Autoregressive Bandits

POLITECNICO MILANO 1863

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

{francesco.bacchiocchi, gianmarco.genalti, davide.maran, marco.mussi,
nicola.gatti, marcello.restelli, albertomaria.metelli}@polimi.it

e l l i s

FUTURE AI RESEARCH

Contributions

- We define **Autoregressive Bandits** to represent **sequential decision-making** problems where the rewards are governed by an **Autoregressive Process**
- We demonstrate that the **optimal policy** is **greedy**
- We propose **AR-UCB**, an **optimistic regret minimization** algorithm, and provide:
  - an upper bound on the **expected policy regret** in the order of $\widetilde{\mathcal{O}}(\sqrt{T})$
  - an extensive numerical validation

Setting - Autoregressive Bandits

Reward Model

At every time $t \in [\![T]\!]$, we select an action a_t and receive a reward x_t composed of:

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

where:

- k is the **order of the AR process**
- $(\gamma_i(a_t))_{i\in\{0,\dots,k\}}$ is an **unknown parameter vector**, characteristic of action $a_t \in \mathcal{A}$ ($|\mathcal{A}| = n$)
- ξ_t is σ^2-subgaussian random noise

Assumptions

a. **(Non-negative coef.)** $\gamma_i(a) \geq 0,\ \forall a \in \mathcal{A},\ i \in [\![0,k]\!]$

b. **(Stability)** $\Gamma := \max_{a\in\mathcal{A}} \sum_{i=1}^{k} \gamma_i(a) < 1$

c. **(Boundedness)** $m := \max_{a\in\mathcal{A}} \gamma_0(a) < +\infty$

Optimal Policy

Under Assumption (a), for every round $t \in \mathbb{N}$, the optimal policy π_t^* satisfies:

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

where:

- $\boldsymbol{\gamma}(a) := (\gamma_0(a), \dots, \gamma_k(a))$ is the **coefficient vector** for action a
- $\mathbf{z}_{t-1} := (1, x_{t-1}, \dots, x_{t-k})$ is a **Markovian state representation of the problem**

Algorithm - Autoregressive Upper Confidence Bound (AR-UCB)

`AutoRegressive Upper Confidence Bound` is an optimistic regret minimization algorithm that **exploits the linear structure of autoregressive processes**.

- `AR-UCB` takes as input:
  - Regularization parameter λ
  - Subgaussianity coefficient σ^2
  - AR process order k
  - Scale of the process m
- For every action a, `AR-UCB` makes use of a **Ridge-Regularized Regression** in order to estimate the corresponding coefficients $\boldsymbol{\gamma}(a)$.
- `AR-UCB` plays the **optimistic action** based on the regression's estimates of AR coefficients and their **uncertainty region**:

Algorithm 1: `AR-UCB`.

Input: $\lambda > 0,\ \sigma^2,\ k,\ m$
Initialize $\mathbf{V}_0(a) = \lambda \mathbf{I}_{k+1}$, $\mathbf{b}_0(a) = \mathbf{0}_{k+1}$, $\widehat{\boldsymbol{\gamma}}_0(a) = \mathbf{0}_{k+1}, \forall a \in \mathcal{A}$
Initialize $\mathbf{z}_0 = (1, 0, \dots, 0)^T$
for $t \in [\![T]\!]$ **do**
  Compute $a_t \in \arg\max_{a\in\mathcal{A}} \mathrm{UCB}_t(a)$
  Play a_t and observe $x_t = \langle \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle + \xi_t$
  for $a \in \mathcal{A}$ **do**
    $\mathbf{V}_t(a) = \mathbf{V}_{t-1}(a) + \mathbf{z}_{t-1}\mathbf{z}_{t-1}^T \mathbb{1}_{\{a=a_t\}}$
    $\mathbf{b}_t(a) = \mathbf{b}_{t-1}(a) + \mathbf{z}_{t-1}x_t \mathbb{1}_{\{a=a_t\}}$
    $\widehat{\boldsymbol{\gamma}}_t(a) = \mathbf{V}_t(a)^{-1}\mathbf{b}_t(a)$
  end
  Update $\mathbf{z}_t = (1, x_t, \dots, x_{t-k+1})^T$
end

$$\|\widehat{\boldsymbol{\gamma}}_t(a) - \boldsymbol{\gamma}(a)\|_{\mathbf{V}_t(a)} \leq \beta_t(a) := \sqrt{\lambda(m^2+1)} + \sigma\sqrt{2\log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(a)}{\lambda^{k+1}}\right)}$$

$$a_t \in \arg\max_{a\in\mathcal{A}} \mathrm{UCB}_t(a) := \arg\max_{a\in\mathcal{A}} \langle \widehat{\boldsymbol{\gamma}}_{t-1}(a), \mathbf{z}_{t-1}\rangle + \beta_{t-1}(a)\|\mathbf{z}_{t-1}\|_{\mathbf{V}_{t-1}(a)^{-1}},$$

Regret Analysis

Regret Decomposition

$$\underbrace{r_t = x_t^* - x_t}_{\text{Instantaneous Policy Regret}} = \underbrace{\sum_{i=1}^{k}\gamma_i(a_t^*)(x_{t-i}^* - x_{t-i})}_{\text{Instantaneous Autoregressive Regret}} + \underbrace{\langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle}_{\text{Instantaneous External Regret}} = \sum_{i=1}^{k}\gamma_i(a_t^*)r_{t-i} + \rho_t$$

External-to-Policy Regret Bound

$$\mathbb{E}\left[R(\boldsymbol{\pi}, T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{k}\gamma_i(a_t^*)r_{t-i} + \rho_t\right]\right] \leq \underbrace{\left(\frac{\Gamma k}{1-\Gamma} + 1\right)}_{\text{External-to-Policy Regret}} \cdot \underbrace{\varrho(\boldsymbol{\pi}, T)}_{\text{External Regret}}$$

Upper Bound on Expected Policy Regret

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

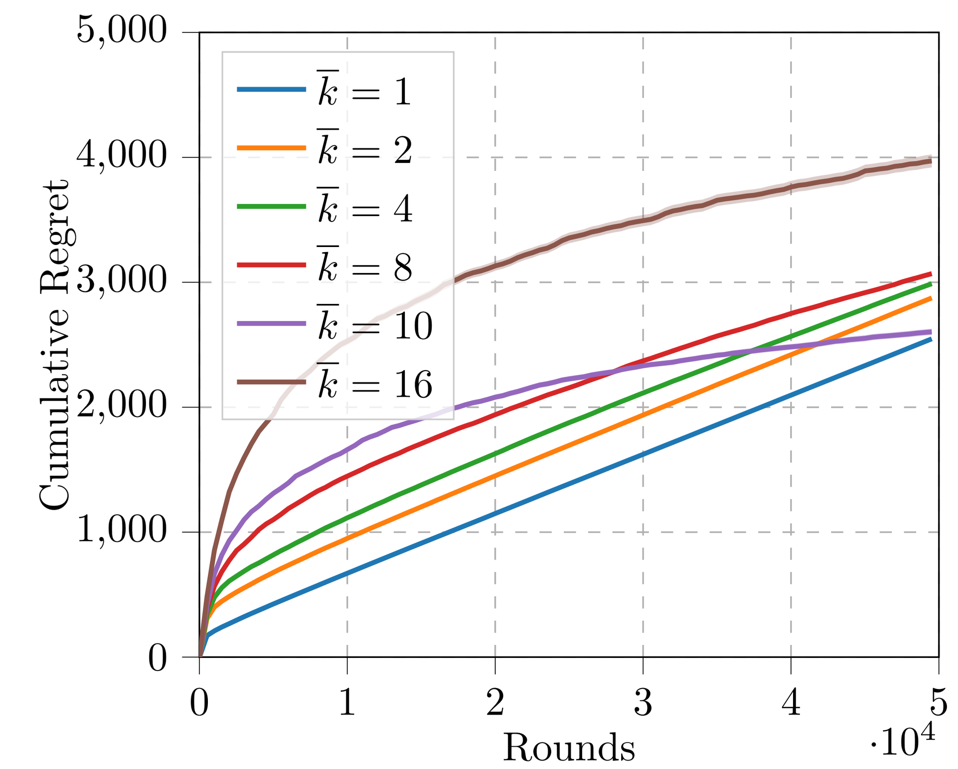
Experimental Validation

AR-UCB vs Bandit Baselines

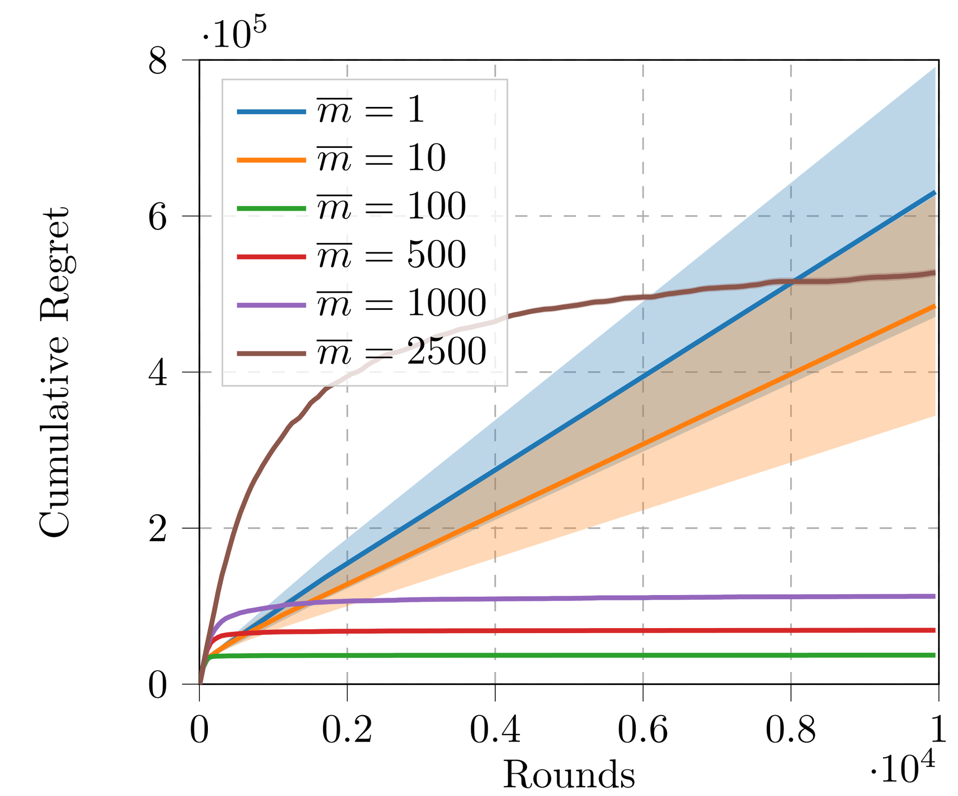
- Time horizon $T = 10^4$
- $\gamma_i(a) \sim \mathcal{U}(0, 1/k),\ \forall a \in \mathcal{A},\ i \in [\![0,k]\!]$

Parameters			Cumulative Regret $(\cdot 10^3)$		
k	m	σ	`AR-UCB`	`UCB1`	`EXP3`
2	1	0.75	**0.58**	3.2	3.6
4	20	1.5	**25**	739	352
4	920	10	**247**	4249	2925

Misspecification of k (real $k = 10$)



Misspecification of m (real $m = 500$)



References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *NeurIPS*, 2011.

S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun. Information theoretic regret bounds for online nonlinear control. *NeurIPS*, 2020.

M. Mussi, A. M. Metelli, and M. Restelli. Dynamical linear bandits. *ICML*, 2023.