Quantifying the Burden of Exploration and the Unfairness of Free Riding*

Christopher Jung[†] Sampath Kannan[‡] Neil Lutz[§]

Abstract

We consider the multi-armed bandit setting with a twist. Rather than having just one decision maker deciding which arm to pull in each round, we have n different decision makers (agents). In the simple stochastic setting, we show that a "free-riding" agent observing another "self-reliant" agent can achieve just $O(1)$ regret, as opposed to the regret lower bound of $\Omega(\log t)$ when one decision maker is playing in isolation. This result holds whenever the self-reliant agent's strategy satisfies either one of two assumptions: (1) each arm is pulled at least $\gamma \ln t$ times in expectation for a constant γ that we compute, or (2) the self-reliant agent achieves $o(t)$ realized regret with high probability. Both of these assumptions are satisfied by standard zero-regret algorithms. Under the second assumption, we further show that the free rider only needs to observe the number of times each arm is pulled by the self-reliant agent, and not the rewards realized.

In the linear contextual setting, each arm has a distribution over parameter vectors, each agent has a context vector, and the reward realized when an agent pulls an arm is the inner product of that agent's context vector with a parameter vector sampled from the pulled arm's distribution. We show that the free rider can achieve $O(1)$ regret in this setting whenever the free rider's context is a small (in L_2-norm) linear combination of other agents' contexts and all other agents pull each arm $\Omega(\log t)$ times with high probability. Again, this condition on the self-reliant players is satisfied by standard zero-regret algorithms like UCB. We also prove a number of lower bounds.

1 Introduction

We consider situations where exploitation must be balanced with exploration in order to obtain optimal performance. Typically there is a single decision maker who does this balancing, in order to minimize a quantity called the regret. In this paper we consider settings where there are many agents and ask how a single agent (the *free rider*) can benefit from the exploration of other *self-reliant* agents. For example, competing pharmaceutical companies might be engaged in research for drug discovery. If one of these companies had access to the research findings of its competitors, it might greatly reduce its own exploration cost. Of course, this is an unlikely scenario since intellectual property is jealously guarded by companies, which points to an important consideration in modeling such scenarios: the amount and type of information that one agent is able to gather about the findings of others.

More realistically, and less consequentially, a recommendation system such as Yelp™ makes user ratings of restaurants publicly available. The assumption underlying such systems is that "the crowd" will explore available options so that we end up with accurate average ratings. Many problems of this sort can be modeled using the formalism of multi-armed bandits. Free riding also arises in online advertising. Each advertiser may be modeled as an agent with a context vector describing its likely customers, and it must choose online niches in which to advertise. A free rider can take advantage of competitors' exploration of niches by monitoring impressions and clickthroughs of their ads. In fact, there are a number of paid services (WhatRunsWhere, Adbeat, SpyFu, etc.) that facilitate this behavior.

Multi-armed bandit problems model decision making under uncertainty [15, 12, 5]. Our focus in this paper will be on the *stochastic bandits model* where there is an unknown reward distribution associated with each arm, and the decision maker has to decide which arm to pull in each round. Her goal is to minimize *regret*, the (expected) difference between the reward of the best arm and the total reward she obtains. In the extension to the *linear contextual bandits model*, each arm i has an unknown parameter vector $\theta_i \in \mathbb{R}^d$ for $i = 1, \ldots, k$, where k is the total number of arms. At round t, a *context* $x^t \in \mathbb{R}^d$ arrives. The expected reward for pulling arm i in round t is the inner product $\langle \theta_i, x^t \rangle$.

In the simple stochastic case, there are two types of relevant information: the other agents' actions and the resulting rewards. In the full-information setting, the free rider has access to both types of information. We also consider a partial-information setting where the free rider can only observe the other agents' actions. For

*Part of this research were done during the *Algorithmic Fairness* and *Fairness* Summer Clusters at the Simons Institute for the Theory of Computing, Berkeley.

†University of Pennsylvania, chrjung@seas.upenn.edu. Research supported in part by NSF grant CCF-1763307.

‡University of Pennsylvania, kannan@cis.upenn.edu. Research supported in part by NSF grant CCF-1763307.

§Iowa State University, nlutz@iastate.edu. This work was done while the author was at the University of Pennsylvania.

linear contextual bandits, the full-information setting also includes the context vectors of the other agents.

In our setting, using Yelp^{TM} as the running example, the k arms correspond to restaurants. Our model differs from standard bandit models in three significant ways. First, there are n decision makers rather than one; in the Yelp^{TM} example, each decision maker corresponds to a diner. Upon visiting a restaurant, a diner samples from a distribution to determine her dining experience. In the stochastic setting, we assume that all diners have identical criteria for assessing their experiences, meaning that identical samples lead to identical rewards. Second, in the linear contextual setting, the contexts in our model are fixed in time and can be regarded as the *types* of the individual decision makers. Each diner's context vector represents the weight she assigns to various features (parameters) of a restaurant, such as innovativeness, decor, noise level, suitability for vegetarians, etc. Third, each arm has a distribution over parameter vectors instead of a fixed parameter vector. When a diner visits a restaurant, her reward is determined by taking the inner product of her context with a parameter vector drawn from the restaurant's distribution, rather than by adding sub-Gaussian noise to the inner product with a fixed parameter vector as in the standard model.

In the standard stochastic or linear contextual bandit setting, a decision-making algorithm is called *zero-regret* if its regret over t rounds is $o(t)$. It is well known that exploration is essential for achieving zero regret [12]. One algorithm that achieves the asymptotically optimal regret bound of $O(\log t)$ over t rounds is the so-called Upper Confidence Bound (UCB) algorithm of Lai and Robbins [12]. In addition to maintaining a sample mean for each arm, this algorithm maintains confidence intervals around these means, where the width of the confidence interval for arm i drops roughly as $1/\sqrt{n_i}$ where n_i is the number of times arm i has been pulled so far. The UCB algorithm then selects the arm with the highest upper limit to its confidence interval. There are many other zero-regret strategies, such as Thompson sampling [16] or one where an initial round-robin exploration phase is followed by an exploitation phase in which the apparently optimal arm is pulled [8].

Our results:

- In the stochastic setting a free rider can achieve $O(1)$ regret under either of two reasonable assumptions, both of which are satisfied by standard zero-regret algorithms:
  - Some self-reliant agent has pulled each arm at least $\gamma \ln t$ times in expectation at all sufficiently large times t, where γ is a constant derived from our analysis (Theorem 4.1).
  - Some self-reliant agent is playing a strategy that with high probability achieves $o(t)$ realized regret. In this case, the free rider can achieve $O(1)$ regret even in the partial-information setting (Theorem 4.2). As a corollary, a free rider can achieve $O(1)$ regret whenever a self-reliant agent plays UCB (Corollary 4.1).
- For linear contextual bandits, a free rider can again achieve $O(1)$ regret in the full-information setting under an assumption similar to the first assumption above (Theorem 5.1).
- As a way of relating the two assumptions in the first bullet above, we prove that if a self-reliant agent achieves $O(t^{1-\epsilon})$ regret, then that agent must pull *each* arm $\Omega(\log t)$ times in expectation (Theorem 3.1) and with a high probability that depends on ϵ (Theorem 3.2).
- There is a deterministic lower bound of $\Omega(\log t)$ on the number of times a UCB agent must pull each arm in the stochastic case (Theorem 3.3).
- To achieve $o(\log t)$ regret in the contextual setting, the free rider must know both the contexts and the observed rewards of the other agents (Theorems 5.2 and 5.3).

Related work: This paper asks how and when an agent may avoid doing their "fair share" of exploration. Several recent works have studied how the cost of exploration in multi-armed bandit problems is distributed, from the perspective of algorithmic fairness. Works by Bastani, Bayati, and Khosravi [2]; Kannan, Morgenstern, Roth, Waggoner, and Wu [9]; and Raghavan, Slivkins, Vaughan, and Wu [14] show that if the data is sufficiently diverse, e.g., if the contexts are randomly perturbed, then exploration may not be necessary. Celis and Salehi [7] consider a model in both the stochastic and the adversarial setting where each agent in the network plays a certain zero-regret algorithm (UCB in the stochastic setting and EXP3 in the adversarial setting) and study how much information an agent can gather from his neighbors.

There is some discussion in the economics literature of free riding in bandit settings. In the model of Bolton and Harris [3], agents choose what fraction of each time unit to devote to a safe action (exploitation) and to a risky action (exploration), and they show that while the attraction of free riding drives agents to select the safe action always, risky action by a agent may enable everyone to converge to the correct posterior belief faster. Keller, Rady, and Cripps [10] consider a very similar setting where a risky arm will generates positive payoff after an exponentially distributed random time; they characterize unique symmetric equilibrium as well as various asymmetric equilibria. Klein [11] gives conditions for complete learning in a two-agent, three-armed bandit setting where there are two negatively-correlated risky arms and a safe arm, with further assumptions about their behavior. It is clear that these models do not support having more than two arms (or three in the case of [11]) and that their goal is maximizing expected reward, not minimizing regret. Moreover, one arm is explicitly designated as the safe arm and the other(s) as risky, *a priori*.

2 Preliminaries

Stochastic Model There are k *arms*, indexed by $[k] = \{1, \dots, k\}$ and n *players* or *agents*, indexed by $[n]$. Arm i has a *reward distribution* D_i supported on $[-1, 1]$ with mean μ_i, and $\mathbf{D} = (D_1, \dots, D_k)$ is the *reward distribution profile*, or the *stochastic bandit*. The arm with the highest mean reward is denoted by

$$i^* = \arg\max_{i \in [k]} \mu_i \,,$$

and we write μ^* for μ_{i^*}; we assume that i^* is unique. An important parameter is

$$\Delta = \mu^* - \max_{i \in [k] \setminus \{i^*\}} \mu_i \,,$$

the *gap* between optimal and suboptimal arms.

In round $t = 1, 2, \dots$, each player p selects an arm $i_p^t \in [k]$ and receives a reward $r_p^t \sim D_{i_p^t}$. We write

$$H^T = ((i_p^t, r_p^t)_{t \in [T]})_{p \in [n]}$$

to denote the *history* of all players' actions and rewards through round T. A *policy* or *strategy* for a player p is a function f_p mapping each history to an arm or to a distribution over the arms; a player p with policy f_p who observes history H^T will pull arm $f_p(H^T)$ in round $T+1$. A *policy profile* is a vector $\mathbf{f} = (f_1, \dots, f_n)$, where each f_p is a policy for player p. Notice that a policy profile and a stochastic bandit together determine a distribution on histories. A policy f_p for player p is *self-reliant* if it depends only on p's own observed actions and rewards. In contrast, a *free-riding* policy may use all players' history.

The *regret* of player p at time T under stochastic bandit $\mathbf{D}$ is

$$R_p^T(\mathbf{D}, \mathbf{f}) = T\mu^* - \sum_{t \in [T]} \mathbb{E}[r_p^t] \,,$$

where the expectation is according to the distribution on histories determined by $\mathbf{D}$ and $\mathbf{f}$.[1] When it will not introduce ambiguity, we simply write R_p^T or, in single-player settings, R^T. We also consider the *realized regret* under a particular history H^T,

$$\hat{R}_p^T(\mathbf{D}, \mathbf{f}) = T\mu^* - \sum_{t \in [T]} r_p^t \,.$$

For any player p, arm i, and time t, the *sample count* is $N_{p,i}^t$, the number of times i has been pulled by the player in the first t rounds, and the *sample mean* is $\mu_{p,i}^t$, the average of all of player p's samples of arm i through time t.

One well-studied self-reliant policy that achieves logarithmic regret in the stochastic setting is called α-*UCB* [12], defined by

$$\alpha\text{-UCB}(H^t) = \arg\max_{i \in [k]} \mu_i^t + \sqrt{\frac{\alpha \ln(t+1)}{2N_i^t}}$$

for all histories H^t. The parameter α calibrates the balance between exploration and exploitation. For each arm i, a player using this policy maintains an *upper confidence bound* on μ_i, and in each round, she pulls the arm with the highest upper confidence bound. The distance from each arm's sample mean to its upper confidence bound depends its sample count.

Linear Contextual Model The linear contextual model generalizes the stochastic model. Now, each arm i has a *feature distribution* F_i supported on the d-dimensional closed unit ball, for some $d \in \mathbb{N}$, and $\mathbf{F} = (F_1, \dots, F_k)$ is the *feature distribution profile* or *contextual bandit*. Each player p has a context $x_p \in \mathbb{R}^d$, and $\mathbf{x} = (x_1, \dots, x_n)$ is the *context profile*. As before, in each round t, each player p selects an arm i_p^t, but now the reward is given by sampling a feature vector $\theta_p^t \sim F_{i_p^t}$, and taking its inner product with x_p, i.e., $r_p^t = \langle \theta_p^t, x_p \rangle$. $D_{p,i}$ is the distribution of rewards from arm i for player p, and the mean of this distribution is

$$\mu_{p,i} = \mathop{\mathbb{E}}_{\theta_i \sim F_i}[\langle \theta_i, x_p \rangle] \,.$$

The optimal arm for player p is

$$i_p^* = \arg\max_{i \in k} \mu_{p,i} \,.$$

Histories, policies, policy profiles, self-reliance, and free riding are defined exactly as in the stochastic setting. The regret of player p through round $T \in \mathbb{N}$

[1] Some sources refer to this quantity as *pseudo-regret* and use *regret* to refer to the realized regret.

under contextual bandit $\mathbf{F}$, context profile $\mathbf{x}$, and policy profile $\mathbf{f}$ is given by

$$R_p^T(\mathbf{F}, \mathbf{x}, \mathbf{f}) = T\mu_p^* - \sum_{t\in[T]} \mathbb{E}[r_p^t] \,,$$

where the expectation is taken according to the distribution determined by $\mathbf{F}$, $\mathbf{x}$, and $\mathbf{f}$. Notice that for a self-reliant player p with context x_p, the contextual bandit $\mathbf{F} = (F_1, \dots, F_k)$ is equivalent to the stochastic bandit $\mathbf{D} = (D_{p,1}, \dots, D_{p,k})$.

3 Lower Bounds on Sample Counts

If a self-reliant player has sampled an arm sufficiently many times, then a free rider with full information can use those samples to find a good estimate of that arm's mean. In this section, we give three lower bounds on the sample counts of each arm. All missing proofs can be found in the appendix.

Theorem 3.1 shows that if a policy guarantees $O(T^{1-\epsilon})$ regret for some positive ϵ, then every arm must be sampled $\Omega(\log T)$ times in expectation. We prove this using the method of Bubeck, Perchet, and Rigollet [6], showing via the Bretagnolle-Huber inequality [4] that the learner cannot rule out any arm's optimality without sampling that arm $\Omega(\log T)$ times.

Theorem 3.1. *Let f be any self-reliant policy such that $R^T = O(T^{1-\epsilon})$ for all stochastic bandits and some $\epsilon > 0$. Then for all stochastic bandits with $\mu^* < 1$ and all $i \in [k]$, f satisfies $\mathbb{E}[N_i^T] = \Omega(\log T)$.*

In addition to a bound on the expected sample count, we sometimes need stronger guarantees on the tail of the sample count distribution. In Theorem 3.2, we use a coupling argument to show that if a policy has regret $O(T^{1-\beta})$ for relatively large β, then the probability that any arm that is sampled too few times is small.

Theorem 3.2. *Let f be any self-reliant policy such that $R^T = O(T^{1-\beta})$ for all stochastic bandits and some $\beta > 0$. Then for all stochastic bandits with $\mu^* < 1$, all $i \in [k]$, and all $\gamma > 0$, f satisfies*

$$\Pr\left(N_i^T \le \gamma \ln T\right) = O\left(T^{\gamma c_i - \beta}\right),$$

where $c_i = \ln\left(\frac{1-\mu^*}{2(1-\mu_i)}\right)$.

Proof. Let $\alpha, \beta, \gamma, t_0 > 0$, and let f be a self-reliant policy such that for all stochastic bandits $\mathbf{D}$ and all $T > t_0$, $R^T(\mathbf{D}, f) \le \alpha T^{1-\beta}$. Let $t_1 \ge t_0$ satisfy $\gamma \ln t_1 < t_1/2$. Assume for contradiction that there is some stochastic bandit $\mathbf{D}$ with $\mu^* < 1$, some arm $i \in [k]$, and some $T > t_1$ such that

$$\Pr\left(N_i^T(\mathbf{D}, f) \le \gamma \ln T\right) > \frac{C\alpha}{T^{\beta+\gamma \ln p_i}} \,, \tag{3.1}$$

where $C = \frac{2}{\min\{\Delta, (1-\mu^*)/2\}}$ and $p_i = \frac{1-\mu^*}{2(1-\mu_i)}$.

Observe that if i is the optimal arm in $\mathbf{D}$, then since $\gamma \ln p_i < 0$, we have

$$\begin{aligned} R^T(\mathbf{D}, f) &> \Delta \cdot (T - \gamma \ln T) \cdot \frac{C\alpha}{T^{\beta+\gamma \ln p_i}} \\ &\ge \Delta \cdot (T - \gamma \ln T) \cdot C\alpha T^{-\beta} \\ &\ge \alpha T^{1-\beta} \,, \end{aligned}$$

contradicting the assumption that $R^T(\mathbf{D}, f) \le \alpha T^{1-\beta}$. Hence, we assume that i is suboptimal.

We now construct a stochastic bandit $\mathbf{D}'$ in which i is optimal. Let $\mathbf{D}' = (D_1, \dots, D_i', \dots, D_k)$, where

$$D_i'(x) = p_i \cdot D_i(x) + 1 - p_i$$

for all $x \in [-1, 1]$. Notice that the mean of D_i' is $\mu_i' = \frac{1+\mu^*}{2}$, and that the gap between optimal and suboptimal arms in $\mathbf{D}'$ is $\Delta' = \mu_i' - \mu^* = \frac{1-\mu^*}{2}$.

We now use a coupling argument to bound $\Pr\left(N_i^T(\mathbf{D}', f) \le \gamma \ln T\right)$. Observe that to sample from D_i', one can sample a reward $x \sim D_i$, keep x with probability p_i, and otherwise output 1. Thus, for any history h in which i is pulled exactly s times,

$$\Pr\left(H^T(\mathbf{D}', f) = h\right) \ge p_i^s \cdot \Pr\left(H^T(\mathbf{D}, f) = h\right) .$$

By summing over all such histories, we have

$$\Pr\left(N_i^T(\mathbf{D}', f) = s\right) \ge p_i^s \cdot \Pr\left(N_i^T(\mathbf{D}, f) = s\right) ,$$

and therefore

$$\begin{aligned} &\Pr\left(N_i^T(\mathbf{D}', f) \le \gamma \ln T\right) \\ &= \sum_{s=0}^{\lfloor \gamma \ln T \rfloor} \Pr\left(N_i^T(\mathbf{D}', f) = s\right) \\ &\ge \sum_{s=0}^{\lfloor \gamma \ln T \rfloor} p_i^s \cdot \Pr\left(N_i^T(\mathbf{D}, f) = s\right) \\ &\ge p_i^{\gamma \ln T} \sum_{s=0}^{\lfloor \gamma \ln T \rfloor} \Pr\left(N_i^T(\mathbf{D}, f) = s\right) \\ &= T^{\gamma \ln p_i} \cdot \Pr\left(N_i^T(\mathbf{D}, f) \le \gamma \ln T\right) . \end{aligned}$$

Combining this bound with inequality (3.1) yields

$$\begin{aligned} \Pr\left(N_i^T(\mathbf{D}', f) \le \gamma \ln(T)\right) &> T^{\gamma \ln p_i} \cdot \frac{C\alpha}{T^{\beta+\gamma \ln p_i}} \\ &= C\alpha T^{-\beta} \,. \end{aligned}$$

Thus,

$$R^T(\mathbf{D}', f) > C\alpha T^{-\beta} \cdot (T - \gamma \ln T) \cdot \frac{1-\mu^*}{2} \geq \alpha T^{1-\beta},$$

which again contradicts the assumed regret bound on f. ∎

Finally, we use a delicate inductive argument to prove the following deterministic guarantee on the sample count for each arm when the arms are pulled according to the α-UCB policy.

THEOREM 3.3. *Let $\alpha > 0$ and $\eta > 2$. There exists a constant t_0 such that for all stochastic bandits, all $t \geq t_0$, and all $i \in [k]$, an agent playing the α-UCB policy must satisfy $N_i^{t-1} \geq \alpha \ln t/(2\eta^2 k^2)$.*

Proof. For every $j \in [k]$ and $t \in \mathbb{N}$, define the set

$$U_j^t = \left\{ i \in [k] : N_i^{t-1} \geq \frac{\alpha \ln t}{2\eta^2 j^2} \right\}.$$

We claim that for all $j \in [k]$ there is a constant t_j such that for all $t \geq t_j$, $|U_j^t| \geq j$. We will prove this claim by induction on j.

For any time t, there is clearly some arm i with $N_i^{t-1} \geq \frac{t-1}{k}$, and we can choose t_1 such that $\frac{t-1}{k} \geq \frac{\alpha \ln t}{2\eta^2 k^2}$ whenever $t \geq t_1$, so the claim holds for $j = 1$.

Now fix $j > 1$, and assume that the claim holds for $j - 1$. Define a function $g_j : \mathbb{N} \to \mathbb{R}$ by

$$g_j(t) = t - (k - j + 1)\frac{\alpha \ln t}{2\eta^2 j^2}.$$

We choose t_j sufficiently large such that for all $t \geq t_j$ we have $g_j(t) > t_{j-1}$ and

$$\text{(3.2)} \qquad \frac{\ln(g_j(t) - 1)}{\ln t} > \left(1 - \frac{1 - 2/\eta}{j}\right)^2.$$

Assume for contradiction that there is some time $t \geq t_j$ such that $|U_j^t| < j$. Since $U_{j-1}^t \subseteq U_j^t$, the inductive hypothesis then implies that $U_j^t = U_{j-1}^t$. Thus, $|U_j^t| = j - 1$, and there are exactly $k - j + 1$ arms outside of U_j^t. Each one of those arms has been pulled at most $\frac{\alpha \ln t}{2\eta^2 j^2}$ times by round $t - 1$, so by the pigeonhole principle there is some $s \in [g_j(t) - 1, t - 1]$ such that an arm from U_j^t is pulled in round s.

Furthermore, inequality (3.2) implies

$$\frac{\alpha \ln s}{2\eta^2 (j-1)^2} > \frac{\alpha \ln t}{2\eta^2 j^2},$$

which guarantees that $U_{j-1}^s \subseteq U_j^t$. Since $s \geq g_j(t_j) - 1 \geq t_{j-1}$, the inductive hypothesis tells us that $|U_{j-1}^s| \geq j - 1$, so we have $U_{j-1}^s = U_j^t$, meaning that the arm pulled in round s is also in U_{j-1}^s.

Now, $N_i^{s-1} \geq \frac{\alpha \ln s}{2\eta^2 (j-1)^2}$ for all $i \in U_{j-1}^s$, so the upper confidence bound of the arm pulled at time s is at most

$$1 + \sqrt{\frac{\alpha \ln s}{2\alpha \ln s/(2\eta^2(j-1)^2)}} = 1 + \eta \cdot (j-1).$$

The upper confidence bound at time s of any arm in $[k] \setminus U_j^t$ is at least

$$-1 + \sqrt{\frac{\alpha \ln s}{2\alpha \ln t/(2\eta^2 j^2)}} = -1 + \eta j \sqrt{\frac{\ln s}{\ln t}}.$$

But since $t \geq t_j$ and $s \geq g_j(t_j)$, inequality (3.2) implies

$$-1 + \eta j \sqrt{\frac{\ln s}{\ln t}} > 1 + \eta \cdot (j-1).$$

This means that all arms outside of U_j^t have higher upper confidence bounds at time s than the arms in U_j^t, contradicting the choice to pull an arm in U_j^t at time s.

By induction, we conclude that the claim holds for all $j \in [k]$, and in particular that the theorem holds with $t_0 = t_k$. ∎

4 Free Riding with Stochastic Bandits

Full-Information Case for Stochastic Bandits Here we describe how a free rider can take advantage of the samples collected by another player p who pulls every arm sufficiently many times in expectation. This free-riding policy, which we call SAMPLEAUGMENTMEANGREEDY, divides time into *epochs* of doubling length. In the j^{th} epoch, the free rider checks whether a given player p has observed at least γj samples of each arm $i \in [k]$, where γ is an appropriate constant. If all sample counts are sufficient, then the free rider uses p's observed rewards to estimate the mean of each arm, committing to the arm with the maximum estimated mean for the remainder of the epoch. Otherwise, the free rider pulls any under-sampled arms, augmenting all sample counts up to at least γj before proceeding. Doing this allows the free rider to circumvent the logarithmic lower bound on regret and achieve $O(1)$ regret. A more detailed description of this policy can be found in Appendix A.2.

THEOREM 4.1. *Fix a stochastic bandit, and suppose some player p plays a self-reliant policy that satisfies $\mathbb{E}[N_i^{t-1}] \geq \gamma \ln t$ for some $\gamma > 2/\Delta^2$, all $i \in [k]$, and all sufficiently large t. Then a free rider can achieve $O(1)$ regret.*

Table 1: Necessary and sufficient information about self-reliant agent(s) p for a free rider seeking to achieve constant regret, where γ is an appropriate constant. In this table "context," "actions," and "rewards" refer to agent p's context x_p, their selected arm i_p^t at each time t, and their observed reward r_p^t at each time t, respectively.

Guarantee on policy of other agent(s)	Stochastic setting	Contextual setting
$\geq \gamma \ln t$ samples of each arm in expectation	actions and rewards	—
$\geq \gamma \ln t$ samples of each arm with high probability	actions and rewards	context, actions, and rewards
$O(t^{1-\epsilon})$ realized regret with high probability	actions	—

Partial-Information Case for Stochastic Bandits We now show that a free rider can achieve constant regret by observing a player who plays any policy that is unlikely to pull suboptimal arms too often. This class of policies includes UCB. We consider a specific, natural free-riding policy $\textsc{CountGreedy}_p$, defined by

$$\textsc{CountGreedy}_p(H^t) = \arg\max_{i \in [k]} N_{p,i}^t ,$$

which always pulls whichever arm i has been pulled most frequently by player p. Notice that this policy does not require the free rider to observe player p's rewards.

One might suspect that it would be sufficient for player p's policy to achieve $R_p^T = o(T)$ in order for the free rider to achieve constant regret under $\textsc{CountGreedy}$, but this turns out not to be the case. It is possible for p to achieve logarithmic regret despite frequently pulling suboptimal arms with non-trivial probability, preventing $\textsc{CountGreedy}$ from achieving constant regret.

Lemma 4.1. *There is a self-reliant policy for player p with $R_p^T(T) = O(\log T)$ such that, if player 1 plays $\textsc{CountGreedy}_p$, then $R_1^T(T) = \Omega(T)$.*

Proof. Consider the policy that, for $j = 0, 1, 2 \ldots$, dictates the following behavior in epochs of tripling length. With probability $1/3^j$, player p "gives up" on rounds 3^j to $3^{j+1} - 1$, choosing arm

$$i_p^t = \arg\min_i \mu_i^{3^j - 1} .$$

Otherwise, with probability $1 - 1/3^j$, player p plays α-UCB during those rounds. This policy is self-reliant, and player p's regret grows at most logarithmically. Notice that whenever player p gives up, $i_p^{3^j}$ will become her most frequently pulled arm by round $2 \cdot 3^j$, so the $\textsc{CountGreedy}_p$-playing free rider will pull this arm at least 3^j times before round 3^{j+1}. It is routine to show that $i_p^{3^j}$ is suboptimal with probability $1 - O(1/3^j)$, so the free rider's regret through round T is at least $R_1^T \geq \sum_{j=0}^{\lfloor \log_3 T \rfloor} (1 - O(1/3^j)) \cdot \Delta \cdot 3^j = \Omega(T)$. □

Notice that by Theorem 3.3, the above policy satisfies the conditions of Theorem 4.1 when α is sufficiently large, and therefore a free rider playing $\textsc{SampleAugmentMeanGreedy}_p$ would achieve constant regret in this situation. Intuitively, this is because the policy of sometimes giving up on entire epochs is not "rational," and $\textsc{SampleAugmentMeanGreedy}$, unlike $\textsc{CountGreedy}$, does not make any implicit assumption of rationality for the self-reliant player.

Since logarithmic regret for player p is not a strong enough assumption, we instead show that if the realized regret $\hat{R}_p^T$ is sublinear with sufficiently high probability, then the free rider achieves constant regret by playing $\textsc{CountGreedy}_p$.

Theorem 4.2. *Fix a stochastic bandit and assume there is some player p such that for all $\epsilon > 0$ there exists a $w > 1$ satisfying $\Pr(\hat{R}_p^T \geq \epsilon T) = O(T^{-w})$. Then a free rider playing $\textsc{CountGreedy}_p$ achieves $O(1)$ regret.*

Audibert, Munos, and Szepesvári [1] showed that α-UCB satisfies the probability bound of Theorem 4.2 in the single-player setting whenever $\alpha > 1/2$. Since α-UCB is a self-reliant policy, this immediately yields the following corollary.

Corollary 4.1. *If some player p's policy is α-UCB for any $\alpha > 1/2$, then a free rider playing $\textsc{CountGreedy}_p$ achieves $O(1)$ regret.*

5 Free Riding with Contextual Bandits

Theorems 4.1 and 4.2 show that free riding is easy in the stochastic case, in which the reward distribution is identical for all players, but the task is more nuanced when players may have diverse contexts. In the linear contextual setting, different players may have different optimal arms, so a simple free-riding strategy like $\textsc{CountGreedy}$ may fail, even when there are strong regret guarantees for the other players. In fact, as we show in Theorems 5.2 and 5.3, successful free riding in this setting requires knowledge of both the contexts and the rewards of other players.

Full-Information Case for Contextual Bandits We now consider the full-information setting where the free rider knows other players' contexts, actions, and rewards. We show that if the free rider's context is a linear combination of the other players' contexts — and if other players pull all arms sufficiently many times — then the free rider can aggregate other players' observations to estimate the means of its own reward distribution profile. In the event that some arm has not been sampled enough by some player, the free rider temporarily acts self-reliantly and chooses arms according to UCB. Under the above assumptions, a player 1 with this free-riding policy, UCBMeanGreedy, achieves $O(1)$ regret. Formally, for every history H^t, let $j = \lfloor \log t \rfloor$, let S_j be the event that $N_{p,i}^{2^j-1} \geq \gamma j$ for all $p \in \{2, \ldots, n\}$ and all $i \in [k]$, and let $\hat{\mu}_{p,i}^s$ denote the average of the first s observed samples of $D_{p,i}$. Then for every $\gamma > 0$ and every $\mathbf{c} = (c_2, \ldots, c_n) \in \mathbb{R}^{n-1}$, we define $\text{UCBMeanGreedy}_{\gamma,\mathbf{c}}(H^t)$ as

$$\arg\max_i \sum_{p=2}^{n} c_p \hat{\mu}_{p,i}^{\lceil \gamma j \rceil}$$

if S_j occurs and

$$\text{2-UCB}\left(\left(i_1^{2^j}, r_1^{2^j}\right), \ldots, \left(i_1^t, r_1^t\right)\right)$$

otherwise.

We will assume that the free rider's context is a linear combination of the others players' contexts. The vector $\mathbf{c}$ will consist of the coefficients of this linear combination. If these coefficients are small, then no individual player's sampling noise can affect the free rider's choices too much.

Notice that when applying UCB this policy treats the bandit as stochastic and "starts from scratch" in each epoch, only considering the free rider's own actions and observed rewards from the current epoch.

Theorem 5.1. *Let* $\mathbf{x}$ *be a context profile and* $\mathbf{c} \in \mathbb{R}^{n-1}$ *be a vector* $(c_2, \ldots, c_n)$ *such that* $\sum_{p=2}^{n} c_p x_p = x_1$. *Fix a contextual bandit, let* Δ *be the gap for player 1, and suppose that* $\epsilon > 0$ *and* $\gamma > \frac{8\langle \mathbf{c}, \mathbf{c} \rangle \ln 2}{\Delta^2}$ *satisfy*

$$\Pr(N_{p,i}^{t-1} < \gamma \log t) = O((\log t)^{-2-\epsilon})$$

for every player $p \in \{2, \ldots, n\}$ *and arm* $i \in [k]$. *Then a free rider playing* $\text{UCBMeanGreedy}_{\gamma,\mathbf{c}}$ *achieves* $O(1)$ *regret.*

Proof. Fix $j \in \mathbb{N}$, let *epoch* j be rounds 2^j through $2^{j+1} - 1$, and let $s_j = \lceil \gamma j \rceil$. For each $i \in [k]$ let

$$\tilde{\mu}_i^j = \sum_{p=2}^{n} c_p \hat{\mu}_{p,i}^{s_j},$$

and let $\tilde{i}^j = \arg\max_i \tilde{\mu}_i^j$. We analyze the free rider's regret incurred during epoch j in each of the following cases: (1) S_j and $\tilde{i}^j = i_1^*$, (2) S_j and $\tilde{i}^j \neq i_1^*$, and (3) $\neg S_j$.

Notice that $\tilde{i}^j$ is defined in the first and second cases. In case (1), the free rider incurs no regret during epoch j because it pulls only the optimal arm i_1^* during that epoch. We now analyze the regret incurred from the other two cases.

For case (2),

$$\Pr(\tilde{i}^j \neq i_1^*) \leq \sum_{i \in [k]} \Pr\left(\left|\tilde{\mu}_i^j - \mu_{1,i}\right| \geq \frac{\Delta}{2}\right).$$

For fixed $i \in [k]$, letting $\hat{r}_{p,i}^s$ denote the value of the s^{th} observed sample of $D_{p,i}$, we have

$$\begin{aligned}\left|\tilde{\mu}_i^j - \mu_{1,i}\right| &= \left|\sum_{p=2}^{n} \frac{c_p}{s_j} \sum_{s=1}^{s_j} \hat{r}_{p,i}^s - \sum_{p=2}^{n} c_p \mu_{p,i}\right| \\ &= \left|\sum_{p=2}^{n} \sum_{s=1}^{s_j} X_p^s\right|,\end{aligned}$$

where $X_p^s = \frac{c_p}{s_j}\left(\hat{r}_{p,i}^s - \mu_{p,i}\right)$. Notice that the X_p^s are independent, and each X_p^s is supported on $[-c_p/s_j, c_p/s_j]$ with $\mathbb{E}[X_p^s] = 0$, so Hoeffding's lemma gives

$$\mathbb{E}[\exp(\lambda X_p^s)] \leq \exp\left(\frac{\lambda^2 c_p^2}{2 s_j^2}\right)$$

for all $\lambda \in \mathbb{R}$. Let $\lambda = \frac{4j \ln 2}{\Delta}$, and apply a Chernoff bound:

$$\begin{aligned}&\Pr\left(\left|\sum_{p=2}^{n} \sum_{s=1}^{s_j} X_p^s\right| \geq \frac{\Delta}{2}\right) \\ &\leq 2\exp\left(-\frac{\lambda\Delta}{2}\right) \mathbb{E}\left[\prod_{p=2}^{n} \prod_{s=1}^{s_j} \exp(\lambda X_p^s)\right] \\ &= 2\exp\left(-\frac{\lambda\Delta}{2}\right) \prod_{p=2}^{n} \prod_{s=1}^{s_j} \mathbb{E}[\exp(\lambda X_p^s)] \\ &\leq 2\exp\left(-\frac{\lambda\Delta}{2} + \sum_{p=2}^{n} \sum_{s=1}^{s_j} \frac{\lambda^2 c_p^2}{2 s_j^2}\right) \\ &\leq 2\exp\left(\frac{\lambda}{2}\left(\frac{\lambda}{\gamma j}\langle \mathbf{c}, \mathbf{c} \rangle - \Delta\right)\right) \\ &= 2\exp\left(2j \ln 2\left(\frac{4\langle \mathbf{c}, \mathbf{c} \rangle \ln 2}{\gamma \Delta^2} - 1\right)\right).\end{aligned}$$

Thus, the contribution of this case to the regret during epoch j is at most

$$k \cdot 2^{1+2j\left(\frac{4\langle \mathbf{c},\mathbf{c}\rangle \ln 2}{\gamma \Delta^2}-1\right)} \cdot 2^{j+1} = k \cdot 2^{2+j\cdot\left(\frac{8\langle \mathbf{c},\mathbf{c}\rangle \ln 2}{\gamma \Delta^2}-1\right)}$$
$$= k \cdot 2^{-\Omega(j)} ,$$

since $\gamma > \frac{8\langle \mathbf{c},\mathbf{c}\rangle \ln 2}{\Delta^2}$.

For the case (3), observe that

$$\Pr(\neg S_j) \le \sum_{p=2}^{n}\sum_{i=1}^{k} \Pr(N_{p,i}^{2^j-1} < \gamma j)$$
$$= O(nkj^{-2-\epsilon}) ,$$

by assumption. In this case, UCBMeanGreedy$_{\gamma,\mathbf{c}}$ resorts to playing α-UCB for 2^j steps, incurring $O(j)$ regret. Thus, the third case contributes $O(nkj^{-1-\epsilon})$ to the expected regret of epoch j. Therefore, for any time horizon $T \in \mathbb{N}$, the regret incurred by the free rider is bounded by

$$\sum_{j=0}^{\infty}\left(k \cdot 2^{-\Omega(j)} + O(nkj^{-1-\epsilon})\right) ,$$

which converges to a constant. □

Partial-Information Cases for Contextual Bandits Now we consider a situation where player 1 must choose a free-riding policy without knowledge of the other players' contexts. We show that this restriction can force the free rider to incur logarithmic regret even given knowledge of the other players' policies, actions, and rewards. Intuitively, this is true because a self-reliant player might behave identically in two different environments, making observations of their behavior useless to the free rider. To prove the theorem, we construct a one-dimensional, two-arm example of two such environments, then appeal to the lower bound technique of Bubeck et al. [6] to show that the free rider must incur $\Omega(\log T)$ regret when acting self-reliantly.

Theorem 5.2. *A free rider without knowledge of the other players' contexts may be forced to incur $R^T = \Omega(\log T)$ regret, regardless of what self-reliant policies the other players employ.*

Similarly, a free rider in the contextual setting needs to know the other players' rewards; knowing their contexts, policies, and actions may not be sufficient to successfully free ride, even when all other players have low realized regret with high probability and are guaranteed to pull all arms frequently. This is in contrast to the stochastic case, as Theorem 4.2 demonstrates. We prove this by describing a self-reliant policy, EpochExploreThenCommit, that again proceeds in doubling epochs. At the beginning of the j^{th} epoch, the player samples each arm $\Theta(j)$ times, then commits to the arm with the highest sample mean for the remainder of the epoch. This policy has strong guarantees on sample count and realized regret, but we construct an example where, with constant probability, the sequence of arm pulls is completely uninformative to the free rider.

Theorem 5.3. *A free rider without knowledge of the other players' rewards may be forced to incur $R^T = \Omega(\log T)$ regret, even when all other players satisfy the conditions of Theorems 4.1 and 4.2.*

6 Conclusion

We have demonstrated that in the linear contextual setting, a free rider can successfully shirk the burden of exploration, achieving constant regret by observing other players engaged in standard learning behavior. Furthermore, we have shown that even with partial information and weaker assumptions on the other players' learning behaviors, the free rider can achieve constant regret in the simple stochastic setting. It would be interesting to examine richer settings. For example, exploring players need not be self-reliant, and both exploring players and free riders could play a range of strategies. As another example, when a free rider in the stochastic setting only sees the actions (and not the rewards) of the self-reliant players and does not know which of them are playing UCB or other zero-regret strategies, can he still achieve constant regret? More realistically, users of a service like Yelp™ cannot be partitioned into self-reliant public learners and selfish free riders who keep their data private. It would be interesting to explore more nuanced player roles and to characterize the equilibria that arise from their interactions. Such a characterization might also suggest mechanisms for the deterrence of free riding or for incentivizing exploration.

References

[1] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[2] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*, 2017.

[3] Patrick Bolton and Christopher Harris. Strategic experimentation. *Econometrica*, 67(2):349–374, 1999.

[4] J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

[5] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed

bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
[6] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT '13), June 12–14, 2013, Princeton University, NJ, USA*, pages 122–134, 2013.
[7] L Elisa Celis and Farnood Salehi. Lean from thy neighbor: Stochastic & adversarial bandits in a network. *arXiv preprint arXiv:1704.04470*, 2017.
[8] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *CoRR*, abs/1602.07182, 2016.
[9] Sampath Kannan, Jamie H. Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2231–2241, 2018.
[10] Godfrey Keller, Sven Rady, and Martin Cripps. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, 2005.
[11] Nicolas Klein. Strategic learning in teams. *Games and Economic Behavior*, 82:636–657, 2013.
[12] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
[13] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms.* Cambridge University Press, 2019. Pre-publication version, revision 8b22b8b6131c37e388d5e3b2eecf0b4ff5d7db92.
[14] Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Proceedings of the 31st Annual Conference On Learning Theory (COLT '18), Stockholm, Sweden, July 6–9 2018.*, pages 1724–1738, 2018.
[15] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
[16] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

A Appendix

A.1 Proof of Theorem 3.1

Proof. Let $k \geq 1$ and $\epsilon > 0$, and let f be any self-reliant policy satisfying $R^T(\mathbf{D}, f) = O(T^{1-\epsilon})$ for all k-arm stochastic bandits $\mathbf{D}$. We prove that for all k-arm stochastic bandits $\mathbf{D}$ with $\mu^* < 1$ and all $i \in [k]$, we have $\mathbb{E}[N_i^T(\mathbf{D}, f)] = \Omega(\ln T)$.

Fix an arm i, and let $X = \mathbb{E}[N_i^T(\mathbf{D}, f)]$. If i is optimal, then $X \geq T - R^T(\mathbf{D}, f)/\Delta$ and the theorem holds. Hence, let i be any suboptimal arm, let $\delta = \min\left\{\Delta, \frac{1-\mu^*}{2}\right\}$, let $p = \frac{1-\mu^*-\delta}{1-\mu_i}$, and let $\mathbf{D}' = (D_1, \ldots, D_i', \ldots, D_k)$, where for all $x \in [-1, 1]$,

$$D_i'(x) = p \cdot D_i(x) + 1 - p\,.$$

Notice that the mean of D_i' is $\mu_i' = \mu^* + \delta$ and that $\mathrm{KL}(D_i, D_i') \leq \ln(1/p)$. Now,

$$\begin{aligned}
&\max\{R^T(\mathbf{D}, f), R^T(\mathbf{D}', f)\} \\
&\geq \frac{1}{2}(R^T(\mathbf{D}, f) + R^T(\mathbf{D}', f)) \\
&\geq \frac{\delta}{2}\sum_{t=1}^{T}(\Pr[i_t(\mathbf{D}, f) = i] + \Pr[i_t(\mathbf{D}', f) \neq i]) \\
&\geq \frac{\delta}{4}\sum_{t=1}^{T}\exp(-\mathrm{KL}(H^{t-1}(\mathbf{D}, f), H^{t-1}(\mathbf{D}', f))) \\
&\geq \frac{\delta T}{4}\exp(-\mathrm{KL}(H^{T-1}(\mathbf{D}, f), H^{T-1}(\mathbf{D}', f)))\,,
\end{aligned}$$

where KL denotes Kullback-Leibler divergence and the second-to-last line follows from Lemma A.2, the Bretagnolle-Huber inequality [4]. By Lemma A.1 [13],

$$\begin{aligned}
&\mathrm{KL}(H^{t-1}(\mathbf{D}, f), H^{t-1}(\mathbf{D}', f) \\
&= \mathrm{KL}(D_i, D_i') \cdot \mathbb{E}[N_i^T(\mathbf{D}, f)] \\
&\leq X\ln(1/p)\,.
\end{aligned}$$

Thus, we have

$$\max\{R^T(\mathbf{D}, f), R^T(\mathbf{D}', f)\} \geq \frac{\delta T}{4}\exp\left(-X\ln(1/p)\right)\,.$$

It follows that $\exp(-X\ln(1/p)) = O(T^{-\epsilon})$ and therefore that $X = \Omega(\ln T)$. ◻

A.2 Proof of Theorem 4.1

Proof. Let $\mathbf{D}$ be a stochastic bandit and let let f be a self-reliant policy. Assume there are constants $\gamma > 2/\Delta^2$ and t_0 such that $\mathbb{E}\left[N_i^{t-1}(\mathbf{D}, f)\right] \geq \gamma \ln t$, for all $i \in [k]$ and all $t \geq t_0$. Let $\mathbf{f} = (f_1, \ldots, f_n)$ be a policy profile with $f_1 = \textsc{SampleAugmentMeanGreedy}_{p,\gamma}$ and $f_p = f$, for some player $p \in \{2, \ldots, n\}$. We prove that $R_1^T(\mathbf{D}, \mathbf{f}) = O(1)$.

Let $t_0' = 2^{\lceil \log t_0 \rceil}$. For all $j \in \mathbb{N}$, define $s_j = \gamma j \ln 2$ and $\tilde{i}_j^* = \arg\max_i \nu_i^j$. Then, letting Δ_i denote $\mu^* - \mu_i$,

$$\begin{aligned}
R_1^T(\mathbf{D}, \mathbf{f}) &= \sum_{i \neq i^*}\mathbb{E}\left[N_{1,i}^T(\mathbf{D}, \mathbf{f})\right]\Delta_i \\
&\leq 2t_0' + \sum_{i \neq i^*}\Delta_i\left(\sum_{t=t_0'}^{T}\Pr[i_1^t = i]\right)\,.
\end{aligned}$$

Now, for each $i \in [k]$,

$$
\begin{aligned}
&\sum_{t=t_0'}^{T} \Pr[i_1^t = i] \\
&< \sum_{j=\lceil \log t_0 \rceil}^{\lceil \log T \rceil} \left(\max\left\{0, s_j - \mathbb{E}\left[N_{p,i}^{2^j-1}(\mathbf{D}, \mathbf{f})\right]\right\} \right. \\
&\left. + \Pr\left[\tilde{i}_j^* = i\right] \cdot 2^j \right) \\
&= \sum_{j=\lceil \log t_0 \rceil}^{\lceil \log T \rceil} \Pr\left[\tilde{i}_j^* = i\right] \cdot 2^j \\
&\le \sum_{j=\lceil \log t_0 \rceil}^{\lceil \log T \rceil} \left(\Pr\left[\nu_i^j - \mu_i > \Delta_i/2\right] \right. \\
&\left. + \Pr\left[\mu^* - \nu_{i^*}^j > \Delta_i/2\right] \right) \cdot 2^j \\
&\le 2 \sum_{j=0}^{\infty} \exp\left(-2\left(\frac{\Delta_i}{2}\right)^2 \gamma j \ln 2\right) \cdot 2^j \\
&= 2 \sum_{j=0}^{\infty} 2^{(1-\Delta^2\gamma/2)j} .
\end{aligned}
$$

Since $\Delta^2\gamma/2 > 1$, this sum converges to a constant, so $R_1^T(\mathbf{D}, \mathbf{f}) = O(1)$. □

Below, $\hat{\mu}_{p,i}^s$ denotes the average of the first s samples of D_i observed by player p.

A.3 Proof of Theorem 4.2

Proof. Let $\mathbf{D}$ be a stochastic bandit, let $p \in \{2, \ldots, n\}$ be a player, and let $\mathbf{f} = (f_1, \ldots, f_n)$ be a policy profile with $f_1 = \text{CountGreedy}_p$. Assume that for all $\epsilon > 0$ there is some $w > 1$ satisfying $\Pr(\hat{R}_p^T(\mathbf{D}, \mathbf{f}) \ge \epsilon T) = O(T^{-w})$. We prove that $R_1^T(\mathbf{D}, \mathbf{f}) = O(1)$.

The free rider pulls a suboptimal arm at each time t if and only if $\text{CountGreedy}_p(H^{t-1}) \ne i^*$, which implies that $N_{p,i^*}^{t-1} \le \frac{t-1}{2}$ and therefore $\hat{R}_p^T(\mathbf{D}, \mathbf{f}) \ge \Delta\frac{t-1}{2}$. Hence, we can bound the free rider's regret by

$$R_1^T(\mathbf{D}, \mathbf{f}) \le 2 \sum_{t=0}^{T-1} \Pr(\hat{R}_p^t(\mathbf{D}, \mathbf{f}) \ge \Delta t/2) .$$

If $w > 1$ satisfies $\Pr(\hat{R}_p^t(\mathbf{D}, \mathbf{f}) \ge \Delta t/2) = O(t^{-w})$, then we have

$$R_1^T(\mathbf{D}, \mathbf{f}) \le 2 \sum_{t=0}^{T-1} O(t^{-w}) = O(1) .$$

□

for $j \in \mathbb{N}$ **do**
  $t = 2^j$
  for $i \in [k]$ **do**
    $N = N_{p,i}^{2^j-1}$
    if $N \ge \gamma j$ **then**
      // p sampled arm i enough prior to this epoch
      $\nu_i^j = \hat{\mu}_{p,i}^{s_j}$
    end
    else
      // the free rider samples arm i up to $\lceil \gamma j \rceil - N$ times
      $\nu_i^j = \hat{\mu}_{p,i}^N \cdot N/(\lceil \gamma j \rceil)$
      while $N < s_j$ *and* $t < 2^{j+1} - 1$ **do**
        $i_1^t = i$
        $\nu_i^j = \nu_i^j + r_1^t/s$
        $N = N + 1$
        $t = t + 1$
      end
    end
  end
  while $t < 2^{j+1} - 1$ **do**
    $i_1^t = \arg\max_i \nu_i^j$
    $t = t + 1$
  end
end

Algorithm 1: SampleAugmentMeanGreedy$_{p,\gamma}$

A.4 Proof of Theorem 5.2

Proof. We prove that there exist a pair of contextual bandits $\mathbf{F}$ and $\mathbf{F}'$ and a pair of two-player context profiles $\mathbf{x}$ and $\mathbf{x}'$ such that, for every time horizon T and every policy profile $\mathbf{f} = (f_1, f_2)$ in which f_1 is independent of player 2's context and f_2 is self-reliant,

(A.1)
$$\max\{R_1^T(\mathbf{F}, \mathbf{x}, \mathbf{f}), R_1^T(\mathbf{F}', \mathbf{x}', \mathbf{f})\} \ge \frac{\ln(T/12) + 1}{2} .$$

We construct a one-dimensional, two-arm, two-player example. Let F_1 be a point mass at 0; let F_2 and F_2' be discrete random variables that take value 1 with probability 1/3 and 2/3, respectively, and value -1 otherwise; and let $\mathbf{F} = (F_1, F_2)$ and $\mathbf{F}' = (F_1, F_2')$. Let $\mathbf{f} = (f_1, f_2)$ be any linear contextual bandit policy profile such that f_2 is self-reliant, and consider a free-riding player 1.

Let $\mathbf{x} = (1, 1)$ and $\mathbf{x}' = (1, -1)$. For $p, i \in [2]$, let $D_{p,i}$ be the reward distribution of arm i for player p under contextual bandit $\mathbf{F}$ and context profile $\mathbf{x}$. Similarly, let $D_{p,i}'$ be the reward distribution of arm i for player p under parameter distribution profile $\mathbf{F}'$

and context profile $\mathbf{x}'$. Observe that $D_{1,1} = D'_{1,1}$, $D_{2,1} = D'_{2,1}$, and $D_{2,2} = D'_{2,2}$, but $D_{1,2} = -D'_{1,2}$.

Informally, the environment $(\mathbf{F}, \mathbf{x})$ is indistinguishable from $(\mathbf{F}', \mathbf{x}')$ from the perspective of player 2. Observing player 2's actions and rewards will therefore be completely uninformative for player 1, who is ignorant of player 2's context. Thus, player 1's task is essentially equivalent to a single-player stochastic bandit problem where the learner must distinguish between reward distribution profiles $(D_{1,1}, D_{1,2})$ and $(D'_{1,1}, D'_{1,2})$. Bubeck et al. [6] showed that the latter task requires the learner to experience logarithmic regret. Adapting their proof to the present situation, we can demonstrate that (A.1) holds. Our situation is almost identical to theirs, except for the presence of an uninformative second player, which requires only minor changes to their proof. We include the details here for the sake of completeness:

Let

$$Q_T = R_1^T(\mathbf{F}, \mathbf{x}, \mathbf{f}) \quad \text{and} \quad Q'_T = R_1^T(\mathbf{F}', \mathbf{x}', \mathbf{f}) .$$

For all $t \in \mathbb{N}$, let $\mathcal{G}_t$ and $\mathcal{G}'_t$ be distributions on $\{(1,0),(2,-1),(2,1)\}^{2t}$ such that

$$H^t(\mathbf{F}, \mathbf{x}, \mathbf{f}) \sim \mathcal{G}_t \quad \text{and} \quad H^t(\mathbf{F}', \mathbf{x}', \mathbf{f}) \sim \mathcal{G}'_t ,$$

and for $i \in \{1, 2\}$, let

$$M_{t,i} = N_{1,i}^t(\mathbf{F}, \mathbf{x}, \mathbf{f}) \quad \text{and} \quad M'_{t,i} = N_i^t(\mathbf{F}', \mathbf{x}', \mathbf{f}) .$$

Observe that

$$\max\{Q_T, Q'_T\} \geq \frac{\mathbb{E}[M_{T,2}]}{3} ,$$

and

$$\begin{aligned}
\max\{Q_T, Q'_T\} &\geq \frac{1}{2}(Q_T + Q'_T) \\
&= \frac{1}{6}\sum_{t=1}^{T}\left(\Pr_{\mathcal{G}_{t-1}}[i_1^t = 1] + \Pr_{\mathcal{G}'_{t-1}}[i_1^t = 2]\right) \\
&\geq \frac{1}{12}\sum_{t=1}^{T}\exp(-\mathrm{KL}(\mathcal{G}_{t-1}, \mathcal{G}'_{t-1})) \\
&\geq \frac{T}{12}\exp(-\mathrm{KL}(\mathcal{G}_T, \mathcal{G}'_T)) ,
\end{aligned}$$

where the second-to-last line follows from the Bretagnolle-Huber inequality (Lemma A.2) [4].

We now calculate $\mathrm{KL}(\mathcal{G}_T, \mathcal{G}'_T)$. For each $t \in \mathbb{N}$ and $p \in \{1, 2\}$, let $\gamma_{p,t}$ and $\gamma'_{p,t}$ be the distributions on (i_p^t, r_p^t) under $(\mathbf{F}, \mathbf{x}, \mathbf{f})$ and $(\mathbf{F}', \mathbf{x}', \mathbf{f})$, respectively. Observe that $\gamma_{2,t} = \gamma'_{2,t}$ for all $t \in \mathbb{N}$. By the chain rule for KL divergence, taking probabilities over $H^{t-1} \sim \mathcal{G}_{t-1}$,

$$\begin{aligned}
&\mathrm{KL}(\mathcal{G}_T, \mathcal{G}'_T) \\
&= \sum_{t=1}^{T}\mathbb{E}[\mathrm{KL}((\gamma_{t,1}, \gamma_{t,2} \mid H^{t-1}), (\gamma'_{t,1}, \gamma'_{t,2} \mid H^{t-1}))] \\
&= \sum_{t=1}^{T}\mathbb{E}[\mathrm{KL}((\gamma_{t,1} \mid H^{t-1}), (\gamma'_{t,1} \mid H^{t-1}))] \\
&= \sum_{t=1}^{T}\Pr[f_1(H^{t-1}) = 2] \cdot \mathrm{KL}(D_2, D'_2) \\
&= \mathrm{KL}(D_2, D'_2)\,\mathbb{E}[M_{T,2}] \\
&= \mathbb{E}[M_{T,2}]/3 .
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\max\{Q_T, Q'_T\} &\geq \frac{1}{2}\left(\frac{\mathbb{E}[M_{T,2}]}{3} + \frac{T}{12}\exp(-\mathbb{E}[M_{T,2}]/3)\right) \\
&\geq \frac{1}{6}\min_{x\in[0,T]}\left(x + \frac{Te^{-x/3}}{4}\right) \\
&= \frac{\ln(T/12) + 1}{2} .
\end{aligned}$$

∎

A.5 Proof of Theorem 5.3

Proof. Consider the following policy that essentially plays the standard explore-then-commit strategy with a doubling time horizon.

```
t = 1
for j ∈ ℕ do
    for i ∈ [k] do
        N_i = 0
        μ̂_i^j = 0
        while N_i < γ(j + 2) and t < 2^{j+1} − 1 do
            i^t = i
            μ̂_i^j = μ̂_i^j + r^t / γ(j+1)
            t = t + 1
        end
    end
    î* = arg max_{i∈[k]} μ̂_i^j
    while t < 2^{j+1} − 1 do
        i^t = î*
        t = t + 1
    end
end
```

Algorithm 2: EpochExploreThenCommit$_\gamma$

Fix $\gamma \geq 2/\Delta^2$, and let

$$g_\gamma = \textsc{EpochExploreThenCommit}_\gamma .$$

We prove that g_γ satisfies the following three properties:

1. For all contextual bandits $\mathbf{F}$, context profiles $\mathbf{x}$, and policy profiles $\mathbf{f} = (f_1, \ldots, f_n)$ with $f_p = g_\gamma$, there is some $t_0 \in \mathbb{N}$ such that $N_{p,i}^T(\mathbf{F}, \mathbf{x}, \mathbf{f}) \geq \gamma \log T$ for all $i \in [k]$ and all $T > t_0$.

2. For all contextual bandits $\mathbf{F}$, context profiles $\mathbf{x}$, policy profiles $\mathbf{f} = (f_1, \ldots, f_n)$ with $f_p = g_\gamma$, and $\epsilon > 0$, there is some $w > 1$ such that $\Pr\left(\hat{R}_p^T(\mathbf{F}, \mathbf{x}, \mathbf{f}) \geq \epsilon T\right) = O(T^{-w})$.

3. There exist a pair of contextual bandits $\mathbf{F}$ and $\mathbf{F}'$ and a context profile $\mathbf{x}$ such that for all policy profiles $\mathbf{f} = (f_1, g_\gamma, g_\gamma)$ such that f_1 is independent of the other players' observed rewards, $\max\left\{R_1^T(\mathbf{F}, \mathbf{x}, \mathbf{f}), R_1^T(\mathbf{F}', \mathbf{x}, \mathbf{f})\right\} = \Omega(\log T)$.

First, let $j_0 \in \mathbb{N}$ satisfy $2^{j_0} \geq k\gamma(j_0 + 2)$, and let $i \in [k]$ be any arm. Then, for all $j \geq j_0$, f satisfies $N_{p,i}^{2^j}(\mathbf{F}, \mathbf{x}, \mathbf{f}) \geq \gamma(j + 1)$, i.e., each arm has been pulled at least $\gamma(j + 1)$ times at the beginning of the epoch. So for every round t in the j^{th} epoch, we have $N_{p,i}^t(\mathbf{F}, \mathbf{x}, \mathbf{f}) \geq \gamma(j + 1) \geq \gamma \log t$. Hence, $N_{p,i}^T(\mathbf{F}, \mathbf{x}, \mathbf{f}) \geq \gamma \log T$ for all $T > 2^{j_0}$, so g_γ satisfies the first property.

Now, for each $j \in \mathbb{N}$, define $\tilde{R}^j = \hat{R}^{2^{j+1}-1} - \hat{R}^{2^j-1}$, the realized regret incurred during epoch j while playing g_γ. At most $2k\gamma(j+2)$ of this realized regret can come from the exploration phase; any further regret in that epoch can only result from committing to a suboptimal arm. By Hoeffding's inequality,

$$\begin{aligned}
&\Pr\left(\tilde{R}^j > 2k\gamma(j + 2)\right) \\
&\leq \sum_{i\in[k]} \exp\left(-\frac{\gamma(j + 2)(\mu^* - \mu_i)^2}{2}\right) \\
&\leq k \cdot \exp\left(-\frac{\gamma(j + 2)\Delta^2}{2}\right) .
\end{aligned}$$

Notice that

$$\sum_{j=0}^{\lfloor\log(\epsilon T)\rfloor-2} \tilde{R}^j \leq \sum_{t=1}^{\epsilon T/4} 2 = \frac{\epsilon T}{2} ,$$

and that $\lceil \log T\rceil - (\lfloor\log(\epsilon T)\rfloor - 2) \leq 4 - \log \epsilon$, so

$$\begin{aligned}
&\Pr\left(\hat{R}^T \geq \epsilon T\right) \\
&\leq \Pr\left(\sum_{j=\lfloor\log(\epsilon T)\rfloor-1}^{\lceil\log T\rceil} \tilde{R}^j \geq \frac{\epsilon T}{2}\right) \\
&\leq \sum_{j=\lfloor\log(\epsilon T)\rfloor-1}^{\lceil\log T\rceil} \Pr\left(\tilde{R}^j \geq \frac{\epsilon T}{8 - 2\log \epsilon}\right) .
\end{aligned}$$

For all sufficiently large T, $2k\gamma(\lceil\log T\rceil + 2) < \frac{\epsilon T}{8-2\log\epsilon}$, so for $j = \lfloor\log(\epsilon T)\rfloor - 1, \ldots, \lceil\log T\rceil$,

$$\begin{aligned}
&\Pr\left(\tilde{R}^j \geq \frac{\epsilon T}{8 - 2\log \epsilon}\right) \\
&\leq \Pr\left(\tilde{R}^j > 2k\gamma(j + 2)\right) \\
&\leq k \cdot \exp\left(-\frac{\gamma(\lfloor\log(\epsilon T)\rfloor + 1)\Delta^2}{2}\right) .
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr\left(\hat{R}^T \geq \epsilon T\right) \\
&\leq (4 - \log \epsilon) \cdot k \cdot \exp\left(-\frac{\gamma(\lfloor\log(\epsilon T)\rfloor + 1)\Delta^2}{2}\right) \\
&= O\left(T^{-\frac{\gamma\Delta^2}{2}}\right) ,
\end{aligned}$$

meaning that $\frac{\gamma\Delta^2}{2} > 1$. So g_γ satisfies the second property.

Finally, let $\mathbf{x} = ((\sqrt{2}/2, \sqrt{2}/2), (1, 0), (0, 1))$ be a three-player context profile, and let $\mathbf{F}$ be the contextual bandit where for each $i \in [3]$, the feature distribution F_i satisfies $F_i(x_i) = 2/3$ and $F_i(-x_i) = 1/3$, and F_4 is a point mass at $(0, 0)$. Define a second contextual bandit $\mathbf{F}' = (F_1', F_2, F_3, F_4)$, where $F_1'(x_1) = 1/3$ and $F_1'(-x_1) = 2/3$. Let f_1 be some policy that is independent of the other players' observed rewards, and consider the policy profile $\mathbf{f} = (f_1, g_\gamma, g_\gamma)$.

Notice that under $\mathbf{x}$, for both $\mathbf{F}$ and $\mathbf{F}'$, we have $i_2^* = 2$ and $i_3^* = 3$. As we showed above, in each epoch $j \in \mathbb{N}$, the probability that the policy g_γ commits to a suboptimal arm is at most $k\exp(-\gamma(j+2)\Delta^2/2)$. Notice also that neither player can commit to any arm for any epoch j in which $2^j \leq k\gamma(j + 2)$. Let j_1 be the first epoch satisfying $2^{j_1} > k\gamma(j_1 + 2)$. Then, for both $\mathbf{F}$ and $\mathbf{F}'$, the probability that either player 2 or player 3 ever commits to a suboptimal arm in any epoch is at most

$$\begin{aligned}
&\sum_{j=j_1}^{\infty} k \exp\left(\frac{-\gamma(j + 2)\Delta^2}{2}\right) \\
&\leq k \int_{j_1}^{\infty} \exp\left(\frac{-\gamma\Delta^2}{2}(x + 1)\right) dx \\
&< ke^{-j_1} .
\end{aligned}$$

Now, $2^{j_1} > k\gamma(j_1 + 2)$ implies that $j_1 < \ln k$, so the above expression converges to some constant $p < 1$. Thus, with probability at least $1 - p > 0$, neither player 2 nor player 3 ever commits to any suboptimal arm. In this situation, their sequences of arm pulls are completely uninformative to the free rider, meaning that the player 1 will have to act self-reliantly to solve

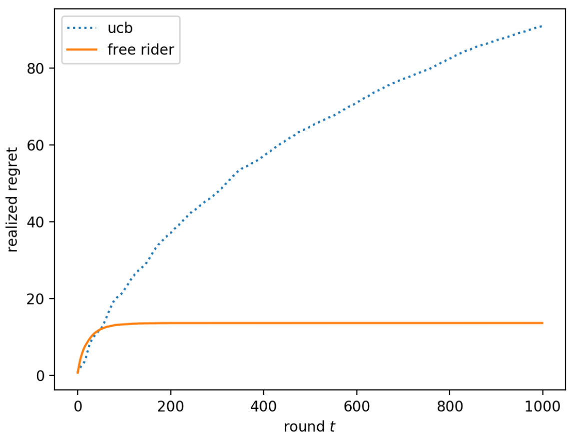

Figure 1: Average realized regret of a 2-UCB player and a free rider over 100 simulations. There are 10 arms whose reward distributions are Bernoulli, with parameters $0.0, 0.1, \ldots, 0.8, 0.9$.

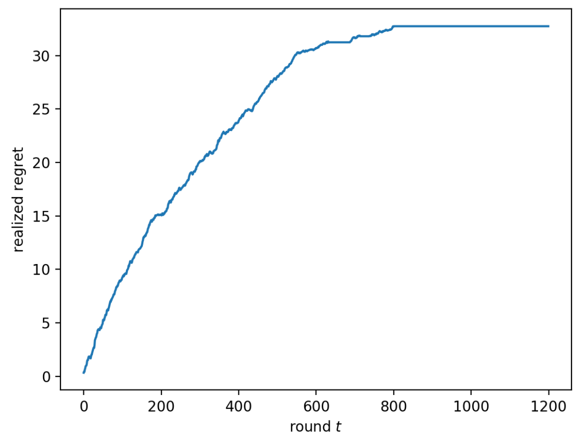

Figure 2: Average realized regret of the free rider over 10 simulations. There are 50 players playing 10-UCB and 30 arms, the dimension of the vectors is 10, and $\Delta = 0.1839$.

essentially the same 1-dimensional instance described in the proof of Theorem 5.2, incurring $\Omega(\log T)$ regret. We conclude that g_γ satisfies the third property. □

A.6 Simulations We present simulation results for both the stochastic and the contextual cases. In the stochastic case, we have the free rider simply pull the most pulled arm of the self-reliant player, and in the contextual case, we have the free rider pull the arm with the highest sample mean calculated by taking the linear combination of the sample means from the other players. In each of the experiments, we refer to $\max_{i\in[k]} \sum_{t=1}^{T} r^{t,i} - \sum_{t=1}^{T} r_p^t$ as the *realized regret* for player p, where $r^{t,i}$ is the reward one would have observed by pulling arm i in round t. For the stochastic case, we consider situations where the reward distribution D_i of each arm i is Bernoulli, with different parameters p. As shown in Figure 1, the realized regret of the free rider flattens out after some constant number of rounds, where this constant depends on the reward distribution profile $\mathbf{D}$.

In the contextual case, the context x_p for each player p and the vector $\mathbf{c}$ (i.e., the coefficients for the linear combination of other players' contexts that gives the free rider's context), are all chosen uniformly at randomly from $[-1, 1]$. The feature distribution F_i for each arm i is a multi-variate normal distribution with covariance matrix is $0.1\mathcal{I}$, where $\mathcal{I}$ is the identity matrix, and the mean vector is once again chosen by sampling each coordinate uniformly from $[-1, 1]$. We normalize x_p and the mean vector of F_i so that the expected reward for each arm i falls within $[-1, 1]$ for every player p.

A.7 Technical Lemmas from Other Sources

LEMMA A.1. (DIVERGENCE DECOMPOSITION [13]) *Let $\mathbf{D} = (D_1, \ldots, D_k)$ and $\mathbf{D}' = (D'_1, \ldots, D'_k)$ be stochastic bandits. For any policy f and time t,*

$$\mathrm{KL}(H^t(\mathbf{D}, f), H^t(\mathbf{D}', f)) = \sum_{i\in[k]} \mathbb{E}[N_i^t(\mathbf{D}, f)]\,\mathrm{KL}(D_i, D'_i)\,.$$

LEMMA A.2. (BRETAGNOLLE-HUBER [13]) *Let P and Q be probability measures on the same measurable space, and let A be any event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(P, Q))\,,$$

where A^c is the complement of A.