Fairness Incentives for Myopic Agents

Sampath Kannan[*] Michael Kearns[†] Jamie Morgenstern [‡] Mallesh Pai [§]
Aaron Roth [¶] Rakesh Vohra [‖] Zhiwei Steven Wu [**]

May 4, 2017

Abstract

We consider settings in which we wish to incentivize myopic agents (such as Airbnb landlords, who may emphasize short-term profits and property safety) to treat arriving clients *fairly*, in order to prevent overall discrimination against individuals or groups. We model such settings in both classical and contextual bandit models in which the myopic agents maximize rewards according to current empirical averages, but are also amenable to exogenous *payments* that may cause them to alter their choices. Our notion of fairness asks that more qualified individuals are never (probabilistically) preferred over less qualified ones [Joseph et al., 2016b].

We investigate whether it is possible to design inexpensive subsidy or payment schemes for a principal to motivate myopic agents to play fairly in all or almost all rounds. When the principal has full information about the state of the myopic agents, we show it is possible to induce fair play on every round with a subsidy scheme of total cost $o(T)$ (for the classic setting with k arms, $\tilde{O}(\sqrt{k^3T})$, and for the d-dimensional linear contextual setting $\tilde{O}(d\sqrt{k^3T})$). If the principal has much more limited information (as might often be the case for an external regulator or watchdog), and only observes the number of rounds in which members from each of the k groups were selected, but not the empirical estimates maintained by the myopic agent, the design of such a scheme becomes more complex. We show both positive and negative results in the classic and linear bandit settings by upper and lower bounding the cost of fair subsidy schemes.

[*]University of Pennsylvania. Email: kannan@cis.upenn.edu
[†]University of Pennsylvania. Email: mkearns@cis.upenn.edu
[‡]University of Pennsylvania. Email: jamiemor@cis.upenn.edu
[§]Rice University. Email: mallesh.pai@rice.edu
[¶]University of Pennsylvania. Email: aaroth@cis.upenn.edu
[‖]University of Pennsylvania. Email: rvohra@seas.upenn.edu
[**]University of Pennsylvania. Email: wuzhiwei@cis.upenn.edu

1 Introduction

Recent uses of machine learning to make decisions of consequence for individual citizens (such as credit, employment and criminal sentencing) have led to concerns about the potential for these techniques to be discriminatory or unfair (Barocas and Selbst [2016], Coglianese and Lehr [2016], Brill [2015]). Existing research has emphasized discriminatory outcomes originating from biases encoded into the data sets on which algorithms are trained.[1] In this paper, we consider a different source of unfairness in a stochastic bandit setting. The key friction we examine is when a forward-looking principal concerned with fairness (such as a regulator or technology platform) is not the one directly making the choices. Instead, a sequence of myopic agents are making the choices. To prevent unfair choices by these agents, the principal may offer targeted monetary rewards to agents to incentivize them to make different choices than they would have in the absence of such payments. Our concern in this paper is how much the principal needs to be prepared to pay in order to incentivize fair decisions by myopic agents.

To help fix ideas and motivate our problem, consider a challenge faced by peer-to-peer (P2P) platforms such as Prosper (P2P lending) or Airbnb (P2P short-stay housing). The platform cannot dictate to their users who to extend loans or rent to. Nevertheless, it may wish to ensure the choices its users make are fair, either to avoid criticism,[2] or to comply with existing regulations. P2P lending, for example, is subject to the Equal Credit Opportunity Act (ECOA). One provision of the Act is a requirement that lenders furnish reasons for adverse lending decisions. This obligation falls on the shoulders of the platform, and is challenging to discharge because the platform aggregates the decisions of many different lenders.[3]

In our model, an agent arrives at each period and must choose amongst a set of available alternatives. (For instance, the agent might be a lender on Prosper choosing to whom they will grant a loan, or an Airbnb host choosing which guest to accept.) We model this as a choice of which arm to pull in a stochastic bandit setting. We consider both the classic and contextual bandit cases. In the classic setting, each arm represents an individual, who will over time repeatedly be considered for service. In the contextual setting, each arm represents a group, and individual members of that group are represented by contexts (i.e. sets of individual features) which change at each round. A stochastic reward from the pull of an arm models the uncertain payoff associated with serving an individual (i.e. extending a loan, or having the individual rent). Each agent is myopic in the sense that they are occasional users of this platform, and thus care only about their current expected payoff. Because of myopia, the agent chooses the arm with the highest empirical mean in the classical case, and the context with the highest predicted reward according to a fixed one-shot learning procedure known to the principal (e.g. ordinary least squares regression, or ridge regression). Each agent is limited to pulling a single arm to model their limited resources (e.g., they may only have the funds to grant one loan, or host one guest on any particular night), and of course this simplifies our analysis.

The platform, motivated by a need for "accountability," would like to be fair. Our formal

[1] For example, Boston's Street Bump program, which uses smartphones to determine where road repairs are needed, results in certain areas being underserved because of the sparsity of smartphones traversing them (O'Leary [2013]).

[2] Airbnb has very recently received scrutiny over both anecdotal reports and systematic studies of racist behavior by landlords on their platform (Edelman et al. [2017]). This study also suggests that myopia may play a role in discrimination, in the sense that landlords with no prior exposure to minority renters were more likely to discriminate.

[3] Lenders on Prosper must agree to comply with the relevant provisions of the ECOA, but there is still evidence of discrimination; see Pope and Snyder [2011].

	Full Information	Partial Information
Classical	$\tilde{O}(\sqrt{k^3T})$ cost	$O(\sqrt{T})$ cost for $k = 2$ $\Omega(T)$ cost for $k \geq 3$ $\tilde{O}(\sqrt{k^3T})$ cost allowing $\tilde{O}(k^2)$ unfair rounds
Linear Contextual	$\tilde{O}(d\sqrt{k^3T})$ cost	$o(T)$ unfair rounds $\rightarrow \Omega(T)$ cost $o(T)$ cost $\rightarrow \Omega(T)$ unfair rounds

Table 1: Summary of cost to incentivize perfect fairness and fairness in all but a limited number of rounds.

definition of fairness (Joseph et al. [2016b]) may be found in Section 2, and can be informally described as follows: Suppose an auditor knew the expected reward of each arm (or more generally, in the contextual case, of each context), and looked back at the platform's decisions. Fairness requires that a worse individual was never favored over a better individual. More precisely, if a platform is fair, then on any day t, the probability p_x that the agent pulls arm x is such that if the expected reward of x is at least that of x', then $p_x \geq p_{x'}$. Fairness as defined in this paper does not address inequities "outside the model." For example, if one group has lower expected payoff for every context than another, perhaps due to historical inequities, and there are no additional features available to the learning algorithm, our notion of fairness permits the agent to favor the higher expected group. In this sense our fairness notion is aligned with (apparent) meritocracy.

Intuitively, there are two impediments to fairness in this model. First, neither the platform nor the agents know the distribution of rewards of each arm. If these were known, the problem would be trivial: it would be both fair and agent-optimal to always pull the arm with highest expected reward. Second, the agents are myopic — they have no incentive to directly invest in fairness or learning.

We examine whether it is possible for the platform, hereafter called the *principal*, to incentivize the agents to make fair choices by offering the agent payments for selecting particular arms. These payments can be randomized. Because the agents behave identically in our model (they are all myopic), we treat them as if they are a single myopic entity, whom we term the *agent*. We investigate how much information a principal needs to incentivize fair behavior. Characterizing the information requirements is important, since sometimes the principal may be an external regulator or other entity tasked with oversight without the full information available to the platform. At one extreme, called partial information, we suppose the principal observes only which decisions were made by the agent in each of the previous rounds but not the rewards. In the P2P context, rewards might be unobservable because the reward an agent experiences is a function of both observed characteristics of the borrower or renter and a private type of the agent. At the other extreme of full information, the principal has the same information as the agent.

We ask each of these questions in both the classic and linear contextual bandits settings. In the classic case, individual i is better than individual j if the mean of distribution i is higher than that of distribution j. In the contextual case, individual i on day t is represented by a set of features, or a context x_i^t, and that individual's expected reward is defined as $f_i(x_i^t)$ for some $f_i \in C$.

1.1 Our Results

Table 1 summarizes our results for characterizing the cost of incentivizing fair play in myopic agents. Informally, we show a stark separation: in the partial information model, any fair payment scheme must cost $\Omega(T)$; while in the full information model there are payment schemes which cost $\tilde{O}(d\sqrt{k^3T})$. In both the classic and contextual settings, the full information upper bounds effectively incentivize the agent to play known fair algorithms. The lower bounds for the partial information model look somewhat different. For the classic setting, our lower bound simply shows that the first round at which the principal doesn't offer a payment of 1 will be unfair with constant probability; for the contextual setting, *every* round must either be unfair or offer $\Omega(1)$ payment.

We additionally show that in partial information model, the classic problem is somewhat easier in two cases. When there are only 2 arms, we give a simple payment scheme that only incurs a cost of $O(\sqrt{T})$ and guarantees with high probability that the agent is fair at every round. Even when $k \geq 3$, a principal who is allowed $\tilde{O}(k^2)$ unfair rounds can design a payment scheme which costs only $\tilde{O}(\sqrt{k^3T})$.[4] In the full information model, we exhibit a payment scheme for the principal which, with high probability, is fair in *every* round, and has cost $\tilde{O}(\sqrt{k^3T})$.

It is interesting to observe that all our payment schemes that guarantee fairness (either in each period, or except at a constant number of periods) and achieve sublinear costs also induce the agent to play so as to experience sublinear regret. If we think of sub-linear regret as a proxy for efficiency, our full information results say we can achieve *both* efficiency and fairness with subsides that grow slowly over time. Under certain conditions, this is also possible in the partial information setting, which does not require the principal to "open the books" of the agent.

1.2 Related Work

Our work is closely related to the literature on incentivizing experimentation in bandit settings. In these papers, a sequence of agents arrives one at a time and each is allowed to select an arm to pull. Each agent "lives" for one period and therefore pulls the arm that has the highest current estimated payoff given the history. The agents' myopia means that they do not explore sufficiently and the patient principal must encourage it.

In this context, Frazier et al. [2014] explore the achievable set between the monetary rewards the principal must pay to incentivize exploration and the time discounted expected reward to the principal. In this paper the history of actions and outcomes is observed by the principal and there is, therefore, no examination of what happens with limited information. More importantly, there is no consideration of fairness, which is our primary interest.

A different string of papers does not explicitly allow for monetary payments, but instead the principal discloses information about past agents' realized rewards to the future agents, see e.g. Kremer et al. [2014], a more general exploration in Mansour et al. [2015, 2016], or an analytical solution in a continous time Poisson bandit setting by Che and Horner [2015] and the same in a discrete time setting in Papanastasiou et al. [2017]. The key point of departure for our work is that for us, the principal is not explicitly interested in the long term reward of the agent, but is instead interested in promoting "fairness" (although this will incidentally have the property of increasing long-term reward of the agent by encouraging experimentation).

[4] In other words, we show that by allowing a constant number of unfair rounds, independent of T, one can achieve sublinear costs.

Joseph et al. [2016b] originally proposed our definition of contextual fairness. They consider the tradeoffs between requiring this form of fairness and achieving no-regret in both classic and contextual bandit settings. This notion was also employed in Joseph et al. [2016a], when specialized to the linear contextual bandits case.

2 Preliminaries

A principal faces a sequence of homogeneous myopic agents, each operating in a *contextual bandit* setting. For simplicity, we view the agents as a single agent repeatedly making myopic choices.

2.1 Contextual Bandits

Let $\{k\}$ refer to a set of arms. In each round $t \in [T]$, an adversary reveals to the agent a *context* $x_j^t \in \mathcal{X}$, where $\mathcal{X}$ is the common domain of contexts, for each arm $j \in [k]$.

Fix some class C of functions of the form $f : \mathcal{X} \to [0,1]$. Associated with each arm j is a function $f_j \in C$, unknown to both the agent and the principal.[5]. An agent who chooses an arm j in period t with context is x_j^t receives a stochastic reward $r_j^t \sim \mathcal{F}_j^{x_j^t}$, where $\mathbb{E}[r_j^t] = f_j(x_j^t)$, for some distribution $\mathcal{F}_j^{x_j^t}$ over $[0,1]$.

Remark 2.1 (Linear Contextual Bandits)**.** *The results in this paper which involve contexts are for the case where the set of contexts* $\mathcal{X} = \{x \in [0,1]^d : \|x\| \leq 1\}$ *for some number* $d > 0$*, and*

$$C = \left\{ f_j : \text{ there exists some } \theta_j \in [0,1]^d, \|\theta_j\| \leq 1, \text{ s.t. } f_j(x_j) = \langle \theta_j, x_j \rangle, \ \forall x_j \in \mathcal{X} \right\}.$$

Remark 2.2 (Classic Bandits)**.** *The* classic *bandits problem is a special case of the contextual bandits problem where the set of possible contexts is a singleton. Then* $\mathcal{F}_j^{\cdot} = \mathcal{F}_j$ *and* $r_j^t \sim \mathcal{F}_j$*.*

In the running example of P2P lending, $\mathcal{X}$ represents possible profiles of attributes that a lender can observe about a potential borrower. The exact relationship between a profile of attributes at time t, denoted x_j^t, and the expected reward earned from extending a loan to a borrower with this profile is $f_j(x_j^t)$; this functional form is unknown to the lender, the platform, and the agent.

2.2 The Myopic Agent and the Principal

In each period t, the principal can offer a vector of payments $p^t \in \mathbb{R}_+^k$ to the agent. Here p_i^t is to be interpreted as the monetary incentive the agent receives from the principal if they selects arm i on top of any reward that would accrue from the arm itself.

We assume the agent makes *myopic* choices facing empirical estimates of the reward from each arm (we describe how these empirical estimates are constructed momentarily). We use $\hat{\mu}_i^t$ to denote the empirical estimated reward for selecting arm i in round t. When the agent receives a proposed subsidy vector $p^t \sim \gamma^t(\cdot)$, they choose the arm which maximizes the sum of empirical expected reward and payment, i.e. chooses $i^t \in \arg\max_i(\hat{\mu}_i^t + p_i^t)$. This is the sense in which the

[5]Often, the contextual bandit problem is defined so that there is a single function f associated with all of the arms. Our model is only more general.

agent is myopic—they maximize (their estimate of) today's net reward plus payment, with no concern for the future. Note that we assume that rewards are expressed in monetary terms, i.e. that they are directly comparable to offered payments.

For concreteness and without loss of generality, we fix a tie-breaking rule—if $M = \arg\max_i(\hat{\mu}_i^t + p_i^t)$ contains multiple elements, the agent chooses uniformly at random amongst the members of M that also maximize payment, i.e. from the set $\arg\max_{i \in M} p_i^t$.[6] The principal experiences cost $p_{i^t}^t$ at round t, and total cost $\sum_{t=1}^{T} p_{i^t}^t$ over the course of T rounds. The payments offered by the principal, and the empirical estimates of the agent, depend on what they know about past choices and outcomes. We define these next.

2.3 Information and Histories

The agent will, in any period t, recall history $h^t \in \left(\mathcal{X}^k \times \mathbb{R}_+^k \times [k] \times [0,1]\right)^{t-1} = \mathcal{H}^t$. This is a record of the previous $t-1$ rounds experienced by the agent: $t-1$ 4-tuples encoding the realization of the contexts for each arm in a given period, the payments the principal offered, the arm chosen, and the realized reward observed.

In the linear contextual case, let $\hat{\theta}_i^t$ represent an estimate of the linear model θ_i based on the history h^t (this could, for example, be the ordinary least-squares or regularized ridge regression estimator — the important thing is that whichever method is used is known to the principal). The myopic decision maker, at day t, when facing contexts $x_1^t, \ldots, x_k^t$ will have empirical estimated reward $\hat{\mu}_i^t = \langle \hat{\theta}_i^t, x_i^t \rangle$. In the classic setting,

$$\hat{\mu}_i^t = \frac{\sum_{t'<t} r_i^{t'}}{|\{t' < t : \ i \text{ played in round } t'\}|}$$

that is, it represents the empirical average for the set of previous rewards observed from arm i in previous rounds.

Note that during the first several rounds, the myopic reward estimates $\hat{\mu}_i^t$ are not necessarily defined, e.g. if in the classic setting, the agent has not yet observed any rewards from arm i, or if in the linear contextual case, the agent has not observed sufficiently many reward/context pairs to uniquely define the OLS estimator. To get around this issue, we assume that the agent has previously observed sufficiently many observations from each arm to make these estimates well defined — i.e. at least one observation per arm in the classic case, and observations corresponding to contexts that combined form a full rank matrix in the linear case.

We consider two information models for the principal. In the *full information* model, the principal observes everything the agent observes, i.e. there is no information asymmetry between the two. In the *partial information* model, the principal observes neither the contexts faced by the agents, nor the realized reward of the arm the agent pulled. We will denote this by $\underline{h}^t \in \left(\mathbb{R}_+^k \times [k]\right)^{t-1} = \underline{\mathcal{H}}^t$. The principal's information scheme is a function of the information they have.

In what follows, we define various notions of performance of algorithm. This is without loss: the payments that the principal offers, and the resulting choices made by the agents choices, taken together, can be thought of as an algorithm making choices in a stochastic bandit setting.

[6] Our results do not depend in any important way on the particulars of the tie-breaking rule—we chose this one to simplify parts of the lower bound proof.

2.4 Fairness and Regret

A standard method for measuring the performance of a bandit algorithm is to measure its *regret*. If one knew $\{f_j\}_{j\in[k]}$, selecting the arm with highest expected reward in each period would be optimal. Fix an algorithm $\mathcal{A}$ and let π^t be the distribution over arms at round t of the algorithm: the **regret** of $\mathcal{A}$ is the difference between the reward of the optimal policy, and the reward of the agent:

$$\text{Regret}(x^1,\dots,x^T) = \sum_t \max_j \left(f_j(x_j^t)\right) - \mathbb{E}_{i^t\sim\pi^t}\left[\sum_t f_{i^t}(x_{i^t}^t)\right].$$

We say that $\mathcal{A}$ satisfies regret bound $R(T)$, if $\max_{x^1,\dots,x^T} \text{Regret}(x^1,\dots,x^T) \le R(T)$.

We denote by $\pi^t_{j|h^t}$ the probability that $\mathcal{A}$ chooses arm j after observing contexts x^t in period t, given h^t. For economy, we will often drop the superscript t on the history when referring to the distribution over arms: $\pi^t_{j|h} \coloneqq \pi^t_{j|h^t}$.

We now define what it means for an algorithm $\mathcal{A}$ to be fair in a particular round t. Informally, this will mean that $\mathcal{A}$ will play arm i with higher probability than arm j in round t only if i has higher *true* expected reward than j in round t.

Definition 2.3 (Round Fairness)**.** *Fix some history h^t. Recall $\pi^t_{j|h^t}$ is the probability that $\mathcal{A}$ plays arm j in round t given the history h^t. We will say $\mathcal{A}$ is fair in round t if, for any context x^t, for all pairs of arms $j, j' \in [k]$,*

$$\pi^t_{j|h} > \pi^t_{j'|h} \text{ only if } f_j(x_j^t) > f_{j'}(x_{j'}^t).$$

Similarly, a payment scheme is fair in round t if the selection by the myopic agent under the payment distribution is fair.

Remark 2.4. *When this definition is specialized to the classic (noncontextual) case, the reward distributions do not vary with time, i.e. $\mathcal{F}_j^t = \mathcal{F}_j$ for all t. Thus, "noncontextual" fairness reduces to guaranteeing that if arm i is played with higher probability than arm j, it must be that the average reward drawn from distribution $\mathcal{F}_i$ is higher than the average reward drawn from distribution $\mathcal{F}_j$.*

Remark 2.5. *To be clear about this definition in the partial information model, and what we mean by probabilities: note that the "algorithm" has access to h^t at the beginning of time t. By this we mean that, the principal has access to $\underline{h}^t$. The principal then offers payments to the agent, possibly randomizing. The agent sees the full history h^t, and the realized payments drawn from a distribution, and makes a choice. The principal's randomization, and then, if there are ties, the agent's randomization in period t, can be amalgamated into a net probability of each arm being selected in period t after history h^t. These are the π's that the definition refers to.*

We now introduce a notion of fairness which holds at every round with high probability over the history of observed rewards.

Definition 2.6 (Contextual Fairness)**.** *$\mathcal{A}(\cdot)$ is* **fair** *if, for any input $\delta \in (0,1)$, for all sequences of contexts, $x^1,\dots,x^t$ and all reward distributions $\mathcal{F}_1^t,\dots,\mathcal{F}_k^t$, with probability at least $1-\delta$ over the realization of the history h, for all rounds $t \in [T]$, $\mathcal{A}(\delta)$ is fair in round t.*

Contextual fairness, introduced in Joseph et al. [2016b], formalizes the idea that highly qualified individuals should be treated at least as well as less qualified individuals. Here, an individual's qualification is measured in terms of their expected reward for $\mathcal{A}$. If two individuals

have different profiles (or contexts) but generate the same expected reward to the learner, this definition enforces that both be played with *equal* probability every round. We also introduce a relaxation of contextual fairness, which allows for an algorithm to have some number of unfair rounds.

Definition 2.7 (g-Contextual Fairness)**.** *$\mathcal{A}(\delta)$ is g-***fair** *if, for any input $\delta \in (0,1)$, for all sequences of contexts, and all reward distributions, with probability at least $1-\delta$ over the realization of the history h, for all but g rounds $t \in [T]$, $\mathcal{A}(\delta)$ is fair in round t.*

Our principal is willing to incentivize the agent's behavior to ensure contextual fairness (and, incidentally, low regret). We investigate what subsidy schemes incentivize fair choices by a myopic agent, and the cumulative cost of such subsidies. We show this answer depends upon the kind of information the principal has access to: incentivizing fair play with partial information is in general very expensive, while incentivizing fair play under full information need not be so.

3 A Principal with Partial Information Cannot Ensure T Fair Rounds

In this section, we give a lower bound on the total payments needed in the partial information setting to ensure contextual fairness in every round. In fact, we don't even need to move to the contextual case: this section focuses on the classic bandit setting (where the context x_j^t is invariant with respect to t for each arm j). We show that in the partial information setting, any principal who incentivizes a myopic agent to satisfy contextual fairness in each round must incentivize uniformly random play in each of the T rounds, which has cumulative cost $\Omega(T)$.

Theorem 3.1. *Suppose $k \geq 3$. There is an instance such that any fair payment scheme in the partial information model must, with probability $1-\delta$, (where δ is the fairness parameter passed to the principal) spend $\Omega(T)$ in payments over T rounds and incur regret $\Omega(T)$.*

The lower bound proceeds from the following idea: at the first round, the principal has no information about what the instance is. Hence, in order to guarantee fairness against all instances, they must proceed cautiously and use a payment scheme that is able to induce uniformly random play (the only distribution that is fair for all instances) for every possible realization of empirical means. Because empirical means can range between 0 and 1, this will cost 1. However, because this payment distribution (by design) induces identical behavior on every possible instance, it does not allow the principal to learn anything about the instance. Thus, in every round before which fair play has been guaranteed, the principal has the same informational disadvantage. By induction, therefore, they must induce uniformly random play at every round, at a cost of 1 per round.

We show that fairness at every round, against all instances is equivalent to the payment scheme in each round being what we term *peaked*. A peaked payment rule is one that can always incentivize the play of some arm regardless of the empirical means the myopic agent currently has. This is equivalent to saying that there is some arm $i \in [k]$ for which $p_i \geq p_{i'} + 1$ for all $i' \neq i$. This will imply the payment scheme must spend $\Omega(1)$ in each round to incentivize fair play, or $\Omega(T)$ in total.

Definition 3.2 (Peaked)**.** *Let $p \in \mathbb{R}^k$. If for some $i \in [k]$, $p_i \geq \max_{i' \neq i} p_{i'} + 1$, we say p is* peaked. *If*

$$\mathbb{P}_{p \sim \mathcal{D}}[p \text{ is peaked}] = 1$$

then we call distribution $\mathcal{D}$ peaked.

Observation 1. *If a principal uses a peaked distribution $\mathcal{D}$ in a round, they learn nothing about the instance the agent faces from the agent's play in that round.*

Proof. By Definition 3.2, every payment scheme drawn from $\mathcal{D}$ is peaked. In other words, for every $p_\ell \sim \mathcal{D}$: there is some i_ℓ such that $p_{i_\ell} \geq \max_{i' \neq i_\ell} p_{i'} + 1$. Thus, the myopic agent will choose i_ℓ when presented with p_ℓ regardless of the instance the agent faces. □

The main idea behind Theorem 3.1 is in proving that any fair payment scheme must be peaked in every round. Technically, we use the fact that the principal learns nothing about the instance from a peaked distribution to allow us freedom to design a lower bound instance as a function of the first distribution $\mathcal{D}^t$ deployed by the principal that is not peaked. Because this distribution, by virtue of being the first non-peaked distribution, cannot be a function of the underlying instance, we are unconstrained in our ability to tailor the instance as a function of $\mathcal{D}^t$. We then show this instance forces an unfair round for $\mathcal{D}^t$; we can conclude that with probability $1 - \delta$, the principal must *never* deploy any distribution over payments that is not peaked.

Lemma 3.3. *For any fairness parameter δ, a fair payment scheme must with probability $1 - \delta$ generate a sequence of payment distributions $\mathcal{D}^1, \ldots, \mathcal{D}^T$ such that each D^t is peaked.*

We now conclude the proof of Theorem 3.1 before presenting the proof of Lemma 3.3.

Proof. Lemma 3.3 implies that a fair payment scheme must with probability $1 - \delta$ generate T peaked payment distributions. Since $\max_i p_i \geq \max_{j \neq i} p_j + 1$, and $\hat{\mu}_i^t \in [0, 1]$ for all i, the payment scheme's largest payment is always at least 1, and is always accepted. Thus, the myopic agent will receive a payment of at least 1 in every round, for a total cost of $\Omega(T)$.

To prove the regret of this payment scheme may be $\Omega(T)$ on some instances, consider each of the k instances in which one arm has mean 1 and the remaining $k - 1$ arms have mean 0. By Observation 1, the principal has no information about which of these instances is realized. Therefore to be to be fair with respect to all of these instances, each arm must be assigned the largest payment with equal probability, which induces uniformly random play amongst all k arms, $\Omega(1)$ regret per round, and cumulative regret $\Omega(T)$. □

We now present the proof of the main lemma for this section: that in order for a payment distribution to be fair, it must be peaked. Informally, we first show that any fair payment distribution must be "invariant under permutation": any coordinate i should have have an equal probability of having the largest payment, and have an equal probability of j being the second-largest payment with margin c, for each value of j and c. We then show in the first round t at which the payment distribution is unpeaked, $\mathcal{D}^t$ is unfair for some instance I constructed as a function of $\mathcal{D}^t$.

Proof of Lemma 3.3. We consider some round t. Suppose that for every round $t' < t$, the payment distribution $\mathcal{D}^{t'}$ was peaked. If the payment distribution $\mathcal{D}^t = \mathcal{D}$ at round t is fair, we show that it too must be peaked. Observe that by Observation 1, $\mathcal{D}$ must be defined independently of the underlying instance I, and because fairness is defined in the worst case over instances, we continue to have complete freedom in choosing I.

We first claim that in round t, if $\mathcal{D}$ is fair, for any two distinct $i, i', \in [k]$ and any $c \in [0, 1]$, that

$$\mathbb{P}_{p \sim \mathcal{D}}\left[p_i \geq \max_{\ell \in [k], \ell \neq i} p_\ell + c\right] = \mathbb{P}_{p \sim \mathcal{D}}\left[p'_i \geq \max_{\ell \in [k], \ell \neq i'} p_\ell + c\right]. \tag{1}$$

Suppose Equation 1 does not hold. We construct an instance for which $\mathcal{D}$ will not be fair, a contradiction. Suppose the left-hand side is larger than the right-hand side. Consider an instance where $\mu_i = \mu_{i'} = 1 - c$, and all other arms (of which there is at least 1) have means $\mu_j = 1$. Suppose further that the distribution over i's reward is deterministic point mass at $1-c$, whereas i''s reward distribution yields reward $1 - c + \epsilon$ with probability $\frac{1}{2}$ and $1 - c - \epsilon$ with probability $\frac{1}{2}$. Then, with probability at least $\frac{1}{4}$, $\hat{\mu}_{i'} < \hat{\mu}_i$.[7] Thus, with probability $\frac{1}{4}$ over the history of rewards observed, i wins with higher probability than i', since i wins whenever i's payment is the largest by c, and i' can only win when i''s payment is the largest by at least c for any history for which $\hat{\mu}_{i'} < \mu_{i'}$. This is a violation to fairness for $\delta < \frac{1}{4}$.

Notice that Equation 1 implies that each arm receives the highest payment with probability $\frac{1}{k}$, and that this also holds conditioning on any gap c between highest and second-highest payments.

Now, since $\mathcal{D}$ is not peaked,

$$\mathbb{P}_{p\sim\mathcal{D}}\left[\exists i : p_i \geq \max_{i'\neq i} p_{i'} + 1\right] < 1.$$

Define c as follows:

$$c = \sup_{y\geq 0} \text{s.t. } \mathbb{P}_{p\sim\mathcal{D}}\left[\exists i : p_i \geq \max_{i'\neq i} p_{i'} + y\right] = 1.$$

Notice, this implies that:

$$\forall \epsilon > 0, \exists \eta > 0 \text{ s.t. } \mathbb{P}_{p\sim\mathcal{D}}\left[\exists i : p_i \geq \max_{i'\neq i} p_{i'} + c + \epsilon\right] \leq 1 - \eta. \tag{2}$$

We now construct an instance as a function of c. There are two cases – either $c > 0$ or $c = 0$.

Case 1: $c > 0$ Consider the following instance, defined in terms of c and a constant $0 < \epsilon < c$: arm 1 has mean $1 - c$ with deterministic rewards, arm 2 has mean $1 - c$ with reward $1 - c - \epsilon$ with probability $\frac{1}{2}$ and reward $1 - c + \epsilon$ with probability $\frac{1}{2}$, and arms $3, \ldots, k$ have a deterministic reward of 1. Note that by definition of c, and the deterministic nature of arm 1's distribution, we have that for every history h, $\pi^t_{1|h} \geq 1/k$. By the fairness constraint, we must therefore also have that for every other arm $i > 1$, $\pi^t_{i|h} \geq 1/k$, since no other arm has lower mean. This implies that for every arm i, it must be that $\pi^t_{i|h} = 1/k$.

Note, as we argued in footnote 7, that with probability at least $\frac{1}{4}$, for any t, $\hat{\mu}_2 < \hat{\mu}_1 = 1 - c$, by construction. In this case, there is some $\epsilon' > 0$ such that arm 2 is not played unless $p_2 > \max_{i\neq 2} p_i + c + \epsilon'$. However, by definition of c, this occurs with probability strictly less than $1/k$, contradicting the assertion that $\mathcal{D}$ is a fair distribution.

Case 2: $c = 0$ Consider the instance in which arms $1, \ldots, k-1$ have mean $\frac{1}{2}$ and deterministic reward distributions, while arm k has mean 1/2, and stochastic rewards that are $\frac{1}{2} - \epsilon$ with probability $\frac{1}{2}$ and $\frac{1}{2} + \epsilon$ with probability $\frac{1}{2}$. Note that in this case, fairness requires that all arms be played with identical probabilities. With probability at least $\frac{1}{4}$, arm k has empirical mean lower than its true mean. Condition on $\hat{\mu}_k < \mu_k$. In this case, since $c = 0$, with arm k must be selected with probability less than $\frac{1}{k}$ since the payment to arm k will be strictly less than $\mu_k - \hat{\mu}_k$ with strictly positive probability (2), and therefore unfair. □

[7] For t odd it is $\frac{1}{2}$, for even t it is $\frac{1}{2}\left(1 - \cdot\binom{t}{t/2}\right) \cdot \frac{1}{2^t} \geq \frac{1}{4}$, achieved at $t = 2$ and increasing in t.

3.1 A Fair Payment Scheme for Two Arms

We now show that having at least three arms is necessary for our lower bound result. Indeed, in the classic stochastic partial information setting with two arms there exists a simple payment scheme that can ensure fairness in every round while achieving sublinear regret and payment.

The key idea in this payment scheme is to maintain confidence intervals around empirical reward means for the two arms. The following lemma tells us how to construct confidence intervals.

Lemma 3.4 (Lemma 1, [Joseph et al., 2016b]). *Suppose arm i has been pulled n_i^t times before round t. Let $\ell_i^t = \hat{\mu}_i^t - \sqrt{\frac{\ln \frac{(\pi(t+1))^2}{3\delta}}{2n_i^t}}$, and $u_i^t = \hat{\mu}_i^t + \sqrt{\frac{\ln \frac{(\pi(t+1))^2}{3\delta}}{2n_i^t}}$. Then, with probability at least $1-\delta$, for every $i \in [k], t \in [T]$, $\ell_i^t \le \mu_i \le u_i^t$.*

In the light of this result, we will define the function ConfidenceWidth as follows, which will also be useful for describing our payment scheme in the following section.

$$\textsc{ConfidenceWidth}(\delta, t, n) = 2\sqrt{\frac{\ln\left(\frac{(\pi \cdot (t+1))^2}{3\delta}\right)}{n}} \tag{3}$$

Given this confidence width function, our payment scheme is the following: in each round t, choose an arm a^t uniformly at random, and offer payment $p(\delta, t, n_1^t, n_2^t)$ for playing arm a^t and offer 0 for playing the other arm, where

$$p(\delta, t, n_1^t, n_2^t) = \textsc{ConfidenceWidth}(\delta, t, n_1^t) + \textsc{ConfidenceWidth}(\delta, t, n_2^t)$$

and n_1^t, n_2^t denote the number of times that the two arms are played before round t. Whenever the agent selects the arm associated with zero payment, the principal will then offer zero payment for both arms in all future rounds.

Theorem 3.5. *Consider the classic case with $k = 2$ arms in the partial information setting. Then the payment scheme above instantiated with parameter δ is fair in every round with probability at least $1-\delta$. Moreover, the incurred total cost and expected regret are at most $\tilde{O}(\sqrt{T})$.*

4 Classic Setting: Sublinear Payments with Only $\tilde{O}(k^2)$ Unfair Rounds

The necessity of linear growth in subsidies (Theorem 3.1) was driven by the requirement that the agent satisfy contextual fairness in each period. It is natural to ask what would happen if one relaxed this requirement. In this section, we describe how to design a payment scheme which will satisfy contextual fairness in all but $\tilde{O}(k^2)$ rounds. We show that it is possible to achieve payments and regret which grow sub-linearly with T.

The rough idea behind this upper bound is inspired by Joseph et al. [2016b] who show that fairness can be achieved by maintaining confidence intervals around empirical arm means, and enforcing the constraint that any pair of arms with overlapping confidence intervals are played with equal probability: in particular, a fair no-regret algorithm can play uniformly at random amongst the set of arms "chained" to the arm with highest upper confidence bound by the confidence intervals, called the *chained set* X.

Denote the confidence interval associated with arm i at round t by $[\ell_i^t, u_i^t]$. Fix a set of confidence intervals at round t, $[\ell_1^t, u_1^t], \ldots, [\ell_k^t, u_k^t]$. We say i is *linked* to j if $[\ell_i^t, u_i^t] \cap [\ell_j^t, u_j^t] \neq \emptyset$, and i is

chained to j if i and j are in the same component of the transitive closure of the linked relation. We refer to the set of arms chained to the arm with highest upper confidence bound as the *chained set* X. We say the sequence of confidence intervals are *valid* if, with probability $1-\delta$, they contain the true and empirical averages of every arm in every round.

In the absence of explicit knowledge of the sample means, the principal does not have sufficient information to incentivize uniformly random play amongst exactly the set of arms chained to the arm with highest upper confidence bound X[8]. The principal does not know the empirical means of the arms, and therefore cannot compute the arms contained in X directly.

The principal can, however, incentivize the myopic agent to play an arm j with a payment vector p^t such that $p_j^t \geq \max_i p_i^t + |\max_i \hat{\mu}_i^t - \hat{\mu}_j^t|$. Unfortunately, the principal neither knows *which* arms belong to X, nor how many arms are in X, nor how far apart the empirical means are in X. Instead, the principal can maintain upper bounds on all of these quantities. Namely, the principal tracks a superset of the chained set X, called the *active set* $\hat{X}$. $|\hat{X}|$ will then act as an upper bound on the size of the chained set, and $|\hat{X}| \cdot x_{\hat{X}}$ will upper bound the difference between the highest arm mean and the lowest chained arm's means, where $x_{\hat{X}}$ is the width of the largest confidence interval of any arm in $\hat{X}$. By offering a payment of $|\hat{X}| \cdot x_{\hat{X}}$ to an arm selected uniformly from $\hat{X}$ (and zero for all other arms), the principal will cause uniformly random play amongst $\hat{X}$ if all arms in $\hat{X}$ have empirical means within $|\hat{X}| \cdot x_{\hat{X}}$ of the best empirical mean.

This is fair if in every round $\hat{X} = X$: all means will then be within $|\hat{X}| \cdot x_{\hat{X}}$ by the definition of chaining and $x_{\hat{X}}$, and so this will induce uniformly random play amongst the chained set, exactly the behavior shown to be fair in Joseph et al. [2016b]. On the other hand, if $\hat{X} \setminus X \neq \emptyset$, this behavior could be unfair, either because not all arms within $\hat{X}$ have empirical means within $|\hat{X}| \cdot x_{\hat{X}}$ of one another (i.e., not all arms in the set are chained together), or because some arms in $\hat{X}$ chain to other arms outside of $\hat{X}$, or because some arms in $\hat{X}$ are "below" arms outside of $\hat{X}$. We will guarantee the latter issues do not occur, by always ensuring $\hat{X}$ contains any arms "above" or chained to any arm in $\hat{X}$. The former issue (that some arms in $\hat{X}$ may not be chained to others in $\hat{X}$, and their empirical means may then not be close enough for the payment to change the myopic agent's behavior in all cases) cannot be entirely avoided. However, we can quickly discover if any arm in $\hat{X}$ has empirical mean less than $|\hat{X}| \cdot x_{\hat{X}}$ below the best empirical mean: in $O(\hat{X}) = \tilde{O}(k)$ rounds, that arm will be offered the subsidy and it won't change the agent's decision. Those $\tilde{O}(k)$ rounds will be unfair, as are several rounds which follow this discovery and update the set $\hat{X}$.

The following lemma, a generalization of the analysis of Joseph et al. [2016b], can be interpreted to mean the following. Fix a definiton of confidence intervals which are all valid over all rounds for all arms with probability $1-\delta$. Consider any set of arms S which (a) contains the "upper chain" (all arms chained to the arm with highest upper confidence bound), (b) contains any arms "above" the confidence intervals of any arm in the set, and (c) is closed under chaining. Then, playing uniformly at random amongst S will satisfy contextual fairness.

Lemma 4.1. *Suppose, with probability $1-\delta$, at every round t and for every arm i, $\mu_i^t \in [\ell_i^t, u_i^t]$. Consider a set S of arms with the k' highest upper confidence bounds for some $k' < k$. Then, it is fair to play uniformly at random over $S \cup \{i$ chained to an arm in $S\}$.*

The pseudo-code in Figure 1 describes the payment scheme, which we analyze thereafter.

The performance of this payment scheme is summarized in the following theorem.

[8] Indeed, the ability to do so would contradict Theorem 3.1 by acheiving sublinear regret with zero unfair rounds.

ALGORITHM 1: $O(k^2 \ln \frac{k}{\delta})$-fair Payment Scheme

Function PlayAll (δ, T)
  $x \leftarrow 1$;
  $\hat{X} \leftarrow \{1, \ldots, k\}$;
  while $t \le T$ **do**
    $(x, \hat{X}) \leftarrow$ ChainedFair$(\delta, x, \hat{X})$;
  end

Function ChainedFair $(\delta, x, \hat{X})$
  Choose $j^t \in_{\text{UAR}} \hat{X}$; // Pick arm to incentivize
  $x \leftarrow \min(x,$ ConfidenceWidth$(\delta, t, \min_{j \in \hat{X}} n_j^t))$;
  Offer $p^t : p_{j^t}^t = 4x \cdot |X|, p_{i' \neq j^t}^t = 0$;
  $i^t \leftarrow$ the myopic player's choice;
  if $i^t \neq j^t$ **then**
    $\hat{X} \leftarrow$ FindChained$(x, \hat{X}, t)$;
  end
  return $(x, \hat{X})$

Function FindChained $(x, \hat{X}, t)$
  Offer $p^t = \vec{0}$;
  $i^t \leftarrow$ the myopic player's choice;
  $R \leftarrow \{i^t\}$;
  Offer $p^t = p^{t-1} + 2 \cdot x \cdot \sum_{i \in \hat{X} \setminus R} e_i$;
  while $i^t \leftarrow$ *the myopic player's choice and* $i^t \notin R$ **do**
    $R = R \cup \{i^t\}$;
    $t \leftarrow t + 1$;
    Offer $p^t = p^{t-1} + 2 \cdot x \cdot \sum_{i \in \hat{X} \setminus R} e_i$; // add $2 \cdot x$ to payments of arms in $\hat{X}$ not yet chosen
  end
  return R

Theorem 4.2. *For any* δ, PlayAll *is* $O(k^2 \ln(k/\delta))$*-fair, and has expected cost and regret*

$$O\left(k \cdot \sum_t \text{ConfidenceWidth}(\delta, t, \frac{t}{k})\right) = O\left(\sqrt{k^3 T \ln \frac{T}{\delta}}\right).$$

We present the proof to this theorem after stating several lemmas describing the behavior of PlayAll. Observation 2 states that using x as a confidence interval width for all arms in $\hat{X}$ yields valid confidence intervals. Hereafter, we use $[\ell_i^t, u_i^t] = [\hat{\mu}_i^t - x, \hat{\mu}_i^t + x]$ as valid confidence intervals for all $i \in \hat{X}$, $t \in [T]$. Lemma 4.3 shows that FindChained outputs a set which contains the upper confidence chain in its output round. Lemma 4.4 states that FindChained's output is closed under chaining (e.g., that every arm in its output is only chained to arms also belonging to the output set) and contains all arms "above" any arms in its output. Lemma 4.5 argues that the empirical means of every arm in the set output by FindChained are within 4 confidence interval widths of some other arm in the set. Lemma 4.6 shows that when this is the case (that the empirical means are within $4x$ of each other, as is the case right after a call to FindChained), that ChainedFair induces uniformly random play amongst $\hat{X}$. Lemma 4.7 upper-bounds the number of rounds before which ChainedFair will discover when it is inducing unfair play. All proofs of these lemmas can be found in Section A.

Observation 2. *With probability* $1 - \delta$*, for all* $t \in [T]$, $i \in \hat{X}^t$, $\mu_i \in [\hat{\mu}_i^t - x, \hat{\mu}_i^t + x]$.

Lemma 4.3. FindChained$(x, \hat{X}, t)$ *contains all arms chained to the arm with highest upper confidence bound in its output round* t'.

Lemma 4.4. *Any arm chained to the set* $R =$ FindChained$(x, \hat{X}, t)$ *belongs to* R*. Moreover, any arm* $i \notin R$ *must have* $u_i^t < \min_{i' \in R} \ell_{i'}^t$.

Lemma 4.5. *Let* t' *be the round in which* $R =$ FindChained$(x, \hat{X}, t')$ *outputs* R*. Then, for any* $j \in R =$ FindChained$(x, \hat{X}, t')$,

$$\hat{\mu}_j^{t'} \ge \min_{j' \in R \setminus \{j\}} \hat{\mu}_{j'}^{t'} - 4 \cdot x.$$

Moreover, $\max_{j\in R}\hat{\mu}_j^{t'} - \min_{j\in R}\hat{\mu}_j^{'t} \le (2|R|+2)\cdot x$.

Lemma 4.6. *Suppose* $\max_{i,j\in\hat{X}}|\hat{\mu}_i^t - \hat{\mu}_j^t| \le 4|\hat{X}|\cdot x$. *Then* ChainedFair$(\delta,\hat{X})$ *induces uniformly random play amongst* $\hat{X}$.

Lemma 4.7. *Whenever there is an arm* i *such that* $\max_{j\in\hat{X}}|\hat{\mu}_j^t - \hat{\mu}_i^t| > 4|\hat{X}|\cdot x$, *with probability* $1-\delta$, FindChained *will be called within* $O(k\cdot\ln(1/\delta))$ *many rounds.*

Proof of Theorem 4.2. We first upper-bound the number of rounds in which PlayAll might violate the fairness condition.

We argue iteratively about the set $\hat{X}$: that (a) all arms chained to the top arm belong to $\hat{X}$, and (b) all arms chained to any arm in $\hat{X}$ belong to $\hat{X}$. This is trivially true initially as $\hat{X} = \{1,\dots,k\}$. $\hat{X}$ is only updated as the result of a call to FindChained. By Lemma 4.3, any arm chained to the top arm will remain in $\hat{X}$. Furthermore, by Lemma 4.4, any arm chained to an arm in its output also belongs to its output. Thus, (a) and (b) hold for $\hat{X}$ for all rounds.

So, in rounds in which ChainedFair induces uniformly random play amongst $\hat{X}$, (a) and (b) imply ChainedFair satisfies the fairness condition. For any round in which $\max_{i,j\in\hat{X}}|\hat{\mu}_i^t - \hat{\mu}_j^t| \le 2|\hat{X}|\cdot x$, Lemma 4.6 implies ChainedFair induces uniformly random play amongst $\hat{X}$. By Lemma 4.4, $\hat{X}$ contains any arms either above or chained to arms in $\hat{X}$. Thus, Lemma 4.1 applies, and these rounds are fair.

We now upper-bound the number of rounds for which ChainedFair does not induce uniformly random play amongst $\hat{X}$. For any particular i and round t such that $\max_{j\in\hat{X}}|\hat{\mu}_j^t - \hat{\mu}_i^t| > 4|\hat{X}|\cdot x$, Lemma 4.7 implies that this will be found in $O(k\ln(1/\delta))$ rounds, and FindChained will be called. In any future round $t' \ge t$, since the confidence intervals are valid, we know that $\max_{j\in\hat{X}}|\hat{\mu}_j^{t'} - \hat{\mu}_i^{t'}| > 4(|\hat{X}|-2)\cdot x$, since either of the two means can change but by at most x each. Lemma 4.5 will return $\hat{X}$ such that $\max_{i,j\in\hat{X}}|\hat{\mu}_i^t - \hat{\mu}_j^t| \le (2|\hat{X}|+2)\cdot x$. Thus, as $|\hat{X}| \ge 2$, then arm i will be removed at the first round in which it was the impetus for FindChained to be called as $\max_{i,j\in\hat{X}}|\hat{\mu}_j^{t'} - \hat{\mu}_i^{t'}| \le 2(|\hat{X}|+2)\cdot x \le 4(|\hat{X}|-2)\cdot x < \max_{j\in\hat{X}}|\hat{\mu}_j^t - \hat{\mu}_i^t|$, a contradiction if $i \in \hat{X}$.

Since x is non-increasing, so is $\hat{X}$: thus, at most k calls to FindChained are made. Thus, the total number of unfair rounds is equal to the number of rounds in which $\max_{i,j\in\hat{X}}|\hat{\mu}_i^t - \hat{\mu}_j^t| > 4|\hat{X}|\cdot x$ plus the number of rounds in FindChained. The former is bounded by $O(k^2\ln(k/\delta))$ (With probability $1-\delta/k$ it will take at most $O(k\ln(k/\delta))$ rounds of unfair play before FindChained is called when this is the case, and each call will reduce the size of $\hat{X}$ so it can be called at most k times. In total, this bound holds for all k rounds with probability $1-\delta$.); the latter by $O(k^2)$ (each call of FindChained uses $O(k)$ rounds, and there are at most $O(k)$ calls to FindChained).

We now upper-bound the cost of this payment scheme and the regret of the agent. In the $O(k^2\ln(k/\delta))$ unfair rounds, the payments might be $\Omega(1)$; similarly, the regret of the algorithm in those rounds might be $\Omega(1)$. In all other rounds, the myopic agent is playing uniformly at random amongst a set of arms whose true means are within $2k\cdot x$ of the best true mean, so $2k\cdot x$ in each fair round is an upper-bound on per-round regret. The maximum payment offered in any round is $4k\cdot x$ as well, so that also upper bounds the cost. The overall upper bound follows from some basic algebra and the fact that each arm in $\hat{X}$ will have been played $\tilde{\Omega}\left(\frac{t}{k}\right)$ times in round t. □

5 Contextual Setting with Partial Information: Linear Payments or Unfair Rounds

In this section, we argue that the partial information model is much harder in the linear contextual case — in *every* round that the principal does not pay $\Omega(1)$, an adversary can force the myopic agent to behave unfairly. This implies that on an adversarially chosen instance, every round is either unfair or has constant cost: thus, either the sum of the payments must be $\Omega(T)$, or the number of unfair rounds must be $\Omega(T)$, or both. This rules out positive results in the partial information model of the sort we were able to obtain in the classic bandits setting. In the following, we assume that the myopic agent is using an ordinary least squares estimator, for simplicity. Identical results can also be proven for other natural estimators, like ridge regression estimators.

Theorem 5.1. *Suppose $k \geq 3$. Consider any payment scheme in the partial information model in the linear contextual bandit setting. For any $\eta \in (0,1)$, there is an instance for which with probability $1-\delta$, in every round, either the round is unfair, or the expected cost for the principal is $\frac{k-1}{k} \cdot (1-\eta)$.*

The proof of the theorem relies on the fact that the principal cannot observe the adversarially chosen contexts; the expected rewards in any round then can be (almost) arbitrary. In the classic case, it was only in the first unpeaked round that we had the freedom to design our lower bound instance arbitrarily – after that, the principal would have learned some information about the instance, and hence the payment distribution could be a function of the instance. In the linear contextual case, we have sufficient freedom to design a lower bound instance at -every- round. Although the principal may have learned a great deal about the underlying linear functions, she by definition has no information about the realized contexts at the current round, which we use to our advantage. As in the classic setting, in any round where the payment scheme is not peaked, the largest payment is strictly less than 1 larger than the other payments with probability more than zero. We will use this to construct an instance over which there is constant probability (over the history) that the myopic agent chooses an unfair distribution over arms. Additional complications arise from the fact that the principal learns about the instance from the set of previous unfair rounds (which, in the classic case, we did not have, since we only argued there had to be a single unfair round if the payment scheme was not peaked). We circumvent this problem by arguing that the principal must deploy a peaked distribution to be fair, even if the principal knows everything about the instance I and even if the principal knows the empirical estimates $\hat{\theta}_i^t$ for all $t \in [T], i \in [k]$.

Proof. Consider the one-dimensional case, where $\theta_i \in \mathbb{R}_{\geq 0}$. We construct an instance I such that even for a principal who has full information about I, and $\hat{\theta}_i^t$ for all $t' \leq t, i$, in order to guarantee that the payment distribution in round t is fair for any set of arriving contexts x^t, the largest payment must be at least $1-\eta$ with probability $\frac{k-1}{k}$. This clearly holds for any round in which a peaked payment distribution is used, and so for the remainder, we assume that the distribution in round t is not peaked.

Let $\theta_i = 1-\eta \in (0,1)$ for all i, and let arm 1 have deterministic rewards equal to their mean, so that $\theta_1 x_1^t = x_1^t$ for all t. Because the rewards are deterministic and the agent is using an ordinary least squares estimator, the myopic agent's prediction $\hat{\theta}_1^t = \theta_1^t$ as well for all t. For all $i \neq 1$, let $\mathcal{D}_{i,x_i^t}^t = U[\theta_i x_i^t - \epsilon, \theta_i x_i^t + \epsilon]$ for some very small ϵ. $\mathbb{P}_{r_i^t \sim \mathcal{D}_{i,x_i^t}^t}\left[r_i^t > \theta_i x_i^t\right] = \frac{1}{2} = \mathbb{P}_{r_i^t \sim \mathcal{D}_{i,x_i^t}^t}\left[r_i^t < \theta_i x_i^t\right]$: the rewards drawn from these distributions have the right expectation but are always larger or smaller

than their expectation, and each with equal probability. Then, again by properties of the ordinary least squares estimator, this will imply that with probability $\frac{1}{2}$ over observations, in any round t and for any $i \in [k] \setminus \{1\}$, $\hat{\theta}_i^t > \theta_i^t$, and with probability $\frac{1}{2}$, $\hat{\theta}_i^t < \theta_i^t$, for any round t. Furthermore, with probability 1, in every round t, every empirical estimate of the coefficients is distinct: $\hat{\theta}_i^t \neq \hat{\theta}_j^t$ for all $i \neq j \in [k]$.

We begin by arguing that every coordinate i must have equal probability of receiving the largest payment in any round t if the round is to be fair (with probability $1-\delta$ over the history). Precisely, fix some history h^t, and let

$$d_i = \mathbb{P}_{p \sim \gamma^t(h^t)}\left[i \text{ wins with payment vector } p, x_i^t = \vec{0}, \forall i | h^t\right].$$

Since for all $i \in [k]$ and any h^t, $\theta_i \cdot x_i^t = 0$, it must be that $d_i = \frac{1}{k}$ for all i if round t is fair for this h^t.

We have assumed the payment scheme is not peaked in round t, conditioned on some particular history h^t. Thus,

$$\mathbb{P}_{p \sim \gamma^t(h^t)}\left[\max_i p_i - \max_{j \neq i} p_j \geq 1\right] < 1.$$

We argue that round t must be unfair conditioned on h^t or that with probability $\frac{k-1}{k}$, $\max_i p_i \geq 1-\eta$. Let

$$c = \sup_c : \mathbb{P}_{p \sim \gamma^t(h^t)}\left[\max_i p_i - \max_{j \neq i} p_j \geq c\right] = 1;$$

we again know that some such $c \geq 0$ must exist, and that $c < 1$ because the payment scheme is unpeaked. Let arm $\bar{i}$ have the largest empirical coefficient: $\hat{\theta}_{\bar{i}}^t > \max_{i' \neq \bar{i}} \hat{\theta}_{i'}^t$ in round t and arm $\underline{i}$ have the smallest empirical coefficient, $\hat{\theta}_{\underline{i}}^t < \min_{i' \neq \underline{i}} \hat{\theta}_{i'}^t$. Further define

$$c_{i'} = \sup_c : \mathbb{P}_{p \sim \gamma^t(\cdot)}\left[p_{i'} - p_{\bar{i}} \geq c | p_{i'} \geq \max_{i'' \neq i'} p_{i''}\right] = 1,$$

e.g. that $c_{i'}$ is the margin by which i' has payment larger than arm $\bar{i}$ when i' has largest payment. Note $c_{i'} \in [0,1]$ for all i'. Let $i_{\max} \in \arg\max_{i'} c_{i'}$ be an arm with largest payment margin over $\bar{i}$ and $i_{\min} = \text{argmin}_{i'} c_{i'}$ be the arm with the smallest payment margin over i. We consider three cases: when $c_{i_{\max}} > 1-\eta$, when $1-\eta \geq c_{i_{\max}} > c_{i_{\min}}$, and when $1-\eta \geq c_{i_{\max}} = c_{i_{\min}}$. In each case, we show that either the largest payment is at least $1-\eta$ with probability at least $\frac{k-1}{k}$, or the round is unfair.

Case 1: $c_{i_{\max}} > 1-\eta$ We claim here that either $c_{i_{\min}} > 1-\eta$ or the round is unfair: this will imply that with probability $\frac{k-1}{k}$, $\max_i p_i \geq 1-\eta$. Suppose the round is fair. Consider the context $x_{\bar{i}}^t = \frac{1-\eta}{\hat{\theta}_{\bar{i}}^t}$ and $x_{i'}^t = 0$ for all $i' \neq \bar{i}$. Then, $\hat{\theta}_{\bar{i}}^t x_{\bar{i}}^t = 1-\eta$, and $\hat{\theta}_i^t x_i^t = 0 = \theta_i x_i^t$ for all other i. Fairness will imply that all $i \neq 1$ should be played with equal probability. Notice that $i_{\max}$ is played with probability $\frac{1}{k}$: precisely when $i_{\max}$ has the largest payment (which must be largest by $c_{i_{\max}} > 1-\eta$). $i_{\min}$ wins only when her payment is largest (which happens with probability $\frac{1}{k}$) and larger than $\bar{i}$'s by at least $1-\eta$. So, if $\pi_{i_{\min}|h^t}^t = \pi_{i_{\max}|h^t}^t = \frac{1}{k}$, it must be that $c_{i_{\min}} \geq 1-\eta$.

Case 2: $1-\eta \geq c_{i_{\max}} > c_{i_{\min}}$ We argue that the round must be unfair if $c_{i_{\max}} > c_{i_{\min}}$.

Choose contexts x_i^t such that $\hat{\theta}_{\bar{i}}^t x_{\bar{i}}^t = c_{i_{\max}} \leq 1-\eta$ and $x_{i'}^t = 0$ for all other i'. Then, since $\theta_i x_i^t = 0$ for all $i \neq \bar{i}$, if this round is to be fair, all arms $i \neq \bar{i}$ must be played with equal probability. Arm $i_{\max}$ wins whenever it has the largest payment, since $p_{i_{\max}} \geq p_{\bar{i}} + c_{i_{\max}}$ whenever $i_{\max}$ has the largest payment. Therefore $i_{\max}$ wins with probability $\frac{1}{k}$.

$i_{\min}$, on the other hand, wins only when they have the largest payment and beat i's payment by $c_{i_{\max}} > c_{i_{\min}}$, which happens with strictly less probability than $i_{\min}$ having largest payment (probability $\frac{1}{k}$) by the definition of $c_{i_{\min}}$. So, $i_{\min}$ wins with probability strictly less than that of $i_{\max}$; this round must be unfair.

Case 3: $1-\eta \geq c_{i_{\max}} = c_{i_{\min}}$ In this case, $c_a = c_b = \beta$ for all $a, b \in [k] \setminus \{\bar{i}\}$. If $\beta \geq 1-\eta$, the claim holds (the largest payment is at least $1-\eta$ with probability at least $\frac{k-1}{k}$, so assume $\beta < 1-\eta$.

Suppose $\beta > 0$. We exhibit a set of contexts for which this payment scheme combined with the agent is unfair. Fix some $D \in (\beta, 1-\eta)$; define the contexts

- $x_{\bar{i}}^t : \hat{\theta}_{\bar{i}}^t x_{\bar{i}}^t = D > \beta$
- $x_j^t : \hat{\theta}_j^t x_j^t = D - \beta > 0$ for j the arm with second-largest empirical coefficient,
- $x_{i'}^t = x_j^t$ for all $i' \notin \{\bar{i}, j\}$.

Then, $\hat{\theta}_{i'}^t x_{i'}^t < \hat{\theta}_j^t x_j^t$, and so arm j is played whenever j has the largest payment, since j (and all other arms) has margin over $\bar{i}$ of at least β when they have the largest payment; thus j is played with probability $\frac{1}{k}$. Since $\theta_j x_j^t = \theta_{i'} x_{i'}^t$ for all $i' \neq \bar{i}$, if this round is fair, each i' must also be played with probability $\frac{1}{k}$, in particular for i' with smallest $\hat{\theta}_{i'}^t$. However, $\hat{\theta}_{i'}^t x_{i'}^t < D - \beta$; i' can only win if her payment is the largest and it beats the payment of i by strictly more than β, which happens with probability strictly less than $\frac{1}{k}$ by definition of β. Thus i' cannot win with probability as large as j and so round t is unfair if $c_a = c_b = \beta > 0$ for all $a, b \neq i$.

Finally, we consider the case where $\beta = 0$ and separately argue that this round cannot be fair. The contexts $x_{i'}^t = 1$ for all i' should prove this: arm $\bar{i}$ will be played with probability $\frac{1}{k}$ (precisely the probability that $\bar{i}$ gets the weakly largest payment), but arms with smaller empirical means will need to have the largest payment by some margin, which happens with strictly less probability than them having the largest payment by the definition of β, so they win with probability less than $\frac{1}{k}$, meaning fairness is violated in this round, since $\theta_{\bar{i}} x_{\bar{i}}^t = 1 - \eta = \theta_{i'} x_{i'}^t$. □

6 Full Information: Perfect Fairness with Sublinear Payments

In this section, we show that a principal with full information about the state of a myopic agent can design a payment scheme which is fair in every round and has sublinear cost for both the classic and linear contextual bandits problems. This contrasts with the partial information model, where for $k \geq 3$ arms, in both the classic and linear contextual settings, in which there must be unfair rounds for any payment scheme with total cost $o(T)$.

Roughly, the fair payment scheme operates as follows. In each round, the scheme knows the empirical estimates of rewards used by the myopic agent. Moreover, the scheme can compute

confidence intervals around these estimates (the scheme knows how many times each arm was pulled, and, in a contextual setting, the contexts for each previous choice). In such a round, the payment scheme then will choose an arm i uniformly at random from the set of arms chained to the arm with highest upper confidence bound, and offer a payment for choosing i equal to the difference between the empirical estimate of i's reward and the empirical estimate of the highest reward in that round. This induces uniformly random play amongst the top set of arms, and by Lemma 4.1, this will be a fair distribution.

We now present the pseudocode in Figure 2 a parametrized family of payment schemes described informally above. A payment scheme in this family is instantiated by giving a method of constructing valid confidence intervals around myopic predictions.

ALGORITHM 2: A Fair Full Information Payment Scheme

Function *Fair-Payments*(δ, T)
  $\hat{X} \leftarrow \{1, \ldots, k\}$;
  while $t \leq T$ **do**
    $i^t = \arg\max_i \hat{\mu}_{it}^t$;
    Let $\hat{X}^t = \{i$ chained to i^t by δ-valid confidence intervals from round $t\}$;
    Choose $j^t \in_{\text{UAR}} \hat{X}$;
    ; // Pick an arm in the upper chain to incentivize
    Offer $p^t : p_{j^t}^t = \hat{\mu}_{i^t}^t - \hat{\mu}_{j^t}^t, p_{i' \neq j^t}^t = 0$;
  end

Theorem 6.1. *Consider an instance of Fair-Payments(δ, T) instantiated with confidence intervals $[\ell_i^t, u_i^t]$ such that $\hat{\mu}_i^t = \frac{u_i^t - \ell_i^t}{2}$, and with probability $1-\delta$, for all $i \in [k], t \in [T]$, $\mu_i^t \in [\ell_i^t, u_i^t]$. Then, Fair-Payments$(\delta, T)$ is fair at every round, and has cost and regret $O(k \sum_t w(t) + \delta T)$, where $w(t)$ is the maximum width of any confidence interval in the top chained set.*

Before proving Theorem 6.1, we mention that this theorem, when combined with standard methods of constructing confidence intervals, implies the existence of fair payment schemes with sublinear cost and regret, both in the classic and linear contextual settings.

Corollary 6.2. *Consider the classic bandits problem. Then, Fair-Payments(δ, T) using the confidence interval for arm i introduced by* ConfidenceWidth(δ, T, n_i^t) *is fair and has cost and regret $O(\sqrt{k^3 T \ln \frac{kT}{\delta}})$.*

Proof. By Lemma 3.4, with probability $1 - \delta$, these confidence intervals are all valid for all $t \in [T], i \in [k]$. So, Theorem 6.1 applies, and states that this payment scheme is fair, and has regret $O(k \cdot \sum_t w(t))$, where $w(t)$ is the maximum width of any arm in the active set at round t. Since the chained set is monotone, at round t every arm in the chained set has been chained for t rounds. Therefore, in expectation each arm in the chain has been pulled $\frac{t}{k}$ times. An additive Chernoff bound implies that any particular arm has, with probability at least $1 - \frac{\delta}{2kt^2}$, been pulled in round t at least $\frac{t}{k} - \sqrt{\frac{t \ln\left(\frac{2t^2 k}{\delta}\right)}{2}}$ times, and so this bound holds for all rounds and all arms with probability at lest $1 - \frac{\delta}{2}$ summing up over all k arms and all t. Then, by Lemma 3 in Joseph et al. [2016b], we know that $w(t) \leq 2\sqrt{\frac{\ln((\pi t)^2/3\delta)}{2\frac{t}{k} - \sqrt{\frac{t \ln\left(\frac{2kt^2}{\delta}\right)}{2}}}}$. Summing over all t we have the desired result. □

Corollary 6.3. *Consider the linear contextual bandits problem. Suppose the myopic agent uses a ridge regression estimator:* $\hat{\theta}_i^t = \left(X_i^T X_i + \lambda I\right)^{-1} X_i^T Y_i$*, where* X_i, Y_i *are the design matrices and observations before round* t*. Then, define*

$$w_i^t = \|x_i^t\|_{\left(X_i^T X_i + \lambda I\right)^{-1}} (m\sqrt{d \ln \frac{1 + t/\lambda}{\delta}} + \sqrt{\lambda},$$

and

$$\ell_i^t = \langle \hat{\theta}_i^t, x_i^t \rangle - w_i^t, \qquad u_i^t = \langle \hat{\theta}_i^t, x_i^t \rangle + w_i^t.$$

Then, Fair-Payments(δ, T) *is fair and has cost and regret* $O\left(md\sqrt{k^3 T} \ln^2 \frac{T^2 k}{d\lambda\delta}\right)$.

Proof. The analysis in the proof of Theorem 2 of Joseph et al. [2016a] shows that these confidence intervals are valid with probability $1 - \delta$. Their analysis upper-bounds $k \sum_t w(t)$ where $w(t)$ is the largest width of any confidence interval in the chained set in round t, by

$$O\left(md\sqrt{k^3 T} \ln^2 \frac{T^2 k}{d\lambda\delta}\right).$$

Thus, the resulting algorithm is fair, and the bounds on cost and regret follow from Theorem 6.1. □

We now proceed with the proof of Theorem 6.1.

Proof. We begin by proving that Fair-Payments(δ, T) is fair in every round. By assumption, with probability $1 - \delta$, for all $i \in [k], t \in [T]$, $\mu_i^t \in [\ell_i^t, u_i^t]$. Thus it follows from Lemma 4.1 that it suffices to show that these payments suffice to induce uniformly random play amongst the set of arms chained to the arm with upper confidence bound. By definition, the top chain in round t is exactly $\hat{X}^t$. The distribution over payments in round t chooses each $j^t \in \hat{X}^t$ with probability $\frac{1}{|\hat{X}^t|}$ and accordingly a p^t such that $p_{j^t}^t = \hat{\mu}_{i^t}^t - \hat{\mu}_{j^t}^t$ and $p_{i'}^t = 0$. This induces the myopic agent to choose j^t in all such cases. Thus, each $j^t \in \hat{X}^t$ is chosen by the myopic agent with probability $\frac{1}{|\hat{X}^t|}$. So, Fair-Payments(δ, T) is fair.

Condition on all confidence intervals being valid. The myopic agent under this payment scheme chooses uniformly at random from the top chain, which has regret in round t bounded by $\sum_{i \in \hat{X}^t} u_i^t - \ell_i^t \leq k w(t)$, where $w(t) = \max_{i \in \hat{X}^t} u_i^t - \ell_i^t$. Thus, in total, the regret is upper bounded by $k \sum_t w(t)$. Moreover, the payment in round t is $\hat{\mu}_{i^t}^t - \hat{\mu}_{j^t}^t \leq u_{i^t}^t - \ell_{j^t}^t \leq \sum_{i \in \hat{X}^t} u_i^t - \ell_i^t \leq k w(t)$, and so the same bound holds for the cost of the payment scheme. With probability δ, the widths of the confidence intervals could be arbitrary, as could the inaccuracy of the sampled means. An additive δT bounds the additional expected regret. □

7 Conclusion and Open Questions

Our interest in this paper is the information that a principal needs to have about the environment before he can cost-effectively incentivize a short-sighted agent to behave "fairly." We focus on two information models: the partial information model—when the principal can only observe the decisions the agent made, but not their rewards (or, in the contextual case, the contexts informing

those actions), and the full information model, where the principal observes everything the agent does. In the full information model, it is possible to have it all—the principal can with sub-linear total cost incentivize the agent to play fairly at every round, and obtain no regret. In the partial information model, things are more difficult. However, despite showing the impossibility of non-trivially guaranteeing fairness at every round in the classic setting, we show that with sub-linear payments, the principal can incentivize that all but a constant number of rounds are fair, and that the agent obtains a no-regret guarantee. In the linear contextual bandit setting, our results in the partial information model are strongly negative—it is not possible to obtain a sub-linear number of unfair rounds with sub-linear payments. There are many open questions, but here we mention two that we find particularly interesting:

1. Our bounds (both upper and lower) in the linear contextual bandit setting are for *adversarially selected contexts*. In the natural case in which contexts are drawn from an (unknown) probability distribution, it may still be possible to obtain positive results in the partial information setting, analogous to the results we obtain for the classic bandits problem. However, our upper bound technique from the classic case does not directly extend to the linear contextual case even when there is a distribution over contexts.

2. The friction to fairness here is that the agent in question has a short horizon for which he is optimizing. We study the extreme case in which he is entirely myopic. How do our results extend in the case in which the agent is not completely myopic, but is instead optimizing with respect to some fixed discount factor bounded away from 1?

References

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *CALIFORNIA LAW REVIEW*, 104:671, 2016.

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

Yeon-Koo Che and Johannes Horner. Optimal design for social learning. Technical report, Cowles Foundation for Research in Economics, Yale University, 2015.

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

Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22, 2017.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22, New York, NY, 2014. ACM, ACM.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. Technical report, University of Pennsylvania, 2016a.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 325–333. Curran Associates, Inc., Red Hook, NY, 2016b.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, New York, NY, 2015. ACM, ACM.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*, page 661, 2016. doi: 10.1145/2940716.2940755.

Daniel O'Leary. Exploiting big data from mobile device sensor-based apps: Challenges and benefits. *MIS Quarterly Executive*, 12(4):179–187, December 2013.

Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Sava. Crowdsourcing exploration. *Management Science*, 2017. Forthcoming.

Devin Pope and Justin Snyder. What's in a picture? evidence of discrimination from prosper.com. *Journal of Human Resources*, 101(1):53–92, 2011.

A Missing Proofs

Proof of Theorem 3.5. First, we will show that the payment scheme will incentivize the agent to select the arms fairly over all rounds with probability $(1-\delta)$ over the realization of the history. By Lemma 3.4, we know that with probability at least $1-\delta$ over the realizations of the rewards, for all rounds t and both arms i,

$$|\hat{\mu}_i^t - \mu_i| < \frac{\textsc{ConfidenceWidth}(\delta, t, n_i^t)}{2}. \tag{4}$$

We will condition on this event for the remainder of the argument. Note that in each round t, there are two cases. In the first case, the empirical mean rewards of the two arms satisfy $|\hat{\mu}_1^t - \hat{\mu}_2^t| < p(\delta, t, n_1^t, n_2^t)$. Then the arm will be selected uniformly at random, so the algorithm is fair.

In the second case, the empirical mean rewards satisfy $|\hat{\mu}_1^t - \hat{\mu}_2^t| \geq p(\delta, t, n_1^t, n_2^t)$. Without loss of generality, let us assume $\hat{\mu}_1^t > \hat{\mu}_2^t$, so the algorithm will deterministically always play arm 1. Then it follows from (4) and the definition of $p(\delta, t, n_1^t, n_2^t)$ that the true mean rewards $\mu_1 > \mu_2$. Therefore, the algorithm is fair in this case (and also in all future rounds).

Next, we will bound the total expected payment made by the principal. As a first step, we can show that with probability at least $(1-1/T)$ that, no arm is played for more than $2\log T$ consecutive

times. Then we can bound the total payment as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} p(\delta, t, n_1^t, n_2^t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\Big(\textsc{ConfidenceWidth}(\delta, t, n_1^t) + \textsc{ConfidenceWidth}(\delta, t, n_2^t)\Big)\right] \\ &\leq (1-1/T)\left(4\log(T)\sum_{t=1}^{T}\textsc{ConfidenceWidth}(\delta, t, t)\right) + (1/T)T\,O(\log(T/\delta)) \\ &= O\Big(\sqrt{T}\ln^2(T/\delta)\Big)\end{aligned}$$

Finally, we will bound the expected cumulative regret incurred by the algorithm. Without loss of generality, we will assume $\mu_1 > \mu_2$ and let $\Delta = \mu_1 - \mu_2$ be the difference between the mean expected rewards. The algorithm incurs an expected regret of Δ whenever it plays the arm 2. It suffices to bound the number of times arm 2 is played.

Note that the algorithm will stop playing arm 2 if at some round t,

$$\frac{\textsc{ConfidenceWidth}(\delta, t, n_1^t)}{2}, \frac{\textsc{ConfidenceWidth}(\delta, t, n_2^t)}{2} \leq \Delta/3$$

This will require the algorithm to play both arms $S = O\Big(\frac{\ln(T/\delta)}{\Delta^2}\Big)$ number of times. By applying Chernoff bounds, we know that with probability at least $(1-1/T)$, it suffices to have $t \geq O(S)$. This will allow us to upper bound the expected regret by

$$(1-1/T)\Delta O\left(\frac{\ln(T/\delta)}{\Delta^2}\right) + T(1/T) = O\left(\frac{\ln(T/\delta)}{\Delta}\right)$$

Also note that the expected regret is also trivially upper bounded by ΔT, so the expected regret is no more than

$$\min\{O\left(\frac{\ln(T/\delta)}{\Delta}, \Delta T\right)\} \leq O\Big(\sqrt{T\ln(T/\delta)}\Big),$$

where the inequality follows from the fact that $\min\{a, b\} \leq \sqrt{ab}$ for any $a, b > 0$. □

Proof of Observation 2. By Lemma 4.1, the width of the confidence intervals produce by ConfidenceWidth have this property in all rounds. Since x is defined to be the either an "old" value of the output of ConfidenceWidth (in which case the validity of the definition of ConfidenceWidth's confidence widths gives us this guarantee), or the largest output of ConfidenceWidth for some $i \in \hat{X}^t$, this property continues to hold. □

Proof of Lemma 4.3. We prove this iteratively over all inputs and outputs of FindChained. The input to the first call to FindChained is $\hat{X} = [k]$, so this is trivially true. Now, suppose the upper chain is included in the input $\hat{X}$ to FindChained: we will argue that it continues to be included in the output $R = \textsc{FindChained}(x, \hat{X}, t)$. We actually prove something stronger: any arm in the input $\hat{X}$ will be in R if its empirical mean in the output round t' is within $2x$ of anything in R's empirical mean in round t', and that the arm with highest upper confidence bound in round t' is in R. These two together imply R contains the upper chain. Let $R^t = \emptyset, R^{t+\ell}$ = the set R in round $t+\ell$ before R is output by FindChained, so $R^{t'} = R$.

Note that $R^{t+1} = \{i_0\}$ where i_0 is the arm with highest empirical mean in $\hat{X}$ with highest empirical mean at round t. Then, either $R^{t+2} = \{i_0\}$ and no arm has empirical mean within $2x$ of

i_0 in round $t+1$ or $R^{t+2} = \{i_0, i_1\}$ for $i_1 \in \hat{X}$ such that $\hat{\mu}_{i_1}^{t+1} \geq \hat{\mu}_{i_0}^{t+1} - 2x$. More generally, in round $t+\ell < t'$, an arm was added in round $t+\ell-1$, and another arm will be added in round $t+\ell$ if if some empirical mean in $\hat{X}$ is within $2x$ of any arm already belonging to $R = \{i_0, i_1, \ldots, i_{t+\ell-1}\}$, since the payments for any arm in $\hat{X}$ but not R increases by $2x$ in this round. Thus, every arm which is not in the output R must be more than $2x$ away from any arm in the output R in round t'.

We argue now that the arm with highest upper confidence bound in round t' belongs to R. Since $\hat{X}$ contained the upper chain in round t by assumption, it in particular contains the arm with the highest upper confidence in round t'. Thus, it suffices to argue that R contains the arm in $\hat{X}$ with highest upper confidence bound in round t'. Either arm i_0 or some arm with empirical mean within $2x$ of i_0 at round t' must be the arm with highest upper confidence bound in round t' amongst those arms in $\hat{X}$, since x is an upper bound on the width of the confidence intervals of the set of arms in $\hat{X}$, and by the previous argument, R contains every arm whose empirical mean in round t' is within $2x$ of i_0's mean. □

Proof of Lemma 4.4. We first prove the first claim. This is true for the first input to FindChained: we argue that for any input $\hat{X}$ which contains every arm chained to an arm in $\hat{X}$, $R = \text{FindChained}(x, \hat{X}, t)$ contains any arm chained to an arm in R. So, suppose $\hat{X}$ contains every arm chained to an arm in $\hat{X}$. By the argument used in the proof of Lemma 4.3, any arm with an empirical mean within $2x$ of any arm in R's empirical mean in round t' belongs to R. So, any arm linked to an arm in R must belong to R and therefore so must any arm chained to R.

Now, we argue that for any arm $i \notin R$ must have $u_i^t < \min_{i' \in R} \ell_{i'}^t$. This is true for the first input to FindChained (the entire set $[k]$ is the initial input). We argue that conditioned on this holding for a particular input to FindChained, the output from FindChained will also satisfy this claim. Notice that every arm in $i' \in R$ was incentivized to be played during this call to FindChained, and those arms no longer in R were not, which means their empirical means were more than $2x$ away from any arm ultimately in R; that for $i \notin R$, $\hat{\mu}_i^t + 2x < \min_{i' \in R} \hat{\mu}_{i'}^t$. Thus, since $\ell_j^t = \hat{\mu}_j^t - x$ and $u_j^t = \hat{\mu}_j^t + x$ for all $j \in \hat{X}$, the claim follows. □

Proof of Lemma 4.5. The empirical mean of any arm in the output of FindChained had to be within $2x$ of some arm in R when it was input. Those empirical means might change (each mean by at most x, since the confidence intervals are valid, by Observation 2), so the empirical differences might change by $2 \cdot x$, but the total difference between these two arms is then increased by at most $2x$. Thus, summing up these distances gives the second claim. □

Proof of Lemma 4.6. Offering payment of $4x|\hat{X}|$ for arm i' and payment 0 for all other arms j in round t will cause a myopic agent to choose i if $\max_{i \in \{k\}} \hat{\mu}_i^t - \hat{\mu}_{i'}^t \leq 4|\hat{X}| \cdot x$. Thus, if $\arg\max_{i \in \{k\}} \hat{\mu}_i^t \in \hat{X}$, each $i' \in \hat{X}$ will be played with equal probability by construction of the payment vectors used in ChainedFair. Lemma 4.3 implies the top chain and therefore the top arm are contained in $\hat{X}$, thus, the claim holds. □

Proof of Lemma 4.7. With probability $\frac{1}{|\hat{X}|} \geq \frac{1}{k}$, arm i^t selected by ChainedFair for payment $4x \cdot |\hat{X}|$. Since $\max_{j \in \hat{X}} |\hat{\mu}_j^t - \hat{\mu}_i^t| > 4|\hat{X}| \cdot x$, the myopic agent will prefer the arm with highest mean, and his choice will therefore cause FindChained to be called. The probability that such an arm i^t is not called for $k \log(1/\delta)$ consecutive rounds is at most δ. □