Autoregressive Bandits in Near-Unstable or Unstable Environment

Uladzimir Charniauski & Yao Zheng*

Department of Statistics, University of Connecticut, Storrs, CT

https://doi.org/10.33697/ajur.2024.116

*Students: uladzimir.charniauski@gmail.com**
Mentor: yao.zheng@uconn.edu

ABSTRACT

AutoRegressive Bandits (ARBs) is a novel model of a sequential decision-making problem as an autoregressive (AR) process. In this online learning setting, the observed reward follows an autoregressive process, whose action parameters are unknown to the agent and create an AR dynamic that depends on actions the agent chooses. This study empirically demonstrates how assigning the extreme values of systemic stability indexes and other reward-governing parameters severely impairs the ARBs learning in the respective environment. We show that this algorithm suffers numerically larger regrets of higher forms under a weakly stable environment and a strictly exponential regret under the unstable environment over the considered optimization horizon. We also test ARBs against other bandit baselines in both weakly stable and unstable systems to investigate the deteriorating effect of dropping systemic stability on their performance and demonstrate the potential advantage of choosing other competing algorithms in case of weakened stability. Finally, we measure the discussed bandit under various assigned values of key input parameters to study how we can possibly improve this algorithm's performance under these extreme environmental conditions.

KEYWORDS

Reinforcement Learning; Machine Learning; Autoregressive Processes; Bandit Algorithms; Non-Stationary Bandits; Online Learning

INTRODUCTION

Multi-Armed Bandits (MABs)[2] is a simplified reinforcement learning problem where an agent must choose among available actions (arms) to maximize the cumulative reward over time. The goal is to develop strategies that minimize regret - the difference between the rewards obtained and those that the agent could achieve by always choosing the optimal arm. Bacchiocchi et al.[1] introduce an AutoRegressive Bandits (ARBs) setting, an extension of traditional MABs and a novel representation of a particular class of continuous reinforcement learning problems, where the reward is determined by the autoregressive (AR) process, whose parameters depend on the actions the agent chooses. The AR process[3] is one of the most widely used class of stochastic processes to model temporal dependencies in real-world phenomena (e.g. stock markets, weather forecasting, etc.).[18, 19] Bacchiocchi et al.[1] employ an optimistic algorithm for online-learning, AutoRegressive Upper Confidence Bounds (AR-UCB), designed to pursue the reward maximizing action sequence, or optimal policy, within the ARBs setting in an online fashion. The AR-UCB can be used to in dynamic e-commerce pricing to model and forecast price changes over time based on historical pricing data, as the AR models assume that the current price is a linear function of past prices and an error term.[22] This algorithm has empirically proven to be advantageous with respect to existing methods in displaying its regret-minimizing abilities when tested under the same conditions(see the *Baselines Comparison* section for details). The authors demonstrated that AR-UCB always suffers the smallest cumulative regret and, unlike its competitors, displays the sublinear behavior, indicating its dominating efficiency in optimizing the sequence of played actions compared to other methods.

Standard bandit algorithms typically assume a stationary environment. However, in many real-world applications, the underlying conditions affecting the reward distribution of the arms may change over time.[20] As Granger et al.[17] note, "Many economic and business time series are non-stationary and, therefore, the type of models that we have studied cannot (directly) be used." This implication spurs the interest in developing bandit algorithms tailored for non-stationary environments.[7, 9] Particularly, in real-world applications of AR processes, unit-root non-stationarity is frequently observed. This type of non-stationarity is characterized by a stochastic trend, where the time series develops indefinitely, without returning to a fixed mean or trend line.[3, 10, 23] Failing to account for unit roots can lead to spurious regressions and invalid statistical inferences, so it is important to address unit-roots before modeling AR processes.[21] However, the important question is how the unit-root non-stationarity influences the quality of learning the optimal policy within the ARB setting and whether the AR-UCB can still remain competitive compared to other bandit algorithms under such conditions.

Existing works mostly focus on addressing non-stationarity in the standard MABs setting, without incorporating autocorrelations in the processes. For example, Besbes et al.[11] propose an algorithm that achieves the optimal regret bound, which adapts to the degree of non-stationarity in the reward process. The non-stationary bandit problem involves scenarios where the environment is dynamic, leading to fluctuations in rewards and potentially altering the optimal strategy over time.[8] Komiyama[12] propose a new bandit algorithm class named the Adaptive Resetting Bandit (ADR-Bandit) that can achieve the optimal performance in stationary and non-stationary environments, accounting for its abrupt or gradual changes, which the author calls "global changes." Furthermore, Liu et al.[13] introduce the Predictive Sampling learning algorithm that can adapt to the degree of non-stationarity in the environment and empirically outperforms Thompson sampling[14] employed in stationary learning. The prevailing consensus highlights the importance of employing specialized methods for bandit learning problems to attain desired adaptive strategies.

Limited studies have proposed more specialized algorithms that can help the ARBs adapt to changes in their learning environments. Trella et al.[9] introduce a new formulation in the context of the non-stationary latent AR bandits,[15, 16] where the reward distributions of the arms follow a latent AR process with a changing according to this AR dynamics state over time. The authors propose an efficient AR OFUL algorithm, a modified version of OFUL algorithm designed for stochastic linear bandits,[5] that is capable of effectively handling changes in non-stationarity. Nonetheless, the algorithmic approach in this paper addresses the unknown nature of the latent process, which fails to encompass the setting addressed by Bacchiocchi et al.,[1] where the information about the AR process is available for the agent. The scarcity of existing literature studying the behavior of AR bandits under extreme environmental conditions has sparked the initiation of the presented study.

The goal of this paper is to investigate the AR-UCB behavior under various degrees of systemic stability, an employed measure[10] for the degree of stationarity of the autoregressive processes, in terms of generated average cumulative regrets to determine the effect of changing these environmental conditions on the algorithm's performance. We analyze the algorithm in three near-unstable, or with an extremely weakened stability that closely replicates the unit-root non-stationarity, and one strictly unstable environment and compare it to results presented in Bacchiocchi et al.[1] containing the original stability indexes. We also test AR-UCB with other baselines from the literature[4–7] in each introduced environment. Importantly, we empirically demonstrate that near-unstable environment drastically worsen and the unstable environment paralyzes the learning process for AR-UCB and other introduced bandits. Lastly, we will empirically evaluate the AR-UCB under selected values of reward-governing parameters, such as Ridge regularization parameter and the boundedness value, whose meaning we discuss along the paper, to illustrate that the algorithm minimizes the generated cumulative regret and improves the learning process for the smallest values of these controlled parameters.

THE AUTOREGRESSIVE BANDITS

Setting

The AutoRegressive Bandits setting[1] considers the sequential interactions between the learner and the environment. It conditions the reward evolution according to the autoregressive (AR) process of the order k (AR(k)).[3] Thus, at every round t, the learner chooses an action $a_t \in \mathcal{A} = [\![n]\!]$ (we define $[\![a, b]\!] = \{a, ..., b\}$ and $[\![b]\!] = \{1, ..., b\}$ for any $a \leq b \in \mathbb{N}$) among the $n \in \mathbb{N}$ available ones and observes the reward x_t of the form described in **Equation 1**.

$$x_t = \gamma_0(a_t) + \Sigma_{i=1}^{k} \gamma_i(a_t) x_{t-i} + \xi_t \qquad \textbf{Equation 1.}$$

We define the reward space as $x_t \in \mathcal{X} \subseteq \mathbb{R}$, the unknown *parameters* depending on an action a as $\gamma_0(a_t) \in \mathbb{R}$ and $(\gamma_i(a_t))_{i \in [\![k]\!]} \in \mathbb{R}^k$, and the zero-mean σ^2-subgaussian random noise conditioned to the past as ξ_t. We can also express the reward evolution in the *inner product* form $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \Sigma_{i=1}^{n} x_i y_i$, where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are any finite real-valued vectors, as shown in **Equation 2**.

$$x_t = \langle \gamma(a_t), \boldsymbol{z}_{t-1} \rangle + \xi_t \qquad \textbf{Equation 2.}$$

In this equation, we have $\boldsymbol{z}_{t-1} = (1, x_{t-1}, ..., x_{t-k})^T \in \mathcal{Z} = \{1\} \times \mathcal{X}^k$ - the *context vector* expressing the past history of reward observations, and $\gamma(a) = (\gamma_0(a), ..., \gamma_k(a))^T \in \mathbb{R}^{k+1}$ - the *parameter vector* for every $a \in \mathcal{A}$.

Every $\gamma_i(a_t)$ parameter fulfills three following assumptions[1] labeled in order: **Assumption 1** (Non-negativity) requires that the coefficients of are non-negative (i.e. $\gamma_i(a) \geq 0$ for every $i \in [\![0, k]\!]$) for producing realistic results in observing the real-world phenomena. **Assumption 2** (Stability) establishes that the sum of action parameters $\Sigma_{i=1}^{k} \gamma_i(a)$ is limited to a stability coefficient value $\Gamma \in [0, 1)$. **Assumption 3** (Boundedness) enforces the boundedness to $\gamma_0(a)$ with a finite value $m = \max_{a \in \mathcal{A}} \gamma_0(a)$.

Assumption 2 and **3** guarantee the inability of the underlying autoregressive processes to diverge in expectation for any action sequence played by the agent.[1] On the other hand, the near-unstable environments with $\Gamma \approxeq 1$ and/or large values of m aggravate the learning process within the ARB setting, requiring more time for the agent to develop the optimal policy. Finally, the unstable environment $\Gamma = 1$ completely eliminates the systemic stability, in which case the learner is unable to choose the most optimal action sequence to maximize its results regardless of values of other presented parameters. In this work, we show the importance of employing these two assumptions in the learning process by empirically demonstrating that the AR-UCB suffers numerically larger regrets due to losses in learning abilities occurring in the environments, where either of these two assumptions (or both) are relaxed.

Policy and Performance

Since our studies concern the empirical analysis of the achieved regret, we provide the formal regret definition. The policy π models the learner's behavior and the regret R imposes the loss of not choosing an optimal action on a learner. The deterministic learner's policy $\pi = (\pi_t)_{t \in \mathbb{N}}$, defined for each round $t \in [\![T]\!]$ as the mapping function from the history of observations $H_{t-1} = (x_0, a_1, x_1, ..., a_{t-1}, x_{t-1})$ to the action space $\mathcal{A}$, demonstrating that $a_t = \pi_t(H_{t-1})$.

The policy performance is evaluated through the expected cumulative reward $J_T(\pi) = \mathbb{E}[\Sigma_{t-1}^{T} x_t]$ over the horizon $T \in \mathbb{N}$ with respect to the random reward noise ξ_t. The learner objective is to minimize the expected cumulative regret $R(\pi, T)$ by playing a policy π against the *optimal policy* π^* satisfying $\pi^*(H_{t-1}) = \arg\max_{a \in \mathcal{A}} \langle \gamma(a), \boldsymbol{z}_{t-1} \rangle$ [1] (from **Assumption 1**) over a learning horizon $T \in \mathbb{N}$, where where $r_t = x_t^* - x_t$ is the *instantaneous policy regret* and $(x_t^*)_{t \in [\![t \in \mathbb{N}]\!]}$ is the sequence of rewards from playing π^* from **Equation 3**.

$$R(\pi, T) = J_T^* - J_T(\pi) = \mathbb{E}[\Sigma_{t=1}^{T} r_t] \quad \textbf{Equation 3.}$$

We find that the presence of weak systemic stability impedes the agent in completing the regret minimization objective. Specifically, we demonstrate the positive trend between dropping the robustness of systemic stability and the number of rounds required to develop the optimal policy under these aggravating conditions.

The AR-UCB

For the ARB setting, we devise AR-UCB that suffers sublinear regret[1] of order $\mathcal{O}\left(\frac{(m+\sigma)(k+1)^{3/2}\sqrt{nT}}{(1-\Gamma)^2}\right)$, where T is the exploration horizon, n is number of actions, m and σ are the $\max \gamma_0(a)$ and noise values, respectively, and Γ is the index of systemic stability. This formula suggests that higher autoregressive orders k, larger values m and stability indexes Γ increase the complexity of learning for this algorithm. Thus relaxing **Assumption 2** and **Assumption 3** exasperate the learning process by allowing higher values for m and Γ parameters.

At each round $t \in [\![T]\!]$, the AR-UCB algorithm computes the *Upper Confidence Bound* for every $a \in \mathcal{A}$ to play this action and observe the reward x_t as in **Equation 2**. The formal computation is defined as follows in the **Equation 4**.

$$a_t \in \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_t(a) := \langle \hat{\gamma}_{t-1}(a), \boldsymbol{z}_{t-1} \rangle + \beta_{t-1}(a) ||\boldsymbol{z}_{t-1}||_{\boldsymbol{V}_t(a)} \quad \textbf{Equation 4.}$$

The AR-UCB computes $\hat{\gamma}_t(a) = \boldsymbol{V}_t(a)^{-1}\boldsymbol{b}_t(a)$, where it consistently updates the required Gram matrix $\boldsymbol{V}_t(a)$ and the vector $\boldsymbol{b}_t(a)$ ($\boldsymbol{V}_0(a) = \lambda \boldsymbol{I}_{k+1}$ and $\boldsymbol{b}_0(a) = \boldsymbol{0}_{k+1}$ at $t = 0$), as $\boldsymbol{V}_t(a) = \boldsymbol{V}_{t-1}(a) + \boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}^T \mathbb{1}_{\{a=a_t\}}$ and $\boldsymbol{b}_t(a) = \boldsymbol{b}_{t-1}(a) + \boldsymbol{z}_{t-1}x_t \mathbb{1}_{\{a=a_t\}}$ for every $a \in \mathcal{A}$ after observing the reward x_t. In this equation, we also define the exploration coefficient $\beta_{t-1}(a) \geq 0$ for every action $a \in \mathcal{A}$ and round $t \in [\![0, T-1]\!]$ as it follows in **Equation 5**.

$$\beta_t(a) = \sqrt{\lambda(m^2+1)} + \sigma\sqrt{2\log(\frac{n}{\delta}) + \log(\frac{\det \boldsymbol{V}_t(a)}{\lambda^{k+1}})} \quad \textbf{Equation 5.}$$

The first term in this equation is the *bias* term and the second one is the *concentration* term.[1] From this form of $\beta_t(a)$, it follows that smaller input values of λ and m reduce the *bias* term in computing $\beta_t(a)$. This way the algorithm more precisely estimates the action played, which helps achieve lower instantaneous regret r_t from playing the computed action a_t, while larger values of these parameters make the AR-UCB produce more biased calculations of a_t. However, because there is a trade-off between two terms of this equation with respect to λ, the value of $\beta_t(a)$ may evolve differently for smaller m under the same λ.

In practice, the actual value of m is unknown, so we may replace this value with a user-specified upper bound $\bar{m}$ to compute $\beta_t(a)$ in **Equation 5**. For example, in the simulation on the AR-UCB performance under different parameters λ, we allow $\bar{m}$ to differ from the actual value m. We will demonstrate that the AR-UCB suffers smaller regrets with $m = 20$ and $\bar{m} = 100$ as values of λ decrease to 0. However, the relationship between the regret and λ behaves quite differently under $m = \bar{m} = 1$. Moreover, we conduct another experiment to investigate the effect of m on the regret evolution when λ is fixed, and to avoid the ambiguity due to possible misspecification, we simply set $\bar{m} = m$.

SIMULATION DESIGN

In our experiments, we measure the AR-UCB and other baselines performance in terms of average cumulative regret over the optimization horizon $T = 10000$ rounds in a range of provided settings distinguished by the assigned systemic stability or the algorithm's key parameters. We evaluate all the experiments within three specific near-unstable environments, each carrying $\Gamma \in \{0.95, 0.98, 0.999\}$ stability indexes, and the unstable environment with a stability index

$\Gamma = 1$. We run a range of Python simulations, where each complies to the values in every respective setting, and provide the graphs of cumulative regrets achieved by the AR-UCB on alone or other introduced bandit baselines in each of discussed settings. All the algorithms are implemented in Python 3.12, and run over an Apple M1 with 8 GB RAM.

AR-UCB Performance

We first observe and analyze the AR-UCB cumulative regret on alone under systemic near-instability and instability. We consider three experimental settings containing autoregressive orders $k \in \{2, 4\}$, number of actions $n \in \{2, 7\}$, actual values $m \in \{1, 20, 920\}$, and noise parameters $\sigma = \{0.75, 1.5, 10\}$, respectively. We also compare our results to cumulative regret plots for every bandit generated under the original systemic stability indexes $\Gamma_{orig} = \{0.5, 0.7, 0.82\}$ using the same parameters to visually demonstrate the impact of weak stability on the AR-UCB performance. To specify the details of the learning process, we select hyper-parameters $\lambda = 1$, a Ridge regularization parameter value, and boundaries $\bar{m} = \{10, 100, 1000\}$ for the equivalent magnitudes of corresponding values of m (i.e. every $\bar{m}$ is sequentially selected for each m). **Table 1** summarizes the settings details.

	Parameters				
Setting	Γ_{orig}	k	n	m	σ
A	0.5	2	2	1	0.75
B	0.82	4	7	20	1.5
C	0.7	4	7	920	10

Table 1. Settings description.

Baselines Comparison

We test and compare AR-UCB performance in near-unstable and unstable environments against the following selected baselines: UCB1, EXP3, B-EXP3, and AR2. UCB1[4] is a widely-adapted solution for stochastic MABs. EXP3[6] is an algorithm designed for adversarial MABs and its extension to finite-memory adaptive adversaries B-EXP3.[5] AR2[7] is an algorithm that operates within a non-stationary MAB framework with an autoregressive reward structure of the first order (AR(1)). For this experiment, we utilize the same parameters as in the study on *AR-UCB Performance*. We also compare our results to cumulative regret plots for every bandit generated by Bacchiocchi et al.[1] under the original systemic stability indexes $\Gamma_{orig} = \{0.5, 0.7, 0.82\}$ using the same parameters to visually demonstrate the gravity of the impact of weak stability on the baselines performance. However, since we experiment on several baselines altogether, we display our results in three settings consisting of five graphs, where each corresponds to a stability index from Γ_{orig} to $\Gamma = 1$. **Table 1** precisely summarizes the parameters utilized in each setting.

On the AR-UCB Performance Under Different Parameters λ

We experimentally measure the AR-UCB performance under different regularization parameters λ. For our experiments on λ, we test this algorithm for selected $\lambda \in \{0.001, 0.01, 0.05, 0.2, 0.6, 1.2, 1.6, 3, 5\}$ against its performance for the originally utilized by Bacchiocchi et al.[1] choice $\lambda = 1$. We first employ $n = 7$, $k = 4$, $m = 20$, and $\sigma = 1.5$ in every near-unstable and unstable setting (**Figure 3**). Then we repeat this experiment with $m = 1$ (**Figure 4**) to analyze trade-off between the *bias* and *concentration* terms from **Equation 5** with respect to λ. We also select $\bar{m} = 1$ and $\bar{m} = 100$ as our hyper-parameter for $m = 1$ and $m = 20$, respectively. The goal of this experiment is to analyze how each assigned λ impacts the AR-UCB regret evolution with respect to its original value, and what regularization value helps the algorithm minimize it.

On the AR-UCB Performance Under Different Parameters m

We test the AR-UCB under several values of m. For our experiments on m, we consider an array of values $m \in \{0, 0.25, 0.5, 1, 10, 100, 500, 1000\}$. We utilize $n = 7$, $k = 4$, $\sigma = 1.5$, and $\lambda = 1$ in every near-unstable and unstable setting. To avoid the misspecification[1] of presented boundary parameters, each $\bar{m} = m$ for every action $a \in \mathcal{A}$ for its respective scenario (Ex. $\bar{m} = 10$ if $m = 10$), since we do not investigate in this study how the erroneous

estimation of the boundary influences the algorithm's performance. We only aim to establish the general relationship between the m value and the achieved AR-UCB cumulative regret in the learning process.

RESULTS

AR-UCB Performance

Figures 1 shows the average cumulative regrets for AR-UCB under different stability indexes. We may observe that the AR-UCB regret rapidly degenerates from the sublinear to higher forms under $\Gamma \in \{0.95, 0.98, 0.999\}$ in every respective scenario. Thus, under falling stability, the AR-UCB requires drastically more time to learn the optimal action sequence for adapting to its environment. This notion is especially highlighted under $\Gamma = 0.999$ in every experiment, where the AR-UCB achieves the exponential regret in the first stages of the simulations under the $\Gamma = 0.999$ stability index. This way we observe that the extremely low systemic stability makes the algorithm temporarily lose the learning ability within the limited exploration.

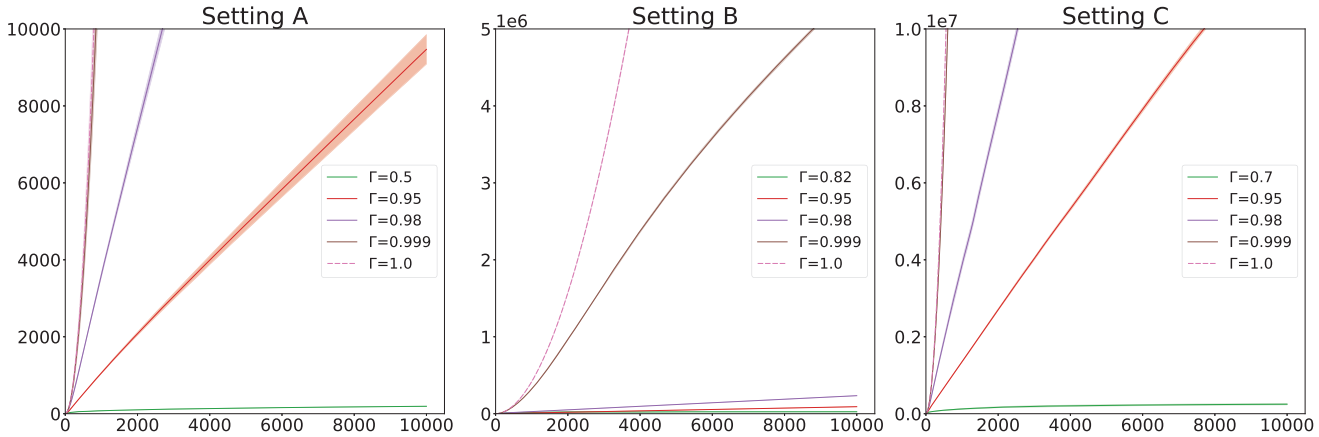

Figure 1. Cumulative regret of AR-UCB across stable, near-unstable, and unstable environments.

The AR-UCB performance under $\Gamma = 1$ is fundamentally different. First, we immediately observe the cumulative regret to be exponentially increasing in every introduced experiment. This indicates that AR-UCB completely loses its ability to acquire information in the unstable environment, making less optimal choices at every round over the learning horizon T. We may similarly observe precisely the same AR-UCB behavior under $\Gamma = 0.999$ at earlier stimulation stages, since this index value closely recreates this effect of the systemic instability on learning. However, it is still possible for the algorithm to achieve the optimal policy even under the infinitely close to 1 value of Γ over a large number of rounds, which is especially seen in a Setting **B**, whereas the unstable environment completely annihilates this possibility by paralyzing the learning process.

The above experiments highlight the importance of employing **Assumption 2** in the learning process. This assumption ensures the efficiency of learning for the algorithm by creating a stable environment s.t. Γ is far within the radius of 1. We observe that weakening this assumption with any $\Gamma \approxeq 1$ consequently weakens the algorithm's ability to process the information from its setting and develop the optimal action sequence that reduces regrets to a sublinear form. This way the more Γ approaches 1, the more this index complicates the ongoing learning for the algorithm. Finally, the unstable $\Gamma = 1$ completely abolishes the learning, so that the algorithm plays sub-optimal actions at all rounds.

Baselines Comparison

Figures 2 illustrates the average cumulative regret of all tested baselines for $\Gamma \in \{0.95, 0.98, 0.999\}$ in three settings. The cumulative regret achieved by many tested bandit baselines progressively degenerates to higher forms with increasing values of Γ. Every bandit suffers precisely the same numerical regret, except for a particular case depicted in the Setting **B** in scenario with $\Gamma = 0.999$, where UCB1 significantly outperforms every other baseline under this stability index. We also see that every baseline achieves the exponential regret under $\Gamma = 0.999$ in every experiment during the

first stages of simulations, which occurs due to the limited exploration. Amongst all the baselines, AR-UCB demonstrates near-identical performance compared to UCB1 in every experiment in all near-unstable scenarios, except for the Setting **B** under $\Gamma = 0.999$. On the other hand, AR2 is able to significantly outperform AR-UCB and other bandits in the experiment in the Setting **C**, although their achieved regrets converge in value across every scenario as the stability weakens. Both EXP3 and B-EXP3 suffer near-identical regret in every presented experiment with near-unstable indexes, achieving the largest cumulative regrets with respect to other baselines.

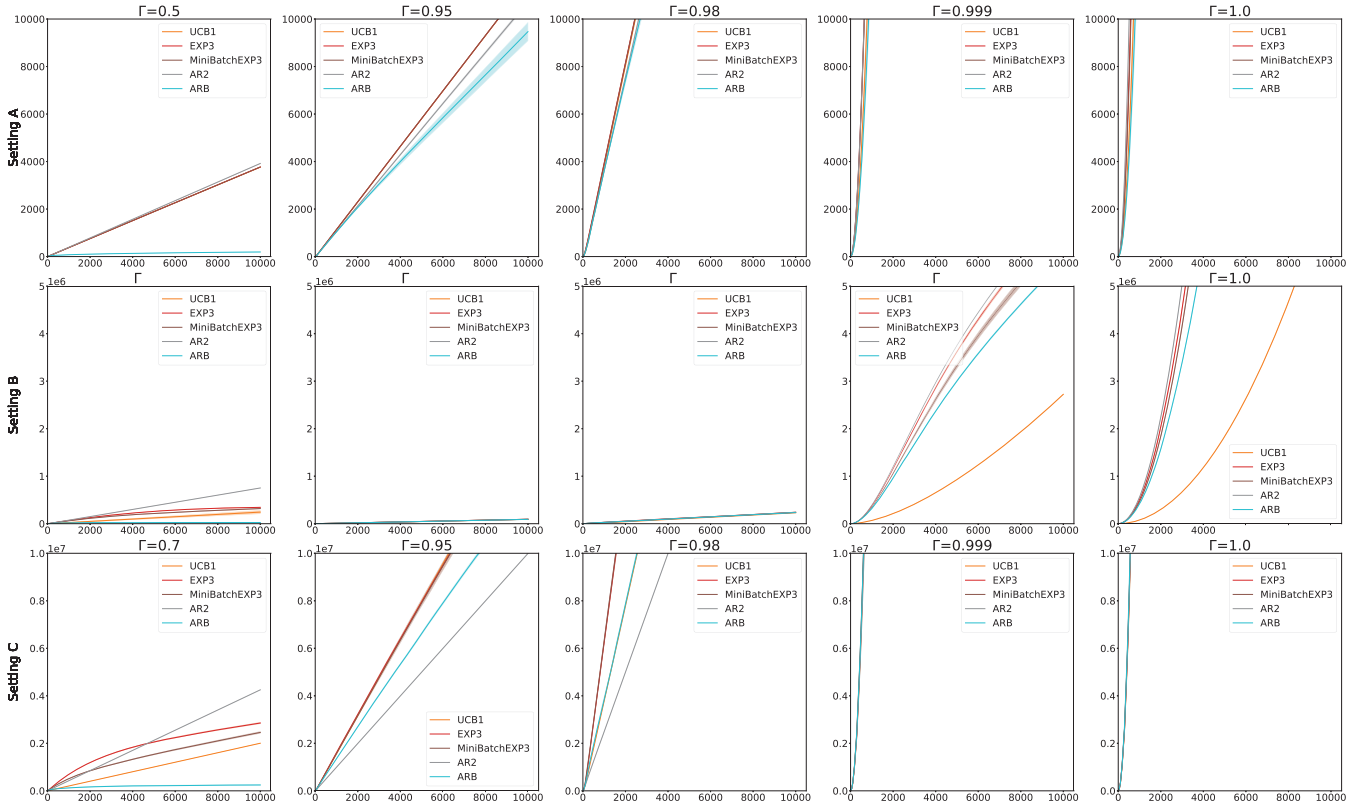

Figure 2. Comparison of cumulative regret for different baselines across stable, near-unstable, and unstable environments.

Unlike in case of near-instability, every experimented baseline displays strictly exponential behavior within the unstable environments in every presented scenario. We may observe that baselines suffer close-in-value exponential cumulative regret in every experiment, except for the one in the Setting **B**, in which UCB1 achieves the lowest regret out of all the bandits, showcasing its least sensibility to the systemic stability. Nonetheless, the AR-UCB again replicates UCB1 in terms of their regret in unstable scenarios in other two experiments (**A** and **C**). Meanwhile, the unstable environment significantly suppressed the AR2 performance, making this baseline to suffer the largest regret in every experiment. Both EXP3 and B-EXP3 again achieve a closely similar exponentially growing cumulative regret under the instability in all presented experiments.

These experiments illustrate the dependence of introduced bandits on the robustness of the systemic stability conditioned by **Assumption 2**. We demonstrated that every baseline develops the same regret behavior depending on the introduced value of a stability index Γ in the respective environment. Thus we showed that all the baselines equivalently lose their learning ability and require more additional time to optimize their action policy under near-unstable $\Gamma \approxeq 1$ values, where . Finally, the unstable index $\Gamma = 1$ disables learning processes for every baseline, making them inefficiently operate within such an environment regardless of other selected parameters.

On the AR-UCB Performance Under Different Parameters λ

Figures 3 and 4 show the average cumulative regret in the near-unstable and unstable environments for each considered λ under $m = 20$ and $m = 1$. We immediately observe that the AR-UCB achieves the best performance under the smallest $\lambda = 0.0001$ for all settings with $m = 20$, whereas the optimal choice of λ in experiments with $m = 1$ varies depending on the stability coefficient: $\lambda = 0.2$ when $\Gamma = \{0.95, 0.98\}$, $\lambda = 0.01$ when $\Gamma = 0.999$, and $\lambda = 1$ when $\Gamma = 1$. In **Figure 3**, we see that the algorithm achieves significantly smaller regrets with choices $\lambda < 1$ with respect to other selections. However, the regrets in **Figure 4** are more clustered, indicating their lesser sensitivity to the choice of λ when m is smaller. We also observe that the AR-UCB enjoys sublinear regret in settings with $\Gamma \leq 0.98$ across all choices of λ. Meanwhile, the severely weakened stability under $\Gamma = 0.999$ conditions the exponential regret evolution at some initial rounds under every λ. Still the algorithm is able to quickly reduce the regret to lower forms over more rounds, especially for optimal choices of λ. It's also worth noting that the AR-UCB can achieve smaller regret in the same fashion across selected λ under $\Gamma = 1$. Nonetheless, because the systemic instability completely disables learning processes, the algorithm does not achieve the sublinear regret in this setting regardless of other parameters.

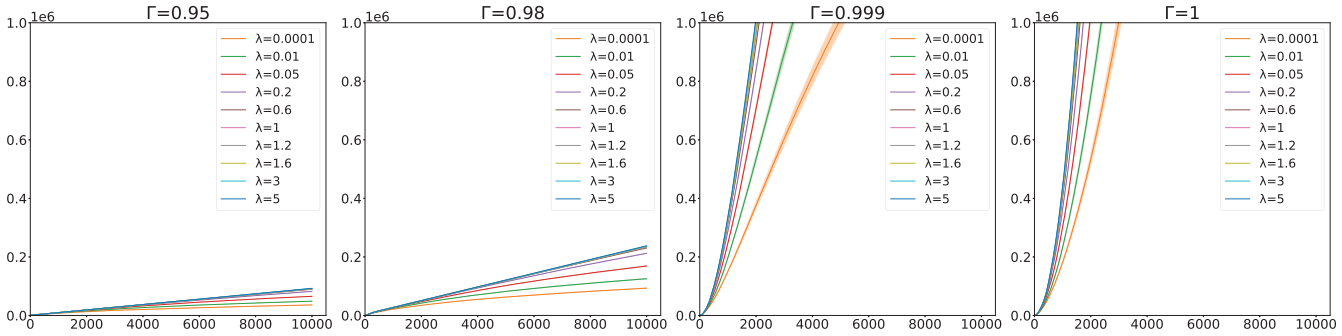

Figure 3. Effect of the choice of λ on the AR-UCB cumulative regret in near-unstable and the unstable environments ($m = 20$).

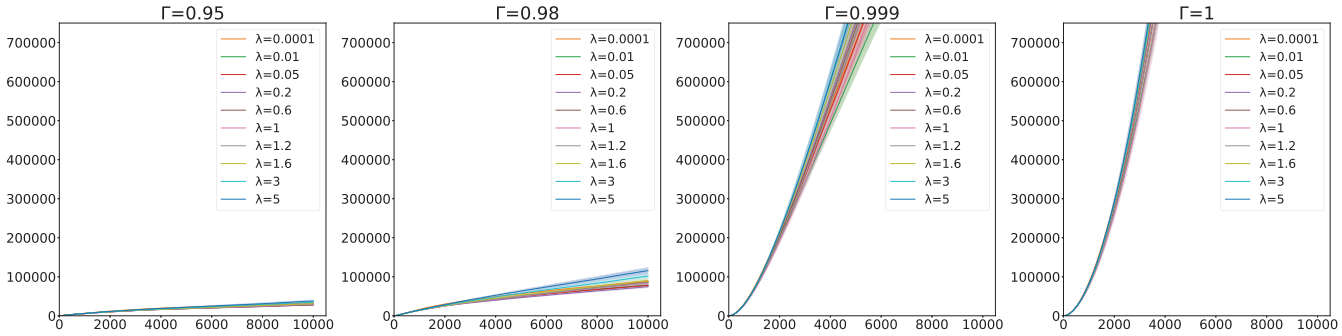

Figure 4. Effect of the choice of λ on the AR-UCB cumulative regret in near-unstable and the unstable environments ($m = 1$).

The experiments on the manipulation of the Ridge regularization parameter λ empirically illustrates the role of $\beta_t(a)$ components from **Equation 5** in the AR-UCB learning process. We can explain the different findings from **Figures 3 and 4** can be explained via the formula for this coefficient. According to this equation, the *bias* term increases with λ, while the concentration term decreases with this parameter. When m is large (ex. $m = 20$), the *bias* will dominate the *concentrtion* term even for extremely small λ choices. Consequently, in **Figure 3**, the smallest $\lambda = 0.0001$ is the most optimal across all the experiment. On the other hand, when m is relatively small, the *bias* term does not necessarily dominate the *concentration* term, and then the optimal choice of λ will depend on the trade-off between the two terms. As a result, in **Figure 4**, the cumulative regret is not monotonic in λ, and the specific optimal choice of lambda depends on the stability parameter.

On the AR-UCB Performance Under Different Parameters m

Figures 5 shows the average cumulative regrets of AR-UCB under different values m. The AR-UCB achieves the minimal regret under every $m = 0$, with a progressive increase as values m get larger across all four experiments. Across all

the experiments, the regrets under $m \in [0, 1]$ are more clustered, while the regret values under $m > 1$ are drastically different. Also, in scenarios with $\Gamma \in \{0.95, 0.98\}$, $m \in [0, 1]$ condition sublinear regret evolution over the entire learning interval, while any $m > 1$ reduces regrets to linear, allowing the sublinear behavior only during the initial stages of simulations. In the scenario with $\Gamma = 0.999$, the AR-UCB seems to exhibit the exponential regret due to severely weakened stability, especially when m is large. Meanwhile, the AR-UCB displays strictly exponential average cumulative regrets for every m within the $\Gamma = 1$ environment. Due to systemic instability, the algorithm will never achieve the optimal policy and continue to suffer this regret regardless of the provided values of m. Thus AR-UCB is only able to minimize the achieved regret with the lowest value of m, as any value $m > 0$ only leads to larger regrets.

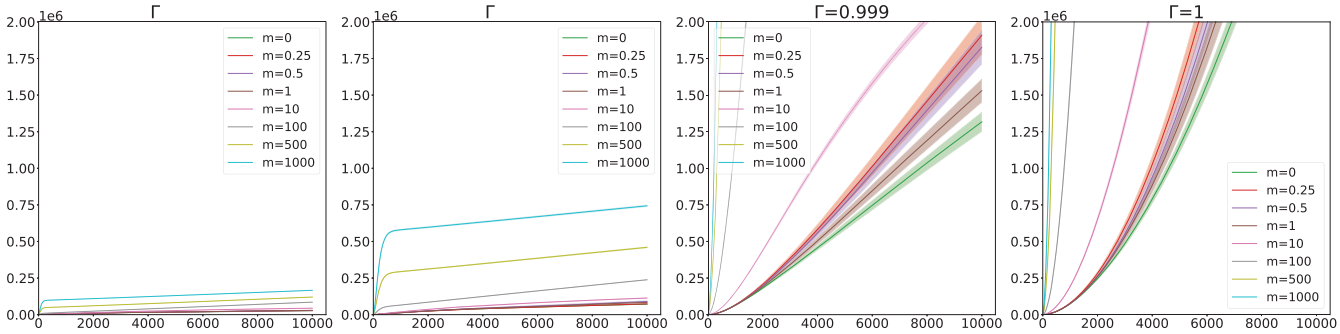

Figure 5. Cumulative regret for AR-UCB in the stable vs. near-unstable and the unstable environments.

These experiments illustrate that larger values of m substantially increase the AR-UCB regrets, while the value $m = 0$ helps produce the smallest regret across our experimental scenarios. However, in the presence of severely weakened stability (i.e $\Gamma = 0.999$), the algorithm under higher values of m can sooner optimize its performance in a trade-off with the size of the regret. Thus we empirically demonstrate that relaxing **Assumption 3** with introducing very high values of m within the weak (or eliminated) systemic stability further reduces the algorithm's performance by enlarging the *bias* in calculating $\beta_t(a)$ for each action, as shown in **Equation 5**.

CONCLUSION AND DISCUSSION

In this research, we experimentally tested the AR-UCB and other bandit baselines' behavior under varying degrees of systemic near-instability $\Gamma \in \{0.95, 0.98, 0.999\}$ and under a definite degree of systemic instability $\Gamma = 1$. First, we measured the AR-UCB performance under the near-instability and instability on alone against its stable performance, demonstrating the differences in the regret evolution within the presented settings. We then repeated our measures under the same selected near-unstable and unstable settings on several other baselines introduced in the original paper. We observed that the presented algorithm achieve precisely the same regret evolution as the AR-UCB and highlighted the performative advantages of some baselines relative to the original one under specific values of systemic near-instability and instability. Finally, we provided a series of experimental measures of AR-UCB under different Ridge regularization parameters λ and values of m to empirically analyze the improvements in the AR-UCB performance by finding the optimal values for these parameters.

Our work highlights the need for a more robust alternative to the AR-UCB algorithm capable of adapting to potentially unstable environments, such as those characterized by (near) unit-root autoregressive processes. Additionally, while our research primarily focuses on numerical investigation, it would be valuable to theoretically analyze the phase transition behavior of the AR-UCB algorithm across stable and unstable regimes. This analysis would provide insights into its capabilities and limitations. Furthermore, it is interesting to extend our study to dynamic linear bandits[24], which assumes an environment evolving according to a stable linear dynamic system. Investigating how the possible violation of stability, characterized by the unit-rootness of the transition matrix, impacts the algorithm's performance remains an open question that warrants further exploration.

REFERENCES

1. Bacchiocchi, F., Genalti, G., Maran, D., Mussi, M., Restelli, M., Gatti, N., & Metelli, A. M. (2024) Autoregressive bandits. In *International Conference on Artificial Intelligence and Statistics*, 937–945. *https://doi.org/10.48550/arXiv.2212.06251*
2. Sutton, R. S., & Barto, A. G. (2005) Reinforcement Learning: An Introduction. 2nd ed., MIT Press, Cambridge, MA.
3. Hamilton, J. D. (1994) *Time Series Analysis*, Princeton university press.
4. Auer, P., Cesa-Bianchi, N., Freund, Y., & R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Processings of IEEE 36th Annual Foundations of Computer Science*, 322–331. *https://doi.org/10.1109/SFCS.1995.492488*
5. Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011) Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems* 24, 2312–2320. *https://doi.org/10.48550/arXiv.2309.14298*
6. Arora, R., Dekel, O., & Tewari, A. (2012) Online Bandit Learning against an Adaptive Adversary: from Regret to Policy Regret. *ICML'12: Proceedings of the 29th International Conference on Machine Learning* 29, 1747–1754.
7. Chen, Q., Golrezaei, N., & Bouneffouf, D. (2023) Non-Stationary Bandits with Auto-Regressive Temporal Dependency. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.
8. Garivier, A., & Moulines, E. (2011) On Upper-Confidence Bound Policies for Switching Bandit Problems. In: Kivinen, J., Szepesvári, C., Ukkonen, E., Zeugmann, T. (eds) *Algorithmic Learning Theory. ALT 2011. Lecture Notes in Computer Science* (vol 6925), 174–183. Springer, Berlin, Heidelberg. *https://doi.org/10.1007/978-3-642-24412-4-16*
9. Trella, A. L., Dempsey, W., Doshi-Velez, F., & Murphy, S. A. (2024) Non-Stationary Latent Auto-Regressive Bandits. *https://doi.org/10.48550/arXiv.2402.03110*
10. Cryer, J. D., & Chan, K.-S. (2011) *Time Series Analysis With Applications In R*. Springer, New York.
11. Besbes, O., Gur, Y., & Zeevi, A. (2019) Optimal exploration-exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4), 319–337. *https://doi.org/10.1287/stsy.2019.0033*
12. Komiyama, J., Fouché, E., & Honda, J. (2021) Finite-time Analysis of Globally Nonstationary Multi-Armed Bandits. *ArXiv, abs/2107.11419*.
13. Liu, Y., Van Roy, B., & Xu, K. (2022) Nonstationary Bandit Learning via Predictive Sampling. *International Conference on Artificial Intelligence and Statistics*, 6215–6244. *https://doi.org/10.48550/arxiv.2205.0197*
14. Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018) A Tutorial on Thompson Sampling. *Foundations and Trends in Machine Learning*, 11(1), 1–96. *https://doi.org/10.1561/2200000070*
15. Zhou, L., & Brunskill, E. (2016) Latent contextual bandits and their application to personalized recommendations for new users. Proceedings of the *Twenty-Fifth International Joint Conference on Artificial Intelligence*, 25, 3646–3653. *https://doi.org/10.48550/arXiv.1604.06743*
16. Hong, J., Kveton, B., Zaheer, M., Chow, Y., Ahmed, A., Ghavamzadeh, M., & Boutilier, C. (2020) Non-Stationary Latent Bandits. *ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2012.00386*
17. Granger, C. W. J., & Swanson, N. R. (1997) An introduction to stochastic unit-root processes. *Journal of Econometrics*, 80(1), 35–62. *https://doi.org/10.1016/S0304-4076(96)00016-4*
18. Z. An, & Z. Feng. (2021) A Stock Price Forecasting Method Using Autoregressive Integrated Moving Average model and Gated Recurrent Unit Network. *2021 International Conference on Big Data Analysis and Computer Science (BDACS)*, 31-34. *https://doi.org/10.1109/bdacs53596.2021.00015*
19. Lai, Y., & Dzombak, D. A. (2020) Use of the Autoregressive Integrated Moving Average (ARIMA) model to forecast Near-Term regional temperature and precipitation. *Weather and Forecasting*, 35(3), 959–976. *https://doi.org/10.1175/waf-d-19-0158.1*
20. Hong, J., Kveton, B., Zaheer, M., Chow, Y., & Ahmed, A. (2021) Non-Stationary Off-Policy Optimization. *International Conference on Artificial Intelligence and Statistics. https://doi.org/10.48550/arXiv.2006.08236*
21. Das, S., & Nason, G. P. (2016) Measuring the degree of non-stationarity of a time series. *Stat*, 5(1), 295–305. *https://doi.org/10.1002/sta4.125*

22. Faehnle, A., & Guidolin, M. (2021) Dynamic Pricing Recognition on E-Commerce Platforms with VAR Processes. *Forecasting*, 3(1), 166–180. *https://doi.org/10.3390/forecast3010011*

23. Zhao, P., Zhang, L., Jiang, Y., & Zhou, Z. (2021) A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 746–755. *https://doi.org/10.48550/arxiv.2103.05324*

24. Mussi, M., Alberto Maria Metelli, & Marcello Restelli. (2023) Dynamical linear bandits. In Proceedings of the *40th International Conference on Machine Learning*, 25563–25587. *https://doi.org/10.48550/arXiv.2211.08997*

ACKNOWLEDGMENT

The authors express gratitude to the editor and reviewers for their insightful feedback, which enabled the authors to refine the manuscript to its current form.

ABOUT THE STUDENT AUTHOR

Uladzimir Charniauski is a second-year student at the University of Connecticut majoring in Applied Mathematics and Statistics at the University of Connecticut. After graduation, he hopes to gain several years of the industry experience before pursuing a Ph.D. in Statistics, focusing on the interdisciplinary studies of Reinforcement Learning and Time Series Analysis.

PRESS SUMMARY

AutoRegressive Upper Confidence Bound (AR-UCB) is a recently proposed online learning algorithm that has been shown to outperform other bandit algorithms in stable autoregressive environments. However, when the environment is weakly stable or unstable, our work reveals that AR-UCB loses its learning ability and may perform worse than other benchmark algorithms. This study highlights the critical dependence of AR-UCB on environmental stability, with important implications for its proper implementation in various real-world domains.