DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

An Optimal Algorithm for Stochastic and Adversarial Bandits

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

Summary

We provide an algorithm that achieves the optimal (up to constants) finite time regret in both adversarial and stochastic multi-armed bandits without prior knowledge of the regime and time horizon. The result provides a negative answer to the open problem of whether extra price has to be paid for the lack of information about the adversariality/stochasticity of the environment. In addition, the proposed algorithm enjoys improved regret guarantees in two intermediate regimes: the moderately contaminated stochastic regime defined by Seldin and Slivkins [2014] and the stochastically constrained adversary studied by Wei and Luo [2018].

Multi-Armed Bandit (MAB)

Stochastic MABs

– Stochastic MABs [Thompson, 1933, Robbins, 1952, Lai and Robbins, 1985] are **sequential decision problems**:

Initialisation: Set of arms $\{1,\dots,K\}$ with unknown distributions p_i over $[0,1]$
Game:
for $t = 1,\dots(,T)$ **do**
Select arm I_t.
Sample loss $\ell_t \sim p_{I_t}$.
Observe and suffer loss ℓ_t.
end for
Target: Minimize $\sum_{s=1}^{T}\ell_s$.

– The **performance** of an algorithm is measured in terms of **simple regret**:

$$\mathbb{E}\left[\overline{\mathrm{Reg}}_T\right] := \mathbb{E}\left[\sum_{t=1}^{T}\ell_t\right] - T\cdot\min_i \mathbb{E}\left[\ell_t|I_t=i\right]$$

– The **difficulty** of a stochastic MAB depends on the **gaps**:

$$\Delta_i = \mathbb{E}\left[\ell_t|I_t=i\right] - \min_j \mathbb{E}\left[\ell_t|I_t=j\right]$$

– The **lower bound** for any consistent algorithm is:

$$\mathbb{E}\left[\overline{\mathrm{Reg}}_T\right] \geq \Omega\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$$

– the lower bound is **matched** by an **upper bound** for algorithms such as UCB:

$$\mathbb{E}\left[\overline{\mathrm{Reg}}_T\right] \leq \mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$$

Adversarial MAB

– Adversarial MABs [Auer et al., 2002] extend bandits to non-stochastic environments.

Initialisation: Set of arms $\{1,\dots,K\}$.
Game:
for $t = 1,\dots(,T)$ **do**
Adversary: Select hidden vector $\ell_t \in [0,1]^K$
Agent: Select arm I_t.
Observe and suffer loss ℓ_{t,I_t}.
end for
Target: Minimize $\sum_{s=1}^{T}\ell_s$.

– The **performance** of an algorithm is measured in terms of **expected regret**:

$$\mathbb{E}\left[\mathrm{Reg}_T\right] := \mathbb{E}\left[\sum_{t=1}^{T}\ell_t - \min_i\sum_{t=1}^{T}\ell_{t,i}\right]$$

– this is always larger than the simple regret:

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \geq \mathbb{E}\left[\overline{\mathrm{Reg}}_T\right]$$

– The **lower bound** for any consistent algorithm is:

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \geq \Omega\left(\sqrt{KT}\right)$$

– the lower bound is **matched** by an **upper bound** for algorithms such as INF:

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \leq \mathcal{O}\left(\sqrt{KT}\right)$$

Problem

Motivation

– in many real world applications, it is unclear if the problem is fully stochastic

– the worst case guarantee for adv. MABs is significantly worse than for stoch. MABs: $\log(T) \ll \sqrt{T}$

– the algorithms achieving optimality in one regime might not be good for the other

Question

Can the same algorithm achieve optimality in both regimes without knowing in which regime it operates?

Previous results

Table 1: Upper bounds for previous algorithms.

Algorithm	Stoch.	Adv.
UCB	$\mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$	T
INF	$\mathcal{O}\left(\sqrt{KT}\right)$	$\mathcal{O}\left(\sqrt{KT}\right)$
EXP++	$\mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)^2}{\Delta_i}\right)$	$\mathcal{O}\left(\sqrt{K\log(K)T}\right)$
BROAD-OMD	$\mathcal{O}\left(K\min_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$	$\mathcal{O}\left(\sqrt{KT\log(T)}\right)$
SAPO*	$\mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$	$\mathcal{O}\left(\sqrt{KT\log(T)}\right)^*$

*requires knowledge of the time horizon T or additional $\log(T)$ on either side

– all algorithms have at least an extra $\log(T)$ term on one of the sides

– it is impossible to have $\mathrm{Reg}_T \leq \mathcal{O}\left(\sqrt{KT}\right)$ with high probability in the adversarial regime if $\mathbb{E}\left[\overline{\mathrm{Reg}}_T\right] \leq \mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$ holds in the stochastic regime [Auer and Chiang, 2016]

– it is impossible to have optimal performance for stochastic and adversarial bandits if we only care about identifying the best arm with the highest probability after T rounds [Abbasi-Yadkori et al., 2018]

Solution

Yes it is possible. The following algorithm:

Algorithm 1: Tsallis Online Mirror Descent.
Initialisation: $L_0 = \mathbf{0}_K$
for $t = 1,\dots$ **do**
choose $w_t = \arg\max_{w\in\Delta_K}\left\{-\langle w, L\rangle + \sum_i\sqrt{w_i t}\right\}$
sample $I_t \sim w_t$
construct $\hat{\ell}_t : \hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}\{I_t=i\}}{w_{t,i}}$
update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
end for

achieves a regret of

Table 2: Upper bounds for TOMD.

Algorithm	Stoch.	Adv.
TOMD	$\mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$	$\mathcal{O}\left(\sqrt{KT}\right)$

Proof

See our paper https://arxiv.org/abs/1807.07623

References

Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on Learning Theory*, 2018.

P. Auer and C.-K. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295, 2014.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. *arXiv preprint arXiv:1801.03265*, 2018.