Lock-In Regret: RR-Decay for Nonstationary Top-k Linear Bandits

Arian Aghamohseni*, Ramtin Moslemi†, Siavash Ahmadi‡
* Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran
Email: arian.aghamohseni19@sharif.edu
† Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Email: ramtin.moslemi@sharif.edu
‡ Electronics Research Institute, Sharif University of Technology, Tehran, Iran
Email: s.ahmadi@sharif.edu

***Abstract*—Decision-making systems in the real world rarely face a stable environment: costs and rewards drift, sometimes abruptly, yet the system must keep acting while only receiving limited feedback. A core difficulty in this non-stationary bandit setting is identifying what remains reliably "best" when the underlying signal changes and may even become temporarily indistinguishable due to noise or near-ties.**

We consider a structured version of this challenge where the learner repeatedly chooses a small subset of options and only observes the total loss of that subset. Although individual losses evolve over time, we assume the identity of an optimal subset stays the same (a realistic regime in applications where the best choices persist but their measured performance fluctuates). The main obstacle is that the instantaneous advantage of the optimal subset can vanish for long periods, making standard margin-based analyses inapplicable.

We resolve this by combining decaying random exploration with a simple closed-form estimator that exploits a known symmetry of the exploration process. Under a weak, time-averaged separation condition that explicitly allows long near-tie segments, we prove a regret bound of order $\widetilde{O}(\mathrm{poly}(d)\sqrt{T})$. Our analysis reveals a "lock-in" effect: after enough exploratory measurements, the algorithm reliably identifies the optimal subset and thereafter incurs essentially no additional exploitation regret.

***Index Terms*—linear bandits, non-stationary bandits, combinatorial bandits**

I. Introduction

Sequential decision-making systems often operate in *drifting* environments: costs or rewards change over time, while the learner must keep acting using limited feedback. Multi-armed bandits formalize this exploration–exploitation tradeoff and admit sharp guarantees in stationary stochastic settings (logarithmic regret under positive gaps, and $\widetilde{O}(\sqrt{T})$ in gap-free/minimax settings) [1]–[3]. Linear bandits extend this framework to structured actions and high-dimensional problems; confidence-based methods such as *Optimism in the Face of Uncertainty for Linear bandits* (OFUL) achieve $\widetilde{O}(d\sqrt{T})$ regret in stationary stochastic models [4], [5]. The non-stationary case is substantially harder, since achievable regret generally depends on how much the environment is allowed to vary [6].

Code repository: https://github.com/MagicWizardMaker/RR-Decay

We study a structured non-stationary linear bandit with *top-k subset actions*: at each round, the learner selects a k-subset of d coordinates (equivalently, a binary vector with exactly k ones), incurs a linear loss, and observes only a noisy scalar feedback. Our benchmark is a *stable optimum*: the loss vector may evolve over time, but there exists a fixed subset $S^\star$ of size k that is never worse than any other subset. For top-k actions, this is equivalent to a possibly non-strict *boundary ordering* between coordinates inside and outside $S^\star$, so ties across the boundary are allowed. These ties are the main obstacle: multiple subsets can be optimal at a given round, and repeated near-ties can make identification ambiguous unless some cumulative separation is present.

We propose *RR-decay*, a simple algorithm that combines decaying uniform top-k exploration with a closed-form ridge-type estimator that exploits the *known* covariance of uniform top-k sampling. The analysis reduces learning to certifying the *pairwise boundary ordering* between coordinates in $S^\star$ and its complement using exploration-only statistics. Under a weak average-separation (identifiability) condition that allows long tie or near-tie segments, we prove an expected static regret bound of order $\widetilde{O}(\mathrm{poly}(d,k)\sqrt{T})$.

A. Contributions

In this structured non-stationary top-k linear bandit setting, we make four contributions:

- **Stable-optimum benchmark with ties.** We formalize a non-stationary top-k linear bandit model with a fixed optimal subset $S^\star$, allowing non-strict boundary ordering (ties) across rounds.
- **RR-decay: decaying uniform exploration + covariance-aware ridge estimation.** We introduce RR-decay, which uses uniform top-k exploration and a ridge-type estimator built from the known second-moment matrix of the exploration distribution.
- **Lock-in via pairwise boundary certification.** We show that pairwise exploration statistics certify the correct boundary ordering between inside and outside coordinates, yielding a finite-sample lock-in guarantee.
- **$\widetilde{O}(\mathrm{poly}(d,k)\sqrt{T})$ regret under weak average separation.** Under a weak identifiability condition that permits

long zero-gap segments, we prove a static regret bound of order $\widetilde{O}(\mathrm{poly}(d,k)\sqrt{T})$.

II. Related Work

Regret guarantees are classical in stationary stochastic bandits: *upper confidence bound* (UCB)-style methods achieve logarithmic regret under positive gaps [1], while gap-free minimax analyses give $\widetilde{O}(\sqrt{T})$ dependence on the horizon [2], [3]. Linear bandits exploit feature structure to obtain $\widetilde{O}(d\sqrt{T})$-type regret in stationary stochastic models via optimism and confidence sets [4], [5]. Our action space is combinatorial (top-k subset selection), closely related to combinatorial bandits and bandit combinatorial optimization [7]; the key distinction here is that uniform top-k exploration has a closed-form covariance, which we exploit directly in both estimation and analysis.

Non-stationarity makes bandit learning harder because performance depends on the allowed temporal variation. Variation-budget formulations and related models characterize this dependence and motivate adaptive exploration mechanisms [6]. Recent non-stationary structured/linear bandit methods typically assume variation control or piecewise stationarity and then adapt to changes [3], [6], [8]. We study a different regime: the loss vector may drift arbitrarily, but subject to a stable-optimum (possibly tied) top-k structure. In this regime, our contribution is a simple lock-in style analysis showing that $\widetilde{O}(\mathrm{poly}(d,k)\sqrt{T})$ static regret is achievable under a weak average-separation condition tailored to the stable-optimum benchmark.

III. Results Overview

Fix integers $d \geq 2$ and $1 \leq k \leq d-1$. Let $\mathbf{1} \in \mathbb{R}^d$ be the all-ones vector and $\mathbf{1}_S \in \{0,1\}^d$ the indicator of $S \subset [d]$. This section states the assumptions and core lemmas used in the regret proof. The key idea is to reduce learning the optimal top-k action to certifying a *strict boundary ordering* of the estimator between coordinates in $S^\star$ and $[d] \setminus S^\star$ after enough exploration (via Assumption 4).

A. Timing model and basic regularity

Assumption 1 (Non-anticipation)**.** There is a filtration $(\mathcal{F}_t)_{t\geq 0}$ such that for each round t, θ_t is $\mathcal{F}_{t-1}$-measurable (chosen before a_t), and the noise η_t is generated after a_t is played.

Assumption 2 (Boundedness and conditional mean-zero noise)**.** There exist constants $B_\theta, B_\eta \geq 0$ such that for all t,

$$\|\theta_t\|_\infty \leq B_\theta, \qquad |\eta_t| \leq B_\eta \text{ a.s.}, \qquad \mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}, a_t, \theta_t] = 0.$$

Consequently,

$$|y_t| = |\langle \theta_t, a_t\rangle + \eta_t| \leq kB_\theta + B_\eta =: B_y \qquad \text{a.s.}$$

B. Stable optimum and boundary ordering

Let $\mathcal{A} = \{a \in \{0,1\}^d : \sum_{i=1}^d a_i = k\}$ and $\ell_t(a) = \langle \theta_t, a\rangle$.

Assumption 3 (Stable optimal top-k set)**.** There exists a fixed subset $S^\star \subset [d]$ with $|S^\star| = k$ such that for all t and all $a \in \mathcal{A}$,

$$\langle \theta_t, \mathbf{1}_{S^\star}\rangle \leq \langle \theta_t, a\rangle.$$

Proposition 1 (Optimality $\Longleftrightarrow$ boundary ordering)**.** *Fix t and a subset $S^\star$ with $|S^\star| = k$. The following are equivalent:*
(P1) $\mathbf{1}_{S^\star} \in \arg\min_{a\in\mathcal{A}} \langle \theta_t, a\rangle$.
(P2) $\max_{j\in S^\star} \theta_{t,j} \leq \min_{i\notin S^\star} \theta_{t,i}$.
Equivalently, the instantaneous boundary gap

$$\gamma_t := \min_{i\notin S^\star} \theta_{t,i} - \max_{j\in S^\star} \theta_{t,j}$$

satisfies $\gamma_t \geq 0$.

Proof. (*P2 $\Rightarrow$ P1*) For any $S \subset [d]$ with $|S| = k$, let $A = S \setminus S^\star$, $B = S^\star \setminus S$, and $m = |A| = |B|$. Then

$$\begin{aligned}\sum_{u\in S} \theta_{t,u} - \sum_{v\in S^\star} \theta_{t,v} &= \sum_{i\in A} \theta_{t,i} - \sum_{j\in B} \theta_{t,j} \\ &\geq m \min_{i\notin S^\star} \theta_{t,i} - m \max_{j\in S^\star} \theta_{t,j} \\ &= m\,\gamma_t \geq 0.\end{aligned}$$

So $\mathbf{1}_{S^\star}$ is optimal.

(*P1 $\Rightarrow$ P2*) If P2 fails, some $j \in S^\star$ and $i \notin S^\star$ satisfy $\theta_{t,i} < \theta_{t,j}$. Swapping j out for i strictly decreases the top-k sum, contradicting P1. □

C. Uniform top-k exploration geometry

Let q be the uniform distribution over $\mathcal{A}$. An "exploration round" t_s means

$$a_{t_s} \sim q \qquad \text{independently of } \mathcal{F}_{t_s-1}.$$

For $a \sim q$, define

$$G := \mathbb{E}_{a\sim q}[aa^\top] \in \mathbb{R}^{d\times d}.$$

Lemma 1 (Closed form and spectrum of G)**.** *Let $p := k/d$ and $\rho := k(k-1)/(d(d-1))$. Then*

$$G_{ii} = p, \qquad G_{ij} = \rho \ (i \neq j), \qquad \textit{hence} \qquad G = (p-\rho)I + \rho\,\mathbf{1}\mathbf{1}^\top.$$

Moreover,

$$\lambda_\parallel = p+(d-1)\rho = \frac{k^2}{d} \quad (\textit{eigenvector } \mathbf{1}), \qquad \lambda_\perp = p-\rho = \frac{k(d-k)}{d(d-1)} \quad ($$

and therefore

$$\alpha := \lambda_{\min}(G) = \lambda_\perp = \frac{k(d-k)}{d(d-1)} > 0.$$

Proof. By symmetry, $\mathbb{P}(a_i = 1) = k/d = p$, so $G_{ii} = p$. For $i \neq j$,

$$\mathbb{P}(a_i = a_j = 1) = \frac{\binom{d-2}{k-2}}{\binom{d}{k}} = \frac{k(k-1)}{d(d-1)} = \rho,$$

so $G_{ij} = \rho$. The eigenvalues follow from the form $aI + b\mathbf{1}\mathbf{1}^\top$ with $a = p - \rho$ and $b = \rho$. □

D. Pairwise moment identities under uniform exploration

Lemma 2 (Unbiased linear moment)**.** *Fix t and suppose $a \sim q$ is drawn independently of $\mathcal{F}_{t-1}$. Let $y = \langle \theta_t, a \rangle + \eta_t$ with $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}, a, \theta_t] = 0$. Then*

$$\mathbb{E}[a\, y \mid \theta_t] = G\theta_t.$$

Proof. Conditioning on θ_t,

$$\mathbb{E}[a\, y \mid \theta_t] = \mathbb{E}[a(a^\top \theta_t) \mid \theta_t] + \mathbb{E}[a\eta_t \mid \theta_t] = \mathbb{E}[aa^\top]\theta_t = G\theta_t,$$

using conditional mean-zero noise. □

Lemma 3 (Pairwise expected signal)**.** *Fix t and suppose $a \sim q$ is drawn independently of $\mathcal{F}_{t-1}$. Let $y = \langle \theta_t, a \rangle + \eta_t$ with $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}, a, \theta_t] = 0$. Then for any $i \neq j$,*

$$\mathbb{E}[(a_i - a_j)\, y \mid \theta_t] = \alpha\, (\theta_{t,i} - \theta_{t,j}),$$

where α is from Lemma 1.

Proof. By Lemma 2,

$$\mathbb{E}[(a_i - a_j)y \mid \theta_t] = (e_i - e_j)^\top G\theta_t.$$

Since $(e_i - e_j)^\top \mathbf{1} = 0$, the vector $(e_i - e_j) \in \mathbf{1}^\perp$, where G acts as αI. Thus $(e_i - e_j)^\top G\theta_t = \alpha(\theta_{t,i} - \theta_{t,j})$. □

E. Uniform-in-prefix identifiability (bursts with slack)

For $i \notin S^\star$ and $j \in S^\star$, define $\Delta_t(i,j) := \theta_{t,i} - \theta_{t,j} \geq 0$ (Proposition 1). Let $t_1 < t_2 < \cdots$ be the (random) exploration times and set

$$D_n(i,j) := \sum_{s=1}^{n} \Delta_{t_s}(i,j).$$

Assumption 4 (Uniform-in-prefix separation with slack)**.** There exist constants $\bar{\gamma} > 0$, $C \geq 0$, and an integer $n_{\min} \geq 1$ such that, with probability at least $1 - \delta_{\mathrm{sep}}$ (over the exploration coin flips), for all $i \notin S^\star$, $j \in S^\star$, and all $n \geq n_{\min}$,

$$D_n(i,j) \;\geq\; n\bar{\gamma} - C.$$

Remark 1. Assumption 4 is a uniform lower-envelope condition along exploration times. It allows long tie segments (many $\Delta_{t_s}(i,j) = 0$) as long as the cumulative deficit is absorbed by C, and it permits an initial burn-in up to $n_{\min}$.

F. Pairwise exploration statistics and concentration

For $i \neq j$, define

$$Z_n(i,j) := \sum_{s=1}^{n} (a_{t_s,i} - a_{t_s,j})\, y_{t_s}.$$

By Lemma 3,

$$\mathbb{E}[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}] = \alpha \sum_{s=1}^{n} \left(\theta_{t_s,i} - \theta_{t_s,j}\right) = \alpha\, D_n(i,j). \tag{1}$$

Lemma 4 (Uniform concentration for $Z_n(i,j)$)**.** *Assume Assumptions 1–2. Fix $\delta \in (0,1)$ and define*

$$L(\delta) := \log\left(\frac{2d^2T}{\delta}\right).$$

With probability at least $1 - \delta$, simultaneously for all $n \in \{1, \ldots, T\}$ and all ordered pairs (i,j) with $i \neq j$,

$$\left| Z_n(i,j) - \mathbb{E}[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}] \right| \leq 2B_y \sqrt{2n\, L(\delta)}.$$

Proof. Fix (i,j) and $n \leq T$. Let

$$X_s := (a_{t_s,i} - a_{t_s,j})\, y_{t_s}, \qquad Z_n(i,j) = \sum_{s=1}^{n} X_s.$$

With

$$\mathcal{H}_s := \sigma(t_{1:s}, \theta_{t_1:t_s}, (a_{t_r}, y_{t_r})_{r \leq s}),$$

define

$$m_s := \mathbb{E}[X_s \mid \mathcal{H}_{s-1}, t_s, \theta_{t_s}], \qquad M_s := \sum_{r=1}^{s} (X_r - m_r).$$

Then (M_s) is a martingale. On exploration rounds, $a_{t_s} \sim q$ is independent of $\mathcal{H}_{s-1}$ given (t_s, θ_{t_s}), and the noise is conditionally mean-zero, so by Lemma 3,

$$m_s = \alpha(\theta_{t_s,i} - \theta_{t_s,j}).$$

Hence $\mathbb{E}[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}] = \sum_{s=1}^{n} m_s$.

Also, $|X_s| \leq B_y$ and $|m_s| \leq B_y$, so $|X_s - m_s| \leq 2B_y$ a.s. Azuma–Hoeffding (e.g., [9]) gives, for any $\delta' \in (0,1)$,

$$\mathbb{P}\Big(|M_n| \geq 2B_y \sqrt{2n \log(2/\delta')}\Big) \leq \delta'.$$

Since $M_n = Z_n(i,j) - \mathbb{E}[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}]$, the same bound holds for the centered statistic. A union bound over $n = 1, \ldots, T$ and at most $d(d-1) \leq d^2$ ordered pairs, with $\delta' = \delta/(d^2T)$, yields the claim. □

G. Covariance-aware ridge estimator: pairwise closed form

After n exploration samples, define

$$b_n := \sum_{s=1}^{n} a_{t_s} y_{t_s}, \qquad \hat{\theta}_n := (\lambda I + nG)^{-1} b_n,$$

for $\lambda > 0$.

Lemma 5 (Pairwise form of $\hat{\theta}_n$)**.** *For any $i \neq j$,*

$$\hat{\theta}_{n,i} - \hat{\theta}_{n,j} = \frac{1}{\lambda + n\alpha}\, Z_n(i,j),$$

where $\alpha = \lambda_{\min}(G)$.

Proof.

$$\hat{\theta}_{n,i} - \hat{\theta}_{n,j} = (e_i - e_j)^\top (\lambda I + nG)^{-1} b_n.$$

Because $(e_i - e_j) \in \mathbf{1}^\perp$ and G acts as αI on $\mathbf{1}^\perp$,

$$(\lambda I + nG)^{-1}(e_i - e_j) = \frac{1}{\lambda + n\alpha}(e_i - e_j).$$

Substituting $b_n = \sum_{s=1}^{n} a_{t_s} y_{t_s}$ gives

$$\hat{\theta}_{n,i} - \hat{\theta}_{n,j} = \frac{1}{\lambda + n\alpha} \sum_{s=1}^{n} (a_{t_s,i} - a_{t_s,j}) y_{t_s} = \frac{1}{\lambda + n\alpha} Z_n(i,j).$$

□

H. Strict boundary ordering implies a unique top-k set

Lemma 6 (Strict boundary ordering $\Rightarrow$ unique top-k set)**.** *Fix $S^\star \subset [d]$ with $|S^\star| = k$ and let $v \in \mathbb{R}^d$. If $v_j < v_i$ for all $j \in S^\star$ and all $i \notin S^\star$, then the indices of the k smallest coordinates of v are uniquely $S^\star$.*

Proof. Every coordinate in $S^\star$ is strictly smaller than every coordinate in $[d] \setminus S^\star$. Thus the k smallest coordinates of v are exactly those indexed by $S^\star$. □

IV. Algorithm

RR-decay alternates between *uniform top-k exploration* and *greedy exploitation*. Under uniform top-k sampling, the exploration covariance $G = \mathbb{E}_{a\sim q}[aa^\top]$ is known in closed form (Lemma 1), and the estimator satisfies the pairwise identity

$$\hat{\theta}_{n,i} - \hat{\theta}_{n,j} = \frac{1}{\lambda + n\alpha} Z_n(i,j) \quad \text{(Lemma 5)}.$$

Thus exploration provides pairwise evidence, and exploitation plays the k smallest estimated coordinates.

A. Exploration schedule and indexing

At round t, the learner draws an independent coin $U_t \sim \mathrm{Unif}[0,1]$ and explores if $U_t \le \varepsilon_t$, where

$$\varepsilon_t := \min\Big\{1, \frac{c}{\sqrt{t}}\Big\}, \qquad c > 0. \tag{2}$$

Let

$$n(t) := \sum_{s=1}^{t} \mathbf{1}\{s \text{ is an exploration round}\}$$

be the number of exploration samples collected by time t. During round t, exploitation uses $\hat{\theta}_{n(t-1)}$; after an exploration update at t, the estimate becomes $\hat{\theta}_{n(t)}$.

B. Uniform exploration and covariance-aware estimator

Let q be the uniform distribution over $\mathcal{A}$. On exploration round t, RR-decay draws $a_t \sim q$ independently of $\mathcal{F}_{t-1}$ and observes scalar feedback y_t.

Let $t_1 < t_2 < \cdots$ be the exploration times, and write $n = n(t)$. RR-decay stores

$$b_n := \sum_{s=1}^{n} a_{t_s} y_{t_s} \in \mathbb{R}^d,$$

and forms the covariance-aware ridge estimate

$$\hat{\theta}_n := (\lambda I + nG)^{-1} b_n, \tag{3}$$

with $\lambda > 0$. Unlike standard ridge regression, this uses the *known* exploration moment matrix G (not an empirical Gram matrix), which makes the pairwise reduction in Section III exact.

Algorithm 1 RR-decay (uniform top-k exploration + covariance-aware ridge)

Input: dimension d, subset size k, horizon T, ridge $\lambda > 0$, schedule constant $c > 0$
Set $p \leftarrow k/d$, $\rho \leftarrow k(k-1)/(d(d-1))$, and $G \leftarrow (p-\rho)I + \rho\,\mathbf{1}\mathbf{1}^\top$
Initialize $n \leftarrow 0$, $b \leftarrow 0 \in \mathbb{R}^d$, $\hat{\theta} \leftarrow 0 \in \mathbb{R}^d$
for $t = 1$ **to** T **do**
  Set $\varepsilon_t \leftarrow \min\{1, c/\sqrt{t}\}$ and draw $U_t \sim \mathrm{Unif}[0,1]$
  if $U_t \le \varepsilon_t$ **then** ▷ Explore
    Sample $a_t \sim q$ uniformly over $\mathcal{A}$
    Receive scalar feedback y_t
    $n \leftarrow n+1, \quad b \leftarrow b + a_t y_t$
    Update $\hat{\theta} \leftarrow (\lambda I + nG)^{-1} b$ (e.g., via (4))
  else ▷ Exploit
    $S \leftarrow \textsc{TopKMin}(\hat{\theta})$
    Play $a_t \leftarrow \mathbf{1}_S$
    Receive scalar feedback y_t (unused)
  end if
end for

C. Greedy exploitation and tie-breaking

For $v \in \mathbb{R}^d$, let $\textsc{TopKMin}(v) \subset [d]$ be the indices of the k smallest coordinates of v, with deterministic tie-breaking (smaller index first). On exploitation rounds, RR-decay plays

$$a_t = \mathbf{1}_{\textsc{TopKMin}(\hat{\theta}_{n(t-1)})}.$$

Tie-breaking only makes the rule deterministic; after lock-in, the boundary ordering is strict with high probability, so the selected top-k set is unique.

D. Fast inverse from the rank-one structure of G

By Lemma 1,

$$G = (p-\rho)I + \rho\,\mathbf{1}\mathbf{1}^\top, \qquad p := \frac{k}{d}, \quad \rho := \frac{k(k-1)}{d(d-1)}.$$

Hence

$$\lambda I + nG = A_n I + B_n\,\mathbf{1}\mathbf{1}^\top, \qquad A_n := \lambda + n(p-\rho), \quad B_n := n\rho.$$

Applying Sherman–Morrison gives

$$(\lambda I + nG)^{-1} = \frac{1}{A_n} I - \frac{B_n}{A_n(A_n + B_n d)}\,\mathbf{1}\mathbf{1}^\top. \tag{4}$$

Therefore

$$\hat{\theta}_n = \frac{1}{A_n}\, b_n - \frac{B_n}{A_n(A_n + B_n d)}\,\mathbf{1}\,(\mathbf{1}^\top b_n),$$

so $\hat{\theta}_n$ can be computed in $O(d)$ time. Selecting $\textsc{TopKMin}(\hat{\theta}_n)$ costs $O(d \log d)$ by sorting (or $O(d)$ expected time via selection).

V. Regret Analysis

We bound the expected regret

$$\mathbb{E}[R_T], \qquad R_T := \sum_{t=1}^{T} \langle \theta_t,\ a_t - a^\star \rangle,$$

where $a^\star = \mathbf{1}_{S^\star}$ is the stable comparator (Assumption 3). Throughout we use Assumptions 1, 2, and 4.

A. Strict lock-in after n_0 exploration samples

Let $t_1 < t_2 < \cdots$ be the (random) exploration times and recall

$$Z_n(i,j) := \sum_{s=1}^{n} (a_{t_s,i} - a_{t_s,j})\, y_{t_s}.$$

By Lemma 5, for every $i \neq j$,

$$\hat{\theta}_{n,i} - \hat{\theta}_{n,j} = \frac{1}{\lambda + n\alpha}\, Z_n(i,j), \qquad \alpha = \lambda_{\min}(G) > 0. \quad (5)$$

Fix $\delta \in (0,1)$ and define

$$L_T := \log\Big(\frac{2d^2T}{\delta}\Big).$$

Let $\mathcal{E}_{\mathrm{conc}}$ be the event from Lemma 4: with probability at least $1-\delta$, for all $n \in \{1,\dots,T\}$ and all ordered pairs (i,j),

$$\Big| Z_n(i,j) - \mathbb{E}\big[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}\big] \Big| \le 2B_y\sqrt{2n\, L_T}. \quad (6)$$

Let $\mathcal{E}_{\mathrm{sep}}$ be the prefix-separation event from Assumption 4: with probability at least $1-\delta_{\mathrm{sep}}$, for all $i \notin S^\star$, $j \in S^\star$, and all $n \ge n_{\min}$,

$$D_n(i,j) := \sum_{s=1}^{n} \big(\theta_{t_s,i} - \theta_{t_s,j}\big) \ \ge\ n\bar{\gamma} - C. \quad (7)$$

Define $\mathcal{E}_{\mathrm{lock}} := \mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{sep}}$, so

$$\mathbb{P}(\mathcal{E}_{\mathrm{lock}}) \ge 1 - (\delta + \delta_{\mathrm{sep}}).$$

Theorem 1 (Strict lock-in after n_0 exploration samples)**.** *Assume Assumptions 1, 2, 3, and 4. Define*

$$n_0 := \max\left\{ n_{\min},\ \left\lceil \frac{4C}{\bar{\gamma}} + \frac{8B_y^2}{\alpha^2\bar{\gamma}^2}\, L_T \right\rceil + 1 \right\}. \quad (8)$$

Then on $\mathcal{E}_{\mathrm{lock}}$, for every $n \in \{n_0,\dots,T\}$,

$$\hat{\theta}_{n,j} < \hat{\theta}_{n,i} \qquad \text{for all } j \in S^\star,\ i \notin S^\star. \quad (9)$$

Hence the greedy top-k set induced by $\hat{\theta}_n$ is uniquely $S^\star$ for all $n \ge n_0$.

Proof. Fix $i \notin S^\star$ and $j \in S^\star$.

Step 1 (mean signal). By Lemma 3,

$$\mathbb{E}\big[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}\big] = \alpha \sum_{s=1}^{n} \big(\theta_{t_s,i} - \theta_{t_s,j}\big) = \alpha D_n(i,j).$$

On $\mathcal{E}_{\mathrm{sep}}$ and for all $n \ge n_{\min}$,

$$\mathbb{E}\big[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}\big] \ge \alpha(n\bar{\gamma} - C). \quad (10)$$

Step 2 (concentration). On $\mathcal{E}_{\mathrm{conc}}$, for all $n \le T$,

$$Z_n(i,j) \ge \mathbb{E}\big[Z_n(i,j) \mid t_{1:n}, \theta_{t_1:t_n}\big] - 2B_y\sqrt{2nL_T}.$$

Combining with (10), on $\mathcal{E}_{\mathrm{lock}}$ and for all $n \ge n_{\min}$,

$$Z_n(i,j) \ge \alpha(n\bar{\gamma} - C) - 2B_y\sqrt{2nL_T}. \quad (11)$$

Step 3 (uniform positivity). Apply Young's inequality to $2B_y\sqrt{2nL_T} = (\sqrt{n})(2B_y\sqrt{2L_T})$ with $\varepsilon = \alpha\bar{\gamma}$:

$$2B_y\sqrt{2nL_T} \le \frac{\alpha\bar{\gamma}}{2}\, n + \frac{4B_y^2}{\alpha\bar{\gamma}} L_T. \quad (12)$$

Substituting into (11) gives

$$Z_n(i,j) \ge \frac{\alpha\bar{\gamma}}{2}\, n - \alpha C - \frac{4B_y^2}{\alpha\bar{\gamma}} L_T. \quad (13)$$

If $n \ge n_0$, then by (8),

$$\frac{\alpha\bar{\gamma}}{2}(n_0 - 1) \ge 2\alpha C + \frac{4B_y^2}{\alpha\bar{\gamma}} L_T,$$

so for all $n \ge n_0$,

$$\frac{\alpha\bar{\gamma}}{2} n > 2\alpha C + \frac{4B_y^2}{\alpha\bar{\gamma}} L_T.$$

Plugging this into (13) yields

$$Z_n(i,j) \ge \alpha C + \frac{\alpha\bar{\gamma}}{2} > 0. \quad (14)$$

Step 4 (strict boundary ordering). By (5) and (14),

$$\hat{\theta}_{n,i} - \hat{\theta}_{n,j} = \frac{1}{\lambda + n\alpha} Z_n(i,j) > 0,$$

so $\hat{\theta}_{n,j} < \hat{\theta}_{n,i}$. Since (6) is uniform over pairs, this holds for all boundary pairs simultaneously, proving (9). The final claim follows from Lemma 6. □

B. How long until n_0 exploration samples are collected?

Define

$$N_t := \sum_{s=1}^{t} \mathbf{1}\{s \text{ is an exploration round}\}, \qquad \tau := \min\{t \in [T] : N_t \ge n_0\}$$

with the convention $\tau = T$ if $N_T < n_0$. Since N_t increases only on exploration rounds, τ is itself an exploration round.

Lemma 7 (Expected number of exploration rounds)**.** *For $\varepsilon_t = \min\{1, c/\sqrt{t}\}$,*

$$\mathbb{E}[N_T] = \sum_{t=1}^{T} \varepsilon_t \le 1 + 2c\sqrt{T}.$$

Proof. By integral comparison,

$$\sum_{t=1}^{T} t^{-1/2} \le 1 + \int_1^T x^{-1/2}\, dx = 2\sqrt{T} - 1.$$

□

Lemma 8 (High-probability upper bound on τ)**.** *Let n_0 be as in* (8) *and define*

$$m_0 := \left\lceil 2n_0 + \frac{n_0^2}{c^2} \right\rceil. \tag{15}$$

Then

$$\mathbb{P}(\tau > m_0) \le \exp(-n_0/4).$$

Proof. Let $\mu_m := \mathbb{E}[N_m] = \sum_{t=1}^{m} \varepsilon_t$. Since for all $t \ge 1$,

$$\varepsilon_t = \min\left\{1, \frac{c}{\sqrt{t}}\right\} \ge \frac{c}{\sqrt{t+c^2}},$$

we get

$$\mu_m \ge c\sum_{t=1}^{m} \frac{1}{\sqrt{t+c^2}} \ge c\int_0^m \frac{dx}{\sqrt{x+c^2}} = 2c(\sqrt{m+c^2} - c).$$

For $m = m_0$, we have $m_0 + c^2 \ge (c + n_0/c)^2$, hence $\mu_{m_0} \ge 2n_0$.

The exploration indicators are independent Bernoulli variables with means ε_t, so N_{m_0} is Poisson–binomial. A multiplicative Chernoff bound gives

$$\mathbb{P}\Big(N_{m_0} \le \frac{\mu_{m_0}}{2}\Big) \le \exp\Big(-\frac{\mu_{m_0}}{8}\Big) \le \exp(-n_0/4).$$

Since $\mu_{m_0}/2 \ge n_0$,

$$\mathbb{P}(\tau > m_0) = \mathbb{P}(N_{m_0} < n_0) \le \exp(-n_0/4).$$

□

C. Main regret bound

Define

$$\Delta_{\max} := \sup_{t\in[T]} \sup_{a\in\mathcal{A}} \langle \theta_t,\ a - a^\star \rangle. \tag{16}$$

Under Assumption 2, $\Delta_{\max} \le 2kB_\theta$.

Theorem 2 (Regret of RR-decay)**.** *Assume Assumptions 1, 2, 3, and 4. Run RR-decay with $\varepsilon_t = \min\{1, c/\sqrt{t}\}$. Let n_0 and m_0 be defined by* (8) *and* (15)*. Then*

$$\mathbb{E}[R_T] \le (1 + 2c\sqrt{T})\,\Delta_{\max} + m_0\,\Delta_{\max} + \Big(\delta + \delta_{\text{sep}} + \exp(-n_0/4)\Big)\,T\,\Delta_{\max}.$$

Moreover, if $\delta = T^{-2}$ and $\delta_{\text{sep}} \le T^{-2}$, then $L_T = O(\log(dT))$ and the failure term is $o(1)\cdot T\Delta_{\max}$. In particular, for fixed $(d, k, B_\theta, B_\eta, \bar{\gamma}, C, \lambda, c)$,

$$\mathbb{E}[R_T] = \widetilde{O}\big(\text{poly}(d,k)\sqrt{T}\big).$$

Proof. Split regret into exploration rounds, exploitation rounds before lock-in, and failure events.

Exploration rounds. Each exploration round contributes at most $\Delta_{\max}$, so by Lemma 7,

$$\mathbb{E}\left[\sum_{t:\,\text{explore}} \langle \theta_t, a_t - a^\star\rangle\right] \le \Delta_{\max}\mathbb{E}[N_T] \le (1 + 2c\sqrt{T})\Delta_{\max}.$$

Exploitation after lock-in. On $\mathcal{E}_{\text{lock}}$, Theorem 1 implies that once $N_t \ge n_0$, every exploitation action equals $a^\star$. Since τ is an exploration round, all exploitation rounds with $t \ge \tau+1$ have zero regret on $\mathcal{E}_{\text{lock}}$.

Exploitation before lock-in. Before τ, each exploitation round contributes at most $\Delta_{\max}$:

$$\sum_{t<\tau:\,\text{exploit}} \langle \theta_t, a_t - a^\star\rangle \le (\tau - 1)\Delta_{\max}.$$

Using Lemma 8,

$$\mathbb{E}[\tau] \le m_0 + T\mathbb{P}(\tau > m_0) \le m_0 + T\exp(-n_0/4),$$

so expected pre-lock-in exploitation regret is at most

$$(m_0 + T\exp(-n_0/4))\Delta_{\max}.$$

Failure events. If either $\mathcal{E}_{\text{conc}}$ or $\mathcal{E}_{\text{sep}}$ fails, we use the trivial bound $R_T \le T\Delta_{\max}$. This contributes at most

$$(\delta + \delta_{\text{sep}})T\Delta_{\max}.$$

Adding the terms proves the claim. □

VI. Simulation Results

We evaluate whether RR-decay matches the main qualitative predictions of the analysis: (i) $\widetilde{O}(\sqrt{T})$-scale regret under $\varepsilon_t = c/\sqrt{t}$, (ii) horizon lock-in under prolonged near-tie bursts, and (iii) robustness to bounded observation noise. Unless noted otherwise, all results are means over 200 seeds with 95% confidence intervals.

A. Setup and metrics

We fix $(d, k) = (12, 4)$ and horizons

$$T \in \{200, 500, 1000, 2000, 5000, 10^4, 2\cdot 10^4, 5\cdot 10^4\}.$$

RR-decay uses $\varepsilon_t = \min\{1, c/\sqrt{t}\}$ and the covariance-aware ridge estimator $\hat{\theta}_n = (\lambda I + nG)^{-1}b_n$, updated only on exploration rounds, with $\lambda = 10^{-2}$ and deterministic top-k tie-breaking (smaller index first).

a) Non-stationary environment and near-tie bursts.: For each seed, we sample a fixed optimal set $S^\star$ of size k. Coordinates in $S^\star$ drift via a reflected random walk in $[0.02, 0.035]$, while coordinates outside $S^\star$ drift in $[0.195, 0.23]$ with per-round drift scale 0.0025. Hence, outside burst segments, the boundary gap satisfies $\gamma_t \ge 0.16$. In the near-tie burst regime, we start a burst of length 500 with probability 5×10^{-5} per round; during a burst, all outside coordinates are pinned to 0.035 while inside coordinates continue drifting in $[0.02, 0.035]$. This creates long contiguous near-tie segments while preserving the stable-optimum benchmark.

b) Noise and metrics.: In the noisy regime, feedback is $y_t = \langle a_t, \theta_t\rangle + \eta_t$ with $\eta_t \sim \text{Unif}[-0.02, 0.02]$; regret is always computed using the true θ_t. We report cumulative regret R_T, normalized regret $R_T/\sqrt{T}$, the number of exploration rounds N_T, and the horizon lock-in time τ (first round after which the greedy top-k set stays equal to $S^\star$ through horizon T, or $\tau = T + 1$ if no lock-in occurs). We also report the wrong-on-exploit rate and the exploration-time proxy

$$\phi_n := \min_{i\notin S^\star,\ j\in S^\star} \frac{1}{n}\sum_{s=1}^{n} \big(\theta_{t_s,i} - \theta_{t_s,j}\big),$$

TABLE I
NEAR-TIE BURSTS AT $T = 50{,}000$, $c = 5$ (PART I: NORMALIZED REGRET AND EXPLORATION COUNT).

Regime	$R_T/\sqrt{T}$	N_T
Bursts	6.12 ± 0.24	2207.5 ± 6.2
Bursts + noise	6.02 ± 0.19	2207.2 ± 6.1

TABLE II
NEAR-TIE BURSTS AT $T = 50{,}000$, $c = 5$ (PART II: LOCK-IN AND WRONG-ON-EXPLOIT).

Regime	τ	lock	wrong expl.
Bursts	2563 ± 379	$200/200$	$3.26\% \pm 0.61\%$
Bursts + noise	2542 ± 333	$200/200$	$3.01\% \pm 0.46\%$

TABLE III
SENSITIVITY TO c AT $T = 50{,}000$ (PART I: NORMALIZED REGRET, EXPLORATION COUNT, AND LOCK COUNT).

Regime	c	$R_T/\sqrt{T}$	N_T	lock
Bursts	3	5.93 ± 0.53	1331.7 ± 4.7	$200/200$
Bursts	5	6.12 ± 0.24	2207.5 ± 6.2	$200/200$
Bursts + noise	3	6.34 ± 0.62	1331.4 ± 4.7	$200/200$
Bursts + noise	5	6.02 ± 0.19	2207.2 ± 6.1	$200/200$

where $t_1 < t_2 < \cdots$ are the exploration times. Because N_T varies across seeds and across c, we report ϕ_{1200}, a fixed exploration index at which all runs below have data from all 200 seeds.

B. Main performance (c=5) under near-tie bursts

Tables I and II summarize performance at $T = 50{,}000$ for $c = 5$. In both noiseless and noisy burst regimes, RR-decay locks in on all seeds ($200/200$), and $R_T/\sqrt{T}$ remains near a constant of order 6. The exploration count is consistent with the $\Theta(c\sqrt{T})$ schedule: at $T = 50{,}000$, $2c\sqrt{T} \approx 2236$ and the observed mean is about 2207.

The burst mechanism does create sustained near-ties: pooling over seeds at $T = 50{,}000$, about 2.53% of rounds satisfy $\gamma_t \leq 10^{-2}$, the longest such run is 500, and the minimum observed gap is 8.7×10^{-9}. Figure 1 shows normalized regret versus horizon in the burst regime.

C. Sensitivity to the exploration constant c and identifiability proxy

Tables III and IV compare $c \in \{3, 5\}$ at $T = 50{,}000$ in the near-tie burst regimes. Both values achieve full horizon lock-in ($200/200$), but $c = 5$ locks earlier (smaller τ) and reduces wrong-on-exploit at the cost of more exploration. Thus, $c = 5$ is a conservative default for difficult near-tie instances, while $c = 3$ is a lighter-exploration alternative that still locks in by this horizon.

We also report the exploration-time proxy ϕ_{1200}. Its value is positive and well away from zero in all settings, which is consistent with the identifiability condition used in the lock-in proof. As expected, ϕ_{1200} is not exactly identical across c (because exploration times differ), but the differences are small here.

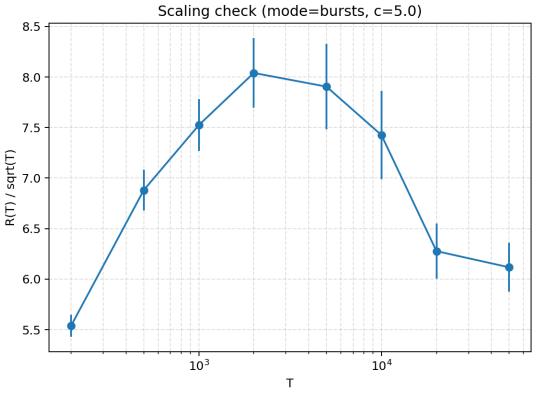

Fig. 1. Near-tie bursts ($c = 5$): normalized regret versus horizon T (mean with 95% CI over 200 seeds).

TABLE IV
SENSITIVITY TO c AT $T = 50{,}000$ (PART II: LOCK-IN TIME, WRONG-ON-EXPLOIT, AND ϕ_{1200}).

Regime	c	τ	wrong expl.	ϕ_{1200}
Bursts	3	5604 ± 875	$7.00\% \pm 1.29\%$	0.17952 ± 0.00051
Bursts	5	2563 ± 379	$3.26\% \pm 0.61\%$	0.17929 ± 0.00079
Bursts + noise	3	5959 ± 951	$7.91\% \pm 1.53\%$	0.17947 ± 0.00051
Bursts + noise	5	2542 ± 333	$3.01\% \pm 0.46\%$	0.17932 ± 0.00078

VII. CONCLUSION

We studied a non-stationary linear bandit with top-k subset actions under a *stable-optimum* benchmark: a fixed subset $S^\star$ minimizes $\langle \theta_t, \mathbf{1}_S \rangle$ over all $|S| = k$ for every round t. We analyzed RR-decay, which mixes decaying uniform top-k exploration with greedy exploitation based on a covariance-aware ridge estimator computed from exploration samples. Using the closed-form structure of the uniform exploration covariance, we reduce learning to recovering (strict) boundary orderings once enough exploration evidence accumulates, and we prove a lock-in phenomenon under a prefix-separation (identifiability) condition allowing long tie segments through an explicit slack term. As a consequence, the expected regret satisfies a $\widetilde{O}(\mathrm{poly}(d, k)\sqrt{T})$ bound under boundedness, non-anticipation, stable-optimum, and the stated identifiability condition.

REFERENCES

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2–3, pp. 235–256, 2002.

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

[3] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[4] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008, pp. 355–366.

[5] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems 24 (NeurIPS)*, 2011.

[6] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Advances in Neural Information Processing Systems 27 (NeurIPS)*, 2014.
[7] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1404–1422, 2012.
[8] C.-Y. Wei and H. Luo, "Non-stationary Reinforcement Learning without Prior Knowledge: An Optimal Black-box Approach," *arXiv preprint arXiv:2102.05406*, 2021.
[9] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
[10] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.
[11] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvári, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 38, 2015, pp. 535–543.
[12] I. Rejwan and Y. Mansour, "Top-k combinatorial bandits with full-bandit feedback," in *Proceedings of Algorithmic Learning Theory (ALT)*, 2020.
[13] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Learning to optimize under non-stationarity," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, 2019.
[14] Y. Russac, O. Cappé, and C. Vernade, "Weighted linear bandits for non-stationary environments," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

APPENDIX A
NOTATION SUMMARY

We summarize the main notation used in the paper.

- d: ambient dimension (number of coordinates/items).
- k: subset size.
- $\mathcal{A} = \{a \in \{0,1\}^d : \sum_i a_i = k\}$: top-$k$ action class.
- $S^\star$: fixed stable-optimal subset, with $|S^\star| = k$.
- $a^\star = \mathbf{1}_{S^\star}$: comparator action.
- $\theta_t \in \mathbb{R}^d$: non-stationary loss vector at round t.
- $y_t = \langle \theta_t, a_t \rangle + \eta_t$: scalar feedback.
- q: uniform distribution over $\mathcal{A}$ (used for exploration).
- $G = \mathbb{E}_{a\sim q}[aa^\top]$: exploration covariance.
- N_t: number of exploration rounds up to time t.
- $t_1, t_2, \ldots$: exploration times.
- $b_n = \sum_{s=1}^{n} a_{t_s} y_{t_s}$: exploration sufficient statistic.
- $\hat{\theta}_n = (\lambda I + nG)^{-1} b_n$: covariance-aware ridge estimator.
- γ_t: instantaneous boundary gap.
- ϕ_n: exploration-time prefix-separation proxy used in simulations.