilon - \beta_2)\right] \\
&\leq \mathbb{P}\left[\eta_{t'} + 1 < 2\epsilon\right] = \epsilon
\end{aligned}$$

where the last equality follows from the fact that $\eta_t + 1 \sim U[0,2]$. The probability that the lower bound is greater than $-\epsilon$ is similarly

$$\mathbb{P}\left[\frac{y_{t'} - 1 - x_{t',1}}{x_{t',2}} > -\epsilon\right] = \mathbb{P}\left[\eta_{t'} > 1 + x_{t',2}(-\epsilon - \beta_2)\right] \leq \epsilon.$$

Thus the probability that any pair $(x_{t'}, y_{t'})$ alters the posterior distribution of β_2 from $U[-\epsilon, \epsilon]$ is at most 2ϵ. [7] It follows that $\mathbb{P}(S \geq t') \geq (1-2\epsilon)^{t'}$, and that the posterior coincides with the prior τ for $\Omega(1/\epsilon)$ steps in expectation.

Now assume that after $t-1$ steps the posterior distribution is equal to τ: we will argue that any non-uniform distribution violates round fairness in round t with probability at least $\frac{3}{4}$. Call two points $a, b \in C_t = [-1,1]^2$ *vertically equivalent* if $a_1 = b_1$, i.e. they agree in their first coordinate. Consider some pair of points $a = (x_1, x_2), b = (x_1, x_3) \in C_t$ which are vertically equivalent with $x_2 > x_3$. Suppose $\mathcal{A}$ plays a with higher probability than b. If $\beta_2 < 0$, then $\langle \beta, a\rangle < \langle \beta, b\rangle$, and $\mathbb{P}[\beta_2 < 0] = 1/2 > \frac{1}{4}$. Thus $\mathcal{A}$ violates round-fairness in round t with probability more than $\frac{1}{4}$. Similarly, if $\beta_2 > 0$, then $\langle \beta, a\rangle > \langle \beta, b\rangle$, and $\mathbb{P}[\beta_2 > 0] = 1/2 > \frac{1}{4}$, so if $\mathcal{A}$ plays b with higher probability than a then $\mathcal{A}$ again violates round-fairness in round t with probability strictly larger than $\frac{1}{4}$. Thus, any two vertically equivalent points must be played with equal probability.

Next, consider any point $b \in C_t$ of the form $\left(x_1 - \alpha, x_2 + \frac{2\alpha}{\epsilon}\right)$ for some $\alpha \in \mathbb{R}$. Call any two points of this form, for fixed (x_1, x_2) and variable $\alpha \in \mathbb{R}$, *diagonally equivalent*. Let $a = (x_1, x_2)$. If $\beta_2 > \epsilon/2$, then $\langle \beta, b\rangle \geq x_1 - \alpha + x_2\beta_2 + \alpha = x_1 + x_2\beta_2 = \langle \beta, a\rangle$. Since $\beta_2 > \epsilon/2$ with probability $\frac{1}{4}$, point b must be played with probability at least that of a to satisfy round-fairness in round t with probability greater than $\frac{3}{4}$. Symmetrically, when $\beta_2 < \frac{-\epsilon}{2}$, which happens with probability $\frac{1}{4}$, a must have at least as much probability of being played as b. Thus, any two diagonally equivalent points must also be played with equal probability.

Given a point $x \in C_t$, let H_x denote the transitive closure under vertical and diagonal equivalence of the point x. Since points that are equivalent must be played with equal probability, by the transitive property all points in H_x must be played with equal probability by $\mathcal{A}$. We now show that when x is a corner of C_t, $H_x = C_t$.

Lemma 10. *Let $x = (1,-1)$. Then $H_x = C_t$.*

Lemma 10 shows that if $\mathcal{A}$ is fair at a given round t with probability at least $\frac{3}{4}$ over the posterior, and the posterior is τ, then $\mathcal{A}$ must play uniformly at random from C_t.

Thus, we have shown for any fair $\mathcal{A}$:

1. With probability at least $1 - 4\delta$, $\mathcal{A}$ must be fair with probability at least $\frac{3}{4}$ at all $t \geq 1$ (Lemma 8)
2. If S is the number of rounds until $\tau \neq \tau|_{h_t}$, $\mathbb{P}(S \geq t) \geq (1-2\epsilon)^t$
3. When $\tau = \tau|_{h_t}$ (i.e. $S \geq t$), and $\mathcal{A}$ is fair with probability $> \frac{3}{4}$ over $\tau|_{h_t}$, then $\mathcal{A}$ must play uniformly at random from C_t

Let $\epsilon < \min(1/2, 1/\log(2/\delta))$ and let the event that $\mathcal{A}$ is fair with probability at least $\frac{3}{4}$ over the posterior at all $t \geq 1$ be denoted by F. Recalling that S denotes the number of rounds required before the posterior distribution of β_2 becomes non-uniform, let the event that $S \geq \frac{\log(1-\delta)}{\log(1-2\epsilon)}$ be denoted by E. Then

$$\mathbb{P}[E] \geq (1-2\epsilon)^{\frac{\log(1-\delta)}{\log(1-2\epsilon)}} = 1 - \delta,$$

so

$$\mathbb{P}[E \cap F] \geq \mathbb{P}[E] + \mathbb{P}[F] - 1 \geq 1 - 5\delta.$$

[7] Note that this bound holds regardless of the particular choice of $x_{t'}$, which is why the probabilities above are over the draw of the rewards $y_{t'}$, conditional on the chosen $x_{t'}$.

We now condition on $F \cap E$ to show that with high probability $\textsc{Regret}(T) = \tilde{\Omega}(\frac{1}{\epsilon})$:

$$\begin{aligned}\mathbb{P}\left[\textsc{Regret}(T) \geq \tilde{\Omega}\left(\frac{1}{\epsilon}\right)\right] &\geq \mathbb{P}\left[\textsc{Regret}(T) \geq \tilde{\Omega}\left(\frac{1}{\epsilon}\right) \mid E \cap F\right] \mathbb{P}\left[E \cap F\right] \\ &\geq \mathbb{P}\left[\textsc{Regret}(T) \geq \tilde{\Omega}\left(\frac{1}{\epsilon}\right) \mid E \cap F\right] (1 - 5\delta)\end{aligned} \tag{5}$$

where the first inequality follows from Bayes' rule. However, we've shown that whenever $E \cap F$ occurs, for at least $\frac{\log(1-\delta)}{\log(1-2\epsilon)} \geq \frac{\log(1/[1-\delta])}{2\epsilon}$ (via $\log(x) \leq x - 1$ for $x > 0$) rounds $\mathcal{A}$ plays uniformly at random from C_t. Let $r_{\mathcal{A}}(t)$ be the regret accrued at round t by uniformly at random play, $\mathbb{E}[r_{\mathcal{A}}(t)] = ||\beta||_1 = \Omega(1) = c$. Then $0 \leq r_{\mathcal{A}}(t) \leq 2(1+\epsilon)$, and the $r_{\mathcal{A}}(t)$ are independent since $\mathcal{A}$ is playing uniformly at random at each t. By Hoeffding's inequality for bounded random variables,

$$\mathbb{P}\left[\sum_{t=1}^{T} r_{\mathcal{A}}(t) \leq T \cdot c - \sqrt{2T\log(2/\delta)}(1+\epsilon)\right] \leq \delta$$

which means

$$\mathbb{P}\left[\sum_{t=1}^{T} r_{\mathcal{A}}(t) \geq T \cdot c - \sqrt{2T\log(2/\delta)}(1+\epsilon)\right] \geq 1 - \delta \tag{6}$$

and when taking $T = \frac{1}{\epsilon}$ we get

$$\mathbb{P}\left[\sum_{t=1}^{T} r_{\mathcal{A}}(t) \geq \frac{1}{\epsilon} \cdot c - \sqrt{\frac{2}{\epsilon}\log(2/\delta)}(1+\epsilon)\right] \geq 1 - \delta$$

or suppressing constants and lower order terms and using the fact that $\epsilon < 1/\log(2/\delta)$, $\mathbb{P}\left[\sum_{t=1}^{T} r_{\mathcal{A}}(t) \geq \tilde{\Omega}(\frac{1}{\epsilon})\right] \geq 1-\delta$. This gives us that $\mathbb{P}\left[\textsc{Regret}(T) \geq \tilde{\Omega}(\frac{1}{\epsilon}) \mid E \cap F\right] \geq \frac{1-\delta}{1-5\delta}$. Hence by Equation 5, $\mathbb{P}\left[\textsc{Regret}(T) \geq \tilde{\Omega}(1/\epsilon)\right] \geq 1 - \delta$, as desired.

□

We now provide the proofs of the lemmas used above.

Proof of Lemma 8. By the definition of fairness, and Lemma 7, we have that

$$\mathbb{P}_{\beta_t \sim \tau|_{h_t}, h_t \sim \mathcal{A}}\left[\exists t' \geq 1 \colon \mathcal{A} \text{ is round-unfair at time } t'\right] \leq \delta.$$

Denote this probability by X. By the above $\mathbb{E}[X] \leq \delta$, and hence by Markov's inequality, $\mathbb{P}\left[X \geq \frac{1}{4}\right] \leq 4\delta$. But then we've shown that, with probability at least $1 - 4\delta$, for all $t \geq 1$ $\mathcal{A}$ is fair with probability at least $\frac{3}{4}$ over $\beta \sim \tau|h_t$. Now if $\exists x, y$ such that $P_{\tau|h_t}(\langle y', \beta\rangle > \langle x', \beta\rangle) > \frac{1}{4}$ but $f_t(x) > f_t(y)$, then the probability that $\mathcal{A}$ is unfair at time t is at least $\mathbb{P}_{\tau|h_t}(\langle y', \beta\rangle > \langle x', \beta\rangle) > \frac{1}{4}$. This proves the claim. □

Proof of Lemma 9. The fact that the posterior distribution of β_2 is uniform on the set of consistent β_2 is immediate via Bayes rule: $\tau(\beta_2|h_t) = p(h_t|\beta_2)\tau(\beta_2)$, where $p(h_t|\beta_2)\tau(\beta_2) \propto 1$ if β_2 is consistent with h_t, and is 0 otherwise. □

Proof of Lemma 10. Choose an arbitrary point $y \in C_t$ with coordinates (z_1, z_2). We want to show $y \in H_x$. Since any two points in C_t with the same x coordinate are vertically equivalent, it suffices to show that there is a point with x-coordinate $z_1 \in H_x$.

Fix $0 < \alpha \leq \min(1, 2\epsilon)$ and suppose $1 - z_1 = k \cdot \alpha$, where $k \in \mathbb{N}$. Note we can guarantee $k \in \mathbb{N}$ by choosing an appropriate α. We now proceed by induction on k.

If $k = 1$, then by diagonal equivalence x is equivalent to $x' = (1-\alpha, -1+2\alpha/\epsilon) = (z_1, 1+2\alpha/\epsilon)$. But by vertical equivalence, $y \in H_{x'}$, and so $y \in H_x$, by transitivity. For the inductive step, construct $x' = (z_1+\alpha, z_2 - 2\alpha/\epsilon)$. Then $1 - x'_1 = 1 - z_1 - \alpha = (k-1)\alpha$. Hence by induction $x' \in H_x$. But since x' is diagonally equivalent to $y = (z_1, z_2)$, then $y \in H_x$ as desired. Since y was arbitrarily chosen, $H_x = C_t$. See Figure 5 for a visualization of these equivalences. □

7.6 Proofs from Section 6

Proof of Theorem 4. Let E_β be the event that given a fixed value of β, $\mathcal{A}$ plays uniformly at random from C_t for all $t \geq 1$. If we can show that for any $\mathcal{A}$ and all β, it is the case that $\mathbb{P}(E_\beta) = \Omega(1)$, this implies the claim, since for any β, T

$$\mathbb{E}\left[\textsc{Regret}\left(T\right)\right] \geq \mathbb{E}\left[\textsc{Regret}\left(T\right) \mid E_\beta\right] \mathbb{P}\left[E_\beta\right] = \Omega(T) \cdot \Omega(1) = \Omega(T),$$

as desired.

By symmetry of S^1, $\mathbb{P}\left[E_\beta\right] = \mathbb{P}\left[E_{\beta'}\right]$ for all $\beta, \beta' \in S^1$. So henceforth we can drop the subscript β, and use E to represent the event that $\mathcal{A}$ plays uniformly at random for all $t \geq 1$. We now exhibit a prior τ such that for any $\mathcal{A}$, $\mathbb{P}\left[E\right] = \Omega(1)$.

Lemmas 7 and 8 both apply; thus, we let $\beta \sim \tau$, where τ is the uniform distribution on $S^1, U(S^1)$, and we assume that at each time t, β is re-drawn from its posterior distribution $\tau|_{h_t}$, as before. Let F again be the event that $\mathcal{A}$ is round-fair with probability at least $\frac{3}{4}$ at each round t, with respect to the posterior distribution $\tau|h_t$. We again analyze the posterior distribution $\tau|_{h_t}$, showing that for any history $h_t, \tau|_{h_t}$ forces $\mathcal{A}$ to play uniformly at random at t, conditioned on F.

As in Section 5 the posterior distribution of $\beta|h_t$ is uniform on the set of $\beta \in S^1$ that are consistent with the observed data. By consistent we again mean in the sense of Lemma 9; the proof is nearly identical and relies on boundedness of the noise η_t, so we do not repeat it here. Denote by $G_t \subset S^1$ the set of consistent β at time t. We will use Lemma 11 to reason about the topology of G_t. We use the relative topology throughout.

Lemma 11. *For any $t \geq 1$ and any history h_t, G_t is a nonempty connected open subset of S^1.*

G_t is an open, non-empty, connected subset of S^1; since we're working in the relative topology, it must be exactly an open interval along the boundary of S^1, as illustrated in Figure 6. Let G_t have length ϵ, and correspondingly $\tau|_{h_t} = U(G_t)$.

Condition on the occurrence of F: that $\mathcal{A}$ must be fair in round t with probability at least $\frac{3}{4}$, with respect to $\tau|_{h_t}$. We claim that this in fact forces $\mathcal{A}$ to play uniformly from S^1 at all time steps t, in an argument similar to Lemma 10.

We say that two points $x, y \in S^1$ are equivalent at time t if $\mathbb{P}_{\beta\sim\tau|_{h_t}}\left[\langle\beta, x\rangle > \langle\beta, y\rangle\right] \in [\frac{1}{4}, \frac{3}{4}]$. Let $S_{x,t}$ be the transitive closure of the set of $y \in S^1$ that are equivalent to x at time t.

Lemma 12. *Let $\tau|_{h_t} \sim U(G_t)$. Then there exists $x \in S^1$ such that $S_{x,t} = S^1$.*

Proof of Lemma 12. By definition, if $\mathbb{P}_{\tau|h_t}\left[\langle\beta, x\rangle < \langle\beta, y\rangle\right] \in [\frac{1}{4}, \frac{3}{4}]$, then $y \in S_{x,t}$. Every point on S^1 can be represented as $(\cos\theta, \sin\theta)$, so let θ_x denote the angle corresponding to x, and let x be the point in G_t such that $\mathbb{P}_{\tau|h_t}\left[\beta < \theta_x\right] = \frac{1}{4}$.

Now let $S_{t,-} = \{z \in G_t : \theta_z \geq \theta_x\}$ and let $S_{t,+} = \{z \in G_t : \theta_z \leq \theta_x\}$. If $\beta \in S_{t,+}$, then for all $z \in S_{t,-}, \beta \cdot z \leq \beta \cdot x$. By construction, $\mathbb{P}_{\tau|h_t}\left[\beta \in S_{t,+}\right] = \frac{1}{4}$, and hence $S_{t,-} \subset G_{x,t}$. But defining x_1 as $\mathbb{P}_{\tau|h_t}\left[\beta > \theta_{x_1}\right] = \frac{1}{4}$, $S'_{t,+}$ as the set $\{z \in G_t : \theta_z > \theta_{x_1}\}$, and $S'_{t,-}$ as $\{z \in G_t : \theta_z < \theta_{x_1}\}$, the same reasoning shows that $S'_{t,-} \subset S_{x_1,t}$. Since $S_{t,-} \cup S'_{t,-} = G_t$, this forces $G_t \subset S_{x_1,t} = S_{x,t}$.

We now show $S_{x,t}$ contains the rest of the boundary of S^1, not just G_t. Let G^1_+ denote the arc of length $\frac{1}{4}\epsilon$ adjoining $S'_{t,+}$ as in Figure 6, and define G^1_- accordingly. Now note that we must have $G^1_+ \in G_{x,t}$, since if $\beta > \theta_{x_1}$ then for all $z \in G^1_+, \beta \cdot z > \beta \cdot x$, and $\mathbb{P}_{\tau|h_t}\left[\beta > \theta_{x_1}\right] = \frac{1}{4}$. Similarly, G^1_- has to be added to $S_{x_1,t} = S_{x,t}$ as well. But then letting the segment $G^1_- \cup G^1_+ \cup G_t$ be denoted by G'_t, we can repeat the argument: we set x', x'_1 to be their initial locations x_1, x translated $\frac{1}{4}\epsilon$ to the right and left respectively, and define G^2_+, G^2_- analogously, as in the Figure 6.

Now we have that $G^2_+ \in S_{x',t}$, since if $\beta \in S'_{t,+}$ then for all $z \in G^2_+, \beta \cdot z > \beta \cdot x'$, and hence $z \in S_{x',t} = S_{x,t}$. The same logic shows that $G^2_+ \subset S_{x'_1,t} = S_{x,t}$.

Since we can keep recursively chaining segments of fixed length $\frac{\epsilon}{4}$ to $S_{x,t}$, and S^1 is of fixed length, a simple induction argument forces $S_{x,t} = S^1$, as desired. □

So Lemma 8 in combination with the above lemma forces the following: when $\mathcal{A}$ is constrained to be fair with probability at least $\frac{3}{4}$ with respect to the posterior distribution of β, for all times $t \geq 1$ and all histories h_t, $\mathcal{A}$ must play uniformly at random from S^1. But then $\mathbb{P}(E) \geq \mathbb{P}(E|F)\mathbb{P}(F) = \mathbb{P}(F) \geq 1 - 4\delta = \Omega(1)$, by Lemma 8. □

Proof of Lemma 11. $C_t \neq \varnothing$ is immediate since, for the true value β, $\beta \in C_t$ for all t. For $\beta \in S^1$ to be consistent with the data, i.e. in C_t, means that $\max_{1 \leq i \leq t} |y_i - \langle \beta, x_i \rangle| < 1$ and $\beta \in S^1$.

We can rephrase this as follows: if $f_i(\beta) = |y - \langle \beta, x_i \rangle|$, and $R_i = \{\beta \in f_i^{-1}(-\infty, 1)\}$, then if we let $C'_t = \bigcap_{i=1}^t R_i$, $C_t = C'_t \cap S^1$. Now we remark that each R_i is the intersection of the two open half spaces $\{\beta : \langle \beta, x_i \rangle < 1 + y_i\}$ and $\{\beta : \langle \beta, x_i \rangle > y_i - 1\}$. Thus C'_t is the intersection of finitely many open half spaces, and is thus an open, connected set (in fact, it is a convex polytope). Since $C_t = S^1 \cap C'_t$, by definition C_t is open and connected in the relative topology on S^1. □

7.7 Experiments

Figure 7.7 depicts experiments conducted in the k-bandit setting. We employ a simple variant of UCB that maintains generic normal confidence intervals around its ongoing estimate of β and uses these to construct confidence intervals for the estimated rewards of the contexts is uses; it then selects all choices with a positive upper confidence bound. We plot *cumulative mistreatments* through $T = 10,000$ rounds, which tracks the cumulative number of individuals who have seen an individual with lower expected quality chosen in a round during which they were not chosen. The plot therefore shows that through 10,000 rounds our version of UCB creates nearly 400 such mistreated people.

Our experiments use $d = 2$ and $\beta \sim U[-1, 1]^2$ for each iteration. In each round we generate $k = 10$ contexts x_i, also from $U[-1, 1]^2$, and generate noisy rewards $\beta \cdot x_i + \eta_{t,i}$ where $\eta_{t,i} \sim N(0, 1)$ is standard normal noise. The results presented are averaged over 100 iterations. For completeness, we present Figure 7.7, which plots cumulative mistreatments for both UCB and FairUCB and empirically validates our theoretical fairness guarantee.

Our second experiment investigates the *structure* of mistreatment in UCB. We use $d = 2, \beta = [1, 0], k = 10$ for each iteration. At each round t with probability $p \in [.8, .95]$ we draw a context (x, x), where $x \sim U[-1, 1]$ and with probability $1 - p$ draw a context from $U[-1, 1]^2$. These two types of contexts naturally encode two populations: in population 1, the two features are perfectly correlated and in population 2 they are independent. However, $\beta = [1, 0]$ crucially means that the second feature *does not affect reward*. Our experiments aim to study how this correlation affects mistreatment rates in the different populations.

For each population we plot the fraction of mistreatment individuals from each population for $T = 1, \ldots 25$, averaging over 1000 iterations. Figure 7.7 shows that for $p \in [.8, .95]$ unfairness accrues at substantially different rates to the two populations. Somewhat counter-intuitively, members of the majority group are significantly more likely to be mistreated than members of the minority group, a natural consequence of UCB-style algorithms favoring minority contexts whose confidence intervals have more uncertainty. While mistreating a majority population may be less obviously unfair than mistreating a minority population, it is still undesirable. In particular, there may be natural practical settings where the group that has faced historical discrimination is the majority population in sample (e.g. criminal sentencing) and so discriminating against the majority is more obviously unfair.

procedure RIDGEFAIR$_m(\delta, T, k, \gamma \geq 1$, ExactBool)
  for $t \geq 1, 1 \leq i \leq k$ **do**
    Let $\mathbf{X}_t, \mathbf{Y}_t$ = design matrix, observed payoffs before round t
    Let C_t be the choice set in round t
    Let $\bar{V}_t = \mathbf{X}_t{}^T \mathbf{X}_t + \gamma I$
    Let $\hat{\beta}_t = (\bar{V}_t)^{-1} \mathbf{X}_t{}^T \mathbf{Y}_t$ ▷ regularized least squares estimator
    Let $\hat{y}_{t,i} = \langle \hat{\beta}_t, x_{t,i} \rangle$ for each $x_{t,i} \in C_t$
    Let $w_{t,i} = ||x_{t,i}||_{(\bar{V}_t)^{-1}} (\sqrt{2d \log(\frac{1+t/\gamma}{\delta})} + \sqrt{\gamma})$
    Let $[\ell_{t,i}, u_{t,i}] = [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$ ▷ Conf. int. for $\hat{y}_{t,i}$
    if ExactBool **then**
      PICK $(m, \{(x_{t,i}, [\ell_{t,i}, u_{t,i}])\})$
    else PICK$_{\leq}$ $(m, \{(x_{t,i}, [\ell_{t,i}, u_{t,i}])\})$
    Update design matrices $\mathbf{X}_{t+1} = \mathbf{X}_t :: X_t, \mathbf{Y}_{t+1} = \mathbf{Y}_t :: Y_t$.
procedure PICK$(m, (x_{t,1}, [\ell_{t,1}, u_{t,1}]), \ldots, (x_{t,k}, [\ell_{t,k}, u_{t,k}]))$
  Let $M = C_t$
  Let $P_t = \emptyset$
  while $|P_t| < m$ **do**
    Let $x_{t,\hat{i}} = \text{argmax}_{x_{t,i} \in M} u_{t,i}$ ▷ Highest UCB not yet selected
    Let S_t be the set of actions in C_t chained to $x_{t,\hat{i}}$ ▷ Highest chain not yet selected
    if $|S_t| \leq m - |P_t|$ **then**
      $P_t = P_t \cup S_t$ ▷ Take the chain with probability 1
      $M = M \setminus S_t$
    else
      Let Q_t be $m - |P_t|$ actions chosen UAR from S_t
      Let $P_t = P_t \cup Q_t$ ▷ fill remaining capacity UAR from the chain
  Play P_t
procedure PICK$_{\leq}(m, (x_{t,1}, [\ell_{t,1}, u_{t,1}]), \ldots, (x_{t,k}, [\ell_{t,k}, u_{t,k}]))$
  Let $P_t = \{$all actions chained to any $x_{t,i} \in C_t$ with $u_{t,i} > 0\ \}$
  Let $M = C_t$
  Let $P_t = \emptyset$
  while $|P_t| < m$ and $u_{t,x_{t,\hat{i}}} > 0$ for $x_{t,\hat{i}} = \text{argmax}_{x_{t,i} \in M} u_{t,i}$ **do**
    Let S_t be the set of actions in C_t chained to $x_{t,\hat{i}}$ ▷ Highest chain not yet selected
    if $|S_t| \leq m - |P_t|$ **then**
      $P_t = P_t \cup S_t$ ▷ Take the chain with probability 1
      $M = M \setminus S_t$
    else
      Let Q_t be $m - |P_t|$ actions chosen UAR from S_t
      Let $P_t = P_t \cup Q_t$ ▷ fill remaining capacity UAR from the chain
  Play P_t

Figure 3: RIDGEFAIR$_m$, a fair no-regret algorithm for picking $\leq m$ actions whose payoffs are linear.

procedure FairGap(δ, C, λ)
  for $t \geq 1$ **do**
    if $2r\ln(2dt/\delta)/\lambda \geq t$ **then**
      Play $\hat{x}_t \sim_{\text{UAR}} C_t$
      Update design matrices $\mathbf{X}_{t+1}, \mathbf{Y}_{t+1}$
    else
      Let $\delta = \min(\delta, 1/t^{1+c})$
      Let $\hat{\beta}_t = ({\mathbf{X}_t}^T\mathbf{X}_t)^{-1}{\mathbf{X}_t}^T\mathbf{Y}_t$ ▷ Least squares estimator
      Let $\kappa = 1 - r\sqrt{2\ln(2dt/\delta)/t\lambda}$
      Let $w_t = \frac{r^2 \cdot R \cdot 2\sqrt{\ln(2t\delta)}}{k\lambda\sqrt{t}}$ ▷ Confidence interval width
      Let $(x_1, x_2) = \text{TopTwo}(C_t, \hat{\beta}_t)$ ▷ Find two ext. pts. maximizing $\langle x, \hat{\beta}_t \rangle$
      Let $U_1 = [\langle \hat{\beta}_t, x_1 \rangle - w_t, \langle \hat{\beta}_t, x_1 \rangle + w_t]$
      Let $U_2 = [\langle \hat{\beta}_t, x_2 \rangle - w_t, \langle \hat{\beta}_t, x_2 \rangle + w_t]$
      if $U_1 \cap U_2 = \emptyset$ **then**
        Let FoundMax $= \{x\}$
        Play $\hat{x}_t = x$ ▷ Play $\hat{x}_t$ once confidence intervals separate
      else
        Play $\hat{x}_t \sim_{\text{UAR}} C_t$
        Update design matrices $\mathbf{X}_{t+1}, \mathbf{Y}_{t+1}$

Figure 4: FairGap, a fair no-regret algorithm for infinite, changing action sets.

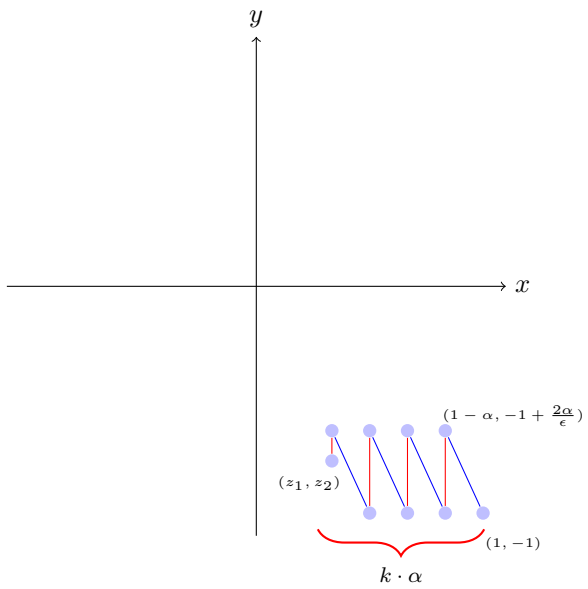

Figure 5: A path connecting $(1, -1)$ to an arbitrary point (z_1, z_2): red segments are vertically equivalent, blue segments are diagonally equivalent.

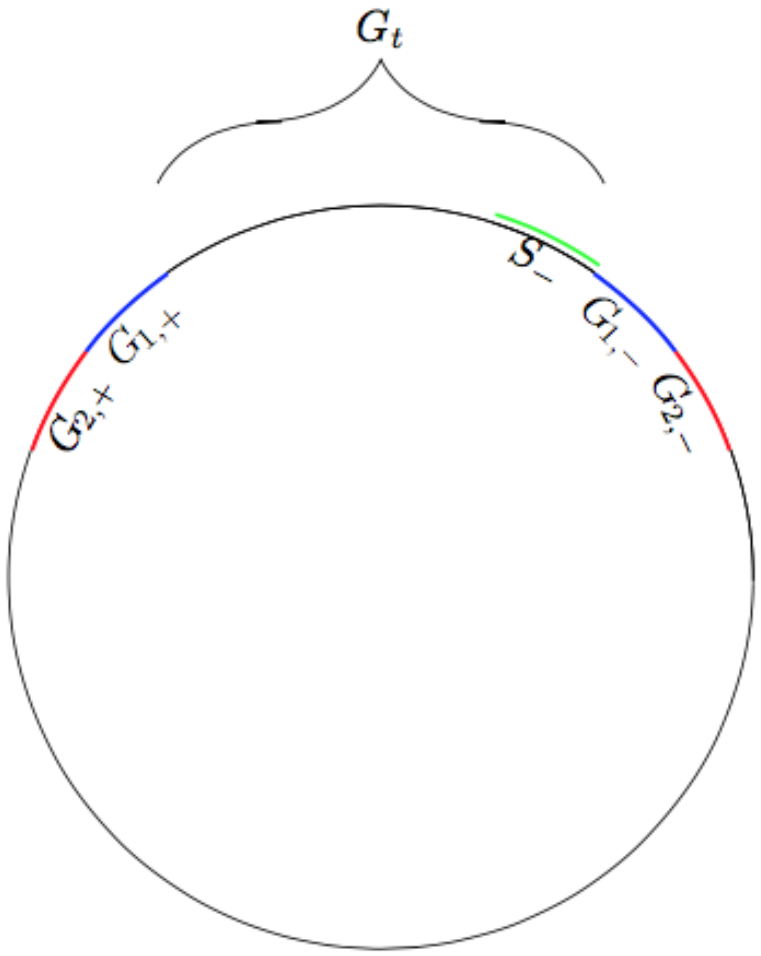

Figure 6: $\mathcal{A}$ must play UAR from $D = S^1$. $|G_t| = \epsilon$; $|S'_{t,-}| = |S_{t,-}| = \frac{3\epsilon}{4}$; $|S'_{t,+}| = |S_{t,+}| = |G_{1,+}| = |G_{1,-}| = |G_{2,+}| = |G_{2,-}| = \epsilon$

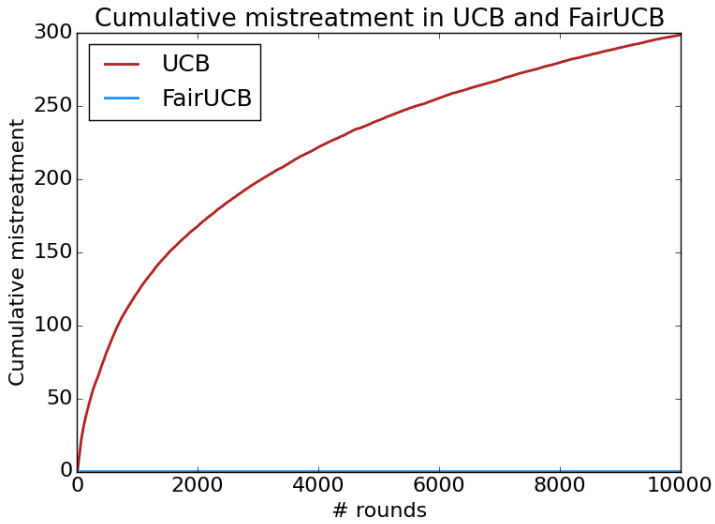

Figure 7: Cumulative mistreatments for UCB and FairUCB.

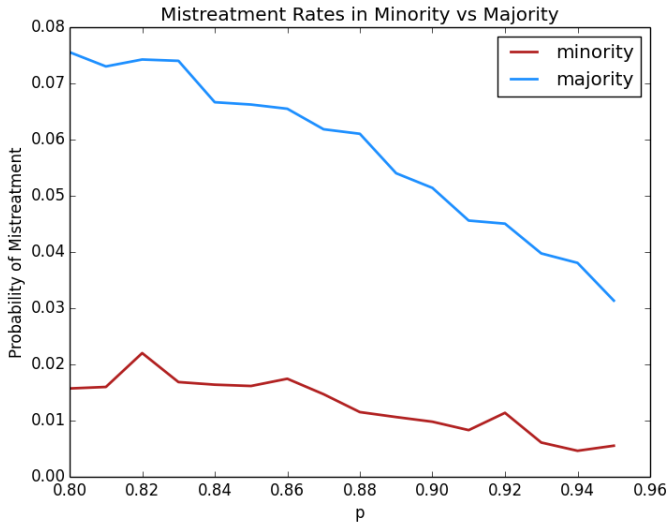

Figure 8: Probability of mistreatment for subpopulations under UCB.