Better Fair Algorithms for Contextual Bandits *

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, Aaron Roth
University of Pennsylvania

Abstract

We study fairness in the linear bandit setting. Starting from the notion of meritocratic fairness introduced in Joseph et al. [11], we introduce a sufficiently more general model in which meritocratic fairness can be imposed and satisfied. We then perform a more fine-grained analysis which achieves better performance guarantees in this more general model. Our work therefore studies fairness for a more general problem and provides tighter performance guarantees than previous work in the simpler setting.

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.

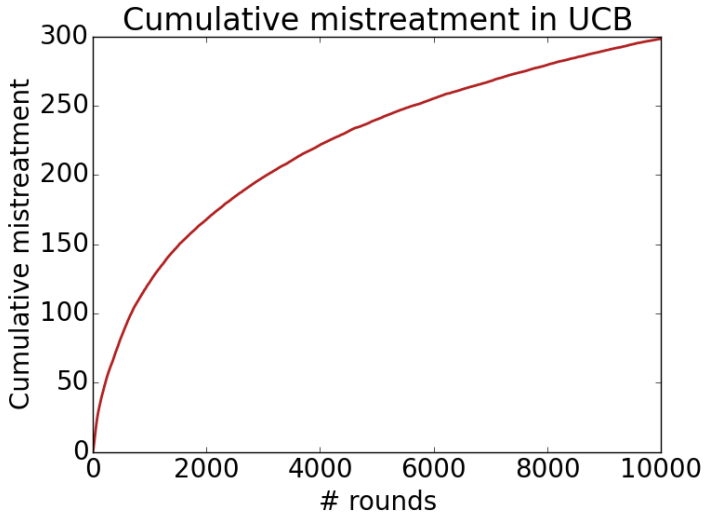
Joseph et al. [11], however, raises the concern that these no-regret policies may be *unfair*: in some rounds, they will choose options with lower expected rewards over options with higher expected rewards, for example choosing less qualified job applicants over more qualified ones. Consider a UCB-like 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, and mistreatment may accrue to different subpopulations at drastically different rates (see Figure 1).

Joseph et al. [11] 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 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. Section 3 shows how to design fair algorithms when at most some finite number k of individuals arrives in any round (which corresponds to the linear contextual bandits problem [2, 4]), as well as when m individuals may be chosen in each round (which corresponds to the "multiple play" introduced and studied absent fairness in Anantharam et al. [3]). Our work therefore both focuses on a much more general model than Joseph et al. [11] and substantially improves upon their black-box regret guarantees for linear bandit problems using a technical analysis specific to the linear setting. We condense our results as follows:

*The full technical version of this paper is available at https://arxiv.org/abs/1610.09559.



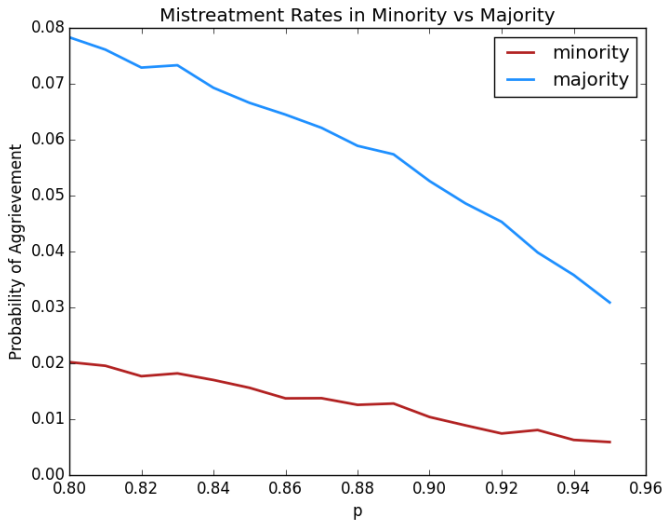


Figure 1: Cumulative mistreatments for UCB.

1. In the m-bandit case where we must select exactly m options in each round, we play all of our chains in descending over, randomizing over the last chain as necessary to select exactly m options, and obtain regret $R(T) = \tilde{O}\left(dkm\sqrt{T}\right)$.

2. In the k-bandit case where we can select any number $\leq k$ of options in each round, we deterministically select every option in every chain with highest UCB > 0 and obtain regret $R(T) = \tilde{O}\left(dk^2\sqrt{T}\right)$.

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. [5] and Corbett-Davies et al. [7] 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. [10] introduced the notion of *equality of opportunity*, which requires that a classifier's predicted outcome should be independent of a protected attribute (such as race) conditioned on the true outcome, and they and Woodworth et al. [13] have studied the feasibility and possible relaxations thereof. Similarly, Zafar et al. [14] analyzed an equivalent concurrent notion of (un)fairness they call *disparate mistreatment*. Separately, Kleinberg et al. [12] and Chouldechova [6] showed that different notions of group fairness may (and sometimes must) conflict with one another.

This paper, like Joseph et al. [11], 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. [8], 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. [11] 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[1] 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.

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

[1] Friedler et al. [9] provide evidence that providing fairness from bias-corrupted data is quite difficult.

from Joseph et al. [11], we work in what we believe to be a 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. [11].

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, i.e. has tails dominated by those of a Gaussian distribution.

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 $P_1, \ldots, P_T$.[2] Then, the expected reward of $\mathcal{A}$ on this sequence is $\mathrm{Rew}(\mathcal{A}) = \mathbb{E}\left[\sum_{t \in [T]} \left[\sum_{x_t \in P_t} y_t\right]\right]$.

Refer to the sequence of feasible choices[3] 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

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

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. [11], 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 $x, x' \in C_t$ if $\langle \beta, x \rangle \geq \langle \beta, x' \rangle$ then $\pi_t(x) \geq \pi_t(x')$. 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 condition.

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)

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

[3] 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.

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. [11]. 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. [11] 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 [1].

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.

References

[1] 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.

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

[3] 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.

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

[5] 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.

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

[7] 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.

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

[9] 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`.

[10] 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`.

[11] 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.

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

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

[14] 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.