DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

An Optimal Algorithm for Stochastic and Adversarial Bandits

Julian Zimmert, Yevgeny Seldin
{zimmert, seldin}@di.ku.dk

Introduction

Stochastic (i.i.d.) and adversarial multi-armed bandits are two fundamental sequential decision making problems in online learning.

When prior information about the nature of environment is available, it is possible to achieve $\mathcal{O}(\sum_{i:\Delta_i>0}\frac{\log(T)}{\Delta_i})$ pseudo-regret in the stochastic case and $\mathcal{O}(\sqrt{KT})$ pseudo-regret in the adversarial case, and both results match the respective lower bounds up to constants.

The challenge in recent years has been to achieve the optimal regret rates without prior knowledge about the nature of the problem. The question of the existence of a universal trade-off preventing optimality in both worlds simultaneously has remained open for a while. We give a concluding answer, showing that a very simple algorithm can be optimal.

Problem setting

- At time $t = 1,\dots,(T)$ the agent chooses an arm $I_t \in \{1,\dots,K\}$
- The environments, oblivious to the agent's action, picks a loss vector $\ell_t \in [0,1]^K$
- The agent observes and suffers only the loss ℓ_{tI_t}
- The agent tries to minimize its pseudo-regret:

$$\overline{Reg}_T = \max_{i\in\{1,\dots,K\}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{tI_t} - \ell_{ti}\right]$$

Stochastically constrained adversary

- Losses satisfy $\mathbb{E}\left[\ell_{ti} - \ell_{ti^*}\right] = \Delta_i > 0$ for all times t
- Absolute mean $\mathbb{E}\left[\ell_{ti^*}\right]$ can be chosen adversarially
- Generalization of stochastic i.i.d. bandits

Online Mirror Descent

Input: $(\Psi_t)_{t=1,2,\dots}$
Initialize: $\hat{L}_0 = \mathbf{0}_K$ (the zero vector of dimension K)
for $t = 1,\dots$ **do**
  choose $w_t = \nabla(\Psi_t + \mathcal{I}_{\Delta^K})^*(-\hat{L}_{t-1})$
  sample $I_t \sim w_t$
  observe ℓ_{tI_t}
  construct $\hat{\ell}_t = \frac{\ell_{tI_t}}{w_{tI_t}}\mathbf{e}_{I_t}$
  update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
end

Algorithm 1: Online Mirror Descent (OMD) for bandits

- $\mathcal{I}_{\Delta^K}(w) = \begin{cases} \infty \text{ if } w \notin \Delta^K \\ 0 \text{ if } w \in \Delta^K \end{cases}$
- $\Psi_t(w) = -\sum_{i=1}^{K} \frac{w_i^\alpha}{\alpha\eta_{ti}}$

What makes the problem so hard?

- The probability of playing sub-optimal arms depends on the loss difference $L_{ti} - L_{ti^*}$
- The difference of loss estimator is in expectation $\mathbb{E}\left[L_{ti} - L_{ti^*}\right] = \Delta_i t$
- Sub-optimal arms cannot be played with probability higher than $\mathcal{O}(\frac{1}{\Delta_i^2 t})$
- The variance of loss estimators is then of order $\Omega(\Delta_i^2 t^2)$
- The signal is of the same magnitude as the standard deviation

Concentration arguments cannot work!

Novel proof

We refine the standard OMD upper bound

$$\overline{Reg}_T \le \square \sum_{t=1}^{T}\sum_{i\neq i^*}\left(\frac{\mathbb{E}\left[w_{ti}\right]^{1-\alpha}}{t^\alpha} + \frac{\mathbb{E}\left[w_{ti}\right]^{\alpha}}{t^{1-\alpha}}\right)$$

We use the explicit form of the regret

$$\overline{Reg}_T = \sum_{t=1}^{t}\sum_{i\neq i^*}\Delta_i\mathbb{E}\left[w_{ti}\right]$$

And take the worst case $\mathbb{E}\left[w_{ti}\right]$ that still satisfies the regret bound (**self-bounding proof**)

$$\max_{\omega_1,\dots\omega_T\in\Delta^K} \square\sum_{t=1}^{T}\sum_{i\neq i^*}\left(\frac{\omega_{ti}^{1-\alpha}}{t^\alpha} + \frac{\omega_{ti}^{\alpha}}{t^{1-\alpha}}\right)$$

$$\text{s.t.} \sum_{t=1}^{t}\sum_{i\neq i^*}\Delta_i\omega_{ti} \le \square\sum_{t=1}^{T}\sum_{i\neq i^*}\left(\frac{\omega_{ti}^{1-\alpha}}{t^\alpha} + \frac{\omega_{ti}^{\alpha}}{t^{1-\alpha}}\right)$$

Choosing α

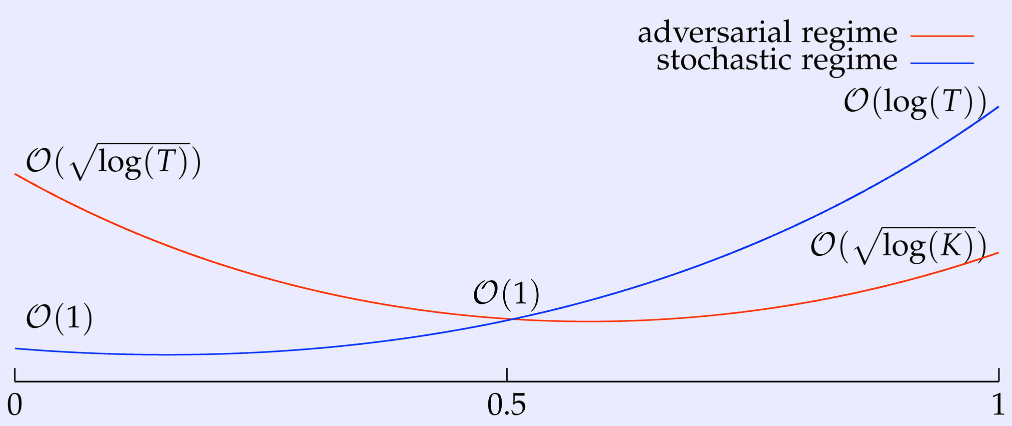
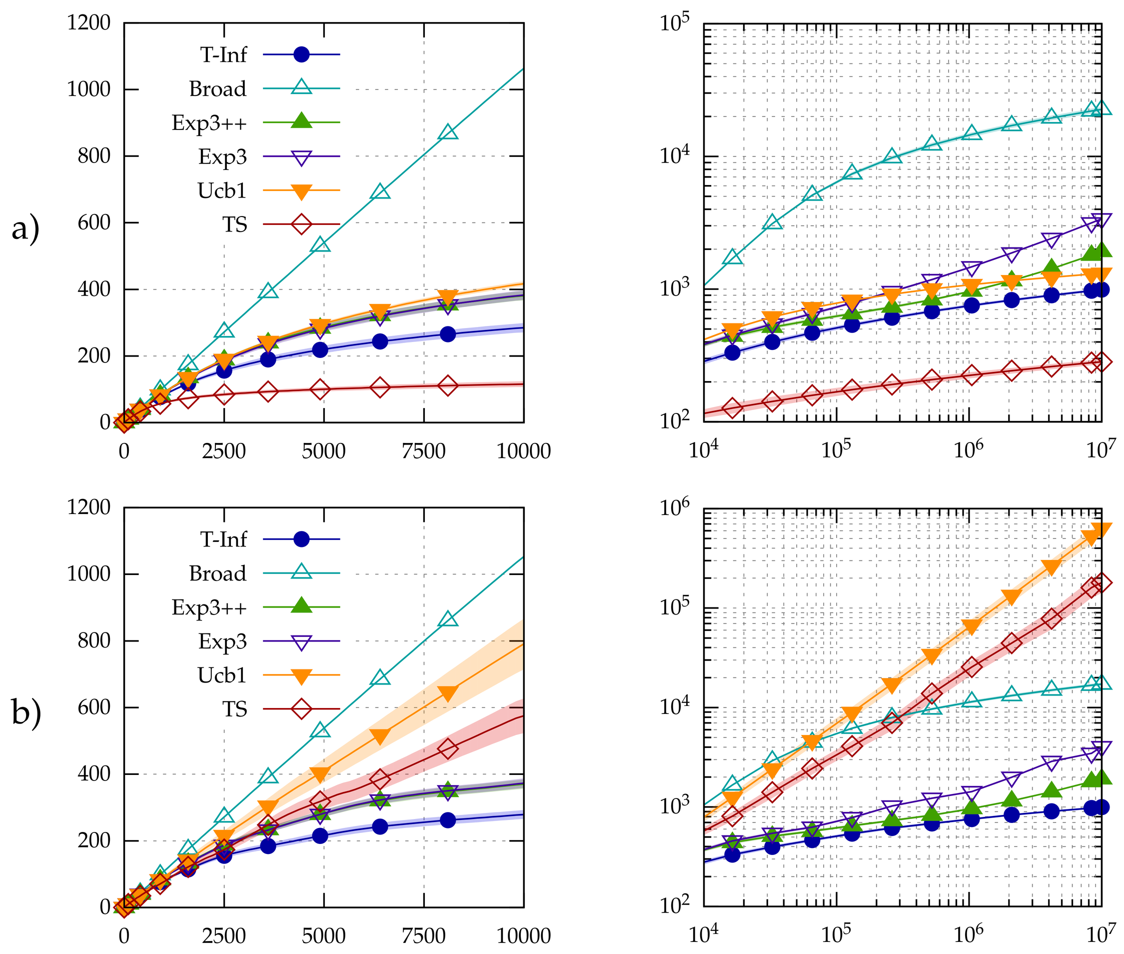


Figure 1: (Upper bound)/(Lower bound) for different values of α.

- $\alpha = \frac{1}{2}$ is superior to EXP3 ($\alpha = 1$) and LogBarrier ($\alpha = 0$)
- Only for $\alpha = \frac{1}{2}$ is the learning rate identical for stochastic and adversarial environments

Empirical evaluation



Comparison of several bandit algorithms with $K = 8$ and $\Delta = 1/8$ under a) stochastic and b) stochastically constrained adversary regime. The left side is in linear scale and the right is in log-log scale.

Application to utility-based dueling bandits

In dueling bandits, the agent selects two actions I_t, J_t every step and receives the result of a "duel".

There are utilities $u_t \in [0,1]^K$ associated with all arms and the probability of arm i winning against arm j is $\frac{1+u_{ti}-u_{tj}}{2}$. In the stochastic case, $u_t = u$ are constant.

The pseudo-regret is defined as

$$\overline{Reg}_T = \max_i \mathbb{E}\left[\sum_{t=1}^{T} 2u_i - u_{I_t} - u_{J_t}\right].$$

When using **sparring**, i.e. running independent algorithms to select I_t and J_t that receive $\mathbb{I}\{I_t/J_t\text{wins}\}$ as a loss, then the pseudo regret is the sum of the individual regrets.

In the stochastic case, the sparring problem is a stochastically constrained adversary, hence our results for MAB apply.

Therefore, we achieve **optimality in both worlds**:

$$\overline{Reg}_T \le \begin{cases} \mathcal{O}\left(\sqrt{KT}\right) & \text{for adversarial environments} \\ \mathcal{O}\left(\sum_{i\neq i^*}\frac{\log(T)}{u_{i^*}-u_i}\right) & \text{for stochastic environments} \end{cases}.$$