BanditQ: Fair Multi-Armed Bandits with Guaranteed Rewards per Arm

Anonymous Author(s)
Affiliation
Address
email

Abstract

Classic no-regret online prediction algorithms, including the Upper Confidence Bound (UCB), Hedge, and EXP3, are inherently unfair by design. Their unfairness stems from their objective of playing the most rewarding arm as frequently as possible while ignoring the less rewarding ones. In this paper, we consider a fair prediction problem in the stochastic setting with a guaranteed minimum rate of accrual of rewards for a subset of arms. We study the problem in both full and bandit feedback settings. Using queueing-theoretic techniques in conjunction with adversarial learning, we propose a new online prediction policy, called BanditQ, that achieves the target reward rates while achieving a regret and target rate violation penalty of $O(T^{3/4})$. In the full-information setting, the regret bound can be further improved to $O(\sqrt{T})$ when considering the average regret over the entire horizon of length T. The proposed BanditQ policy is efficient and admits a black-box reduction from the fair prediction problem to the standard MAB problem with a carefully constructed reward sequence involving state variables that resemble queue lengths. The design and analysis of the BanditQ policy involve recent scale-free second-order regret bounds and a new self-bounding inequality for the reward gradients, which are of independent interest.

1 Introduction

A vast majority of the Multi-armed Bandit (MAB) algorithms deployed in practice are designed to maximize the cumulative rewards (*e.g.,* the number of clicks on ads). Consequently, they could end up discriminating against a subset of arms (which could represent *e.g.,* users belonging to certain categories) the algorithm finds less rewarding [Sweeney, 2013]. In a typical case of algorithmic discrimination, Facebook was sued for targeting ads on housing, credit and employment by race, gender, and religion - all protected classes under US law [Hao, 2019]. A similar problem of fair allocation of resources arises in wireless settings, where schedulers maximizing the total throughput could result in not serving a subset of users having relatively poor channels. As we discuss below, a number of papers have proposed a solution to the fairness problem by guaranteeing a minimum frequency of pulls of each arm. However, in many problems of practical interest, one is primarily interested in guaranteeing a minimum rate of *reward accrual* for each arm - not just a minimum frequency at which the arms are to be pulled. For example, in online ad allocation, the advertisers are interested in maximizing the click-through rates rather than the number of times their ad was displayed on a webpage. In wireless scheduling problems, the users are primarily interested in guaranteed data rates (in Mbps) across a time horizon rather than the number of times they are scheduled - a low-level metric transparent to the users. In online crowd sourcing platforms, workers are primarily interested in the amount of money they make over a given span of time compared to the number of times they are assigned a job. Clearly, the rate of reward accruals depends on the unknown reward distribution, which needs to be learned along the way. In this paper, we solve

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

this fair prediction problem in the stochastic setting through a black-box reduction to an adversarial MAB problem by making use of a natural queueing dynamics. Although we consider stochastic i.i.d. rewards, we will see in the sequel that the use of adversarial MAB algorithms is essential to account for the target reward rate constraints.

1.1 Related Works

There is a vast literature on the classic Multi-armed Bandits problem, where the objective is to sequentially pull an arm at each round from a given set of arms of unknown qualities to maximize the cumulative reward at the end of a time horizon. With the bandit feedback of observed rewards, the problem involves an exploration vs exploitation trade-off. See Cesa-Bianchi and Lugosi [2006], Bubeck et al. [2012], Lattimore and Szepesvári [2020] for text-book treatments on MAB. The fair prediction problem belongs to a class of MAB problems with global constraints. Several authors have considered variants of the fair prediction problem in MAB with different definitions for fairness [Joseph et al., 2016, Gillen et al., 2018, Bechavod et al., 2020, Hossain et al., 2021, Huang et al., 2022]. Closer to our setting, the papers by Patil et al. [2021], Claure et al. [2020], and Li et al. [2019] considered a stochastic MAB problem while requiring the minimum *fraction* of pulls of each arm to exceed a given threshold. Celis et al. [2019] considered a similar problem in the personalized recommendation setting where both the minimum and the maximum fraction of pulls are constrained in order to avoid polarization of views. Technically, Li et al. [2019] used a virtual queueing recursion to handle the fairness constraints. However, their UCB-based policy yielded a regret bound which varies *linearly* with the horizon length [Li et al., 2019, Theorem 2]. Chen et al. [2020] considered the above problem in the contextual bandit setting and proposed a no-regret policy with a known context distribution. Cai et al. [2018] considered a related stochastic MAB problem with a long-term constraint on an auxiliary (level-2) reward process which is assumed to be *independent* of the main (level-1) rewards of the arms. On the other hand, in our problem, the so-called level-1 and level-2 reward processes are identical, and hence, these results do not apply. Constrained bandit problems have also been studied in contexts other than fairness as well. Badanidiyuru et al. [2018], Immorlica et al. [2022], and Xia et al. [2015] considered a related Bandits with Knapsack (`BwK`) problem in stochastic and adversarial environments. In this problem, a given resource budget is allocated to the arms at the beginning, and the game continues until one of the arms finishes all of its budgets. Immorlica et al. [2022] used a Lagrangian-based technique to propose a no-regret policy for the `BwK` problem. Similar to ours, they also employed an adversarial bandit policy in the stochastic setting.

1.2 Our contributions

In contrast with a major line of work in the literature on fair MABs, which is mainly concerned with guaranteeing a minimum fraction of pulls for each arm, in this paper, we initiate the study of a class of problems guaranteeing a minimum *rate* of reward accruals for each arm. Compared to the standard MAB problem, here the difficulty stems from the fact that in addition to pulling the unknown best arm sufficiently many times, other arms with unknown mean rewards also need to be pulled frequently enough so as to satisfy the given fairness constraints. Consequently, our proposed algorithm and its analysis are very different from that of the prior works. In particular, we claim the following major contributions in this paper:

1. We propose an online fair prediction policy, called `BanditQ`, via a *black-box* reduction to the standard adversarial MAB problem. The proposed `BanditQ` policy keeps track of the global reward rate constraints with the help of an auxiliary queueing process, which is used to define the rewards for the standard MAB problem.

2. On the technical side, we use second-order regret bounds to design no-regret online prediction policies with long-term constraints. The key to our results is a new *self-bounding* inequality that implicitly controls the growth of the reward gradients (see Eqs. (14) and (19)). The techniques introduced in this paper could be useful for analyzing other constrained sequential learning problems as well.

3. We complement our theoretical results with illustrative numerical simulations.

2 Problem Formulation

We consider an online prediction problem in the stochastic setting with an additional fairness constraint that requires that each arm belonging to a given subset $\mathcal{P}$ (called *protected class*) must attain pre-specified reward accrual rates, assumed to be feasible. Formally, let there be a set of N arms, which on round t receives a reward vector $\boldsymbol{r}(t) \in [0,1]^N$. We consider a stochastic setting where the rewards $\boldsymbol{r}(t)$ are generated i.i.d. with an unknown expectation vector $\boldsymbol{\mu}$. On round t, an online policy causally predicts a probability distribution $\boldsymbol{x}(t) \in \Delta_N$, where Δ_N is the set of all probability vectors on N arms. The algorithm then randomly samples an arm $I_t \in [N]$ from the distribution $\boldsymbol{x}(t)$[1]. Depending on the feedback structure, either the entire reward vector $\boldsymbol{r}(t)$ (in the case of full-information) or the reward of the sampled arm $r_{I_t}(t)$ (in the case of bandit feedback) is fed to back to the policy at the end of round t. The above process continues for a given time horizon of length T.

Fairness constraints: Due to the action of the policy, the selected arm I_t receives a random reward of value $r_{I_t}(t)$. Hence, if on round t, the prediction policy selects arms according to the distribution $\boldsymbol{x}(t)$, the i^{th} arm receives a (conditional) expected reward of $x_i(t)\mathbb{E}r_i(t) = x_i(t)\mu_i$, and the online policy receives an overall (conditional) expected reward of $\langle \boldsymbol{x}(t), \boldsymbol{\mu} \rangle$. Let $\vec{\boldsymbol{\lambda}}$ be a given vector of target reward rates. Our fairness constraint requires that the long-term rate of rewards accrued by arm $i \in \mathcal{P}$ must be at least $\lambda_i, \forall i \in \mathcal{P}$. We may assume $\lambda_i = 0, \forall i \in [N] \setminus \mathcal{P}$.

Offline Benchmark and Performance Metric: We compare the performance of an online policy against any fixed prediction distribution $\boldsymbol{x}^* \in \Delta_N$ that meets the target reward rates. In other words, our comparator class, denoted by the set $\Omega \subseteq \Delta_N$, is defined as follows:

$$\Omega(\vec{\boldsymbol{\lambda}}) = \{\boldsymbol{x}^* \subseteq \Delta_N : x_i^* \mu_i \geq \lambda_i, \ \forall i, \text{ and } \sum_i x_i^* = 1, \ \boldsymbol{x}^* \geq \boldsymbol{0}\}. \tag{1}$$

Hence, in order for the rate vector $\vec{\boldsymbol{\lambda}}$ to be feasible (*i.e.,* $\Omega(\vec{\boldsymbol{\lambda}}) \neq \varnothing$) it is necessary and sufficient that

$$\sum_i \frac{\lambda_i}{\mu_i} \leq 1. \tag{2}$$

See Section 7.1 in the Appendix for a brief discussion on the feasibility assumption. The set of all offline benchmarks $\Omega(\vec{\boldsymbol{\lambda}})$ is closed and convex with a Euclidean diameter of $D = \sqrt{2}$. Our goal is to design a prediction policy $\{\boldsymbol{x}(t)\}_{t \geq 1}$ that achieves a sublinear (pseudo)-regret against any $\boldsymbol{x}^* \in \Omega$, where

$$\text{Regret}_T(\boldsymbol{x}^*) \equiv \langle \boldsymbol{x}^*, \boldsymbol{\mu} \rangle T - \mathbb{E} \sum_{t=1}^{T} \langle \boldsymbol{x}(t), \boldsymbol{r}(t) \rangle, \tag{3}$$

while meeting the long-term reward rate constraints that we formalize next[2]. Formally, for any time interval $\mathcal{I} \subseteq [T]$, the asymptotic rate constraint requires:

$$\liminf_{|\mathcal{I}| \to \infty} |\mathcal{I}|^{-1} \mathbb{E}\Big[\sum_{t \in \mathcal{I}} r_i(t) x_i(t) \Big] \geq \lambda_i, \ \forall i \in \mathcal{P}. \tag{4}$$

Note that Eq. (4) requires the minimum reward rate guarantee to hold *uniformly* across the time horizon for any sufficiently long interval of time. In other words, we require that no individual arm is starved for a long period of time - a problem left open by Patil et al. [2021]. Furthermore, following Cai et al. [2018], we also define a non-asymptotic *rate violation penalty* as follows:

$$\mathbb{V}(T) = \max_{i \in \mathcal{P}} \mathbb{E}\Big[\sum_{t=1}^{T} \big(\lambda_i - r_i(t) x_i(t) \big) \Big]. \tag{5}$$

In brief, we seek an online prediction policy for which Regret_T and $\mathbb{V}(T)$ increase sub-linearly in T, and satisfies (4). We note two fundamental differences between the above problem and the standard online learning framework [Orabona, 2019]. First, contrary to the online learning setting, where the set of benchmarks Ω is specified a priori (independent of the rewards), in this problem, the set of benchmarks (1) depends on the unknown reward distributions. Second, unlike the online learning setting, the action taken by the policy on a round is not restricted to the set $\Omega(\vec{\boldsymbol{\lambda}})$ provided that the long-term target rates are met. Note that when the vector $\vec{\boldsymbol{\lambda}}$ is set to zero, the above problem reduces to the classic MAB problem. We now propose the `BanditQ` policy that solves the above problem.

[1] The prediction policy could be deterministic where $\boldsymbol{x}(t)$ is supported on only one arm (*e.g.,* the `UCB` policy).

[2] In the case of the worst-case regret, we drop the argument in the regret definition (3).

3 `BanditQ` policy in the full-information setting

For simplicity of exposition, we first consider the full-information setup when the entire reward vector is revealed to the learner at the end of each round. Apart from a technical result on the diameter of an auxiliary random process (Proposition 7 in the Appendix), the extension to the bandit setup requires no essential change in the analysis and will be dealt with in the following section. On a high level, the `BanditQ` policy first defines a natural queueing dynamics to take into account the target reward rates. It then extends the *drift-plus-penalty* framework of Neely [2010, Chapter 4] to simultaneously achieve a small regret and meet the long-term rate constraints. However, to make this work, we must adapt the asymptotic stochastic setting of Neely [2010] to the non-asymptotic adversarial setup with online/bandit information. This extension is highly non-trivial and requires a new proof and algorithmic technique, which are very different from that of the `Max-Weight` policy in Neely [2010].

We associate a non-negative state variable $Q_i(t)$ to each protected arm $i \in \mathcal{P}$. Under the action of an online policy $\pi = \{\boldsymbol{x}(t)\}_{t\geq 1}$, the state variables in the set $i \in \mathcal{P}$ evolve according to the following queueing dynamics, known as the Lindley recursion [Lindley, 1952]:

$$Q_i(t) = \big(Q_i(t-1) + \lambda_i - r_i(t)x_i(t)\big)^+, \; Q_i(0) = 0, \tag{6}$$

where we adopt the standard notation $(y)^+ \equiv \max(0, y)$. We set $Q_i(t) = 0, \forall t, \forall i \notin \mathcal{P}$. To get an intuition for Eq. (6), imagine that on every round t, a fixed deterministic amount of work λ_i arrives at the queue Q_i. Then, under the action $\boldsymbol{x}(t)$ of an online policy, $\min(Q(t-1) + \lambda_i, r_i(t)x_i(t))$ amount of work departs from Q_i. It is intuitive that to stabilize the queues the long-term service rates must be at least as large as the long-term arrival rates. Thus, any online policy stabilizing the queues would automatically satisfy the target rate requirements. However, since we are also interested in achieving a small regret, meeting the rate constraints alone is not enough (*c.f.* Huang et al. [2023]). Our online policy must also perform competitively in terms of cumulative rewards against every feasible stationary actions given by (1). Towards this, let us define the following quadratic potential function (*a.k.a.* Lyapunov function in the queueing theory literature):

$$\Phi(t) = \sum_{i\in\mathcal{P}} Q_i^2(t). \tag{7}$$

We now upper bound the change of potential under the action of a policy. From (6), we have

$$Q_i^2(t) \leq \big(Q_i(t-1) + \lambda_i - r_i(t)x_i(t)\big)^2 \leq Q_i^2(t-1) + \lambda_i + x_i(t) + 2Q_i(t-1)(\lambda_i - r_i(t)x_i(t)),$$

where, in the last inequality, we have used the fact that $0 \leq \lambda_i, r_i(t), x_i(t) \leq 1, \forall i, t$. Summing up the above inequality, we have the following upper bound for the change of potential on round t:

$$\Phi(t) - \Phi(t-1) \leq 2 + 2\sum_{i\in\mathcal{P}} Q_i(t-1)\big(\lambda_i - r_i(t)x_i(t)\big), \tag{8}$$

where we have used the fact that $\sum_i \lambda_i \leq 1, \sum_i x_i(t) \leq 1$. Motivated by the drift-plus-penalty framework of Neely [2010], we now define an instance of the standard online linear optimization (OLO) problem Ξ with action set Δ_N, where the reward of the i^{th} arm on round t is defined as:

$$r_i'(t) \equiv \big(Q_i(t-1) + V_t\big)r_i(t), \;\; \forall i \in [N]. \tag{9}$$

In the above, $\{V_t\}_{t\geq 1}$ is a sequence of non-negative parameters. In our theoretical results, we will primarily consider a constant sequence $V_t = V = \Theta(\sqrt{T}), \forall t$. Intuitively, the surrogate rewards $\boldsymbol{r}'(t)$ strike a balance between attaining the target rates (through the first term) and achieving a small regret (through the second term). However, the definition (9) leads to two significant technical challenges for learning the surrogate rewards online. First, due to the presence of the queue variables, the reward vectors $\boldsymbol{r}'(t)$ need not be bounded *a priori*, which critically affects the regret bound of the surrogate problem Ξ. Second, although the original reward sequence $\{\boldsymbol{r}(t)\}_{t\geq 1}$ is i.i.d., the reward sequence $\{\boldsymbol{r}'(t)\}_{t\geq 1}$ for the auxiliary problem is *not* i.i.d. any more, again due to the presence of the queue variables, which are highly correlated via Eq. (6). The second difficulty prompts us to use an adversarial online learning algorithm for the auxiliary OLO problem Ξ, as discussed below.

The `BanditQ` policy: Our proposed `BanditQ` policy uses an adaptive no-regret policy with a second-order regret bound, *e.g.,* Online Gradient Ascent (`OGA`) [Orabona, 2019] or `Squint` [Koolen and Van Erven, 2015], for the auxiliary problem Ξ for choosing the prediction distribution $\boldsymbol{x}(t)$.

In this paper, we use the OGA policy due to its simplicity. Recall that the OGA policy updates the prediction distribution on each round via a gradient step with an adaptive step size as follows:

$$\boldsymbol{x}(t+1) \leftarrow \Pi_{\Delta_N}\left(\boldsymbol{x}(t) + \frac{\boldsymbol{r}'(t)}{\sqrt{2\sum_{\tau=1}^{t}\|\boldsymbol{r}'(\tau)\|_2^2}}\right). \tag{10}$$

In the above, $\Pi_{\Delta_N}(\cdot)$ denotes the Euclidean projection operator on the standard simplex Δ_N, which can be efficiently implemented in $O(N\log N)$ time [Wang and Carreira-Perpinán, 2013]. The complete BanditQ policy in the full-information setting is summarized in Algorithm 1.

Algorithm 1 BanditQ Policy in the Full-Information setting

Input: Target reward rate vector $\vec{\boldsymbol{\lambda}}$, Euclidean projection oracle $\Pi_{\Delta_N}(\cdot)$ onto the simplex Δ_N.
$\boldsymbol{Q} \leftarrow \boldsymbol{0}, \boldsymbol{x} \leftarrow [1/N, 1/N, \ldots, 1/N], V \leftarrow \sqrt{T}, S \leftarrow 0.$ ▷ *Initialization*
for each round $t = 1 : T$: **do**
  Sample an arm $I_t \in [N]$ from the distribution $\boldsymbol{x}$.
  Observe the *entire* reward vector $\boldsymbol{r}(t)$ ▷ *Full-information feedback*
  $Q_i = \big(Q_i + \lambda_i - r_i(t)x_i\big)^+, \ \forall i \in \mathcal{P}.$ ▷ *Updating the queue lengths*
  $r'_i(t) \leftarrow \big(Q_i + V\big)r_i(t), \ \forall i \in [N]$ ▷ *Computing the surrogate rewards*
  $S \leftarrow S + \|\boldsymbol{r}'(t)\|^2.$ ▷ *Accumulating the norm of past gradients*
  $\boldsymbol{x} \leftarrow \Pi_{\Delta_N}\left(\boldsymbol{x} + \frac{\boldsymbol{r}'(t)}{\sqrt{2S}}\right)$ ▷ *Implementing the online gradient ascent step*
end for each

In our analysis, we use the following standard second-order regret bound achieved by the OGA policy.

Theorem 1 (Theorem 4.14 of Orabona [2019]). *Let $X \subseteq \mathbb{R}^d$ be a convex set with an Euclidean diameter D. Consider a sequence of linear reward functions with gradients $\{\boldsymbol{g}_t\}_{t\geq 1}$. Assume that the Online Gradient Ascent policy is run with step sizes $\eta_t = \frac{D}{\sqrt{2\sum_{\tau=1}^{t}\|\boldsymbol{g}_\tau\|_2^2}}, 1 \leq t \leq T$. Then the regret of the OGA policy can be upper-bounded as follows:*

$$\mathit{Regret}_T \leq D\sqrt{2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_2^2}. \tag{11}$$

It is important to note that the above bound is *scale-free*, *i.e.,* no *a priori* bounds on the gradients are needed for the above result [Putta and Agrawal, 2022, Hadiji and Stoltz, 2023]. Specializing Theorem 1 to our surrogate problem Ξ, we obtain the following second-order regret bound:

$$\mathrm{Regret}_t^{\Xi} \leq 2\sqrt{\sum_{\tau=1}^{t}\sum_i (Q_i(\tau-1)+V_\tau)^2 r_i(t)^2} \leq 2\sqrt{2\sum_{\tau=1}^{t}\sum_i Q_i^2(\tau)} + 2\sqrt{2N\sum_{\tau=1}^{t}V_\tau^2}, \tag{12}$$

where we have used the fact that $0 \leq r_i(t) \leq 1, \forall t, i$, and the elementary inequality $\sqrt{x+y} \leq \sqrt{x}+\sqrt{y}$.

3.1 Analysis and Regret Bounds

Unlike the analysis in [Patil et al., 2021, Cai et al., 2018], which proceeds by constructing UCB-like stochastic confidence intervals for the mean rewards of each arm, we directly make use of the regret bound (12) via an "adversarial-style" analysis, which revolves around a new *self-bounding* inequality. Since the state variables $\{\boldsymbol{Q}(t)\}_{t\geq 1}$ evolve according to the recursion (6), we do not immediately have an explicit control on the regret bound (12). Hence, to control the regret, we take an indirect approach. Fix any distribution $\boldsymbol{x}^* \in \Omega$. From Eq. (8), we have

$$\Phi(\tau) - \Phi(\tau-1) - 2V_\tau\sum_i r_i(\tau)x_i(\tau) \leq 2 + 2\sum_i Q_i(\tau-1)\lambda_i - 2\sum_i \underbrace{(Q_i(\tau-1)+V_\tau)r_i(\tau)}_{r'_i(\tau)}x_i(\tau).$$

Summing up the above inequality from $\tau = 1$ to $\tau = t$ and recalling that $\Phi(t) = \sum_i Q_i^2(t), \Phi(0) = 0$,

$$\sum_i Q_i^2(t) + 2\sum_{\tau=1}^{t}V_\tau\sum_i r_i(\tau)(x_i^* - x_i(\tau)) \leq 2t + 2\sum_{\tau=1}^{t}\sum_i Q_i(\tau-1)\big(\lambda_i - r_i(\tau)x_i^*\big) + 2\mathrm{Regret}_t^{\Xi}, \tag{13}$$

where Regret_t^{Ξ} denotes the worst-case regret for the surrogate problem (defined similarly as Eq. (3)). Note that, in the above, the regret bound on the RHS is random as it depends on the magnitude of the random queue process $\{\boldsymbol{Q}(\tau)\}_\tau$. In our analysis, we will exclusively consider a constant $\{V_t\}_{t\geq 1}$ sequence, where $V_t = V, \forall t \geq 1$ for some $V \geq 0$ to be fixed later. Let $\{\mathcal{F}_\tau\}_{\tau\geq 0}$ be the natural filtration generated by the sequence of rewards $\{\boldsymbol{r}(\tau)\}_{\tau\geq 0}$. We now have the following series of inequalities:

$$
\begin{aligned}
& \sum_i \mathbb{E}Q_i^2(t) + 2V\text{Regret}_t(\boldsymbol{x}^*) \\
= \;& \sum_i \mathbb{E}Q_i^2(t) + 2V \sum_{\tau=1}^{t} \mathbb{E}\sum_i r_i(\tau)(x_i^* - x_i(\tau)) \\
\overset{(a)}{\leq} \;& 2t + 2\sum_{\tau=1}^{t} \mathbb{E}\sum_i Q_i(\tau-1)\big(\lambda_i - x_i^*\mathbb{E}[r_i(\tau)|\mathcal{F}_{\tau-1}]\big) + 2\mathbb{E}\big[\text{Regret}_t^{\Xi}\big] \\
\overset{(b)}{\leq} \;& 2t + 2\sum_{\tau=1}^{t} \mathbb{E}\sum_i Q_i(\tau-1)\big(\lambda_i - \mu_i x_i^*\big) + 2\mathbb{E}\big[\text{Regret}_t^{\Xi}\big] \\
\overset{(c)}{\leq} \;& 2t + 2\mathbb{E}\big[\text{Regret}_t^{\Xi}\big] \\
\overset{(d)}{\leq} \;& 2t + 4\sqrt{2\sum_{\tau=1}^{t}\sum_i \mathbb{E}Q_i^2(\tau)} + 4V\sqrt{2Nt},
\end{aligned} \tag{14}
$$

where in (a), we have taken the expectation of both sides of (13) with respect to the i.i.d. reward process $\{\boldsymbol{r}(t)\}_{t\geq 1}$, and used the law of iterated expectation; in (b) we have used the i.i.d. nature of the reward generation process; in (c) we have used the feasibility condition of the benchmark $\boldsymbol{x}^*$ from Eq. (1); in (d) we have used the second-order regret bound from Eq. (12) in conjunction with Jensen's inequality used for the square root function. Inequality (14) constitutes the key step in our analysis. This result shows that the queue-length process $\{\boldsymbol{Q}(t)\}_{t\geq 1}$ under the `BanditQ` policy possesses a *self-bounding* property in the sense that the expected queue-length squared at any time t is bounded by the square root of the sum of expected queue-length squared up to time t plus other auxiliary terms. Inequality (14) leads to the following bound on the second moments of the queue variables.

Proposition 1. *Setting $V_t = V = \Theta(\sqrt{T}), \forall t \geq 1$, we obtain $\mathbb{E}Q_i^2(t) = O(\sqrt{N}T^{3/2}), 1 \leq t \leq T$.*

Proof. From Eq. (14), we have that for all $i \in \mathcal{P}$ and all $t \geq 1$:

$$
\mathbb{E}Q_i^2(t) \leq 2(V+1)t + 4\sqrt{2\sum_{\tau=1}^{t}\sum_i \mathbb{E}Q_i^2(\tau)} + 4V\sqrt{2Nt}, \tag{15}
$$

where we have used the fact that $\sum_i r_i(\tau)(x_i(\tau) - x_i^*) \leq 1, \forall \tau$. Which implies that $\text{Regret}_t(\boldsymbol{x}^*) \geq -t$. Furthermore, since $\lambda_i \leq 1$, from Eq. (6), we trivially have $Q_i(\tau) \overset{a.s.}{\leq} \tau, \forall i, \tau$ We now improve upon this trivial upper bound on the queue lengths by substituting it on the RHS of Eq. (15), which yields:

$$
\mathbb{E}Q_i^2(t) \leq O(T^{3/2}) + O\Big(\sqrt{N\sum_{\tau=1}^{t}\tau^2}\Big) + O(\sqrt{N}T) = O(\sqrt{N}T^{3/2}), \forall i, t \in [T].
$$

□

As a consequence of Proposition 1, the following result shows that the under the action of the `BanditQ` policy, the target reward accrual rates are met while incurring a reward violation penalty of $O(T^{3/4})$. Our result improves upon the $O(T^{5/6})$ violation penalty established by Cai et al. [2018, Theorem 1] under independence assumptions.

Proposition 2. *Upon setting $V_t = V = \Theta(\sqrt{T})$, for any interval $\mathcal{I} \subseteq [T]$ such that $T^{3/4} = o(|\mathcal{I}|)$, the `BanditQ` policy in the full-information setting yields:*

$$
\liminf_{|\mathcal{I}|\to\infty} |\mathcal{I}|^{-1}\mathbb{E}\sum_{t\in\mathcal{I}} r_i(t)x_i(t) \geq \lambda_i, \;\; \forall i \in \mathcal{P}, \;\; \textit{and} \;\; \mathbb{V}(T) = O(N^{1/4}T^{3/4}).
$$

See Appendix 7.2 for the proof. Using Proposition 1 once again, we now derive a sublinear regret bound achieved by the `BanditQ` policy.

Proposition 3. *Upon setting $V_t = V = \Theta(\sqrt{T}), 1 \le t \le T$, the* `BanditQ` *policy achieves a minimax regret bound of $O((NT)^{3/4})$ in the full-information setting.*

See Appendix 7.3 for the proof. Note that, unlike the standard MAB problem, in this case, the worst-case regret could be negative on some rounds. This stems from the fact that, unlike the offline benchmark, the `BanditQ` policy is not *required* to always take actions from the set Ω, which is unknown to the policy. This poses a technical difficulty in proving an $O(\sqrt{T})$ worst-case regret bound starting from Eq. (14). Nevertheless, our next result shows that the proposed `BanditQ` policy admits a substantially stronger bound for the *average* regret, averaged over the entire time horizon T.

Proposition 4. *Under the* `BanditQ` *policy with full feedback with $V_t = V = \Theta(\sqrt{T}), \forall t$, we have that $\frac{1}{T}\sum_{t=1}^{T} Regret_t(\boldsymbol{x}^*) = O(\sqrt{NT})$, for any $\boldsymbol{x}^* \in \Omega$.*

See Appendix 7.4 for the proof. The reader should compare the above bound with the unconstrained case, where the min-max regret is lower bounded by $\Omega(\sqrt{T \log N})$ [Lattimore and Szepesvári, 2020]. Finally, if one is only interested in achieving the target rates while disregarding the cumulative rewards altogether, the following result shows that the queue-length bound in Proposition 1 can be further improved to $O(\sqrt{t})$ by setting $V_t = 0, \forall t \in [T]$.

Proposition 5. *Setting $V_t = 0, \forall t \ge 1$, the second and first moments of the state variables $\{\boldsymbol{Q}(t)\}_{t\ge 1}$ can be bounded as:*

$$\mathbb{E}Q_i^2(t) \le 64Nt, \text{ which implies } \mathbb{E}Q_i(t) \le 8\sqrt{Nt}, \ \forall i \in \mathcal{P}, \forall t \ge 1.$$

See Appendix 7.5 for the proof. This would imply corresponding improved bounds in Proposition 2.

4 `BanditQ` policy with Bandit feedback

Recall that in the case of bandit feedback, only the reward of the selected arm , *i.e.,* $r_{I_t}(t)$, is revealed to the policy at the end of round t. The reader should compare this with the previous full-information setup where the entire reward vector $\boldsymbol{r}(t)$ is revealed to the policy irrespective of its action. To deal with the resulting in the *exploration-exploitation* trade-off in the limited information setup, we replace the full-information `OGA` policy (10) with a recent adversarial MAB policy, proposed by Putta and Agrawal [2022], that enjoys a *scale-free* second-order regret bound similar to Eq. (11). Their *Follow-the-regularized-leader* (FTRL)-based MAB policy uses the usual inverse propensity score to estimate the reward vectors and employs a log-barrier regularizer in the FTRL step with a carefully chosen learning rate schedule. The arms are finally selected by mixing a uniform exploration component with the distribution from the FTRL step. For completeness, we describe the `BanditQ` policy in the bandit information setting in Appendix 7.8. Putta and Agrawal [2022] showed that their proposed MAB policy works for *any* real loss vector (unlike, *e.g.,* EXP3, which requires non-negative losses) and enjoys the following regret bound.

Theorem 2 (Putta and Agrawal [2022]). *MAB Algorithm 1 of Putta and Agrawal [2022], run with linear reward sequence with coefficient vectors $\{\boldsymbol{g}_t\}_{t=1}^{T}$, enjoys the following scale-free regret bound:*

$$Regret_T = \tilde{O}\left(\sqrt{N\sum_{t=1}^{T}||\boldsymbol{g}_t||_2^2} + \max_{t\in[T]}||\boldsymbol{g}_t||_\infty\sqrt{NT}\right). \tag{16}$$

It can be seen that the only essential difference between the above regret bound and that of the `OGA` regret bound given in Eq. (11) is the presence of an additional $\tilde{O}(\max_{t\in[T]}||\boldsymbol{g}_t||_\infty\sqrt{NT})$ term in the former. However, we will see that with the help of an additional technical result (Proposition 7 in the Appendix), our previous arguments go through with minimal changes. In the following, we outline the main changes necessary to go from the full-information setting to the bandit setup.

Notation: Let us denote the random arm selected on round t by the one-hot encoded vector $\boldsymbol{X}(t) = \big(X_1(t), X_2(t), \ldots, X_N(t)\big) \in \{0,1\}^N$ such that $\mathbb{P}(X_i(t) = 1|\mathcal{F}_{t-1}) = 1 - \mathbb{P}(X_i(t) = 0|\mathcal{F}_{t-1}) = x_i(t), \forall i, t$. Hence the marginal distributions satisfy the relation $\mathbb{E}(X_i(t)|\mathcal{F}_{t-1}) = x_i(t), \forall i, t$.

Queueing recursion and the auxiliary MAB problem: Note that the queueing recursion (6) for the full-feedback setting does not work in the case of Bandit feedback because the rewards of the

unobserved arms are not revealed. However, it is straightforward to modify the recursion (6) by replacing the prediction probabilities $\boldsymbol{x}(t)$ with the corresponding random realizations $\boldsymbol{X}(t)$. Hence, in the bandit setting, the modified queueing evolution reads:

$$Q_i(t) = \big(Q_i(t-1) + \lambda_i - r_i(t)X_i(t)\big)^+,\ Q_i(0) = 0. \tag{17}$$

Eq. (17) is well-defined in the bandit feedback setting as $X_i(t) = 0$ if $i \neq I_t$. Next, analogous to the full-information setting (Eq. (9)), the `BanditQ` policy defines an instance of an adversarial MAB problem $\Xi^{\texttt{Bandit}}$ where the reward of the i^{th} arm on round t is defined as

$$r'_i(t) \equiv \big(Q_i(t-1) + V_t\big)r_i(t),\ \forall i \in [N]. \tag{18}$$

As before, the reward components could be large with no a priori non-trivial upper bounds.

4.1 Analysis and Regret Bounds

As before, the components of the reward gradients are given by $\boldsymbol{g}_{t,i} = r'_i(t) = \big(Q_i(t-1) + V_t\big)r_i(t)$. Set $V_t = V = \Theta(\sqrt{T}), \forall t$. Using the same quadratic potential function $\Phi(\cdot)$ in Eq. (7), and working identically up to step (c) of Eq. (14), we have the following self-bounding inequality:

$$\begin{aligned}
&\sum_i \mathbb{E}Q_i^2(t) + 2V\sum_{\tau=1}^{t}\mathbb{E}\sum_i r_i(\tau)\big(x_i^* - X_i(\tau)\big) \\
\leq\ & 2t + 2\mathbb{E}\big[\text{Regret}_t^{\Xi^{\texttt{Bandit}}}\big] \\
\overset{(a)}{\leq}\ & 2t + \tilde{O}\Big(\sqrt{N\sum_{\tau=1}^{t}\sum_i \mathbb{E}Q_i^2(\tau)} + NV\sqrt{T} + V\sqrt{NT} + \sqrt{NT}\mathbb{E}\big[\max_{i,t\in[T]}(Q_i(t))\big]\Big) \qquad (19) \\
\overset{(b)}{\leq}\ & 2t + \tilde{O}\Big(\sqrt{N\sum_{\tau=1}^{t}\sum_i \mathbb{E}Q_i^2(\tau)} + NV\sqrt{T} + \sqrt{N}T^{3/2}\Big), \qquad (20)
\end{aligned}$$

where, in step (a), we have substituted the regret bound from Theorem 2 and in step (b), we have used the trivial bound $Q_i(t) \leq T, \forall t \in [T]$. Hence, using the fact that $\sum_i r_i(\tau)(X_i(\tau) - x_i^*) \leq 1, \forall \tau$, similar to Eq. (15), we have that for all $t \geq 1$:

$$\mathbb{E}Q_i^2(t) \leq 2(V+1)t + \tilde{O}\Big(\sqrt{N\sum_{\tau=1}^{t}\sum_i \mathbb{E}Q_i^2(\tau)} + NT + \sqrt{N}T^{3/2}\Big) = \tilde{O}(NT^{3/2}), \tag{21}$$

where, in the last inequality, we have again used the trivial bound $Q_i(t) \leq T, \forall t \in [T]$ on the RHS. Substituting the above bound back in the RHS of Eq. (21), we get an improved bound

$$\mathbb{E}Q_i^2(t) = \tilde{O}(\sqrt{N}T^{3/2}),\ \forall t \in [T]. \tag{22}$$

Eq. (22) yields the following counterpart to Proposition 2 with an identical proof, which we omit.

Proposition 6. *Setting $V_t = V = \Theta(\sqrt{T})$, for any interval $\mathcal{I} \subseteq [T]$ such that $T^{3/4} = o(|\mathcal{I}|)$, the `BanditQ` policy in the bandit information setting yields:*

$$\liminf_{|\mathcal{I}|\to\infty} |\mathcal{I}|^{-1}\mathbb{E}\sum_{t\in\mathcal{I}} r_i(t)x_i(t) \geq \lambda_i,\ \forall i \in \mathcal{P},\ \ \textit{and}\ \ \mathbb{V}(T) = O(N^{1/4}T^{3/4})$$

Our final result is the following sublinear regret bound for the `BanditQ` policy in the bandit setting. However, the proof is more technical as we now need to strengthen Eq. (22) to bound the *diameter* of the queueing processes $\{\boldsymbol{Q}(t)\}_{t\geq 1}$, *i.e.,* $\mathbb{E}(\max_{i,t} Q_{i,t})$. See Proposition 7 in the Appendix for the derivation of this bound using Martingale methods. Using this result, we now establish the following regret bound for the `BanditQ` policy under bandit feedback. See Appendix 7.7 for the proof.

Theorem 3. *Upon setting $V_t = V = \Theta(\sqrt{T}), t \in [T]$, the `BanditQ` policy achieves a regret bound of $\tilde{O}(N^{5/4}T^{3/4})$ in the bandit feedback setting.*

Remarks: When all target rates are zero (*i.e.,* $\vec{\boldsymbol{\lambda}} = \vec{\mathbf{0}}$), the fair prediction problem reduces to the classic MAB problem, which is known to have a minimax regret bound of $O(\sqrt{NT})$ [Lattimore and Szepesvári, 2020]. Evidently, the regret bound given by Proposition 3 could be improved further. We leave the question of the tightness of the above regret bound as an interesting open problem.

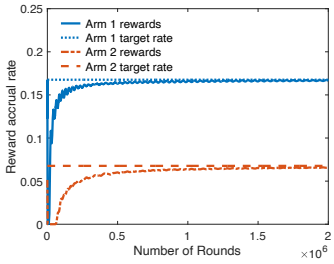

Figure 1: Reward accrual rates in the full-information setting

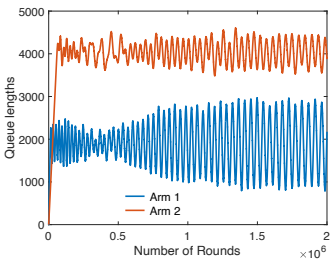

Figure 2: Queue lengths under in the full-information setting

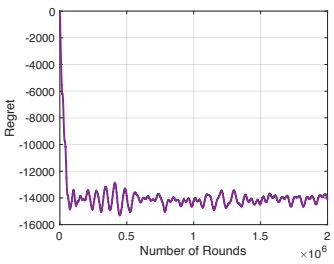

Figure 3: Regret of `BanditQ` in the full-information setting

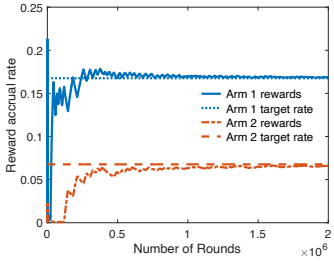

Figure 4: Reward accrual rates in the bandit feedback

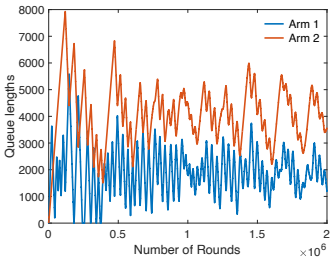

Figure 5: Queue lengths in bandit feedback setting

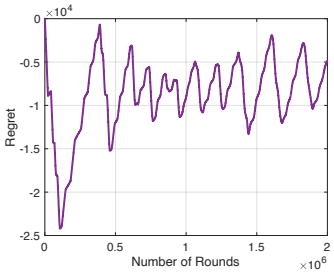

Figure 6: Regret of `BanditQ` in the bandit feedback setting

5 Numerical Simulations

Simulation Setup: We consider a problem instance with $N = 5$ arms and $k = 2$ protected classes consisting of the first and the second arm. We set the mean reward vector of the arms to $\boldsymbol{\mu} = (0.335, 0.203, 0.241, 0.781, 0.617)$, and the target reward rates for the first and the second arm to $\lambda_1 = 0.167$ and $\lambda_2 = 0.067$ respectively. From Eq. (2), it can be easily verified that the required rates are feasible for this problem. Clearly, Arm # 4 is the most rewarding among the five arms. We simulate the `BanditQ` policy for $T = 2 \times 10^6$ rounds upon setting the parameter $V = \sqrt{T} \approx 1414$. We write a custom optimizer, described in Appendix 7.9, for efficiently implementing the optimization steps in the `BanditQ` policy.

Discussion: Figures 1, 2, and 3 show the performance of the `BanditQ` policy in the full-information set-up. Figure 1 shows that the protected arms, Arm 1 and Arm 2, asymptotically meet their target rates. Note that since both Arm 1 and Arm 2 have sub-optimal expected rewards, they would have received asymptotically zero reward rates under the action of an unfair prediction policy. Figure 2 shows the evolution of the surrogate queue length variables, and Figure 3 shows the regret of the `BanditQ` policy in the full-information setting. Negative regret suggests that the `BanditQ` policy achieves a cumulative reward that exceeds the reward achieved by the static benchmark policy (which is forced to take actions from the restricted set Ω on all rounds). Figures 4, 5, and 6 show the corresponding plots in the bandit feedback setting. As expected, in the case of bandit feedback, the variables exhibit greater variance compared to their full-information counterpart due to the limited availability of information. However, the `BanditQ` policy achieves the target rates in this case as well. A comparison of the `BanditQ` policy with an Oracle LFG policy is shown in Appendix 7.10.

6 Conclusion and Open Problems

Since we use adversarial MAB policies as subroutines, it is reasonable to conjecture that the proposed `BanditQ` policy is robust and would work for adversarial rewards as well. However, proving this statement, or designing fair learning policies in the adversarial setting, are beyond our scope. Improving the regret and the queue length bounds and coming up with instance-dependent regret bounds for the fair learning problem would be interesting. Finally, designing an anytime version of the policy that does not need to know the horizon length T in advance would be practically useful.

References

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

Karen Hao. Facebook's ad-serving algorithm discriminates by gender and race. *MIT Technology Review*, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31, 2018.

Yahav Bechavod, Christopher Jung, and Steven Z Wu. Metric-free individual fairness in online learning. *Advances in neural information processing systems*, 33:11214–11225, 2020.

Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34:24005–24017, 2021.

Wen Huang, Lu Zhang, and Xintao Wu. Achieving counterfactual fairness for causal bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6952–6959, 2022.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *The Journal of Machine Learning Research*, 22(1):7885–7915, 2021.

Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 299–308, 2020.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 160–169, 2019.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pages 181–190. PMLR, 2020.

Kechao Cai, Xutong Liu, Yu-Zhen Janice Chen, and John CS Lui. An online learning approach to network application optimization with guarantee. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 2006–2014. IEEE, 2018.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3960–3966. AAAI Press, 2015. ISBN 9781577357384.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Michael J Neely. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks*, 3(1):1–211, 2010.

David V Lindley. The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289. Cambridge University Press, 1952.

Jiatai Huang, Leana Golubchik, and Longbo Huang. Queue scheduling with adversarial bandit learning. *arXiv preprint arXiv:2303.01745*, 2023.

Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175. PMLR, 2015.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Sudeep Raja Putta and Shipra Agrawal. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 910–930. PMLR, 2022.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k–armed bandits. *Journal of Machine Learning Research*, 24(13):1–33, 2023.

Sheldon M Ross. *Stochastic processes*. John Wiley & Sons, 1995.

Joseph L Doob. *Stochastic processes*. John Wiley & Sons, 1953.

Lester E Dubins and Gideon Schwarz. A sharp inequality for sub-martingales and stopping-times. *Astérisque*, 157(158):129–145, 1988.

Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2011.