TIGHT REGRET BOUNDS FOR MEAN-REVERTING LINEAR BANDITS VIA RECURSIVE STATE ESTIMATION

Linghang Meng[⋆] *and Jie Wu*[⋆†]

[⋆]Cloud Computing Research Institute, China Telecom
[†]Department of Computer and Information Sciences, Temple University

ABSTRACT

This paper investigates the linear bandit problem in nonstationary environments where the unknown environment parameter θ_t evolves according to a mean-reverting process. Unlike existing works that only assume a bounded amount of parameter variation without exploiting its temporal structure, our formulation leverages the mean-reverting dynamics to enable more precise estimation and efficient learning. We establish regret lower bounds for two scenarios: (i) a linear mean-reverting process, and (ii) a more practical case where θ_t is constrained within a bounded domain around its long-term mean. For both settings, we propose filtering-based state estimation algorithms and show that they achieve regret upper bounds that nearly match our lower bounds. Our theoretical analysis demonstrates that the steady-state regret of the proposed algorithm for the linear case is optimal with respect to the environment evolution noise, yielding tight bounds. Simulation results further validate the theoretical findings and highlight the effectiveness of the proposed methods.

Index Terms— Linear Bandits, Mean Reverting Process, Regret Bounds

1. INTRODUCTION

Sequential decision making under uncertainty is a fundamental problem in machine learning, statistics, and control theory. In many real-world scenarios, a decision maker must repeatedly take actions and receive feedback, aiming to maximize cumulative rewards over time. One of the most prominent models in this area is the multi-armed bandit (MAB) problem, which captures the core trade-off between exploration, gathering information about the environment, and exploitation, leveraging existing knowledge to make high-reward decisions. The bandit formulation has found wide applicability in areas such as recommendation systems, online advertising, adaptive clinical trials, and dynamic pricing [1, 2, 3, 4, 5, 6, 7].

The classical bandit literature predominantly focuses on stationary environments. However, many practical applications involve environments that evolve dynamically. In such non-stationary settings, the optimal decision may shift over time, and algorithms must adapt accordingly [8, 9, 10, 11, 12]. Despite the prevalence of non-stationarity in practical problems, most existing works on non-stationary bandits adopt time-structure-agnostic assumptions. These approaches typically impose constraints on the total amount of change (*e.g.*, bounded total variation or limited number of shifts) and design algorithms that either discount old observations or detect and adapt to abrupt changes [13, 14, 15, 16, 17, 18]. While effective in certain settings, such methods overlook the fact that in many applications, the temporal evolution of the environment is stochastic yet not entirely arbitrary. The evolution often follows underlying dynamics that can be exploited. A particularly relevant and practically motivated dynamic is the mean-reverting process, in which the environment parameter evolves stochastically but tends to revert towards a long-term equilibrium. Intuitively, this process captures systems where short-term fluctuations occur around a stable equilibrium. For example, in wireless communication or control systems, channel conditions or system states may deviate due to noise or disturbances but are stabilized by physical constraints, pulling them back towards nominal operating points [19]. In financial applications, asset returns often exhibit temporary deviations driven by shocks but revert toward a long-run equilibrium dictated by market fundamentals [20].

Motivated by these observations, this paper focuses on the linear bandit problem under mean-reverting dynamics. This formulation provides a principled way to capture structured nonstationarity commonly observed in practice. Our study considers two complementary settings: (i) a linear mean-reverting model, and (ii) a bounded nonlinear variant where the parameter is constrained within a fixed domain around the equilibrium. For both cases, we analyze the regret lower bound, and develop filtering-based algorithms—Kalman filtering for the linear case and particle filtering for the bounded case. The proposed method's regret upper bound matches the lower bound for the linear case. Our results establish that the proposed algorithms attain optimal steady-state regret with respect to the environment evolution noise, yielding tight theoretical guarantees.

Our work connects to a growing literature on non-stationary bandits that incorporate temporal dynamics. For instance, Mussi *et al.* [21] studies controllable linear dynamical systems and aims to maximize cumulative reward by strategically choosing the control input. Trella *et al.* [22] considers multi-armed bandits with reward changes induced by an unobserved latent state evolving according to an auto-regressive process. Bacchiocchi *et al.* [23] models the current reward as a linear combination of rewards from previous rounds, thereby capturing temporal dependencies in a non-parametric fashion. Other works [24, 25] examine non-stationarity driven by seasonal or periodic changes. The most closely related work is [26], which considers a multi-armed bandit problem and models reward means as an autoregressive process but assumes a state-informed setting where the latent state is the same as the reward and is directly observed at each step. By contrast, our setting is state-oblivious: the latent parameter is unobserved and must be inferred from noisy reward feedback. Moreover, unlike prior works that focus on specific algorithmic heuristics, our contribution lies in establishing tight regret lower and upper bounds under mean-reverting dynamics, thereby providing a sharp theoretical characterization of this important class of nonstationary bandits.

2. PROBLEM FORMULATION

We consider a sequential decision-making problem in the linear bandit setting. At each discrete time step $t = 1, 2, \ldots, T$, the learner selects an action $x_t \in \mathcal{D}$, where $\mathcal{D}$ is a known subset of $\mathbb{R}^d$ and $L = \max_{x \in \mathcal{D}} \|x\|$. After choosing the action x_t, the learner observes a scalar reward $y_t \in \mathbb{R}$, which is generated according to a linear model $y_t = x_t^\top \theta_t + \eta_t$, where $\theta_t \in \mathbb{R}^d$ is the unknown, time-varying parameter (also referred to as the system state), and $\eta_t \sim \mathcal{N}(0, \sigma_e^2)$ is independent zero-mean Gaussian noise. The goal of the learner is to sequentially choose actions x_t to maximize the cumulative reward, under the evolving environment parameter θ_t.

In contrast to classical stationary bandit problems, we consider a non-stationary setting in which the environment parameter θ_t evolves over time. Motivated by practical scenarios where systems tend to revert toward a stable long-term equilibrium, we assumed the state evolution follows a mean-reverting stochastic process. Specifically, θ_t is governed by the following dynamics

$$\theta_{t+1} = (1-\rho)\theta_t + \rho\mu + \xi_t, \tag{1}$$

where $\mu \in \mathbb{R}^d$ is the long-term mean, $\rho \in (0,1)$ is the reversion parameter, and $\xi_t \sim \mathcal{N}(0, \sigma_w^2 I)$ is the Gaussian noise. This model captures a structured evolution of the system state, wherein θ_t gradually drifts toward the equilibrium, subject to random perturbations.

In many real-world applications, the system state deviation from the long term equilibrium is bounded in norm, due to physical or domain-specific constraints. To account for this, we further consider a nonlinear extension, where the state vector will be projected onto the sphere centered at μ with radius S if the deviation exceeds S:

$$\theta_{t+1} = \Pi_{\mu,S}\left((1-\rho)\theta_t + \rho\mu + \xi_t\right), \tag{2}$$

where $\Pi_{\mu,S}(\cdot)$ denotes the Euclidean projection onto the ball $\mathcal{B}(\mu, S) := \{\theta \in \mathbb{R}^d : \|\theta - \mu\|_2 \leq S\}$.

At each round t, the learner selects an action $x_t \in \mathcal{D}$, while the optimal action is $x_t^* := \arg\max_{x \in \mathcal{D}} x^\top \theta_t$. Then the instantaneous regret at round t is $r_t := x_t^{*\top}\theta_t - x_t^\top \theta_t$. This represents the loss in expected reward incurred by selecting x_t instead of the optimal action x_t^*. The cumulative regret up to time T is defined as $R_T := \sum_{t=1}^T r_t$. An effective bandit algorithm aims to minimize R_T, ideally achieving sublinear regret as $T \to \infty$, even in the presence of a time-varying environment. However, as we will show in Section 3, in the specific setting considered in this paper, the regret of any proper bandit algorithm will grow linearly with respect to T. Therefore, it is more reasonable to consider the rate of this linear growth in the long term. To this end, we define the steady-state instantaneous regret as the expected regret incurred when $t \to \infty$, i.e.

$$\bar{r} := \lim_{t \to \infty} x_t^{*\top}\theta_t - x_t^\top \theta_t.$$

One can check that this definition is equivalent to the growth rate of R_T in the long term, *i.e.*, $\bar{r} = \lim_{T \to \infty} R_T/T$.

3. REGRET LOWER BOUND

In this section, we establish lower bounds for the mean-reverting bandit problem. These lower bounds show that for any admissible learning algorithm, the steady-state instantaneous regret cannot be arbitrarily small and is fundamentally constrained by the stochastic nature of the environment.

3.1. Regret lower bound for the linear case

Since the system state evolves according to a mean-reverting stochastic process, it admits a unique stationary distribution, and the distribution of θ_t converges to this stationary regime as $t \to \infty$. The long-term instantaneous regret defined previously corresponds to the expected regret under this stationary distribution. We first consider the unconstrained process in Eq.(1). Assume without loss of generality that the long-term mean is zero, then the steady-state distribution is $\theta_t \sim \mathcal{N}(0, \Sigma_\infty)$, where $\Sigma_\infty = \sigma_w^2 I/(2\rho - \rho^2)$. For this mean reversion process, we provide the following regret lower bound.

Theorem 1. *For the unconstrained process in Eq.*(1)*, for any proper bandit algorithm, the expected long-term instantaneous regret has the following lower bound:* $\mathbb{E}[\bar{r}] \geq \sqrt{2d/\pi}\sigma_w L\sqrt{\rho/(2-\rho)}$.

Proof Sketch. We can consider the structured action set $\mathcal{D} = \{\pm L/\sqrt{d}\}^d$, which corresponds to the 2^d vertices of a hypercube centered at the origin. By this construction, every action has norm $\|x\|_2 = L$, and for any direction of the system state θ_t, there exists an action in $\mathcal{D}$ that aligns optimally with it. In the mean-reverting linear process, the noise ξ_t can cause the sign of certain coordinates of θ_t to flip from one time step to the next. This implies that the optimal action may shift to a different orthant of the space. However, since the noise realization ξ_t is unobservable to the learner at t, no algorithm can perfectly anticipate these sign changes. Therefore, any algorithm will, with non-zero probability, select a suboptimal action that lies in a non-aligned orthant. Let s_t denote the vector of coordinate-wise signs of θ_t. Given knowledge of only the previous state θ_{t-1}, the learner will select $s_{t-1}L/\sqrt{d}$, while the optimal action is $s_t L/\sqrt{d}$. Then the instantaneous regret is $r_t = L/\sqrt{d} \cdot \theta_t^\top (s_t - s_{t-1})$. Hence, we have $\mathbb{E}[r_t] = L\sqrt{d} \cdot \mathbb{E}[\theta_{t,1}(s_{t,1} - s_{t-1,1})] = 4L\sqrt{d}\int_{\theta_{t,1}>0, \theta_{t-1,1}<0} \theta_{t,1}\, d\mathrm{P}(\theta_{t,1}, \theta_{t-1,1})$. In the steady state, $(\theta_{t-1,i}, \theta_{t,i})$ follows a bivariate Gaussian distribution with correlation coefficient $1-\rho$. By evaluating the above expectation via distributional integration, the proof can be completed. □

3.2. Regret lower bound for the bounded case

For the bounded evolve model in Eq.(2), the distribution of the system state is centered at μ and exhibits central symmetry. We assume $\mu = 0$ to simplify the analysis, without loss of generality. As $t \to \infty$ the process converges to a unique stationary distribution π_{ss}, which consists of two components, a continuous density supported in the interior of the ball, and a distribution supported on the boundary of the ball.

Using the symmetry, the density within the ball only depends on $\|\theta - \mu\|$, and the distribution on the boundary is uniform. The relative mass between the interior and the boundary can be determined through a steady-state balance condition that matches the incoming probability flow from the linear process and the projection mechanism at the boundary. Let p_s represent the total probability mass on the sphere, and $1 - p_s$ be the probability mass inside the sphere. The following theorem provides an analytical expression for p_s.

Theorem 2. *For the state evolve process in Eq.*(2)*, the steady state probability mass on the surface is* $p_s(\rho) = A(\rho)/\left(1 + A(\rho) - B(\rho)\right)$. *Let* $Q_{d/2}(a,b)$ *denote the generalized Marcum Q-function, which represents the right-tail probability of a non-central* χ^2 *distribution. Then* $A(\rho) = \mathbb{E}_{r|\mathit{inside}}\left[Q_{d/2}\left((1-\rho)r/\sigma_w, S/\sigma_w\right)\right]$, $B(\rho) = Q_{d/2}\left((1-\rho)S/\sigma_w, S/\sigma_w\right)$.

Proof Sketch. For a given state θ with magnitude $r = \|\theta\|$, the next state before projection is given by $(1-\rho)\theta + \xi \sim \mathcal{N}((1-\rho)\theta, \sigma_w^2 I)$. This state remains inside the sphere if its magnitude is less than S, otherwise, it is projected onto the spherical surface. Therefore, the conditional probability of the next state being on the surface is $\Pr(\|(1-\rho)\theta + \eta\| \geq S \mid \|\theta\| = r) = Q_{d/2}\left((1-\rho)r/\sigma_w, S/\sigma_w\right)$. In the steady state, the probability mass on the surface is equal to the probability of a state being projected onto the surface in one step, starting from the current steady-state distribution. By separating the

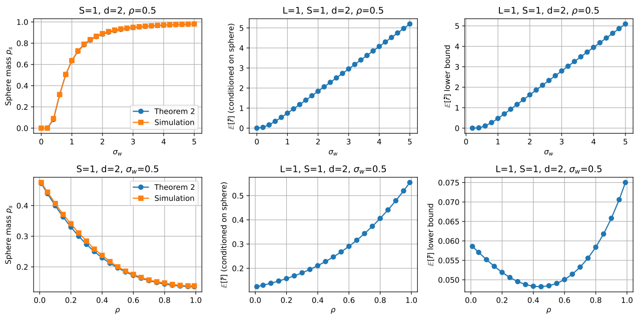

Fig. 1. Sphere probability mass, conditioned regret, and total regret lower bound under different values of σ_w and ρ.

contributions from states inside the sphere and those on the surface, we can establish that $p_s = (1-p_s)A(\rho) + p_s B(\rho)$, where $A(\rho)$ and $B(\rho)$ are defined in the theorem. Solving for the fixed point, we can complete the proof. □

Since $Q_{d/2}(a,b)$ is a monotonically increasing function with respect to a, $p_s(\rho)$ decreases monotonically with ρ. Besides, as $\rho \to 0^+$, the system state distribution approaches the sphere, and a "boundary layer" emerges in the vicinity of the boundary. Intuitively, the probability distribution "clings to the boundary", with the spherical term dominating.

Following a similar line of reasoning to Theorem 1, we can get the following lower bound for the evolution process in Eq.(2).

Theorem 3. *For the nonlinear bounded process in Eq.*(2)*, for any proper bandit algorithm, the expected steady state instantaneous regret satisfies* $\mathbb{E}[\bar{r}] \geq p_s \mathrm{R_s}$*, and* $\mathrm{R_s}$ *is the expected instantaneous regret conditioned on the event that* θ_t *is on the sphere,* $\mathrm{R_s} = 2L\sqrt{d}\int_0^S \left[\sigma_w \phi(\frac{\alpha u}{\sigma_w}) - \alpha u \Phi(-\frac{\alpha u}{\sigma_w})\right] f(u)du$*, where* $\alpha = 1-\rho$*,* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *is the PDF and CDF of the standard normal distribution, respectively, and* $f(u) = 2\Gamma(\frac{d}{2})/\sqrt{\pi}S\Gamma(\frac{d-1}{2}) \cdot (1 - u^2/S^2)^{\frac{d-3}{2}}$ *is the single dimension PDF of the uniform distribution on sphere* $\mathbb{S}^{d-1}$.

Proof Sketch. We again consider the action set $\mathcal{D} = \{\pm L/\sqrt{d}\}^d$. The optimal action at time $t+1$ is $a_{t+1}^* = L/\sqrt{d}\ \mathrm{sign}(\theta_{t+1})$, while the oracle selects the action based on information from the previous step: $a_t = L/\sqrt{d}\,\mathrm{sign}(\theta_t)$. Let $X = \theta_{t,i}$ and $Y = \theta_{t+1,i} = \alpha X + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_w^2)$. The regret is written as $\mathbb{E}[r_t] = \mathbb{E}[\sum_{i=1}^d r_i] = \sum_{i=1}^d L/\sqrt{d}\ \mathbb{E}[Y(\mathrm{sign}(Y) - \mathrm{sign}(X))] = 2L\sqrt{d}\ \mathbb{E}[|Y| \cdot \mathbf{1}\{XY < 0\}]$. We focus on the regret caused by the probability mass p_s of the system state being uniformly distributed on the spherical surface. Under this assumption, the probability density for a single dimension X is $f(u) = 2\Gamma(\frac{d}{2})/\sqrt{\pi}S\Gamma(\frac{d-1}{2}) \cdot (1 - \frac{u^2}{S^2})^{(d-3)/2}$. Due to symmetry, we have $\mathbb{E}\left[|Y|\mathbf{1}\{Y<0\}\big| |X| = u\right] = \mathbb{E}\left[-Y\mathbf{1}\{Y<0\}\right] = \sigma_w\phi(k) - \alpha u \Phi(-k)$, where $k = \alpha u/\sigma_w$. By integrating this expression over the distribution of $|X| = U$ and multiplying by p_s, the proof can be completed. □

This theorem presents the conditional regret in the form of an integral. To simplify the form, we may consider the scenario where the dimension d is relatively large. Then the single coordinate of the uniform distribution on the sphere is almost concentrated in a Gaussian distribution with a variance of S^2/d, and the corresponding regret is given by $\mathrm{R_{sphere}} \approx 2L(\sqrt{d}\sigma_2\phi(\gamma) - \alpha S\Phi(-\gamma))$, where $\gamma = \alpha S/(\sigma_w\sqrt{d})$. This regret increases with the growth of σ_w and ρ. In Fig. 1, we plot the sphere probability mass p_s in Theorem 2 (left column), conditional regret R_s (middle column), and the lower bound of total regret (right column) under different parameters. From the three upper subfigures, it can be seen that when σ_w is large, the sphere probability p_s approaches 1, and the regret lower bound exhibits a linear growth trend with respect to σ_w. From the three lower subfigures, it is shown that the regret lower bound is relatively small when $\rho \approx 0.5$, whereas it becomes larger when ρ is close to 0 or 1.

Algorithm 1 Recursive State Estimation for Linear Bandits

```
Require: Action set $\mathcal{D}$, time horizon T
Initialize prior belief $\hat{\theta}_0$ and filtering parameters
for $t = 1, \dots, T$ do
    Estimate state $\hat{\theta}_t$ using Eq.(3) or Eq.(5)
    Select action $x_t = \arg\max_{x \in \mathcal{D}} x^\top \hat{\theta}_t$
    Play action $x_t$, observe reward $y_t$
    Update the state estimate using Eq.(4) or Eq.(6)
```

4. PROPOSED METHOD AND REGRET BOUND

We address the mean-reverting bandit problem using a unified algorithmic framework that leverages recursive state estimation via filtering techniques. The central idea is to maintain an online estimate of the state θ_t, and to use this estimate to select actions that approximate the optimal ones at each round. The proposed procedure is summarized in Algorithm 1.

This approach can adapt to both the unconstrained linear evolution process in Eq.(1) and the constrained nonlinear process in Eq.(2). For the linear evolution process in Eq.(1), the optimal state estimator is the Kalman filter. Let $\hat{\theta}_t$ be the estimate of θ_t and $\hat{\Sigma}_t$ be the corresponding covariance matrix. The prediction and update equations are

$$\hat{\theta}_t = (1-\rho)\bar{\theta}_{t-1} + \rho\mu, \hat{\Sigma}_t = (1-\rho)^2\bar{\Sigma}_{t-1} + \sigma^2 I, \tag{3}$$

$$\bar{\theta}_t = \hat{\theta}_t + K_t(y_t - x_t^\top \hat{\theta}_t), \bar{\Sigma}_t = (I - K_t x_t^\top)\hat{\Sigma}_t, \tag{4}$$

where $K_t = \hat{\Sigma}_t x_t/(x_t^\top \hat{\Sigma}_t x_t + \sigma^2)$. Fot the constrained nonlinear process in Eq.(2), we adopt the particle filter to approximate the posterior distribution of θ_t. Assume that we use N particles $\{\theta_t^{(i)}\}_{i=1}^N$ with weights $\omega_t^{(i)}$, then the state prediction is done by

$$\theta_t^{(i)} = \Pi((1-\rho)\theta_{t-1}^{(i)} + \rho\mu + n_{t-1}^{(i)}), \hat{\theta}_t = \sum_{i=1}^N w_t^{(i)}\theta_t^{(i)}, \tag{5}$$

where $n_t^{(i)} \sim \mathcal{N}(0, \sigma^2 I)$. After observing y_t, we first update particle weights via

$$w_t^{(i)} \propto w_{t-1}^{(i)} \cdot \exp(-(y_t - x_t^\top \theta_t^{(i)})^2/2\sigma_e^2), \tag{6}$$

and then normalize the weights. If necessary, we will also resample the particles.

4.1. Regret bound for the linear evolution case

Under the state evolution model in Eq.(1), if we address this problem using Algorithm 1 combined with the Kalman filter, in the steady-state regime, the estimation error $\theta_t - \hat{\theta}_t$ becomes a stationary stochastic process with bounded second moment. Leveraging this property, we can control the instantaneous regret in the steady state.

Theorem 4. *Suppose the system state evolves according to Eq.*(1)*, and Algorithm 1 uses a Kalman filter to estimate* θ_t*. Assume* $\sigma_w = \sigma_e = \sigma$*, then after a sufficient number of time steps, the algorithm's expected instantaneous regret satisfies* $\mathbb{E}[\bar{r}] \leq C \cdot \sigma L\sqrt{d}/\sqrt{2\rho - \rho^2}$ *for some universal constant* $C > 0$ *with high probability.*

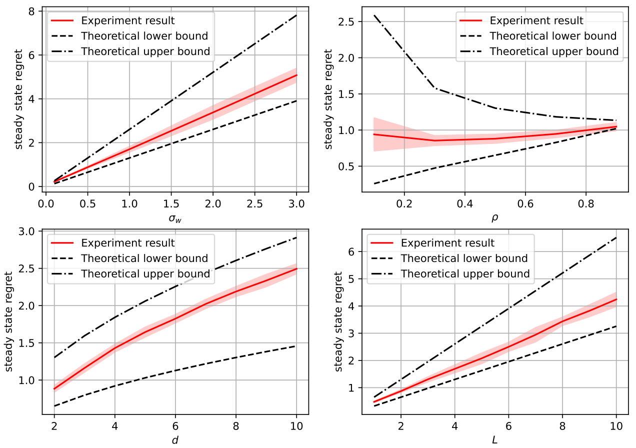

Fig. 2. Comparison of experimental and theoretical results.

Proof Sketch. At round t, the instantaneous regret is $r_t = \theta_t^\top A_t^* - \theta_t^\top A_t = \theta_t^\top A_t^* - \hat{\theta}_t^\top A_t + (\hat{\theta}_t - \theta_t)^\top A_t \leq (\theta_t - \hat{\theta}_t)^\top (A_t^* - A_t) \leq \|A_t^* - A_t\|_{\hat{\Sigma}_t} \|\hat{\theta}_t - \theta_t\|_{\hat{\Sigma}_t^{-1}} \leq 2L\sqrt{\lambda_{\max}(\hat{\Sigma}_t)} \|\hat{\theta}_t - \theta_t\|_{\hat{\Sigma}_t^{-1}}$, where $\lambda_{\max}(\hat{\Sigma}_t)$ denotes the largest eigen value of $\hat{\Sigma}_t$. The first inequality follows from the action-selection rule of the algorithm, namely that for any $a \in \mathcal{D}$, it holds that $\hat{\theta}_t^\top A_t \geq \hat{\theta}_t^\top a$, therefore we have $\hat{\theta}_t^\top A_t \geq \hat{\theta}_t^\top A_t^*$. The second inequality holds from Cauchy–Schwarz inequality. Using Kalman filtering for state estimation, the estimation error satisfies $\|\hat{\theta}_t - \theta_t\|_{\hat{\Sigma}_t^{-1}} \leq \sqrt{\beta(\delta)}$ with probability at least $1 - \delta$, where $\beta(\delta)$ corresponds to the $(1-\delta)$-quantile of the chi-squared distribution χ_d^2. Based on existing research, we have $\beta(\delta) \leq d + 2\sqrt{d \log(1/\delta)} + 2\log(1/\delta)$. This result implies that $\beta(\delta)$ is on the order of d, or $\beta(\delta) \sim \mathcal{O}(d)$. On the other hand, during the Kalman filtering process, the state evolution Equation (3) tends to enlarge the state covariance, while the measurement update Equation (4) reduces it. To derive an upper bound on $\lambda_{\max}(\hat{\Sigma}_t)$, we consider only the evolution process in Equation (4), which yields $\lambda_{\max}(\hat{\Sigma}_t) \leq (1-\rho)^2 \lambda_{\max}(\hat{\Sigma}_{t-1}) + \sigma^2$. After sufficiently many iterations, this recursion stabilizes, and we obtain $\lambda_{\max}(\hat{\Sigma}_t) \leq \frac{\sigma^2}{2\rho - \rho^2}$. Therefore, with high probability, the regret is bounded by $r_t \leq 2L\sigma\sqrt{\beta(\delta)}/\sqrt{2\rho - \rho^2}$. □

Theorem 4 shows that the steady state regret is linearly related to σ, L, and $\sqrt{d}$. Furthermore, this regret decreases as ρ grows. Compared to the lower bound in Theorem 1, this regret bound is $1/\rho$ times the lower bound established Theorem 1. This demonstrates that the regret of Algorithm 1 is optimal with respect to σ, d, and L. When ρ is large, this regret upper bound is also very close to the lower bound, which is on the theoretically optimal level. Only when $\rho \to 0^+$, the regret of Theorem 1 will be significantly larger than the theoretical lower bound. However, for the linear evolution process described in Eq. (1), as $\rho \to 0^+$, the steady-state variance of the state distribution σ_∞ tends to infinity, which implies that the system state diverges. This scenario would not occur in practice.

5. EXPERIMENTS

We compare simulation results against the lower bound from Theorem 1 and the upper bound from Theorem 4. Specifically, we set the time horizon to $T = 10,000$ and use the average instantaneous regret over the final $2,000$ time steps as a measure of the steady-state regret. The default parameters for our experiments are $\sigma = 0.5$, $\rho = 0.5$, $d = 2$, and $L = 2$. We vary σ from 0.1 to 3, ρ from 0.1 to 0.9, d from 2 to 10, and L from 1 to 10. We perform 100 independent runs for each parameter setting and record the steady-state regret per time step. The results are presented in Fig. 2. The figure clearly shows that the empirical results fall consistently between the theoretical lower and upper bounds. Furthermore, the regret exhibits a linear relationship with σ_w, L, and $\sqrt{d}$. These findings are in complete agreement with our theoretical predictions. Additionally, we observe that when ρ is large, the regret upper bound is very close to the lower bound. This indicates that Algorithm 1 performs nearly optimally when the mean reversion rate is high.

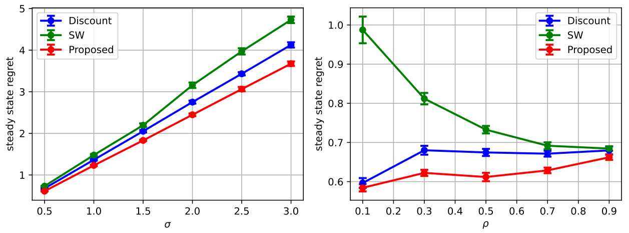

Fig. 3. Regret comparison under linear dynamics (Eq.(1)).

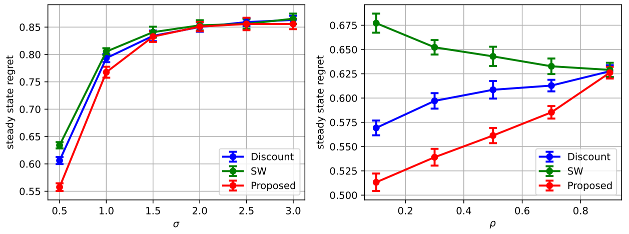

Fig. 4. Regret comparison under nonlinear dynamics (Eq.(2)).

We also compare the proposed method against two baseline methods that incorporate no knowledge of the underlying system dynamics: sliding window method [13] (SW) and discount method [14] (Discount). The SW method estimates the system state using only the most recent W observations, discarding all earlier data. The Discount method applies exponentially decaying weights to past observations, assigning higher importance to recent data. The discount factor controls the rate of memory decay. The regret of each method is computed over 5000 time steps and averaged over 50 repeated trials. For the linear system state evolution process in Eq.(1), Fig. 3 presents the regret curves under various values of ρ and σ. For the nonlinear bounded evolution model in Eq.(2), similar results are given in Fig. 4. It's shown that the proposed method consistently outperforms the baselines across all scenarios, and its advantage becomes more pronounced for large σ and small ρ.

6. CONCLUSION

This paper studied the nonstationary linear bandit problem where the latent parameter evolves according to a mean-reverting process. We analyzed the fundamental difficulty of this setting and established regret lower bounds for both the linear mean-reverting model and a nonlinear variant. We propose filtering-based algorithms that effectively track the latent parameter. For the linear case, we show that the resulting regret upper bound matches the lower bound and is optimal with respect to the environment evolution noise. These results provide tight characterization of regret in mean-reverting linear bandits. Simulation results demonstrate that our approach outperforms existing baselines.

7. REFERENCES

[1] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al., "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[2] Eric M Schwartz, Eric T Bradlow, and Peter S Fader, "Customer acquisition via display advertising using multi-armed bandit experiments," *Marketing Science*, vol. 36, no. 4, pp. 500–522, 2017.

[3] Sofía S Villar, Jack Bowden, and James Wason, "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 30, no. 2, pp. 199, 2015.

[4] Tor Lattimore, Koby Crammer, and Csaba Szepesvári, "Linear multi-resource allocation with semi-bandit feedback," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

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

[6] I-Hong Hou, "Distributed no-regret learning for multi-stage systems with end-to-end bandit feedback," in *Proceedings of the Twenty-fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2024, pp. 41–50.

[7] Nida Zamir and I-Hong Hou, "Deep index policy for multi-resource restless matching bandit and its application in multi-channel scheduling," in *Proceedings of the Twenty-fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2024, pp. 71–80.

[8] Quang Minh Nguyen and Eytan Modiano, "Learning to schedule in non-stationary wireless networks with unknown statistics," in *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2023, pp. 181–190.

[9] Chen-Yu Wei and Haipeng Luo, "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach," in *Conference on learning theory*. PMLR, 2021, pp. 4300–4354.

[10] Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford, "Efficient contextual bandits in non-stationary worlds," in *Proceedings of the 31st Conference On Learning Theory*, Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, Eds. 06–09 Jul 2018, vol. 75 of *Proceedings of Machine Learning Research*, pp. 1739–1776, PMLR.

[11] Lihong Li, Wei Chu, John Langford, and Robert E Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[12] Zhenshan Bing, David Lerch, Kai Huang, and Alois Knoll, "Meta-reinforcement learning in non-stationary and dynamic environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3476–3491, 2022.

[13] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu, "Learning to optimize under non-stationarity," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Kamalika Chaudhuri and Masashi Sugiyama, Eds. 16–18 Apr 2019, vol. 89 of *Proceedings of Machine Learning Research*, pp. 1079–1087, PMLR.

[14] Yoan Russac, Claire Vernade, and Olivier Cappé, "Weighted linear bandits for non-stationary environments," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou, "A simple approach for non-stationary linear bandits," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Silvia Chiappa and Roberto Calandra, Eds. 26–28 Aug 2020, vol. 108 of *Proceedings of Machine Learning Research*, pp. 746–755, PMLR.

[16] Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec, "Efficient change-point detection for tackling piecewise-stationary bandits," *Journal of Machine Learning Research*, vol. 23, no. 77, pp. 1–40, 2022.

[17] Reda Alami, "Bayesian change-point detection for bandit feedback in non-stationary environments," in *Asian Conference on Machine Learning*. PMLR, 2023, pp. 17–31.

[18] Lilian Besson and Emilie Kaufmann, "The generalized likelihood ratio test meets klucb: an improved algorithm for piecewise non-stationary bandits," in *Proceedings of Machine Learning Research*, 2019, vol. 1, p. 35.

[19] Eric Chin, David Chieng, Victor Teh, Marek Natkaniec, Krzysztof Loziak, and Janusz Gozdecki, "Wireless link prediction and triggering using modified ornstein–uhlenbeck jump diffusion process," *Wireless networks*, vol. 20, no. 3, pp. 379–396, 2014.

[20] Jessica A Wachter, "Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets," *Journal of financial and quantitative analysis*, vol. 37, no. 1, pp. 63–91, 2002.

[21] Marco Mussi, Alberto Maria Metelli, and Marcello Restelli, "Dynamical linear bandits," in *International Conference on Machine Learning*. PMLR, 2023, pp. 25563–25587.

[22] Anna L. Trella, Walter H. Dempsey, Asim Gazi, Ziping Xu, Finale Doshi-Velez, and Susan Murphy, "Non-stationary latent auto-regressive bandits," in *Reinforcement Learning Conference*, 2025.

[23] Francesco Bacchiocchi, Gianmarco Genalti, Davide Maran, Marco Mussi, Marcello Restelli, Nicola Gatti, and Alberto Maria Metelli, "Autoregressive bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 937–945.

[24] Parth Thaker, Vineet Gattani, Vignesh Tirukkonda, Pouria Saidi, and Gautam Dasarathy, "Non-stationary bandits with periodic behavior: Harnessing ramanujan periodicity transforms to conquer time-varying challenges," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7790–7794.

[25] Giuseppe Di Benedetto, Vito Bellini, and Giovanni Zappella, "A linear bandit for seasonal environments," *arXiv preprint arXiv:2004.13576*, 2020.

[26] Qinyi Chen, Negin Golrezaei, and Djallel Bouneffouf, "Non-stationary bandits with auto-regressive temporal dependency," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7895–7929, 2023.