



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 $\tilde{O}(\sqrt{T})$
 - an extensive numerical validation

SETTING - AUTOREGRESSIVE BANDITS

REWARD MODEL

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

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

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

- (Non-negative coef.)** $\gamma_i(a) \geq 0, \forall a \in \mathcal{A}, i \in \llbracket 0, k \rrbracket$
- (Stability)** $\Gamma := \max_{a \in \mathcal{A}} \sum_{i=1}^k \gamma_i(a) < 1$
- (Boundedness)** $m := \max_{a \in \mathcal{A}} \gamma_0(a) < +\infty$

OPTIMAL POLICY

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

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

where:

- $\gamma(a) := (\gamma_0(a), \dots, \gamma_k(a))$ 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 **L²-Penalized Least Squares Regression** in order to estimate the corresponding coefficients $\gamma(a)$.
- AR-UCB plays the **optimistic action** based on the regression's estimates of AR coefficients and their **uncertainty region**:

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

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

Algorithm 1: AR-UCB.

Input: $\lambda > 0, \sigma^2, k, m$

Initialize $\mathbf{V}_0(a) = \lambda \mathbf{I}_{k+1}, \mathbf{b}_0(a) = \mathbf{0}_{k+1},$
 $\hat{\gamma}_0(a) = \mathbf{0}_{k+1}, \forall a \in \mathcal{A}$

Initialize $\mathbf{z}_0 = (1, 0, \dots, 0)^T$

for $t \in \llbracket T \rrbracket$ **do**

 Compute $a_t \in \arg \max_{a \in \mathcal{A}} \text{UCB}_t(a)$

 Play a_t and observe $x_t = \langle \gamma(a_t), \mathbf{z}_{t-1} \rangle + \xi_t$

for $a \in \mathcal{A}$ **do**

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

$\mathbf{b}_t(a) = \mathbf{b}_{t-1}(a) + x_t \mathbb{1}_{\{a=a_t\}}$

$\hat{\gamma}_t(a) = \mathbf{V}_t(a)^{-1} \mathbf{b}_t(a)$

end

 Update $\mathbf{z}_t = (1, x_t, \dots, x_{t-k+1})^T$

end

REGRET ANALYSIS

REGRET DECOMPOSITION

$$\underbrace{r_t = x_t^* - x_t}_{\text{Instantaneous policy regret}} = \underbrace{\sum_{i=1}^k \gamma_i(a_t^*)(x_{t-i}^* - x_{t-i})}_{\text{Instantaneous autoregressive regret}} + \underbrace{\langle \gamma(a_t^*) - \gamma(a_t), \mathbf{z}_{t-1} \rangle}_{\text{Instantaneous external regret}} = \sum_{i=1}^k \gamma_i(a_t^*) r_{t-i} + \rho_t$$

EXTERNAL-TO-POLICY REGRET BOUND

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

UPPER BOUND ON EXPECTED POLICY REGRET

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

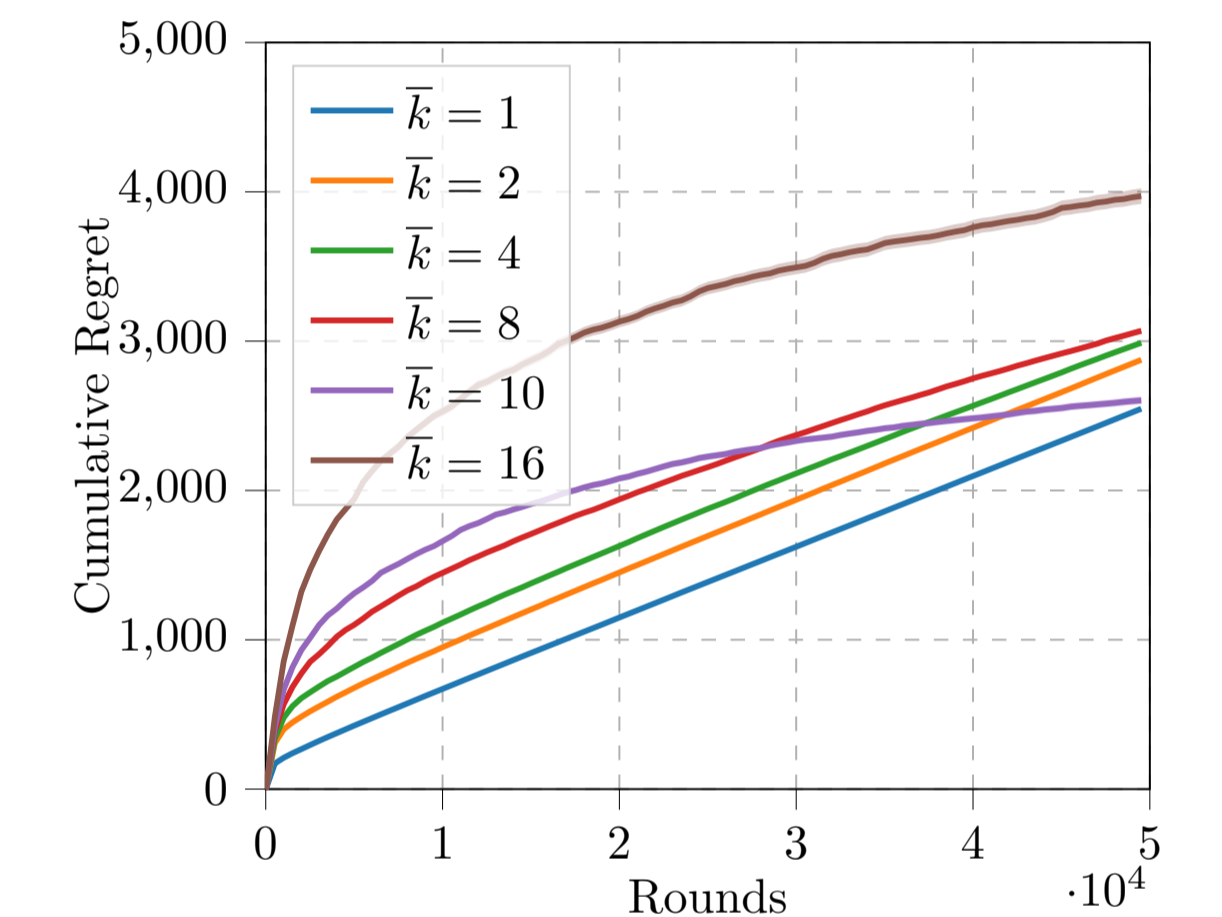
EXPERIMENTAL VALIDATION

AR-UCB VS BANDIT BASELINES

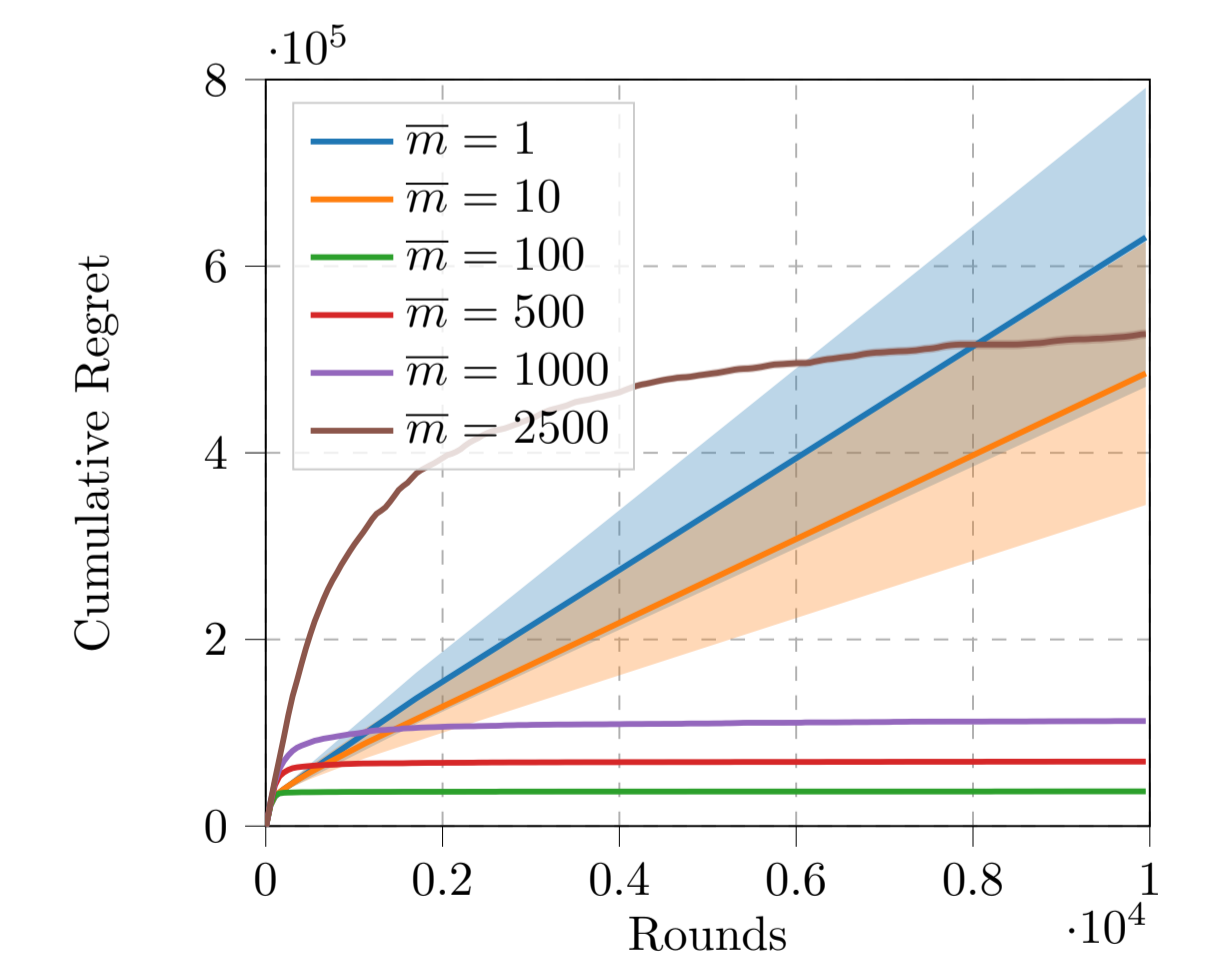
- Time horizon $T = 10^4$
- $\gamma_i(a) \sim \mathcal{U}(0, 1/k), \forall a \in \mathcal{A}, i \in \llbracket 0, k \rrbracket$

k	m	σ	Cumulative Regret ($\cdot 10^3$)		
			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

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *NeurIPS*, 2011.
- Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *NeurIPS*, 2020.
- Marco Mussi, Alberto Maria Metelli, and Marcello Restelli. Dynamical linear bandits. *ICML*, 2023.