Rawlsian Fairness for Machine Learning

Matthew Joseph * Michael Kearns * Jamie Morgenstern * Seth Neel †
Aaron Roth *

April 18, 2017

Abstract

Motivated by concerns that automated decision-making procedures can unintentionally lead to discriminatory behavior, we study a technical definition of fairness modeled after John Rawls' notion of "fair equality of opportunity". In the context of a simple model of online decision making, we give an algorithm that satisfies this fairness constraint, while still being able to learn at a rate that is comparable to (but necessarily worse than) that of the best algorithms absent a fairness constraint. We prove a regret bound for fair algorithms in the linear contextual bandit framework that is a significant improvement over our technical companion paper [16], which gives black-box reductions in a more general setting. We analyze our algorithms both theoretically and experimentally. Finally, we introduce the notion of a "discrimination index", and show that standard algorithms for our problem exhibit structured discriminatory behavior, whereas the "fair" algorithms we develop do not.

*majos, mkearns, jamiemor, aaroth@cis.upenn.edu. Department of Computer and Information Sciences, University of Pennsylvania.

†sethneel@wharton.upenn.edu. Department of Statistics, The Wharton School, University of Pennsylvania.

Contents

1 Introduction

Automated techniques from statistics and machine learning are increasingly being used to make important decisions that directly affect the lives of individuals, including hiring [25], lending [8], policing [30], and criminal sentencing [6]. These high-stakes uses of machine learning have led to increasing concern in law and policy circles about the potential for (often opaque) machine learning techniques to be discriminatory or unfair [11, 5, 26]. These concerns are more than hypothetical. For example, a 2016 ProPublica study [4] of the COMPAS Recidivism Algorithm (used to inform criminal sentencing decisions by attempting to predict recidivism) found that the algorithm was significantly more likely to incorrectly label black defendants as recidivism risks compared to white defendants, despite similar overall rates of prediction accuracy between groups.[1] Despite the recognized importance of this problem, very little is known about technical solutions to the problem of algorithmic unfairness,[2] or the extent to which "fairness" is in conflict with the goals of learning, which typically emphasize predictive accuracy and computational efficiency.

In this paper, we propose a mathematically precise definition of fairness in machine learning for sequential decision making, and analyze its implications on the quality of learning, both theoretically and experimentally. As a running expository example, we shall consider a scenario in which a machine learning algorithm is making decisions about whom to grant loans to, based on the data provided about applicants. There are several extant definitions of fairness in machine learning, most of which formalize some notion of "group fairness", a constraint that binds at the group level (see e.g. [9, 24, 18, 13, 14] for a sample of papers studying such definitions). One such notion is that of *statistical parity*, which requires that the fraction of individuals granted loans by the algorithm should be equal across "protected" groups (e.g. race, ethnicity, or gender).

However, as discussed in Dwork et al. [12], these group-level definitions often fail at both fairness and accurate learning. If two groups actually have different proportions of individuals who are able to pay back their loans, then the accuracy of any learning algorithm will obviously suffer when constrained to predict an equal proportion of paybacks for the two groups. Furthermore, such definitions do not guarantee that a truly creditworthy individual from one group has an equal chance of being given a loan as a similarly creditworthy individual from another group. With these problems in mind, Dwork et al. [12] advocate that technical definitions of fairness should focus on *individual* fairness, rather than fairness at the group level.

Our work explores the consequences of one such definition of individual fairness, and can be viewed as a mathematical formalization of Rawls' notion of "fair equality of opportunity"[3]:

> *"... assuming there is a distribution of natural assets, those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system."* [27]

This Rawlsian view of fairness is consistent with much of the legal doctrine on non-discrimination. For instance, Title VII of the 1964 Civil Rights Act prohibits "disparate impact discrimination" —

[1] The company that sells COMPAS has raised methodological concerns about the ProPublica study, but it is apparent that not only are some of the objections raised valid in the case of COMPAS, they are more generally unavoidable [19].

[2] In a recent speech FTC Commissioner Julie Brill [17] observed, "... a lot remains unknown about how big data-driven decisions may or may not use factors that are proxies for race, sex, or other traits that U.S. laws generally prohibit from being used in a wide range of commercial decisions ... What can be done to make sure these products and services–and the companies that use them treat consumers fairly and ethically?"

[3] An interesting recent paper [15] gives a different formalization of the idea of "equality of opportunity" in a classification setting. The definition suggested by [15] is a "group fairness" constraint on the *average* behaviour of the algorithm over all individuals of the same quality. In contrast, in our work, we impose a stronger constraint that binds at the individual level.

and therefore forbids not only explicit discrimination, in which choices are directly based on race or other protected attributes, but also discrimination resulting from policies that put members of a protected group at a disadvantage. Interpreting Title VII, Supreme Court Chief Justice Warren Burger writes in *Griggs v. Duke Power Co. (1971)*:

> *"Nothing in the Act precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance. Congress has not commanded that the less qualified be preferred over the better qualified simply because of minority origins. Far from disparaging job qualifications as such, Congress has made such qualifications the controlling factor, so that race, religion, nationality, and sex become irrelevant. What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract."* [1]

We formalize our Rawlsian definition of fairness in a sequential decision-making framework that is known as the *contextual bandits* setting in machine learning. Continuing with our loan example, on each day t we are given k loan applications, one each from a known set of k distinct groups.[4] The application for the individual from group i comes in the form of a vector $x_{t,i} \in \mathbb{R}^d$, which summarizes their salient properties (e.g. $x_{t,i}$ might include attributes such income, credit history, employment history, educational background, etc.). We assume that for each group i there is some (unknown) function $f_i : \mathbb{R}^d \to \mathbb{R}$ which maps these attribute vectors to the "quality" of an applicant. In our example, this could represent the probability of repaying a loan. Note that this framework explicitly allows that the functions f_i mapping attributes to quality might be different for different groups— which in fact might be *necessary* to achieve fairness. For example, while having a college degree might be strongly predictive of creditworthiness in the overall group, it might be less so for individuals from a protected subgroup that generally had few financial resources growing up. For this group, attributes related to employment history might be more predictive. Insisting on one model for the entire group could lead to discriminating against highly creditworthy members of the protected group simply because they did not attend college.

Our notion of fairness asks that at *every* step, a learning algorithm must never "favor" (that is, choose with higher probability) an applicant whose true quality (as determined by their attributes and their group's mapping) is lower than that of another applicant. Note that if the mappings f_i are known perfectly, it would always be fair to simply choose the candidate of highest quality at each step. Thus fairness and optimal decision-making are perfectly aligned (see Burger's comments above on choosing the most qualified candidates). The challenge arises from the fact that the f_i are initially unknown — they must be learned from data — but we ask that a learning algorithm be fair *at every step*, not just asymptotically. This requirement is especially important in an era in which large-scale machine learning systems are perpetually learning and refining their models, and thus should not be "forgiven" for unfair decisions made during their training process.

Our primary interest is in designing learning algorithms that can (provably) converge to optimal decision-making, while being (provably) fair at every step. We seek to quantify the frictions or trade-offs between fairness and fast convergence to optimality, and thus compare (both theoretically and experimentally) our fair learning algorithms with standard algorithms that are unconstrained by fairness.

We focus on the case in which the underlying quality of each individual is governed by the classic ordinary least squares model: each group i has an underlying weight vector β_i on its observable

[4]Note that both the assumption that only one loan is given and exactly one applicant from each group arrives on each day are for simplicity of exposition. Both can be relaxed in our results with mild degradation in performance guarantees.

attributes, so that each individual x in the group has expected true quality $\beta_i \cdot x$. The decision maker (bank) then learns (approximate) β_i by choosing individuals (granting loans) from the various groups and receiving noisy feedback about their true quality (observing repayment). Crucially, the decision maker *does not* observe the quality of the individuals not served; this models the fact that a bank does not observe whether an individual denied a loan *would have* paid it back. This setting is a generalization of the classical "multi-armed bandit" problem [28, 29, 21] called the *contextual bandit* problem [23, 10].

Our main theoretical results are positive: we describe simple learning algorithms that are provably fair, and for which the cost of fairness is small in the sense that their rate of convergence to optimal decisions is only a factor $k\sqrt{d}$ worse than the best non-fair algorithms, where d is the context dimensionality and k number of distinct groups. We complement these theoretical results with an empirical evaluation and comparison of our fair algorithms with standard non-fair approaches. Our empirical results support the theory's prediction that after a period of "fair exploration" based on the notion of *chained* confidence intervals, our algorithms become competitive with the non-fair algorithms in terms of predictive decision-making performance. In our technical companion paper [16], we generalize the framework presented here beyond linear functions, and relax several other simplifying assumptions made here (at the cost of less tight quantitative bounds).

2 Definitions: Learning and Fairness

In this section, we describe the formal model that we study. A problem instance is defined by a domain $\mathcal{X}$ from which the salient features of each individual are drawn, which we take to be $\mathcal{X} = \mathbb{R}^d$, and a set of k different groups, indexed by $i \in \{1, \ldots, k\}$. Each group j is endowed with an unknown function $f_i : \mathcal{X} \to \mathbb{R}$ mapping the features of an individual from group i to their true "quality". In this paper, these functions have a linear form: that is, there is some unknown vector of coefficients $\beta_i \in \mathbb{R}$ such that $f_i(x) = \beta_i \cdot x$.

In rounds $t = 1, \ldots, T$, an individual from each of the k groups arrives, and the salient features $x_{t,i}$ of each are observed by the learning algorithm $\mathcal{A}$. In this paper, we assume that the features of the individual from group i that arrives at time t are drawn independently from a distribution $\mathcal{D}_i$, which may be different for each group: $x_{t,i} \sim \mathcal{D}_i$. The algorithm must then *choose* one of the individuals i_t (e.g. to grant a loan to), and observes a *reward* r_{t,i_t} from the individual it chose (e.g. the payoff on the granted loan). The reward is stochastically generated and equal to the quality of an individual plus noise generated from a standard Gaussian distribution: $r_{t,i} = f_i(x_{t,i}) + e_{t,i}$, where $e_{t,i} \sim N(0, 1)$. In other words, observed individual qualities follow a standard ordinary least squares model, which may have different parameters for each group. Crucially, the algorithm does *not* observe the reward for those individuals not chosen. This leads to the classical tension between *exploration* (choosing potentially suboptimal individuals in order to learn more about their groups) and *exploitation* (serving those individuals who seem to be best qualified given current knowledge about the groups).

We measure the performance of learning algorithms via *regret*, the difference between the expected reward of the *optimal policy* and the expected reward of the algorithm. The (omniscient) optimal policy chooses the individual with highest expected reward every day, and so if p_t denotes the distribution over choices at round t for an algorithm $\mathcal{A}$, we define the regret of $\mathcal{A}$ by:

$$\text{Regret}(x_1, \ldots, x_T) = \sum_t \max_i \left(f_i(x_{t,i})\right) - \sum_t \mathbb{E}_{\hat{i}_t \sim p_t}\left[f_{\hat{i}_t}(x_{t,\hat{i}_t})\right].$$

We say that $\mathcal{A}$ satisfies regret bound $R(T)$ if

$$\max_{\mathcal{D}1,\ldots,\mathcal{D}_k} \mathbb{E}_{x_{t,1}\sim\mathcal{D}_1,\ldots,x_{t,k}\sim\mathcal{D}_k}\left[\text{Regret}(x_1,\ldots,x_T)\right] \leq R(T).$$

Let the history $h_t \in \left(\mathcal{X}^k \times [k] \times \mathbb{R}\right)^{t-1}$ be a record of the $t-1$ rounds experienced by $\mathcal{A}$ up until round t. Thus h_t encodes, for each of the $t-1$ rounds, 3 things: the attribute vectors of all k individuals observed, the index of the individual chosen, and the chosen reward observed. The history is a sufficient statistic to determine the distribution of the algorithm's choices at the next round. We write $\pi^t_{j|h_t}$ to denote the probability that $\mathcal{A}$ chooses individual j after observing features $x_{t,i}$, given h_t. For notational simplicity, we will often drop the superscript t on the history when referring to the distribution over individuals: $\pi^t_{i|h} = \pi^t_{i|h_t}$.

We now define what it means for an algorithm to be fair. Informally, this will mean that with high probability (with respect to the randomness inherent in noisy observations, and any probabilistic decisions made by the algorithm), at every round, and for any set of individuals the algorithm is presented at that round, the algorithm will choose a higher quality individual with probability at least that with which it chooses a lower quality individual. More formally:

Definition 1 (Fairness)**.** *Algorithm $\mathcal{A}$ is* **fair** *if, for any input $\delta > 0$, with probability at least $1-\delta$ over the realization of the history h, for all rounds $t \in [T]$ and all pairs of choices $i, j \in [k]$, if $f_i(x_{t,i}) > f_j(x_{t,j})$ then $\pi^t_{i|h} \geq \pi^t_{j|h}$.*

This definition requires that, with high probability, the algorithm does not make a discriminatory decision at *any* round t, regardless of the specific individuals being compared at each round. A weaker condition would require that the fraction of discriminatory rounds is small. Under this alternative definition, it is permissible for an algorithm to always be discriminatory under certain infrequent but specific conditions — such as always preferring less qualified white male lawyers to more qualified female scientists when such comparisons arise — so long as it is not discriminatory on a large fraction of rounds. In contrast, our (stronger) definition requires fair treatment of the specific individuals who arrive at *every* round, thus precluding this type of rare-subgroup discrimination.

3 Provably Fair and No-Regret Algorithms

One immediate consequence of our fairness definition is that the optimal policy which deterministically plays the highest quality individual at each round satisfies our fairness constraint. This suggests that the goal of fairness is not intrinsically at odds with the goal of accuracy (i.e. regret minimization). When designing fair algorithms, then, a natural starting place is the class of well-studied bandit algorithms that do not obey a fairness condition, but instead purely aim for accuracy [29]. A defining characteristic of this class of algorithms is that they balance *exploration* (choosing individuals from groups about which the algorithm has high uncertainty to learn more about that group) and *exploitation* (choosing individuals that have the highest estimated quality based on the algorithm's observations so far). As the algorithm gradually learns the true parameters, the *exploration* steps become increasingly infrequent, until the algorithm is essentially playing optimally. Optimal learning algorithms thus delicately balance exploration and exploitation.

However, both of these steps can result in unfair play: exploration steps can be unfair because they implicitly favor uncertainty rather than quality, and exploitation steps can be unfair if the algorithm misidentifies the best individual. In contrast, a fair algorithm must carry out this exploration and exploitation without violating fairness. A key observation here is that uniformly random

selection is fair, as all individuals are selected with the *same* probability. Since, as mentioned above, optimal selection is also fair, this suggests a general outline for a fair algorithm: begin by exploring uniformly at random and gradually transition to optimal exploitation, while preserving fairness throughout. This raises a central challenge in the design of fair no-regret algorithms: in order to both satisfy fairness and guarantee good performance, an algorithm must *quickly* transition from uniformly random to near-optimal play. Much of the technical difficulty in designing fair algorithms stems from managing this transition such that the algorithm does not violate the fairness constraint at any step along the way.

We now present two closely related algorithms, IntervalChaining and RidgeFair, both with accompanying regret bounds. We present IntervalChaining first for its simplicity, and will use IntervalChaining in our experiments. We then present RidgeFair, which obtains a superior theoretical regret guarantee even in the more general setting in which contexts are adversarially chosen. IntervalChaining achieves regret

$$R(T) = \tilde{O}\left(\sqrt{k^3 d} \cdot T^{2/3} + \sqrt{d^3 k^3}\right)$$

in many settings but depends on a distribution dependent constant that can in principle be arbitrarily large. We postpone the proof to the supplement, since RidgeFair achieves a stronger regret bound with fewer assumptions. For RidgeFair, a sharper technical analysis yields a regret bound of

$$R(T) = \tilde{O}(d\sqrt{k^3 T})$$

In our companion paper [16] we prove a lower bound of $R(T) = \Omega(k^3)$ for fair algorithms in the multi-armed bandit setting (i.e. in the special case in which contexts are unchanging between rounds). A lower bound of $R(T) = \Omega(\sqrt{T})$ is also known for this setting, even absent a fairness constraint [7]. Since that simpler setting is a special case of the contextual bandit framework, it follows that our dependence on k and T is optimal up to logarithmic factors. In [3] an $R(T) = \Omega(\sqrt{Td})$ lower bound is proven for linear contextual bandit algorithms absent a fairness constraint, along with an algorithm enjoying a matching $\tilde{O}(\sqrt{Td})$ upper bound. Taken together, these results show that the regret bound achieved by RidgeFair has optimal dependence (for fair algorithms) on T and k, with the possibility of at most a $\sqrt{d}$ improvement in d. The *cost* of fairness, or the degradation of the regret guarantee due to the fairness constraint, is thus between $\tilde{O}(k)$ and $\tilde{O}(k\sqrt{d})$. Finally, we note that this is a significant improvement over the $R(T) = \tilde{O}(T^{4/5}k^{6/5}d^{3/5})$ bound for the linear contextual bandit setting given in our companion paper [16], which is derived as a corollary of a general black box reduction rather than with a specialized analysis.

3.1 IntervalChaining

We now present IntervalChaining, a provably fair algorithm with strong performance guarantees. In IntervalChaining, round t is chosen independently with probability η_t to be an *exploration* round, in which the algorithm chooses uniformly at random among all individuals to learn better estimates of the k linear models. All other rounds are *exploitation* rounds, in which the algorithm uses those estimates to choose high-quality individuals, to the extent possible subject to the fairness constraint (the algorithm also improves its model estimates from the data gathered in these rounds). In exploitation rounds, IntervalChaining uses OLS estimators $\hat{\beta}_{t,i}$ for each group i in each round t to compute estimated qualities $\hat{y}_{t,i} = \hat{\beta}_{t,i} \cdot x_{t,i}$ for each individual $x_{t,i}$. The algorithm also computes *confidence interval widths* $w_{t,i}$ around these estimates, such that the true qualities lie within these confidence intervals: $y_{t,i} \in [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$, with probability $1 - \delta$, for all i and t. Using these confidence intervals, the algorithm finds the individual with highest upper confidence bound i^t_*.

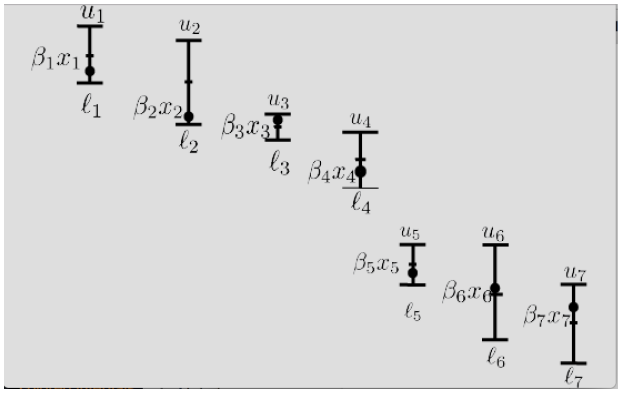


Figure 1: An example of the confidence intervals around the estimated values (represented by $-$), which contain the true values $\mu_i = \beta_i x_i$ for each group i. In this example, individuals 1 and 2, 2 and 3, 3 and 4, 5 and 6, and 6 and 7 overlap. The set of arms $\{1, 2, 3, 4\}$ are all chained to the top interval individual, individual 1.

procedure IntervalChaining(δ, T, η)
  for t from 1 to T **do**
    Let $r \in_R [0, 1]$
    if $r < \eta_t$ **then**
      Play $\hat{i}^t \in_R \{1, \ldots, k\}$
    else
      for i from 1 to k **do**
        Let $\hat{\beta}_{t,i}$ = OLS estimate of β_i
        Let $\hat{y}_{t,i}$ = OLS estimate of $\beta_i \cdot x_{t,i}$
        Set confidence interval width $w_{t,i}$
        Let $[\ell_i^t, u_i^t] = [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$
      Let $i_*^t = \arg\max_i u_i^t$
      Let S_t be the set of actions chained to i_*^t
      Choose and play $\hat{i}^t \in_R S_t$
    Let $y_{t,\hat{i}^t}$ be the observed feedback

Figure 2: IntervalChaining, a fair algorithm for the linear contextual bandit problem. $X_{t,i}$ refers to the set of observed feature vectors $x_{t',i}$ for action i at times $t' < t$ for which the algorithm observed $y_{t',i}$; $Y_{t,i}$ to these observed $y_{t',i}$ values.

Standard no-regret algorithms would simply choose i_*^t, but this will not in general be fair, since the individual with highest upper confidence bound is not necessarily the highest quality individual. Instead, our algorithm chooses uniformly at random amongst S_t, the set of individuals *chained* to i_*^t, defined as follows. For any individual $x_{t,i}$ whose confidence interval *overlaps* with i_*^t's ($[\ell_i^t, u_i^t] \cap [\ell_{i_*^t}^t, u_{i_*^t}^t] \neq \emptyset$), the algorithm does not have enough data to know with confidence which of $x_{t,i}$ or x_{t,i_*^t} has higher quality. The algorithm therefore behaves conservatively with respect to these estimates to guarantee fairness with respect to any qualities consistent with these confidence intervals: it will treat any pair of individuals whose confidence intervals overlap equally (namely, will choose them with equal probability). This motivates the following definition of chaining, which is the transitive closure of the "overlapping" relation. If the confidence intervals around i and j overlap, then we will say that i and j are chained. Further, if i and j are chained, and j and j' are chained, then i and j' are chained (even if the confidence intervals around i and j' do *not* overlap). See Figure 1 for a visual representation of chaining, and Figure 2 for the pseudo-code for IntervalChaining.

IntervalChaining, is fair and satisfies the following performance guarantee, when implemented with $\eta_t, w_{t,i}$ as defined in the formal statement and proof in the supplement.

Theorem 1. IntervalChaining *is fair, and has regret*

$$R(T) = O\left(\sqrt{k^3 d \cdot \frac{\ln \frac{2kT}{\delta}}{\ell}} T^{2/3} + \left(\frac{dkL}{\ell}\left(\ln^2 \frac{2kT}{\delta} + \ln d\right)\right)^{3/2}\right)$$

where $\ell = \min_i \mathbb{E}_{x_{t,i}\sim\mathcal{D}_i}\left[\lambda_{min}\left(\sum_{t\in[d]} x_{t,i}^T x_{t,i}\right)\right]$ *(the minimum over* i *of the expectation of the minimum eigenvalue of the random design matrix of* d *observations from group* i*) and* $L \geq \max_t \lambda_{max}\left(x_{t,i}^T x_{t,i}\right)$*, assuming* $||x_{t,i}||, ||\beta_i|| = O(1)$.

We note briefly that, in many interesting settings, the regret bound given in Theorem 1 reduces to $\tilde{O}\left(\sqrt{k^3 d} \cdot T^{2/3} + \sqrt{d^3 k^3}\right)$.[5] If, for example, the distribution over feature vectors has $||x_{t,i}|| = 1$, $\mathbb{E}[x_{t,i,j}] = 0$, and $x_{t,i,j}$ and $x_{t,i,j}$ are independent for all i, t and j, j', the regret bound has this form.

We now present a high-level outline of the techniques used to bound the regret of IntervalChaining (and RidgeFair).

Proof Sketch: We define the confidence interval widths $w_{t,i}$ such that, with probability $1-\delta$, for all rounds t and all groups i, $y_{t,i} \in [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$. So, for any pair of groups i, j, if $y_{t,i} > y_{t,j}$, then $\hat{y}_{t,i} + w_{t,i} \geq \hat{y}_{t,j} - w_{t,j}$: if j is chained to i_*^t, then so will be i. Since the algorithm chooses uniformly from the set of individuals chained to the individual with highest upper confidence interval, for every pair of individuals, either they are selected with identical probabilities, or else one of them is selected with probability 0 (but has lower quality unless the confidence intervals have failed).

The regret guarantee follows by calculating, for each group, a lower bound on the number of rounds for which the algorithm has selected a member of that group (and hence has obtained a data point with which to update its OLS estimator). Given that lower bound, we upper bound the width of the confidence interval around individuals from that group. Since chains can have length at most k, any individual chained to the individual with highest upper confidence bound has quality that differs from the quality of the best individual by a term that is at most k times this upper bound on the confidence width. Summing up over all rounds yields the desired result. □

3.2 RidgeFair

In this section we define RidgeFair, observe that it is fair by the same argument we saw for IntervalChaining, and prove its regret bound. RidgeFair estimates β using confidence regions centered at the ℓ_2-regularized least squares estimator, and then translates these regions into confidence intervals for each group payoff at each time step. The algorithm then follows a similar chaining procedure, and plays a group uniformly at random from the set of arms chained to the top arm at every time t. The regularized least squares estimator given a design matrix X, response vector y, and regularization parameter $\lambda \geq 1$, is of the form $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$. Using this ridge estimator decreases the variance in each groups's estimated quality, but as a result the estimated payoff is no longer an unbiased estimate for the true payoff mean. Consequently valid

[5] $\tilde{O}$ hides constants and logarithmic factors.

confidence intervals are harder to derive than for the simple OLS estimator, and rely on martingale techniques borrowed from [2], which derives similar bounds absent a fairness constraint.

RidgeFair thus possesses a number of theoretical advantages over IntervalChaining. First, its narrower confidence intervals allow us to derive our tightest regret bound for a fair algorithm, $\tilde{O}(d\sqrt{k^3T})$, under far less restrictive assumptions. We no longer need to assume normally distributed noise, instead requiring only that noise is mean zero and R sub-Gaussian. Moreover, we can assume contexts are selected adversarially from a bounded set, rather than drawn i.i.d. from a distribution, as we no longer require the Matrix Chernoff bound that powers the regret bounds for IntervalChaining. For the analysis below we assume for simplicity that the noise is R sub-Gaussian with $R = 1$, and that all contexts lie in the unit ball. This is without loss of generality up to scaling — any scaling parameters will appear in the final bound.

procedure RidgeFair($\delta, T, \lambda \geq 1$)
  for $t \geq 1, 1 \leq i \leq k$ **do**
    Let X_i, Y_i = design matrix, observed payoffs for i
    Let $x_{t,i}$ = feature vector for arm i in round t
    Let $\bar{V}_{it} = X_i^T X_i + \lambda I$
    Let $\hat{\beta}_{it} = (\bar{V}_{it})^{-1} X_i^T Y_i$ ▷ regularized least squares estimator
    Let $\hat{y}_{t,i} = x_{t,i}^T \hat{\beta}_{it}$
    Let $w_{t,i} = ||x_{t,i}||_{\bar{V}_{it}^{-1}} (R\sqrt{d \log(\frac{1+t/\lambda}{\delta})} + \sqrt{\lambda})$
    Let $[\ell_i^t, u_i^t] = [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$ ▷ Conf. int. for $\hat{y}_{t,i}$
    Let $i_*^t = \arg\max_i u_i^t$
    Let S_t be the set of actions chained to i_*^t
    Play uniformly at random among all arms in S_t
    Update design matrices X_i, Y_i.

Figure 3: RidgeFair, a fair algorithm for the linear contextual bandit problem with provable sublinear regret.

We now formally state that RidgeFair is both δ-fair and has a sublinear regret guarantee.

Theorem 2. *Assume that all for all contexts $||x_{t,i}|| \leq 1$, that $||\beta_i|| \leq 1$ for all i, $\lambda = 1$, and the noise is R sub-Gaussian with $R = 1$. Then* RidgeFair *is fair, and has regret $\tilde{O}(d\sqrt{k^3T \ln \frac{1}{\delta}})$ for all T.*

Remark. Note that the $\tilde{O}$ hides logarithmic factors in $1/\delta$, and T. The assumptions of at most unit norm for contexts and parameters are only for convenience, and the argument works equally well for both lying in a bounded set.

Before proceeding with the proof, we state a theorem we will use in its proof.

Theorem 3 (From [2])**.** *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\forall t \geq 0, \;\; ||X_i^T \eta_i||_{\bar{V}_{it}^{-1}} \leq R\sqrt{d \log\left(\frac{1 + t/\lambda}{\delta}\right)}$$

We now prove our tightest regret bound for any fair algorithm.

Proof of Theorem 2. The fact that RidgeFair is fair follows from an identical argument as the one for IntervalChaining, and the fact that the confidence intervals derived hold over all t with probability $1-\delta$. Toward proving the regret bound, we prove a slightly stronger result. In our definition of regret in the preliminaries, we measure the *expected* regret: the definition takes an expectation over the randomization of the algorithm. In fact, we can show that our regret bound holds with high probability $(1-\delta)$; not only in expectation. We now define

$$R(T) = \text{Regret}(x_1, \ldots, x_T) = \sum_t \max_i (f_i(x_{t,i})) - \sum_t f_{\hat{i}_t}(x_{t,\hat{i}_t}).$$

Taking $\delta = O(\frac{1}{\sqrt{T}})$, and noting that $R(T) \leq T$, we see that an $O(\sqrt{T})$ bound with probability $1-\delta$ on $R(T)$, also implies an $O(\sqrt{T})$ bound on the expected regret. We now proceed with bounding $R(T)$ with high probability.

We adopt the notation in Abbasi-yadkori et al. [2]: let $\bar{V}_{it} = X_i^T X_i + \lambda I$, where X_i is the design matrix at time t corresponding to group $i, \lambda \geq 1$. Let $\hat{\beta}_{it} = \bar{V}_{it}^{-1} X_i^T Y_i$ be the regularized least squares estimator for group i at time t. Consider the feature vector $x_{t,i}$ at time t. For a d-dimensional vector z and a $d \times d$ p.d. matrix A, let $\langle z, z \rangle_A$ denote $z^t A z$. Letting η_i be the noise sequence corresponding to group i, we have $\hat{\beta}_{it} = \bar{V}_{it}^{-1} X_i^T (X\beta_i + \eta_i)$. Then some matrix algebra from [2] shows:

$$x_{t,i} \cdot (\hat{\beta}_{it} - \beta_i) = x_{t,i}^T \bar{V}_{it}^{-1} X_i^T \eta_i - \lambda x_{t,i}^T \bar{V}_{it}^{-1} \beta_i,$$

which using the above notation gives

$$x_{t,i} \cdot (\hat{\beta}_{it} - \beta_i) = \langle x_{t,i}, X_i^T \eta_i \rangle_{\bar{V}_{it}^{-1}} - \lambda \langle x_{t,i}, \beta_i \rangle_{\bar{V}_{it}^{-1}}$$

Applying Cauchy-Schwarz,

$$|x_{t,i} \cdot (\hat{\beta}_i - \beta_i)| \leq ||x_{t,i}||_{\bar{V}_{it}^{-1}} (||X_i^T \eta_i||_{\bar{V}_{it}^{-1}} + \sqrt{\lambda})$$

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

By Theorem 3, and combining the inequalities we get that over all rounds $t \geq 0$ with probability $1-\delta$:

$$x_{t,i} \cdot (\hat{\beta}_i - \beta_i)| \leq ||x_{t,i}||_{\bar{V}_{it}^{-1}} \left(R\sqrt{d \log \left(\frac{1 + t/\lambda}{\delta} \right)} + \sqrt{\lambda} \right) \tag{1}$$

Finally, in the proof of Lemma 11 in [2] it is noted that:

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

We now have all of the tools in hand to analyze the chaining algorithm. Let r_i^j be the number of times group i is active (and thus picked uniformly at random among) between pulls $j, j+1$. Conditioning on the event that all payoff means lie in their respective confidence intervals (an event with probability mass $1-\delta$ by construction) the total regret incurred by RidgeFair up to time T, is upper bounded by the sum of the widths of all the confidence intervals around the means for each group i, over all time steps that group i is active. Let w_i^j be the width of the confidence interval around group i after j pulls. Let $R_i(T)$ denote the sum of the widths of the confidence

interval around the payoff for group i, summed over all times $t \leq T$ where group i is active, and let $n_i(T)$ be the number of times group i is pulled up to time T. Note that the regret $R(T) \leq \sum_i R_i(T)$ with probability $1 - \delta$. Then

$$R_i(T) \leq \sum_{j=1}^{n_i^T} w_i^j \cdot r_i^j$$

and hence,

$$R(T) \leq \sum_{i=1}^{k} \sum_{j=1}^{n_i^T} w_i^j \cdot r_i^j.$$

Now when group i is active, it has at least a $1/k$ chance of being pulled uniformly at random. Thus with high probability, $r_i^j = \tilde{O}(k)$. More formally, $\mathbb{P}(r_i^j \geq h) = (1 - 1/k)^h \leq e^{-h/k}$. Letting $h = k \log(j^2 k/\delta)$ gives $\mathbb{P}(r_i^j \geq h) \leq \delta/kj^2$, thus we can assume over all time steps $t \leq T$ that $r_i^j \leq k \log(T^2 k/\delta)$. Under the constraint $\sum_i n_i^T = T$, $R(T)$ is maximized at $n_i^T = T/k$, which gives

$$\begin{aligned}
R(T) &\leq k^2 \log(T^2 k/\delta) \sum_{t=1}^{T/k} w_i^t \\
&\leq 2k^2 \log(T^2 k/\delta) \sum_{t=1}^{T/k} ||x_t||_{\bar{V}_{it}^{-1}} (R\sqrt{d \log\left(\frac{1 + t/\lambda}{\delta}\right)} + \sqrt{\lambda}) \\
&\leq 2k^2 \log(T^2 k/\delta) \sqrt{\sum_{t=1}^{T/k} ||x_t||^2_{\bar{V}_{it}^{-1}}} \\
&\cdot \sqrt{\sum_{t=1}^{T/k} R^2 d \log\left(\frac{1 + t/\lambda}{\delta}\right) + T\lambda/k}
\end{aligned}$$

where the last equality follows from Cauchy-Schwarz. Combining this bound with that from Equation 2 gives that with probability $1 - \delta$,

$$R(T) \leq 2k^2 \log(T^2 k/\delta) \sqrt{2d \log\left(1 + \frac{T}{d\lambda}\right)} \cdot \sqrt{\frac{T}{k}\left(R^2 d \log\left(\frac{1 + T/k\lambda}{\delta}\right) + \lambda\right)}$$

or $R(T) = \tilde{O}(d\sqrt{Tk^3})$ for $\lambda = 1$, as desired. □

4 Experimental Results

In this section we provide an empirical evaluation of IntervalChaining. Recalling that the confidence intervals for RidgeFair hold in greater generality, when we specialize to the case of normal noise they are wider than the confidence intervals of IntervalChaining, which are derived using normality explicitly. Since our experiments feature normal noise, it follows that IntervalChaining reliably outperforms RidgeFair empirically in spite of RidgeFair's superior theoretical regret bound. We therefore focus exclusively on IntervalChaining in simulations.

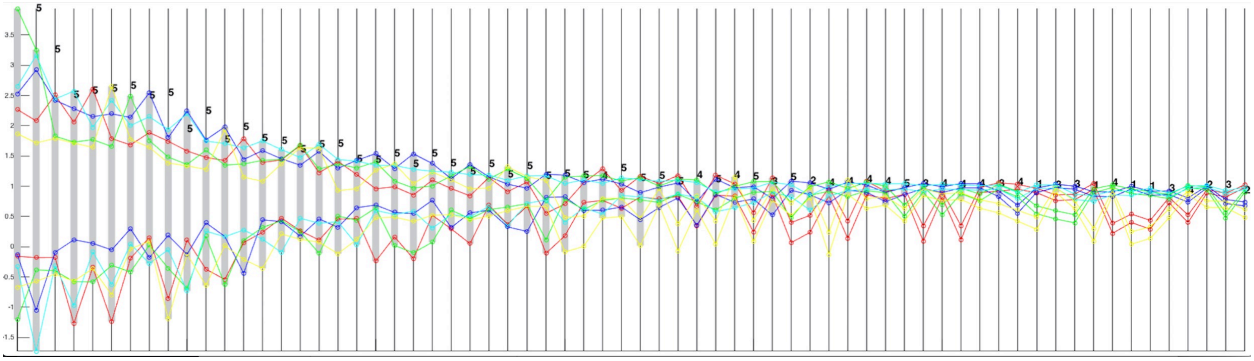


Figure 3: Chaining behavior (y axis) of IntervalChaining from one run of the algorithm over 10^6 rounds (x axis, on a log scale). See text for discussion.

Before turning to a more systematic set of experiments, we first give an illustration of the effects and variability of chaining over the course of a single run of IntervalChaining. In Figure 3, we visualize the chaining behavior of IntervalChaining over a run of $T = 10^6$ rounds on a log scale of the x axis, which measures t. There are $k = 5$ groups, each represented by a different color, and the upper and lower confidence bounds for each of the groups at each t are plotted using colored dots connected by lines. The number above each round's confidence intervals indicates the number of groups chained to the top group, while the gray bars show the union of the intervals chained to the top interval. In early rounds we see that we often have full chaining, with the gray bars encompassing all the colored dots; in later rounds the gray bars generally include only an upper subset of the colored intervals. Both the chaining number and the width of the grey bars shrink on average, but not montonically due to the variability of contexts in each round.

4.1 Empirical Cost of Fairness: Regret

We now present experimental evaluations of the regret of IntervalChaining compared to TopInterval[6] for various settings of d, k, and T. TopInterval uses the same OLS estimated qualities and confidence intervals as IntervalChaining, but chooses only the interval with highest upper confidence bound (rather than needing to play amongst all individuals chained to the top individual). We performed three kinds of experiments:

(a) Varying T: fixing $d = k = 2$, we measured the average regret of IntervalChaining as a function of increasing T.

(b) Varying k: fixing $d = 2$ and $T = 1000$, we measured the average regret of IntervalChaining as a function of increasing k.

(c) Varying d: fixing $k = 2$ and $T = 1000$, we measured the average regret of IntervalChaining as a function of increasing d.

The resulting plots are collected in Figure 4. Each plot contains four lines, one for each value of a parameter c (whose effects are discussed below) controlling the distribution of $\beta \sim U[0, c]^d$ for each group. In all cases the values presented were averaged over 1000 trials, with contexts drawn

[6] TopInterval is a variant of the standard linear bandits algorithm "LinUCB" [22], simplified to take advantages of the assumptions in our model.

uniformly at random from $[0,1]^d$ and standard Gaussian noise. In each case we plot the difference in regret between TopInterval and IntervalChaining. We note that our implementation of IntervalChaining does not include the random sampling component given in its formal presentation in the previous section, as we have found that this improves empirical performance over the theoretical regret guarantees presented above. However, we emphasize that this modification does not alter the fairness guarantee of IntervalChaining.

(a) **Varying T:** As T increases, the performance of IntervalChaining relative to TopInterval improves with time. This is consistent with our theory showing IntervalChaining has sublinear regret. As c varies, another trend also emerges: larger c increases both the range of possible values for each β and the regret incurred by playing randomly. This is reflected, for example, by the early spike in regret when $c = 10$. Since IntervalChaining explores randomly for more rounds than TopInterval, the prefix for which IntervalChaining continues to explore extends beyond the point where TopInterval begins to exploit. For $c = 10$, this results in a short period of high regret (the spike) before IntervalChaining begins exploiting. Conversely, once IntervalChaining begins exploiting, a larger c value corresponds to a higher signal-to-noise ratio for uncovering the βs (as standard Gaussian noise is relatively smaller compared to larger normed β). The βs also become more separated as c increases. For these reasons, $c = 10$ produces the lowest asymptotic regret and $c = 1$ produces the highest: when $c = 10$, distinguishing groups is easy, and chaining is not an issue for IntervalChaining; the opposite is true for $c = 1$.

(b) **Varying k:** As k increases but $T = 1000, d = 2$ are held constant, we observe a general linear increase in regret with k, improving upon the $O(k^{3/2})$ dependence expected from Theorem 1. We also see a general pattern in k's effect on regret fixing c. $c = 5$ is a particularly instructive case. For small k, slightly increasing k does not increase regret substantially: random additional βs are still well-separated and don't cause significant chaining (or regret) for IntervalChaining. For similar reasons, for small k, regret is low and nearly flat. However, as the number of groups increases, the βs populate the $[0, c]^d$ hypercube more densely, and IntervalChaining chains become more frequent and longer, leading to the apparent linear increase of regret in k for the middle range of k. Finally, for large values of k the cube is densely populated with groups and chaining forces IntervalChaining to play essentially at random, leading to the high regret plateau on the plot's right side, and a general S-shape for the curve as a whole. Varying c then changes the size and scale of the resulting S curve, but not its underlying shape: a smaller cube fills (and therefore chains) more quickly with increasing k, but the smaller cube also prevents the best group from being much better than a random group (constraining regret from growing too large). The plateaus of regret therefore begin at values of k which increase with c.

(c) **Varying d:** As d increases, we observe a general linear increase in regret, in accordance with the $O(d)$ dependence from Theorem 1. Once again, varying c interacts with increases in d in different ways. For small c such as $c = 1$, the lower signal-to-noise ratio makes estimation difficult; this is compounded by the increase in d, resulting in the high regret shown in the plot. In contrast, for large c such as $c = 10$ the effect of noise is relatively small, and the small number of groups makes distinguishing between groups relatively easy, even for large dimension.

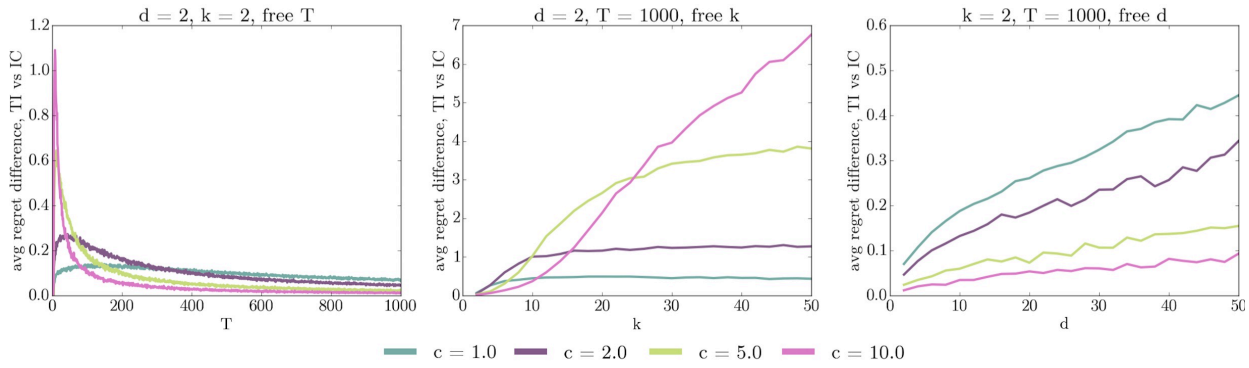


Figure 4: Plots of average regret with free variables T, k, and d. See text for discussion.

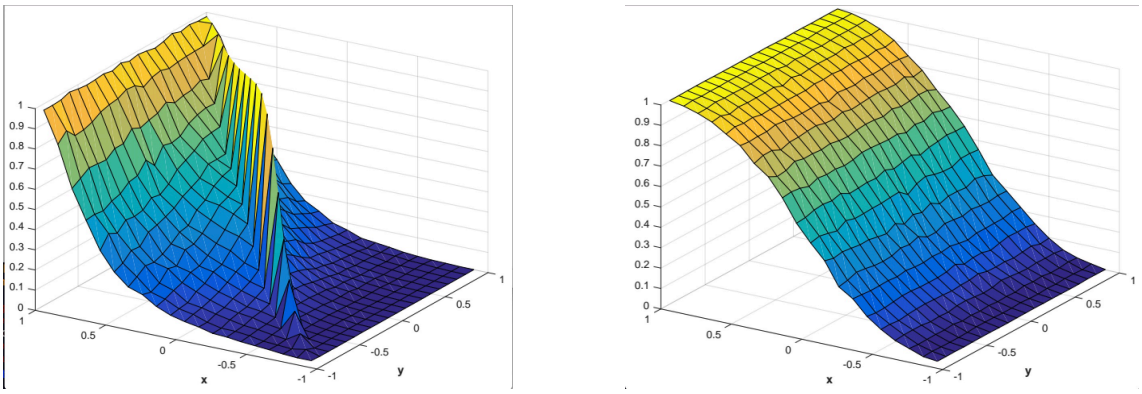
4.2 Empirical Unfairness of Standard Algorithms

Standard algorithms like TopInterval lack IntervalChaining's formal fairness guarantee, but it is not obvious that these algorithms actually exhibit discriminatory behaviour. Indeed, one might suspect that the harm that is caused by those rounds in which TopInterval selects a sub-optimal individual would be spread uniformly across all groups and all members of each group, since there is no intentional discrimination built into the algorithm. In this section, we demonstrate empirically that this is not the case: in particular, the costs of the incorrect decisions made by TopInterval can accrue disproportionately to certain groups and to structured subgroups within a given group, which we term *structural unfairness*.

Concepts of structural unfairness. To talk about structural unfairness, we introduce some notation. We will be interested in studying how the cost of those rounds at which the algorithm does not select the most qualified individual accrues to different subsets of individuals. Fix a group i, and consider two disjoint subgroups $P_1, P_2 \subset \mathcal{X}$ within that group (if the groups represent e.g. income brackets, the subgroups might represent racial background). For any round at which a sub-optimal decision is made, a given individual may have either *benefited* (if they were chosen despite not being the best individual), or have been *victimized* (if they were not chosen, despite being the best individual). Let T_i^v be the set of rounds in which an individual from P_i was *victimized*, and let T_i^b be the set rounds in which an individual from P_i *benefited*. We can now define the *discrimination index* with respect to P_i as $d_i = \mathbb{E}\left[\frac{|T_i^v|}{|T_i^v|+|T_i^b|}\right]$ – the fraction of times that individuals from P_1 were victimized, as a proportion of the times they were involved in sub-optimal decisions by the algorithm.

An algorithm $\mathcal{A}$ then exhibits *structural discrimination* with respect to P_2 relative to P_1 if $d_1 \ll d_2$. Intuitively, the discrimination index quantifies how the burden of suboptimal decisions is distributed: a subgroup with high discrimination index is *victimized* in a disproportionately high number of suboptimal decisions, while a subgroup with low discrimination index is *benefited* in a disproportionately high number of suboptimal decisions.

Empirical verification of structural unfairness. We now discuss an illustrative 2-dimensional instance on which TopInterval exhibits structural discrimination. Informally, there will be two groups. Group 1 contains two structured subgroups, while group 2 is homogeneous. In group 1's majority subgroup, the two features describing an individual are perfectly correlated. For individuals in group 1's minority subgroup, however, the two features are uncorrelated. Moreover, all individuals in group 1 have quality entirely determined by feature 1 – a fact only apparent from observations of individuals in the minority subgroup. The algorithm will therefore have increased



(a) TopInterval (b) IntervalChaining

Figure 5: Discrimination index for $T = 25$, averaged over 10^6 trials, for TopInterval and IntervalChaining. The sharp ridge along the line $y = x$ in panel (a) shows that TopInterval exhibits structural unfairness to P_1 relative to P_2, while the fact that the surface depends only on x in panel (b) demonstrates that IntervalChaining is not structurally unfair. See text for details and discussion.

uncertainty about the quality of individuals from the minority subgroup. As we will show, because TopInterval implicitly favors uncertainty, on this instance TopInterval exhibits structural discrimination in favor of the minority subgroup.

Formally, we consider an instance in which $d = k = 2$, and $\mathcal{X} = [-1, 1]^2$. The qualities of groups 1 and 2 are determined by the unknown coefficient vectors $\beta_1 = (1, 0)$, and $\beta_2 = (.5, .5)$ respectively. In group 1, we encode the majority subgroup by drawing 90% of individuals from the diagonal ($x = y$) and draw the remaining minority 10% of individuals uniformly off the diagonal ($x \neq y$). In group 2 contexts are drawn uniformly in each coordinate. Note that the quality of an individual from both the majority and minority subgroup of group 1 are identically distributed uniformly random in x, so there is no "reason" for discrimination.

We ran 10^6 simulations of TopInterval and IntervalChaining, for $T = 25$ rounds. We found that TopInterval exhibited strong structural discrimination in favor of the minority subgroup among group 1. Put simply, when a sub-optimal decision is made, individuals from group 1's majority subgroup are nearly 7 times more likely to be victimized than individuals from group 1's minority subgroup. See Figure 5 which plots the average discrimination index — the sharp ridge along $y = x$ represents a spike in discrimination index and corresponds to structural discrimination against group 1's majority subgroup. Finally, 59.6% of the sub-optimal decisions victimized group 1, and 40.4% of the sub-optimal decisions victimized group 2. Thus TopInterval is both structurally unfair in its unequal treatment of subgroups in group 1 and unfair in its unequal treatment of groups 1 and 2.

In Figure 5 we also show the discrimination index of IntervalChaining. A potential critique of our definition of fairness is that it binds between all individuals in a single round, but does not bind between individuals from the same group across time. As a result, there is no mathematical guarantee that algorithms that satisfy our fairness definition do not exhibit structural discrimination. For example, our fairness definition would in principle allow for the following kind of discrimination against a structured subgroup: when an individual arrives from the majority subgroup and is the most qualified, the algorithm gives them a loan deterministically, but when an individual arrives from the minority sub-group and is the most qualified, the algorithm chooses

uniformly at random. Note that this discriminatory strategy would not have a diminishing regret guarantee as our algorithms do – but nothing in our definition itself rules out this kind of behavior early in the learning process. However, figure 5 shows empirically that IntervalChaining does *not* exhibit structural discrimination on the same instance on which TopInterval does. This is an interesting empirical finding that is not explained by our theory, and one which we feel merits further investigation.

Acknowledgements

We thank Glen Weyl for pointing out the connection between our fairness definition and the ideas of John Rawls, and Adel Boyarsky for help researching the legal interpretations of disparate impact discrimination.

References

[1] Griggs v. Duke Power Co., 1971.

[2] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2312–2320. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4417-improved-algorithms-for-linear-stochastic-bandits.pdf`.

[3] Alekh Agarwal, Daniel J. Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1638–1646, 2014.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Propublica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[5] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016. Available at SSRN: http://ssrn.com/abstract=2477899.

[6] Anna Maria Barry-Jester, Ben Casselman, and Dana Goldstein. The new science of sentencing. *The Marshall Project*, August 8 2015. URL `https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing`. Retrieved 4/28/2016.

[7] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[8] Nanette Byrnes. Artificial intolerance. *MIT Technology Review*, March 28 2016. URL `https://www.technologyreview.com/s/600996/artificial-intolerance/`. Retrieved 4/28/2016.

[9] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[10] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 208–214, 2011.

[11] Cary Coglianese and David Lehr. Regulating by robot: Administrative decision-making in the machine-learning era. *Georgetown Law Journal*, 2016. Forthcoming.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[13] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268, 2015.

[14] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. *SIAM International Symposium on Data Mining*, 2016.

[15] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[16] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139*, 2016.

[17] FTC Commisioner Julie Brill. Navigating the "trackless ocean": Fairness in big data research and decision making. Keynote Address at the Columbia University Data Science Institute, April 2015.

[18] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 643–650. IEEE, 2011.

[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[20] Chung-Ming Kuan. Classical least squares theory. *Available on: http://homepage. ntu. edu. tw/˜ ckuan/pdf/et01/et_Ch3. pdf*, 2004.

[21] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[22] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[23] Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *AISTATS*, pages 485–492, 2010.

[24] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510. ACM, 2011.

[25] Clair C Miller. Can an algorithm hire better than a human? *The New York Times*, June 25 2015. URL `http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html`. Retrieved 4/28/2016.

[26] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

[27] John Rawls. *A Theory of Justice*. Harvard university press, 2009.

[28] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[29] Herbert Robbins. A sequential decision problem with a finite memory. *Proceedings of the National Academy of Sciences*, 42(12):920–923, 1956.

[30] Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine*, August 2013. URL `http://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-\patterns-of-crime/`. Retrieved 4/28/2016.

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

5 Supplement

5.1 Complete Description of Interval Chaining

procedure IntervalChaining(δ, T)
  for t from 1 to T **do**
    With probability $\frac{1}{t^{1/3}}$, play $\hat{i}_t \in_R \{1, \ldots, k\}$
    Else
    for i from 1 to k **do**
      Let $\hat{\beta}_{t,i} = (X_{t,i}^T X_{t,i})^{-1} X_{t,i}^T Y_{t,i}$ ▷ OLS est. of β_i
      Let $x_{t,i}$ be the feature vector for arm i at t
      Let $\hat{y}_{t,i} = \hat{\beta}_{t,i} \cdot x_{t,i}$ ▷ OLS est. of $\beta_i \cdot x_{t,i}$
      Let $\mathcal{F}_{t,i} = \mathcal{N}(0, \sigma^2 x_{t,i}(X_{t,i}^T X_{t,i})^{-1} x_{t,i}^T)$
      ▷ Width of $\frac{\delta}{2Tk}$confidence interval for $\hat{y}_{t,i}$
      Let $w_{t,i} = Q_{\mathcal{F}_{t,i}}(\frac{\delta}{2kT})$
      Let $[\ell_i^t, u_i^t] = [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$
    Let $i_*^t = \arg\max_i u_i^t$
    Let S_t be the set of actions chained to i_*^t
    Choose and play $\hat{i}^t \in_R S_t$
    Let $y_{t,\hat{i}^t}$ be the observed feedback

Figure 6: IntervalChaining, a fair algorithm for the linear contextual bandit problem. $X_{t,i}$ refers to the set of observed feature vectors $x_{t',i}$ for action i at times $t' < t$ for which the algorithm observed $y_{t',i}$; $Y_{t,i}$ to these observed $y_{t',i}$ values.
Let Q_F be the quantile function of a distribution F: $Q_F(x)$ represents the value at which the CDF of F equals x.

5.2 Proof of the Fairness of Interval Chaining

Proof of fairness in Theorem 1. Fix some $t \in \{1, \ldots, T\}], i \in \{1, \ldots, k\}$. Let π_i^t represent the probability the algorithm places on action i in round t in this particular run of the algorithm. If t is a round in which uniformly random play occurs, then $p_{t,i} = p_{t,i'}$ for all $i, i' \in [k]$, and so the algorithm satisfies the fairness condition in any such round.

Now, suppose t is a round in which the OLS estimates are used. By standard properties of OLS estimators (see, e.g. Kuan [20, Theorem 3.7]), $\hat{\beta}_{t,i} \sim \mathcal{N}(\beta_{t,i}, \sigma^2(X_{t,i}^T X_{t,i})^{-1})$. Then, for any fixed $x_{t,i}$, we have that

$$\hat{\beta}_{t,i} \cdot x_{t,i} \sim \mathcal{N}(\beta_{t,i} \cdot x_{t,i}, x_{t,i}^T \sigma^2 (X_{t,i}^T X_{t,i})^{-1} x_{t,i}) \tag{3}$$

by the properties of normally distributed random variables.

By the definition of the quantile function and the fact that the normal distribution is symmetric, with probability at least $1 - \frac{\delta}{Tk}$,

$$|\hat{\beta}_{t,i} \cdot x_{t,i} - \beta_{t,i} \cdot x_{t,i}| \leq w_{t,i}. \tag{4}$$

That is, $\beta_{t,i} \cdot x_{t,i} \notin [\ell_i^t, u_i^t]$ with probability at most $\frac{\delta}{kT}$. Thus, the probability that this fails to hold for any i at any time t is at most $kT \cdot \frac{\delta}{kT} = \delta$. Thus, we condition on this event for the

remainder of the argument and show that for all arms i, j and all rounds t, $\pi_i^t \geq \pi_j^t$ whenever $\beta_{t,i} \cdot x_{t,i} \geq \beta_{t,j} \cdot x_{t,j}$ with probability 1 when it holds.

For all fixed pair of actions i, j and a fixed round t, if $\beta_{t,i} \cdot x_{t,i} \geq \beta_{t,j} \cdot x_{t,j}$, then as $u_i^t \geq \beta_{t,i} \cdot x_{t,i}$ and $\ell_j^t \leq \beta_{t,j} \cdot x_{t,j}$, we have that

$$u_i^t \geq \beta_{t,i} \cdot x_{t,i} \geq \beta_{t,j} \cdot x_{t,j} \geq \ell_j^t$$

and so either (a) $\ell_i^t \geq u_j^t$, in which case the two intervals don't overlap, or (b) $\ell_i^t \leq u_j^t$. In either case, if j is chained to i_*^t, then so will be i. So, either both will be played with probability $\frac{1}{|S_t|}$, or only i will be played with that probability and j with probability 0, or both will be played with probability 0. In all such cases, $\pi_i^t \geq \pi_j^t$. □

Proof of Regret in Theorem 1. First, recall $L \geq \max_t \lambda_{max}\left(x_{t,i}^T x_{t,i}\right)$. Since $\max_t \lambda_{max}\left(x_{t,i}^T x_{t,i}\right) \leq \max_t ||x_{t,i}||^2$, and we assume $x_{t,i}$ bounded, we can always choose finite L. Also for each i let

$$\lambda_{min_{d,i}} = \mathbb{E}_{x_{t,i} \sim \mathcal{D}_i}\left[\lambda_{min}\left(\sum_{t \in [d]} x_{t,i}^T x_{t,i}\right)\right]$$

be the expected minimum eigenvector of a random design matrix made up of d observations from group i.

The entire regret of IntervalChaining can be broken into three components: the regret in exploration rounds, the regret of exploitation rounds before the estimators have enough samples to have concentrated, and the regret of exploitation rounds once the estimators are sufficiently accurate. We will bound the regret of IntervalChaining for each of these three phases, e.g., for any $T_1 \leq T$:

$$\begin{aligned}
&\text{Regret}(\textsc{IntervalChaining}, T) \\
&\quad = \sum_{t\text{:t is an exploit round}} \text{Regret}(t) \\
&\quad + \sum_{t\text{:t is an exploit round and } t < T_1} \text{Regret}(t) \\
&\quad + \sum_{t\text{:t is an exploit round and } t \geq T_1} \text{Regret}(t) \\
&\leq \sum_{t \leq T} \frac{1}{t^{1/3}} + T_1 + \sum_{t\text{:t is an exploit round and } t \geq T_1} \text{Regret}(t)
\end{aligned}$$

Let T_1 be defined as follows:

$$T_1 = \Theta\left(\min_i \left(\frac{dkL}{\lambda_{min_{d,i}}}\left(\ln^2 \frac{2}{\delta'} + \ln d\right)\right)^{3/2}\right). \tag{5}$$

Informally, T_1 is the number of rounds after which our estimators will be computed on sufficiently many samples such that the estimates are well-concentrated.

Let $p_t = \frac{1}{t^{1/3}}$ denote the probability that round t is an exploration round. Then, for any t, we have that

$$\sum_{t' < t} p_{t'} = \Theta(t^{2/3}) \tag{6}$$

and for any $t \geq T_1$, we have

$$\sum_{t'<t} p_{t'} = \Omega\left(\min_i\left(\frac{dk^2L}{\lambda_{min_{d,i}}}\left(\ln^2\frac{2}{\delta'} + \ln d\right)\right)\right) \tag{7}$$

Let δ denote the probability that some estimator we use falls outside its $1-\delta'$ probability bound; we will choose $\delta = O\left(\frac{1}{T^{1/3}}\right)$. Then, we have

$$\begin{aligned} R(T) &\leq \sum_{t=1}^{T} p_t \cdot 1 + T_1 + \sum_{t=T_1}^{T} regret(t, A) + \delta T \\ &\leq O\left(T^{2/3} + T_1 + \sum_{\substack{t>T_1, \\ t \text{ is an exploit round}}}^{T} regret(t, A)\right). \end{aligned} \tag{8}$$

We now upper-bound the regret in the exploitation rounds after round T_1, namely the terms above corresponding to $regret(t, A)$. We will upper-bound $w_{t,i}$, the width of the confidence intervals, which will satisfy $\beta_i \cdot x_{t,i} \in [\hat{y}_{t,i} - w_{t,i}, \hat{y}_{t,i} + w_{t,i}]$ for all groups i and rounds t, with probability $1-\delta$. We will then condition on the event that over all rounds and all arms this holds. Recall that $w_{t,i} = Q_{\mathcal{N}(0, x_{t,i}(X_{t,i}^T X_{t,i})^{-1} x_{t,i}^T)}(\frac{\delta}{2kT})$. Trivially, this confidence bound fails to hold for a fixed i and t with probability at most $\frac{\delta}{Tk}$: thus, for any times and any groups, some confidence intervals will fail with probability at most δ.

Now, we condition on the confidence intervals holding for all t, i. Let $\hat{i}_t$ be the optimal arm in round t, and recall that i_*^t is the arm round with highest upper confidence interval in round t and $\hat{i}^t$ the arm chosen by IntervalChaining. Since the confidence intervals are valid, it must be that $[\ell_{i_*^t}^t, u_{i_*^t}^t] \cap [\ell_{\hat{i}_t}^t, u_{\hat{i}_t}^t] \neq \emptyset$, so the instantaneous regret for any chained action is at most

$$regret(t) \leq 4\sum_{i \in S_t} w_{t,i} \leq 4k \max_{i \in S_t} w_{t,i} \tag{9}$$

To bound this term, we will focus on bounding $\max_{i \in S_t} w_{t,i} = Q_{\mathcal{N}(0, x_{t,i}(X_{t,i}^T X_{t,i})^{-1} x_{t,i})}\left(\frac{2kT}{\delta}\right)$. We first bound

$$\begin{aligned} & x_{t,i}(X_{t,i}^T X_{t,i})^{-1} x_{t,i} \\ & \leq ||x_{t,i}|| \lambda_{max}\left((X_{t,i}^T X_{t,i})^{-1}\right) \\ & = ||x_{t,i}|| \frac{1}{\lambda_{min}\left(X_{t,i}^T X_{t,i}\right)} \\ & \leq \frac{1}{\lambda_{min}\left(X_{t,i}^T X_{t,i}\right)} \end{aligned} \tag{10}$$

where the last inequality holds as $||x_{t,i}|| \leq 1$ for all t, i. We now bound $\lambda_{min}\left(X_{t,i}^T X_{t,i}\right)$. Let us use the notation

$$\mathbb{E}\left[\lambda_{min}\right] = \mathbb{E}\left[\lambda_{min}\left(X_{t,i}^t X_{t,i}\right)\right]$$

Let $G_{t,i}$ be the number of observations of action i with contexts drawn uniformly from the distribution for action i prior to round t, and let $L \geq \max_t \lambda_{max}\left(x_{t,i}^T x_{t,i}\right)$. Then for any $\alpha \in [0,1]$, by the superadditivity of minimum eigenvalues for PSD matrices and linearity of expectations, we get

$$\mathbb{E}\left[\lambda_{min}\right] \geq \frac{G_{t,i}}{d}\lambda_{min_{d,i}} \geq \lfloor \frac{G_{t,i}}{d} \rfloor \lambda_{min_{d,i}}.$$

This implies that

$$\begin{aligned}
&\mathbb{P}_{X_{t,i}}\left[\lambda_{min}\left(X_{t,i}^T X_{t,i}\right) \leq \alpha \lfloor \frac{G_{t,i}}{d} \rfloor \lambda_{min_{d,i}}\right] \\
&\leq \mathbb{P}_{X_{t,i}}\left[\lambda_{min}\left(X_{t,i}^T X_{t,i}\right) \leq \alpha \mathbb{E}\left[\lambda_{min}\right]\right] \\
&\leq \mathbb{P}_{X_{t,i}}\left[\lambda_{min}\left(X_{t,i}^t X_{t,i}\right) \leq \alpha \lambda_{min}\left(\mathbb{E}\left[X_{t,i}^t X_{t,i}\right]\right)\right] \\
&\leq d e^{-(1-\alpha)^2 \lambda_{min}\left(\mathbb{E}\left[X_{t,i}^t X_{t,i}\right]\right)/2L} \\
&\leq d e^{-(1-\alpha)^2 \mathbb{E}[\lambda_{min}]/2L} \\
&\leq d e^{-(1-\alpha)^2 \lfloor \frac{G_{t,i}}{d} \rfloor \lambda_{min_{d,i}}/2L}
\end{aligned} \tag{11}$$

where the second and fourth inequalities follow from Jensen's inequality (which implies that $\mathbb{E}\left[\lambda_{min}\right] \leq \lambda_{min}\left(\mathbb{E}\left[X_{t,i}^t X_{t,i}\right]\right)$) and the third inequlity follows from a matrix Chernoff bound (see e.g. Tropp et al. [31]). Then, taking logs and rearranging, with probability $1-\delta$,

$$\lambda_{min}\left(X_{t,i}^T X_{t,i}\right) \geq \alpha \lfloor \frac{G_{t,i}}{d} \rfloor \lambda_{min_{d,i}} \tag{12}$$

whenever

$$G_{t,i} \geq d\left(\frac{L}{(1-\alpha)^2 \lambda_{min_{d,i}}}\right)\left(\ln \frac{1}{\delta} + \ln d\right). \tag{13}$$

A standard multiplicative Chernoff bound implies, for any fixed t, with probability $1-\delta'$, the number of exploration rounds G_t prior to round t will satisfy

$$|G_t - \sum_{t'<t} p_{t'}| \leq \sqrt{\ln \frac{2}{\delta'} \sum_{t'<t} p_{t'}}. \tag{14}$$

Similarly, after G_t rounds of exploration, $G_{t,i}$, the number of exploration rounds in which a fixed action i was explored, with probability $1-\delta'$, a Chernoff bound implies

$$|G_{t,i} - \frac{G_t}{k}| \leq \sqrt{\ln \frac{2}{\delta'} \frac{G_t}{k}}. \tag{15}$$

Combining Equation 14 and 15, with probability at least $1-2\delta'$, for a fixed i and t, if $\sum_{t'<t} p_{t'} \geq 36k \ln^2 \frac{2}{\delta'}$ we have that

$$|G_{t,i} - \frac{\sum_{t'<t} p_{t'}}{k}| \leq \frac{\sum_{t'<t} p_{t'}}{2k} \tag{16}$$

Thus, with probability $1-\delta'$, Equation 13 and therefore Equation 12 hold for any t such that $\sum_{t'<t} p_{t'} \geq 36k \ln^2 \frac{2}{\delta'}$ and

$$\frac{\sum_{t'<t} p_t}{2k} \geq d\left(\frac{L}{(1-\alpha)^2 \lambda_{min_{d,i}}}\right)\left(\ln \frac{1}{\delta} + \ln d\right). \tag{17}$$

Both Equation 16 and 17 hold by our assumption that $t > T_1$ and Equation 7. In total, we now upper-bound the sum of the instantaneous regrets from the exploitation rounds in the remaining $T - T_1$ rounds.

$$
\begin{aligned}
&\sum_{\substack{t>T_1,\\ t \text{ is an exploit round}}} regret(t, A) \\
&\leq 4k \sum_{t>T_1} \max_i Q_{\mathcal{N}(0,\lambda_{max}\left((X_{t,i}^T X_{t,i})\right)^{-1})}\left(\frac{\delta}{2kT}\right) \\
&\leq 4k \sum_{t>T_1} Q_{\mathcal{N}(0,\frac{1}{\min_i \lambda_{min}\left(\left(X_{t,i}^T X_{t,i}\right)\right)})}\left(\frac{\delta}{2kT}\right) \\
&\leq 4k \sum_{t>T_1} Q_{\mathcal{N}(0,\frac{1}{\min_i \alpha\lfloor\frac{G_{t,i}}{d}\rfloor\lambda_{min_{d,i}}})}\left(\frac{\delta}{2kT}\right) + \delta T \\
&\leq 4k \sum_{t>T_1} \sqrt{\frac{\ln\frac{2kT}{\delta}}{\min_i \alpha\lfloor\frac{G_{t,i}}{d}\rfloor\lambda_{min_{d,i}}}} + 2\delta T \\
&= O\left(k \sum_{t>T_1} \sqrt{d \cdot \frac{\ln\frac{2kT}{\delta}}{\min_i G_{t,i}\lambda_{min_{d,i}}}} + 2\delta T\right) \\
&= O\left(k\sqrt{d \cdot \frac{\ln\frac{2kT}{\delta}}{\min_i \lambda_{min_{d,i}}}} \sum_{t>T_1} \sqrt{\frac{1}{\min_i G_{t,i}}} + 2\delta T\right) \\
&= O\left(k\sqrt{d \cdot \frac{\ln\frac{2kT}{\delta}}{\min_i \lambda_{min_{d,i}}}} \sum_{t>T_1} \sqrt{\frac{k}{\sum_{t'<t} p_{t'}}} + 2\delta T\right) \\
&= O\left(k\sqrt{d \cdot \frac{\ln\frac{2kT}{\delta}}{\min_i \lambda_{min_{d,i}}}} \sum_{t>T_1} \sqrt{\frac{k}{t^{2/3}}} + 2\delta T\right) \\
&= O\left(k^{3/2}\sqrt{d \cdot \frac{\ln\frac{2kT}{\delta}}{\min_i \lambda_{min_{d,i}}}} \sum_{t\in[T_1,T]} \frac{1}{t^{1/3}} + 2\delta T\right) \\
&= O\left(k^{3/2}\sqrt{d \cdot \frac{\ln\frac{2kT}{\delta}}{\min_i \lambda_{min_{d,i}}}} T^{2/3} + \delta T\right)
\end{aligned}
\tag{18}
$$

where the third inequality follows from Equation 12, the fourth from a Chernoff bound for sub-gaussian random variables, the fifth and sixth from basic algebra and choosing $\alpha = \frac{1}{2}$, the seventh from Equation 16, the eighth from Equation 6, and the ninth and tenth from basic algebra.

The final regret bound follows from Equation 8, substituting in the value for T_1 in Equation 5 and the upper bound on the exploitation rounds in times after T_1 steps given by Equation 18, using $\delta' = \min\left(\frac{1}{3kT}, \frac{1}{T^{1/3}}\right)$:

$$R(T) = T^{2/3} + O\left(k^{3/2}\sqrt{d \cdot \frac{\ln \frac{2kT}{\delta}}{\min_i \lambda_{min_{d,i}}}}T^{2/3} + T^{2/3}\right) +$$
$$\Theta\left(\min_i \left(\frac{dkL}{\lambda_{min_{d,i}}}\left(\ln^2 \frac{2TK}{\delta} + \ln d\right)\right)^{3/2}\right)$$

as desired. □