Linear Ensemble Sampling with Smaller Ensembles

Taehyun Hwang
Seoul National University
th.hwang@snu.ac.kr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

Ensemble sampling offers a practical approach to randomized exploration by maintaining a collection of models, but how small an ensemble can be while retaining strong regret guarantees remains unresolved. In particular, the existing guarantees use an ensemble size of $\Theta(d \log T)$, leaving a logarithmic gap in the horizon T relative to the intrinsic $\Omega(d)$ ensemble-size barrier. We aim to narrow this gap by proposing an ensemble sampling algorithm that refreshes the ensemble only when the regularized Gram matrix changes substantially. This mechanism localizes the perturbation analysis to epochs with controlled Gram-matrix drift and reduces the sufficient ensemble size to $\Theta(d \log d + d \log \log T)$, while preserving the state-of-the-art $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret for ensemble sampling with arbitrary bounded arm sets. We further show that, when the arm set is finite of cardinality K, the proposed algorithm achieves the sharper regret bound $\widetilde{\mathcal{O}}(d\sqrt{T \log K})$. To the best of our knowledge, this is the first ensemble-sampling guarantee that simultaneously recovers both canonical regret scalings known for randomized linear bandit algorithms: the $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ rate for arbitrary bounded arm sets and the $\widetilde{\mathcal{O}}(d\sqrt{T \log K})$ rate for finite arm sets. The algorithm also admits an anytime implementation without resetting past data, and experiments show that it remains competitive with baselines while using substantially smaller ensembles.

1 Introduction

Randomized exploration is a key principle in sequential decision making. In linear contextual bandits, algorithms such as Thompson sampling (TS) and perturbed-history exploration (PHE) translate uncertainty about the unknown parameter into randomized actions, leading to strong regret guarantees and strong empirical performance [23, 4, 2, 12–14]. However, exact posterior sampling can be difficult to implement beyond conjugate models in TS, and repeatedly applying perturbations to all past observations at every round can be computationally costly in PHE.

To address these challenges, *ensemble sampling* (ES) offers an attractive alternative [16]. Instead of sampling from an exact posterior distribution or resampling perturbations for all past observations, ES maintains a finite collection of perturbed models and acts greedily with respect to one randomly selected ensemble member. This makes ES incremental and compatible with how ensembles are used in practice, including in exploration for deep reinforcement learning and online recommendation systems [18, 17, 19, 20, 26, 25]. These practical advantages have motivated a growing theoretical literature [17, 21, 9, 15, 22, 10] studying whether ES can match the statistical guarantees of canonical randomized exploration methods such as TS. Despite recent progress, two open questions remain.

The first open question concerns the *ensemble complexity*. Reducing ensemble complexity while maintaining strong regret guarantees has been a central theme in the ES literature. As summarized in Table 1, the sufficient ensemble size has decreased from $\Theta(dT)$ in early work [21] to $\Theta(d \log T)$ in the sharpest existing analysis of linear ES [10], with the latter achieving the canonical regret bound $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. Despite this progress, the multiplicative $\log T$ factor can still make the sufficient

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of regret bounds and ensemble complexities for linear ensemble sampling. Here d is the feature dimension, K is the number of arms, T is the horizon, and δ is the failure probability in high-probability regret guarantees. The dependence on δ is omitted for results stated in Bayesian regret, where no high-probability failure parameter is used. Logarithmic factors other than those displayed are suppressed in the regret bounds.

Paper	Ensemble size	Regret	Setting
Lu and Van Roy [16]	$\sqrt{KT}$	N/A	Finite fixed arms
Qin et al. [21]	dT	$\sqrt{dT}$ †	Finite fixed arms
Janz et al. [9]	$d\log T + d\log\frac{1}{\delta}$	$d^{5/2}\sqrt{T}$	Arbitrary bounded arms
Lee and Oh [15]	$K\log T + \log\frac{1}{\delta}$	$d^{3/2}\sqrt{T}$	Finite fixed arms
Sun et al. [22]	$K\log T + \log\frac{1}{\delta}$	$d^{3/2}\sqrt{T} + d^{9/2}$	Finite fixed arms, GLM
Janz et al. [10]	$d\log T + d\log\frac{1}{\delta}$	$d^{3/2}\sqrt{T}$	Arbitrary bounded arms
This work (Theorem 1)	$d\log d + d\log\log T + \log\frac{1}{\delta}$	$d^{3/2}\sqrt{T}$	Arbitrary bounded arms
This work (Theorem 2)	$d\log d + d\log\log T + \log\frac{1}{\delta}$	$d\sqrt{T\log K}$	Finite arms
TS [4, 2] / PHE [13, 15]‡	–	$d^{3/2}\sqrt{T}$	Arbitrary bounded arms
	–	$d\sqrt{T\log K}$	Finite arms

† The regret bound in Qin et al. [21] is Bayesian, whereas all the other regret bounds are frequentist.

‡ For linear PHE, Kveton et al. [13] established a $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ regret bound for finite arms, while Lee and Oh [15] later proved a $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret bound for arbitrary bounded arm sets.

ensemble size substantially larger than the dimension, particularly in linear bandit instances where the horizon T is much larger than d. On the other hand, an $\Omega(d)$ lower bound on the ensemble size has also been established [10], implying that sublinear-in-d ensembles cannot suffice in general. This leaves a natural question: *Can ES approach the intrinsic $\Omega(d)$ ensemble-size barrier while preserving sharp regret?*

The second open question concerns the regret landscape of ES in the finite-arm setting. For randomized linear bandit algorithms such as TS and PHE, the regret benchmark is not a single regret rate. For arbitrary bounded arm sets, including infinite ones, the *dimension-dominated* rate is $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. When the arm set is finite with cardinality K and $K \lesssim e^d$, which we refer to as the *finite-arm-dominated* regime, a regret guarantee of $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ is available for both TS and PHE.[1] This $\log K$-dependent finite-arm rate is meaningful for two reasons: (i) in finite-arm-dominated regimes, it is sharper than $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$; and (ii) the analysis only needs to control the perturbed reward estimates on the K candidate arms, rather than uniformly over all directions in $\mathbb{R}^d$.

However, existing ES analyses do not fully recover the bounds in these two regimes. They obtain the dimension-dominated regret bound of $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$, but they do not yield the finite-arm-dominated regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$. The main analytical obstacle stems from the persistent nature of ensemble perturbations. In TS and PHE, the random perturbation is independently resampled at each round, or can be analyzed as conditionally independent randomness given the past. Hence, for each fixed arm, the perturbed score has a clean conditional tail bound, and a union bound over the K arms yields the desired $\sqrt{\log K}$ dependence. In ES, by contrast, perturbations are persistent: each ensemble member accumulates perturbations over time, and the actions used to generate those perturbations are themselves selected using earlier ensemble values. Consequently, the perturbations and the subsequent action sequence are adaptively coupled, precluding the standard conditional tail-bound argument.

In this work, we address both open problems through an epoch-refresh ensemble sampling algorithm. The proposed algorithm refreshes the ensemble only when the regularized Gram matrix changes by a prescribed multiplicative factor. This localizes the perturbation analysis to epochs over which the regularized Gram matrices remain uniformly comparable. As a result, in the dimension-

[1] The relevant comparison of regret bounds depends on the relation between the arm cardinality K and the dimension d. We refer to the regime $K \lesssim e^d$ as the *finite-arm-dominated regime*, since in this regime the finite-arm rate is sharper than the dimension-dominated rate. We refer to the complementary regime $K \gtrsim e^d$ as the *dimension-dominated regime*; infinite arm sets also fall under the dimension-dominated regimes. Note that this comparison corresponds to whether K is smaller or larger than e^d, up to constants and logarithmic factors.

dominated regime, which includes infinite arm sets, we recover the $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret bound with only $\Theta(d\log d + d\log\log T)$ ensemble members, narrowing the gap between the previously known sufficient ensemble size $\Theta(d\log T)$ [9, 10] and the known $\Omega(d)$ lower bound whenever $d < T/\log T$, a common parameter regime in linear bandit problems. Moreover, in the finite-arm-dominated regime, the same localized analysis enables an arm-wise concentration argument and yields the sharper regret bound $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ with the same reduced ensemble complexity. Our main contributions are summarized as follows:

- We propose an ES algorithm for stochastic linear contextual bandits. For arbitrary bounded arm sets, including infinite ones, we prove that the proposed algorithm attains the sharpest known regret order for ensemble sampling [10, 15], $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$, where d is the feature dimension and T is the horizon. While matching the previously sharpest regret bound, we prove the ensemble complexity of $\Theta(d\log d + d\log\log T)$. This narrows the gap between the previously known sufficient ensemble size, $\Theta(d\log T)$ [9, 10], and the lower bound of $\Omega(d)$ on the ensemble complexity whenever $d < T/\log T$.
- We further show that, when the arm sets are chosen by an oblivious adversary and each has cardinality K, the proposed algorithm achieves a regret bound of order $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ with the same reduced ensemble complexity $\Theta(d\log d + d\log\log T)$. To the best of our knowledge, this is the first $\log K$-dependent regret bound for ES that scales linearly in d. For instances with $K \lesssim e^d$, i.e., the finite-arm-dominated regime, $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ can be significantly sharper than $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$, depending on the number of arms.
- Importantly, the same ES algorithm simultaneously recovers the two canonical regret bounds known for randomized linear bandit algorithms such as TS and PHE: the dimension-dominated bound $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ and the finite-arm-dominated bound $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$.
- We empirically show that the proposed algorithm achieves regret comparable to existing randomized baselines while using substantially fewer ensemble members. This reduction in ensemble size also leads to lower total runtime, with the improvement becoming more pronounced as the feature dimension increases. In addition, to the best of our knowledge, our results provide the first empirical evidence that ensemble methods can offer a computationally efficient alternative to TS and PHE.
- Our algorithm admits an anytime implementation without resetting past data. When the horizon T is unknown, prior ensemble sampling methods typically rely on the doubling trick, increasing the ensemble size at each doubling epoch and restarting the algorithm. In contrast, our epoch-refresh design allows the ensemble size to be increased online only as needed, while retaining all previously collected observations. This makes the procedure substantially more data-efficient in practical scenarios.

1.1 Related Work

Randomized exploration in linear bandits. Linear bandits have been studied extensively through both optimism-based methods, such as `OFUL` [1] and `LinUCB` [6], and randomized exploration methods. Among the latter, TS [23, 4, 2, 7] is a classical posterior-sampling method that achieves the canonical regret rate $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ for general arm sets, and sharper $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ bounds are available when $|\mathcal{X}| = K$. However, TS requires sampling from, or accurately approximating, the posterior distribution, which can be computationally demanding beyond simple conjugate models. Another representative randomized approach is PHE [12–14], which avoids explicit posterior sampling by injecting random perturbations into the observed history. PHE can also attain the finite-arm rate $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$. These two regret scalings serve as the benchmark for randomized linear bandit algorithms and motivate our goal of recovering them for ensemble sampling.

Beyond these standard guarantees, sharper regret bounds of order $\widetilde{\mathcal{O}}(d\sqrt{T})$ have also been established for randomized linear-bandit algorithms under additional structural or computational assumptions. For example, Abeille et al. [3] obtain an $\mathcal{O}(d\sqrt{T}\log T)$ regret bound for Gaussian LinTS when the fixed action space is smooth and strongly convex. Minimax-rate guarantees have also been obtained through modified posterior sampling, such as Feel-Good Thompson Sampling [24], and through approximate posterior sampling based on Langevin Monte Carlo under additional regularity

conditions [8]. Related randomized approaches based on information-directed sampling have also been studied [11].

Ensemble sampling. ES [16] was introduced as a computationally attractive alternative to posterior sampling. Instead of maintaining an exact posterior distribution, ES keeps a finite collection of perturbed models and acts greedily with respect to one randomly selected ensemble member. This structure is appealing in settings where incremental model updates are cheap, and it has motivated algorithms in deep reinforcement learning, online recommendation, and other sequential decision-making problems [18, 17, 19, 20, 26, 25]. However, despite its empirical success, ES has been difficult to analyze theoretically because the ensemble models are trained on data collected adaptively by the ensemble itself. In contrast to TS or PHE, the randomness used for exploration is not freshly sampled independently of the past at every round; rather, the perturbations persist and interact with the adaptively chosen arms.

Early theoretical guarantees for ES required ensemble sizes that were too large for the method to be practically meaningful. Qin et al. [21] obtained Bayesian regret guarantees under ensemble sizes that scale at least linearly with the horizon. Janz et al. [9] gave the first rigorous frequentist analysis of ES in stochastic linear bandits, showing that a symmetrized version of linear ES with $\Theta(d \log T)$ ensemble members achieves $\widetilde{\mathcal{O}}(d^{5/2}\sqrt{T})$ regret and, importantly, also applies to infinite arm sets. This result established that sublinear-in-T ensembles can suffice, but the regret remained worse than the canonical TS rate by a factor of roughly d. At the same time, Lee and Oh [15] improved the regret analysis of linear ES and showed that ES can achieve the canonical regret rate $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ with ensemble size logarithmic in T. However, their ES guarantee requires the ensemble size to scale with the number of arms, namely $\Omega(K \log T)$. Thus, while their result is sharp for finite and moderately sized arm sets, it becomes unsuitable for large finite arm sets and does not directly address general infinite arm sets. Janz et al. [10] recently provided a sharp analysis of linear ES for general arm sets. They showed that Gaussian linear ES with $\Theta(d \log T)$ ensemble members achieves the regret rate $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$, matching the canonical rate of TS for general linear bandits. Their analysis reduces the main difficulty to a time-uniform self-normalized exceedance problem and solves it through a Brownian-motion embedding argument. They also established a lower-bound result showing that ensembles of size at most $d/2$ can suffer linear regret on some instances, indicating that the ensemble size cannot be pushed far below the dimension in general.

2 Preliminaries

2.1 Notation

For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \mathbf{y}$. For a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, define $\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}$, and let $\|\mathbf{A}\|_{\mathrm{op}}$ denote its operator norm. We use $\mathbb{1}\{\cdot\}$ for the indicator function. The notation $\mathcal{O}(\cdot)$ denotes the standard big-$\mathcal{O}$ notation with respect to the problem parameters, such as d and T, and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in these parameters. We denote the unit Euclidean ball in $\mathbb{R}^d$ by $\mathcal{B}_2^d := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le 1\}$.

2.2 Problem setup

We consider a stochastic linear contextual bandit problem. At each round $t \in [T]$, the agent observes a compact arm set $\mathcal{X}_t \subseteq \mathbb{R}^d$, selects an arm $\mathbf{x}_t \in \mathcal{X}_t$, and observes a real-valued reward $y_t \in \mathbb{R}$. The reward is generated according to $y_t = \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle + \eta_t$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is an unknown parameter and η_t is noise. Let $\mathcal{H}_t := \sigma(\mathcal{X}_1, \mathbf{x}_1, y_1, \dots, \mathcal{X}_{t-1}, \mathbf{x}_{t-1}, y_{t-1}, \mathcal{X}_t, \mathbf{x}_t)$ be the σ-algebra generated by the past observations, the current arm set, and the selected arm at round t. We assume that η_t is conditionally zero-mean and 1-sub-Gaussian with respect to $\mathcal{H}_t$, that is, for all $s \in \mathbb{R}$, $\mathbb{E}[\eta_t \mid \mathcal{H}_t] = 0$, $\mathbb{E}[\exp(s\eta_t) \mid \mathcal{H}_t] \le \exp(s^2/2)$. The agent's objective is to maximize the cumulative reward over T rounds, or equivalently, to minimize the cumulative regret

$$\mathrm{Regret}_T := \sum_{t=1}^{T} \left(\langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle \right) ,$$

where $\mathbf{x}_t^* \in \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle$ is an optimal arm at round t.

3 Main Results

3.1 Algorithm: LinES-ER

Algorithm 1 Linear Ensemble Sampling with Epoch Refresh (LinES-ER)

Inputs: horizon $T \in \mathbb{N}$, ensemble size $m \in \mathbb{N}$, regularization parameter $\lambda > 0$, inflation parameter $\gamma > 0$, confidence radii $\{\beta_t\}_{t\geq 0}$, and epoch refresh parameter $\alpha > 0$
Initialize $\mathbf{V}_0 = \lambda \mathbf{I}_d$, $\hat{\boldsymbol{\theta}}_0 = \mathbf{0}_d$, and set the current epoch start time $\tau = 0$
Sample $\mathbf{g}^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and set $\mathbf{s}_0^j = \mathbf{V}_0^{1/2}\mathbf{g}^j$, $\boldsymbol{\theta}_0^j = \mathbf{V}_0^{-1}\mathbf{s}_0^j$ for each $j \in [m]$
for $t = 1, \ldots, T$ **do**
  if $\mathbf{V}_{t-1} \not\preceq (1+\alpha)\mathbf{V}_\tau$ **then** ▷ *Refresh the ensemble perturbations*
    Set $\tau \leftarrow t-1$
    For each $j \in [m]$, sample $\mathbf{g}^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and set

$$\mathbf{s}_{t-1}^j = \mathbf{V}_\tau^{1/2}\mathbf{g}^j\,, \quad \boldsymbol{\theta}_{t-1}^j = \hat{\boldsymbol{\theta}}_{t-1} + \gamma\beta_{t-1}\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j$$

  end if
  Observe $\mathcal{X}_t$, sample $J_t \sim \mathrm{Unif}([m])$, choose $\mathbf{x}_t \in \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}_t}\langle \mathbf{x}, \boldsymbol{\theta}_{t-1}^{J_t}\rangle$ and observe y_t
  Update $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$ and $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1}\sum_{s=1}^t \mathbf{x}_s y_s$
  For each $j \in [m]$, sample $\xi_t^j \sim \mathcal{N}(0,1)$ and update

$$\mathbf{s}_t^j = \mathbf{s}_{t-1}^j + \mathbf{x}_t\xi_t^j\,, \quad \boldsymbol{\theta}_t^j = \hat{\boldsymbol{\theta}}_t + \gamma\beta_t\mathbf{V}_t^{-1}\mathbf{s}_t^j$$

end for

We now describe our ensemble sampling algorithm. The algorithm maintains m perturbed least-squares models. At each round, it samples one ensemble member uniformly at random and acts greedily with respect to the corresponding perturbed parameter. The key difference from prior ensemble sampling methods [9, 15, 22, 10] is that the ensemble is refreshed only when the regularized Gram matrix changes substantially.

Let $\alpha > 0$ be the epoch-refresh parameter. Each epoch starts at some round τ. During the epoch, the algorithm monitors the regularized Gram matrix and starts a new epoch whenever

$$\mathbf{V}_{t-1} \not\preceq (1+\alpha)\mathbf{V}_\tau.$$

Thus, throughout an epoch, the regularized Gram matrix remains comparable to its value at the epoch start. This comparability is the key property used in our analysis.

At the beginning of an epoch, each ensemble perturbation is initialized according to the current Gram matrix. Specifically, for each $j \in [m]$, the algorithm samples $\mathbf{g}^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and sets $\mathbf{s}_\tau^j = \mathbf{V}_\tau^{1/2}\mathbf{g}^j$ and $\boldsymbol{\theta}_\tau^j = \hat{\boldsymbol{\theta}}_\tau + \gamma\beta_\tau\mathbf{V}_\tau^{-1}\mathbf{s}_\tau^j$. At round t, the algorithm observes $\mathcal{X}_t$, samples an ensemble index $J_t \sim \mathrm{Unif}([m])$, and chooses

$$\mathbf{x}_t \in \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}_t}\langle \mathbf{x}, \boldsymbol{\theta}_{t-1}^{J_t}\rangle.$$

After observing reward y_t, the algorithm updates $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$ and $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1}\sum_{s=1}^t \mathbf{x}_s y_s$. Then, for each $j \in [m]$, it samples $\xi_t^j \sim \mathcal{N}(0,1)$ and updates $\mathbf{s}_t^j = \mathbf{s}_{t-1}^j + \mathbf{x}_t\xi_t^j$ and $\boldsymbol{\theta}_t^j = \hat{\boldsymbol{\theta}}_t + \gamma\beta_t\mathbf{V}_t^{-1}\mathbf{s}_t^j$.

The epoch-refresh mechanism has two roles. First, it repeatedly adjusts the perturbation distribution to match the current Gram matrix. Second, it restricts the uniform control of the ensemble to a sequence of epochs within which the regularized Gram matrices remain comparable. This localization is what allows us to replace the previous $\Theta(d\log(T/\delta))$ ensemble-size requirement for ensemble sampling [9, 10] by $\Theta(d\log d + d\log\log T + \log(1/\delta))$.

3.2 Regret Bound in the Dimension-Dominated Regime

We present our main theoretical guarantees. The first result considers arbitrary bounded arm sets, including infinite ones. It shows that the proposed algorithm matches the best known regret rate for ensemble sampling while using a smaller ensemble.

Assumption 1 (Boundedness)**.** *For every $t \in [T]$, the arm set $\mathcal{X}_t$ is closed and satisfies $\|\mathbf{x}\|_2 \le 1$ for all $\mathbf{x} \in \mathcal{X}_t$. We also assume that there exists $S > 0$ such that $\|\boldsymbol{\theta}^*\|_2 \le S$.*

Theorem 1 (Regret bound for `LinES-ER`)**.** *Suppose Assumption 1 holds and $T \ge \max\{2, d\}$. Fix any $\delta \in (0, 1/4)$, and choose the parameters in Algorithm 1 as $\lambda \ge 80$, $\gamma = 40$, $\alpha = 1$,*

$$\beta_t = \sqrt{d \log\left(1 + \frac{t}{d\lambda}\right) + 2 \log \frac{1}{\delta}} + \sqrt{\lambda}\, S, \quad \textit{and} \quad m \ge C \left(d \log\left(d + \log \frac{\bar{R}_\alpha}{\delta} \right) + \log \frac{\bar{R}_\alpha}{\delta} \right),$$

where $C > 0$ is an absolute constant and $\bar{R}_\alpha := 1 + \left\lfloor d \log_{1+\alpha}\left(1 + \frac{T}{\lambda d}\right) \right\rfloor$. Then, with probability at least $1 - 4\delta$, Algorithm 1 incurs cumulative regret $\mathrm{Regret}_T = \widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{T}\right)$.

Remark 1 (Ensemble size)**.** *When $\alpha = 1$, it follows that $\log \bar{R}_1 = \mathcal{O}(\log d + \log\log T)$. Hence, the sufficient condition on the ensemble size in Theorem 1 gives*

$$m = \Theta\left(d \log d + d \log\log T + \log \frac{1}{\delta} \right)$$

up to universal constants and lower-order logarithmic terms in d, λ, δ.

Discussion of Theorem 1. Theorem 1 shows that Algorithm 1 achieves a regret bound of order $\widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{T}\right)$ in the dimension-dominated regime, with an ensemble size smaller than that required by prior ES analyses [9, 10]. In particular, when $\alpha = 1$, Remark 1 implies that it suffices to take $m = \mathcal{O}\left(d \log d + d \log\log T + \log \frac{1}{\delta}\right)$, up to lower-order logarithmic terms. Thus, our ES algorithm matches the sharpest existing analysis of linear ensemble sampling [10], while reducing the sufficient ensemble size from order $d \log(T/\delta)$ to roughly $d \log d + d \log\log T + \log(1/\delta)$.

This improvement is particularly meaningful in view of the known ensemble-size barrier for ES. Janz et al. [10] showed that, although an ensemble size of order $d \log T$ is sufficient to obtain $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret, there exist instances on which ES with only $m = \Theta(d)$ ensemble members suffers linear regret. Our result therefore moves the sufficient ensemble size significantly closer to this lower-bound regime, while preserving the same regret rate as the best known ES bound for general arm sets.

Remark 2 (Anytime implementation without reset)**.** *Our theoretical choice of the ensemble size depends on the horizon T. When T is unknown, we can use a doubling schedule [5] and increase the ensemble size whenever the current horizon estimate doubles. Unlike prior ES methods [9, 22, 10], this does not require restarting the algorithm or discarding past observations. Because our algorithm already refreshes ensemble perturbations according to the current Gram matrix, we can simply start a new epoch and initialize the enlarged ensemble using the current Gram matrix. The regularized Gram matrix and the ridge estimator are kept unchanged, so all past data continue to be used. Thus, our epoch-refresh mechanism yields an anytime implementation without data reset, preserving the regret guarantees up to logarithmic factors and providing a more data-efficient procedure in practice. Further details of the anytime implementation are provided in Appendix E.3.*

3.3 Regret Analysis

In this section, we present the main ingredients of the regret analysis for Theorem 1. The proof follows the standard randomized-optimism route for linear bandits [2, 13, 9, 15, 10], but the key technical difficulty is to show that a sufficiently large fraction of ensemble members is optimistic uniformly over time and over all directions [10]. Our epoch-refresh construction makes this possible with a smaller ensemble size by localizing the perturbation analysis to epochs.

3.3.1 Deterministic Upper Bound on the Number of Epochs

The number of epochs generated by Algorithm 1 is random, since epoch refreshes depend on the realized Gram matrices, which in turn are determined by the algorithm's randomization and the observed data. We first derive a deterministic upper bound on the total number of epochs. Let R_α denote the total number of epochs generated by Algorithm 1. Since each completed epoch is triggered by a multiplicative growth of the regularized Gram matrix, a determinant-growth argument gives

$$R_\alpha \le 1 + \frac{d \log\left(1 + \frac{T}{\lambda d}\right)}{\log(1+\alpha)} \le 1 + \left\lfloor \frac{d \log\left(1 + \frac{T}{\lambda d}\right)}{\log(1+\alpha)} \right\rfloor =: \bar{R}_\alpha\,. \tag{1}$$

This bound is deterministic and holds for every possible realization of the algorithm. The proof of Eq. (1) is provided in Lemma 4.

3.3.2 Self-Normalized Exceedance Frequencies Inside One Epoch

The central object in the analysis is the self-normalized exceedance frequency

$$E_{t,m}(\mathbf{u}, c) := \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{\frac{\langle \mathbf{u}, \mathbf{s}_{t-1}^{j}\rangle}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} \geq c\right\}, \qquad \mathbf{u} \in \mathbb{S}^{d-1}.$$

This quantity measures the fraction of ensemble perturbations that are large enough in direction $\mathbf{u}$ at round t. This exceedance frequency is also the central quantity in the analysis of Janz et al. [10]. Their proof controls a fixed collection of m persistent ensemble members over the entire horizon by reducing the problem to a time-uniform exceedance event for the corresponding m independent Brownian motions. At a high level, for each fixed direction $\mathbf{u}$, the projected perturbation process $\mathbf{M}_t := (\langle \mathbf{s}_{t-1}^{j}, \mathbf{u}\rangle)_{j=1}^{m}$ can be represented through Brownian motions whose time parameter is the self-normalized clock $\|\mathbf{u}\|_{\mathbf{V}_{t-1}}^2$. Since their ensemble is not refreshed, this clock must be controlled over the full range induced by the Gram-matrix growth from the beginning to the end of the horizon. After further making this control uniform over directions, their analysis requires an ensemble size of order $d \log T$.

The main advantage of epoch refreshes is that the regularized Gram matrix varies only within a controlled multiplicative range during each epoch. Indeed, if round t belongs to the r-th epoch starting at τ_r, then the refresh rule guarantees $\mathbf{V}_{\tau_r} \preceq \mathbf{V}_{t-1} \preceq (1+\alpha)\mathbf{V}_{\tau_r}$. Thus, for any fixed direction $\mathbf{u}$, the Brownian clock associated with the projected perturbation process varies only over the local interval $\left[\|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}^2, (1+\alpha)\|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}^2\right]$. Consequently, when applying the Brownian exceedance theorem (Theorem 5.4 in [10]), the relevant clock ratio is controlled by $1+\alpha$, rather than by the global Gram-matrix growth over the entire horizon.

Lemma 1 (Fixed-direction exceedance inside one epoch)**.** *Fix an epoch r and a direction $\mathbf{u} \in \mathbb{S}^{d-1}$. For $\alpha > 0$, let $\gamma := 40$, $\kappa_\alpha := \lceil 250 \log(1+\alpha)\rceil$. Then for every $\delta' \in (0,1)$, if $m \geq 40 \log \frac{\kappa_\alpha}{\delta'}$, then with probability at least $1-\delta'$, $\forall t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$, we have $E_{t,m}(\mathbf{u}, 2/\gamma) \geq \frac{1}{10}$.*

Note that the epoch-refresh construction allows us to represent the projected perturbations by Brownian motions whose clocks vary only over the local multiplicative range of that epoch. This makes it possible to invoke the Brownian exceedance theorem with a uniformly bounded clock ratio, rather than the horizon-dependent ratio arising in the global persistent-ensemble analysis [10].

3.3.3 Self-Normalized Exceedance Frequencies Across Epochs

The fixed-direction bound is not sufficient for bounded arm sets that may be infinite, because the optimistic direction can depend on both the round and the unknown optimal arm. We therefore extend the fixed-direction guarantee uniformly over all directions on the unit sphere. The first ingredient is a local Lipschitz property of the normalized map induced by the Gram-matrix change within an epoch. For an epoch r and a round t in that epoch, define

$$\mathbf{B}_{r,t} := \mathbf{V}_{t-1}^{1/2}\mathbf{V}_{\tau_r}^{-1/2}, \qquad \boldsymbol{\phi}_{r,t}(\mathbf{w}) := \frac{\mathbf{B}_{r,t}\mathbf{w}}{\|\mathbf{B}_{r,t}\mathbf{w}\|_2}, \quad \mathbf{w} \in \mathbb{S}^{d-1}.$$

Since $\mathbf{V}_{t-1} \preceq (1+\alpha)\mathbf{V}_{\tau_r}$ inside the epoch, this map is Lipschitz (Lemma 7):

$$\|\boldsymbol{\phi}_{r,t}(\mathbf{w}) - \boldsymbol{\phi}_{r,t}(\mathbf{w}')\|_2 \leq 2\sqrt{1+\alpha}\,\|\mathbf{w} - \mathbf{w}'\|_2.$$

This allows us to transfer the exceedance guarantee from a nearby net point to an arbitrary direction. This is a key point where our epoch-refresh analysis differs from prior persistent-ensemble analyses [9, 10]. In the analysis of Janz et al. [10], the corresponding normalized map must be controlled over the full sequence of Gram-matrix geometries along the entire horizon. Since the Gram matrix may grow by a factor polynomial in T, the Lipschitz constant can scale as large as order $\sqrt{T}$, which forces a much finer sphere net and contributes a $d \log T$ term to the required ensemble size. In contrast, our refresh rule ensures that, within each epoch, the Gram matrix changes by at most a factor $1+\alpha$.

Consequently, the Lipschitz constant of $\phi_{r,t}$ is bounded by $2\sqrt{1+\alpha}$, which is independent of T for constant α. This local Lipschitz control is one of the main reasons why the uniform exceedance argument can be carried out with a smaller ensemble.

The second ingredient is a uniform upper bound on the normalized ensemble perturbations. Define

$$\Gamma_\delta := \sqrt{d} + \sqrt{2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)} + \sqrt{d\log(1+\alpha) + 2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)}.$$

Then, with high probability, we have

$$\forall r \le R_\alpha,\ \forall j \in [m],\ \forall t \in \{\tau_r+1, \dots, \tau_{r+1}\}: \qquad \left\|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\right\|_2 \le \Gamma_\delta\,.$$

This bound controls the loss incurred when passing from a net point to a nearby direction. Combining the local Lipschitz property with this perturbation bound, we choose a sphere net of radius proportional to $(\sqrt{1+\alpha}\gamma\Gamma_\delta)^{-1}$. Putting these ingredients together, and taking a union bound over all epochs and all points in the sphere net, yields the following uniform exceedance guarantee:

Lemma 2 (Uniform exceedance count across epochs)**.** *Fix $\gamma := 40$, $\kappa_\alpha := \lceil 250\log(1+\alpha)\rceil$, $N_\delta := \left(6\sqrt{1+\alpha}\,\gamma\Gamma_\delta\right)^d$. If*

$$m \ge 40\log\left(\frac{2\kappa_\alpha \bar{R}_\alpha N_\delta}{\delta}\right), \tag{2}$$

then with probability at least $1-\delta$, $\min\limits_{t\in[T]}\ \inf\limits_{\mathbf{u}\in\mathbb{S}^{d-1}} E_{t,m}(\mathbf{u}, 1/\gamma) \ge \dfrac{1}{10}$.

Furthermore, the sufficient ensemble-size condition in Eq. (2) can be simplified as follows:

Corollary 1 (Sufficient ensemble size)**.** *Suppose $0 < \alpha \le 1$. There exists a universal constant $C > 0$ such that the sufficient condition on the ensemble size in Eq.* (2) *is satisfied by choosing*

$$m = \left\lceil C\left[d\log\left(d + \log\frac{\bar{R}_\alpha}{\delta}\right) + \log\frac{\bar{R}_\alpha}{\delta}\right]\right\rceil.$$

Combining these ingredients with the standard regret-analysis framework for optimistic randomized algorithms for linear bandits yields the desired regret bound. The proof is deferred to Appendix A.

3.4 Regret Bound in the Finite-Arm-Dominated Regime

Our second result considers the finite-arm-dominated regime. In this case, the regret analysis only needs arm-wise concentration over the finite set of arms presented at each round, rather than uniform concentration over all directions. This makes it possible to replace the dimension-dependent uniform control by an arm-wise concentration argument, leading to the sharper finite-arm regret scaling.

For randomized linear bandit algorithms such as TS [4, 2] and PHE [13], such a $\sqrt{\log K}$-type bound is relatively natural. In these algorithms, the randomization used for arm selection is either resampled at each round or conditionally independent of the past. Hence, conditional on the history and the current finite arm set, the perturbed reward estimate of each presented arm admits a sub-Gaussian tail bound. A union bound over the K arms then yields an arm-wise confidence scale of order $\sqrt{\log K}$.

However, this argument does not directly apply to ES. In ES, the perturbation vector of each ensemble member is accumulated over time and reused across rounds. Moreover, the selected arms are themselves chosen using past ensemble perturbations. As a result, the perturbation scores at later rounds are not conditionally independent random variables given the past. Therefore, a direct Hoeffding-type argument over arms and rounds does not yield the desired arm-wise concentration bound. Our epoch-refresh construction addresses this difficulty by reinitializing the perturbations at epoch boundaries, thereby limiting the interval over which this adaptive dependence accumulates.

Lemma 3 (Uniform arm-wise concentration)**.** *Suppose Assumption 1 holds, and suppose that the arm sets $\{\mathcal{X}_t\}_{t=1}^T$ are chosen by an oblivious adversary with $|\mathcal{X}_t| = K \ge 2$ for all $t \in [T]$. For any $\alpha > 0$ and $\delta \in (0,1)$, let $L_\alpha := \log\frac{8KmT\bar{R}_\alpha}{\delta}$. Then, with probability at least $1-\delta$, we have*

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left|\frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}}\right| \le C\left[\sqrt{L_\alpha} + \sqrt{\alpha L_\alpha} + \alpha(\sqrt{d} + \sqrt{L_\alpha}) + \alpha\sqrt{\alpha(d+L_\alpha)}\right],$$

where $C > 0$ is a universal constant. In particular, if a deterministic quantity $\Lambda_* \geq 1$ satisfies $L_\alpha \leq C_\Lambda \Lambda_*$ for some universal constant $C_\Lambda > 0$, and if $\alpha = \min\{1, \sqrt{\Lambda_*/d}\}$, then we have

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left| \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1} \mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}} \right| \leq \widetilde{C}\sqrt{\Lambda_*}\,,$$

where $\widetilde{C} > 0$ is a universal constant.

Proof sketch of Lemma 3. Fix an epoch r and an ensemble member $j \in [m]$, and let τ_r denote the starting round of the r-th epoch. For a round $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$, the quantity we need to control is the normalized perturbation score for an arm $\mathbf{x} \in \mathcal{X}_t$, defined as

$$Z_{r,t}^j(\mathbf{x}) := \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1} \mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}}\,.$$

The ensemble perturbation $\mathbf{s}_{t-1}^j$ can be written as

$$\mathbf{s}_{t-1}^j = \mathbf{V}_{\tau_r}^{1/2} \mathbf{g}_r^j + \sum_{s=\tau_r+1}^{t-1} \mathbf{x}_s \xi_s^j = \mathbf{V}_{\tau_r}^{1/2} \left(\mathbf{g}_r^j + \mathbf{n}_{r,t-1}^j \right), \quad \text{where} \quad \mathbf{n}_{r,t-1}^j := \sum_{s=\tau_r+1}^{t-1} \mathbf{V}_{\tau_r}^{-1/2} \mathbf{x}_s \xi_s^j\,.$$

Writing $\mathbf{A}_{r,t} := \mathbf{V}_{\tau_r}^{-1/2} \mathbf{V}_{t-1} \mathbf{V}_{\tau_r}^{-1/2}$ and $\mathbf{y}_r(\mathbf{x}) := \mathbf{V}_{\tau_r}^{-1/2} \mathbf{x}$, the normalized perturbation score can be decomposed as

$$Z_{r,t}^j(\mathbf{x}) = \frac{\mathbf{y}_r(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1} \mathbf{g}_r^j}{\sqrt{\mathbf{y}_r(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1} \mathbf{y}_r(\mathbf{x})}} + \frac{\mathbf{y}_r(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1} \mathbf{n}_{r,t-1}^j}{\sqrt{\mathbf{y}_r(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1} \mathbf{y}_r(\mathbf{x})}}\,.$$

Moreover, by the epoch-refresh rule, we have $\mathbf{I}_d \preceq \mathbf{A}_{r,t} \preceq (1+\alpha)\mathbf{I}_d$. Hence, when α is small, all eigenvalues of $\mathbf{A}_{r,t}$ are close to one. Consequently, $\mathbf{A}_{r,t}^{-1}$ only mildly distorts both the direction $\mathbf{y}_r(\mathbf{x})$ and the corresponding normalizing factor. More precisely, for any $\mathbf{v} \in \mathbb{R}^d$,

$$\left| \frac{\mathbf{y}_r(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1} \mathbf{v}}{\sqrt{\mathbf{y}_r(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1} \mathbf{y}_r(\mathbf{x})}} - \left\langle \frac{\mathbf{y}_r(\mathbf{x})}{\|\mathbf{y}_r(\mathbf{x})\|_2}, \mathbf{v} \right\rangle \right| \lesssim \alpha \|\mathbf{v}\|_2\,.$$

In words, the self-normalized projection under the current Gram matrix is close to the Euclidean projection after normalization by the epoch-start Gram matrix, and the discrepancy is proportional to the within-epoch Gram-matrix drift α. Applying this comparison with $\mathbf{v} = \mathbf{g}_r^j$ and $\mathbf{v} = \mathbf{n}_{r,t-1}^j$ gives

$$|Z_{r,t}^j(\mathbf{x})| \lesssim \underbrace{\left| \left\langle \frac{\mathbf{y}_r(\mathbf{x})}{\|\mathbf{y}_r(\mathbf{x})\|_2}, \mathbf{g}_r^j \right\rangle \right|}_{(i)} + \underbrace{\left| \left\langle \frac{\mathbf{y}_r(\mathbf{x})}{\|\mathbf{y}_r(\mathbf{x})\|_2}, \mathbf{n}_{r,t-1}^j \right\rangle \right|}_{(ii)} + \underbrace{\alpha \|\mathbf{g}_r^j\|_2}_{(iii)} + \underbrace{\alpha \|\mathbf{n}_{r,t-1}^j\|_2}_{(iv)}\,.$$

Thus, the desired finite-arm concentration reduces to controlling four terms: (i) the epoch-start Gaussian projection, (ii) the scalar within-epoch martingale projection, (iii) the norm of the epoch-start Gaussian vector, and (iv) the norm of the within-epoch martingale vector. The first two terms give the main stochastic fluctuations, while the last two are drift terms caused by the change of the Gram matrix within the epoch and are multiplied by the refresh parameter α.

It remains to control these four terms uniformly over all epochs, ensemble members, and arms. The nontrivial terms are (ii) and (iv), since they are accumulated within the epoch and are coupled with the adaptively selected arms. The scalar martingale projection in (ii) is controlled by an exponential-martingale argument and Ville's inequality. The vector martingale norm in (iv) is controlled by combining the same martingale tail bound with a sphere-net argument. Combining these estimates proves Lemma 3; the full proof is deferred to Appendix B. This concentration result, together with the standard optimistic-randomization regret argument, gives the following regret bound.

Theorem 2 (Regret in the finite-arm-dominated regime)**.** *Suppose Assumption 1 holds, and suppose that the arm sets $\{\mathcal{X}_t\}_{t=1}^T$ are chosen by an oblivious adversary with $|\mathcal{X}_t| = K \geq 2$ for all $t \in [T]$.*

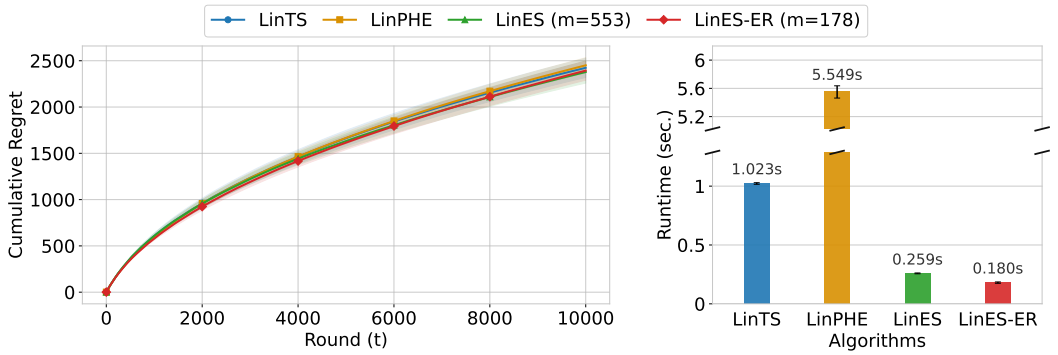


Figure 1: Cumulative regret and runtime on stochastic linear bandits over $\mathcal{B}_2^d$ with $d = 30$ and $T = 10{,}000$.

Fix $\delta \in (0, 1/4)$, and choose $\lambda \geq 80$, $\gamma = 40$, and $\{\beta_t\}_{t\geq 0}$ as in Theorem 1. For $\alpha_0 = 1/\sqrt{d}$ and $\bar{R}_0 := \bar{R}_{\alpha_0}$, define

$$M_0 = \left\lceil C_m \left[d \log\left(d + \log \frac{\bar{R}_0}{\delta} \right) + \log \frac{\bar{R}_0}{\delta} \right] \right\rceil , \quad \Lambda_0 := \max\left\{ 1,\, C_\Lambda \log\left(\frac{K M_0 T \bar{R}_0}{\delta} \right) \right\} ,$$

for some constants $C_m, C_\Lambda > 0$. Choose $\alpha := \min\{1, \sqrt{\Lambda_0/d}\}$ and $m := M_0$. Then, with probability at least $1 - 4\delta$, Algorithm 1 incurs cumulative regret $\mathrm{Regret}_T = \widetilde{\mathcal{O}}(d\sqrt{T \log K})$.

4 Numerical Experiments

We evaluate the proposed ensemble algorithm `LinES-ER` against `LinTS` [2], `LinPHE` [13], and `LinES` [10]. The experiments are conducted on stochastic linear bandits over the unit ball $\mathcal{B}_2^d$ with Gaussian noise of variance proxy 0.3. Further experimental details are provided in Appendix E.

Figure 1 reports cumulative regret and runtime for $d = 30$ and $T = 10{,}000$. For `LinES`, we use the theoretical ensemble size $m = \lceil d \log(T/\delta) \rceil$, while for `LinES-ER`, we use $m = \lceil d \log d + d \log\log T + \log(1/\delta) \rceil$, with $\delta = 1/T$. The proposed method, `LinES-ER`, achieves regret comparable to `LinES` while using only about one-third as many ensemble members (178 vs. 553). The runtime comparison shows that this reduction in ensemble size also lowers the computational cost. In particular, `LinES-ER` takes 0.180 seconds, whereas `LinES` takes 0.259 seconds; thus, `LinES-ER` reduces the total runtime by approximately $\frac{0.259 - 0.180}{0.259} \times 100 \approx 30.5\%$ relative to `LinES`.

5 Conclusion

We proposed an ensemble sampling algorithm for stochastic linear bandits that achieves the same regret order as existing ensemble sampling methods with fewer ensemble members. In particular, the proposed algorithm attains $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret for arbitrary bounded arm sets with ensemble size $\Theta(d \log d + d \log\log T)$, and further achieves the sharper $\widetilde{\mathcal{O}}(d\sqrt{T \log K})$ regret when the arm sets are chosen by an oblivious adversary and each has cardinality K. The key technical idea is to refresh the ensemble only when the regularized Gram matrix changes substantially, thereby localizing the perturbation analysis to epochs with controlled covariance drift. The algorithm also admits an anytime implementation without resetting past data, making it more data-efficient in practical scenarios. A promising direction for future work is to characterize the statistical and computational behavior of ensemble sampling in more complex nonlinear reward settings.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023- 00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008).

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[2] Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, PMLR, 20–22 Apr 2017.

[3] Marc Abeille, David Janz, and Ciara Pike-Burke. When and why randomised exploration works (in linear bandits). In *36th International Conference on Algorithmic Learning Theory*, 2025.

[4] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[5] Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

[6] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[7] Nima Hamidi and Mohsen Bayati. On frequentist regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

[8] Tom Huix, Matthew Zhang, and Alain Durmus. Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 8749–8770. PMLR, 2023.

[9] David Janz, Alexander Litvak, and Csaba Szepesvari. Ensemble sampling for linear bandits: small ensembles suffice. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[10] David Janz, Arya Akhavan, and Csaba Szepesvári. Sharp analysis of linear ensemble sampling. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3716–3750. PMLR, 2026.

[11] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

[12] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *International Joint Conference on Artificial Intelligence*, 2019.

[13] Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Uncertainty in Artificial Intelligence*, pages 530–540. PMLR, 2020.

[14] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

[15] Harin Lee and Min-hwan Oh. Improved regret of linear ensemble sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[16] Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

[17] Xiuyuan Lu, Zheng Wen, and Branislav Kveton. Efficient online recommendation via low-rank ensemble sampling. In *Proceedings of the 12th ACM conference on recommender systems*, pages 460–464, 2018.

[18] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

[19] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[20] Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

[21] Chao Qin, Zheng Wen, Xiuyuan Lu, and Benjamin Van Roy. An analysis of ensemble sampling. *Advances in Neural Information Processing Systems*, 35:21602–21614, 2022.

[22] Jiazheng Sun, Weixin Wang, and Pan Xu. Provable anytime ensemble sampling algorithms in nonlinear contextual bandits. *arXiv preprint arXiv:2510.10730*, 2025.

[23] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[24] Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

[25] Jie Zhou, Botao Hao, Zheng Wen, Jingfei Zhang, and Will Wei Sun. Stochastic low-rank tensor bandits for multi-dimensional online decision making. *Journal of the American Statistical Association*, 120(549):198–211, 2025.

[26] Zheqing Zhu and Benjamin Van Roy. Deep exploration for recommendation systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 963–970, 2023.

Contents of Appendix

A Regret Analysis in the Dimension-Dominated Regime **14**

A.1 Deterministic Upper Bound on the Number of Epochs 14

A.2 Self-Normalized Exceedance Frequencies Inside One Epoch 15

A.3 Self-Normalized Exceedance Frequencies Across Epochs 17

A.4 Proof of Theorem 1 22

B Regret Analysis in the Finite-Arm-Dominated Regime **26**

B.1 Proof of Lemma 3 27

B.2 Proof of Theorem 2 33

C Technical Lemmas **36**

C.1 Proof of Lemma 5 36

C.2 Proof of Lemma 11 37

C.3 Proof of Lemma 12 38

D Auxiliary Lemmas **38**

E Experimental Details **39**

E.1 Experimental Setup 39

E.2 Additional Experiments 39

E.3 Anytime Implementation 40

E.4 Finite-Arm Experiments 42

A Regret Analysis in the Dimension-Dominated Regime

In this section, we provide the detailed proof of Theorem 1. We first establish a deterministic upper bound on the number of epochs, as discussed in the main text. We then prove a self-normalized exceedance guarantee within a single epoch and extend it uniformly across all epochs. Finally, we combine these ingredients with the standard regret-analysis framework for optimistic randomized algorithms to complete the proof of Theorem 1.

A.1 Deterministic Upper Bound on the Number of Epochs

The number of epochs generated by Algorithm 1 is random, since epoch refreshes depend on the realized Gram matrices, which in turn are determined by the algorithmic randomization and the observed data. In this section, we derive a deterministic upper bound on the total number of epochs.

Define the epoch start times recursively by

$$\tau_0 := 0, \qquad \tau_{r+1} := \min\left\{\inf\left\{t > \tau_r : \ \mathbf{V}_t \not\preceq (1+\alpha)\mathbf{V}_{\tau_r}\right\}, T\right\},$$

where we adopt the convention that $\inf \emptyset = \infty$. Then, the r-th epoch then consists of the rounds

$$\tau_r < t \le \tau_{r+1}\,.$$

The total number of epochs is bounded as follows.

Lemma 4 (Number of epochs)**.** *For $\alpha > 0$, let R_α denote the total number of epochs. Under the assumption that $\|\mathbf{x}\|_2 \le 1$ for all $\mathbf{x} \in \mathcal{X}$, we have*

$$R_\alpha \le 1 + \frac{d\log\left(1 + \frac{T}{\lambda d}\right)}{\log(1+\alpha)}\,.$$

Consequently, the deterministic quantity

$$\bar{R}_\alpha := 1 + \left\lfloor \frac{d\log\left(1 + \frac{T}{\lambda d}\right)}{\log(1+\alpha)} \right\rfloor$$

satisfies $R_\alpha \le \bar{R}_\alpha$.

Proof of Lemma 4. If epoch r is completed before the horizon, then by definition,

$$\mathbf{V}_{\tau_{r+1}} \not\preceq (1+\alpha)\mathbf{V}_{\tau_r}\,.$$

Let

$$\mathbf{A}_r := \mathbf{V}_{\tau_r}^{-1/2}\mathbf{V}_{\tau_{r+1}}\mathbf{V}_{\tau_r}^{-1/2}\,.$$

Since $\mathbf{V}_{\tau_{r+1}} \succeq \mathbf{V}_{\tau_r}$, all eigenvalues of $\mathbf{A}_r$ are at least 1. Moreover, since $\mathbf{V}_{\tau_{r+1}} \not\preceq (1+\alpha)\mathbf{V}_{\tau_r}$ by construction, at least one eigenvalue of $\mathbf{A}_r$ is strictly larger than $1+\alpha$. Hence,

$$\det(\mathbf{A}_r) > 1+\alpha \quad \Longrightarrow \quad \frac{\det(\mathbf{V}_{\tau_{r+1}})}{\det(\mathbf{V}_{\tau_r})} > 1+\alpha\,. \tag{3}$$

Since there are R_α epochs in total, exactly $R_\alpha - 1$ epochs are completed before the horizon, namely those indexed by $r = 0, 1, \ldots, R_\alpha - 2$. Multiplying the inequalities in (3) over these indices yields

$$\frac{\det(\mathbf{V}_{\tau_{R_\alpha-1}})}{\det(\lambda\mathbf{I}_d)} = \prod_{r=0}^{R_\alpha-2} \frac{\det(\mathbf{V}_{\tau_{r+1}})}{\det(\mathbf{V}_{\tau_r})} > (1+\alpha)^{R_\alpha-1}\,.$$

Finally, since $\tau_{R_\alpha} = T$ and $\mathbf{V}_t$ is increasing in t with respect to the positive semidefinite order, we have

$$\mathbf{V}_T \succeq \mathbf{V}_{\tau_{R_\alpha-1}},$$

which implies

$$\det(\mathbf{V}_T) \ge \det(\mathbf{V}_{\tau_{R_\alpha-1}})\,.$$

Therefore,

$$(1+\alpha)^{R_\alpha - 1} < \frac{\det(\mathbf{V}_{\tau_{R_\alpha - 1}})}{\det(\lambda \mathbf{I}_d)} \le \frac{\det(\mathbf{V}_T)}{\det(\lambda \mathbf{I}_d)}.$$

This proves that

$$R_\alpha \le 1 + \frac{\log\big(\det(\mathbf{V}_T)/\det(\lambda \mathbf{I}_d)\big)}{\log(1+\alpha)}.$$

Also, the determinant-trace inequality (Lemma 13) gives

$$\det(\mathbf{V}_T) \le \left(\lambda + \frac{T}{d}\right)^d,$$

and hence

$$\frac{\det(\mathbf{V}_T)}{\det(\lambda \mathbf{I}_d)} \le \left(1 + \frac{T}{\lambda d}\right)^d.$$

Combining the above displays yields the final bound

$$R_\alpha \le 1 + \frac{d \log\big(1 + \frac{T}{\lambda d}\big)}{\log(1+\alpha)}.$$

Since R_α is an integer, this further implies

$$R_\alpha \le 1 + \left\lfloor \frac{d \log\big(1 + \frac{T}{\lambda d}\big)}{\log(1+\alpha)} \right\rfloor = \bar{R}_\alpha.$$

□

A.2 Self-Normalized Exceedance Frequencies Inside One Epoch

The central object in the analysis is the self-normalized exceedance frequency, which measures the fraction of ensemble perturbations that are sufficiently large in a given direction at a given round. In the setting of Janz et al. [10], for each fixed direction, the projected perturbation process can be represented by Brownian motions whose clock is the corresponding self-normalized norm. Since the ensemble is not refreshed, this clock must be controlled over the full Gram-matrix growth along the horizon. Our epoch-refresh construction localizes this difficulty: within each epoch, the Gram matrix changes by at most a factor $1+\alpha$, so the associated Brownian clock varies only over a local multiplicative range. Therefore, the Brownian exceedance theorem can be applied with clock ratio $1+\alpha$, rather than the global Gram-matrix growth. This is the key reason the fixed-direction exceedance guarantee can be obtained with fewer ensemble members.

Lemma 1 (Fixed-direction exceedance inside one epoch)**.** *Fix an epoch r and a direction $\mathbf{u} \in \mathbb{S}^{d-1}$. For $\alpha > 0$, let $\gamma := 40$, $\kappa_\alpha := \lceil 250 \log(1+\alpha) \rceil$. Then for every $\delta' \in (0,1)$, if $m \ge 40 \log \frac{\kappa_\alpha}{\delta'}$, then with probability at least $1-\delta'$, $\forall t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$, we have $E_{t,m}(\mathbf{u}, 2/\gamma) \ge \frac{1}{10}$.*

Proof of Lemma 1. Fix an epoch r and condition on the history up to time τ_r. By construction,

$$\forall j \in [m]: \qquad \mathbf{s}^j_{\tau_r} = \mathbf{V}^{1/2}_{\tau_r} \mathbf{g}^j_r, \qquad \mathbf{g}^j_r \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d).$$

Hence, for every $t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$ and every $j \in [m]$,

$$\langle \mathbf{u}, \mathbf{s}^j_{t-1} \rangle = \left\langle \mathbf{u}, \mathbf{V}^{1/2}_{\tau_r} \mathbf{g}^j_r \right\rangle + \sum_{s=\tau_r+1}^{t-1} \langle \mathbf{u}, \mathbf{x}_s \rangle \, \xi^j_s.$$

Note that for $t \in \{\tau_r+1, \dots, \tau_{r+1}\}$, the arm $\mathbf{x}_t$ is selected before ξ^j_t is sampled. Hence, the coefficient $\langle \mathbf{u}, \mathbf{x}_t \rangle$ is predictable with respect to the filtration immediately before the perturbation noise ξ^j_t is drawn. Therefore, the projected process is a Gaussian martingale transform with predictable scalar coefficients.

On the other hand, by the epoch-refresh construction, all ensemble perturbations are reinitialized at time τ_r, so within epoch r they evolve from a common Gaussian restart point. This differs from

the Brownian representation in Janz et al. [10], where the process is tracked globally over the entire horizon starting from time 0. Here, by conditioning on the history up to τ_r, we obtain an *epoch-wise* Brownian representation whose clock starts from the local covariance matrix $\mathbf{V}_{\tau_r}$. We formalize this observation in the following lemma, which gives an epoch-wise Brownian representation for a diagonal martingale transform of Gaussian noise. Its proof is deferred to Appendix C.1.

Lemma 5 (Epoch-wise scalar-clock Brownian representation)**.** *Fix an epoch r and a direction $\mathbf{u} \in \mathbb{S}^{d-1}$. Let*

$$\ell_r := \tau_{r+1} - \tau_r \,.$$

For $k = 0, 1, \ldots, \ell_r - 1$, define the m-dimensional vector

$$\mathbf{M}_k := \big(\langle \mathbf{u}, \mathbf{s}^1_{\tau_r+k}\rangle, \ldots, \langle \mathbf{u}, \mathbf{s}^m_{\tau_r+k}\rangle\big) \in \mathbb{R}^m \,.$$

Then $(\mathbf{M}_k)_{k=0}^{\ell_r-1}$ is a diagonal martingale transform of a standard m-dimensional Gaussian noise sequence with scalar coefficients

$$D_0 := \|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}, \qquad D_k := \langle \mathbf{u}, \mathbf{x}_{\tau_r+k}\rangle \,, \qquad k = 1, \ldots, \ell_r - 1 \,.$$

Consequently, on an extension of the probability space, there exist independent standard Brownian motions $W^1, \ldots, W^m$ such that for every $k = 0, \ldots, \ell_r - 1$ and every $j \in [m]$,

$$\mathbf{M}^j_k = W^j(A^2_k) \,, \qquad A^2_k := D_0^2 + \sum_{s=1}^{k} D_s^2 = \|\mathbf{u}\|^2_{\mathbf{V}_{\tau_r+k}} \,.$$

Equivalently, for every $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$ and every $j \in [m]$,

$$\langle \mathbf{u}, \mathbf{s}^j_{t-1}\rangle = W^j\big(\|\mathbf{u}\|^2_{\mathbf{V}_{t-1}}\big) \,.$$

Applying Lemma 5, on an extension of the probability space, there exist independent standard Brownian motions $W^1, \ldots, W^m$ such that for all $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$ and all $j \in [m]$,

$$\langle \mathbf{u}, \mathbf{s}^j_{t-1}\rangle = W^j\big(\|\mathbf{u}\|^2_{\mathbf{V}_{t-1}}\big).$$

Therefore,

$$E_{t,m}(\mathbf{u}, 2/\gamma) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{ \frac{W^j(\|\mathbf{u}\|^2_{\mathbf{V}_{t-1}})}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} \geq \frac{2}{\gamma} \right\}.$$

Now, for every $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$, we have

$$\mathbf{V}_{\tau_r} \preceq \mathbf{V}_{t-1} \preceq (1+\alpha)\mathbf{V}_{\tau_r} \,,$$

which implies

$$\|\mathbf{u}\|^2_{\mathbf{V}_{t-1}} \in \big[\|\mathbf{u}\|^2_{\mathbf{V}_{\tau_r}},\ (1+\alpha)\|\mathbf{u}\|^2_{\mathbf{V}_{\tau_r}}\big] \,.$$

Thus, the Brownian clock ratio is at most $1 + \alpha$.

We now invoke the time-uniform exceedance count result for Brownian motions along a fixed direction due to Janz et al. [10].

Lemma 6 (Theorem 5.4 in Janz et al. [10])**.** *Let $W^1, \ldots, W^m$ be independent standard Brownian motions. Fix $0 < \tau \leq \tau' < \infty$, $c > 0$, and $0 < p < p_0(c) := \frac{1}{4}(1 - \Phi(c))$. Then there exists $h \in (0, 1]$ such that, for any $\delta \in (0, 1)$, if*

$$m \geq \frac{4}{p} \log\left(\frac{\lceil \log(\tau'/\tau)/h \rceil}{\delta} \right),$$

then with probability at least $1 - \delta$,

$$\inf_{t \in [\tau, \tau']} \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{ \frac{W^j(t)}{\sqrt{t}} \geq c \right\} \geq p.$$

Moreover, if $c \leq 1/20$ and $p \leq 1/10$, then $h = 1/250$ is admissible.

Applying Lemma 6 on the interval

$$[\tau, \tau'] = \left[\|\mathbf{u}\|^2_{\mathbf{V}_{\tau_r}},\ (1+\alpha)\|\mathbf{u}\|^2_{\mathbf{V}_{\tau_r}}\right],$$

with $p = 1/10$, $c = 1/20$, and the admissible choice $h = 1/250$, we obtain the desired conclusion provided that

$$m \geq \frac{4}{p}\log\left(\frac{\lceil \log(\tau'/\tau)/h \rceil}{\delta'}\right) = 40\log\left(\frac{\lceil 250\log(\tau'/\tau)\rceil}{\delta'}\right).$$

Since $\tau'/\tau \leq 1+\alpha$, it suffices to require

$$m \geq 40\log\frac{\kappa_\alpha}{\delta'}, \qquad \kappa_\alpha := \lceil 250\log(1+\alpha)\rceil.$$

This concludes the proof. □

A.3 Self-Normalized Exceedance Frequencies Across Epochs

The fixed-direction bound is not sufficient for general bounded arm sets, because the optimistic direction can depend on both the round and the unknown optimal arm. We therefore extend the fixed-direction guarantee uniformly over all directions on the unit sphere. The first ingredient is a local Lipschitz property of the normalized map induced by the Gram-matrix change within an epoch.

Lemma 7 (Local Lipschitz map within one epoch)**.** *Fix an epoch r and a round $t \in \{\tau_r+1, \dots, \tau_{r+1}\}$. Let*

$$\mathbf{B}_{r,t} := \mathbf{V}_{t-1}^{1/2}\mathbf{V}_{\tau_r}^{-1/2}.$$

Define a map $\boldsymbol{\phi}_{r,t} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ by

$$\boldsymbol{\phi}_{r,t}(\mathbf{w}) := \frac{\mathbf{B}_{r,t}\mathbf{w}}{\|\mathbf{B}_{r,t}\mathbf{w}\|_2}, \qquad \mathbf{w} \in \mathbb{S}^{d-1}.$$

Then, for any $\mathbf{w}, \mathbf{w}' \in \mathbb{S}^{d-1}$,

$$\|\boldsymbol{\phi}_{r,t}(\mathbf{w}) - \boldsymbol{\phi}_{r,t}(\mathbf{w}')\|_2 \leq 2\sqrt{1+\alpha}\,\|\mathbf{w}-\mathbf{w}'\|_2.$$

Proof of Lemma 7. Since t lies in epoch r, by the definition of the epoch refresh rule we have

$$\mathbf{V}_{\tau_r} \preceq \mathbf{V}_{t-1} \preceq (1+\alpha)\mathbf{V}_{\tau_r}.$$

Multiplying on the left and right by $\mathbf{V}_{\tau_r}^{-1/2}$, we obtain

$$\mathbf{I}_d \preceq \mathbf{V}_{\tau_r}^{-1/2}\mathbf{V}_{t-1}\mathbf{V}_{\tau_r}^{-1/2} = \mathbf{B}_{r,t}^\top\mathbf{B}_{r,t} \preceq (1+\alpha)\mathbf{I}_d.$$

Therefore,

$$\|\mathbf{B}_{r,t}\|_{\mathrm{op}} \leq \sqrt{1+\alpha}, \qquad \inf_{\|\mathbf{w}\|_2=1}\|\mathbf{B}_{r,t}\mathbf{w}\|_2 \geq 1.$$

Now let

$$\mathbf{a} := \mathbf{B}_{r,t}\mathbf{w}, \qquad \mathbf{b} := \mathbf{B}_{r,t}\mathbf{w}',$$

where $\mathbf{w}, \mathbf{w}' \in \mathbb{S}^{d-1}$. Since $\|\mathbf{a}\|_2, \|\mathbf{b}\|_2 \geq 1$, we may use the elementary inequality

$$\left\|\frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2}\right\|_2 \leq \frac{2\|\mathbf{a}-\mathbf{b}\|_2}{\min\{\|\mathbf{a}\|_2, \|\mathbf{b}\|_2\}}.$$

Applying this with $\mathbf{a} = \mathbf{B}_{r,t}\mathbf{w}$ and $\mathbf{b} = \mathbf{B}_{r,t}\mathbf{w}'$, and using $\min\{\|\mathbf{a}\|_2, \|\mathbf{b}\|_2\} \geq 1$, we obtain

$$\|\boldsymbol{\phi}_{r,t}(\mathbf{w}) - \boldsymbol{\phi}_{r,t}(\mathbf{w}')\|_2 \leq 2\|\mathbf{B}_{r,t}(\mathbf{w}-\mathbf{w}')\|_2 \leq 2\|\mathbf{B}_{r,t}\|_{\mathrm{op}}\|\mathbf{w}-\mathbf{w}'\|_2 \leq 2\sqrt{1+\alpha}\,\|\mathbf{w}-\mathbf{w}'\|_2.$$

□

To make the net argument quantitative, we also need a uniform bound on the normalized ensemble perturbations over all epochs. This guarantees that transferring exceedance from a nearby net point to an arbitrary direction incurs only a controlled loss, uniformly over time and across epochs.

Lemma 8 (Uniform bound on normalized perturbations)**.** *For a given $\delta \in (0,1)$, define*

$$\Gamma_\delta := \underbrace{\sqrt{d} + \sqrt{2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)}}_{\textit{refreshed Gaussian term}} + \underbrace{\sqrt{d\log(1+\alpha) + 2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)}}_{\textit{within-epoch self-normalized term}}.$$

With probability at least $1-\delta$, we have

$$\forall r \le R_\alpha,\ \forall j \in [m],\ \forall t \in \{\tau_r+1, \dots, \tau_{r+1}\}: \qquad \|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\|_2 \le \Gamma_\delta.$$

Proof of Lemma 8. Fix an epoch r and an ensemble index $j \in [m]$. For each $t \in \{\tau_r+1, \dots, \tau_{r+1}\}$, the construction of the algorithm gives

$$\mathbf{s}_{t-1}^j = \mathbf{V}_{\tau_r}^{1/2}\mathbf{g}_r^j + \sum_{s=\tau_r+1}^{t-1} \mathbf{x}_s \xi_s^j.$$

Hence,

$$\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j = \underbrace{\mathbf{V}_{t-1}^{-1/2}\mathbf{V}_{\tau_r}^{1/2}\mathbf{g}_r^j}_{\text{refreshed Gaussian term}} + \underbrace{\mathbf{V}_{t-1}^{-1/2}\sum_{s=\tau_r+1}^{t-1}\mathbf{x}_s\xi_s^j}_{\text{within-epoch martingale term}}.$$

We bound these two terms separately.

Step 1: The refreshed Gaussian term. Since $\mathbf{V}_{t-1} \succeq \mathbf{V}_{\tau_r}$, we have

$$\|\mathbf{V}_{t-1}^{-1/2}\mathbf{V}_{\tau_r}^{1/2}\|_{\mathrm{op}} \le 1.$$

Therefore,

$$\|\mathbf{V}_{t-1}^{-1/2}\mathbf{V}_{\tau_r}^{1/2}\mathbf{g}_r^j\|_2 \le \|\mathbf{g}_r^j\|_2.$$

Moreover, by the concentration inequality for the Euclidean norm of a standard Gaussian vector (Lemma 15),

$$\mathbb{P}\Big(\|\mathbf{g}_r^j\|_2 > \sqrt{d} + \sqrt{2x}\Big) \le e^{-x}.$$

Taking

$$x = \log\left(\frac{4m\bar{R}_\alpha}{\delta}\right),$$

and applying a union bound over all pairs (r, j), we conclude that, with probability at least $1-\delta/2$,

$$\forall r \le \bar{R}_\alpha,\ \forall j \in [m]: \qquad \|\mathbf{g}_r^j\|_2 \le \sqrt{d} + \sqrt{2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)}.$$

Step 2: The within-epoch martingale term.

For a fixed epoch r and ensemble index $j \in [m]$, define

$$\mathbf{M}_t^{(r,j)} := \sum_{s=\tau_r+1}^{t-1} \mathbf{x}_s \xi_s^j, \qquad t \in \{\tau_r+1, \dots, \tau_{r+1}\}.$$

Conditional on the history up to time τ_r, namely $\mathcal{F}_{\tau_r}$, the process $\{\mathbf{M}_t^{(r,j)}\}_{t=\tau_r+1}^{\tau_{r+1}}$ is a vector-valued martingale transform with conditionally Gaussian increments and predictable coefficients. More precisely, for each $s \in \{\tau_r+1, \dots, \tau_{r+1}-1\}$, the increment

$$\mathbf{x}_s \xi_s^j$$

has conditional mean zero given the past, the scalar noise ξ_s^j is conditionally standard Gaussian, and the coefficient vector $\mathbf{x}_s$ is predictable.

We may therefore apply the standard self-normalized inequality for vector-valued martingales with conditionally sub-Gaussian increments (Lemma 16), with initial matrix $\mathbf{V}_{\tau_r}$. This implies that for every $\eta \in (0,1)$,

$$\mathbb{P}\left(\exists\, t \in \{\tau_r+1,\dots,\tau_{r+1}\} : \|\mathbf{V}_{t-1}^{-1/2}\mathbf{M}_t^{(r,j)}\|_2 > \sqrt{2\log(1/\eta) + \log\frac{\det(\mathbf{V}_{t-1})}{\det(\mathbf{V}_{\tau_r})}} \;\middle|\; \mathcal{F}_{\tau_r}\right) \le \eta. \tag{4}$$

Next, we control the determinant ratio uniformly within the epoch. Since t lies in epoch r, the epoch refresh rule implies

$$\mathbf{V}_{t-1} \preceq (1+\alpha)\mathbf{V}_{\tau_r}.$$

Hence,

$$\frac{\det(\mathbf{V}_{t-1})}{\det(\mathbf{V}_{\tau_r})} \le (1+\alpha)^d.$$

Substituting this bound into the Eq. (4), we obtain

$$\mathbb{P}\left(\sup_{t\in\{\tau_r+1,\dots,\tau_{r+1}\}} \|\mathbf{V}_{t-1}^{-1/2}\mathbf{M}_t^{(r,j)}\|_2 > \sqrt{d\log(1+\alpha) + 2\log(1/\eta)}\right) \le \eta.$$

Finally, taking

$$\eta = \frac{\delta}{2m\bar{R}_\alpha},$$

and applying a union bound over all epochs $r \le \bar{R}_\alpha$ and all ensemble indices $j \in [m]$, we conclude that, with probability at least $1-\delta/2$, $\forall r \le \bar{R}_\alpha$, $\forall j \in [m]$, and $\forall t \in \{\tau_r+1,\dots,\tau_{r+1}\}$, it holds:

$$\|\mathbf{V}_{t-1}^{-1/2}\mathbf{M}_t^{(r,j)}\|_2 \le \sqrt{d\log(1+\alpha) + 2\log\left(\frac{2m\bar{R}_\alpha}{\delta}\right)}.$$

Intersecting the events from Steps 1 and 2 and applying the triangle inequality, we conclude that, with probability at least $1-\delta$,

$$\|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\|_2 \le \Gamma_\delta$$

for all $r \le \bar{R}_\alpha$, all $j \in [m]$, and all $t \in \{\tau_r+1,\dots,\tau_{r+1}\}$, where

$$\Gamma_\delta = \left(\sqrt{d} + \sqrt{2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)}\right) + \left(\sqrt{d\log(1+\alpha) + 2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)}\right).$$

This concludes the proof. □

We are now ready to combine the preceding ingredients. The fixed-direction epochwise exceedance bound provides control at a single net point within a single epoch. The local Lipschitz lemma shows that nearby directions induce nearby normalized geometries, while the uniform perturbation bound controls the loss incurred when passing from a net point to an arbitrary direction. Taking a union bound over the sphere net and over all epochs then yields a uniform exceedance lower bound simultaneously for all times and all directions.

Lemma 2 (Uniform exceedance count across epochs)**.** *Fix* $\gamma := 40$, $\kappa_\alpha := \lceil 250\log(1+\alpha)\rceil$, $N_\delta := \left(6\sqrt{1+\alpha}\,\gamma\Gamma_\delta\right)^d$. *If*

$$m \ge 40\log\left(\frac{2\kappa_\alpha\bar{R}_\alpha N_\delta}{\delta}\right), \tag{2}$$

then with probability at least $1-\delta$, $\min\limits_{t\in[T]}\ \inf\limits_{\mathbf{u}\in\mathbb{S}^{d-1}} E_{t,m}(\mathbf{u}, 1/\gamma) \ge \dfrac{1}{10}$.

Proof of Lemma 2. Fix an epoch r. For each nonzero $\mathbf{u} \in \mathbb{R}^d$, define

$$\mathbf{w}(\mathbf{u}) := \frac{\mathbf{V}_{\tau_r}^{1/2}\mathbf{u}}{\|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}} \in \mathbb{S}^{d-1}.$$

Then for every $t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$ and every $j \in [m]$, since

$$\frac{\mathbf{V}_{t-1}^{1/2}\mathbf{u}}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} = \frac{\mathbf{V}_{t-1}^{1/2}\mathbf{V}_{\tau_r}^{-1/2}\mathbf{w}(\mathbf{u})}{\|\mathbf{V}_{t-1}^{1/2}\mathbf{V}_{\tau_r}^{-1/2}\mathbf{w}(\mathbf{u})\|_2} = \frac{\mathbf{B}_{r,t}\mathbf{w}(\mathbf{u})}{\|\mathbf{B}_{r,t}\mathbf{w}(\mathbf{u})\|_2} = \boldsymbol{\phi}_{r,t}(\mathbf{w}(\mathbf{u})),$$

we have

$$\frac{\langle \mathbf{u}, \mathbf{s}_{t-1}^j \rangle}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} = \frac{\left\langle \mathbf{V}_{t-1}^{1/2}\mathbf{u}, \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} = \left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}(\mathbf{u})), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle. \tag{5}$$

Now let $\mathcal{N}_r$ be a deterministic ε-net of $\mathbb{S}^{d-1}$ with $\varepsilon := 1/(2\sqrt{1+\alpha}\,\gamma\Gamma_\delta)$. By the standard covering bound, we may choose $\mathcal{N}_r$ so that

$$|\mathcal{N}_r| \le \left(\frac{3}{\varepsilon}\right)^d = \left(6\sqrt{1+\alpha}\,\gamma\Gamma_\delta\right)^d =: N_\delta\,.$$

For each fixed $\mathbf{w}' \in \mathcal{N}_r$, choose a nonzero vector $\mathbf{u}(\mathbf{w}')$ such that $\mathbf{w}(\mathbf{u}(\mathbf{w}')) = \mathbf{w}'$. Such a vector exists because $\mathbf{V}_{\tau_r}$ is positive definite, and hence $\mathbf{V}_{\tau_r}^{1/2}$ is invertible. Therefore, the map $\mathbf{u} \mapsto \mathbf{V}_{\tau_r}^{1/2}\mathbf{u}/\|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}$ is surjective onto $\mathbb{S}^{d-1}$.

Applying Lemma 1 with $\delta_{r,\mathbf{w}'} := \delta/(2\bar{R}_\alpha N_\delta)$, and using condition (2) to ensure that

$$m \ge 40 \log \frac{\kappa_\alpha}{\delta_{r,\mathbf{w}'}} = 40 \log \frac{2\kappa_\alpha \bar{R}_\alpha N_\delta}{\delta}\,,$$

we conclude that, with probability at least $1 - \delta_{r,\mathbf{w}'}$,

$$\forall t \in \{\tau_r + 1, \dots, \tau_{r+1}\}: \qquad E_{t,m}(\mathbf{u}(\mathbf{w}'), 2/\gamma) \ge \frac{1}{10}.$$

Define the event

$$\mathcal{E}_{\mathrm{net}} := \left\{ \forall r \le \bar{R}_\alpha,\ \forall \mathbf{w}' \in \mathcal{N}_r,\ \forall t \in \{\tau_r + 1, \dots, \tau_{r+1}\}: \quad E_{t,m}(\mathbf{u}(\mathbf{w}'), 2/\gamma) \ge \frac{1}{10} \right\}.$$

Taking a union bound over all epochs $r \le \bar{R}_\alpha$ and all net points $\mathbf{w}' \in \mathcal{N}_r$, we obtain $\mathbb{P}(\mathcal{E}_{\mathrm{net}}) \ge 1 - \delta/2$.

Next, define

$$\mathcal{E}_{\mathrm{state}} := \left\{ \forall r \le \bar{R}_\alpha,\ \forall j \in [m],\ \forall t \in \{\tau_r + 1, \dots, \tau_{r+1}\}: \quad \|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\|_2 \le \Gamma_\delta \right\}.$$

By Lemma 8, we have $\mathbb{P}(\mathcal{E}_{\mathrm{state}}) \ge 1 - \delta/2$. Hence, by the union bound, $\mathbb{P}(\mathcal{E}_{\mathrm{net}} \cap \mathcal{E}_{\mathrm{state}}) \ge 1 - \delta$.

We now work on the event $\mathcal{E}_{\mathrm{net}} \cap \mathcal{E}_{\mathrm{state}}$. Fix an epoch r, a round $t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$, an ensemble index $j \in [m]$, and $\mathbf{w}, \mathbf{w}' \in \mathbb{S}^{d-1}$. By Lemma 7,

$$\|\boldsymbol{\phi}_{r,t}(\mathbf{w}) - \boldsymbol{\phi}_{r,t}(\mathbf{w}')\|_2 \le 2\sqrt{1+\alpha}\,\|\mathbf{w} - \mathbf{w}'\|_2.$$

Since $\|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\|_2 \le \Gamma_\delta$ on $\mathcal{E}_{\mathrm{state}}$, we obtain

$$\begin{aligned}
\left|\left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle - \left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}'), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle\right| &= \left|\left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}) - \boldsymbol{\phi}_{r,t}(\mathbf{w}'), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle\right| \\
&\le \|\boldsymbol{\phi}_{r,t}(\mathbf{w}) - \boldsymbol{\phi}_{r,t}(\mathbf{w}')\|_2\, \|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\|_2 \\
&\le 2\sqrt{1+\alpha}\,\Gamma_\delta \|\mathbf{w} - \mathbf{w}'\|_2.
\end{aligned}$$

Now fix any $t \in [T]$ and any $\mathbf{u} \in \mathbb{S}^{d-1}$, and let r be the epoch containing round t. Choose $\mathbf{w}' \in \mathcal{N}_r$ such that $\|\mathbf{w}(\mathbf{u}) - \mathbf{w}'\|_2 \le \varepsilon$. Then, for every $j \in [m]$,

$$\left|\left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}(\mathbf{u})), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle - \left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}'), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j \right\rangle\right| \le 2\sqrt{1+\alpha}\,\Gamma_\delta \varepsilon = \frac{1}{\gamma}.$$

Hence, if

$$\left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}'), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^{j} \right\rangle \geq \frac{2}{\gamma},$$

then necessarily

$$\left\langle \boldsymbol{\phi}_{r,t}(\mathbf{w}(\mathbf{u})), \mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^{j} \right\rangle \geq \frac{1}{\gamma}.$$

Using (5), this is equivalent to

$$\frac{\langle \mathbf{u}(\mathbf{w}'), \mathbf{s}_{t-1}^{j} \rangle}{\|\mathbf{u}(\mathbf{w}')\|_{\mathbf{V}_{t-1}}} \geq \frac{2}{\gamma} \quad \Longrightarrow \quad \frac{\langle \mathbf{u}, \mathbf{s}_{t-1}^{j} \rangle}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} \geq \frac{1}{\gamma}.$$

Therefore,

$$\mathbb{1}\left\{ \frac{\langle \mathbf{u}(\mathbf{w}'), \mathbf{s}_{t-1}^{j} \rangle}{\|\mathbf{u}(\mathbf{w}')\|_{\mathbf{V}_{t-1}}} \geq \frac{2}{\gamma} \right\} \leq \mathbb{1}\left\{ \frac{\langle \mathbf{u}, \mathbf{s}_{t-1}^{j} \rangle}{\|\mathbf{u}\|_{\mathbf{V}_{t-1}}} \geq \frac{1}{\gamma} \right\}.$$

Summing over $j \in [m]$ and dividing by m, we obtain

$$E_{t,m}(\mathbf{u}, 1/\gamma) \geq E_{t,m}(\mathbf{u}(\mathbf{w}'), 2/\gamma).$$

Since $\mathcal{E}_{\mathrm{net}}$ guarantees that $E_{t,m}(\mathbf{u}(\mathbf{w}'), 2/\gamma) \geq 1/10$, it follows that

$$E_{t,m}(\mathbf{u}, 1/\gamma) \geq \frac{1}{10}.$$

As $t \in [T]$ and $\mathbf{u} \in \mathbb{S}^{d-1}$ were arbitrary, we conclude that

$$\forall t \in [T],\ \forall \mathbf{u} \in \mathbb{S}^{d-1}: \qquad E_{t,m}(\mathbf{u}, 1/\gamma) \geq \frac{1}{10}.$$

Since this holds on $\mathcal{E}_{\mathrm{net}} \cap \mathcal{E}_{\mathrm{state}}$, whose probability is at least $1 - \delta$, this concludes the proof. □

Now, we show that the sufficient condition in Eq. (2) can be simplified as follows:

Corollary 1 (Sufficient ensemble size)**.** *Suppose $0 < \alpha \leq 1$. There exists a universal constant $C > 0$ such that the sufficient condition on the ensemble size in Eq.* (2) *is satisfied by choosing*

$$m = \left\lceil C \left[d \log\left(d + \log \frac{\bar{R}_\alpha}{\delta} \right) + \log \frac{\bar{R}_\alpha}{\delta} \right] \right\rceil.$$

Proof of Corollary 1. Let

$$z := \log \frac{\bar{R}_\alpha}{\delta}.$$

Since $0 < \alpha \leq 1$, we have $\sqrt{1+\alpha} \leq \sqrt{2}$, $\log(1+\alpha) \leq \log 2$, which implies

$$\kappa_\alpha = \lceil 250 \log(1+\alpha) \rceil \leq \lceil 250 \log 2 \rceil.$$

Thus all explicit α-dependent prefactors in κ_α and N_δ are bounded by universal constants.

Recall that

$$N_\delta = \left(6\sqrt{1+\alpha}\, \gamma \Gamma_\delta \right)^d.$$

Moreover, by the definition of Γ_δ,

$$\Gamma_\delta = \sqrt{d} + \sqrt{2 \log\left(\frac{4m\bar{R}_\alpha}{\delta} \right)} + \sqrt{d \log(1+\alpha) + 2 \log\left(\frac{4m\bar{R}_\alpha}{\delta} \right)}.$$

Using $0 < \alpha \leq 1$, we get

$$\Gamma_\delta \leq C_1 \sqrt{d + \log m + z}$$

for a universal constant $C_1 > 0$. Therefore,

$$\log N_\delta = d \log\left(6\sqrt{1+\alpha}\, \gamma \Gamma_\delta \right) \leq C_2 d \log(d + \log m + z)$$

for another universal constant $C_2 > 0$.

Substituting this bound into Eq. (2), we see that it is enough to ensure

$$m \geq C_3 \left[z + d \log(d + \log m + z) \right] \tag{6}$$

for a universal constant $C_3 > 0$. Now define

$$M := d \log(d + z) + z.$$

We claim that choosing $m = \lceil C_4 M \rceil$ with a sufficiently large universal constant $C_4 > 0$ satisfies the inequality in Eq. (6). Suppose that $m = \lceil C_4 M \rceil$. Indeed, since

$$M = d \log(d + z) + z \leq d(d + z) + z \leq (d + z)^2 .$$

Then, we obtain

$$\log m \leq C_5 + \log M \leq C_6 \log(d + z)$$

for universal constants $C_5, C_6 > 0$. Hence

$$d + \log m + z \leq d + z + C_6 \log(d + z) \leq C_7 (d + z)$$

for a universal constant $C_7 > 0$, and therefore

$$d \log(d + \log m + z) \leq d \log(C_7 (d + z)) \leq C_8 d \log(d + z)$$

for a universal constant $C_8 > 0$. It follows that

$$C_3 \left[z + d \log(d + \log m + z) \right] \leq C_3 \left[z + C_8 d \log(d + z) \right] \leq C_9 \left[d \log(d + z) + z \right] = C_9 M .$$

Choosing $C_4 \geq C_9$ proves that $m = \lceil C_4 M \rceil$ implies (2). □

A.4 Proof of Theorem 1

In the previous section, we established that our epoch-refresh technique yields time-uniform exceedance control with a substantially smaller ensemble size. Accordingly, Theorem 1 follows by replacing the uniform exceedance control component in the proof of Janz et al. [10] with our Lemma 2. For completeness, we provide the full proof below.

Theorem 1 (Regret bound for `LinES-ER`)**.** *Suppose Assumption 1 holds and $T \geq \max\{2, d\}$. Fix any $\delta \in (0, 1/4)$, and choose the parameters in Algorithm 1 as $\lambda \geq 80$, $\gamma = 40$, $\alpha = 1$,*

$$\beta_t = \sqrt{d \log\left(1 + \frac{t}{d\lambda}\right) + 2 \log \frac{1}{\delta}} + \sqrt{\lambda}\, S, \quad \textit{and} \quad m \geq C \left(d \log\left(d + \log \frac{\bar{R}_\alpha}{\delta} \right) + \log \frac{\bar{R}_\alpha}{\delta} \right),$$

where $C > 0$ is an absolute constant and $\bar{R}_\alpha := 1 + \left\lfloor d \log_{1+\alpha}\left(1 + \frac{T}{\lambda d}\right) \right\rfloor$. Then, with probability at least $1 - 4\delta$, Algorithm 1 incurs cumulative regret $\mathrm{Regret}_T = \widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{T}\right)$.

Proof of Theorem 1. We begin by recalling several standard lemmas used in the regret analysis of linear bandit algorithms. The first is the well-known concentration result for the ridge estimator.

Lemma 9 (Theorem 2 of Abbasi-Yadkori et al. [1])**.** *With probability at least $1 - \delta$, for all $t \geq 0$, $\boldsymbol{\theta}^*$ lies in the set*

$$\left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|_{\mathbf{V}_t} \leq \sqrt{d \log\left(1 + \frac{t}{d\lambda}\right) + 2 \log \frac{1}{\delta}} + \sqrt{\lambda} S \right\} .$$

Next, we recall the general regret analysis framework of Janz et al. [9] for linear bandit algorithms that select arms greedily with respect to an estimated parameter. Although the original result is stated for a fixed arm set, its proof only requires that the available arm set at each round be measurable with respect to the history before the arm is selected. Therefore, the same argument applies to time-varying arm sets under this measurability condition.

Lemma 10 (Adapted from Theorem 2 of Janz et al. [9]). *Fix $\delta \in (0,1]$, $\lambda \geq 1$, and $T \in \mathbb{N}$. Let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration, and let $\{\mathcal{A}_t\}_{t\geq 0}$ be a filtration on the same probability space such that*

$$\mathcal{F}_{t-1} \subset \mathcal{A}_{t-1} \subset \mathcal{F}_t \qquad \text{for all } t \geq 1.$$

Assume that, for each $t \geq 1$, the sigma-field $\mathcal{F}_t$ makes $\mathcal{X}_1, \mathbf{x}_1, y_1, \ldots, \mathcal{X}_t, \mathbf{x}_t, y_t$ as well as all learner randomizations used up to round t measurable.

Let $\{\mathcal{X}_t\}_{t\geq 1}$ be a sequence of closed subsets of the unit ball $\mathcal{B}_2^d$, each $\mathcal{A}_{t-1}$-measurable. Let $\{\boldsymbol{\theta}_t\}_{t\geq 1}$ be an $\mathcal{F}_t$-adapted $\mathbb{R}^d$-valued sequence, and suppose that

$$\forall t \geq 1, \qquad \mathbf{x}_t \in \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}_t \rangle \,.$$

For each $t \geq 1$, let

$$\mathbf{V}_{t-1} := \lambda \mathbf{I}_d + \sum_{s=1}^{t-1} \mathbf{x}_s \mathbf{x}_s^\top,$$

and let $\hat{\boldsymbol{\theta}}_{t-1}$ be the usual ridge regression estimate of $\boldsymbol{\theta}^*$, which we assume to be $\mathcal{F}_{t-1}$-measurable. Let $\{b_{t-1}\}_{t\geq 1}$ be a nonnegative $\mathcal{F}_t$-predictable sequence such that:

1. *the event $\mathcal{E} = \{\forall t \leq T,\ \|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \leq b_{t-1}\}$ satisfies $\mathbb{P}(\mathcal{E}^{\complement}) \leq \delta$; and*
2. *the event $\mathcal{E}_* = \{\forall t \leq T,\ \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \leq b_{t-1}\}$ satisfies $\mathbb{P}(\mathcal{E}_*^{\complement}) \leq \delta$.*

For each $t \geq 1$, let

$$\mathbf{x}_t^* \in \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle$$

be an $\mathcal{A}_{t-1}$-measurable choice, and define the $\mathcal{A}_{t-1}$-conditional probability of optimism by

$$p_{t-1} := \mathbb{P}(\langle \mathbf{x}_t, \boldsymbol{\theta}_t \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle \,|\, \mathcal{A}_{t-1}) \,.$$

Then, on an event $\bar{\mathcal{E}} \subset \mathcal{E} \cap \mathcal{E}_*$ with $\mathbb{P}(\bar{\mathcal{E}}) \geq 1 - 3\delta$, for all $\tau \in [T]$, the τ-step regret

$$\mathrm{Regret}_\tau := \sum_{t=1}^{\tau} \langle \mathbf{x}_t^* - \mathbf{x}_t, \boldsymbol{\theta}^* \rangle$$

satisfies

$$\mathrm{Regret}_\tau \leq 2 \max_{i\in[\tau]} \frac{b_{i-1}}{p_{i-1}} \left(2\sqrt{2d\tau \log\Big(1 + \frac{\tau}{d\lambda}\Big)} + \sqrt{2(4\tau/\lambda + 1) \log\left(\frac{\sqrt{4\tau/\lambda + 1}}{\delta} \right)} \right) .$$

Note that Lemma 10 follows by a direct extension of Theorem 2 of Janz et al. [9]. Indeed, suppose that each $\mathcal{X}_t \subset \mathcal{B}_2^d$ is nonempty, compact, and measurable before the arm selection step at round t. Then the fixed support function $J(\boldsymbol{\theta}) = \max_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{x}, \boldsymbol{\theta} \rangle$ used in [9] can be replaced by the time-dependent support function $J_t(\boldsymbol{\theta}) = \max_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta} \rangle$, and the optimistic region can be replaced by $\Theta_t^{\mathrm{OPT}} = \{\boldsymbol{\theta} : J_t(\boldsymbol{\theta}) \geq J_t(\boldsymbol{\theta}^*)\}$. The proof is otherwise unchanged, since it only uses, round by round, the fact that $\mathbf{x}_t \in \operatorname{argmax}_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}_t \rangle$, which implies $\mathbf{x}_t \in \partial J_t(\boldsymbol{\theta}_t)$, together with the confidence events $\boldsymbol{\theta}_t, \boldsymbol{\theta}^* \in \Theta_{t-1}$ and the elliptical potential bound for the selected arms. Hence, the same regret bound holds for the time-varying regret $\mathrm{Regret}_\tau = \sum_{s=1}^{\tau} \Big(\max_{\mathbf{x}\in\mathcal{X}_s} \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_s, \boldsymbol{\theta}^* \rangle \Big)$, with p_{t-1} defined using Θ_t^{OPT} instead of a fixed Θ^{OPT}.

In the remaining part of the proof, we first set up the notation and verify the conditions required to apply Lemma 10. For $t \in [T]$, define

$$b_{t-1} := \gamma \Gamma_\delta \beta_{t-1} \,.$$

Also, we define

$$\mathcal{E} := \Big\{ \forall t \leq T :\ \|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \leq b_{t-1} \Big\} \,.$$

Since

$$\boldsymbol{\theta}_t = \hat{\boldsymbol{\theta}}_{t-1} + \gamma\beta_{t-1}\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^{J_t},$$

we have

$$\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} = \gamma\beta_{t-1}\|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^{J_t}\|_2.$$

By Lemma 8, with probability at least $1-\delta$,

$$\forall t \le T,\ \forall j \in [m]: \qquad \|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^{j}\|_2 \le \Gamma_\delta.$$

Hence $\mathbb{P}(\mathcal{E}) \ge 1-\delta$.

Next, let

$$\mathcal{E}_* := \left\{\forall t \le T:\ \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \le b_{t-1}\right\}.$$

Since $\Gamma_\delta \ge \sqrt{d}$, we have $\gamma\Gamma_\delta \ge 1$, we have $b_{t-1} \ge \beta_{t-1}$. By Lemma 9 and our choice of β_t, we have $\mathbb{P}(\mathcal{E}^*) \ge 1-\delta$.

We now define filtrations adapted to the epoch-refresh construction. For each epoch-start time τ_r, let $\mathbf{g}_{\tau_r}^1, \dots, \mathbf{g}_{\tau_r}^m$ denote the refreshed Gaussian vectors sampled at time τ_r. Define

$$\mathcal{F}_t := \sigma(\{\mathbf{g}_{\tau_r}^1, \dots, \mathbf{g}_{\tau_r}^m\} : \tau_r \le t) \vee \sigma(\{\mathcal{X}_s, \mathbf{x}_s, y_s, J_s\} : s \le t) \vee \sigma(\{\xi_s^1, \dots, \xi_s^m\} : s < t), \quad (7)$$

and

$$\mathcal{A}_t := \sigma(\{\mathbf{s}_t^1, \dots, \mathbf{s}_t^m\}) \vee \sigma(\mathcal{X}_{t+1}) \vee \mathcal{F}_t \quad (8)$$

where $\vee$ denotes the join of σ-algebras. By construction, $\hat{\boldsymbol{\theta}}_t$, $\mathbf{V}_t$, and β_t are $\mathcal{F}_t$-measurable, while $\{\mathbf{s}_t^j\}_{j=1}^m$ are $\mathcal{A}_t$-measurable. Moreover, conditional on $\mathcal{A}_{t-1}$, the index J_t is uniform on $[m]$.

For each $t \in [T]$, define the $\mathcal{A}_{t-1}$-conditional probability of optimism by

$$p_{t-1} := \mathbb{P}(\langle \mathbf{x}_t, \boldsymbol{\theta}_t\rangle \ge \langle \mathbf{x}_t^*, \boldsymbol{\theta}^*\rangle \mid \mathcal{A}_{t-1}),$$

where

$$\mathbf{x}_t^* \in \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}^*\rangle.$$

Fix $t \in [T]$. If $\mathbf{x}_t^* = \mathbf{0}_d$, then trivially

$$\langle \mathbf{x}_t, \boldsymbol{\theta}_t\rangle = \max_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}_t\rangle \ge 0 = \langle \mathbf{x}_t^*, \boldsymbol{\theta}^*\rangle,$$

and hence $p_{t-1} = 1$.

Assume henceforth that $\mathbf{x}_t^* \ne \mathbf{0}_d$, and define

$$\mathbf{u}_t^* := \mathbf{V}_{t-1}^{-1}\mathbf{x}_t^*.$$

Then $\|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}} = \|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}}$. On the other hand, by Lemma 9, we have

$$\langle \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle \le \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}}\|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}} \le \beta_{t-1}\|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}}.$$

Therefore, if

$$\langle \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle \ge \beta_{t-1}\|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}},$$

then

$$\begin{aligned}
\langle \boldsymbol{\theta}_t, \mathbf{x}_t^*\rangle &= \langle \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle + \langle \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle \\
&\ge \langle \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle + \beta_{t-1}\|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}} \\
&\ge \langle \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle + \langle \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle = \langle \boldsymbol{\theta}^*, \mathbf{x}_t^*\rangle.
\end{aligned}$$

Since $\mathbf{x}_t \in \operatorname{argmax}_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}_t\rangle$, this implies

$$\langle \mathbf{x}_t, \boldsymbol{\theta}_t\rangle \ge \langle \mathbf{x}_t^*, \boldsymbol{\theta}^*\rangle.$$

Now,

$$\langle \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^*\rangle = \gamma\beta_{t-1}\left\langle \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^{J_t}, \mathbf{x}_t^*\right\rangle = \gamma\beta_{t-1}\left\langle \mathbf{s}_{t-1}^{J_t}, \mathbf{V}_{t-1}^{-1}\mathbf{x}_t^*\right\rangle = \gamma\beta_{t-1}\langle \mathbf{s}_{t-1}^{J_t}, \mathbf{u}_t^*\rangle.$$

Hence the sufficient condition

$$\frac{\langle \mathbf{s}_{t-1}^{J_t}, \mathbf{u}_t^* \rangle}{\|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}}} \geq \frac{1}{\gamma}$$

implies

$$\langle \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}, \mathbf{x}_t^* \rangle \geq \beta_{t-1} \|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}} = \beta_{t-1} \|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}},$$

and therefore implies optimism:

$$\frac{\langle \mathbf{s}_{t-1}^{J_t}, \mathbf{u}_t^* \rangle}{\|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}}} \geq \frac{1}{\gamma} \quad \Longrightarrow \quad \langle \mathbf{x}_t, \boldsymbol{\theta}_t \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle .$$

Since $\mathbf{x}_t^* \neq \mathbf{0}_d$, we have $\mathbf{u}_t^* \neq \mathbf{0}_d$. Define $\bar{\mathbf{u}}_t^* := \frac{\mathbf{u}_t^*}{\|\mathbf{u}_t^*\|_2} \in \mathbb{S}^{d-1}$. Moreover, since the self-normalized ratio is invariant under positive rescaling of $\mathbf{u}_t^*$, we have

$$\frac{\langle \mathbf{s}_{t-1}^{J_t}, \mathbf{u}_t^* \rangle}{\|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}}} = \frac{\langle \mathbf{s}_{t-1}^{J_t}, \bar{\mathbf{u}}_t^* \rangle}{\|\bar{\mathbf{u}}_t^*\|_{\mathbf{V}_{t-1}}} .$$

Therefore, using the conditional uniformity of J_t given $\mathcal{A}_{t-1}$,

$$\begin{aligned} p_{t-1} &:= \mathbb{P}\left(\langle \mathbf{x}_t, \boldsymbol{\theta}_t \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle \,|\, \mathcal{A}_{t-1}\right) \\ &\geq \mathbb{P}\left(\langle \mathbf{s}_{t-1}^{J_t}, \mathbf{u}_t^* \rangle / \|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}} \geq 1/\gamma \,\middle|\, \mathcal{A}_{t-1} \right) \\ &= \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\left\{ \frac{\langle \mathbf{s}_{t-1}^{j}, \bar{\mathbf{u}}_t^* \rangle}{\|\bar{\mathbf{u}}_t^*\|_{\mathbf{V}_{t-1}}} \geq \frac{1}{\gamma} \right\} \\ &= E_{t,m}\left(\bar{\mathbf{u}}_t^*, 1/\gamma\right) . \end{aligned}$$

By Lemma 2 and Corollary 1, our choice of the ensemble-size condition gives, with probability at least $1-\delta$,

$$\min_{t \in [T]} \inf_{\mathbf{u} \in \mathbb{S}^{d-1}} E_{t,m}(\mathbf{u}, 1/\gamma) \geq \frac{1}{10} ,$$

which implies

$$\forall t \in [T] : \qquad p_{t-1} \geq \frac{1}{10} .$$

We have therefore verified the key ingredients for Lemma 10: the sampled parameter lies in a b_{t-1}-ellipsoid around $\hat{\boldsymbol{\theta}}_{t-1}$ on an event of probability at least $1-\delta$, the true parameter lies in a b_{t-1}-ellipsoid around $\hat{\boldsymbol{\theta}}_{t-1}$ on an event of probability at least $1-\delta$, and, by Lemma 2 together with Corollary 1, the conditional optimism probabilities satisfy $p_{t-1} \geq 1/10$ uniformly over t on an event of probability at least $1-\delta$. Applying Lemma 10 and intersecting with the uniform optimism event, the following final regret bound holds with probability at least $1-4\delta$:

$$\begin{aligned} \mathrm{Regret}_T &\leq 2 \max_{t \in [T]} \frac{b_{t-1}}{p_{t-1}} \left(2\sqrt{2dT \log\left(1 + \frac{T}{d\lambda}\right)} + \sqrt{2(4T/\lambda + 1) \log\left(\frac{\sqrt{4T/\lambda + 1}}{\delta} \right)} \right) \\ &\leq \widetilde{\mathcal{O}}\Big(b_T \sqrt{dT} \Big) . \end{aligned}$$

Finally, since $b_T = \gamma \Gamma_\delta \beta_T$, $\Gamma_\delta = \widetilde{\mathcal{O}}(\sqrt{d})$, and $\beta_T = \widetilde{\mathcal{O}}(\sqrt{d})$, we conclude that

$$\mathrm{Regret}_T = \widetilde{\mathcal{O}}(d^{3/2}\sqrt{T}).$$

□

B Regret Analysis in the Finite-Arm-Dominated Regime

In this section, we provide the regret analysis of our algorithm in the finite-arm-dominated regime. First, we revisit the regret analyses of previous randomized linear bandit algorithms in the finite-arm-dominated regime. Let $\widetilde{\boldsymbol{\theta}}_t$ denote the perturbed estimator used for exploration at round t. In TS [2], $\widetilde{\boldsymbol{\theta}}_t$ is given by

$$\widetilde{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} + \beta_{t-1}\mathbf{V}_{t-1}^{-1/2}\mathbf{g}_t, \qquad \mathbf{g}_t \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d).$$

Then, for each candidate arm $\mathbf{x} \in \mathcal{X}_t$ at round t, we have

$$\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1}) = \beta_{t-1}\mathbf{x}^\top\mathbf{V}_{t-1}^{-1/2}\mathbf{g}_t \sim \mathcal{N}\Big(0, \beta_{t-1}^2\|\mathbf{x}\|^2_{\mathbf{V}_{t-1}^{-1}}\Big).$$

Therefore, for each fixed arm $\mathbf{x}$, the quantity $|\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1})|$ can be controlled by a Gaussian tail bound at the scale $\beta_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}$. On the other hand, in PHE [13], $\widetilde{\boldsymbol{\theta}}_t$ is given by

$$\widetilde{\boldsymbol{\theta}}_t := \mathbf{V}_{t-1}^{-1}\left[\sum_{s=1}^{t-1}\mathbf{x}_s(y_s + Z_{t,s})\right], \qquad \{Z_{t,s}\}_{s=1}^{t-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

so that

$$\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1}) = \mathbf{x}^\top\mathbf{V}_{t-1}^{-1}\left[\sum_{s=1}^{t-1}\mathbf{x}_s Z_{t,s}\right].$$

Since the perturbations $\{Z_{t,s}\}_{s=1}^{t-1}$ are sampled freshly at round t, conditional on the history up to round $t-1$, the above quantity is again a centered Gaussian random variable with variance of order $\|\mathbf{x}\|^2_{\mathbf{V}_{t-1}^{-1}}$. Hence, for each candidate arm $\mathbf{x} \in \mathcal{X}_t$, one obtains an arm-wise concentration bound at the same self-normalized scale.

The key point is that, in both TS and PHE, the randomized score of each arm admits a clean *arm-wise* concentration bound conditional on the past. More precisely, for every $\mathbf{x} \in \mathcal{X}_t$,

$$|\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1})| \lesssim \beta_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}$$

holds with high probability for a fixed arm, and thus, by taking a union bound over the K candidate arms, one obtains

$$\max_{\mathbf{x}\in\mathcal{X}_t}|\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1})| \lesssim \beta_{t-1}\sqrt{\log K}\max_{\mathbf{x}\in\mathcal{X}_t}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}.$$

This is precisely the mechanism by which the finite-arm assumption yields a $\sqrt{\log K}$ factor instead of a $\sqrt{d}$ factor in the regret analysis. In other words, the $\log K$ improvement becomes available because the perturbation can be controlled separately for each of the K arms, and the only price for uniformity over the current arm set is a union bound over K events.

By contrast, this argument does not directly apply to ensemble sampling. In ensemble sampling, *the perturbation used at round t is not freshly generated from scratch, but is accumulated along the realized arm sequence.* More specifically, for each ensemble member j, the perturbation takes the form

$$\mathbf{s}_{t-1}^j = \sqrt{\lambda}\,\boldsymbol{\zeta}^j + \sum_{s=1}^{t-1}\mathbf{x}_s\xi_s^j,$$

and the perturbed estimator is given by

$$\widetilde{\boldsymbol{\theta}}_t^j = \hat{\boldsymbol{\theta}}_{t-1} + \gamma\beta_{t-1}\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j.$$

The key difficulty is that the arm sequence $\{\mathbf{x}_s\}_{s=1}^{t-1}$ is itself selected using the ensemble, and therefore depends on the same perturbations that define $\mathbf{s}_{t-1}^j$. As a result, after conditioning on the history up to round $t-1$, the quantity

$$\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t^j - \hat{\boldsymbol{\theta}}_{t-1}) = \gamma\beta_{t-1}\mathbf{x}^\top\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j$$

can no longer be viewed as a centered Gaussian random variable with variance proportional to $\|\mathbf{x}\|^2_{\mathbf{V}_{t-1}^{-1}}$. In other words, the perturbation and the current Gram matrix are coupled through the past trajectory, and thus the simple arm-wise conditional concentration argument used in TS and PHE breaks down.

This distinction is crucial. In TS and PHE, the perturbation at round t remains conditionally well-behaved for each arm after fixing the past, so a union bound over the K candidate arms yields a $\sqrt{\log K}$ factor. In ensemble sampling, however, the perturbation is persistent and evolves together with the action sequence that it helps generate. Therefore, the projected perturbation along a given arm does not admit the same clean arm-wise tail bound after conditioning on the past. Consequently, one cannot directly obtain

$$\max_{\mathbf{x}\in\mathcal{X}} |\mathbf{x}^\top(\widetilde{\boldsymbol{\theta}}_t^j - \hat{\boldsymbol{\theta}}_{t-1})| \lesssim \beta_{t-1}\sqrt{\log K}\, \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}$$

by the same argument as in TS or PHE.

Instead, previous analyses of ensemble sampling control the perturbation in a more global manner, for example through bounds on self-normalized directional processes or on the vector norm $\|\mathbf{V}_{t-1}^{-1/2}\mathbf{s}_{t-1}^j\|$. Such bounds are sufficient to establish optimism and sublinear regret, but they naturally lead to a $\sqrt{d}$-scale complexity rather than a $\sqrt{\log K}$ one. This is the main reason why the standard finite-arm $\log K$ improvement available in TS and PHE does not automatically carry over to ensemble sampling.

However, our epoch-refresh construction makes an arm-wise analysis possible. Within each epoch, the refresh rule ensures that the Gram matrix remains close to its epoch-start value, so the self-normalized perturbation score of each fixed arm changes only mildly during the epoch. This allows us to control the perturbations directly over the finite set of K arms, rather than uniformly over all directions. Lemma 3 formalizes this idea by showing that the ensemble perturbations admit a uniform arm-wise concentration bound at essentially the $\sqrt{\log K}$ scale, which is the key ingredient for the regret analysis for the finite-arm-dominated regime.

B.1 Proof of Lemma 3

Lemma 3 (Uniform arm-wise concentration)**.** *Suppose Assumption 1 holds, and suppose that the arm sets $\{\mathcal{X}_t\}_{t=1}^T$ are chosen by an oblivious adversary with $|\mathcal{X}_t| = K \geq 2$ for all $t \in [T]$. For any $\alpha > 0$ and $\delta \in (0,1)$, let $L_\alpha := \log \frac{8KmT\bar{R}_\alpha}{\delta}$. Then, with probability at least $1-\delta$, we have*

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left| \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}} \right| \leq C\left[\sqrt{L_\alpha} + \sqrt{\alpha L_\alpha} + \alpha(\sqrt{d} + \sqrt{L_\alpha}) + \alpha\sqrt{\alpha(d+L_\alpha)}\right],$$

where $C > 0$ is a universal constant. In particular, if a deterministic quantity $\Lambda_* \geq 1$ satisfies $L_\alpha \leq C_\Lambda \Lambda_*$ for some universal constant $C_\Lambda > 0$, and if $\alpha = \min\{1, \sqrt{\Lambda_*/d}\}$, then we have

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left| \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}} \right| \leq \widetilde{C}\sqrt{\Lambda_*},$$

where $\widetilde{C} > 0$ is a universal constant.

Proof of Lemma 3. Fix an epoch r and a round $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$ played in epoch r. Recall that by construction,

$$\mathbf{s}_{t-1}^j = \mathbf{V}_{\tau_r}^{1/2}\mathbf{g}_r^j + \sum_{s=\tau_r+1}^{t-1} \mathbf{x}_s \xi_s^j.$$

Fix a candidate arm $\mathbf{x} \in \mathcal{X}_t$ at round t. The case $\mathbf{x} = \mathbf{0}_d$ is trivial, so we assume $\mathbf{x} \neq \mathbf{0}_d$. We define

$$\mathbf{A}_{r,t} := \mathbf{V}_{\tau_r}^{-1/2}\mathbf{V}_{t-1}\mathbf{V}_{\tau_r}^{-1/2}, \qquad \mathbf{y}_{r,t}(\mathbf{x}) := \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}, \qquad \mathbf{n}_{r,t-1}^j := \mathbf{V}_{\tau_r}^{-1/2}\sum_{s=\tau_r+1}^{t-1} \mathbf{x}_s\xi_s^j.$$

Then we have

$$\mathbf{V}_{\tau_r}^{-1/2}\mathbf{s}_{t-1}^j = \mathbf{g}_r^j + \mathbf{n}_{r,t-1}^j.$$

Moreover, since $\mathbf{V}_{t-1}^{-1} = \mathbf{V}_{\tau_r}^{-1/2}\mathbf{A}_{r,t}^{-1}\mathbf{V}_{\tau_r}^{-1/2}$, for $\mathbf{x} \neq \mathbf{0}_d$, we have

$$Z_{r,t}^j(\mathbf{x}) := \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}} = \frac{\mathbf{y}_{r,t}(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1}(\mathbf{g}_r^j + \mathbf{n}_{r,t-1}^j)}{\sqrt{\mathbf{y}_{r,t}(\mathbf{x})^\top \mathbf{A}_{r,t}^{-1}\mathbf{y}_{r,t}(\mathbf{x})}}\,.$$

In the following proof, we establish a uniform bound on $|Z_{r,t}^j(\mathbf{x})|$ over all $t \in [T]$, $j \in [m]$, and $\mathbf{x} \in \mathcal{X}_t$.

Note that since t is played in epoch r, the refresh rule guarantees

$$\mathbf{I}_d \preceq \mathbf{A}_{r,t} \preceq (1+\alpha)\mathbf{I}_d\,.$$

Next, we introduce a lemma which quantifies how little the self-normalized projection changes when the Gram matrix drifts only mildly within an epoch. In particular, when $\mathbf{A}$ remains within a multiplicative factor $1+\alpha$ of the identity, the normalized linear functional induced by $\mathbf{A}^{-1}$ stays uniformly close to the corresponding Euclidean projection. A proof of Lemma 11 is provided in Appendix C.2.

Lemma 11 (Stability of normalized projections under small matrix drift)**.** *Let $\alpha \geq 0$, and suppose*

$$\mathbf{I}_d \preceq \mathbf{A} \preceq (1+\alpha)\mathbf{I}_d\,.$$

Then, for every $\mathbf{y} \neq \mathbf{0}_d$ and every $\mathbf{v} \in \mathbb{R}^d$,

$$\left|\frac{\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{v}}{\sqrt{\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y}}} - \left\langle \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \mathbf{v}\right\rangle\right| \leq 2\alpha\|\mathbf{v}\|_2\,.$$

Applying Lemma 11 with

$$\mathbf{y} = \mathbf{y}_{r,t}(\mathbf{x}), \qquad \mathbf{A} = \mathbf{A}_{r,t}, \qquad \mathbf{v} = \mathbf{g}_r^j + \mathbf{n}_{r,t-1}^j,$$

we obtain

$$|Z_{r,t}^j(\mathbf{x})| \leq \left|\left\langle \frac{\mathbf{y}_{r,t}(\mathbf{x})}{\|\mathbf{y}_{r,t}(\mathbf{x})\|_2}, \mathbf{g}_r^j\right\rangle\right| + \left|\left\langle \frac{\mathbf{y}_{r,t}(\mathbf{x})}{\|\mathbf{y}_{r,t}(\mathbf{x})\|_2}, \mathbf{n}_{r,t-1}^j\right\rangle\right| + 2\alpha\|\mathbf{g}_r^j\|_2 + 2\alpha\|\mathbf{n}_{r,t-1}^j\|_2\,. \tag{9}$$

We now control the four terms on the right-hand side of Eq.(9).

Term 1: epoch-start Gaussian projection. Define

$$\mathbf{u}_{r,t}(\mathbf{x}) := \frac{\mathbf{y}_{r,t}(\mathbf{x})}{\|\mathbf{y}_{r,t}(\mathbf{x})\|_2}.$$

We first control the epoch-start Gaussian projection. By the oblivious-arm-set assumption, the sequence $\{\mathcal{X}_t\}_{t=1}^T$ is generated independently of the algorithm's internal randomization. Moreover, the epoch-start Gaussian perturbation $\mathbf{g}_r^j$ is sampled independently of the history available at the beginning of epoch r. Hence, for each nonzero $\mathbf{x} \in \mathcal{X}_t$, the direction $\mathbf{u}_{r,t}(\mathbf{x})$ is independent of $\mathbf{g}_r^j$. Since $\|\mathbf{u}_{r,t}(\mathbf{x})\|_2 = 1$, it follows that

$$\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{g}_r^j\rangle \sim \mathcal{N}(0,1)\,.$$

Thus, for every $z > 0$,

$$\mathbb{P}\big(\big|\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{g}_r^j\rangle\big| \geq z\big) \leq 2\exp(-z^2/2)\,.$$

Taking a union bound over epochs $r \leq R_\alpha$, rounds $t \in [T]$, ensemble members $j \in [m]$, and arms $\mathbf{x} \in \mathcal{X}_t$, and using $|\mathcal{X}_t| \leq K$ and $R_\alpha \leq \bar{R}_\alpha$, we obtain

$$\mathbb{P}\left(\max_{r\leq R_\alpha,\, t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \big|\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{g}_r^j\rangle\big| \geq z\right) \leq 2KmT\bar{R}_\alpha \exp(-z^2/2)\,.$$

Therefore, taking

$$z = \sqrt{2\log\left(\frac{8KmT\bar{R}_\alpha}{\delta}\right)}$$

gives, with probability at least $1 - \delta/4$,

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left|\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{g}_{r(t)}^j\rangle\right| \leq \sqrt{2\log\left(\frac{8KmT\bar{R}_\alpha}{\delta}\right)} = \sqrt{2L_\alpha}\,. \tag{E1}$$

Term 2: scalar within-epoch martingale. Fix an epoch r, an ensemble index $j \in [m]$, a round $t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$, and a nonzero candidate arm $\mathbf{x} \in \mathcal{X}_t$. Define

$$\mathcal{M}^j_{r,\tau_r,t}(\mathbf{x}) := 0, \qquad \mathcal{M}^j_{r,q,t}(\mathbf{x}) := \sum_{s=\tau_r+1}^{q} \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}_s \right\rangle \xi^j_s, \quad q \in \{\tau_r + 1, \dots, t-1\}.$$

Then

$$\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{n}^j_{r,t-1} \rangle = \mathcal{M}^j_{r,t-1,t}(\mathbf{x}).$$

We now specify the filtration used in the martingale argument. Let

$$\mathcal{G}^{\mathcal{X}}_t := \sigma(\mathcal{X}_1, \dots, \mathcal{X}_t)$$

be the sigma-algebra generated by the sequence of arm sets. By the oblivious-arm-set assumption, $\mathcal{G}^{\mathcal{X}}_t$ is independent of the algorithm's internal randomization. For $q \in \{\tau_r, \dots, t-1\}$, define

$$\mathcal{F}^{(r,t)}_q := \mathcal{G}^{\mathcal{X}}_t \vee \sigma(\mathbf{g}^1_\ell, \dots, \mathbf{g}^m_\ell : \ell \le r) \vee \sigma(\mathbf{x}_s, y_s, J_s : s \le q) \vee \sigma(\xi^1_s, \dots, \xi^m_s : s < q)\,.$$

This filtration contains the realized arm set $\mathcal{X}_t$, the selected arm $\mathbf{x}_q$, and all perturbation noises generated before round q, but not ξ^j_q. Hence, by the oblivious-arm-set assumption and the independence of the perturbation noises, $\xi^j_q \sim \mathcal{N}(0,1)$ is independent of $\mathcal{F}^{(r,t)}_q$.

On the other hand, since $\mathbf{x}_q$ is $\mathcal{F}^{(r,t)}_q$-measurable, therefore the scalar $a_{r,q,t}(\mathbf{x}) := \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}_q \right\rangle$ is also $\mathcal{F}^{(r,t)}_q$-measurable.

Define the enlarged filtration

$$\widehat{\mathcal{F}}^{(r,t)}_{\tau_r} := \mathcal{F}^{(r,t)}_{\tau_r}, \qquad \widehat{\mathcal{F}}^{(r,t)}_q := \mathcal{F}^{(r,t)}_q \vee \sigma(\xi^1_q, \dots, \xi^m_q), \quad q \in \{\tau_r + 1, \dots, t-1\}.$$

Then

$$\widehat{\mathcal{F}}^{(r,t)}_{q-1} \subseteq \mathcal{F}^{(r,t)}_q \subseteq \widehat{\mathcal{F}}^{(r,t)}_q.$$

Furthermore, we have

$$\begin{aligned}
\mathcal{M}^j_{r,q,t}(\mathbf{x}) &= \sum_{s=\tau_r+1}^{q} \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}_s \right\rangle \xi^j_s \\
&= \sum_{s=\tau_r+1}^{q-1} \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}_s \right\rangle \xi^j_s + \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}_q \right\rangle \xi^j_q \\
&= \mathcal{M}^j_{r,q-1,t}(\mathbf{x}) + a_{r,q,t}(\mathbf{x})\xi^j_q\,.
\end{aligned}$$

Since ξ^j_q is independent of $\mathcal{F}^{(r,t)}_q$ and has mean zero,

$$\mathbb{E}\Big[\mathcal{M}^j_{r,q,t}(\mathbf{x}) \mid \mathcal{F}^{(r,t)}_q\Big] = \mathcal{M}^j_{r,q-1,t}(\mathbf{x}) + a_{r,q,t}(\mathbf{x})\mathbb{E}\left[\xi^j_q \mid \mathcal{F}^{(r,t)}_q\right] = \mathcal{M}^j_{r,q-1,t}(\mathbf{x})\,.$$

By the tower property,

$$\mathbb{E}\Big[\mathcal{M}^j_{r,q,t}(\mathbf{x}) \mid \widehat{\mathcal{F}}^{(r,t)}_{q-1}\Big] = \mathbb{E}\left[\mathbb{E}\Big[\mathcal{M}^j_{r,q,t}(\mathbf{x}) \mid \mathcal{F}^{(r,t)}_q\Big] \mid \widehat{\mathcal{F}}^{(r,t)}_{q-1}\right] = \mathcal{M}^j_{r,q-1,t}(\mathbf{x}).$$

Thus, $\{\mathcal{M}^j_{r,q,t}(\mathbf{x})\}_{q=\tau_r}^{t-1}$ is a martingale with conditionally Gaussian increments with respect to $\{\widehat{\mathcal{F}}^{(r,t)}_q\}_{q=\tau_r}^{t-1}$. We also define the variance process

$$A_{r,\tau_r,t}(\mathbf{x}) := 0, \qquad A_{r,q,t}(\mathbf{x}) := \sum_{s=\tau_r+1}^{q} \left(a_{r,s,t}(\mathbf{x})\right)^2, \quad q \in \{\tau_r + 1, \dots, t-1\}.$$

Since t belongs to epoch r, for every $q \le t-1$,

$$\mathbf{V}_q \preceq (1+\alpha)\mathbf{V}_{\tau_r}.$$

Hence for any $q \in \{\tau_r + 1, \dots, t-1\}$, we obtain

$$A_{r,q,t}(\mathbf{x}) = \sum_{s=\tau_r+1}^{q} \left(\left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{V}_{\tau_r}^{-1/2} \mathbf{x}_s \right\rangle \right)^2$$
$$= \mathbf{u}_{r,t}(\mathbf{x})^\top \mathbf{V}_{\tau_r}^{-1/2} (\mathbf{V}_q - \mathbf{V}_{\tau_r}) \mathbf{V}_{\tau_r}^{-1/2} \mathbf{u}_{r,t}(\mathbf{x}) \le \alpha \,. \tag{10}$$

Now, for a fixed $\lambda_0 > 0$, we define the exponential process

$$\mathcal{E}_q(\lambda_0) := \exp\left(\lambda_0 \mathcal{M}_{r,q,t}^j(\mathbf{x}) - \frac{\lambda_0^2}{2} A_{r,q,t}(\mathbf{x}) \right), \qquad q \in \{\tau_r, \dots, t-1\} \,.$$

We claim that $\{\mathcal{E}_q(\lambda_0)\}_{q=\tau_r}^{t-1}$ is a nonnegative martingale with respect to $\{\widehat{\mathcal{F}}_q^{(r,t)}\}_{q=\tau_r}^{t-1}$. Indeed, $\mathcal{M}_{r,q,t}^j(\mathbf{x})$ is $\hat{\mathcal{F}}_q^{(r,t)}$-measurable by construction. Moreover, since $\mathcal{M}_{r,q-1,t}^j(\mathbf{x})$ and $a_{r,q,t}(\mathbf{x})$ are $\mathcal{F}_q^{(r,t)}$-measurable and $\xi_q^j \sim \mathcal{N}(0,1)$ is independent of $\mathcal{F}_q^{(r,t)}$, we have

$$\mathbb{E}\left[\mathcal{E}_q(\lambda_0) \mid \mathcal{F}_q^{(r,t)} \right] = \mathcal{E}_{q-1}(\lambda_0) \mathbb{E}\left[\exp\left(\lambda_0 a_{r,q,t}(\mathbf{x}) \xi_q^j - \frac{\lambda_0^2}{2} (a_{r,q,t}(\mathbf{x}))^2 \right) \mid \mathcal{F}_q^{(r,t)} \right] = \mathcal{E}_{q-1}(\lambda_0)$$

where the last equality follows from the Gaussian moment generating function. Then, the tower property again yields

$$\mathbb{E}\left[\mathcal{E}_q(\lambda_0) \mid \hat{\mathcal{F}}_{q-1}^{(r,t)} \right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{E}_q(\lambda_0) \mid \mathcal{F}_q^{(r,t)} \right] \mid \hat{\mathcal{F}}_{q-1}^{(r,t)} \right] = \mathcal{E}_{q-1}(\lambda_0) \,.$$

Thus, $\{\mathcal{E}_q(\lambda_0)\}_{q=\tau_r}^{t-1}$ is a nonnegative martingale. Now, for any $z \ge 0$, we define an event

$$E_z := \left\{ \max_{q \in \{\tau_r+1, \dots, t-1\}} \mathcal{M}_{r,q,t}^j(\mathbf{x}) \ge z \right\} .$$

On E_z, there exists $q^* \in \{\tau_r + 1, \dots, t-1\}$ such that $\mathcal{M}_{r,q^*,t}^j \ge z$. By Eq. (10), we have

$$A_{r,q^*,t}(\mathbf{x}) \le \alpha \,.$$

Therefore,

$$\mathcal{E}_{q^*}(\lambda_0) = \exp\left(\lambda_0 \mathcal{M}_{r,q^*,t}^j(\mathbf{x}) - \frac{\lambda_0^2}{2} A_{r,q^*,t}(\mathbf{x}) \right) \ge \exp\left(\lambda_0 z - \frac{\lambda_0^2}{2} \alpha \right) .$$

Hence, we have the following relationship:

$$E_z \subset \left\{ \max_{q \in \{\tau_r, \dots, t-1\}} \mathcal{E}_q(\lambda_0) \ge \exp\left(\lambda_0 z - \frac{\lambda_0^2}{2} \alpha \right) \right\} .$$

By Ville's inequality, we have

$$\mathbb{P}(E_z) \le \mathbb{P}\left(\max_{q \in \{\tau_r, \dots, t-1\}} \mathcal{E}_q(\lambda_0) \ge \exp\left(\lambda_0 z - \frac{\lambda_0^2}{2} \alpha \right) \right) \le \exp\left(-\lambda_0 z + \frac{\lambda_0^2}{2} \alpha \right) . \tag{11}$$

Note that Eq. (11) holds for any $\lambda_0 > 0$, for specific choicse $\lambda_0 = z/\alpha$, we have

$$\mathbb{P}(E_z) \le \exp\left(-\frac{z^2}{2\alpha} \right) .$$

Applying the same argument to $-\mathcal{M}_{r,q,t}^j(\mathbf{x})$, we obtain

$$\mathbb{P}\left(\left| \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{n}_{r,t-1}^j \right\rangle \right| \ge z \right) \le \mathbb{P}\left(\max_{q \in \{\tau_r, \dots, t-1\}} \left| \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{n}_{r,q}^j \right\rangle \right| \ge z \right) \le 2 \exp\left(-\frac{z^2}{2\alpha} \right) . \tag{12}$$

Finally, taking a union bound over epochs $r \le R_\alpha$, rounds $t \in [T]$, ensemble members $j \in [m]$, and arms $\mathbf{x} \in \mathcal{X}_t$, and using $|\mathcal{X}_t| \le K$ and $R_\alpha \le \bar{R}_\alpha$, we get

$$\mathbb{P}\left(\max_{\substack{r \le R_\alpha,\ t \in [T], \\ j \in [m],\ \mathbf{x} \in \mathcal{X}_t}} \left| \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{n}_{r,t-1}^j \right\rangle \right| \ge z \right) \le 2KmT\bar{R}_\alpha \exp\left(-\frac{z^2}{2\alpha} \right) .$$

Thus, taking $z = \sqrt{2\alpha L_\alpha}$ yields, with probability at least $1 - \delta/4$,

$$\max_{t \in [T],\, j \in [m],\, \mathbf{x} \in \mathcal{X}_t} \left| \left\langle \mathbf{u}_{r,t}(\mathbf{x}), \mathbf{n}_{r,t-1}^j \right\rangle \right| \le \sqrt{2\alpha L_\alpha} \,. \tag{E2}$$

Term 3: epoch-start Gaussian norm. Fix an epoch $r \le \bar{R}_\alpha$ and an ensemble index $j \in [m]$. Since $\mathbf{g}_r^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, the standard concentration inequality for the Euclidean norm of a Gaussian vector (Lemma 15) gives, for every $u > 0$,

$$\mathbb{P}\Big(\|\mathbf{g}_r^j\|_2 \ge \sqrt{d} + \sqrt{2u}\Big) \le \exp(-u)\,.$$

Taking a union bound over all epochs $r \le \bar{R}_\alpha$ and ensemble members $j \in [m]$, we obtain

$$\mathbb{P}\left(\max_{r \le \bar{R}_\alpha,\, j \in [m]} \|\mathbf{g}_r^j\|_2 \ge \sqrt{d} + \sqrt{2u}\right) \le m\bar{R}_\alpha \exp(-u)\,.$$

Choosing

$$u := \log \frac{4m\bar{R}_\alpha}{\delta}$$

gives $m\bar{R}_\alpha \exp(-u) = \delta/4$. Hence, with probability at least $1 - \delta/4$,

$$\max_{r \le \bar{R}_\alpha,\, j \in [m]} \|\mathbf{g}_r^j\|_2 \le \sqrt{d} + \sqrt{2\log\left(\frac{4m\bar{R}_\alpha}{\delta}\right)} \le \sqrt{d} + \sqrt{2L_\alpha}\,. \tag{E3}$$

Term 4: vector within-epoch martingale. Fix an epoch r and an ensemble index $j \in [m]$. Define

$$\mathbf{a}_{r,q} := \mathbf{V}_{\tau_r}^{-1/2}\mathbf{x}_q, \qquad q \in \{\tau_r + 1, \dots, \tau_{r+1} - 1\}.$$

Then, for every $t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$,

$$\mathbf{n}_{r,t-1}^j = \mathbf{V}_{\tau_r}^{-1/2} \sum_{s=\tau_r+1}^{t-1} \mathbf{x}_s \xi_s^j = \sum_{s=\tau_r+1}^{t-1} \mathbf{a}_{r,s}\xi_s^j.$$

Moreover,

$$\sum_{s=\tau_r+1}^{t-1} \mathbf{a}_{r,s}\mathbf{a}_{r,s}^\top = \mathbf{V}_{\tau_r}^{-1/2}(\mathbf{V}_{t-1} - \mathbf{V}_{\tau_r})\mathbf{V}_{\tau_r}^{-1/2} \preceq \alpha \mathbf{I}_d,$$

because t belongs to epoch r and hence $\mathbf{V}_{t-1} \preceq (1+\alpha)\mathbf{V}_{\tau_r}$. Consequently, for every $\mathbf{u} \in \mathbb{S}^{d-1}$,

$$\sum_{s=\tau_r+1}^{t-1} \big(\langle \mathbf{u}, \mathbf{a}_{r,s}\rangle\big)^2 \le \alpha.$$

Applying the scalar martingale bound from Term 2 in Eq. (11) to the process

$$\sum_{s=\tau_r+1}^{t-1} \langle \mathbf{u}, \mathbf{a}_{r,s}\rangle \xi_s^j = \langle \mathbf{u}, \mathbf{n}_{r,t-1}^j\rangle,$$

we obtain, for every fixed $\mathbf{u} \in \mathbb{S}^{d-1}$ and every $z \ge 0$,

$$\mathbb{P}\left(\max_{t \in \{\tau_r+1, \dots, \tau_{r+1}\}} \left|\langle \mathbf{u}, \mathbf{n}_{r,t-1}^j\rangle\right| \ge z\right) \le 2\exp\left(-\frac{z^2}{2\alpha}\right). \tag{13}$$

Let $\mathcal{N}_{1/2} \subset \mathbb{S}^{d-1}$ be a $1/2$-net satisfying $|\mathcal{N}_{1/2}| \le 5^d$. We use the standard net argument: for every $\mathbf{v} \in \mathbb{R}^d$,

$$\|\mathbf{v}\|_2 \le 2 \max_{\mathbf{u} \in \mathcal{N}_{1/2}} |\langle \mathbf{u}, \mathbf{v}\rangle|.$$

Indeed, the claim is trivial if $\mathbf{v} = \mathbf{0}_d$. Otherwise, let $\mathbf{w} := \mathbf{v}/\|\mathbf{v}\|_2$. By the definition of the net, there exists $\mathbf{u}^* \in \mathcal{N}_{1/2}$ such that $\|\mathbf{u}^* - \mathbf{w}\|_2 \le 1/2$. Then

$$|\langle \mathbf{u}^*, \mathbf{v}\rangle| \ge |\langle \mathbf{w}, \mathbf{v}\rangle| - |\langle \mathbf{u}^* - \mathbf{w}, \mathbf{v}\rangle| \ge \|\mathbf{v}\|_2 - \frac{1}{2}\|\mathbf{v}\|_2 = \frac{1}{2}\|\mathbf{v}\|_2,$$

which proves the claim.

Applying this claim with $\mathbf{v} = \mathbf{n}^j_{r,t-1}$, we have

$$\|\mathbf{n}^j_{r,t-1}\|_2 \le 2 \max_{\mathbf{u}\in\mathcal{N}_{1/2}} \left|\langle \mathbf{u}, \mathbf{n}^j_{r,t-1}\rangle\right|.$$

Therefore,

$$\left\{\max_{t\in\{\tau_r+1,\dots,\tau_{r+1}\}} \|\mathbf{n}^j_{r,t-1}\|_2 \ge 2z\right\} \subseteq \bigcup_{\mathbf{u}\in\mathcal{N}_{1/2}} \left\{\max_{t\in\{\tau_r+1,\dots,\tau_{r+1}\}} \left|\langle \mathbf{u}, \mathbf{n}^j_{r,t-1}\rangle\right| \ge z\right\}.$$

Using Eq. (13) and taking a union bound over $\mathcal{N}_{1/2}$, we get

$$\mathbb{P}\left(\max_{t\in\{\tau_r+1,\dots,\tau_{r+1}\}} \|\mathbf{n}^j_{r,t-1}\|_2 \ge 2z\right) \le 2|\mathcal{N}_{1/2}| \exp\left(-\frac{z^2}{2\alpha}\right) \le 2\cdot 5^d \exp\left(-\frac{z^2}{2\alpha}\right).$$

Taking another union bound over all $r \le \bar{R}_\alpha$ and $j \in [m]$, we obtain

$$\mathbb{P}\left(\max_{\substack{r\le \bar{R}_\alpha,\, j\in[m],\\ t\in\{\tau_r+1,\dots,\tau_{r+1}\}}} \|\mathbf{n}^j_{r,t-1}\|_2 \ge 2z\right) \le 2m\bar{R}_\alpha 5^d \exp\left(-\frac{z^2}{2\alpha}\right).$$

Choosing

$$z^2 = 2\alpha\left(d\log 5 + \log\frac{8m\bar{R}_\alpha}{\delta}\right),$$

we conclude that, with probability at least $1-\delta/4$,

$$\max_{\substack{r\le \bar{R}_\alpha,\, j\in[m],\\ t\in\{\tau_r+1,\dots,\tau_{r+1}\}}} \|\mathbf{n}^j_{r,t-1}\|_2 \le 2\sqrt{2\alpha\left(d\log 5 + \log\frac{8m\bar{R}_\alpha}{\delta}\right)}.$$

Finally, since $K, T \ge 1$ and $L_\alpha = \log\frac{8KmT\bar{R}_\alpha}{\delta}$, the logarithmic factor above is bounded by $d\log 5 + L_\alpha$. Hence, for a universal constant $C > 0$,

$$\max_{\substack{r\le \bar{R}_\alpha,\, j\in[m],\\ t\in\{\tau_r+1,\dots,\tau_{r+1}\}}} \|\mathbf{n}^j_{r,t-1}\|_2 \le C\sqrt{\alpha(d+L_\alpha)}. \tag{E4}$$

Combining the four events. By a union bound, the events in Eqs. (E1)–(E4) hold simultaneously with probability at least $1-\delta$. On this event, Eq. (9) implies that, for all $t \in [T]$, $j \in [m]$, and nonzero $\mathbf{x} \in \mathcal{X}_t$,

$$\left|\frac{\mathbf{x}^\top \mathbf{V}^{-1}_{t-1}\mathbf{s}^j_{t-1}}{\|\mathbf{x}\|_{\mathbf{V}^{-1}_{t-1}}}\right| \le C\left[\sqrt{L_\alpha} + \sqrt{\alpha L_\alpha} + \alpha(\sqrt{d}+\sqrt{L_\alpha}) + \alpha\sqrt{\alpha(d+L_\alpha)}\right], \tag{14}$$

where $C > 0$ is a universal constant. This proves the first part of the theorem.

For the second part, suppose that a deterministic quantity $\Lambda_* \ge 1$ satisfies $L_\alpha \le C_\Lambda \Lambda_*$ for some universal constant $C_\Lambda > 0$, and choose $\alpha = \min\{1, \sqrt{\Lambda_*/d}\}$. We show that every term on the right-hand side of Eq. (14) is bounded by a universal constant times $\sqrt{\Lambda_*}$.

First, since $L_\alpha \le C_\Lambda\Lambda_*$, we have

$$\sqrt{L_\alpha} \le \sqrt{C_\Lambda}\sqrt{\Lambda_*}.$$

Also, since $\alpha \le 1$,

$$\sqrt{\alpha L_\alpha} \le \sqrt{L_\alpha} \le \sqrt{C_\Lambda}\sqrt{\Lambda_*}.$$

Next, we control the drift term $\alpha(\sqrt{d}+\sqrt{L_\alpha})$. By the definition of α, we have

$$\alpha\sqrt{d} = \min\left\{\sqrt{d}, \sqrt{\Lambda_*}\right\} \le \sqrt{\Lambda_*}.$$

Moreover,

$$\alpha\sqrt{L_\alpha} \le \sqrt{L_\alpha} \le \sqrt{C_\Lambda}\sqrt{\Lambda_*}.$$

Therefore,

$$\alpha(\sqrt{d} + \sqrt{L_\alpha}) \le (1 + \sqrt{C_\Lambda})\sqrt{\Lambda_*}.$$

It remains to bound the last term. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we get

$$\alpha\sqrt{\alpha(d + L_\alpha)} \le \alpha\sqrt{\alpha d} + \alpha\sqrt{\alpha L_\alpha}.$$

Since $\alpha \le 1$,

$$\alpha\sqrt{\alpha d} \le \alpha\sqrt{d} \le \sqrt{\Lambda_*},$$

and similarly,

$$\alpha\sqrt{\alpha L_\alpha} \le \sqrt{L_\alpha} \le \sqrt{C_\Lambda}\sqrt{\Lambda_*}.$$

Thus,

$$\alpha\sqrt{\alpha(d + L_\alpha)} \le (1 + \sqrt{C_\Lambda})\sqrt{\Lambda_*}.$$

Combining the above bounds in Eq. (14), and absorbing universal constants into a new constant $\widetilde{C} > 0$, we obtain

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left| \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1} \mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}} \right| \le \widetilde{C}\sqrt{\Lambda_*}.$$

□

B.2 Proof of Theorem 2

In this section, we provide a regret bound for Algorithm 1 in the finite-arm setting.

Theorem 2 (Regret in the finite-arm-dominated regime)**.** *Suppose Assumption 1 holds, and suppose that the arm sets $\{\mathcal{X}_t\}_{t=1}^T$ are chosen by an oblivious adversary with $|\mathcal{X}_t| = K \ge 2$ for all $t \in [T]$. Fix $\delta \in (0, 1/4)$, and choose $\lambda \ge 80$, $\gamma = 40$, and $\{\beta_t\}_{t\ge 0}$ as in Theorem 1. For $\alpha_0 = 1/\sqrt{d}$ and $\bar{R}_0 := \bar{R}_{\alpha_0}$, define*

$$M_0 = \left\lceil C_m \left[d \log\left(d + \log \frac{\bar{R}_0}{\delta} \right) + \log \frac{\bar{R}_0}{\delta} \right] \right\rceil, \quad \Lambda_0 := \max\left\{ 1,\, C_\Lambda \log\left(\frac{K M_0 T \bar{R}_0}{\delta} \right) \right\},$$

for some constants $C_m, C_\Lambda > 0$. Choose $\alpha := \min\{1, \sqrt{\Lambda_0/d}\}$ and $m := M_0$. Then, with probability at least $1 - 4\delta$, Algorithm 1 incurs cumulative regret $\mathrm{Regret}_T = \widetilde{\mathcal{O}}(d\sqrt{T \log K})$.

Proof of Theorem 2. We first verify that the parameter choices stated in Theorem 2 satisfy the technical requirements needed for the proof. Since $\Lambda_0 \ge 1$, we have

$$\alpha = \min\left\{ 1, \sqrt{\frac{\Lambda_0}{d}} \right\} \ge \frac{1}{\sqrt{d}} = \alpha_0.$$

Since the deterministic epoch bound $\bar{R}_\alpha$ is nonincreasing in α, we have

$$\bar{R}_\alpha \le \bar{R}_{\alpha_0} = \bar{R}_0.$$

Therefore, by the definition of M_0 and Corollary 1, the choice $m = M_0$ is sufficient for the uniform exceedance guarantee of Lemma 2 for the actual value of α.

Moreover, since $m = M_0$ and $\bar{R}_\alpha \le \bar{R}_0$, the logarithmic factor in Lemma 3 satisfies

$$L_\alpha = \log \frac{8KmT\bar{R}_\alpha}{\delta} \le \log \frac{8KM_0T\bar{R}_0}{\delta} \le C\Lambda_0$$

for a universal constant $C > 0$, by the definition of Λ_0. Since $\alpha = \min\{1, \sqrt{\Lambda_0/d}\}$, the second part of Lemma 3, applied with $\Lambda_* = \Lambda_0$, gives

$$\max_{t\in[T],\, j\in[m],\, \mathbf{x}\in\mathcal{X}_t} \left| \frac{\mathbf{x}^\top \mathbf{V}_{t-1}^{-1} \mathbf{s}_{t-1}^j}{\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}} \right| \le \widetilde{C}\sqrt{\Lambda_0}$$

with probability at least $1 - \delta$, where $\widetilde{C} > 0$ is a universal constant.

We now introduce the finite-arm version of the randomized-optimism master regret bound.

Lemma 12 (Finite-arm regret bound for randomized algorithms)**.** *Suppose that the arm sets $\{\mathcal{X}_t\}_{t=1}^T$ are chosen by an oblivious adversary with $|\mathcal{X}_t| = K$ for all $t \in [T]$. Under the assumptions of Lemma 10, the same conclusion continues to hold if the event*

$$\mathcal{E} = \left\{\forall t \le T : \|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \le b_{t-1}\right\}$$

is replaced by the arm-wise event

$$\widetilde{\mathcal{E}} = \left\{\forall t \le T,\ \forall \mathbf{x} \in \mathcal{X}_t : \left|\mathbf{x}^\top(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1})\right| \le b_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}\right\}.$$

Define the following events:

$$\begin{aligned}
\mathcal{E}_{\mathrm{LS}} &:= \left\{\forall t \in [T] : \|\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\|_{\mathbf{V}_{t-1}} \le \beta_{t-1}\right\}, \\
\mathcal{E}_{\mathrm{arm}} &:= \left\{\forall t \in [T],\ \forall j \in [m],\ \forall \mathbf{x} \in \mathcal{X}_t : \left|\mathbf{x}^\top \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j\right| \le \widetilde{C}\sqrt{\Lambda_0}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}\right\}, \\
\mathcal{E}_{\mathrm{exc}} &:= \left\{\forall t \in [T] : \inf_{\mathbf{u} \in \mathbb{S}^{d-1}} E_{t,m}(\mathbf{u}, 1/\gamma) \ge \frac{1}{10}\right\}.
\end{aligned}$$

By Lemma 9, Lemma 3, and Lemma 2, respectively,

$$\mathbb{P}(\mathcal{E}_{\mathrm{LS}}) \ge 1 - \delta, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{arm}}) \ge 1 - \delta, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{exc}}) \ge 1 - \delta.$$

For each $j \in [m]$, define the j-th ensemble parameter at round t by

$$\boldsymbol{\theta}_t^j := \hat{\boldsymbol{\theta}}_{t-1} + \gamma\beta_{t-1}\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j.$$

On $\mathcal{E}_{\mathrm{arm}}$, for all $t \in [T]$, $j \in [m]$, and $\mathbf{x} \in \mathcal{X}_t$,

$$\left|\mathbf{x}^\top(\boldsymbol{\theta}_t^j - \hat{\boldsymbol{\theta}}_{t-1})\right| = \gamma\beta_{t-1}\left|\mathbf{x}^\top\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j\right| \le \gamma\widetilde{C}\sqrt{\Lambda_0}\,\beta_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}.$$

Define

$$b_{t-1} := \left(1 + \gamma\widetilde{C}\sqrt{\Lambda_0}\right)\beta_{t-1}.$$

Then $b_{t-1} \ge \beta_{t-1}$. Hence, on $\mathcal{E}_{\mathrm{LS}}$,

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \le \beta_{t-1} \le b_{t-1}$$

for all $t \in [T]$. Moreover, on $\mathcal{E}_{\mathrm{arm}}$, the arm-wise event $\widetilde{\mathcal{E}}$ in Lemma 12 holds for the sampled ensemble parameter $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^{J_t}$, since

$$\left|\mathbf{x}^\top(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1})\right| \le \gamma\widetilde{C}\sqrt{\Lambda_0}\,\beta_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}} \le b_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}.$$

It remains to verify the optimism probability lower bound. Fix $t \in [T]$, and let

$$\mathbf{x}_t^* \in \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}_t}\langle \mathbf{x}, \boldsymbol{\theta}^*\rangle.$$

If $\mathbf{x}_t^* = \mathbf{0}_d$, then

$$\langle \mathbf{x}_t, \boldsymbol{\theta}_t\rangle = \max_{\mathbf{x} \in \mathcal{X}_t}\langle \mathbf{x}, \boldsymbol{\theta}_t\rangle \ge 0 = \langle \mathbf{x}_t^*, \boldsymbol{\theta}^*\rangle,$$

and hence $p_{t-1} = 1$.

Assume now that $\mathbf{x}_t^* \ne \mathbf{0}_d$, and define

$$\mathbf{u}_t^* := \mathbf{V}_{t-1}^{-1}\mathbf{x}_t^*, \qquad \bar{\mathbf{u}}_t^* := \frac{\mathbf{u}_t^*}{\|\mathbf{u}_t^*\|_2} \in \mathbb{S}^{d-1}.$$

Since $\mathcal{E}_{\mathrm{exc}}$ holds,

$$E_{t,m}(\bar{\mathbf{u}}_t^*, 1/\gamma) \ge \frac{1}{10}.$$

Equivalently, at least a $1/10$-fraction of ensemble members $j \in [m]$ satisfy

$$\frac{\langle \mathbf{s}_{t-1}^j, \bar{\mathbf{u}}_t^*\rangle}{\|\bar{\mathbf{u}}_t^*\|_{\mathbf{V}_{t-1}}} \ge \frac{1}{\gamma}.$$

By positive rescaling invariance of the self-normalized ratio, this is equivalent to

$$\frac{\langle \mathbf{s}_{t-1}^j, \mathbf{u}_t^* \rangle}{\|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}}} \geq \frac{1}{\gamma}.$$

Since $\mathbf{u}_t^* = \mathbf{V}_{t-1}^{-1}\mathbf{x}_t^*$, we have

$$\|\mathbf{u}_t^*\|_{\mathbf{V}_{t-1}} = \|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}},$$

and therefore every such j satisfies

$$\langle \mathbf{x}_t^*, \boldsymbol{\theta}_t^j - \hat{\boldsymbol{\theta}}_{t-1} \rangle = \gamma\beta_{t-1}(\mathbf{x}_t^*)^\top \mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j \geq \beta_{t-1}\|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}}.$$

On the other hand, on $\mathcal{E}_{\mathrm{LS}}$,

$$\langle \mathbf{x}_t^*, \hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^* \rangle \geq -\beta_{t-1}\|\mathbf{x}_t^*\|_{\mathbf{V}_{t-1}^{-1}}.$$

Combining the two displays gives

$$\langle \mathbf{x}_t^*, \boldsymbol{\theta}_t^j \rangle = \langle \mathbf{x}_t^*, \hat{\boldsymbol{\theta}}_{t-1} \rangle + \langle \mathbf{x}_t^*, \boldsymbol{\theta}_t^j - \hat{\boldsymbol{\theta}}_{t-1} \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle.$$

Let $\mathbf{x}_t(j) \in \operatorname{argmax}_{\mathbf{x}\in\mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}_t^j \rangle$ be the arm that would be selected if ensemble member j were sampled. Then

$$\langle \mathbf{x}_t(j), \boldsymbol{\theta}_t^j \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}_t^j \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle.$$

Since J_t is conditionally uniform on $[m]$ given $\mathcal{A}_{t-1}$, it follows that, on $\mathcal{E}_{\mathrm{LS}} \cap \mathcal{E}_{\mathrm{exc}}$,

$$p_{t-1} := \mathbb{P}(\langle \mathbf{x}_t, \boldsymbol{\theta}_t \rangle \geq \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle \,|\, \mathcal{A}_{t-1}) \geq \frac{1}{10}$$

for all $t \in [T]$.

We now apply Lemma 12. Its confidence conditions are verified on $\mathcal{E}_{\mathrm{arm}}$ and $\mathcal{E}_{\mathrm{LS}}$, and the lower bound $p_{t-1} \geq 1/10$ holds on $\mathcal{E}_{\mathrm{LS}} \cap \mathcal{E}_{\mathrm{exc}}$. Therefore, applying the finite-arm master regret lemma and intersecting with the uniform exceedance event gives an event of probability at least $1 - 4\delta$ on which

$$\mathrm{Regret}_T \leq 20 \max_{t\in[T]} b_{t-1} \left(2\sqrt{2dT \log\left(1 + \frac{T}{d\lambda}\right)} + \sqrt{2(4T/\lambda + 1)\log\left(\frac{\sqrt{4T/\lambda + 1}}{\delta}\right)} \right).$$

Finally, by monotonicity of β_t,

$$\max_{t\in[T]} b_{t-1} \leq \left(1 + \gamma\widetilde{C}\sqrt{\Lambda_0}\right)\beta_T.$$

Since $\beta_T = \widetilde{O}(\sqrt{d})$ and

$$1 + \gamma\widetilde{C}\sqrt{\Lambda_0} = \widetilde{O}(\sqrt{\Lambda_0}),$$

we obtain

$$\max_{t\in[T]} b_{t-1} = \widetilde{O}(\sqrt{d\Lambda_0}).$$

Consequently,

$$\mathrm{Regret}_T = \widetilde{O}\left(d\sqrt{T\Lambda_0}\right).$$

Since

$$\Lambda_0 = \max\left\{1,\, C_\Lambda \log\left(\frac{8KM_0T\bar{R}_0}{\delta}\right)\right\},$$

and M_0 and $\bar{R}_0$ depend only logarithmically on T, d, and $1/\delta$, we have

$$\mathrm{Regret}_T = \widetilde{O}\left(d\sqrt{T\log K}\right).$$

□

C Technical Lemmas

C.1 Proof of Lemma 5

Lemma 5 (Epoch-wise scalar-clock Brownian representation)**.** *Fix an epoch r and a direction $\mathbf{u} \in \mathbb{S}^{d-1}$. Let*

$$\ell_r := \tau_{r+1} - \tau_r \,.$$

For $k = 0, 1, \ldots, \ell_r - 1$, define the m-dimensional vector

$$\mathbf{M}_k := \big(\langle \mathbf{u}, \mathbf{s}^1_{\tau_r+k}\rangle, \ldots, \langle \mathbf{u}, \mathbf{s}^m_{\tau_r+k}\rangle\big) \in \mathbb{R}^m \,.$$

Then $(\mathbf{M}_k)_{k=0}^{\ell_r-1}$ is a diagonal martingale transform of a standard m-dimensional Gaussian noise sequence with scalar coefficients

$$D_0 := \|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}, \qquad D_k := \langle \mathbf{u}, \mathbf{x}_{\tau_r+k}\rangle \,, \qquad k = 1, \ldots, \ell_r - 1 \,.$$

Consequently, on an extension of the probability space, there exist independent standard Brownian motions $W^1, \ldots, W^m$ such that for every $k = 0, \ldots, \ell_r - 1$ and every $j \in [m]$,

$$\mathbf{M}^j_k = W^j(A^2_k) \,, \qquad A^2_k := D_0^2 + \sum_{s=1}^{k} D_s^2 = \|\mathbf{u}\|^2_{\mathbf{V}_{\tau_r+k}} \,.$$

Equivalently, for every $t \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$ and every $j \in [m]$,

$$\langle \mathbf{u}, \mathbf{s}^j_{t-1}\rangle = W^j\big(\|\mathbf{u}\|^2_{\mathbf{V}_{t-1}}\big) \,.$$

Proof of Lemma 5. Let $\mathcal{G}_{-1}$ be the σ-field generated by all randomness prior to drawing the refreshed Gaussian vectors at the start of epoch r.

For each $j \in [m]$, define

$$Z^j_r := \frac{\left\langle \mathbf{u}, \mathbf{V}^{1/2}_{\tau_r} \mathbf{g}^j_r \right\rangle}{\|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}}.$$

Since $\mathbf{g}^j_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $\|\mathbf{u}\|_{\mathbf{V}_{\tau_r}} > 0$, the variables $Z^1_r, \ldots, Z^m_r$ are *i.i.d.* $\mathcal{N}(0,1)$, and are independent of $\mathcal{G}_{-1}$.

For $k = 1, \ldots, \ell_r - 1$, let

$$\varepsilon_k := (\xi^1_{\tau_r+k}, \ldots, \xi^m_{\tau_r+k}) \in \mathbb{R}^m,$$

and set

$$\varepsilon_0 := (Z^1_r, \ldots, Z^m_r) \in \mathbb{R}^m.$$

Then $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{\ell_r-1}$ are *i.i.d.* $\mathcal{N}(\mathbf{0}_m, \mathbf{I}_m)$.

Now define

$$\mathbf{U}_0 := D_0 \mathbf{I}_m, \qquad \mathbf{U}_k := D_k \mathbf{I}_m \,, \quad 1 \le k \le \ell_r - 1 \,.$$

For $k \ge 1$, the coefficient

$$D_k = \langle \mathbf{u}, \mathbf{x}_{\tau_r+k}\rangle$$

is measurable with respect to the randomness up to just before ε_k is sampled, so $\mathbf{U}_k$ is predictable; $\mathbf{U}_0$ is $\mathcal{G}_{-1}$-measurable.

Moreover, for every $k = 0, \ldots, \ell_r - 1$,

$$\mathbf{M}_k = \mathbf{U}_0 \varepsilon_0 + \sum_{s=1}^{k} \mathbf{U}_s \varepsilon_s.$$

Indeed, for each $j \in [m]$,

$$\begin{aligned}
\mathbf{M}^j_k &= \langle \mathbf{u}, \mathbf{s}^j_{\tau_r+k}\rangle \\
&= \left\langle \mathbf{u}, \mathbf{V}^{1/2}_{\tau_r} \mathbf{g}^j_r \right\rangle + \sum_{s=\tau_r+1}^{\tau_r+k} \langle \mathbf{u}, \mathbf{x}_s\rangle \, \xi^j_s \\
&= D_0 Z^j_r + \sum_{s=1}^{k} D_s \xi^j_{\tau_r+s} \,.
\end{aligned}$$

Thus $(\mathbf{M}_k)$ is a diagonal martingale transform of a standard m-dimensional Gaussian noise sequence, with diagonal coefficient matrices that are scalar multiples of the identity. Applying the diagonal Gaussian embedding theorem of Janz et al. [10] (Lemma 14), we obtain, on an extension of the probability space, independent standard Brownian motions $W^1, \dots, W^m$ such that

$$\mathbf{M}_k^j = W^j(A_{k,j}^2), \qquad A_{k,j}^2 := \sum_{s=0}^{k} D_{s,j}^2 \,.$$

Since every coefficient matrix is scalar, the clocks are the same for all j, so

$$A_{k,j}^2 = A_k^2 = D_0^2 + \sum_{s=1}^{k} D_s^2 \,.$$

Finally,

$$A_k^2 = \|\mathbf{u}\|_{\mathbf{V}_{\tau_r}}^2 + \sum_{s=1}^{k} \langle \mathbf{u}, \mathbf{x}_{\tau_r+s} \rangle^2 = \|\mathbf{u}\|_{\mathbf{V}_{\tau_r+k}}^2,$$

because

$$\mathbf{V}_{\tau_r+k} = \mathbf{V}_{\tau_r} + \sum_{s=1}^{k} \mathbf{x}_{\tau_r+s} \mathbf{x}_{\tau_r+s}^\top \,.$$

Substituting $k = t - 1 - \tau_r$ gives

$$\langle \mathbf{u}, \mathbf{s}_{t-1}^j \rangle = W^j\big(\|\mathbf{u}\|_{\mathbf{V}_{t-1}}^2\big)$$

for all $t \in \{\tau_r + 1, \dots, \tau_{r+1}\}$ and all $j \in [m]$. □

C.2 Proof of Lemma 11

Lemma 11 (Stability of normalized projections under small matrix drift)**.** *Let $\alpha \geq 0$, and suppose*

$$\mathbf{I}_d \preceq \mathbf{A} \preceq (1+\alpha)\mathbf{I}_d \,.$$

Then, for every $\mathbf{y} \neq \mathbf{0}_d$ and every $\mathbf{v} \in \mathbb{R}^d$,

$$\left| \frac{\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{v}}{\sqrt{\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y}}} - \left\langle \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \mathbf{v} \right\rangle \right| \leq 2\alpha \|\mathbf{v}\|_2 \,.$$

Proof of Lemma 11. Let

$$\mathbf{u} := \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \qquad \mathbf{B} := \mathbf{A}^{-1} \,.$$

Then we have

$$\frac{\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{v}}{\sqrt{\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y}}} = \left\langle \frac{\mathbf{B}\mathbf{u}}{\sqrt{\mathbf{u}^\top \mathbf{B}\mathbf{u}}}, \mathbf{v} \right\rangle \,.$$

Therefore,

$$\left| \frac{\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{v}}{\sqrt{\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y}}} - \langle \mathbf{u}, \mathbf{v} \rangle \right| = \left| \left\langle \frac{\mathbf{B}\mathbf{u}}{\sqrt{\mathbf{u}^\top \mathbf{B}\mathbf{u}}} - \mathbf{u}, \mathbf{v} \right\rangle \right| \leq \left\| \frac{\mathbf{B}\mathbf{u}}{\sqrt{\mathbf{u}^\top \mathbf{B}\mathbf{u}}} - \mathbf{u} \right\|_2 \|\mathbf{v}\|_2 \,,$$

where the inequality uses Cauchy-Schwarz inequality.

Since $\mathbf{I}_d \preceq \mathbf{A} \preceq (1+\alpha)\mathbf{I}_d$, we have $\frac{1}{1+\alpha}\mathbf{I}_d \preceq \mathbf{B} \preceq \mathbf{I}_d$. Hence we get

$$\|\mathbf{B} - \mathbf{I}_d\|_{\mathrm{op}} \leq 1 - \frac{1}{1+\alpha} = \frac{\alpha}{1+\alpha} \,, \qquad \mathbf{u}^\top \mathbf{B}\mathbf{u} \in \left[\frac{1}{1+\alpha}, 1 \right] \,.$$

Thus

$$\begin{aligned}
\left\| \frac{\mathbf{B}\mathbf{u}}{\sqrt{\mathbf{u}^\top \mathbf{B}\mathbf{u}}} - \mathbf{u} \right\|_2 &\leq \frac{\|(\mathbf{B} - \mathbf{I}_d)\mathbf{u}\|_2}{\sqrt{\mathbf{u}^\top \mathbf{B}\mathbf{u}}} + \left| \frac{1}{\sqrt{\mathbf{u}^\top \mathbf{B}\mathbf{u}}} - 1 \right| \\
&\leq \frac{\alpha}{1+\alpha} \sqrt{1+\alpha} + \left(\sqrt{1+\alpha} - 1 \right) \\
&= \frac{\alpha}{\sqrt{1+\alpha}} + \frac{\alpha}{\sqrt{1+\alpha}+1} \\
&\leq \alpha + \alpha = 2\alpha \,,
\end{aligned}$$

which concludes the proof. □

C.3 Proof of Lemma 12

Lemma 12 (Finite-arm regret bound for randomized algorithms)**.** *Suppose that the arm sets $\{\mathcal{X}_t\}_{t=1}^T$ are chosen by an oblivious adversary with $|\mathcal{X}_t| = K$ for all $t \in [T]$. Under the assumptions of Lemma 10, the same conclusion continues to hold if the event*

$$\mathcal{E} = \left\{\forall t \le T : \|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \le b_{t-1}\right\}$$

is replaced by the arm-wise event

$$\widetilde{\mathcal{E}} = \left\{\forall t \le T,\ \forall \mathbf{x} \in \mathcal{X}_t : \left|\mathbf{x}^\top(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1})\right| \le b_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}\right\}.$$

Proof of Lemma 12. In the proof of Lemma 10 in Janz et al. [9], the only role of the event

$$\mathcal{E} = \left\{\forall t \le T :\ \|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\|_{\mathbf{V}_{t-1}} \le b_{t-1}\right\}$$

is to imply, by Cauchy–Schwarz, that for any arm $\mathbf{x} \in \mathcal{X}_t$,

$$\left|\mathbf{x}^\top(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1})\right| \le b_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}.$$

However, in the regret argument this implication is only invoked for two specific arms: the played arm $\mathbf{x}_t$ and the optimal arm $\mathbf{x}_t^*$ at round t. Indeed, on the optimism event

$$\langle \boldsymbol{\theta}_t, \mathbf{x}_t\rangle \ge \langle \boldsymbol{\theta}^*, \mathbf{x}_t^*\rangle,$$

the instantaneous regret satisfies

$$\begin{aligned}\langle \mathbf{x}_t^* - \mathbf{x}_t, \boldsymbol{\theta}^*\rangle &\le \langle \mathbf{x}_t, \boldsymbol{\theta}_t - \boldsymbol{\theta}^*\rangle \\ &= \langle \mathbf{x}_t, \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\rangle + \langle \mathbf{x}_t, \hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\rangle.\end{aligned}$$

Thus the proof only needs upper bounds on

$$\left|\langle \mathbf{x}_t, \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\rangle\right| \quad \text{and} \quad \left|\langle \mathbf{x}_t, \hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\rangle\right|.$$

Similarly, to show that a sampled parameter is optimistic whenever it is sufficiently large in the optimal-arm direction, the proof only uses the corresponding inequalities at $\mathbf{x}^*$:

$$\left|\langle \mathbf{x}_t^*, \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1}\rangle\right| \quad \text{and} \quad \left|\langle \mathbf{x}_t^*, \hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\rangle\right|.$$

Therefore, the full confidence-set event $\mathcal{E}$ is stronger than necessary. It is enough to assume the armwise event

$$\widetilde{\mathcal{E}} = \left\{\forall t \le T,\ \forall \mathbf{x} \in \mathcal{X}_t : \left|\mathbf{x}^\top(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t-1})\right| \le b_{t-1}\|\mathbf{x}\|_{\mathbf{V}_{t-1}^{-1}}\right\}.$$

Since $\mathbf{x}_t \in \mathcal{X}_t$ and $\mathbf{x}_t^* \in \mathcal{X}_t$, this replaced event provides exactly the same inequalities used in the proof of Lemma 10. All subsequent steps, including the optimism-probability argument and the elliptical-potential bound, are unchanged. Hence the same regret bound follows. □

D Auxiliary Lemmas

Lemma 13 (Lemma 10 in Abbasi-Yadkori et al. [1])**.** *Suppose $X_1, \dots, X_t \in \mathbb{R}^d$ and for any $1 \le s \le t$, $\|X_s\|_2 \le L$, Let $\mathbf{V}_t = \lambda \mathbf{I}_d + \sum_{s=1}^t X_s X_s^\top$ for some $\lambda > 0$. Then,*

$$\det(\mathbf{V}_t) \le (\lambda + tL^2/d)^d.$$

Lemma 14 (Theorem 5.3 in Janz et al. [10])**.** *Fix a horizon $N \in \mathbb{N}$. Assume*

$$M_t = \sum_{s=0}^{t} \operatorname{diag}(D_s)\,\xi_s, \qquad 0 \le t \le N-1,$$

is a diagonal martingale transform of a standard m-dimensional Gaussian noise. Define the coordinate clocks

$$A_{t,j}^2 := \sum_{s=0}^{t} D_{s,j}^2, \qquad j \in [m].$$

Then, on an extension of the probability space, there exist independent standard Brownian motions $W^1, \dots, W^m$ such that for all $t \in \{0, \dots, N-1\}$ and all $j \in [m]$,

$$M_{t,j} = W^j(A_{t,j}^2).$$

Lemma 15 (Lemma 7 in Lee and Oh [15])**.** *If $Z \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ is a d-dimensional multivariate Gaussian vector, then for any $\delta \in (0, 1]$,*

$$\mathbb{P}\left(\|Z\|_2 \geq \sqrt{d} + \sqrt{2 \log \frac{1}{\delta}} \leq \delta \,. \right)$$

Lemma 16 (Theorem 1 in Abbasi-Yadkori et al. [1])**.** *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,

$$\|S_t\|_{\overline{V}_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right).$$

E Experimental Details

E.1 Experimental Setup

We evaluate all methods on synthetic stochastic linear bandit instances over the unit Euclidean ball in $\mathbb{R}^d$. For each seed, the unknown parameter $\boldsymbol{\theta}^*$ is sampled from a standard Gaussian distribution and normalized to unit norm. We compare LinES [10], `LinES-ER` (ours) against randomized baselines LinTS [2] and LinPHE [13]. Unless otherwise specified, we use parameter $\lambda = 1$, failure probability $\delta = 10^{-4}$, Gaussian noise level $\sigma = 0.5$, and average results over 10 independent seeds. The ensemble sizes are set according to the theoretical prescriptions: $m = \lceil d \log(T/\delta) \rceil$ for `LinES` and $m = \lceil d \log d + d \log \log T + \log(1/\delta) \rceil$ for `LinES-ER`. Runtime is measured as the cumulative wall-clock time spent in arm selection and model updates. Experiments were run on a MacBook Pro with an Apple M4 Pro chip, 12 CPU cores, and 24 GB memory, running macOS 26.3.1.

For `LinES-ER`, we implement epoch refreshes using an infrequent scheduled check rather than testing the refresh condition at every round. Specifically, the refresh criterion is evaluated only at geometrically spaced checkpoints within each horizon scale, which avoids adding a per-round eigenvalue computation to the runtime. This implementation preserves the intended epoch-reset behavior while substantially reducing overhead; empirically, the resulting scheduled refreshes do not noticeably degrade performance in our synthetic experiments.

E.2 Additional Experiments

Figure 2 reports cumulative regret and runtime for $T = 10{,}000$ with $d \in \{10, 20, 50\}$. Across all dimensions, `LinES-ER` achieves regret comparable to `LinES` while using substantially fewer ensemble members. The computational benefit becomes more visible as the dimension increases. For example, when $d = 50$, `LinES-ER` uses 316 ensemble members compared with 922 for `LinES`, and reduces the runtime from 0.498 seconds to 0.316 seconds, corresponding to a runtime reduction of about $\frac{0.498 - 0.316}{0.498} \times 100 \approx 36.5\%$. These results support the main practical implication of our theory: reducing the required ensemble size preserves the statistical performance of ensemble sampling while improving computational efficiency.

Table 2: Comparison of theoretical ensemble sizes and runtime improvement for different dimensions. The baseline ensemble size m_{base} is computed using the prescription of Janz et al. [10], $m_{\text{base}} = \lceil d \log(T/\delta) \rceil$, whereas our ensemble size is computed as $m_{\text{ours}} = \lceil d \log d + d \log \log T + \log(1/\delta) \rceil$. Runtime improvement is measured relative to `LinES`.

T	d	δ	m_{ours}	m_{base}	$m_{\text{base}}/m_{\text{ours}}$	Runtime improvement
10^4	10	10^{-4}	55	185	3.36	9.65%
10^4	20	10^{-4}	114	369	3.24	24.71%
10^4	30	10^{-4}	178	553	3.10	30.5%
10^4	50	10^{-4}	316	922	2.92	36.55%

Table 2 summarizes how the theoretical ensemble sizes translate into runtime gains. For all dimensions, `LinES-ER` uses substantially fewer ensemble members than `LinES`; the baseline ensemble size is about 3.36, 3.24, 3.12 and 2.92 times larger for $d = 10, 20, 30, 50$, respectively. Despite this reduction, Figure 2 shows that `LinES-ER` maintains comparable cumulative regret. The smaller ensemble size also leads to lower runtime, with the improvement increasing as the dimension grows: from 9.65% at $d = 10$ to 36.55% at $d = 50$. These results support the computational benefit of the proposed reduced-ensemble design.

E.3 Anytime Implementation

Algorithm description. We now describe an anytime implementation of `LinES-ER` for the case where the horizon T is unknown. The main idea is to combine the proposed epoch-refresh mechanism with a doubling schedule [5]. Let H_0, H_1, $H_2, \ldots$ be a sequence of horizon proxies satisfying $H_{\ell+1} = 2H_\ell$. At the beginning of each doubling phase, the algorithm updates the target ensemble size according to the current horizon proxy H_ℓ. However, unlike standard ensemble-sampling methods [9, 22, 10], this does not require restarting the entire algorithm. The regularized Gram matrix $\mathbf{V}_t$, the ridge estimator $\hat{\boldsymbol{\theta}}_t$, and all previously collected observations are retained. Only the ensemble perturbations are reinitialized according to the current Gram matrix.

This is possible because the epoch-refresh construction already represents the ensemble perturbations relative to the current Gram matrix. Hence, when the ensemble size needs to be increased, the algorithm can simply start a new epoch and initialize the enlarged ensemble using the current Gram matrix $\mathbf{V}_{t-1}$. In contrast, prior ensemble-sampling methods without such a refresh mechanism typically need to restart under the doubling trick, since new ensemble members cannot be made consistent with the previous perturbation history without replaying or discarding past data. The anytime implementation of `LinES-ER` is described in Algorithm 2.

We briefly justify that Algorithm 2 preserves the fixed-horizon regret guarantee up to logarithmic factors. Let $H_\ell = 2^\ell H_0$, and define

$$I_0 := \{1, \ldots, H_0\}, \qquad I_\ell := \{H_{\ell-1} + 1, \ldots, H_\ell\} \quad (\ell \geq 1),$$

with $n_\ell := |I_\ell|$. Allocate

$$\delta_\ell := \frac{6\delta}{\pi^2(\ell+1)^2}, \qquad \sum_{\ell=0}^{\infty} \delta_\ell \leq \delta.$$

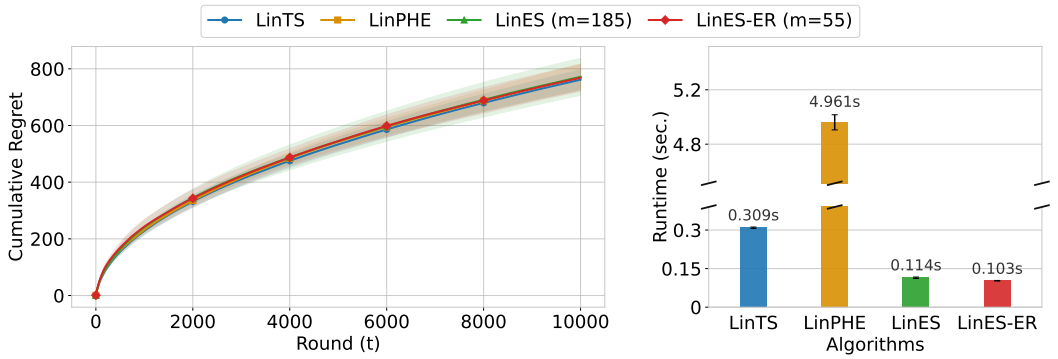
Fix a phase I_ℓ, and condition on the history before this phase. Then $\mathbf{V}_{a_\ell}$ and $\hat{\boldsymbol{\theta}}_{a_\ell}$ are fixed, and the refreshed perturbations satisfy

$$\mathbf{s}_{a_\ell}^j = \mathbf{V}_{a_\ell}^{1/2} \mathbf{g}_\ell^j, \qquad \mathbf{g}_\ell^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d).$$

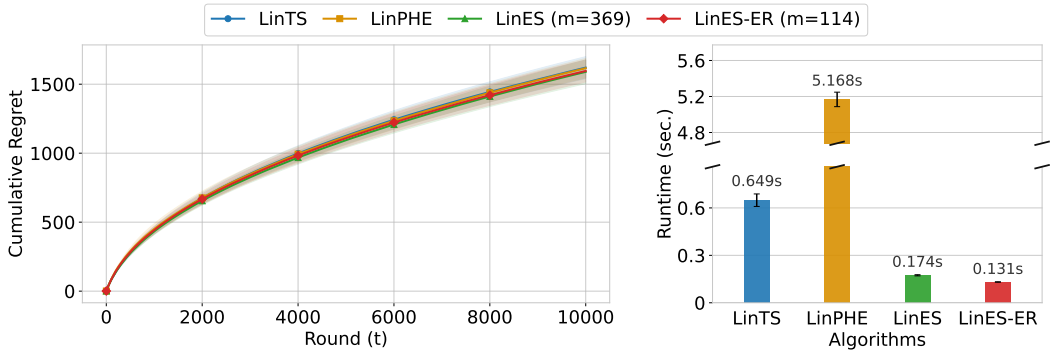
Since the Gram matrix and ridge estimator retain all past data, the global least-squares confidence event remains valid. Conditioning on the phase-start history merely fixes the initial Gram matrix for the phase-local perturbation analysis. Hence, the fixed-horizon perturbation and optimism arguments apply after a time shift, with $\mathbf{V}_{a_\ell}$ replacing $\lambda \mathbf{I}_d$.

Since $\mathbf{V}_{a_\ell} \succeq \lambda \mathbf{I}_d$, the number of refresh epochs inside I_ℓ is bounded by

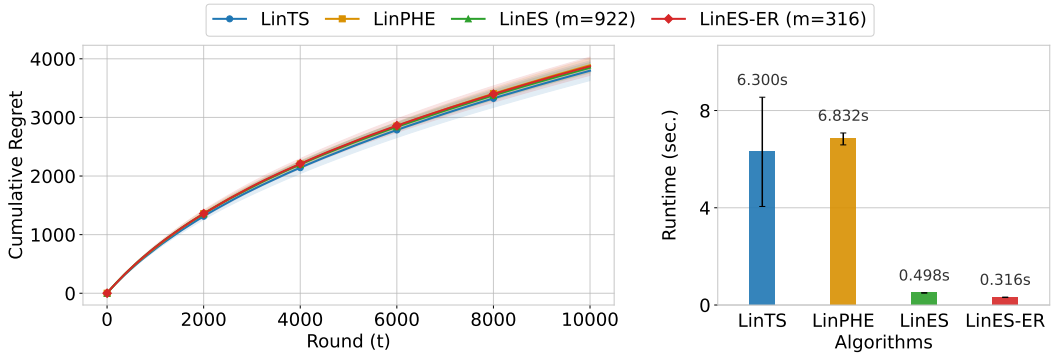
$$1 + \frac{d \log(1 + n_\ell/(\lambda d))}{\log(1+\alpha)}.$$



(a) $T = 10{,}000$ and $d = 10$



(b) $T = 10{,}000$ and $d = 20$



(c) $T = 10{,}000$ and $d = 50$

Figure 2: Cumulative regret and runtime on stochastic linear bandits over $\mathcal{B}_2^d$ with $T = 10{,}000$ and $d \in \{10, 20, 50\}$.

Therefore, choosing $m_\ell = M(H_\ell)$ according to the fixed-horizon rule with confidence level δ_ℓ, the proof of Theorem 1 gives, with probability at least $1 - \delta_\ell$,

$$\mathrm{Regret}_{I_\ell} \le \widetilde{\mathcal{O}}\left(d^{3/2}\sqrt{n_\ell}\right).$$

The bound also holds for every prefix of I_ℓ.

By a union bound, all phase-wise bounds hold simultaneously with probability at least $1 - \delta$. For any $T \ge 1$, let $L(T) := \min\{\ell : T \le H_\ell\}$. Then

$$\mathrm{Regret}_T \le \sum_{\ell=0}^{L(T)} \mathrm{Regret}_{I_\ell \cap [T]} \le \widetilde{\mathcal{O}}\left(d^{3/2} \sum_{\ell=0}^{L(T)} \sqrt{n_\ell}\right).$$

Algorithm 2 Anytime Linear Ensemble Sampling with Epoch Refresh

Inputs: initial horizon proxy $H_0 \in \mathbb{N}$, regularization parameter $\lambda > 0$, inflation parameter $\gamma > 0$, confidence radii $\{\beta_t\}_{t\geq 0}$, epoch refresh parameter $\alpha > 0$, ensemble-size rule $M(\cdot)$
Set $H \leftarrow H_0$, $m \leftarrow M(H)$
Initialize $\mathbf{V}_0 = \lambda \mathbf{I}_d$, $\hat{\boldsymbol{\theta}}_0 = \mathbf{0}_d$, and set the current epoch start time $\tau = 0$
Sample $\mathbf{g}^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and set $\mathbf{s}_0^j = \mathbf{V}_0^{1/2}\mathbf{g}^j$, $\boldsymbol{\theta}_0^j = \mathbf{V}_0^{-1}\mathbf{s}_0^j$ for each $j \in [m]$
for $t = 1, 2, \ldots$ **do**
    if $t > H$ **then** ▷ *Doubling update*
        Set $H \leftarrow 2H$, $m \leftarrow M(H)$, and $\tau \leftarrow t - 1$
        For each $j \in [m]$, sample $\mathbf{g}^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and set

$$\mathbf{s}_{t-1}^j = \mathbf{V}_\tau^{1/2}\mathbf{g}^j, \qquad \boldsymbol{\theta}_{t-1}^j = \hat{\boldsymbol{\theta}}_{t-1} + \gamma\beta_{t-1}\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j$$

    end if
    if $\mathbf{V}_{t-1} \not\preceq (1+\alpha)\mathbf{V}_\tau$ **then** ▷ *Refresh the ensemble perturbations*
        Set $\tau \leftarrow t - 1$
        For each $j \in [m]$, sample $\mathbf{g}^j \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and set

$$\mathbf{s}_{t-1}^j = \mathbf{V}_\tau^{1/2}\mathbf{g}^j, \qquad \boldsymbol{\theta}_{t-1}^j = \hat{\boldsymbol{\theta}}_{t-1} + \gamma\beta_{t-1}\mathbf{V}_{t-1}^{-1}\mathbf{s}_{t-1}^j$$

    end if
    Observe $\mathcal{X}_t$, sample $J_t \sim \mathrm{Unif}([m])$, choose $\mathbf{x}_t \in \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}_t}\langle \mathbf{x}, \boldsymbol{\theta}_{t-1}^{J_t}\rangle$, and observe y_t
    Update $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$ and $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1}\sum_{s=1}^t \mathbf{x}_s y_s$
    For each $j \in [m]$, sample $\xi_t^j \sim \mathcal{N}(0,1)$ and update

$$\mathbf{s}_t^j = \mathbf{s}_{t-1}^j + \mathbf{x}_t\xi_t^j, \qquad \boldsymbol{\theta}_t^j = \hat{\boldsymbol{\theta}}_t + \gamma\beta_t\mathbf{V}_t^{-1}\mathbf{s}_t^j$$

end for

Since $n_0 = H_0$ and $n_\ell = H_{\ell-1}$ for $\ell \geq 1$,

$$\sum_{\ell=0}^{L(T)} \sqrt{n_\ell} = \mathcal{O}(\sqrt{T})$$

for fixed H_0. Hence

$$\mathrm{Regret}_T = \widetilde{\mathcal{O}}\Big(d^{3/2}\sqrt{T}\Big)$$

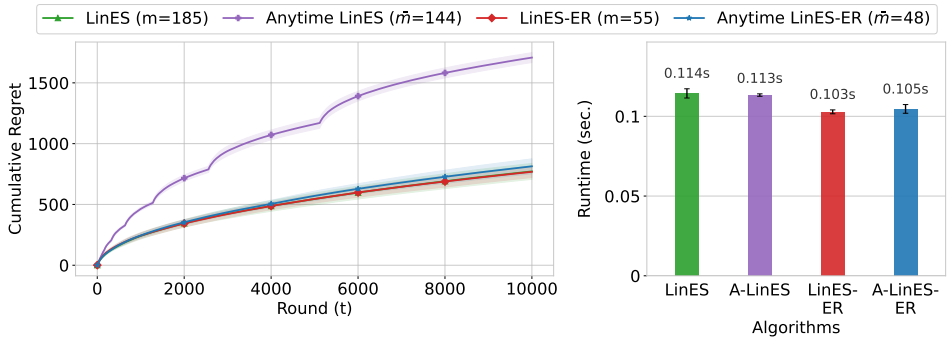
simultaneously for all $T \geq 1$, up to the standard logarithmic factors from the doubling schedule [5].

Experiments. We further evaluate the anytime variants under the doubling schedule. In the anytime version of `LinES`, the algorithm is restarted at each doubling epoch and the ensemble size is updated according to the new horizon proxy. In contrast, the anytime version of `LinES-ER` (Algorithm 2) retains all past observations and only updates its ensemble representation through the epoch-refresh mechanism.

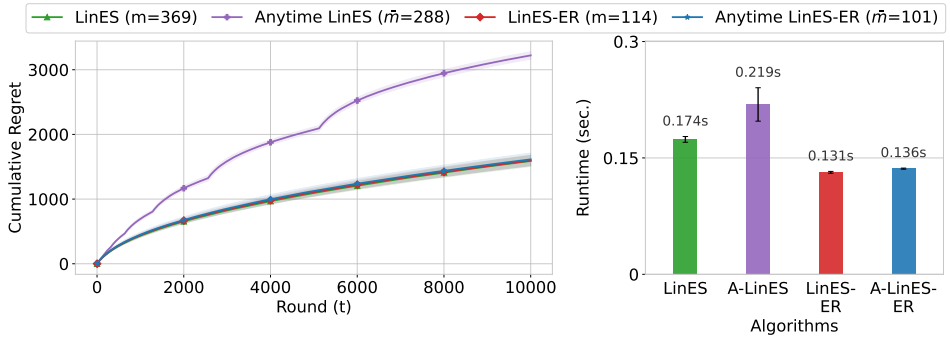
Figure 3 shows the cumulative regret and runtime for $T = 10{,}000$ and $d \in \{10, 20, 30, 50\}$. Across all dimensions, `Anytime LinES` suffers a substantial increase in regret due to repeated restarts. By contrast, `Anytime LinES-ER` remains close to the fixed-horizon `LinES-ER` baseline, demonstrating that our no-reset implementation preserves the statistical efficiency of the fixed-horizon algorithm. Interestingly, `Anytime LinES-ER` uses fewer ensemble members on average than the fixed-horizon `LinES-ER`, since its ensemble size is chosen according to the local doubling horizon rather than the final horizon. Its runtime can nevertheless be slightly larger, because the ensemble is refreshed at every doubling epoch in addition to the usual Gram-matrix-triggered refreshes. Thus, the anytime version trades a small amount of additional refresh cost for the ability to retain all past observations and avoid full restarts. Here, $\bar{m}$ denotes the average ensemble size used across doubling epochs.

E.4 Finite-Arm Experiments

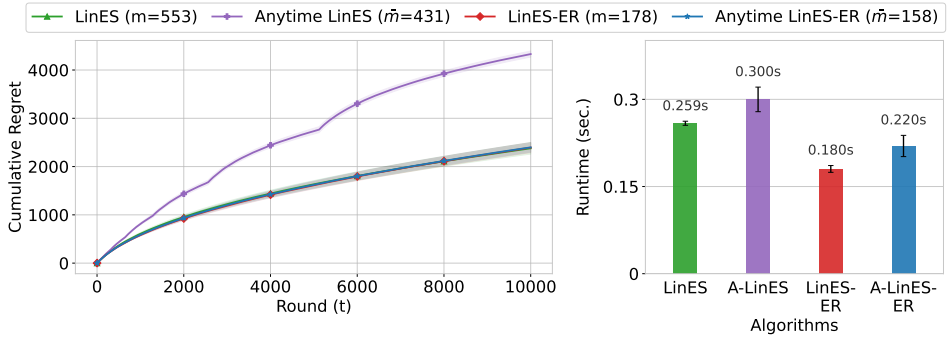
We also evaluate the algorithms in a finite-arm setting, where each instance is generated by sampling K arms uniformly from the unit sphere in $\mathbb{R}^d$. Figure 4 reports cumulative regret for several values



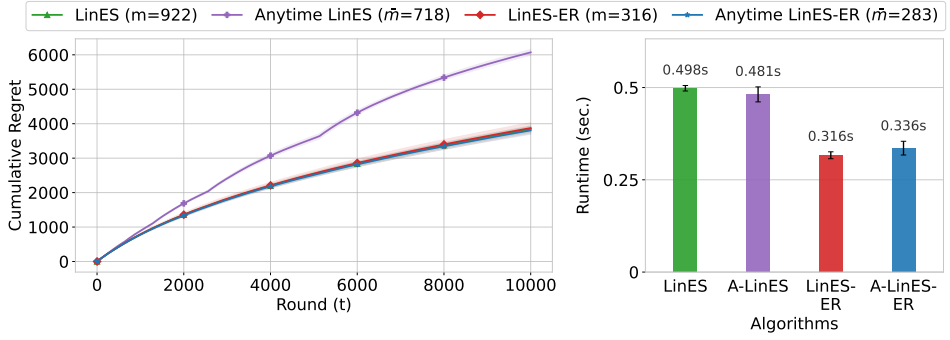
(a) $T = 10{,}000$ and $d = 10$



(b) $T = 10{,}000$ and $d = 20$



(c) $T = 10{,}000$ and $d = 30$

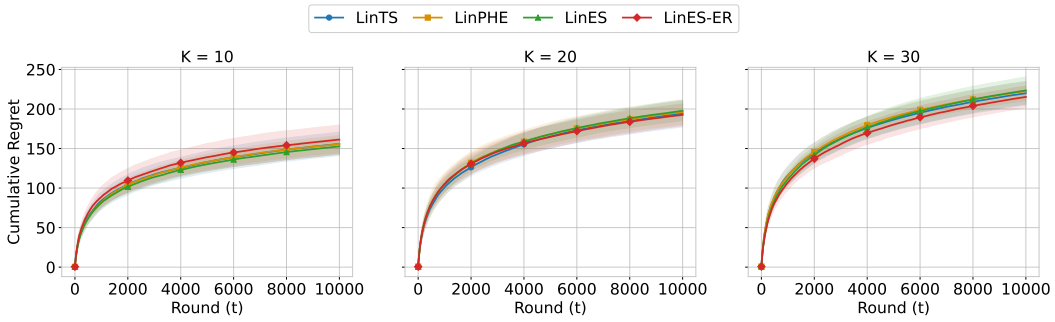


(d) $T = 10{,}000$ and $d = 50$

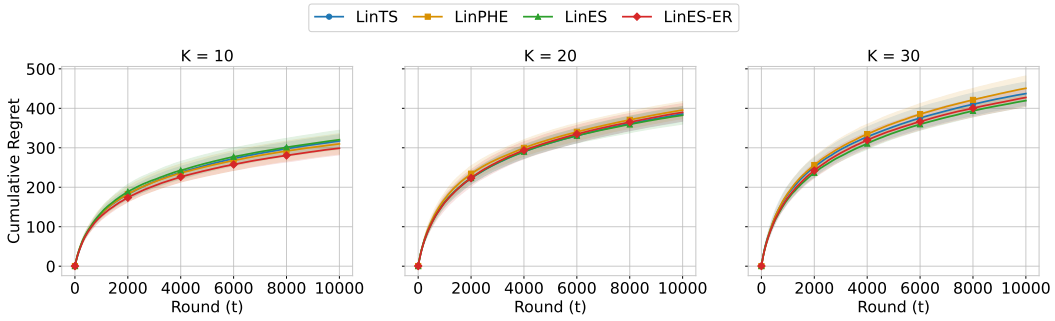
Figure 3: Anytime implementation under the doubling schedule for $T = 10{,}000$ and $d \in \{10, 20, 30, 50\}$. Left panels: cumulative regret. Right panels: total runtime.

of K. Across all tested settings, `LinES-ER` remains competitive with `LinES`, `LinTS`, and `LinPHE`.

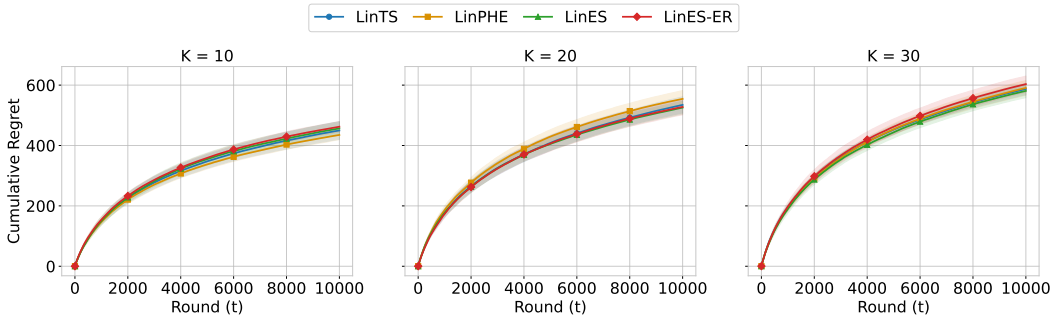
This indicates that the proposed epoch-refresh mechanism does not degrade empirical performance in finite-arm problems, despite using a smaller ensemble than `LinES`. The results are consistent with our finite-arm theory, which shows that arm-wise concentration can yield the sharper $\widetilde{\mathcal{O}}(d\sqrt{T \log K})$ regret scaling.



(a) $T = 10{,}000$ and $d = 10$



(b) $T = 10{,}000$ and $d = 20$



(c) $T = 10{,}000$ and $d = 30$

Figure 4: Finite-arm experiments with randomly generated arms on the unit sphere.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's main contributions: a reduced-ensemble linear ensemble sampling algorithm, regret guarantees for both general bounded arm sets and finite-arm settings. The stated claims are consistent with the formal theorems and experimental evaluation presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and scope conditions of the proposed results. In particular, the finite-arm regret guarantee requires arm sets to be given obliviously, and the theoretical analysis is developed for stochastic linear bandits. The paper also identifies extending the analysis to more complex nonlinear reward settings as an important direction for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for each theoretical result. Complete proofs of the main regret bounds and auxiliary concentration results are provided in the main text and appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the experimental settings needed to reproduce the main results, including the bandit environments, arm-set generation, noise distribution, parameter choices, ensemble-size prescriptions, number of independent runs, runtime measurement protocol, and hardware information. Additional experimental details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code for reproducing the main experimental results will be made publicly available in the supplemental material, together with instructions for running the experiments and generating the reported figures. The experiments use synthetic data generated from the stochastic linear bandit model described in the paper, so no external dataset is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments use synthetic stochastic linear bandit instances rather than train/test splits. The paper specifies the environment generation, reward noise model, regularization parameter, ensemble-size choices, confidence parameters, number of independent runs, and runtime measurement protocol. Since the algorithms are online bandit methods, no separate optimizer or offline training procedure is used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The experimental results are averaged over multiple independent runs, and the plots report variability across runs using shaded error regions or standard deviations where appropriate. The paper also specifies the number of independent simulations used to compute these statistics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute environment used for the experiments, including the processor type, number of CPU cores, memory, and operating system. The experiments are lightweight synthetic linear bandit simulations and do not require specialized hardware such as GPUs. Runtime is measured and reported as wall-clock time for arm selection and model updates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical and experimental work on stochastic linear bandit algorithms using synthetic data. It does not involve human subjects, sensitive personal data, or deployed decision systems, and we are not aware of any aspect that conflicts with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release high-risk data or models such as pre-trained language models, image generators, or scraped datasets. The experiments use synthetic stochastic linear bandit instances generated according to the model described in the paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new datasets, pretrained models, or other reusable assets requiring separate documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology, theoretical analysis, or experiments. Any LLM usage, if any, is limited to writing, editing, or formatting assistance and does not affect the scientific content or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.