Better Best-of-Both-Worlds Bounds for Bandits with Switching Costs

Guy Azov[1], Idan Amir[1], Tomer Koren[1,2], Roi Livni[1]

[1] Tel-Aviv University [2] Google

NeurIPS 2022

TEL AVIV UNIVERSITY

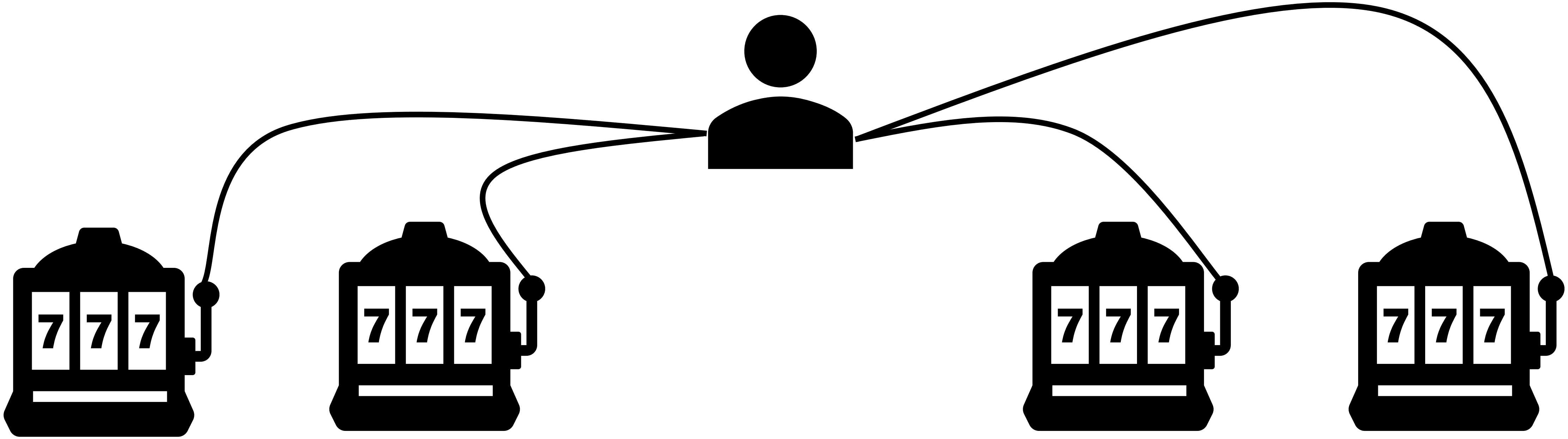
Google Research

Multi-Armed Bandits

Arms (actions) $\{1, \ldots, K\}$

At each time step $t = 1,2, \ldots, T$:

- A loss vector $\ell_t \in [0,1]^K$ is generated by the environment
- Player generates $p_t \in \Delta^K$ and samples $I_t \sim p_t$.
- Player incurs and observes loss ℓ_{t,I_t}.



Multi-Armed Bandits

- Adversarial (oblivious) regime - $\ell_1, \dots, \ell_t$ may be entirely arbitrary.
- Stochastically-constrained adversarial regime - $\mathbb{E}[\ell_{t,i} - \ell_{t,i^*}] = \Delta_i$
  - Generalizes the stochastic regime where losses are generated in an i.i.d manner.
  - For $K = 2$: $\Delta_i \triangleq \Delta$

Player's goal : minimize the pseudo-regret:

$$\boxed{\overline{\mathcal{R}_T} = \sum_{t\in[T]} \ell_{t,I_t} - \min_{i\in[K]} \sum_{t\in[T]} \mathbb{E}[\ell_{t,i}]}$$

If $\overline{\mathcal{R}_T} = \mathrm{o}(\mathrm{T})$ -> player is learning

Switching Cost

The player incurs an extra (switching) cost $\lambda > 0$ when she switches actions between rounds.

Switching cost pseudo-regret:

$$\overline{\mathcal{R}_T^{\lambda}} = \sum_{t\in[T]} \ell_{t,I_t} - \min_{i\in[K]} \sum_{t\in[T]} \mathbb{E}[\ell_{t,i}] + \boxed{\sum_{t\in[T]} \lambda \cdot (\mathbb{1}\{I_t \neq I_{t-1}\})}$$

In this **presentation** : $\lambda = 1, K = 2$

Best-of-Both-Worlds : Bandits with Switching Cost

Stochastic setting

Algorithms: BaSE (Gao et al, 2019)
Batched Arm Elimination (Esfandiari et al, 2021)

Optimal regret: $O\left(\frac{\ln(T)}{\Delta}\right)$

Adversarial setting

Algorithm: EXP3's variant (Arora et al, 2012)

Regret : $O\left(T^{2/3}\right)$

Lower Bound: $\widetilde{\Omega}\left(T^{2/3}\right)$ (Dekel et al, 2014)

Follow the Regularized Leader-based approach

Rouyer et al (2021) proposed *Tsallis-Switch* - a batched version of *Tsallis-INF* (Zimmert & Seldin, 2019).

Oblivious Adversarial Setting:

$$\mathbb{E}[\overline{\mathcal{R}_T^{\lambda=1}}] \leq O\left(T^{2/3}\right)$$

Tight

Stochastically Constrained Setting:

$$\mathbb{E}[\overline{\mathcal{R}_T^{\lambda=1}}] \leq O\left(\frac{T^{1/3} + \log T}{\Delta}\right)$$

Tight ?

Can we do better?

Our Main Results

We designed an algorithm that obtain the following regret bounds :

- Oblivious Adversarial Setting:

$$\mathbb{E}[\overline{\mathcal{R}_T^{\lambda=1}}] \leq O\left(T^{2/3}\right)$$

- Stochastically Constrained Setting:

$$\mathbb{E}[\overline{\mathcal{R}_T^{\lambda=1}}] \leq O\left(\min\left\{\left(\frac{logT}{\Delta^2} + \frac{logT}{\Delta}\right), T^{2/3}\right\}\right)$$

$\boxed{\text{Potentially improves by a factor of } \tilde{O}(\mathrm{T}^{1/3}\Delta)}$

Algorithm

Key observation:

Under the stochastically constrained setting, the number of switches, S, is bounded by:

$$S \leq O\left(\frac{\overline{\mathcal{R}_T}}{\Delta}\right)$$

Switch Tsallis, Switch!

- Start playing the original *Tsallis-INF* (Zimmert & Seldin, 2019).
- If $S \geq O\left(T^{2/3}\right)$: *If we made too many switches – we are in the adversarial regime*
  - Play *Tsallis-INF* over blocks of size $O\left(T^{1/3}\right)$

Can we do even better?

Our Main Results

Lower Bound

Given a randomized player in the multi-armed bandits game with $\mathbb{E}[\overline{\mathcal{R}_T^{\lambda=1}}] \leq O\left(T^{2/3}\right)$ under the adversarial regime, for every $\Delta > 0$ there exists a sequence of stochastically constrained losses $\ell_1, \dots, \ell_t$ with a minimal gap Δ, such that the player incurs:

$$\overline{\mathcal{R}_T^{\lambda=1}} = \tilde{\Omega}\left(\min\left\{\frac{1}{\Delta^2}, T^{2/3}\right\}\right)$$

For $K > 2$ - there is an interesting gap (check the paper for more information).

Takeaways

We presented **Switch Tsallis, Switch!**

- Simple and effective algorithm
- Achieve the minimax regret in the *oblivious adversarial* setting (up to logarithmic factors) of $O\left(T^{2/3}\right)$.
- In the *stochastically constrained* setting obtain the upper bound of $O\left(\min\left\{\left(\frac{logT}{\Delta^2}+\frac{logT}{\Delta}\right),T^{2/3}\right\}\right)$.

Potentially improves by a factor of $\tilde{O}(\mathrm{T}^{1/3}\Delta)$ **.**

We provided a lower bound which demonstrates that

$$\tilde{\Omega}\left(\min\left\{\frac{1}{\Delta^2},T^{2/3}\right\}\right)$$

Switching cost pseudo regret Is unavoidable in the stochastically-constrained case for algorithms with $O\left(T^{2/3}\right)$ worst-case switching cost pseudo regret.

For $K>2$ ***- there is an interesting gap between the bounds.***

Thank You!