Online Learning in Autoregressive Dynamics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Autoregressive processes naturally arise in a large variety of real-world scenarios, including stock markets, sales forecasting, weather prediction, advertising, and pricing. When facing a sequential decision-making problem in such a context, the temporal dependence between consecutive observations should be properly accounted for guaranteeing convergence to the optimal policy. In this work, we propose a novel online learning setting, namely, Autoregressive Bandits (ARBs), in which the observed reward is governed by an autoregressive process of order k, whose parameters depend on the chosen action. We show that, under mild assumptions on the reward process, the optimal policy can be conveniently computed. Then, we devise a new optimistic regret minimization algorithm, namely, `AutoRegressive Upper Confidence Bound` (`AR-UCB`), that suffers sublinear regret of order $\widetilde{\mathcal{O}}\left(\frac{(k+1)^{3/2}\sqrt{nT}}{(1-\Gamma)^2}\right)$, where T is the optimization horizon, n is the number of actions, and $\Gamma < 1$ is a stability index of the process. Finally, we empirically evaluate our algorithm in both synthetic and real-world domains, illustrating its advantages w.r.t. relevant bandit baselines.

1 Introduction

In a large variety of sequential decision-making problems, a learner is required to choose an action that, when executed, determines: (i) the immediate reward and (ii) the behavior of an underlying process that will influence, in some unknown manner, the future rewards. This process is influenced by the course of actions the agent performs and generates a temporal dependence between the sequence of observed rewards. A class of stochastic processes widely employed to model the temporal dependencies in real-world phenomena are the *autoregressive* (AR) processes (Hamilton, 2020). In this paper, we model the reward of a sequential decision-making problem as an AR process whose parameters depend on the action selected by the agent at every round. This scenario can be represented as a particular class of continuous *reinforcement learning* problems (Sutton and Barto, 2018) where an AR process governs the temporal structure of the observed rewards through the action-dependent AR parameters that are unknown to the agent. It is worth mentioning that such a scenario displays notable differences compared to more traditional *non-stationary* learning problems. Indeed, in the scenario we address, the environment does not change, and the reward dynamics depend on the agent's course of actions only. Let us consider a motivating real-world example.

Example *Consider the problem of finding the optimal price for a given product. A pricing strategy aims at maximizing a certain index,* e.g.*, volumes, turnover, or profit. Usually, pricing algorithms focus on the* one-step *performance (Mueller et al., 2019; Mussi et al., 2022). These solutions, however, fail in modeling the* long-term *phenomena that a pricing strategy inherently presents. Indeed, with one-step pricing algorithms, we fail (i) to model the long-term effect of our pricing strategy on customer loyalty and (ii) to capture the different demands of loyal and not-loyal customers. This problem, even if ubiquitous in the real world (Bowen and Chen, 2001), is unexplored in the literature, as existing approaches struggle to correctly deal with these autoregressive dynamics.*

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

Original Contribution In this work, motivated by the real-world examples presented above, we propose a novel setting, named *AutoRegressive Bandit* (ARB), in which the reward follows an AR process of order k whose parameters depend on the agent's actions. Importantly, we show that the optimal policy, differently from many bandit models, is *stationaty* and *closed-loop*, as the optimal action depends on the previously observed rewards (Section 2). Then, we devise a new optimistic algorithm, namely `AutoRegressive Upper Confidence Bound` (`AR-UCB`), to learn an optimal policy in an online fashion (Section 3), and we show that it suffers sublinear regret of order $\widetilde{\mathcal{O}}\left(\frac{(k+1)^{3/2}\sqrt{nT}}{(1-\Gamma)^2}\right)$, where T is the optimization horizon, n is the number of actions, and $\Gamma < 1$ is a stability index of the process (Section 4). Finally, we empirically evaluate `AR-UCB` on both synthetic and real-world data, comparing its performance with several bandit baselines with competitive results and illustrating its notable robustness w.r.t. the misspecification of key parameters (Section 5).

2 Problem Formulation

In this section, we introduce the ARB setting, formalize the learning problem, how the learner interacts with the environment, assumptions, policies and definition of regret (Section 2.1). Subsequently, we derive a closed-form solution for the optimal policy of an ARB (Section 2.2).

2.1 Setting

We study the sequential interaction between a learner and an environment. At every round $t \in \mathbb{N}$, the learner chooses an action $a_t \in \mathcal{A} := [\![n]\!]$, among the $n \in \mathbb{N}$ available ones.[1] In the ARB setting, the reward evolves according to an *autoregressive process* of order k (AR(k), Hamilton, 2020). Thus, the learner observes a noisy reward x_t of the form:

$$x_t = \gamma_0(a_t) + \sum_{i=1}^{k} \gamma_i(a_t) x_{t-i} + \xi_t, \tag{1}$$

where $x_t \in \mathcal{X} \subseteq \mathbb{R}$ is the reward space, $\gamma_0(a_t) \in \mathbb{R}$ and $(\gamma_i(a_t))_{i \in [\![k]\!]} \in \mathbb{R}^k$ are the unknown *parameters* depending on chosen action a_t, and ξ_t is a zero-mean σ^2-subgaussian random noise, conditioned to the past. The reward evolution can be expressed in an alternative form as follows: [2]

$$x_t = \langle \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle + \xi_t, \tag{2}$$

where $\mathbf{z}_{t-1} := (1, x_{t-1}, \dots, x_{t-k})^T \in \mathcal{Z} := \{1\} \times \mathcal{X}^k$ is the *vector of past rewards* expressing past history, and $\boldsymbol{\gamma}(a) := (\gamma_0(a), \dots, \gamma_k(a))^T \in \mathbb{R}^{k+1}$ is the *parameter vector*, defined for all the actions $a \in \mathcal{A}$. It is worth noting that when $\gamma_i(a) = 0$ for all $i \in [\![k]\!]$ and $a \in \mathcal{A}$, the ARB setting reduces to a standard MAB (Auer et al., 2002a).

We introduce the assumption that we employ in the paper and comment on its role.

Assumption 1. *The parameters $(\gamma_i(a))_{i \in [\![0,k]\!]}$ fulfill the following conditions:*

a. (Non-negative coefficients) $\gamma_i(a) \geq 0$ for every $a \in \mathcal{A}$, $i \in [\![0, k]\!]$;
b. (Stability) $\Gamma := \max_{a \in \mathcal{A}} \sum_{i=1}^{k} \gamma_i(a) < 1$;
c. (Boundedness) $m := \max_{a \in \mathcal{A}} \gamma_0(a) < +\infty$.

First, Assumption 1.a requires that the coefficients of the AR process are non-negative. This scenario is ubiquitous in real-world AR phenomena (*e.g.*, pricing, stock markets, digital advertising), where processes violating such an assumption will generate unrealistic sign alternation behaviours. An extensive discussion and a graphical elaboration about this assumption are provided in Appendix D. Assumption 1.b requires that the sum of $(\gamma_i(a))_{i \in [\![k]\!]}$ is limited to a value $\Gamma \in [0, 1)$ and Assump-

[1] Along the paper, we use the following notation. Let $a, b \in \mathbb{N}$ with $a \leq b$, we denote with $[\![a, b]\!] := \{a, \dots, b\}$, and with $[\![b]\!] := \{1, \dots, b\}$. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be real-valued vectors, we denote with $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i$ the inner product. For a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we denote with $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^T \mathbf{A} \mathbf{x}$ the weighted 2-norm. A zero-mean random variable ξ is σ^2-subgaussian if $\mathbb{E}[e^{\lambda \xi}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$, for every $\lambda \in \mathbb{R}$.

[2] Although the linear structure might resemble the *contextual linear bandits* (Chu et al., 2011), the two settings are non-comparable. Indeed, in our ARBs the vector $\mathbf{z}_{t-1}$ is not sampled independently at every round, but, instead, follows a sequential process depending on the past, making the decision problem way more challenging.

tion 1.c enforces the boundedness of $\gamma_0(a)$. These latter assumptions guarantee that the AR process does not diverge in expectation regardless of the sequence of the actions played.

Policies and Regret The learner's behavior is modeled by a deterministic policy $\boldsymbol{\pi} = (\pi_t)_{t\in\mathbb{N}}$ defined, for every round $t \in \mathbb{N}$ as $\pi_t : \mathcal{H}_{t-1} \to \mathcal{A}$, mapping the history of observations $H_{t-1} = (x_0, a_1, x_1, \dots, a_{t-1}, x_{t-1}) \in \mathcal{H}_{t-1}$ to an action $a_t = \pi_t(H_{t-1}) \in \mathcal{A}$ where $\mathcal{H}_{t-1} = \mathcal{X} \times (\mathcal{A} \times \mathcal{X})^{t-1}$ is the set of histories of length $t-1$. The performance of a policy $\boldsymbol{\pi}$ is evaluated in terms of the *expected cumulative reward* over the horizon $T \in \mathbb{N}$, defined as:

$$J_T(\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{t=1}^{T} x_t\right] \qquad \text{with} \qquad \begin{cases} x_t = \langle \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle + \xi_t \\ a_t = \pi_t(H_{t-1}) \end{cases}, \tag{3}$$

where the expectation is taken w.r.t. the randomness of the reward noise ξ_t. A policy $\boldsymbol{\pi}^*$ is *optimal* if it maximizes the expected average reward, *i.e.*, $\boldsymbol{\pi}^* \in \arg\max_{\boldsymbol{\pi}} J_T(\boldsymbol{\pi})$, whose performance is denoted as $J_T^* := J_T(\boldsymbol{\pi}^*)$. The goal of the learner is to minimize the *expected cumulative (policy) regret* by playing a policy $\boldsymbol{\pi}$, competing against the optimal policy $\boldsymbol{\pi}^*$ over a *learning horizon* $T \in \mathbb{N}^+$:

$$R(\boldsymbol{\pi}, T) = J_T^* - J_T(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right], \tag{4}$$

where $r_t := x_t^* - x_t$ is the instantaneous policy regret and $(x_t^*)_{t\in[\![T]\!]}$ is the sequence of rewards observed by playing the optimal policy $\boldsymbol{\pi}^*$.

Example (cont.) *The problem of optimal pricing presented in Section 1 can be easily mapped to the ARB setting. Consider the scenario in which we want to maximize the volumes over time. The volumes are our* reward x_t*, and the history of our rewards* $x_{t-1}, \dots, x_{t-k}$ *provides an indication of the loyal customer pool over* k *units of time (*e.g.*, days, weeks). The ARB setting allows modeling a reward which is the contribution of both new customers (via* γ_0*) and the loyal customer pool (via* $\gamma_1, \dots, \gamma_k$*). Specifically, the price, our* action a_t*, induces different values of the coefficient* $\boldsymbol{\gamma}(a_t)$*, to represent the different demand curves that loyal and new customers might have.*

2.2 Optimal Policy

In this section, we derive a closed-form expression for the optimal policy $\boldsymbol{\pi}^*$ for the expected cumulative reward of Equation (3), under Assumption 1.a.

Theorem 1 (Optimal Policy)**.** *Under Assumption 1.a, for every round* $t \in \mathbb{N}$*, the optimal policy* $\pi_t^*(H_{t-1})$ *satisfies:*

$$\pi_t^*(H_{t-1}) \in \arg\max_{a\in\mathcal{A}} \langle \boldsymbol{\gamma}(a), \mathbf{z}_{t-1}\rangle. \tag{5}$$

This result deserves some comments. First, the optimal action depends on the vector of past rewards $\mathbf{z}_{t-1}$ and, thus, on the most recent k rewards $x_{t-1}, \dots, x_{t-k}$ only. Thus, the optimal policy $\boldsymbol{\pi}^*$ is non-Markovian with memory k or, equivalently, Markovian w.r.t. the state representation $\mathbf{z}_{t-1}$.[3] Second, the optimal action maximizes, at every round $t \in \mathbb{N}$, the *expected instantaneous reward* $\mathbb{E}[x_t|H_{t-1}] = \langle \boldsymbol{\gamma}(a), \mathbf{z}_{t-1}\rangle$. This is a consequence of the non-negativity of the parameters $\gamma_i(a)$ (Assumption 1.a), which enforces a meaningful evolution of the AR process, compatible with our real-world motivating scenarios. This way, the action maximizing the expected *immediate* reward (*i.e.*, a *myopic* policy) is optimal for the expected *cumulative* reward too. The proof can be found in Appendix B.

3 AutoRegressive Upper Confidence Bounds

In this section, we present `AutoRegressive Upper Confidence Bound` (`AR-UCB`), an optimistic regret minimization algorithm for the ARB setting whose pseudo-code is reported in Algorithm 1. `AR-UCB` leverages the myopic optimal policy for ARBs (Theorem 1) and implements an incremental regularized least squares procedure to estimate the unknown parameters $\boldsymbol{\gamma}(a)$, for every action $a \in \mathcal{A}$

[3] We can look at the ARB as a particular *Markov decision processes* (Puterman, 2014) with $\mathbf{z}_{t-1} \in \mathcal{Z}$ as state representation.

Algorithm 1: `AR-UCB`.

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

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

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

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

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

5 Update for all $a \in \mathcal{A}$:

6 $\mathbf{V}_t(a) = \mathbf{V}_{t-1}(a) + \mathbf{z}_{t-1}\mathbf{z}_{t-1}^T \mathbb{1}_{\{a=a_t\}}$
$\mathbf{b}_t(a) = \mathbf{b}_{t-1}(a) + \mathbf{z}_{t-1} x_t \mathbb{1}_{\{a=a_t\}}$

7 Compute $\widehat{\boldsymbol{\gamma}}_t(a) = \mathbf{V}_t(a)^{-1}\mathbf{b}_t(a)$

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

9 $t \leftarrow t + 1$

10 **end**

independently. The algorithm requires the knowledge of the order k of the AR process, although this knowledge can be replaced with the one of an upper bound $\overline{k} > k$ of the AR order.[4]

`AR-UCB` starts by initializing for all the actions $a \in \mathcal{A}$ the Gram matrix $\mathbf{V}_0(a) = \lambda \mathbf{I}_{k+1}$, where $\lambda > 0$ is the Ridge regularization parameter, the vectors $\mathbf{b}_0(a) = \widehat{\boldsymbol{\gamma}}_0(a) = \mathbf{0}_{k+1}$, and the observations vector $\mathbf{z}_0 = (1, 0, \dots, 0)^T$ (Line 1).[5] Then, for each round $t \in [\![T]\!]$, `AR-UCB` computes the *Upper Confidence Bound* (UCB) index (Line 3) for every $a \in \mathcal{A}$ and the optimistic action a_t:

$$a_t \in \arg\max_{a\in\mathcal{A}} \ \mathrm{UCB}_t(a) \coloneqq \langle \widehat{\boldsymbol{\gamma}}_{t-1}(a), \mathbf{z}_{t-1}\rangle + \beta_{t-1}(a) \left\|\mathbf{z}_{t-1}(a)\right\|_{\mathbf{V}_{t-1}(a)^{-1}}, \tag{6}$$

where $\widehat{\boldsymbol{\gamma}}_{t-1}(a)$ is the most recent estimate of the parameter vector $\boldsymbol{\gamma}(a)$, $\mathbf{z}_{t-1} = (1, x_{t-1}, \dots, x_{t-k})^T$ is the observations vector, and $\beta_{t-1}(a) \geq 0$ is an exploration coefficient that will be defined later (Section 4). The index $\mathrm{UCB}_t(a)$ is designed to be optimistic, *i.e.*, $\langle \boldsymbol{\gamma}_{t-1}(a), \mathbf{z}_{t-1}\rangle \leq \mathrm{UCB}_t(a)$ with high probability for all $a \in \mathcal{A}$. Then, action a_t is executed (Line 4) and the new reward x_t is observed. This sample is employed to update the Gram matrix estimate $\mathbf{V}_t(a_t)$, the vector $\mathbf{b}_t(a_t)$, and the estimate $\widehat{\boldsymbol{\gamma}}_t(a_t)$ of the played action only (lines 6-7).

4 Regret Analysis

In this section, we present the analysis of the regret of `AR-UCB`. We start providing a self-normalized concentration inequality for estimating the AR parameters $\boldsymbol{\gamma}(a)$ (Section 4.1). Then, we derive a decomposition of the regret (Section 4.2) that is useful to complete the analysis and, finally, we present the bound on the expected cumulative (policy) regret (Section 4.3). The complete proofs of the theorems stated in this section can be found in Appendix B.

4.1 Concentration Inequality for the Parameter Vectors

We start by providing a concentration result for the estimates $\widehat{\boldsymbol{\gamma}}_t(a)$ of the true parameter vector $\boldsymbol{\gamma}(a)$, for every action $a \in \mathcal{A}$, as performed in Algorithm 1. At the end of each round $t \in \mathbb{N}$, where the chosen action is $a_t \in \mathcal{A}$, we solve the Ridge-regularized linear regression problem and update the coefficient vector estimate $\widehat{\boldsymbol{\gamma}}_t(a_t)$ associated to a_t:

$$\widehat{\boldsymbol{\gamma}}_t(a_t) = \arg\min_{\widetilde{\boldsymbol{\gamma}}\in\mathbb{R}^{k+1}} \sum_{l\in\mathcal{O}_t(a_t)} (x_l - \langle \widetilde{\boldsymbol{\gamma}}, \mathbf{z}_{l-1}\rangle)^2 + \lambda \left\|\widetilde{\boldsymbol{\gamma}}\right\|_2^2 = \mathbf{V}_t(a_t)^{-1}\mathbf{b}_t(a_t),$$

where $\mathcal{O}_t(a)$ is the set of rounds where action a has been chosen, *i.e.*, $\mathcal{O}_t(a) \coloneqq \{\tau \in [\![t]\!] : a_\tau = a\}$. The following result shows how the estimate $\widehat{\boldsymbol{\gamma}}(a)$ concentrates around the true parameters $\boldsymbol{\gamma}(a)$ over the rounds.

Lemma 2 (Self-Normalized Concentration)**.** *Let $a \in \mathcal{A}$ be an action, let $(\widehat{\boldsymbol{\gamma}}_t(a))_{t\in\mathcal{O}_\infty(a)}$ be the sequence of solutions to the Ridge regression problems computed by Algorithm 1. Then, for every*

[4] Indeed, any AR process of order k can be regarded as an AR process of order $\overline{k} > k$ setting $\gamma_i(a) = 0$ for $i \in [\![k+1, \overline{k}]\!]$. Empirical validation of the misspecification of k in Section 5.4.

[5] We assume to know the initial observations vector $\mathbf{z}_0$. If this is not the case, we can play an arbitrary action for the first k rounds to observe $(x_t)_{t\in[\![k]\!]}$ with just an additional constant loss term.

regularization parameter $\lambda > 0$, confidence $\delta \in (0,1)$, simultaneously for every round $t \in \mathbb{N}$ and action $a \in \mathcal{A}$, with probability at least $1-\delta$ it holds that:

$$\|\widehat{\boldsymbol{\gamma}}_t(a) - \boldsymbol{\gamma}(a)\|_{\mathbf{V}_t(a)} \leq \sqrt{\lambda}\|\boldsymbol{\gamma}(a)\|_2 + \sigma\sqrt{2\log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(a)}{\lambda^{k+1}}\right)}.$$

Lemma 2 resembles the self-normalized concentration inequality of (Abbasi-Yadkori et al., 2011, Theorem 1). However, contrary to `LIN-UCB` (Abbasi-Yadkori et al., 2011), the exploration coefficients $\beta_t(a)$ are different for every action $a \in \mathcal{A}$. Lemma 2 allows properly defining the exploration coefficients $\beta_t(a)$ employed in Algorithm 1, defined for every action $a \in \mathcal{A}$ and round $t \in [\![0, T-1]\!]$:

$$\beta_t(a) \coloneqq \sqrt{\lambda(m^2+1)} + \sigma\sqrt{2\log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(a)}{\lambda^{k+1}}\right)}. \tag{7}$$

This formula contains two terms. The first one is a *bias* term that increases with m (*i.e.*, the maximum value of the largest $\gamma_0(a)$ over the actions $a \in \mathcal{A}$, see Assumption 1.c) and with the regularization parameter of the Ridge regression $\lambda > 0$. The second one is the *concentration* term and increases with the subgaussian parameter σ of the noise, the number of actions n, and the determinant of the design matrix $\mathbf{V}_t(a)$, but decreases in λ. It is worth noting that $\beta_t(a)$ is obtained from Lemma 2, by observing that, under Assumptions 1.b and 1.c, we have $\|\boldsymbol{\gamma}(a)\|_2 \leq \sqrt{m^2+\Gamma^2} \leq \sqrt{m^2+1}$. Thus, the exploration coefficient $\beta_t(a)$ ensures that, with probability $1-\delta$, the following inequality holds simultaneously for all actions $a \in \mathcal{A}$ and rounds $t \in [\![0, T-1]\!]$:

$$\|\widehat{\boldsymbol{\gamma}}_t(a) - \boldsymbol{\gamma}(a)\|_{\mathbf{V}_t(a)} \leq \beta_t(a). \tag{8}$$

We observe that $\beta_t(a)$ (see Equation 7) and `AR-UCB` do not require the knowledge of the maximum sum Γ of the parameters $\gamma_i(a)$ over the actions (see Assumption 1.b). This is a desirable feature of our algorithm as Γ is often unknown in practice and difficult to upper bound or estimate. Nevertheless, Γ appears in the regret analysis in Section 4.2. Differently, the value of m, needed to compute the optimistic coefficient $\beta_t(a)$, can be easily replaced with an upper bound $\overline{m} > m$ when unknown. An empirical analysis of the effect of the misspecification of such a parameter is provided in Section 5.3.

4.2 Regret Decomposition

In this section, we present a novel *decomposition* of the regret that will be employed in the final bound of Section 4.3. The contents of this section are of independent interest and applicable to any learner's policy $\boldsymbol{\pi}$, beyond `AR-UCB`. From a technical perspective, the analysis is composed of two steps: (i) we decompose the instantaneous (policy) regret r_t in terms of the instantaneous *external regret* ρ_t (Lemma 3); (ii) we bound the cumulative expected (policy) regret $R(\boldsymbol{\pi}, T) = \mathbb{E}[\sum_{t=1}^T r_t]$ in terms of the expected cumulative external regret $\varrho(\boldsymbol{\pi}, T) = \mathbb{E}[\sum_{t=1}^T \rho_t]$ (Lemma 4).

We start with step (i), by recalling that the definition of cumulative expected (policy) regret $R(\boldsymbol{\pi}, T)$ in Equation (4) compares the sequence of rewards $(x_t^*)_{t\in[\![T]\!]}$ when executing the optimal policy $\boldsymbol{\pi}^*$ with the sequence of rewards $(x_t)_{t\in[\![T]\!]}$ when executing the learner's policy $\boldsymbol{\pi}$. However, in our ARB setting, the observed reward x_t depends on the past history H_{t-1}. Thus, the instantaneous (policy) regret $r_t \coloneqq x_t^* - x_t$ can be decomposed in two terms: (a) the dissimilarity between the past history H_{t-1}^* when executing the optimal policy and the learner's observed history H_{t-1}; (b) the instantaneous *external regret* (Dekel et al., 2012) $\rho_t \coloneqq \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle$ representing the loss of executing the learner action a_t instead of the optimal one $a_t^* = \pi_t^*(H_{t-1}^*)$ *assuming* that such actions are applied to the observations vector $\mathbf{z}_{t-1}$ generated by the execution of the learner's policy. The following result formalizes the instantaneous regret decomposition.

Lemma 3 (Policy Regret Decomposition)**.** *Let $(x_t^*)_{t\in[\![T]\!]}$ be the sequence of rewards by executing the optimal policy $\boldsymbol{\pi}^*$ and let $(x_t)_{t\in[\![T]\!]}$ be the sequence of rewards by executing the learner's policy $\boldsymbol{\pi}$. Then, for every $t \in [\![T]\!]$ it holds that:*

$$r_t = x_t^* - x_t = \sum_{i=1}^k \gamma_i(a_t^*)(x_{t-i}^* - x_{t-i}) + \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle = \sum_{i=1}^k \gamma_i(a_t^*) r_{t-i} + \rho_t, \tag{9}$$

where $r_t \coloneqq x_t^* - x_t$ is the instantaneous policy regret, $\rho_t \coloneqq \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle$ is the instantaneous external regret, $a_t^* = \pi_t^*(H_{t-1}^*)$, and $r_{t-i} = 0$ if $i \geq t$.

The decomposition in Equation (9) comprises two terms. The second one ρ_t is the instantaneous external regret discussed above. The first one defines a recurrence relation of order k on the instantaneous policy regret r_t. We now move to step (ii) with the following result that shows that the contribution of the recurrence can be reduced to a term depending on Γ and k that multiplies the cumulative external regret.

Lemma 4 (External-to-Policy Regret Bound)**.** *Let $\boldsymbol{\pi}$ be the learner's policy and $T \in \mathbb{N}$ be the horizon. Under Assumptions 1.a and 1.b, it holds that:*

$$R(\boldsymbol{\pi}, T) = \mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{k}\gamma_i(a_t^*)r_{t-i} + \rho_t\right]\right] \leq \left(\frac{\Gamma k}{1-\Gamma} + 1\right)\varrho(\boldsymbol{\pi}, T), \tag{10}$$

where $\varrho(\boldsymbol{\pi}, T) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\rho_t\right]$ is the cumulative expected external regret.

Lemma 4 provide us a bound on the cumulative expected (policy) regret $R(\boldsymbol{\pi}, T)$ achieved by `AR-UCB` (or any algorithm playing in an ARB) by bounding the cumulative expected external regret $\varrho(\boldsymbol{\pi}, T)$. The order of the regret bound w.r.t. T is governed by the external regret, while the effect of a *weaker* history (*i.e.*, the sub-optimal actions of the past) emerges as an instance-specific constant. Such a constant is 1 whenever $k = 0$ or $\Gamma = 0$, *i.e.*, when the ARB reduces to a standard MAB. In all other cases, the bigger the value of k or Γ, the more visible the AR effects are, and, consequently, the more the sub-optimal choices of the past get amplified. Finally, we point out that the multiplicative factor $\frac{\Gamma k}{1-\Gamma} + 1$ to pass from external to policy regret is tight since there exists a sequence of external regrets in which the inequality of Lemma 4 holds with equality (see Appendix B).

4.3 Regret Bound

In the following, we present a bound on the expected policy regret bound for `AR-UCB`.

Theorem 5. *Let $\delta = (2T)^{-1}$. Under Assumptions 1.a, 1.b, and 1.c, `AR-UCB` suffers a cumulative expected (policy) regret bounded by (highlighting the dependence on m, σ, k, Γ, n, and T only):*

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

Some observations are in order. First, when we set $k = 0$ and $\Gamma = 0$, *i.e.*, we reduce the ARB to a standard MAB, we obtain a regret rate of $\widetilde{\mathcal{O}}((m+\sigma)\sqrt{nT})$, which is tight for standard MABs. The quantity $\frac{m+\sigma}{1-\Gamma}$ is the maximum value that rewards can achieve, as proven in Lemma 6. As intuition suggests, the ARB learning problem becomes more challenging as the AR order k increases and when the bound on the sum of the parameters Γ approaches one. This is witnessed in Theorem 5 with the dependence of the regret on $(k+1)^{3/2}$ and $(1-\Gamma)^{-1}$. The interplay between k and $(1-\Gamma)^{-1}$ is showing that even if two instances have the same sum of parameters (*i.e.*, Γ), the one with fewer coefficients (*i.e.*, k) is more easily learnable. This is explained by the fact that our algorithm learns the individual parameters by means of a regression procedure learning to a $\sqrt{k+1}$ in the regret. Finally, suppose we run `AR-UCB` with a larger AR order $\overline{k} > k$. In such a case, the dependence on $(k+1)^{3/2}$ becomes $(k+1)(\overline{k}+1)^{1/2}$, since the factor due to passing from external to policy regret (Lemma 4) will always contain the true k, while $\overline{k}$ appears because of the estimation process. Similarly, if we execute `AR-UCB` with a value $\overline{m} > m$, the regret bound still holds by replacing m with $\overline{m}$.

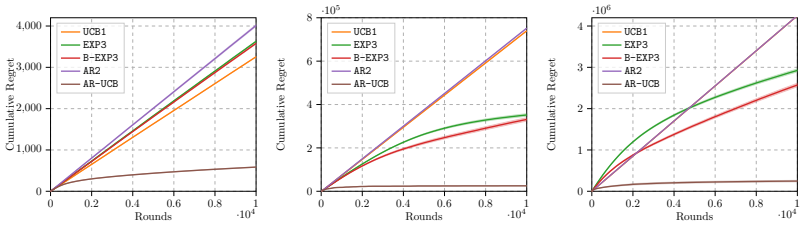
5 Experimental Validation

In this section, we first provide (Section 5.1) a numerical validation of `AR-UCB` compared with other bandit baselines in synthetically-generated domains. Then, we discuss (Section 5.2) the importance of exploiting the noise in this setting, and, subsequently, we analyze the sensitivity of `AR-UCB` to the misspecification of the two most important parameters, *i.e.*, m (Section 5.3) and k (Section 5.4). Finally, in Section E, we conduct validation in a setting generalized from real-world data.[6]

[6]The code to run the experiments is available in the supplementary material. It will be published in a public repository after the acceptance of the paper.

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

(a) Setting descriptions.



(b) Setting A. (c) Setting B. (d) Setting C.

Figure 1: Settings and cumulative regret of AR-UCB and multiple baselines (100 runs, mean $\pm$ std).

5.1 AR-UCB vs Bandit Baselines

Setting We evaluate AR-UCB in three scenarios that differ in the properties of the autoregressive processes that govern the rewards. The competing algorithms are evaluated in terms of cumulative regret w.r.t. to the setting-specific clairvoyant. The three settings have their AR(k) process order $k \in \{2, 4\}$, respectively, and $m \in \{1, 20, 920\}$. The values of $\boldsymbol{\gamma}(a)$ have been sampled from uniform probability distributions for each action $a \in \mathcal{A}$ and for each setting. The environments are noisy with a standard deviation $\sigma \in \{0.75, 1.5, 10\}$. We chose to set the hyper-parameters of AR-UCB as follows: $\lambda = 1$, while $\overline{m} \in \{10, 100, 1000\}$, that is equivalent to chose $\overline{m}$ of the same magnitude of the true value m, in a pessimistic fashion. Figure 1a summarizes the details of the three environments.

Baselines AR-UCB will compete with several bandit baselines. First, it is compared with UCB1 (Lai and Robbins, 1985; Auer et al., 2002a), a widely adopted solution for stochastic MABs. Second, we consider EXP3, designed for adversarial MABs (Auer et al., 1995, 2002b) and its extension to finite-memory adaptive adversaries B-EXP3 (Dekel et al., 2012). Lastly, we compare AR-UCB with AR2 (Chen et al., 2021), an algorithm for managing AR(1) processes. The hyper-parameters chosen for the baselines are the ones proposed in the original papers.

Results Figure 1 shows the average cumulative regrets for AR-UCB and the other bandit baselines. We observe that AR-UCB suffers the smallest cumulative regret in these scenarios, always displaying a sublinear behavior. Both EXP3 and B-EXP3 in two scenarios out of three (B and C) achieve sublinear regret. On the other hand, both UCB1 and AR2 are not able to achieve sublinear regret in the presented scenarios. This is not surprising since we require them to learn more complex processes than those they are designed for (*i.e.*, models with $k = 0$ and $k = 1$ for UCB1 and AR2, respectively).

5.2 On the Effect of Stochasticity

The optimal policy (Theorem 1) for the ARB setting exploits the contribution of the noise to increase the collected reward. In this section, we provide experimental evidence of this phenomenon. We first introduce a notion of *optimal policy without noise*. Then, we conduct an experiment to highlight the variations between the two policies in environments presenting different noise magnitudes.

Optimal Policy without Noise The optimal policy, when no noise is involved, is *constant* and corresponds, for sufficiently large T, to playing the action $a^+ \in \mathcal{A}$ that brings the system to the most profitable steady state.[7] Such an action a^+ is the one maximizing the *steady-state reward*, namely:

$$a^+ \in \arg\max_{a \in \mathcal{A}} \frac{\gamma_0(a)}{1 - \sum_{i=1}^{k} \gamma_i(a)}. \tag{11}$$

It is worth noting the role of Assumption 1.b which guarantees the existence of the inverse $(1 - \sum_{i=1}^{k} \gamma_i(a))^{-1} \geq (1 - \Gamma)^{-1}$ for each action $a \in \mathcal{A}$. The proof can be found in Appendix C.

Setting To demonstrate the importance of the noise in this setting, we consider the two clairvoyant policies defined above. We compare the optimal Stochastic policy (Equation 5) and the optimal policy for the Deterministic setting (Equation 11). The setting selected is challenging and made of $k = 2$ actions, a_1 and a_2, that are very close in terms of expected steady-state reward:

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

where $\rho = 0.5$, $\epsilon = 0.02$ and the noise is Gaussian with $\sigma \in \{0,\ 0.1,\ 0.5,\ 1.0,\ 2.0\}$.

[7] The request for large T is to make transient effects neglectable.

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)

Table 1: Cumulative reward of the Clairvoyant `Stochastic` and `Deterministic` policies (100 runs, mean (std)).

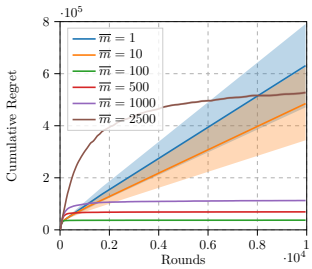


Figure 2: Effect of the choice of parameter $\overline{m}$ on the `AR-UCB` cumulative regret (100 runs, mean $\pm$ std).

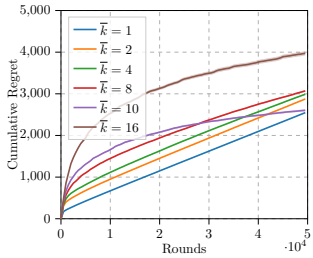
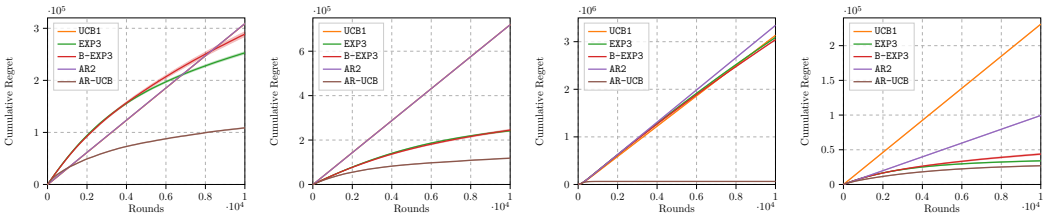


Figure 3: Effect of the choice of parameter $\overline{k}$ on the `AR-UCB` cumulative regret (100 runs, mean $\pm$ std).



(a) Product 1 (b) Product 2 (c) Product 3 (d) Product 4

Figure 4: `AR-UCB`, `UCB1`, `EXP3`, `B-EXP3` , and `AR2` in the experiment from real-world data (100 runs, mean $\pm$ std).

Results Table 1 shows the performance of the two policies in terms of cumulative reward. First, with no noise (*i.e.*, $\sigma = 0$), the performances of the two policies are equivalent. However, when we consider a stochastic setting (*i.e.*, $\sigma > 0$), the `Stochastic` policy can exploit the beneficial effect of the noise in order to increase the average reward. Indeed, the optimal `Deterministic` policy retrieves almost the same reward for all the tested values of σ, while `Stochastic` policy increases its average reward as much as the system is noisy (since it can exploit it).

5.3 On the Knowledge of Parameter m

A fundamental parameter of `AR-UCB` is the value $m = \max_{a \in \mathcal{A}} \gamma_0(a)$. In this experiment, we empirically show that any choice in the same order of magnitude as the actual value will let the algorithm achieve a sublinear regret, while severe underestimation prevents the algorithm from achieving a sublinear cumulative regret.

Setting We run multiple simulations varying the value of parameter $\overline{m}$. We chose $n = 7$, $k = 4$ and $\gamma_0(a) = 500$ for every action $a \in \mathcal{A}$ (*i.e.*, $m = 500$). The autoregressive parameters $\gamma_i(a)$ have been sampled from a uniform probability distribution with support in $[0, 1/4 - \epsilon]$, where $\epsilon > 0$ is an arbitrarily small value. For this experiment, we test values $\overline{m} \in \{1, 10, 100, 500, 1000, 2500\}$.

Results In Figure 2, we report the cumulative regrets of `AR-UCB` under different choices of $\overline{m}$. First, it is worth noting how choosing values of $\overline{m} \geq m$ always results in a sublinear cumulative regret, with a progressive increase as $\overline{m}$ gets larger. This is highlighted when comparing, for instance, the scenario where $\overline{m} = 2500$ to the one where $\overline{m} \in \{500, 1000\}$. When $\overline{m}$ is underestimated, we empirically observe two facts. When $\overline{m}$ is in the same order of magnitude as the true value m (*e.g.*, $\overline{m} = 100$), we empirically get a smaller sublinear cumulative regret (even if no theoretical guarantees are present). Finally, a severe underestimation of the parameter leads to a linear cumulative regret, as clearly visible for $\overline{m} \in \{1, 10\}$, although, in these settings, the cumulative regret is lower w.r.t. the other settings in the very first stages of the simulations (due to a more limited exploration).

5.4 On the Knowledge of the Autoregressive order k

As stated in Section 4, `AR-UCB` can also run under a misspecified parameter $\overline{k} \neq k$. In this section, we provide an empirical analysis of the effect of misspecifying such a value.

Setting We consider a configuration with $n = 7$, $k = 10$, $\gamma_0(a) = 1$ and $\gamma_i(a)$ for $i \geq 1$ sampled from a uniform distribution having support in $[0, 10^{-2} \cdot 2i)$ for every action $a \in \mathcal{A}$. AR-UCB is run varying the parameter $\overline{k} \in \{1, 2, 4, 8, 10, 16\}$.

Results Figure 3 reports the average cumulative regret for the considered values of $\overline{k}$. On the one hand, an underestimation of parameter k (*i.e.*, $\overline{k} \in \{1, 2, 4\}$) results in an asymptotically linear cumulative regret. This effect is justified since AR-UCB is not able to learn the actual AR dynamics due to underfitting, *i.e.*, the considered models are too simple. On the other hand, AR-UCB achieves sublinear cumulative regret when $\overline{k} \geq k$ (*i.e.*, $\overline{k} \in \{10, 16\}$). In particular, when $\overline{k} > k$, the linear models use more parameters than required, resulting in slower learning. However, as the samples increase, the algorithm learns that the exceeding coefficients are not significant. A particular case is when $\overline{k}$ is close to k but strictly lower (*i.e.*, $\overline{k} = 8$). Here, the cumulative regret degenerates to linear, but if the coefficients $\gamma_j(a)$ for $j \in [\![\overline{k} + 1, k]\!]$ are not very large, the performance of AR-UCB with misspecified $\overline{k}$ results, in practice, close to the one obtained with the true k.

5.5 Validation using Real-World Data

We evaluate AR-UCB over the *dynamic pricing* task in e-commerce. The problem of sequentially choosing the price while dealing with the exploration-exploitation dilemma is a well-known task in the literature (Kleinberg and Leighton, 2003). We show that AR-UCB is able to find the pricing schedule that maximizes the total sales while accounting for loyalty dynamics, using a simulation environment generated from real-world data.

Setting Configuration We have the possibility to access a dataset of transactions generated from a real e-commerce website selling consumables.[8] We focused on the top 4 best-selling products: for each product, we have weekly records of the number of units sold and the related price. we discretize the prices into $n = 8$ price bands (*i.e.*, our actions) and we build the simulation environment considering a maximum horizon of the effect of the past prices on the customer of $k = 8$ weeks. The choice of $k = 8$ (*i.e.*, two months) is ruled by business logic that is characteristic of the market in analysis. For each price band a, we estimated the parameters $\boldsymbol{\gamma}(a)$ through standard regression techniques. We used the historical sales data as the response variable predicted with the observed conversion rates for any group of customers (grouped using the number of weeks passed since their last purchase) with respect to any given price band a_t. In this experiment, we compare AR-UCB and the other bandit baselines presented in Section 5.1.

Results Figure 4 shows that only AR-UCB achieves sublinear regret for all the four products. EXP3 and B-EXP3 achieve sublinear regret for 3 out to 4 products, although their cumulative regret is always larger than that of AR-UCB, making the latter the best performing algorithm over the competitors. Lastly, both UCB1 and AR2 suffer linear regret for all the products under analysis.

6 Discussion and Conclusions

In this work, we faced the online sequential decision-making problem where an autoregressive temporal structure between the observed rewards is present. First, we formally introduced the ARB setting and defined the notion of optimal policy, demonstrating that, under certain circumstances, a myopic policy is optimal also to optimize the total reward, regardless of the target time horizon, and that the optimal policy is not constant over time and depends on the most recent observed rewards. Then, we proposed an optimistic bandit algorithm, AR-UCB, to learn online the parameters of the underlying process for each action. We demonstrated that the presented algorithm enjoys sublinear regret, depending on the AR order k and on an index of the speed at which the system reaches a stable condition (Γ). Finally, we provided an experimental campaign to validate the proposed solution demonstrating the effectiveness of AR-UCB w.r.t. several bandit baselines on both synthetic and real-world scenarios, and we analyzed the behavior of AR-UCB when key parameters are misspecified. Short-term future directions should focus on fully understanding the complexity of learning in the ARB setting, deriving tight lower bounds, and matching algorithms. Long-term future directions include the extension of the presented setting to continuous actions.

[8] We cannot share the original dataset due to an NDA.

References

James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Jonas W Mueller, Vasilis Syrgkanis, and Matt Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

Marco Mussi, Gianmarco Genalti, Francesco Trovò, Alessandro Nuara, Nicola Gatti, and Marcello Restelli. Pricing the long tail by explainable product aggregation and monotonic bandits. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3623–3633, 2022.

John T Bowen and Shiang-Lih Chen. The relationship between customer loyalty and customer satisfaction. *International journal of contemporary hospitality management*, 2001.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *JMLR Proceedings*, pages 208–214, 2011.

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

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Conference on Machine Learning*, 2012.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 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, 1995.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Qinyi Chen, Negin Golrezaei, and Djallel Bouneffouf. Dynamic bandits with temporal structure. *Available at SSRN 3887608*, 2021.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27, 2014.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pages 214–228. Springer, 2012.

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

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *J. Mach. Learn. Res.*, 23:32–1, 2022.

Jonas Umlauft and Sandra Hirche. Learning stable stochastic nonlinear dynamical systems. In *International Conference on Machine Learning*, pages 3502–3510. PMLR, 2017.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 757–762. IEEE, 2021.

Sartaj Sahni. Computationally related problems. *SIAM Journal on computing*, 3(4):262–279, 1974.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, 2008.

Fahad Albalawi, Zihang Dong, and David Angeli. Regret-based robust economic model predictive control for nonlinear dissipative systems. In *2021 European Control Conference (ECC)*, pages 1105–1111. IEEE, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

A Related Works

In this section, we discuss and compare the works that share similarities with the ARBs, focusing on MABs and online learning in non-linear systems.

Multi-Armed Bandits In the more classical Multi-Armed Bandit (MAB) setting, the learning problem does not involve temporal dependencies between successive rewards. The MAB setting has been studied under the assumptions of both *stochastic* and *adversarial* noise models. In the former case, `UCB1` (Lai and Robbins, 1985; Auer et al., 2002a) represents the parent algorithm. Instead, when adversarial noise is involved `EXP3` (Auer et al., 1995, 2002b) is usually employed. This algorithm has been extended by `REXP3` (Besbes et al., 2014) to handle with the *non-stationary* setting. Differently from both the adversarial and non-stochastic setting, we assume that the rewards are not preselected by an adversary or nature but, instead, they change as an effect of the actions played. Indeed, the underlying autoregressive process (affected by a stochastic noise) is such that the current action impacts the future rewards. Therefore, importing the adversarial MAB terminology, the ARBs can be reduced to an adversary setting with an *adaptive* (or non-oblivious) adversary (Dekel et al., 2012). In particular, the $\mathcal{O}(\sqrt{nT})$ regret guarantees of `EXP3` are not achievable in the ARB setting as `EXP3` competes against the best constant policy while the optimal policy for ARBs is not constant (Theorem 1). As we shall see empirically in Section 5, a constant policy suffers a linear regret.

Moreover, our setting presents similarities with MABs with *delayed* feedback (*e.g.*, Pike-Burke et al., 2018). However, in ARB the effect of the actions is propagated (not exactly delayed). Markov (Ortner et al., 2012) and restless (Tekin and Liu, 2012) bandits, instead, consider underlying processes that influence the rewards. However, these processes are not supposed to be controlled by the action history. In Chen et al. (2021), the authors study the problem of learning and control in a setting that considers temporal structure in the feedback, modeled as an AR(1) autoregressive process. However, the model poses significant constraints (*e.g.*, reflective boundaries), that our method does not require.

Online Learning in Non-Linear Systems The ARB setting is a specific case of a non-linear dynamical system. Although the literature related to this setting is wide, no work faces all problems that the ARB setting presents, including learning to control with regret guarantees. Mania et al. (2022) focus on learning the parameters of a particular class of non-linear systems. However, the approach is limited to estimation and no control algorithm is proposed. Similarly, Umlauft and Hirche (2017) deal with learning the system parameters with stability guarantees without the chance to control it. Several recent works (Kakade et al., 2020; Lale et al., 2021) focus on the learning and control of non-linear systems with regret guarantees. However, these works make use of an oracle to solve a complex optimization problem to perform optimistic planning (*i.e.*, optimal policy given an optimistic estimate of the system). This problem in a non-linear setting, however, is proven to be NP-hard (Sahni, 1974; Dani et al., 2008). Furthermore, the class of non-linear systems considered in these works does not include the ARB setting. Other works (*e.g.*, Albalawi et al., 2021) overcome the request for the oracle by searching in the restricted space of constant policies, leading to the best equilibrium. However, this solution can be suboptimal in several cases, including ARBs (*e.g.*, Section 5.2).

B Omitted Proofs

Theorem 1 (Optimal Policy)**.** *Under Assumption 1.a, for every round $t \in \mathbb{N}$, the optimal policy $\pi_t^*(H_{t-1})$ satisfies:*

$$\pi_t^*(H_{t-1}) \in \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \mathbf{z}_{t-1} \rangle. \tag{5}$$

Proof. We first prove an intermediate result auxiliary to get to the final statement. Let us denote with $J_T^*(\mathbf{z})$ the expected cumulative reward when the initial observations vector is $\mathbf{z} = (1, x_0, x_{-1}, \dots, x_{-k+1})$. Let us denote with $\succeq$ the element-wise inequality. We show that for every $T \in \mathbb{N}$, if $\mathbf{z} \succeq \overline{\mathbf{z}}$, then $J_T^*(\mathbf{z}) \geq J_T^*(\overline{\mathbf{z}})$.

We proceed by induction.

For $T = 1$, we have $J_1^*(\mathbf{z}) = \max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \mathbf{z} \rangle = \langle \boldsymbol{\gamma}(a_1^*), \mathbf{z} \rangle$, where $a_1^* \in \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \mathbf{z} \rangle$ and $J_1^*(\overline{\mathbf{z}}) = \max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \overline{\mathbf{z}} \rangle = \langle \boldsymbol{\gamma}(\overline{a}_1^*), \mathbf{z}' \rangle$, where $\overline{a}_1^* \in \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \overline{\mathbf{z}} \rangle$. Thus, we have:

$$J_1^*(\mathbf{z}) = \langle \boldsymbol{\gamma}(a_1^*), \mathbf{z} \rangle \geq \langle \boldsymbol{\gamma}(\overline{a}_1^*), \mathbf{z} \rangle \overset{\text{(a)}}{\geq} \langle \boldsymbol{\gamma}(\overline{a}_1^*), \overline{\mathbf{z}} \rangle = J_1^*(\overline{\mathbf{z}}),$$

where inequality (a) follows from Assumption 1.a.

Suppose the statement holds for $T-1$, we prove it for $T>1$. To this end, we consider the *transition operator* $P : \mathcal{Z} \times \mathcal{A} \times \mathbb{R} \to \mathcal{Z}$, defined for every observations vector $\mathbf{z}_t = (1, x_{t-1}, x_{t-2}, \dots, x_{t-k}) \in \mathcal{Z}$, action $a \in \mathcal{A}$, and noise $\xi \in \mathbb{R}$ as follows:

$$P(\mathbf{z}_t, a, \xi) = P\left(\begin{pmatrix} 1 \\ x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-k} \end{pmatrix}, a, \xi\right) = \begin{pmatrix} 1 \\ x_t \\ x_{t-1} \\ \vdots \\ x_{t-k+1} \end{pmatrix} = \mathbf{z}_{t+1}, \qquad \text{where} \qquad x_t = \langle \boldsymbol{\gamma}(a), \mathbf{z}_t \rangle + \xi.$$

Thus, we can look at the stochastic process as a Markov decision process (Puterman, 2014) with $\mathbf{z}_t$ as state representation. We immediately observe that if $\mathbf{z} \succeq \overline{\mathbf{z}}$, we have that $P(\mathbf{z}, a, \xi) \succeq P(\overline{\mathbf{z}}, a, \xi)$, for every action $a \in \mathcal{A}$ and noise $\xi \in \mathbb{R}$. By applying the Bellman equation, we obtain:

$$J_T^*(\mathbf{z}) = \max_{a \in \mathcal{A}} \left\{ \langle \boldsymbol{\gamma}(a), \mathbf{z} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{z}, a, \xi_T)) \right] \right\} = \langle \boldsymbol{\gamma}(a_T^*), \mathbf{z} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{z}, a_T^*, \xi_T)) \right],$$

$$J_T^*(\overline{\mathbf{z}}) = \max_{a \in \mathcal{A}} \left\{ \langle \boldsymbol{\gamma}(a), \overline{\mathbf{z}} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{z}}, a, \xi_T)) \right] \right\} = \langle \boldsymbol{\gamma}(\overline{a}_T^*), \overline{\mathbf{z}} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{z}} s, \overline{a}_T^*, \xi_T)) \right],$$

where the actions are defined as $a_T^* \in \arg\max_{a \in \mathcal{A}} \left\{ \langle \boldsymbol{\gamma}(a), \mathbf{z} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{z}, a, \xi_T)) \right] \right\}$ and $\overline{a}_T^* \in \arg\max_{a \in \mathcal{A}} \left\{ \langle \boldsymbol{\gamma}(a), \overline{\mathbf{z}} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{z}}, a, \xi_T)) \right] \right\}$. Thus, we have:

$$\begin{aligned} J_T^*(\mathbf{z}) &= \langle \boldsymbol{\gamma}(a_T^*), \mathbf{z} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{z}, a_T^*, \xi_T)) \right] \\ &\geq \langle \boldsymbol{\gamma}(\overline{a}_T^*), \mathbf{z} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{z}, \overline{a}_T^*, \xi_T)) \right] \\ &\overset{(b)}{\geq} \langle \boldsymbol{\gamma}(\overline{a}_T^*), \overline{\mathbf{z}} \rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{z}}, \overline{a}_T^*, \xi_T)) \right] = J_T^*(\overline{\mathbf{z}}), \end{aligned}$$

where (b) follows from Assumption 1.a when bounding $\langle \boldsymbol{\gamma}(\overline{a}_T^*), \mathbf{z} \rangle \geq \langle \boldsymbol{\gamma}(\overline{a}_T^*), \overline{\mathbf{z}} \rangle$ and by observing that $P(\mathbf{z}, \overline{a}_T^*, \xi_1) \succeq P(\overline{\mathbf{z}}, \overline{a}_T^*, \xi_T)$ and, then, exploiting the inductive hypothesis.

We conclude that the optimal policy is the myopic one by observing that both $\langle \boldsymbol{\gamma}(a), \boldsymbol{z} \rangle$ and $J_{T-1}^*(P(\mathbf{z}, a, \xi))$ are simultaneously maximized by $\arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\gamma}(a), \boldsymbol{z} \rangle$. □

Lemma 2 (Self-Normalized Concentration)**.** *Let $a \in \mathcal{A}$ be an action, let $(\widehat{\boldsymbol{\gamma}}_t(a))_{t \in \mathcal{O}_\infty(a)}$ be the sequence of solutions to the Ridge regression problems computed by Algorithm 1. Then, for every regularization parameter $\lambda > 0$, confidence $\delta \in (0,1)$, simultaneously for every round $t \in \mathbb{N}$ and action $a \in \mathcal{A}$, with probability at least $1-\delta$ it holds that:*

$$\|\widehat{\boldsymbol{\gamma}}_t(a) - \boldsymbol{\gamma}(a)\|_{\mathbf{V}_t(a)} \leq \sqrt{\lambda}\|\boldsymbol{\gamma}(a)\|_2 + \sigma\sqrt{2\log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(a)}{\lambda^{k+1}}\right)}.$$

Proof. We consider an action at a time; then, the final result is obtained with a union bound over $\mathcal{A} = [\![n]\!]$. Let $a \in \mathcal{A}$. We first observe that the estimates of action a change only when a is pulled. Let $l \in \mathbb{N}$ be an index and let $t_l(a) \in \mathbb{N}$ be the round in which action a is pulled for the l-th time, *i.e.*, $\{t_l(a) : l \in \mathbb{N}\} = \mathcal{O}_\infty(a)$. Thus, we have:

$$\begin{aligned} \boldsymbol{\gamma}_{t_l}(a) &= \mathbf{V}_{t_l(a)}^{-1}(a)\mathbf{b}_{t_l(a)}^{-1}(a) \\ &= \left(\lambda \mathbf{I}_{k+1} + \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1}\mathbf{z}_{t_j(a)-1}^T\right)^{-1} \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1} x_{t_j} \\ &= \left(\lambda \mathbf{I}_{k+1} + \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1}\mathbf{z}_{t_j(a)-1}^T\right)^{-1} \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1} \left(\langle \boldsymbol{\gamma}(a), \mathbf{z}_{t_j(a)-1} \rangle + \xi_{t_j(a)}\right) \\ &\overset{(a)}{=} \boldsymbol{\gamma}(a) - \lambda \left(\lambda \mathbf{I}_{k+1} + \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1}\mathbf{z}_{t_j(a)-1}^T\right)^{-1} \boldsymbol{\gamma}(a) + \\ &\qquad + \left(\lambda \mathbf{I}_{k+1} + \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1}\mathbf{z}_{t_j(a)-1}^T\right)^{-1} \sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1} \xi_{t_j(a)} \end{aligned}$$

$$= \boldsymbol{\gamma}(a) - \lambda \mathbf{V}_{t_l(a)}^{-1}(a)\boldsymbol{\gamma}(a) + \mathbf{V}_{t_l(a)}^{-1}(a) \underbrace{\sum_{j=1}^{l} \mathbf{z}_{t_j(a)-1}\xi_{t_j(a)}}_{\mathbf{s}_{t_l(a)}},$$

where the passage (a) derives from the observation that $\sum_{j=1}^{l} \mathbf{z}_{t_j-1}(\langle \boldsymbol{\gamma}(a), \mathbf{z}_{t_j-1}\rangle) = \sum_{j=1}^{l} \mathbf{z}_{t_j-1}\mathbf{z}_{t_j-1}^T \boldsymbol{\gamma}(a)$. Thus, we have:

$$\left\|\boldsymbol{\gamma}_{t_l(a)}(a) - \boldsymbol{\gamma}(a)\right\|_{\mathbf{V}_{t_l(a)}(a)} \leq \sqrt{\lambda}\|\boldsymbol{\gamma}(a)\|_2 + \|\mathbf{s}_{t_l(a)}\|_{\mathbf{V}_{t_l(a)}^{-1}(a)}.$$

Let us denote with $\mathcal{F}_{t_l(a)} = \sigma(\mathbf{z}_0, a_1, \mathbf{z}_1, a_2, \ldots, \mathbf{z}_{t_l(a)-1}, a_{t_l(a)})$ be the filtration generated by all events realized at round $t_l(a)$. Let us now consider the stochastic processes $(\xi_{t_l(a)})_{l\in\mathbb{N}}$ and $(\mathbf{z}_{t_l(a)-1})_{l\in\mathbb{N}}$. We observe that $\xi_{t_l(a)}$ is $\mathcal{F}_{t_l(a)}$-measurable and conditionally σ^2-subgaussian and that $\mathbf{z}_{t_l(a)-1}$ is $\mathcal{F}_{t_l(a)-1}$-measurable. By applying Theorem 1 of Abbasi-Yadkori et al. (2011), we have that simultaneously for all $l \in \mathbb{N}$, w.p. $1-\delta$:

$$\|\mathbf{s}_{t_l(a)}\|_{\mathbf{V}_{t_l(a)}^{-1}(a)} \leq \sigma\sqrt{2\log\frac{1}{\delta} + \log\frac{\det \mathbf{V}_{t_l(a)}(a)}{\lambda^{k+1}}}.$$

Clearly, this hold for the rounds $t \in \mathbb{N}$ in which the action a is not pulled, since the corresponding estimates do not change. □

Lemma 3 (Policy Regret Decomposition)**.** *Let $(x_t^*)_{t\in[\![T]\!]}$ be the sequence of rewards by executing the optimal policy $\boldsymbol{\pi}^*$ and let $(x_t)_{t\in[\![T]\!]}$ be the sequence of rewards by executing the learner's policy $\boldsymbol{\pi}$. Then, for every $t \in [\![T]\!]$ it holds that:*

$$r_t = x_t^* - x_t = \sum_{i=1}^{k} \gamma_i(a_t^*)(x_{t-i}^* - x_{t-i}) + \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle = \sum_{i=1}^{k} \gamma_i(a_t^*) r_{t-i} + \rho_t, \quad (9)$$

where $r_t \coloneqq x_t^* - x_t$ is the instantaneous policy regret, $\rho_t \coloneqq \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle$ is the instantaneous external regret, $a_t^* = \pi_t^*(H_{t-1}^*)$, and $r_{t-i} = 0$ if $i \geq t$.

Proof. Let $t \in [\![T]\!]$ and let us denote with $\mathbf{z}_{t-1}^* = (1, x_{t-1}^*, \ldots, x_{t-k}^*)^T$ the observations vector associated with the execution of the optimal policy and with $\mathbf{z}_{t-1} = (1, x_{t-1}, \ldots, x_{t-k})^T$ the observations vector associated with the execution of the learner's policy. We have:

$$\begin{aligned}
r_t &= x_t^* - x_t \\
&= \langle \boldsymbol{\gamma}(a_t^*), \mathbf{z}_{t-1}^*\rangle - \langle \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle \\
&= \langle \boldsymbol{\gamma}(a_t^*), \mathbf{z}_{t-1}^*\rangle - \langle \boldsymbol{\gamma}(a_t^*), \mathbf{z}_{t-1}\rangle + \langle \boldsymbol{\gamma}(a_t^*), \mathbf{z}_{t-1}\rangle - \langle \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle \\
&= \langle \boldsymbol{\gamma}(a_t^*), \mathbf{z}_{t-1}^* - \mathbf{z}_{t-1}\rangle + \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle \\
&= \sum_{i=1}^{k} \gamma_i(a_t^*) \underbrace{(x_{t-i}^* - x_{t-i})}_{r_{t-i}} + \underbrace{\langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle}_{\rho_t},
\end{aligned}$$

where in expanding the inner product we made the summation start from $i = 1$ as the two vectors $\mathbf{z}_{t-1}^*$ and $\mathbf{z}_{t-1}$ have the same first component equal to 1. □

Lemma 4 (External-to-Policy Regret Bound)**.** *Let $\boldsymbol{\pi}$ be the learner's policy and $T \in \mathbb{N}$ be the horizon. Under Assumptions 1.a and 1.b, it holds that:*

$$R(\boldsymbol{\pi}, T) = \mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{k} \gamma_i(a_t^*) r_{t-i} + \rho_t\right]\right] \leq \left(\frac{\Gamma k}{1-\Gamma} + 1\right) \varrho(\boldsymbol{\pi}, T), \quad (10)$$

where $\varrho(\boldsymbol{\pi}, T) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \rho_t\right]$ is the cumulative expected external regret.

Proof. We start from the decomposition of Lemma 3. To prove the result we employ the so-called "superposition principle", which allows us to decompose the linear recurrence as follows:

$$r_t = \sum_{i=1}^{k} \gamma_i(a_t^*) r_{t-i} + \rho_t = \sum_{\tau=0}^{+\infty} \rho_\tau \widetilde{r}_{t,\tau},$$

where if $\tau > t$ we set $\widetilde{r}_{t,\tau} = 0$ and if $\tau \leq t$ we have that $\widetilde{r}_{t,\tau}$ is given by the recurrence:

$$\widetilde{r}_{t,\tau} = \sum_{i=1}^{k} \boldsymbol{\gamma}_i(a_t^*)\widetilde{r}_{t-i,\tau} + \delta_{t,\tau} \qquad \text{where} \qquad \delta_{t,\tau} := \begin{cases} 1 & t = \tau \\ 0 & t \neq \tau \end{cases} .$$

This way, we decompose the exogenous term ρ_τ as a linear combination of unitary impulses. Then by Assumption 1.a and 1.b, recalling that $\widetilde{r}_{t,\tau} = 0$ if $\tau > t$ and that $\widetilde{r}_{\tau,\tau} = 1$, we have that for every $t > \tau$ it holds that:

$$\widetilde{r}_{t,\tau} \leq \Gamma \max_{i \in [\![k]\!]} \widetilde{r}_{t-i,\tau} \leq \Gamma^2 \max_{i \in [\![k]\!]} \max_{j \in [\![k]\!]} \widetilde{r}_{t-i-j,\tau} \leq \cdots \leq \Gamma^{\lceil (t-\tau)/k \rceil},$$

since we will encounter the $1 = \delta_{\tau,\tau}$ after $\lceil (t-\tau)/k \rceil$ steps of unfolding.

Now, we can manipulate this formula to have an expression of the full regret:

$$\begin{aligned}
\sum_{t=1}^{T} r_t &\leq \sum_{t=1}^{T} \left(\rho_t + \sum_{\tau=1}^{t-1} \Gamma^{\lceil (t-\tau)/k \rceil} \rho_\tau \right) \\
&= \sum_{\tau=1}^{T} \left(1 + \rho_\tau \sum_{t=\tau+1}^{T} \Gamma^{\lceil (t-\tau)/k \rceil} \right) \\
&\overset{(a)}{\leq} \sum_{\tau=1}^{T} \rho_\tau \left(1 + \sum_{s=1}^{+\infty} \Gamma^{\lceil s/k \rceil} \right) \\
&\overset{(b)}{=} \sum_{\tau=1}^{T} \rho_\tau \left(1 + \sum_{l=1}^{+\infty} k\Gamma^{l} \right) \\
&= \left(1 + \frac{\Gamma k}{1-\Gamma} \right) \sum_{\tau=1}^{T} \rho_\tau,
\end{aligned}$$

where (a) follows from bounding the summation with the series and changing the index $s = t - \tau$ and (b) is obtained by observing that the exponent $\lceil s/k \rceil$ changes only when s is divisible by k. □

Counterexample to show that this bound is tight There are k arms

$$\boldsymbol{\gamma}(a_1) := [\Gamma, 0 \ldots 0], \qquad \boldsymbol{\gamma}(a_2) := [0, \Gamma, 0 \ldots 0], \;\; \ldots \;\; \boldsymbol{\gamma}(a_k) := [0, \ldots 0, \Gamma].$$

All these arms have nonnegative coeffcients whose sum is bounded by Γ. If the sequence of internal regrets is

$$\rho_t = \begin{cases} 1 & t = 1 \\ 0 & t > 1 \end{cases}$$

and the sequence of arms is $a_1^* = 1$, and $a_t^* = a_{t-1 \ (mod\ k)+1}$ (which means $a_1, a_2, ..., a_k, a_1, a_2 \ldots$), we have

$$r_1 = 1, r_2 = \Gamma, r_3 = \Gamma, \ldots r_{k+1} = \Gamma$$

and then, we start again with the same sequence of arms

$$r_{k+2} = \Gamma^2, r_{k+3} = \Gamma^2, \ldots r_{2k+1} = \Gamma^2.$$

Making the sum of this terms for t from one to infinity we get

$$\sum_{t=1}^{\infty} r_t = 1 + k \sum_{t=1}^{\infty} \Gamma^t = 1 + \frac{k\Gamma}{1-\Gamma},$$

which is exactly the bound we get.

Lemma 6. *Let $(\mathbf{z}_t)_{t\in[\![T]\!]}$ be the sequence of observations vectors observed by executing the learner's policy. If $\mathbf{z}_0 = (1, 0, \dots, 0)^T$, then, for every $\delta \in (0, 1)$, with probability at least $1-\delta$, simultaneously for all $t \in [\![T]\!]$, it holds that:*

$$\|\mathbf{z}_{t-1}\|_2 \leq \sqrt{1 + k\left(\frac{m+\eta}{1-\Gamma}\right)^2},$$

where $\eta = \sqrt{2\sigma^2 \log(T/\delta)}$.

Proof. Let $(\xi_t)_{t\in[\![T]\!]}$ be the sequence of noises. We consider the event $\mathcal{E} = \bigcap_{t=1}^{T} \{|\xi_t| \leq \eta\}$ prescribing that all noises are smaller than η in absolute value. By union bound, knowing that all the noises are independent σ^2-subgaussian random variables we, can bound the probability of event $\mathcal{E}$:

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}\left(\bigcap_{t=1}^{T} \{|\xi_t| \leq \eta\}\right) \geq 1 - Te^{-\frac{\eta^2}{2\sigma^2}} = 1 - \delta,$$

having set $\eta = \sqrt{2\sigma^2 \log(T/\delta)}$. Under event $\mathcal{E}$ and when $\mathbf{z}_0 = (1, 0, \dots, 0)^T$, we prove by induction that all rewards x_t are bounded in absolute value by $\frac{m+\eta}{1-\Gamma}$, regardless the actions played. For $T = 1$, the statement is trivial since $x_1 = \gamma_0(a_1) + \eta_1$ and, thus, $|x_1| \leq \gamma_0(a_1) + |\eta_1| \leq m + \eta \leq \frac{m+\eta}{1-\Gamma}$. Suppose the statement holds for all $s < t$, we prove it for t. We have:

$$\begin{aligned} x_t = \gamma_0(a_t) + \sum_{i=1}^{k} \gamma_i(a_t) x_{t-i} + \eta_t \quad \implies |x_t| &\leq \gamma_0(a_t) + \sum_{i=1}^{k} \gamma_i(a_t)|x_{t-i}| + |\eta_t| \\ &\leq m + \Gamma\frac{m+\Gamma}{1-\Gamma} + \eta = \frac{m+\eta}{1-\Gamma}, \end{aligned}$$

where the first inequality uses Assumption 1.a, the second inequality follows from the inductive hypothesis and by Assumptions 1.b and 1.c. Passing to the observations vector, we have:

$$\|\mathbf{z}_{t-1}\|_2^2 = 1 + \sum_{i=1}^{k} x_{t-i}^2 \leq 1 + k\left(\frac{m+\eta}{1-\Gamma}\right)^2.$$

□

For deriving the regret bound, we make use of the following result, known as *Elliptic Potential Lemma* (Lattimore and Szepesvári, 2020, Lemma 19.4).

Lemma 7 (Elliptic Potential Lemma)**.** *Let $\mathbf{V}_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix and let $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^d$ be a sequence of vectors such that $\|\mathbf{a}_t\|_2 \leq L < +\infty$ for all $t \in [\![n]\!]$. Let $\mathbf{V}_t = \mathbf{V}_0 + \sum_{s=1}^{t} \mathbf{a}_s\mathbf{a}_s^T$, Then:*

$$\sum_{t=1}^{n} \min\{1, \|\mathbf{a}_s\|_{\mathbf{V}_{t-1}^{-1}}\} \leq 2d \log\left(\frac{\mathrm{tr}(\mathbf{V}_0) + nL^2}{d\det(\mathbf{V}_0)^{1/d}}\right).$$

Theorem 5. *Let $\delta = (2T)^{-1}$. Under Assumptions 1.a, 1.b, and 1.c, `AR-UCB` suffers a cumulative expected (policy) regret bounded by (highlighting the dependence on m, σ, k, Γ, n, and T only):*

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

Proof. We denote with $(x_t^*)_{t\in[\![T]\!]}$ the sequence of rewards generated by playing the optimal policy and with $(x_t)_{t\in[\![T]\!]}$ the sequence of rewards generated by playing `AR-UCB`. Thanks to Lemma 4, we have to bound the external regret only. Let $\delta \in (0, 1)$, and define, as in the main paper, for every round $t \in [\![T]\!]$ and action $a \in \mathcal{A}$:

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

Let us define the confidence set $\mathcal{C}_t(a) \coloneqq \{\boldsymbol{\gamma} \in \mathbb{R}^{k+1} : \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{t-1}(a)\|_{\mathbf{V}_{t-1}(a)} \leq \beta_{t-1}(a)\}$ and the optimistic estimate of the true parameter vector $\boldsymbol{\gamma}(a)$:

$$\widetilde{\boldsymbol{\gamma}}_t(a) \in \arg\max_{\boldsymbol{\gamma} \in \mathcal{C}_t(a)} \langle \boldsymbol{\gamma}, \mathbf{z}_{t-1} \rangle,$$

By Theorem 2, we have that, for every action $a \in \mathcal{A}$ and round $t \in [\![T]\!]$, the true parameter vector satisfies $\boldsymbol{\gamma}(a) \in \mathcal{C}_t(a)$ with probability at least $1 - \delta$. Therefore, with the same probability, we have:

$$\begin{aligned}
\langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle &= \underbrace{\langle \boldsymbol{\gamma}(a_t^*) - \widetilde{\boldsymbol{\gamma}}_t(a_t), \mathbf{z}_{t-1} \rangle}_{\leq 0} + \langle \widetilde{\boldsymbol{\gamma}}_t(a_t) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle \\
&\leq \langle \widetilde{\boldsymbol{\gamma}}_t(a_t) - \widehat{\boldsymbol{\gamma}}_{t-1}(a_t), \mathbf{z}_{t-1} \rangle + \langle \widehat{\boldsymbol{\gamma}}_{t-1}(a_t) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle \\
&\leq 2\beta_{t-1}(a_t) \|\mathbf{z}_{t-1}\|_{\mathbf{V}_{t-1}(a)^{-1}},
\end{aligned}$$

where the first inequality follows from the optimism and in the last passage we have used Cauchy-Schwartz inequality, recalling that for every couple of vectors $\mathbf{v}, \mathbf{w}$ it holds $\langle \mathbf{v}, \mathbf{w} \rangle \leq \|\mathbf{v}\|_{\mathbf{V}_{t-1}(a)} \|\mathbf{w}\|_{\mathbf{V}_{t-1}(a)^{-1}}$, and having observed that $\boldsymbol{\gamma}(a_t), \widetilde{\boldsymbol{\gamma}}_t(a_t) \in \mathcal{C}_t(a_t)$.

Furthermore, we observe that the external regret $\rho_t = \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle \leq \|\boldsymbol{z}_{t-1}\|_2 + m$, since the coefficients γ_j for $j \neq 0$ have a sum bounded by $\Gamma < 1$ and get multiplied by $\mathbf{z}_{t-1}$, while γ_0, which is bounded by m gets multiplied by 1, then we have $\rho_t \leq L + m = \mathcal{O}(m)$. By Lemma 6 with probability of at least $1 - \delta$ we have:

$$\|\boldsymbol{z}_t\|_2 \leq \sqrt{1 + k \left(\frac{m + \eta}{1 - \Gamma} \right)^2} =: L,$$

where $\eta = \sqrt{2\sigma^2 \log(T/\delta)}$ and, consequently:

$$\rho_t \leq m + L =: C_1.$$

At this point, we proceed as follows:

$$\rho_t \leq 2 \min\{C_1, \beta_{t-1}(a_t) \|\mathbf{z}_{t-1}\|_{\mathbf{V}_{t-1}(a_t)^{-1}}\} \leq 2 \max\{C_1, \beta_{t-1}(a_t)\} \min\{1, \|\mathbf{z}_{t-1}\|_{\mathbf{V}_{t-1}(a_t)^{-1}}\}.$$

Summing over $t \in [\![T]\!]$, we obtain a bound on the cumulative external regret:

$$\begin{aligned}
\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t &= \sum_{t=1}^{T} 1 \cdot \rho_t \\
&\leq \sqrt{T \sum_{t=1}^{T} \rho_t^2} \\
&\leq 2 \max\{C_1, \beta_{T-1}\} \sqrt{T \sum_{t=1}^{T} \min\{1, \|\mathbf{z}_{t-1}\|^2_{\mathbf{V}_{t-1}(a_t)^{-1}}\}}
\end{aligned}$$

where

$$\beta_{T-1} \coloneqq \max_{a \in \mathcal{A}} \beta_{T-1}(a),$$

where the first inequality follows from an application of Cauchy-Schwartz inequality and the last passage holds since the sequence $\beta_t(a_t)$ is non-decreasing, and so we can bound each of them with their value at $t = T$. Now, we are finally able to use the *Elliptic Potential Lemma* (Lemma 7):

$$\begin{aligned}
\sum_{t=1}^{T} \min\{1, \|\mathbf{z}_{t-1}\|^2_{\mathbf{V}_{t-1}(a_t)^{-1}}\} &= \sum_{a \in \mathcal{A}} \sum_{l \in \mathcal{O}_T(a)} \min\{1, \|\mathbf{z}_{l-1}\|^2_{\mathbf{V}_{l-1}(a)^{-1}}\} \\
&\leq \sum_{a \in \mathcal{A}} 2(k+1) \log \left(\frac{\lambda(k+1) + |\mathcal{O}_T(a)| L^2}{\lambda(k+1)} \right) \\
&\leq 2n(k+1) \log \left(1 + \frac{TL^2}{n\lambda(k+1)} \right),
\end{aligned}$$

where the first inequality follows from an application of the elliptic potential lemma for each action $a \in \mathcal{A}$ observing that $\mathbf{V}_0 = \lambda \mathbf{I}_{k+1}$ and, consequently, $\mathrm{tr}(\mathbf{V}_0) = \lambda(k+1)$ and $\det(\mathbf{V}_0)^{1/(k+1)} = \lambda$. The second inequality follows by observing that $\sum_{a \in \mathcal{A}} |\mathcal{O}_T(a)| = T$ and since the log is a concave function, the worst allocation of pulls is the uniform one. Now that we have bounded the inner

summation, we can state that:

$$\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t \leq 2\max\{C_1, \beta_{T-1}\}\sqrt{2Tn(k+1)\log\left(1 + \frac{TL^2}{n\lambda(k+1)}\right)}.$$

To conclude, we bound the term β_{T-1} as follows:

$$\begin{aligned}\beta_{T-1} &= \sqrt{\lambda(m^2+1)} + \sigma \max_{a\in\mathcal{A}} \sqrt{2\log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_{T-1}(a)}{\lambda^{k+1}}\right)} \\ &\leq \sqrt{\lambda(m^2+1)} + \sigma\sqrt{2\log\left(\frac{n}{\delta}\right) + (k+1)\log\left(\frac{\lambda(k+1)+TL^2}{\lambda(k+1)}\right)}.\end{aligned}$$

Therefore, by highlighting the dependences on m, k, σ, and Γ, we have:

$$\beta_{T-1} = \widetilde{O}\left(m + \sigma\sqrt{k+1}\right), \qquad C_1 = \widetilde{O}\left(1 + \sqrt{k}\frac{m+\sigma}{1-\Gamma}\right).$$

These results hold with probability $1-2\delta$. We set $\delta = (2T)^{-1}$. Putting all together, we obtain:

$$\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t \leq \widetilde{O}\left(\frac{(m+\sigma)\sqrt{n(k+1)T}}{1-\Gamma}\right),$$

and, applying the previous Lemma 4, this results in:

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

□

C Optimal Policy without Noise

In the case of no noise, our system writes:

$$x_t = \gamma_0(a_t) + \sum_{i=1}^{k} \gamma_i(a_t) x_{t-i}. \tag{12}$$

In this case, the process evolution is deterministic. Therefore, even if it is still true that the optimal policy is given by Theorem 1, it is possible to say that there is a constant policy that is asymptotically optimal, in the sense that its cumulative regret is bounded by a constant. This policy is given by:

$$a^* \in \arg\max_{a\in\mathcal{A}} \frac{\gamma_0(a_t)}{1 - \sum_{i=1}^{k}\gamma_i(a_t)}. \tag{13}$$

This result is not surprising. In fact, this action makes the process converge to the highest possible stationary reward, which is of course $\arg\max_{a\in\mathcal{A}} \frac{\gamma_0(a_t)}{1-\sum_{i=1}^{k}\gamma_i(a_t)}$. Precisely, the following result holds.

Theorem 8. *Let us consider the problem formulation of Equation* (12)*. Define:*

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

as in Equation (13)*. Then, there exist no policy $\boldsymbol{\pi}$ (even non-constant) such that:*

$$\limsup_{t\to+\infty} x_t^{\boldsymbol{\pi}} - x_t^* > 0$$

(where $x_t^{\boldsymbol{\pi}}$ denotes the sequence obtained with policy $\boldsymbol{\pi}$, while x_t^* is the one relative to a^*). Moreover, the cumulative regret with respect to the actual optimal policy is bounded by:

$$\gamma_0(a^*)\frac{k}{(1-\Gamma)^2}.$$

Proof. If we play always a^*, we have:

$$\limsup_{t\to+\infty} x_t^* = \frac{\gamma_0(a^*)}{1-\sum_{i=1}^k \gamma_i(a^*)},$$

by imposing the condition of stationarity. For the rest of the proof, let us denote:

$$x^* := \frac{\gamma_0(a^*)}{1-\sum_{i=1}^k \gamma_i(a^*)}.$$

Now, we prove that, for any policy $\boldsymbol{\pi}$, we cannot achieve an $x_t > x^*$. By contradiction, if $\limsup_{t\to\infty} x_t^{\boldsymbol{\pi}} - x_t^* > 0$, then the set $\{t \in \mathbb{N} : x_t > x^*\}$ is non-empty. Let $t_0 = \min\{t \in \mathbb{N} : x_t > x^*\}$. Then, by definition:

$$x_{t_0} = \gamma_0(a_{t_0}) + \sum_{i=1}^k \gamma_i(a_{t_0}) x_{t_0-i}.$$

Recalling that t_0 is the first time in which we surpass x^*, we have:

$$x^* < x_{t_0} = \gamma_0(a_{t_0}) + \sum_{i=1}^k \gamma_i(a_{t_0}) x_{t_0-i} \leq \gamma_0(a_{t_0}) + \sum_{i=1}^k \gamma_i(a_{t_0}) x^*.$$

This inequality entails that:

$$\Big(1 - \sum_{i=1}^k \gamma_i(a_{t_0})\Big) x^* < \gamma_0(a_{t_0}),$$

and, therefore:

$$\frac{\gamma_0(a^*)}{1-\sum_{i=1}^k \gamma_i(a^*)} = x^* < \frac{\gamma_0(a_{t_0})}{1-\sum_{i=1}^k \gamma_i(a_{t_0})},$$

which contradicts the definition of a^*.

For the second part, we start considering that the regret obtained by using constant action a^* is bounded by:

$$\sum_{t=1}^{+\infty} x^* - x_t,$$

since x^* is the maximum instantaneous reward that every policy can achieve. Now, note that $\gamma_0(a^*) > 0$, otherwise it could not be the optimal action. At this point, we have for $0 < t \leq k$ that $x_t \geq \gamma_0(a^*)$, by simply using the fact that all the coefficients of the autoregressive model are non-negative. From this fact we have for $k < t \leq 2k$ that $x_t \geq \gamma_0(a^*)(1 + \sum_{i=1}^k \gamma_i(a^*))$; and generalizing:

$$\forall j > 0 \qquad \text{and} \qquad jk - k < t \leq jk : \qquad x_t \geq \gamma_0(a^*)\Big(\sum_{\ell=0}^{j} (\Gamma^*)^\ell\Big), \qquad \Gamma^* = \sum_{i=1}^k \gamma_i(a^*).$$

Therefore, we have $x_t \geq \gamma_0(a^*)\frac{1-\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma}$, which means:

$$\begin{aligned}
R_t &\leq \sum_{t=1}^{+\infty} x^* - x_t \\
&\leq \sum_{t=1}^{+\infty} x^* - \gamma_0(a^*)\frac{1-\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma} \\
&= \gamma_0(a^*) \sum_{t=1}^{+\infty} \frac{1}{1-\Gamma} - \frac{1-\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma} \\
&= \gamma_0(a^*) \sum_{t=1}^{+\infty} \frac{\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma}
\end{aligned}$$

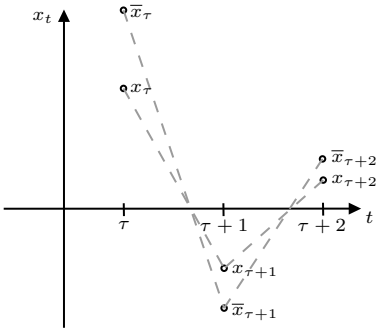


Figure 5: An illustration of the effect of a negative $\gamma_1(a)$ over time.

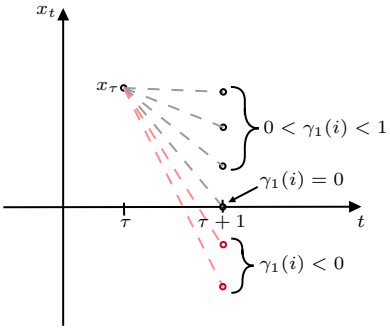


Figure 6: The effect of $\gamma_1(a)$ in the evolution of the state x_t, in the case of a non-negative one (in black), and a negative one (in red).

$$= \gamma_0(a^*)\frac{k}{(1-\Gamma)^2}.$$

□

D Discussion on Assumption 1.a

In this section, we further detail the meaning of Assumption 1.a related to non-negative coefficients governing the AR process (Assumption 1.a). Even if, theoretically, this setting is less general than the one which considers all possible values for the parameters, we argue that, for the real-world applications of interest, considering negative coefficients is not meaningful.

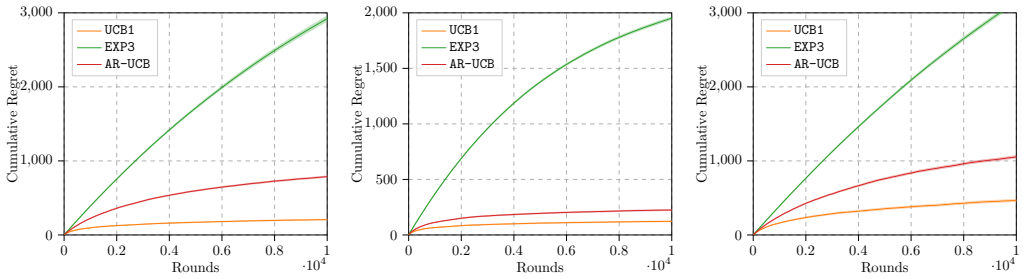
Before introducing our example, let us remark on the meaning of x_t in practice. This value represents the sales volume in the case of pricing, the value of a stock in the stock market, the number of customers that an e-commerce website may have, and so on. In all these real-world scenarios, the quantity x_t is meaningful whenever we consider non-negative values that we want to maximize. We argue that when Assumption 1.a is not fulfilled (*i.e.*, at least one $\gamma_i(a)$ is negative), the positivity of x_t is no longer ensured.

Consider now the example presented in Figure 5, where we present a general scenario in which, at time τ, we are in a given with a certain positive x_τ. Consider, for the sake of simplicity, a noiseless setting with $k = 1$ (*i.e.*, an AR(1) process) and, for a given action i, we have $\gamma_0(a) = 0$. Consider now $\gamma_1(a) < 0$. Figure 5 shows what will happen in this case. The value of x_t continuously changes its sign at each time step, and this behavior is not compatible with the real-world phenomena of our interest. This is even more unrealistic if we think about the scenario in which we have another value of the state $\overline{x}_\tau > x_\tau$. In this scenario, after performing the same action i, we will observe that the best-starting state $\overline{x}_\tau$ leads to a worst next state $\overline{x}_{\tau+1} < x_{\tau+1}$. This behavior has no practical meaning in the applications of our interest. Given these considerations, we can derive that the worst possible effect of a given action is to *reset* the state, which corresponds to have $\gamma_1(a) = 0$. A representation of this phenomenon is drawn in Figure 6. From this figure, it is possible to notice how a process can always decrease as an effect of an action, even for $\gamma_1(a) > 0$.

This consideration trivially generalizes for any $k > 1$ given a generic state representation $\mathbf{z}_\tau$.

E Additional Experimental Results

In this section, we provide additional experimental results. In Section E.1, we assert the effectiveness of `AR-UCB` in the classic stochastic bandit problem by comparing its performances with two standard baselines from the literature. In Section E.2, we stress the effect of misspecifying parameter $\overline{k}$ in the standard multi-armed bandit problem. Finally, in Section E.3, we provide experimental results in the particular case of autoregressive processes of order 1 (*i.e.*, $k = 1$).



(a) Setting A. (b) Setting B. (c) Setting C.

Figure 7: Cumulative regret of AR-UCB, UCB1, and EXP3 in the case of $k = 0$ (100 runs, mean $\pm$ std).

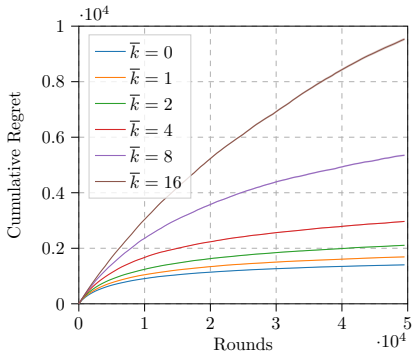


Figure 8: Cumulative regret of AR-UCB in the case of $k = 0$, in with $\overline{k}$ parameter misspecified (100 runs, mean $\pm$ std).

E.1 Stochastic Bandit Problem

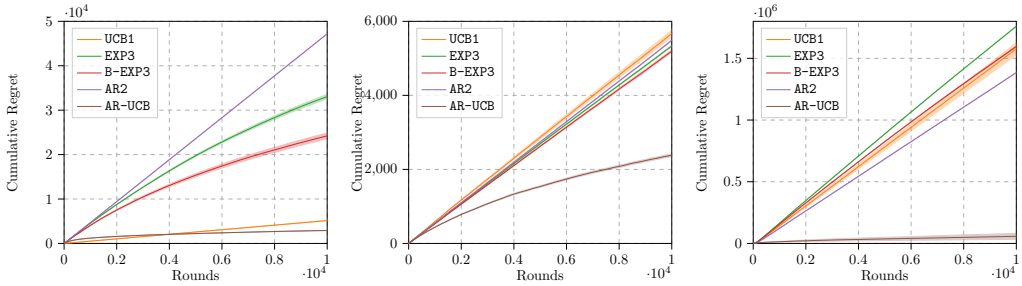
Setting We evaluate AR-UCB in the special case $k = 0$. This problem is equivalent to solving a standard stochastic bandit problem. This experiment compares the performances of AR-UCB in this setting against well-known gold standards: UCB1 and EXP3. The competing algorithms are evaluated in terms of cumulative regret w.r.t. to the setting-specific clairvoyant. The three settings differ in the values of $m \in \{2, 7.5\}$ (*i.e.*, the maximum arms' expected reward) and the values of $\sigma \in \{0.9, 1.25, 2\}$, the noise's standard deviation.

Results Figure 7 shows the average cumulative regrets for AR-UCB, UCB1, and EXP3. We immediately observe that all the algorithms suffer sublinear cumulative regret, as expected since they are all able to provide no-regret theoretical guarantees in this setting. In all the experiments, UCB1 outperforms all the other algorithms since it is specifically designed for the scenario under analysis. AR-UCB, as expected, performs properly in this setting since, as already discussed in Section 4.3, its regret is asymptotically optimal when $k = 0$.

E.2 On the Misspecification of k in Stochastic Bandit Problem

Setting We evaluate AR-UCB in the special case $k = 0$. This problem is equivalent to solving a standard stochastic bandit problem. This experiment compares the performances of AR-UCB under different values of the parameter $\overline{k}$. In particular, this experiment aims to highlight the performances of AR-UCB under a misspecification of the process memory in the special case where the true underlying process does not present a dynamic temporal structure. The parameters $\boldsymbol{\gamma}_0(a)$ have been sampled by a uniform distribution having support $[6, 7]$, and m is set to 10. The noise's standard deviation σ is set to 1.

Results Figure 8 shows the average cumulative regrets for AR-UCB under different values of $\overline{k}$, when the true value is $k = 0$. The figure shows that AR-UCB is capable of achieving sublinear cumulative regret even when the misspecification is severe (*e.g.*, $\overline{k} = 16$), coherently with the theoretical results, the performance degrades as the misspecification grows.



(a) Setting A. (b) Setting B. (c) Setting C.

Figure 9: Cumulative regret of AR-UCB and the others bandit baselines in the case of $k = 1$ (100 runs, mean $\pm$ std).

E.3 AR(1) Bandit Problem

AR(1) processes are the simplest autoregressive processes. Therefore, we will present a specific analysis of this setting to show how AR-UCB and the baselines perform when the complexity given by the dynamic temporal structure is minimal. Results show how even the minimal autoregressive contribution can lead all the algorithms (except for AR-UCB) to linear cumulative regret.

Setting We evaluate AR-UCB in the case $k = 1$. This is the simplest setting in which an autoregressive component contributes to the reward. This experiment compares the performances of AR-UCB in this setting against the same baselines as Section 5.1. The competing algorithms are evaluated in terms of cumulative regret w.r.t. to the setting-specific clairvoyant. The three settings differ in the values of $m \in \{2, 8, 10\}$ (*i.e.*, the maximum arms' expected reward) and the values of $\sigma \in \{1, 1.25, 2\}$, the noise's standard deviation. The values of the $\gamma_1(a)$ parameters have been sampled from uniform distributions having their sampling ranges inside $[0, 1)$.

Results Figure 9 shows the average cumulative regrets for all the competing algorithms. We immediately observe that the only algorithms able to achieve sublinear regret are AR-UCB (in all three settings), B-EXP3 (first and third experiments), and EXP3 (first experiment only). Such a result is unsurprising since none of the baselines has specific theoretical guarantees in the Autoregressive Bandit problem, even in the simple scenario when $k = 1$. Even though, we decided to adopt these algorithms as baselines since they represent the gold standard algorithms in the bandit literature (UCB1, EXP3) and the algorithms that solve problems near to ours (B-EXP3 , AR2), respectively.