Autoregressive Bandits

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

Abstract

Autoregressive processes naturally arise in a large variety of real-world scenarios, including e.g., stock markets, sell forecasting, weather prediction, advertising, and pricing. When addressing a sequential decision-making problem in such a context, the temporal dependence between consecutive observations should be properly accounted for converge to the optimal decision policy. In this work, we propose a novel online learning setting, named Autoregressive Bandits (ARBs), in which the observed reward follows an autoregressive process of order k, whose parameters depend on the action the agent chooses, within a finite set of n actions. Then, we devise an optimistic regret minimization algorithm `AutoRegressive Upper Confidence Bounds` (`AR-UCB`) that suffers regret of order $\widetilde{\mathcal{O}}\left(\frac{(k+1)^{3/2}\sqrt{nT}}{(1-\Gamma)^2}\right)$, being T the optimization horizon and $\Gamma < 1$ an index of the stability of the system. Finally, we present a numerical validation in several synthetic and one real-world setting, in comparison with general and specific purpose bandit baselines showing the advantages of the proposed approach.

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 next reward. This process, of arbitrary nature in general, is influenced by the 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 is that of the *autoregressive* (AR, Hamilton, 2020) processes. In this paper, we model the reward of a sequential decision-making problem as an AR process where its parameters depend on the action selected by the agent at every round. This scenario can be regarded as an extension of the *multi-armed bandit* (MAB, Lattimore & Szepesvári, 2020) problem, in which an AR process governs the temporal structure of the observed rewards that is, 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 MABs (Gur et al., 2014). Indeed, in the presented scenario, we can exploit the knowledge that the underlying process is AR and, more importantly, that such a dynamic depends on the agent's action.

Motivation Numerous real-world phenomena can be effectively represented as a sequential decision-making process under an AR model of the reward. Let us consider, for instance, the optimal *pricing* problem. Our task consists in deciding at which price to sell a given so as to maximize the seller's overall revenue. The pricing policy generates two effects. The first one is immediate and driven by the *demand curve*, which determines the probability of a sale given a price and is represented by a non-increasing function of the price (Mussi et al., 2022b). The second effect, instead, becomes apparent in the long term and consists of *customer loyalty*, which is highly influenced by the sequence of prices the customer observes in the recent past. As intuition suggests, customer loyalty is a desired achievement since a customer which buys a good at a convenient price will be more prone to come back in the future, with a possible increase in future revenue. An effective trade-off between immediate sales and customer loyalty can dramatically increase the overall revenue (Bowen & Chen, 2001). In this scenario, the current revenue (our reward) displays an unknown temporal structure, governed by both the immediate price (our action) and the previous sales and revenues. Thus, it can be effectively represented as an AR process influenced by the sequence of prices. To the best of our knowledge, no work in the dynamic pricing literature faces in an online fashion the degree of loyalty of the customers.

Original Contribution In this work, we propose a novel setting, named *AutoRegressive Bandit* (ARB), in which

[*]Equal contribution
[1]Politecnico di Milano, Milan, Italy.
Correspondence to: Alberto Maria Metelli
<albertomaria.metelli@polimi.it>.

Preprint. Under Review.

the reward follows an AR process, whose parameters depending on the agent's actions. In Section 2, we introduce the setting and the assumptions. Then, we derive the notion of optimal policy, showing that the best action varies based on model parameters and previously observed rewards. Then, in Section 3, we propose an optimistic algorithm, named `AutoRegressive Upper Confidence Bounds` (`AR-UCB`), to learn an optimal policy online, and, in Section 4, we present and discuss its regret guarantees. In Section 6, we test our solution in various synthetic environments to validate it in comparison with several bandit baselines, and we analyze the robustness of `AR-UCB` w.r.t. the misspecification of key parameters.

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. Problem Formulation

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

2.1. Setting

We consider the sequential interaction between a learner and an environment. At each round $t \in \mathbb{N}$, the learner chooses an action $a_t \in \mathcal{A} := [\![n]\!]$, among the $n \in \mathbb{N}$ available ones. 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:

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

where $\mathbf{z}_{t-1} := (1, x_{t-1}, \dots, x_{t-k})^T \in \mathcal{Z} := \{1\} \times \mathcal{X}^k$ is the *context vector* 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 presented setting reduces to a standard MAB (Auer et al., 2002a).

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

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

a. (Monotonicity) $\gamma_i(a) \geq 0$ for every $i \in [\![0, k]\!]$;
b. (Stability) $\sum_{i=1}^{k} \gamma_i(a) \leq \Gamma$ for some $\Gamma < 1$;
c. (Boundedness) $\gamma_0(a) \leq m$ for some $m < +\infty$.

Assumption 1.a enforces a *monotonic* evolution of the AR process when a constant action is played. This condition is typical of "accumulation" processes (Hamilton, 2020) and will play a relevant role in the derivation of the optimal policy (Section 2.2). Assumptions 1.b requires that the sum of $(\gamma_i(a))_{i \in [\![k]\!]}$ is bounded to a value $\Gamma \in [0, 1)$ and 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 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:

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

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 $\boldsymbol{\pi}^*$.

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.

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

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

The result deserves some comments. First, we observe that

the optimal action depends on the context vector $\mathbf{z}_{t-1}$ and, thus, on the most recent k observed rewards $x_{t-1}, \dots, x_{t-k}$ only. Therefore, the optimal policy $\boldsymbol{\pi}^*$ is a non-Markovian policy with memory k or, in a different view, a Markovian policy w.r.t. the state representation $\mathbf{z}_{t-1}$.[1] 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 monotonic evolution of the AR process. This way, the action maximizing the expected *immediate* reward (i.e., a *myopic* policy) is optimal for the expected *cumulative* reward too. This would not hold in the presence of negative parameters $\gamma_i(a)$ that would require planning in the future (i.e., a *far-sighted* policy). The proof can be found in Appendix A.

3. AutoRegressive Upper Confidence Bounds

In this section, we present `AutoRegressive Upper Confidence Bounds` (`AR-UCB`), an optimistic regret minimization algorithm for the AutoRegressive Bandit 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 $a \in \mathcal{A}$ independently. The algorithm requires the knowledge of the order k of the reward AR model, although this condition can be easily replaced with the knowledge of an upper bound $\overline{k} > k$ on the true AR order.[2] An empirical validation of the misspecification of such a parameter is postponed to Section 6.4.

`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 context vector $\mathbf{z}_0 = (1, 0, \dots, 0)^T$ (line 1).[3] 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 context 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).

Algorithm 1: `AR-UCB`.

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

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

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

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**

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 A.

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:

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

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.*

[1] 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.

[2] Indeed, any AR process of order k can be regarded as an AR process of order $\overline{k} > k$ setting to zero the additional parameters, i.e., $\gamma_i(a) = 0$ for $i \in [\![k+1, \overline{k}]\!]$.

[3] We assume to know the initial context 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.

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:

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

From a technical perspective, Lemma 2 is obtained by an adaptation of 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 $t \in [\![0, T-1]\!]$:

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

This formula contains two terms. The first one is a *bias* term that increases with m (Assumption 1.c) and with $\lambda > 0$, the regularization parameter of the Ridge regression. 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 $1-\delta$, simultaneously for all actions $a \in \mathcal{A}$ and rounds $t \in [\![0, T-1]\!]$, it holds that:

$$\|\widehat{\boldsymbol{\gamma}}_t(a) - \boldsymbol{\gamma}(a)\|_{\mathbf{V}_t(a)} \leq \beta_t(a). \tag{7}$$

We observe that thanks to the form of $\beta_t(a)$, AR-UCB, remarkably, does not require the knowledge of the upper bound Γ on the sum of the parameters (as in Assumption 1.b), although Γ will appear in the analysis (Section 4.2). Indeed, Γ is often unknown in practice and cannot be easily estimated from the data. Furthermore, AR-UCB requires, in order to compute the optimistic coefficient $\beta_t(a)$ the value of m, i.e., an upper-bound to the value of the largest $\gamma_0(a)$ over the actions $a \in \mathcal{A}$ (as in Assumption 1.c). An empirical analysis of the effect of the misspecification of such a parameter is provided in Section 6.3.

4.2. Regret Decomposition

In this section, we present a *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).

For the sake of the analysis and w.l.o.g., we compare the execution of the optimal policy $\boldsymbol{\pi}^*$ and the execution of the learner's policy $\boldsymbol{\pi}$ over the same sequence of noise realizations $(\xi_t)_{t \in [\![T]\!]}$. 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 $\boldsymbol{\pi}^*$ and 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: (i) the dissimilarity between the past history H_{t-1}^* when executing the optimal policy and the learner's observed history H_{t-1}; (ii) 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 context vector $\mathbf{z}_{t-1}$ generated by the execution of the learner's policy. The following result formalizes the 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:*

$$\begin{aligned} r_t &= x_t^* - x_t \\ &= \sum_{i=1}^k \gamma_i(a_t^*)(x_{t-i}^* - x_{t-i}) + \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1} \rangle \\ &= \sum_{i=1}^k \gamma_i(a_t^*) r_{t-i} + \rho_t, \end{aligned} \tag{8}$$

where $r_t \coloneqq x_t^* - x_t$ is the instantaneous policy regret, $\rho_t \coloneqq \langle \boldsymbol{\gamma}(a_t^*) - \boldsymbol{\gamma}(\widehat{a}_t), \mathbf{z}_{t-1} \rangle$ 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 (8) is made of two terms. The second one ρ_t is the instantaneous external regret whose meaning is discussed above. The first one, instead, defines a recurrence relation of order k on the instantaneous policy regret r_t. The following result shows that the contribution of this term can be reduced to a constant 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:*

$$R(\boldsymbol{\pi}, T) \leq \left(1 + \frac{\Gamma k}{1-\Gamma}\right) \varrho(\boldsymbol{\pi}, T), \tag{9}$$

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

Thanks to Lemma 4, we are able to provide 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 $\varrho(\boldsymbol{\pi}, T)$ only. 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. As expected, 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 autoregressive effects are, and, consequently, the more the sub-optimal choices of the past get amplified.

4.3. Regret Bound

This section is devoted to the derivation of the final regret bound for AR-UCB. Before providing the bound, we introduce a lemma ensuring that the norm of the context vectors $\boldsymbol{z}_{t-1}$ is bound in high probability over the whole horizon T.

Lemma 5. *Let $(\mathbf{z}_t)_{t\in[\![T]\!]}$ be the sequence of context 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 $\eta = \sqrt{2\sigma^2 \log(T/\delta)}$.

Finally, thanks to Lemma 4 and 5 we can prove that the following theorem holds.

Theorem 6. *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):*

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

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 the tight regret rate $\widetilde{O}(\sqrt{nT})$ (in that case m, the maximum expected reward, is usually set to 1). 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 6 with the dependence on $(k+1)^{3/2}$ and $(1-\Gamma)^{-1}$. The interplay between k and $(1-\Gamma)^{-1}$ is interesting showing that even if two instances have the same sum of parameters (i.e., Γ), the one with less coefficients is more easily learnable. Finally, suppose we run AR-UCB with a larger AR order $\overline{k} > k$. In such a case, the dependence on $(k+1)^{3/2}$ will be replaced by $(k+1)(\overline{k}+1)^{1/2}$, since the factor due to passing from external to policy regret (Lemma 4) will always contain the true order k, while $\overline{k}$ appears because of the estimation processes.

5. 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 et al., 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 $O(\sqrt{nT})$ 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 6, a constant policy suffers a linear regret. Nevertheless, by observing that the memory of the system is k, one can import the approach of (Dekel et al., 2012) for dealing with a finite-memory adaptive adversary, leading to $\widetilde{\mathcal{O}}((p+1)(CT)^{1/(2-q)})$ policy regret, where $\widetilde{\mathcal{O}}(CT^q)$ is the external regret of the original adversarial algorithm and p is the memory of the adversary. This implies that by employing EXP3, we can ensure a regret of $\widetilde{\mathcal{O}}\left((k+1)n^{1/3}T^{2/3}\right)$.

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 & Liu, 2012) bandits, instead, consider an 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 & 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 6.2).

6. Experimental Validation

In this section, we first provide (Section 6.1) a numerical validation of AR-UCB in comparison with other bandit baselines in synthetically-generated and real-world domains. Then, we discuss (Section 6.2) 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 6.3) and k (Section 6.4). Finally, in Section 6.5, we conduct validation in a setting generalized from real-world data. The code to reproduce the experiments can be found at github.com/gianmarcogenalti/autoregressive-bandits.

6.1. 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 (A, B, and C).

Baselines AR-UCB will compete with several bandit baselines. First, it is compared with UCB1 (Lai et al., 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 immediately observe that AR-UCB suffers the smallest cumulative regret in the presented 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 scenarios presented in this experiment. 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).

6.2. 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*, i.e., when $\xi_t = 0$ for every $t \in [\![T]\!]$. Then, we conduct an experiment to highlight the variations in terms of the average reward for the two policies in environments presenting different magnitudes of noise.

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.[4] Such an action a^+ is the one maximizing the *steady-state reward*, namely:

$$a^+ \in \arg\max_{a \in \mathcal{A}} \frac{\gamma_0(a)}{1 - \sum_{i=1}^{k} \gamma_i(a)}. \tag{10}$$

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 of the optimal policy without noise can be found in Appendix B.

Setting To demonstrate the importance of 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 10). The setting selected with $k = 2$ is deliberately challenging and made of just two actions, a_1 and a_2, that are very close in terms of expected steady-state reward:

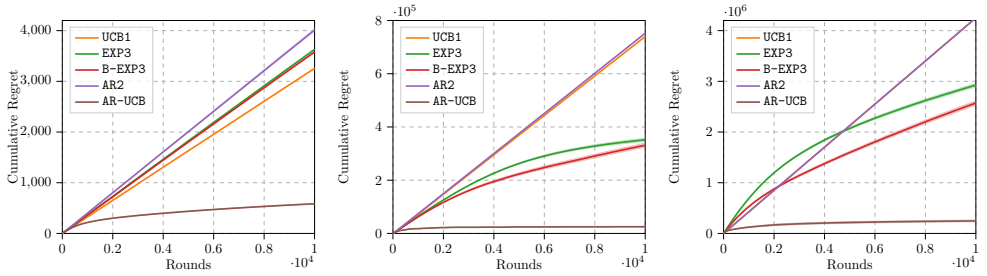
$$\boldsymbol{\gamma}(a_1) = (1,\ \rho,\ 0)^T \qquad \boldsymbol{\gamma}(a_2) = (1,\ 0,\ \rho - \epsilon)^T,$$

where $\rho = 0.5$ and $\epsilon = 0.02$ and several values of zero-mean Gaussian noise are considered with $\sigma \in \{0,\ 0.1,\ 0.5,\ 1.0,\ 2.0\}$. The simulations last for $T = 10000$

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

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 description and cumulative regret of AR-UCB and multiple baselines (100 runs, mean ± std).

in order to make the transient effects neglectable.

Results Table 1 shows the performance of the two policies in terms of average 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 is able to 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 is able to exploit it).

6.3. 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 small cumulative 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 $|\mathcal{A}| = 7$, $k = 4$ 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).

6.4. On the Knowledge of the Autoregressive order k

While the order k of the AR process is assumed to be known to AR-UCB, the algorithm 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, setting them to 0. A particular case is when $\overline{k}$ is close to k but strictly lower(i.e., $\overline{k} = 8$). In this scenario, 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.

6.5. 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 & Leighton,

σ	Stochastic	Deterministic
0	**1.9994 (0)**	**1.9994 (0)**
0.1	**2.0167 (2.03e-5)**	1.9998 (2.04e-5)
0.5	**2.2049 (1.02e-4)**	2.0012 (1.02e-4)
1.0	**2.4504 (2.04e-4)**	2.0030 (2.04e-4)
2.0	**2.9428 (4.09e-4)**	2.0067 (4.08e-4)

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

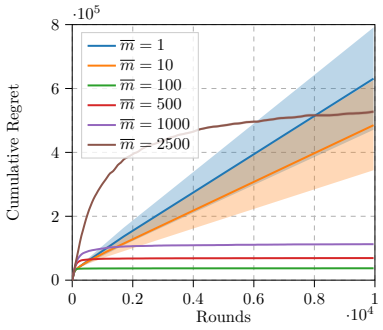


Figure 2: Effect of the choice of parameter $\overline{m}$ on the AR−UCB cumulative regret (100 runs, mean $\pm$ std).

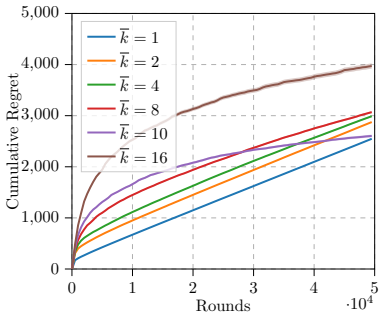
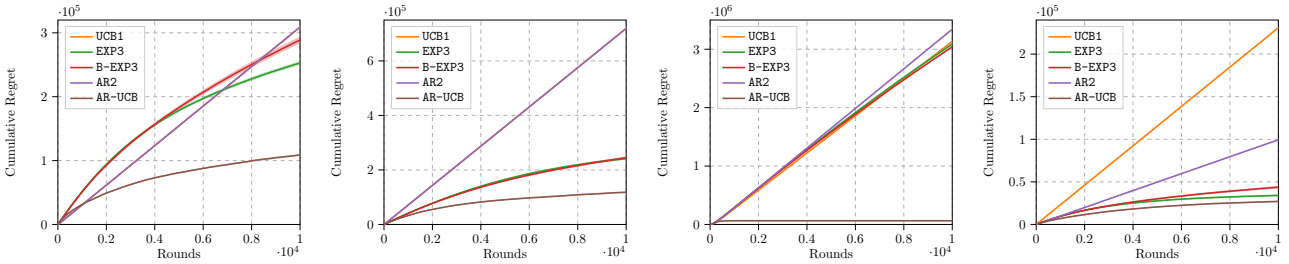


Figure 3: Effect of the choice of parameter $\overline{k}$ on the AR−UCB cumulative regret (100 runs, mean $\pm$ std).



(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).

2003; Mussi et al., 2022a). 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.[5] For each product, we have weekly records of the number of units sold and the related price. We focused on the top 4 best-selling products, estimating the hidden autoregressive parameters governing the sales through standard regression methods. In particular, we discretize the prices into $n = 8$ price bands (i.e., our actions) and we build the simulation environment considering a maximum delay 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. We test AR−UCB and the other bandit baselines presented in Section 6.1.

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.

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

7. 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. Future directions include the study of the complexity of learning in the ARB setting with the goal of deriving regret lower bounds and the extension of the presented setting to continuous actions.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Albalawi, F., Dong, Z., and Angeli, D. Regret-based robust economic model predictive control for nonlinear dissipative systems. In *2021 European Control Conference (ECC)*, pp. 1105–1111. IEEE, 2021.

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

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

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

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

Bowen, J. T. and Chen, S.-L. The relationship between customer loyalty and customer satisfaction. *International journal of contemporary hospitality management*, 2001.

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

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pp. 355–366, 2008.

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

Gur, Y., Zeevi, A., and Besbes, O. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, pp. 199–207, 2014.

Hamilton, J. D. *Time series analysis*. Princeton university press, 2020.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Kleinberg, R. and Leighton, T. 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.*, pp. 594–605. IEEE, 2003.

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

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 757–762. IEEE, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

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

Mussi, M., Genalti, G., Nuara, A., Trovò, F., Restelli, M., and Gatti, N. Dynamic pricing with volume discounts in online settings. *arXiv preprint arXiv:2211.09612*, 2022a.

Mussi, M., Genalti, G., Trovò, F., Nuara, A., Gatti, N., and Restelli, M. 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*, pp. 3623–3633, 2022b.

Ortner, R., Ryabko, D., Auer, P., and Munos, R. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pp. 214–228. Springer, 2012.

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

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

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

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

Umlauft, J. and Hirche, S. Learning stable stochastic nonlinear dynamical systems. In *International Conference on Machine Learning*, pp. 3502–3510. PMLR, 2017.

A. 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 context 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 hold 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 context 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:*

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

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:

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

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

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

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. $1-\delta$:

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

Clearly, this hold for the rounds $t \in \mathbb{N}$ in which the action a is not pulled, since the corresponding estimated 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:*

$$
\begin{aligned}
r_t &= x^*_t - x_t \\
&= \sum_{i=1}^{k} \gamma_i(a^*_t)(x^*_{t-i} - x_{t-i}) + \langle \boldsymbol{\gamma}(a^*_t) - \boldsymbol{\gamma}(a_t), \mathbf{z}_{t-1}\rangle \\
&= \sum_{i=1}^{k} \gamma_i(a^*_t) r_{t-i} + \rho_t, \qquad (8)
\end{aligned}
$$

where $r_t \coloneqq x^*_t - x_t$ is the instantaneous policy regret, $\rho_t \coloneqq \langle \boldsymbol{\gamma}(a^*_t) - \boldsymbol{\gamma}(\widehat{a}_t), \mathbf{z}_{t-1}\rangle$ 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 context 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 context 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:*

$$R(\boldsymbol{\pi}, T) \leq \left(1 + \frac{\Gamma k}{1 - \Gamma}\right) \varrho(\boldsymbol{\pi}, T), \tag{9}$$

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" that allows us to decompose the linear recurrence as follows:

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

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 for 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. □

Lemma 5. *Let $(\mathbf{z}_t)_{t\in[\![T]\!]}$ be the sequence of context 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 hold 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 context 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 & 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 6. *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):*

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

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 $\mathcal{C}_t(a) := \{\boldsymbol{\gamma} \in \mathbb{R}^{k+1} : \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{t-1}(a)\|_{\mathbf{V}_{t-1}(a)} \leq \beta_{t-1}(a)\}$ 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 $\rho_t \leq \|\boldsymbol{z}_{t-1}\|_2(\|\boldsymbol{\gamma}(a_t^*)\|_2 + \|\boldsymbol{\gamma}(a_t)\|_2) \leq 2\sqrt{m^2+1}\|\boldsymbol{z}_{t-1}\|_2$. By Lemma 5 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:

$$\rho_t \leq 2L\sqrt{m^2+1} =: C_1.$$

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:

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

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 nondecreasing, 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:

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

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

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

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

$$\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t \leq \widetilde{O}\left(\frac{(m^2+\sigma)\sqrt{n(k+1)T}}{1-\Gamma}\right),$$

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

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

□

B. Optimal Policy without Noise

In the case of no noise, our system writes:

$$x_t = \gamma_0(a_t) + \sum_{i=1}^{k} \gamma_i(a_t) x_{t-i}. \tag{11}$$

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:

$$a^* \in \arg\max_{a\in\mathcal{A}} \frac{\gamma_0(a_t)}{1-\sum_{i=1}^{k}\gamma_i(a_t)}. \tag{12}$$

This result should not surprise. 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 (11). 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 (12). 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:

$$\begin{aligned}
R_t &\leq \sum_{t=1}^{+\infty} x^* - x_t \\
&\leq \sum_{t=1}^{+\infty} x^* - \gamma_0(a^*)\frac{1-\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma} \\
&= \gamma_0(a^*) \sum_{t=1}^{+\infty} \frac{1}{1-\Gamma} - \frac{1-\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma} \\
&= \gamma_0(a^*) \sum_{t=1}^{+\infty} \frac{\Gamma^{\lfloor t/k \rfloor}}{1-\Gamma} \\
&= \gamma_0(a^*) \frac{k}{(1-\Gamma)^2}.
\end{aligned}$$

□