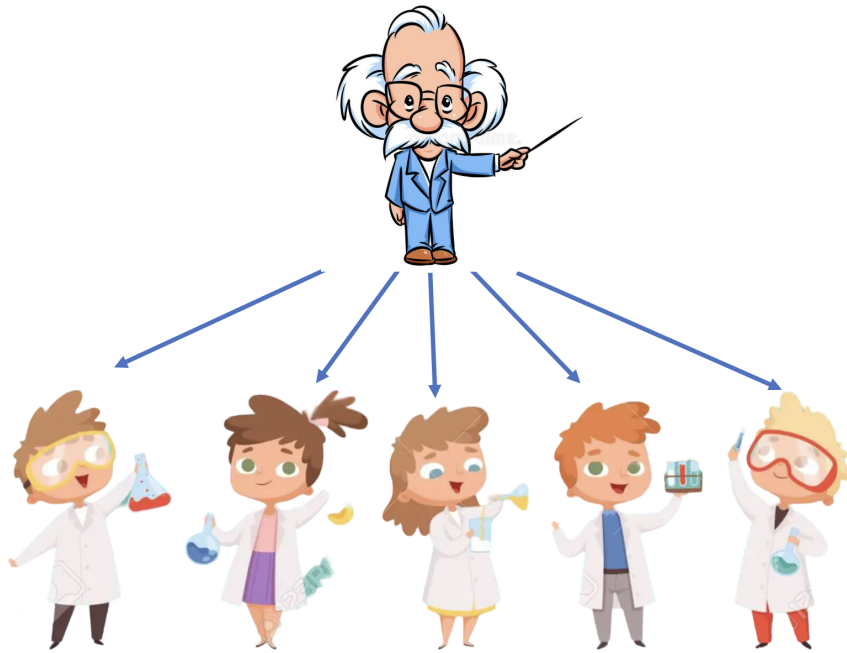
BanditQ: Fair Bandits with Guaranteed Rewards

Abhishek Sinha

Tata Institute of Fundamental Research
Mumbai, India

Fun Problem: Fair allocation of Research Projects

- A Professor receives grants for research projects sequentially over time and wants to assign them *fairly* to his N PhD students



Problem solving probabilities

- The skill levels of the students are initially unknown to the Professor. It can only be learned from their past performances.

Setup

- The students remain in the school for T years
- To meet the degree requirements, student i must solve on the average at least λ_i problems per year
  - The minimum requirements are known to be feasible in expectation
- The advisor wants to solve as many problems as possible subject to each student meeting the minimum requirements

> Question: What is an optimal problem assignment strategy for the Professor?

Applications: Recommendation systems, Online ad allocation, Crowdsourcing, Wireless scheduling, ...

The Fair Bandit Problem

- We are given a set of N arms and a target reward rate vector $\boldsymbol{\lambda} \in \mathbb{R}_N^+$
- On round t, an online policy π selects an arm $I_t \in [N]$ to play
- Both the policy and the selected arm I_t receive an i.i.d. random reward of value $r_{I_t}(t) \in [0, 1]$
- The goal of the policy π is to maximize the rewards while ensuring that the long-term reward accrual rate of arm i is at least $\lambda_i, \forall i \in [N]$.
- The problem is interesting in both Full-information and Bandit feedback settings

Performance Metric

- Let $\boldsymbol{\mu}$ denote the (unknown) mean rewards of the arms and $\boldsymbol{x}(t)$ denote the sampling probabilities of the arms on round t
- Let $\Omega(\boldsymbol{\lambda})$ denote the set of all static feasible distributions of pulls,

$$\Omega(\boldsymbol{\lambda}) = \left\{\boldsymbol{x}^* : x_i^* \mu_i \geq \lambda_i, \forall i, \sum_i x_i^* = 1, \boldsymbol{x}^* \geq \mathbf{0}\right\}.$$

Performance Metric

- Let $\boldsymbol{\mu}$ denote the (unknown) mean rewards of the arms and $\boldsymbol{x}(t)$ denote the sampling probabilities of the arms on round t
- Let $\Omega(\boldsymbol{\lambda})$ denote the set of all static feasible distributions of pulls,

$$\Omega(\boldsymbol{\lambda}) = \left\{\boldsymbol{x}^* : x_i^* \mu_i \geq \lambda_i, \forall i, \sum_i x_i^* = 1, \boldsymbol{x}^* \geq \mathbf{0}\right\}.$$

- Our objective is to design an online policy $\pi = \{\boldsymbol{x}(t)\}_{t=1}^T$ which achieves a sublinear regret against all *static feasible distributions*:

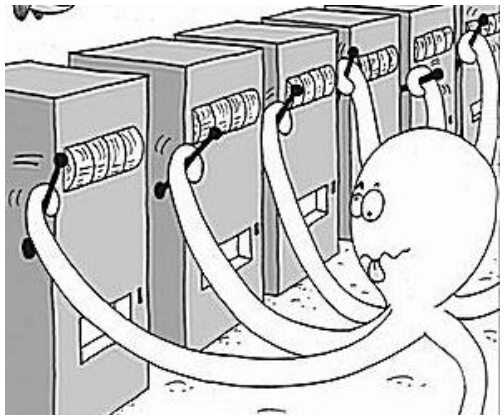
$$\mathrm{Regret}_T(\boldsymbol{x}^*) \equiv \langle \boldsymbol{x}^*, \boldsymbol{\mu} \rangle T - \mathbb{E}^\pi \sum_{t=1}^T \sum_i r_i(t) \mathbb{1}(I_t = i), \ \boldsymbol{x}^* \in \Omega(\boldsymbol{\lambda})$$

$$\text{s.t.} \quad \liminf_{T\to\infty} T^{-1} \sum_{t=1}^T \mathbb{E} r_i(t) \mathbb{1}(I_t = i) \geq \lambda_i, \ \forall i \in [N].$$

- The degree of violation is captured using *violation regret* metric

$$\mathbb{V}(T) = \max_i \mathbb{E}\Big[\sum_{t=1}^T (\lambda_i - r_i(t) x_i(t))\Big]$$

Recap: Stochastic Multi-Armed Bandits



- There is a set of N arms. A policy pulls one arm at a time.
- The reward $r_i(t)$ of the i^{th} arm is sampled i.i.d. from a distribution in $[0, 1]$ with an unknown mean $\mu_i, i \in [N]$.
- The objective is to sequentially pull the arms $\{I_t\}_{t=1}^{T}$ so as to obtain a sublinear pseudo-regret:

$$\text{Regret}_T \equiv T \max_i \mu_i - \mathbb{E} \sum_{t=1}^{T} \sum_i r_i(t) \mathbb{1}(I_t = i)$$

Difference with Stochastic MAB Problem

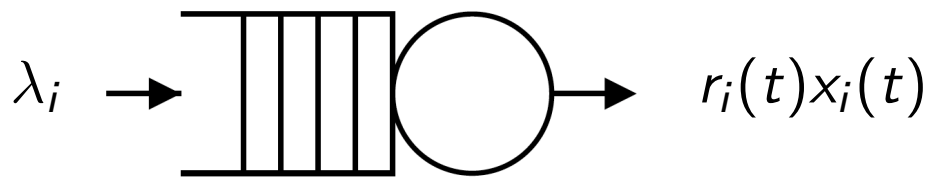
- The standard MAB problem aims to pull the most rewarding arms the maximum number of times while ignoring the rest.
- This strategy does not work in the `Fair Bandit` problem as the suboptimal arms receive negligible rewards
  - Maximizing rewards per round is not the right objective!
  - Instead of the best arm, the policy needs to learn the the best distribution of arms
- Note that, when $\vec{\lambda} = \mathbf{0}$, the `Fair Bandit` problem reduces to the usual stochastic MAB problem.

Main Result: Blackbox reduction to the adversarial MAB problem with simultaneously $O(T^{3/4})$ regret and cumulative violations.

Algorithmic ideas for BanditQ (Full-information setup)

- We keep track of the current rate of reward accruals through queueing recursions

$$Q_i(t) = \big(Q_i(t-1) + \underbrace{\lambda_i}_{\text{target rate}} - \underbrace{r_i(t)x_i(t)}_{\text{expected rate for round } t}\big)^+, Q_i(0) = 0, \forall i$$



- Intuitively, stabilizing the queues implies that the long-term target rates are satisfied
  - For us, only the rate stability of the queues will suffice, *i.e.*, $T^{-1}\mathbb{E}Q_i(T) \to 0, \forall i$.

Incorporating the Backlogs into Rewards

- To minimize the regret, we add a (scaled) reward of each arm to the queues and define a overall reward vector

$$r_i'(t) = (Q_i(t-1) + V) r_i(t), \ \forall i \in [N].$$

$V = \Theta(\sqrt{T})$ depends only on the horizon-length.

- This defines an instance of a MAB problem which we solve using

(Online Gradient Ascent Policy) The `BanditQ` policy selects the arms by running the OGA policy with adaptive step sizes:

$$\boldsymbol{x}(t) \leftarrow \mathrm{Proj}_{\Delta^N} \left[\boldsymbol{x}(t-1) + \frac{\boldsymbol{r}'(t-1)}{\sqrt{2 \sum_{\tau=1}^{t-1} ||\boldsymbol{r}'(\tau)||_2^2}} \right].$$

Second-Order Regret bound of OGA

Theorem 1 (Orabona (2019))

Let $\{\boldsymbol{r}'_t\}_{t\geq 1}$ be the sequence of reward vectors (which could be adversarially chosen). Then the above `OGA` policy achieves the following regret bound:

$$\text{Regret}_T \leq \sqrt{2\sum_{t=1}^{T} ||\boldsymbol{r}'_t||_2^2}.$$

Upshot: The regret bound scales with the norm of the reward vectors.

Analytical challenges

Regret analysis of the BanditQ policy is technically challenging.

- The rewards are no longer iid due to the presence of the recursively defined queue variables $\boldsymbol{Q}(t)$
  - Solution: Take recourse to results on adversarial bandits
- The rewards are no longer bounded as the queues could grow unboundedly
  - Solution: Make use of data-dependent second-order regret bounds
- The dynamics of the queueing process $\boldsymbol{Q}(t)$ depends on the online learning policy, whose actions are tightly coupled with $\boldsymbol{Q}(t)$
  - Solution: Joint analysis of the online learning policy and the queueing dynamics using a Lyapunov (potential) function

Proof technique : Solving an Integral Inequality

- Step 1: Define the quadratic potential function

$$\Phi(t) = \sum_i Q_i^2(t).$$

- Step 2: Using the queueing dynamics and the OGA regret bound, derive a non-linear recursive inequality in queue lengths for each round $t \geq 1$:

$$\sum_i \mathbb{E}Q_i^2(t) + 2V\mathrm{Regret}(\boldsymbol{x}^*) \leq 2t + 4\sqrt{2\sum_{\tau=1}^{t}\sum_i \mathbb{E}Q_i^2(\tau)} + 4V\sqrt{2Nt}.$$

- Step 3: Show that with $V = \Theta(\sqrt{T})$, the above inequality implies that $\{\mathbb{E}Q_i(t)\}_{t\geq 1}$ must grow sub-linearly ($\sim O(N^{1/4}T^{3/4})$), which, in turn, implies $\mathrm{Regret}(\boldsymbol{x}^*) \sim (NT)^{3/4}$.

Corollaries

- By playing with the previous self-bounding inequality further, we can get *almost* optimal regret bounds:

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{Regret}(\boldsymbol{x}^*) = O(\sqrt{NT}).$$

- If we only care about meeting the target rates (i.e., disregard the rewards), we can get $\mathbb{E}Q_i(t) \leq 8\sqrt{Nt}, \forall t$, by setting $V = 0$.

BanditQ in the Bandit feedback setting

- With the bandit feedback, only the reward of the selected arm (*i.e.*, $r_{I_t}(t)$) is revealed.
- This leads to a slightly modified queueing recursion:

$$Q_i(t) = \left(Q_i(t-1) + \lambda_i - r_i(t)X_i(t)\right)^+,$$

  where $X_i(t) = \mathbb{1}(I_t = i)$.
- The OGA policy is replaced with an adversarial MAB policy [Putta and Agarwal 2022], which enjoys a second-order regret bound

$$\text{Regret} = \tilde{O}\left(\sqrt{N\sum_{t=1}^{T}||\boldsymbol{r}_t'||^2 + \max_{t\in[T]}||r_t'||_\infty\sqrt{NT}}.\right).$$

- The proof of the regret bound involves a similar self-bounding inequality as in the full-information case.

Results

- The second term of the bound, involving a max norm over the entire time horizon, requires a careful Martingale-based analysis

The following is our main result:

Theorem 2 (BanditQ Regret Bounds)
With $V = \sqrt{T}$, the BanditQ policy with bandit feedback achieves

$$\mathrm{Regret}(\boldsymbol{x}^*) = \tilde{O}(N^{5/4}T^{3/4}),\ \mathbb{V}(T) = O(N^{1/4}T^{3/4}).$$

Generalization: OCO with Instantaneous Constraints

- On round t, algorithm chooses an action x_t
- Then the adversary reveals a convex cost function f_t AND a constraint $g_t(x) \leq 0$, with convex g_t
- The goal is to design a *parameter-free* policy to achieve both sublinear regret and cumulative constraint violation penalty
- The problem was conjectured to be impossible to solve

> Our result: We design a *parameter-free* online policy with $O(T^{\frac{3}{4}})$ regret and $O(T^{\frac{3}{4}})$ cumulative constraint violation.

- Better bounds are possible with more assumptions/structures.

Experiments: Full-Information Setup

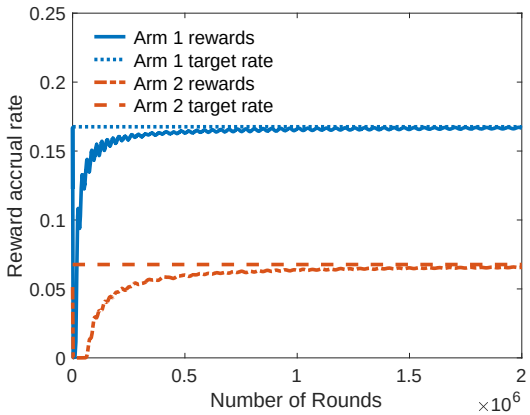


Figure: Reward accrual rates

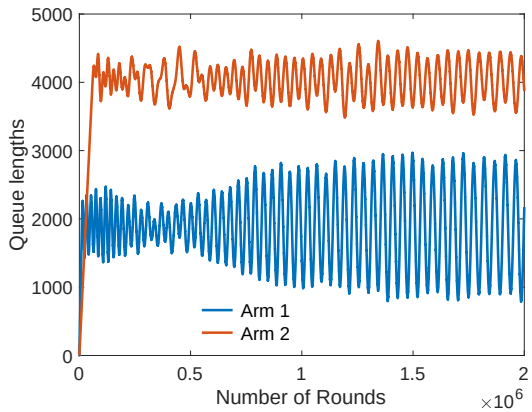


Figure: Queue lengths under `BanditQ`

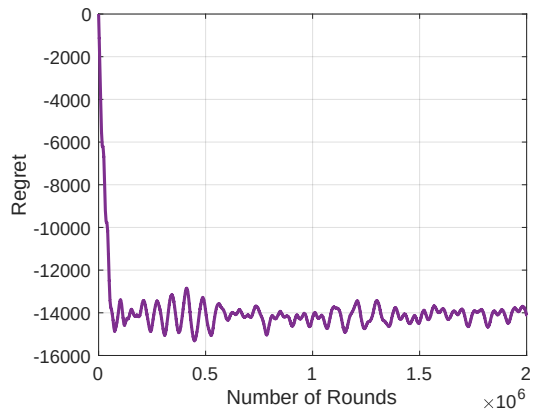


Figure: Regret of `BanditQ`

Experiments: Bandit Information Setup

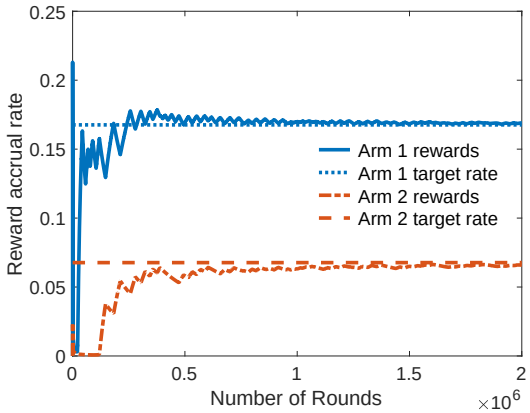


Figure: Reward accrual rates

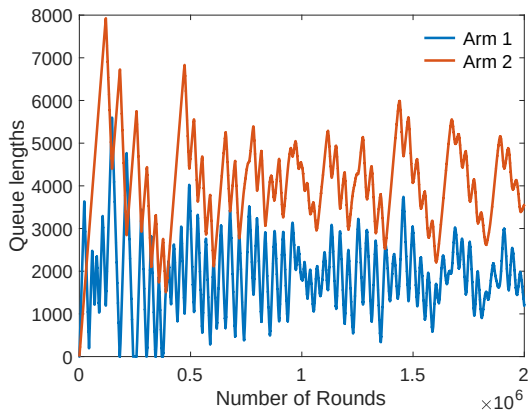


Figure: Queue lengths under `BanditQ`

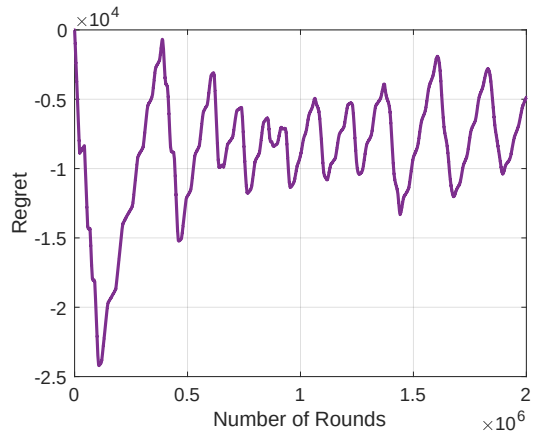


Figure: Regret of `BanditQ`

Open Problems

- Are the regret and constraint violation rates achieved by `BanditQ` optimal?
- How to extend the algorithm to the much more challenging `Bandit Convex Optimization` (BCO) Setup?
- Can we learn non-linear utility function while simultaneously satisfying the long-term constraints?

References

1. Sinha, Abhishek. "`BanditQ`: Fair Multi-Armed Bandits with Guaranteed Rewards per Arm." arXiv preprint arXiv:2304.05219 (2023).

2. Sinha, Abhishek, Joshi, A., Bhattacharjee, R., Musco, C., Hajiesmaili, M. (2023). No-regret Algorithms for Fair Resource Allocation. arXiv preprint arXiv:2303.06396.