Autoregressive Bandits

Francesco Bacchiocchi[*1], Gianmarco Genalti[*1], Davide Maran[*1], Marco Mussi[*1], Nicola Gatti[1], Marcello Restelli[1] and Alberto Maria Metelli[1]

[1] Politecnico di Milano [*] Equal Contribution

RL3 Meeting, Milano. 9th February 2023

In the customary **Multi-Armed Bandit** framework, we consider a problem where:

- We have K arms, each representing an action
- The actions are independent each other
- The effect of the actions lasts for one time step only
- There is no notion of state

- At every time step we see a noisy realization of the expected value of the action we perform:

$$x_t = \mu_{a_t} + \eta_t$$

where η_t is an i.i.d. zero-mean noise

- The goal is to minimize the **regret**:

$$R_T = T\mu_{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} x_t\right]$$

where μ_{a^*} is the expected value of the optimal action

Several works manages bandit with **complex feedbacks**, like:

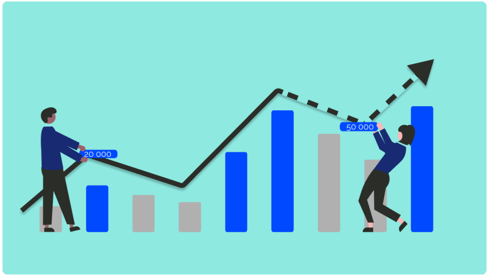
- Bandit with Delayed feedbacks (Pike-Burke et al., 2018)
- Restless and Markov Bandits (Ortner et al., 2012)
- Rested Bandits (Tekin and Liu, 2012; Levine et al., 2017)

In this work, we consider bandits in which the **actions of the previous time steps also influences the future rewards**

We consider that the previous states are a conseguence of the actions (in a **rested** fashion) and influence the next rewards thanks to an **unknown autoregressive** dynamic

Autoregressive Processes (AR) are present in a plenty of scenarios in the real world:

Stocks Market



Sales Forecasting

At every time $t \in \{1, \dots, T\}$, we select an action a_t and we receive a reward x_t:

$$\underbrace{x_t}_{\text{Reward at time } t} = \underbrace{\gamma_0(a_t)}_{\text{Exp. reward of } a_t} + \underbrace{\sum_{i=1}^{k} \gamma_i(a_t) x_{t-i}}_{\text{Contribution of the past}} + \underbrace{\xi_t}_{\text{Subgaussian Noise}}$$

where k is the **length of the history** we consider, and $(\gamma_i(a_t))_{i \in \{0,\dots,k\}}$ is an **unknown parameter vector** characterizing action a_t

The parameters $(\gamma_i(a))_{i \in \{0,\dots,k\}}$ fulfill the following conditions:

- (**Monotonicity**) $\gamma_i(a) \geq 0$ for every $a \in \mathcal{A}$, $i \in \{0, \dots, k\}$
- (**Stability**) $\Gamma := \max_{a \in \mathcal{A}} \sum_{i=1}^{k} \gamma_i(a) < 1$
- (**Boundedness**) $m := \max_{a \in \mathcal{A}} \gamma_0(a) < +\infty$

The goal is to minimize the **expected cumulative policy regret**:

$$R(\boldsymbol{\pi}, T) = J_T^* - J_T(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

where:

$$r_t = x_t^* - x_t$$

Theorem (Optimal Policy)

For every round $t \in \mathbb{N}$, *the optimal policy* $\pi_t^*(H_{t-1})$ *satisfies:*

$$\pi_t^*(H_{t-1}) \in \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \boldsymbol{z}_{t-1} \rangle$$

where $\boldsymbol{z}_{t-1} := (1, x_{t-1}, \dots, x_{t-k})^T$.

- Maximizing the **immediate expected** reward implies maximizing the **cumulative reward**
- We do not need to do action **planning**
- The **optimal policy is Markovian** w.r.t. the state representation $\boldsymbol{z}_{t-1}$

Algorithm 1: `AR-UCB`.

Input : Regularization parameter $\lambda > 0$, autoregressive order $k \in \mathbb{N}$, exploration coefficients $(\beta_{t-1})_{t \in [\![T]\!]}$

Initialize $t \leftarrow 1$, $\mathbf{V}_0(a) = \lambda \mathbf{I}_{k+1}$, $\mathbf{b}_0(a) = \mathbf{0}_{k+1}$, $\widehat{\gamma}_0(a) = \mathbf{0}_{k+1}$ for all $a \in \mathcal{A}$, $\mathbf{z}_0 = (1, 0, \dots, 0)^T$

for $t \in [\![T]\!]$ **do**

- Compute $a_t \in \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_t(a) := \langle \widehat{\gamma}_{t-1}(a), \mathbf{z}_{t-1} \rangle + \beta_{t-1}(a) \, \|\mathbf{z}_{t-1}(a)\|_{\mathbf{V}_{t-1}(a)^{-1}}$
- Play action a_t and observe $x_t = \langle \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle + \xi_t$
- Update for all $a \in \mathcal{A}$:

  $\mathbf{V}_t(a) = \mathbf{V}_{t-1}(a) + \mathbf{z}_{t-1} \mathbf{z}_{t-1}^T \mathbb{1}_{\{a = a_t\}}$

  $\mathbf{b}_t(a) = \mathbf{b}_{t-1}(a) + \mathbf{z}_{t-1} x_t \mathbb{1}_{\{a = a_t\}}$
- Compute $\widehat{\gamma}_t(a) = \mathbf{V}_t(a)^{-1} \mathbf{b}_t(a)$
- Update $\mathbf{z}_t = (1, x_t, \dots, x_{t-k+1})^T$
- $t \leftarrow t + 1$

end

Theorem

Let $\delta = (2T)^{-1}$. AR-UCB suffers a cumulative expected policy regret bounded by (highlighting the dependence on m, σ, k, Γ, n, and T only):

$$R(\textit{AR-UCB}, T) \leq \widetilde{\mathcal{O}}\left(\frac{(m^2 + \sigma)(k+1)^{3/2}\sqrt{nT}}{(1-\Gamma)^2} \right).$$

We perform an experimental validation of AR-UCB:

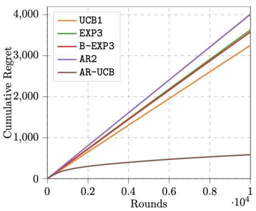
- In comparison with several bandit baselines
- By studying its sensitivity w.r.t. key parameters

We consider as baselines:

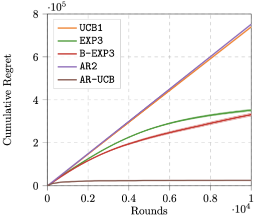
- `UCB1` (Auer et al., 2002a)
- `EXP3` (Auer et al., 1995, 2002b)
- `B-EXP3` (Dekel et al., 2012)
- `AR2` (Chen et al., 2021)

Experimental Validation
Comparison with the other baselines

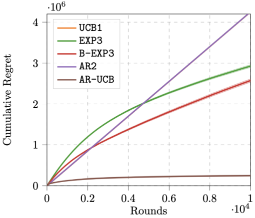
Setting	k	m	σ
	Parameters		
A	2	1	0.75
B	4	20	1.5
C	4	920	10



Setting A.



Setting B.



Setting C.

We recall that the optimal policy $\pi_t^*(H_{t-1})$ for the ARB setting is:

$$\pi_t^*(H_{t-1}) \in \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \boldsymbol{z}_{t-1} \rangle$$

where $\boldsymbol{z}_{t-1} := (1, x_{t-1}, \dots, x_{t-k})^T$.

The optimal policy, when **no noise** is involved, is **constant** and corresponds to playing the action $a^+ \in \mathcal{A}$:

$$a^+ \in \arg\max_{a \in \mathcal{A}} \frac{\gamma_0(a)}{1 - \sum_{i=1}^k \gamma_i(a)}$$

Setting

- $k = 2$
- $T = 10000$
- $|\mathcal{A}| = 2$ (two actions, a_1 and a_2)

$$\gamma(a_1) = (1,\ \rho,\ 0)^T$$
$$\gamma(a_2) = (1,\ 0,\ \rho - \epsilon)^T$$

where $\rho = 0.5$ and $\epsilon = 0.02$

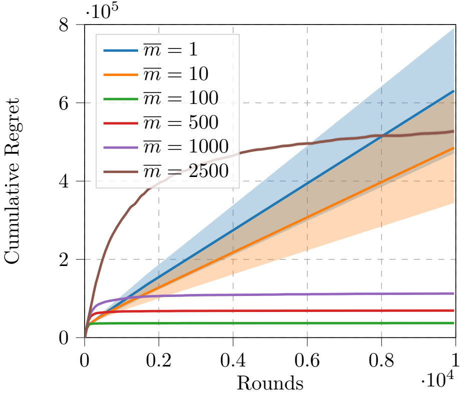
We study the **cumulative regret** at **different values of noise** $\sigma \in \{0,\ 0.1,\ 0.5,\ 1.0,\ 2.0\}$

σ	`Stochastic`	`Deterministic`
0	**19994 (0)**	**19994 (0)**
0.1	**20167 (0.20)**	19998 (2.04)
0.5	**22049 (1.02)**	20012 (1.02)
1.0	**24504 (2.04)**	20030 (2.04)
2.0	**29428 (4.09)**	20067 (4.08)

Setting

- $k = 4$
- $|\mathcal{A}| = 7$
- $m = 5000$
- $\gamma_i(a)$ sampled from uniform distribution

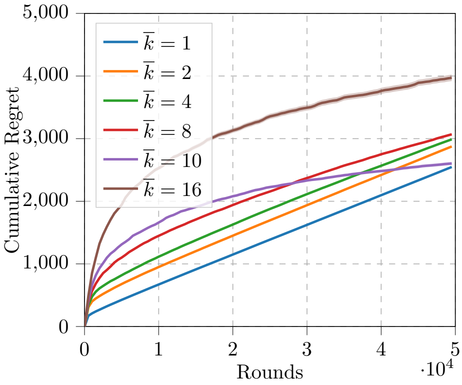
We want to minimize the **cumulative regret** over $T = 10000$ samples



Setting

- $k = 10$
- $|\mathcal{A}| = 7$
- $\gamma_i(a)$ sampled from uniform distribution

We want to minimize the **cumulative regret** over $T = 10000$ samples

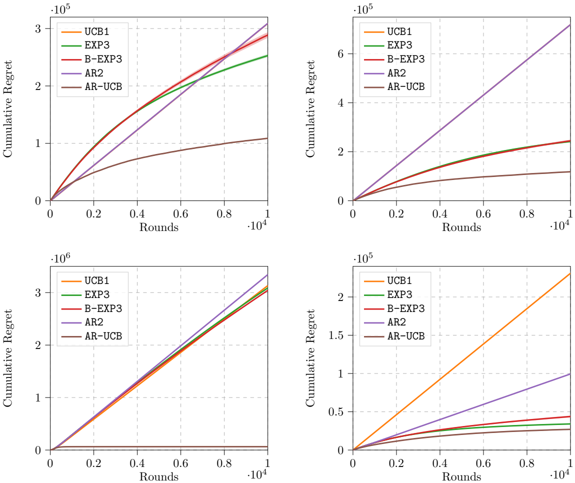


Real-World Scenario

We want to price 4 products of an e-commerce website.
We generalize a synthetic model from real-world data selecting:

- $k = 8$ (2 months)
- $|\mathcal{A}| = 8$

We study the **cumulative regret** over $T = 10000$ samples



Conclusions

- We presented the **Autoregressive Bandits**, a setting to handle autoregressive processes using bandit algorithm
- We defined the notion of **optimal policy**, and we demonstrates that the myopic policy is optimal also over long time horizon
- We presented **AR-UCB**, a regret minimization algorithm for handling autoregressive processes
- We have **theoretically characterized** the algorithm, and we conducted an experimental campaign over synthetic and real-world data

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE.

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

Chen, Q., Golrezaei, N., and Bouneffouf, D. (2021). Dynamic bandits with temporal structure. *Available at SSRN 3887608*.

Dekel, O., Tewari, A., and Arora, R. (2012). Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Conference on Machine Learning*.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. *Advances in neural information processing systems*, 30.

Ortner, R., Ryabko, D., Auer, P., and Munos, R. (2012). Regret bounds for restless markov bandits. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 214–228. Springer.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. (2018). Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR.

Tekin, C. and Liu, M. (2012). Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611.