Tsallis-INF for Decoupled Exploration and Exploitation in Multi-armed Bandits

Chloé Rouyer, Yevgeny Seldin

DIKU, University of Copenhagen

Contact Information:
Chloé Rouyer
Email: `chloe@di.ku.dk`

Introduction

The multi-armed bandit problem is a central a framework for studying the exploration-exploitation trade-off. In the multi-armed bandit game a player repeatedly chooses actions from a set of K actions and observes and suffers the loss of the selected action. The losses may be generated adversarially or stochastically, depending on problem setup.

The goal of the learner is to find an action selection strategy minimizing the regret, which is the difference between the cumulative loss of the player and of the best fixed action in hindsight.

We focus on a variation of the multi-armed bandit problem introduced by Avner et al. [2012], in which at each round the learner is allowed to choose one action to play blindly and one action to observe without suffering its loss. The two actions are allowed, but not required to be different. Thus, exploration is decoupled from exploitation.

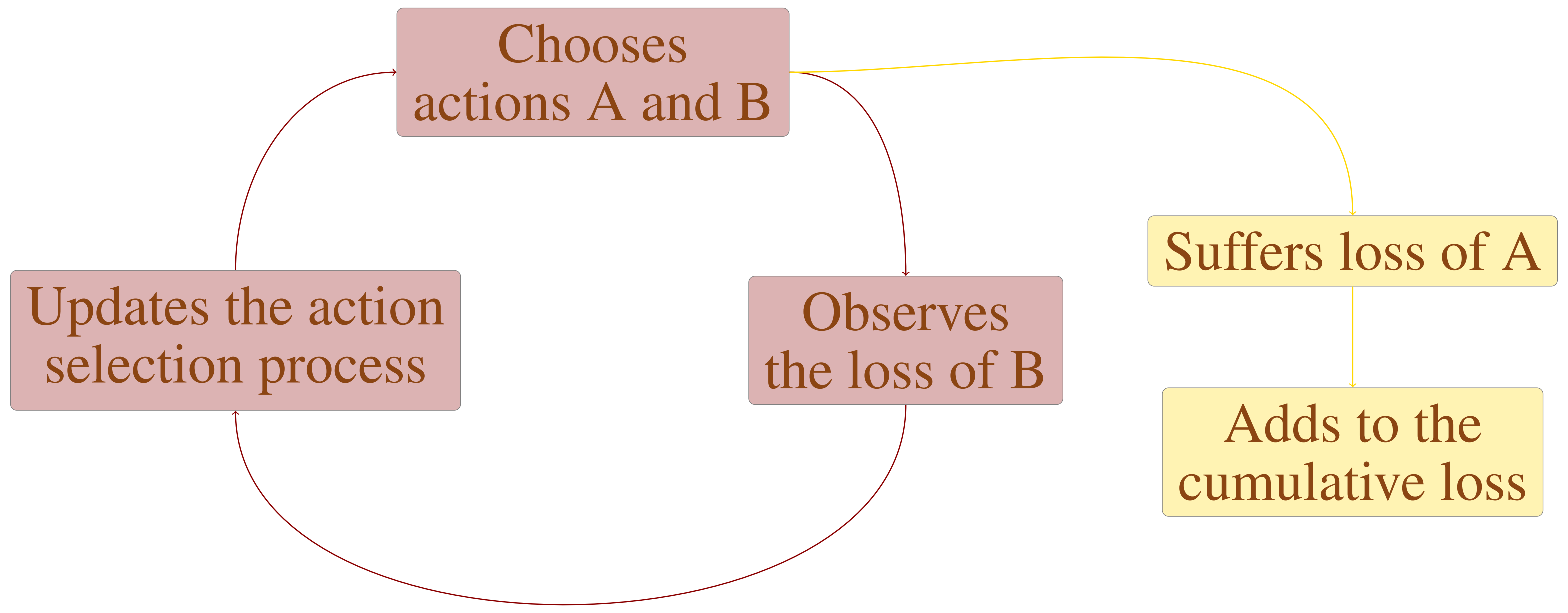
Chooses actions A and B
Suffers loss of A
Updates the action selection process
Observes the loss of B
Adds to the cumulative loss

Figure 1: Representation of one round of the process for the learner. Her goal is to minimize her cumulative loss. Action A should be picked in order to get a small loss, and action B to gather some useful information about an arm.

Problem Setting

We consider a repeated game with K arms. At each round $t = 1, 2, \dots$ of the game the environment picks a loss vector $\ell_t \in [0,1]^K$ and the learner picks an action A_t to exploit and an action B_t to explore. The two actions are allowed to be different, but may also be identical. Then the learner blindly suffers ℓ_{t,A_t} and observes ℓ_{t,B_t} without suffering its loss.

We want our algorithm to adapt to the following types of losses:

- The oblivious adversarial setting, where the environment chooses ℓ_t arbitrarily prior to the beginning of the game,
- and the stochastically constrained adversarial setting where the losses are drawn from distributions with fixed gaps, that we express in terms of the best arm i^*. I.e. for all i we have $\mathbb{E}\left[\ell_{t,i} - \ell_{t,i^*}\right] = \Delta_i$ independently of t. In the analysis we require that the best arm is unique.

Pseudo-regret

We measure the performance of an algorithm in terms of pseudo-regret:

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i}\right],$$

where $i^*_T = \arg\min_i \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i}\right]$ is the best action in hindsight.

In the stochastically constrained adversarial setting we rewrite the pseudo-regret in terms of the suboptimality gaps as

$$\mathcal{R}_T = \sum_{t=1}^T \sum_{i \neq i^*} \mathbb{E}\left[p_{t,i}\right] \Delta_i,$$

where $p_{t,i}$ is the probability that arm i is played at round t.

Algorithm

```
Algorithm 1: Decoupled-Tsallis-INF
Input: Learning rates $\eta_1 \geq \eta_2 \geq \dots > 0$.
Initialize: $\tilde{L}_0 = \mathbf{0}_K$
for $t = 1, 2, \dots$ do
    $p_t = \arg\min_{p \in \Delta^{K-1}} \left\{ \langle p, \tilde{L}_{t-1} \rangle - \frac{1}{\eta_t} \sum_{i=1}^K \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)} \right\}$
    Construct exploration distribution $q_t = \arg\min_{q_t} \sum_{i=1}^K \frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}$
    Sample $A_t$ according to $p_t$, play it and suffer $\ell_{t,A_t}$.
    Sample $B_t$ according to $q_t$ and observe $\ell_{t,B_t}$.
    $\forall i \in [K]: \quad \tilde{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\{B_t = i\}}{q_{t,i}} = \begin{cases} \frac{\ell_{t,i}}{q_{t,i}}, & \text{if } B_t = i, \\ 0, & \text{otherwise.} \end{cases}$
    $\forall i \in [K]: \quad \tilde{L}_t(i) = \tilde{L}_{t-1}(i) + \tilde{\ell}_{t,i}$.
end
```

We note that:

- The distribution p_t, from which we draw the action to exploit A_t is a typical FTRL with regularization by α-Tsallis entropy [Zimmert and Seldin, 2019].
- The distribution q_t, from which we draw the action to explore B_t is chosen to minimize our regret bound. It can be expressed in closed for solution as

$$\forall t \in [T], i \in [K], \quad q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{j=1}^K (p_{t,j})^{1-\alpha/2}}.$$

Results

The main result of our paper is the following.

Corollary 1. *For $\alpha = 2/3$ and $\eta_t = \frac{2K^{-1/6}}{\sqrt{t}}$ the regret of Decoupled-Tsallis-INF satisfies*

$$\mathcal{R}_T \leq 5\sqrt{KT} + 1,$$

in the adversarial regime and

$$\mathcal{R}_T \leq 100 \frac{K}{\Delta_{\min}} + 13\sqrt{K},$$

in the stochastically constrained adversarial regime with a unique best arm i^*. ***The two regret bounds hold simultaneously and with no need in prior knowledge of the regime.***

Intuition for $\alpha = 2/3$

In the regular Multi-armed bandits problem, Zimmert and Seldin [2019] achieve best-of-both world optimal results using Tsallis-Inf with $\alpha = 1/2$. In the stochastic regime, this gives a minimax optimal bound in $\sum_{i \neq i^*} \frac{\log T}{\Delta_i})$ and any choice of a different α seem to lead to exploring or exploiting too much, resulting in a bound which is polynomial in T.

In our problem, because the loss estimates are constructed based on q_t, it is possible to choose a value for α to be larger so p_t puts more more weight on the actions that have the best cumulative loss estimates so far, while the distribution q_t stays more spread out, allowing to still have sufficient exploration. In other words, **decoupling exploration and exploitation allows us to both explore and exploit more, which allows us to bypass the lower bound in $\Omega(\sum_{i \neq i^*} \frac{\log T}{\Delta_i})$ that regular Multi-armed bandits have, and achieve a time independent bound without requiring a larger amount of observations.**

Conclusions

- We derived bounds for our algorithm that are optimal up to constants in the adversarial regime, and time independent in the stochastically constrained adversarial regime.
- The scaling of the lower bound in the stochastically constrained adversarial regime is still unknown, so we may be still suboptimal in this regime by a factor $\frac{K}{\log K}$.

Forthcoming Research

There are two directions for following work. First, deriving a proper lower bound in the stochastically constrained adversarial regime, which we conjecture to be $\Omega(\sum_{i \neq i^*} \frac{1}{\Delta_i})$, and then closing the gap if needed. It should also be possible to work on getting tighter constants in the bound, following the work of Zimmert and Seldin [2019] to reduce the variance of the unbiased estimates. Another direction is to generalize the results of this work to different settings, in particular in prediction with limited feedback and pure exploration.

References

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. `https://arxiv.org/pdf/1807.07623`.

Links

The full version of this work has been published in the proceedings of COLT 2020. A short video presenting this work is also available.