Weighted Sequential Bayesian Inference for Non-Stationary Linear Contextual Bandits

Nicklas Werge[1] **Yi-Shan Wu**[1,2] **Abdullah Akgül**[1] **Melih Kandemir**[1]

[1]Department of Mathematics and Computer Science, University of Southern Denmark
[2]Research Center for Information Technology Innovation, Academia Sinica

Abstract

In non-stationary linear contextual bandits, existing efficient algorithms typically rely on the Weighted Regularized Least-Squares (WRLS) estimator. Because WRLS only provides point estimates, previous methods typically construct surrogate distributions when aiming to perform Bayesian-like randomized exploration. To more properly establish the Bayesian principles, we introduce *Weighted Sequential Bayesian* (WSB) inference, which forms a sequence of posteriors over a sequence of non-stationary reward parameters. This Bayesian take allows us to isolate the influence of initial beliefs into a dynamic prior penalty evaluated through the posterior covariance, which typically decreases over time. Building on this framework, we instantiate three WSB-based algorithms for exploration: `WSB-LinUCB`, `WSB-RandLinUCB`, and `WSB-LinTS`. By extending a refined drift analysis to randomized exploration without requiring local norms, we establish frequentist regret guarantees that match state-of-the-art WRLS-based baselines. Empirically, WSB's dynamic prior penalty reduces overconservatism, allowing our algorithms to consistently match or exceed their WRLS-based counterparts. Lastly, we also provide a simplified proof for the time-uniform concentration of vector-valued martingales, a critical subroutine used throughout the literature, that might be of independent interest.

1 INTRODUCTION

Bandits provide a foundational framework for sequential decision-making under uncertainty, with applications spanning recommendation systems, clinical trials, and adaptive control (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020; Robbins, 1952). Contextual bandits extend this framework by incorporating side information, like user characteristics, patient histories, or sensor readings, to better inform action selection (Krause and Ong, 2011; Li et al., 2010). Among contextual bandits, their linear variant are particularly well-studied (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010).

Many real-world applications involve *non-stationarity*, in which the reward function evolves over time (Besbes et al., 2014). To cope with this, three main strategies have been proposed: *restart*, which periodically reset the learning process (Zhao et al., 2020); *sliding-window*, which focus on a fixed-size window of recent data (Cheung et al., 2019); and *weighted*, which apply decaying weights to older data (Russac et al., 2019; Wang et al., 2023). Among these, weighted strategies are particularly appealing due to their smooth, continuous adaptation without requiring resets or fixed memory budgets. However, they have historically faced analytical challenges, and it remains an open question whether they can achieve optimal regrets (Faury et al., 2021b; Touati and Vincent, 2020; Wang et al., 2023; Zhao and Zhang, 2021).

A fundamental challenge in bandit learning is the exploration-exploitation tradeoff. To balance this, exploration strategies generally fall into one of two paradigms: a *frequentist* or a *Bayesian* approach. The most common frequentist choices usually rely on the *Weighted Regularized Least-Squares* (WRLS) estimator, which admits closed-form updates with per-round complexity $\mathcal{O}(d^2)$ where d is the dimension, but such a point estimate lacks a native mechanism for uncertainty quantification. A Bayesian alternative can use, for instance, a *Gaussian Process* (GP) objective (Chowdhury and Gopalan, 2017; Deng et al., 2022; Srinivas et al., 2010). While GPs provide calibrated uncertainty estimates in a *non-parametric* framework, they require maintaining and inverting kernel matrices that scale with time t, resulting in a per-round complexity of $\mathcal{O}(t^3)$. This limits their use in long-horizon or streaming problems. There have been approaches trying to retain the $\mathcal{O}(d^2)$ computation

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7314–7340.

complexity inspired by Thompson Sampling (TS) (Agrawal and Goyal, 2013; Kim and Tewari, 2020; Vaswani et al., 2020), which, however, did not take full advantage of the Bayesian power. In particular, the algorithms build surrogate posterior distributions from the WRLS point estimates, which are forced to inject artificial Gaussian noise scaled by frequentist covariance matrices.

To bridge this gap, we revisit the non-stationary linear bandit problem from a parametric Bayesian perspective. Specifically, we introduce *Weighted Sequential Bayesian* (WSB) inference, which copes with non-stationarity by building a sequence of generalized posteriors over the finite-dimensional, time-varying reward parameters $\theta_t \in \mathbb{R}^d$. While we show that the means of WSB posteriors algebraically coincide with the WRLS point estimates under an uninformative prior, our framework natively provides the posterior covariance required to properly ground randomized exploration. This allows WSB posteriors to retain the $\mathcal{O}(d^2)$ efficiency of WRLS estimators without relying on artificial surrogate distributions. Furthermore, by carefully incorporating this Bayesian structure into our concentration analysis, we show that the prior's influence shows up as a dynamic prior penalty term. In contrast to WRLS-based analyses, where the initialization penalty is bounded by a fixed worst-case constant, this WSB penalty is evaluated through the time-varying posterior covariance. This yields a data-dependent counterpart of the usual static prior penalty, which can be substantially smaller as the posterior covariance contracts.

Contributions. The main contributions of this paper are as follows:

- First, we establish a time-uniform confidence bound for WSB posteriors (Lemma 2), which decomposes the estimation error into three distinct components: a drift term, a noise term, and a prior term. While this decomposition resembles standard techniques (Russac et al., 2019; Wang et al., 2023), our Bayesian approach allows us to explicitly isolate the influence of the initial beliefs into a dynamic prior penalty term (Π_t).
- Second, while the self-normalized bound used to control the noise term is a weighted corollary to classical martingale results (Abbasi-Yadkori et al., 2011), we provide a simplified proof of this vector-valued concentration inequality. Inspired by the advances in modern martingale theory (Howard et al., 2020), we directly apply Ville's inequality for non-negative super-martingales (Ville, 1939), bypassing the complex stopping-time constructions traditionally used in the bandit literature.
- Third, building on the WSB framework, we propose three exploration strategies: Upper Confidence Bound (UCB), randomized UCB (RandUCB), and Thompson Sampling (TS), yielding `WSB-LinUCB`, `WSB-RandLinUCB`, and `WSB-LinTS`, respectively. We establish frequentist regret guarantees for all three algorithms, summarized in Table 1. While the $d^{1/8}$ regret improvement over earlier baselines relies on the refined drift analysis of Wang et al. (2023), which had previously been applied only to UCB-based approaches, we extend these guarantees to randomized exploration strategies. Furthermore, our experiments demonstrate the empirical benefits of this Bayesian treatment across different dimensions. Since WSB evaluates the prior penalty through the time-varying posterior covariance rather than a fixed worst-case constant, its arm-selection criteria tends to be less conservative.

Table 1: Regrets for non-stationary linear contextual bandits; d is the dimension, K the number of actions, B_T the non-stationarity, T the horizon.

ALGORITHM	REGRET
`D-LinUCB` (Russac et al., 2019)	$\tilde{\mathcal{O}}(d^{7/8}B_T^{1/4}T^{3/4})$
`LB-WeightUCB` (Wang et al., 2023)	$\tilde{\mathcal{O}}(d^{3/4}B_T^{1/4}T^{3/4})$
`WSB-LinUCB` (**Ours**, Algorithm 1)	$\tilde{\mathcal{O}}(d^{3/4}B_T^{1/4}T^{3/4})$
`D-RandLinUCB` (Kim and Tewari, 2020)	$\tilde{\mathcal{O}}(d^{7/8}B_T^{1/4}T^{3/4})$
`WSB-RandLinUCB` (**Ours**, Algorithm 2)	$\tilde{\mathcal{O}}(d^{3/4}B_T^{1/4}T^{3/4})$
`D-LinTS` (Kim and Tewari, 2020)	$\tilde{\mathcal{O}}(d^{7/8}\log(K)^{3/8}B_T^{1/4}T^{3/4})$
`WSB-LinTS` (**Ours**, Algorithm 3)	$\tilde{\mathcal{O}}(d^{3/4}\log(K)^{3/8}B_T^{1/4}T^{3/4})$

Notations. For $x, y \in \mathbb{R}^d$, let $\langle x, y\rangle$ denote the standard inner product and $\|x\|_2 = \sqrt{\langle x, x\rangle}$ the Euclidean norm. Given a matrix $M \in \mathbb{R}^{d\times d}$, we define the weighted inner product $\langle x, y\rangle_M = x^\top My$ and weighted norm $\|x\|_M = \sqrt{x^\top Mx}$. Let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the smallest and largest eigenvalues of M, respectively. We write $M \succ 0$ for positive definite and $M \succeq 0$ for positive semi-definite matrices. Let $\mathbb{I}_d$ be the $d \times d$ identity matrix.

2 BACKGROUND

Before our Bayesian treatment, we first present the non-stationary linear contextual bandit problem. We then review the widely used *Weighted Regularized Least-Squares* (WRLS) estimator, which forms the basis of many existing algorithms. In particular, we discuss how WRLS is used in both (deterministic) UCB-based and randomized exploration strategies, and highlight recent analytical refinements that improve regrets.

2.1 PROBLEM FORMULATION

The non-stationary linear contextual bandit problem is defined as follows. At each round t, the learner observes a set of K context-dependent actions $\mathcal{X}_t \subseteq \mathbb{R}^d$, which may change. Based on the history from the previous $t-1$ rounds, denoted by $\mathcal{H}_{t-1} = \{(X_s, r_s)\}_{s=1}^{t-1}$, the learner

selects an action $X_t \in \mathcal{X}_t$ and receives a noisy reward $r_t = \langle X_t, \theta_t^* \rangle + \varepsilon_t$, where $\theta_t^* \in \mathbb{R}^d$ is the unknown time-varying reward parameter and ε_t is conditionally σ-sub-Gaussian given the σ-algebra $\mathcal{F}_{t-1} = \sigma(\mathcal{H}_{t-1}, X_t)$; that is, for all $\nu \in \mathbb{R}$, $\mathbb{E}[\exp(\nu\varepsilon_t)|\mathcal{F}_{t-1}] \leq \exp(\nu^2\sigma^2/2)$ almost surely. We make the following standard boundedness assumption throughout:

Assumption 1. There exist $S, L \geq 0$ such that for all t, $\|\theta_t^*\|_2 \leq S$ and $\|x\|_2 \leq L$ for any $x \in \mathcal{X}_t$.

The degree of non-stationarity is measured by the total variation budget $B_T = \sum_{t=1}^{T-1} \|\theta_t^* - \theta_{t+1}^*\|_2$, which accounts for both gradual drifts and abrupt changes in the underlying reward parameters (Besbes et al., 2014; Garivier and Moulines, 2011). Let $X_t^* = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta_t^* \rangle$ denote the best action in round t. The learner's goal is to minimize the dynamic (pseudo-)regret

$$R_T = \sum_{t=1}^{T} \langle X_t^*, \theta_t^* \rangle - \sum_{t=1}^{T} \langle X_t, \theta_t^* \rangle.$$

2.2 WEIGHTED REGULARIZED LEAST-SQUARES

In this section, we review the WRLS estimator, its role in UCB-based exploration under the refined analysis of Wang et al. (2023), and the construction of randomized exploration strategies upon it. Later, in Section 3, this WRLS estimator will be replaced by its Bayesian analogue, which retains a similar recursive form while incorporating prior beliefs.

WRLS estimator. The WRLS estimator is a popular choice in non-stationary settings because it adapts to evolving parameters by down-weighting past observations through exponential discounting (Kim and Tewari, 2020; Russac et al., 2019; Wang et al., 2023). It is defined as:

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \left(\lambda \|\theta\|_2^2 + \sum_{s=1}^{t} \gamma^{t-s} (\langle X_s, \theta \rangle - r_s)^2 \right), \quad (1)$$

where $\lambda > 0$ is a regularization parameter and $\gamma \in (0, 1)$ is a discounted factor. If $\gamma = 1$, this reduces to the Regularized Least-Squares (RLS) estimator, which is widely used in stationary settings (Abbasi-Yadkori et al., 2011; Abeille and Lazaric, 2017; Agrawal and Goyal, 2013; Vaswani et al., 2020).

The WRLS estimator admits a closed-form solution given by $\hat{\theta}_t = V_t^{-1} b_t$, where $V_t = \lambda \mathbb{I}_d + \sum_{s=1}^{t} \gamma^{t-s} X_s X_s^\top$ and $b_t = \sum_{s=1}^{t} \gamma^{t-s} X_s r_s$. The matrix $V_t \in \mathbb{R}^{d \times d}$ is positive definite by construction, and both V_t and b_t can be updated recursively; $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda \mathbb{I}_d$ and $b_t = \gamma b_{t-1} + X_t r_t$, with initial values $V_1 = \lambda \mathbb{I}_d$ and $b_1 = \mathbf{0}$. Consequently, the WRLS estimator can be updated recursively as $\hat{\theta}_t = V_t^{-1} b_t$, initialized with $\hat{\theta}_1 = \mathbf{0}$, and thereby avoids the need to store past observations.

Error decomposition. Existing analyses of WRLS-based algorithms typically decompose the estimation error $(\hat{\theta}_{t-1} - \theta_t^*)$ into two distinct parts: a *drift part*, capturing non-stationarity in the reward parameters, and a *noise part*, reflecting randomness in the observed rewards. The index shift arises because the decision at round t is based on $\hat{\theta}_{t-1}$, computed from observations available only up to round $t-1$. The *drift part* is bounded by the bias term α_t^{WRLS}, while the *noise part* is captured by the confidence radius $\beta_t^{\mathrm{WRLS}}(\delta)$.

Earlier work, such as Cheung et al. (2022); Kim and Tewari (2020); Russac et al. (2019), controlled the *drift part* using sliding-window techniques (Cheung et al., 2019), which retain only a fixed-length subset of recent observations. While intuitive, this approach introduces unnecessary complexity. In contrast, Wang et al. (2023, Lemma 4) proposed a simpler analysis that yields a fully deterministic bound on the drift without relying on artificial windows.

For the *noise part*, high-probability concentration inequalities are used. In the stationary setting, Abbasi-Yadkori et al. (2011) derived a self-normalized Martingale tail bound for the RLS estimator, later generalized to non-stationary problems by Russac et al. (2019) via the *local norm*, a construct adopted in several subsequent studies (Kim and Tewari, 2020; Touati and Vincent, 2020). However, this *local norm* is merely a technical artifact of the analysis rather than a fundamental necessity; see, e.g., Wang et al. (2023, Lemma 5).

Concentration bounds for WRLS. The bounds on these two components, drift and noise, together control the estimation error $(\hat{\theta}_{t-1} - \theta_t^*)$. These bounds are central for constructing exploration strategies.

Lemma 1 (Wang et al. (2023, Lemma 1))**.** *For any $\delta \in (0, 1)$, with probability at least $1-\delta$, the following inequality holds for all $t \in \mathbb{N}_+$:*

$$\|\hat{\theta}_{t-1} - \theta_t^*\|_{V_{t-1}} \leq \alpha_{t-1}^{\mathrm{WRLS}} + \beta_{t-1}^{\mathrm{WRLS}}(\delta)$$

and $\forall x \in \mathcal{X}_t$, $|\langle x, \hat{\theta}_{t-1} - \theta_t^* \rangle| \leq (\alpha_{t-1}^{\mathrm{WRLS}} + \beta_{t-1}^{\mathrm{WRLS}}(\delta)) \|x\|_{V_{t-1}^{-1}}$, where

- $\alpha_t^{\mathrm{WRLS}} = L\sqrt{d} \sum_{k=1}^{t} \sqrt{\gamma^t} \sqrt{\frac{\gamma^{-k}-1}{1-\gamma}} \|\theta_k^* - \theta_{k+1}^*\|_2$ *is the drift-induced bias and*
- $\beta_t^{\mathrm{WRLS}}(\delta) = \sigma \sqrt{2\log(\frac{1}{\delta}) + d\log(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)})} + \sqrt{\lambda} S$ *is the confidence level.*

Lemma 1 can be further simplified using $\|x\|_2 \leq L$ (Assumption 1), and the property that for a positive definite and symmetric matrix $M \in \mathbb{R}^{d \times d}$ and any vector $x \in \mathbb{R}^d$, $\|x\|_{M^{-1}} \leq \|x\|_2 / \sqrt{\lambda_{\min}(M)}$, yielding $\alpha_{t-1}^{\mathrm{WRLS}} \|x\|_{V_{t-1}^{-1}} \leq L\alpha_{t-1}^{\mathrm{WRLS}} / \sqrt{\lambda}$ that is independent of x.

2.3 DETERMINISTIC EXPLORATION WITH WRLS

Wang et al. (2023) use Lemma 1 to form their UCB-based algorithm, `LB-WeightUCB`. Specifically, they take

$$X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle x, \hat{\theta}_{t-1} \rangle + \beta_{t-1}^{\mathrm{WRLS}}(\delta) \|x\|_{V_{t-1}^{-1}}\}.$$

Notably, `LB-WeightUCB` requires maintaining only a single covariance matrix V_t. An earlier UCB-based algorithm, `D-LinUCB` by Russac et al. (2019), constructs upper confidence bounds using a more complex *local norm*. Specifically, they set

$$X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle x, \hat{\theta}_{t-1} \rangle + \beta_{t-1}^{\mathrm{WRLS}}(\delta) \|x\|_{V_{t-1}^{-1} \tilde{V}_{t-1} V_{t-1}^{-1}}\},$$

where $\tilde{V}_t = \lambda \mathbb{I}_d + \sum_{s=1}^{t} \gamma^{2(t-s)} X_s X_s^\top$ is an additional covariance matrix in $\mathbb{R}^{d \times d}$. Obviously, `D-LinUCB` is both computationally and memory-wise more demanding than `LB-WeightUCB`, since it requires maintaining and inverting two $d \times d$ matrices, whereas `LB-WeightUCB` works with just one. For instance, Wang et al. (2023, Figure 1) report that `LB-WeightUCB` achieves over a $1.5\times$ speedup relative to `D-LinUCB`, mainly due to eliminating the *local norm* and relying solely on V_t.

Regret guarantees. The established confidence bound is central to the analysis of UCB-style algorithms that use WRLS. As summarized in Table 1, `LB-WeightUCB` (Wang et al., 2023) achieves a regret of $\tilde{\mathcal{O}}(d^{3/4} B_T^{1/4} T^{3/4})$, while the earlier `D-LinUCB` (Russac et al., 2019) has a regret of $\tilde{\mathcal{O}}(d^{7/8} B_T^{1/4} T^{3/4})$. This improved regret is a direct consequence of the refined treatment described above.

2.4 RANDOMIZED EXPLORATION WITH WRLS

We now review two randomized exploration strategies based on the WRLS estimator: `D-RandLinUCB` and `D-LinTS` (Kim and Tewari, 2020). Like `D-LinUCB`, both build on WRLS estimates and use *local norms* in their confidence bounds, which require maintaining two covariance matrices (see Section 2.3). Unlike deterministic UCB-based algorithms that address uncertainty via the Optimism-in-the-Face-of-Uncertainty (OFU) principle, the randomized approaches introduce randomness into the action selection process: `D-RandLinUCB` perturbs the confidence level, while `D-LinTS` samples randomized parameter estimates. We describe each method in detail below.

Randomized UCB. Randomized UCB was originally proposed by Vaswani et al. (2020) and later extended to the non-stationary setting by Kim and Tewari (2020). In this approach, the confidence level $\beta_t^{\mathrm{WRLS}}(\delta)$ in `D-LinUCB` is replaced by a random variable η_t, sampled from a fixed, easy-to-sample distribution with confidence level $a > 0$. For example, Kim and Tewari (2020, Section 4) use a truncated univariate Gaussian distribution that assigns probability mass only to $[0, \infty)$ in their numerical experiments. At each round t, a sample $\eta_t \sim \mathcal{N}(0, a^2)$ is drawn,[1] and the action is selected according to

$$X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle x, \hat{\theta}_{t-1} \rangle + \eta_t \|x\|_{V_{t-1}^{-1} \tilde{V}_{t-1} V_{t-1}^{-1}}\}.$$

Thompson Sampling. The idea of injecting noise into the arm-selection criteria is not new. One of the most well-known ways of perturbing the estimates is Thompson Sampling (TS) (Thompson, 1933). For stationary linear contextual bandits, TS is known as `LinTS` (Agrawal and Goyal, 2013), while `D-LinTS` (Kim and Tewari, 2020) is its non-stationary counterpart. The `D-LinTS` algorithm follows the recipe of Abeille and Lazaric (2017) that builds a sequence of posteriors with similar components as UCB-based methods. In particular, at each round t, a randomized parameter is sampled as $\tilde{\theta}_{t-1} = \hat{\theta}_{t-1} + V_{t-1}^{-1} \tilde{V}_{t-1}^{1/2} \eta_t$, where $\eta_t \sim \mathcal{N}(0, a^2 \mathbb{I}_d)$. The action is then selected according to

$$X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle x, \tilde{\theta}_{t-1} \rangle\}.$$

Coupled and decoupled randomization. A key distinction between `D-RandLinUCB` and `D-LinTS` lies in how random perturbations are applied when selecting actions. The arm-selection criterion of `D-LinTS` can be written as

$$\langle x, \tilde{\theta}_{t-1} \rangle = \langle x, \hat{\theta}_{t-1} \rangle + x^\top V_{t-1}^{-1} \tilde{V}_{t-1}^{1/2} \eta_t,$$

which equivalently can be expressed as

$$\langle x, \hat{\theta}_{t-1} \rangle + \eta_{t,x} \|x\|_{V_{t-1}^{-1} \tilde{V}_{t-1} V_{t-1}^{-1}}, \quad \eta_{t,x} \sim \mathcal{N}(0, a^2).$$

In `D-RandLinUCB`, the random perturbation is *coupled*, i.e., the same random variable is used for all arms in a given round. This means that the randomness affects all actions simultaneously, resulting in correlated exploration across the action set. In contrast, `D-LinTS` uses *decoupled* perturbations, where each arm receives an independent random perturbation in every round. This leads to greater variability in the exploration process, but also results in a slightly higher regret bound due to the increased variance.

Regret guarantees. As outlined in Table 1, the regret of `D-RandLinUCB` is $\tilde{\mathcal{O}}(d^{7/8} B_T^{1/4} T^{3/4})$ and `D-LinTS` is $\tilde{\mathcal{O}}(d^{7/8} \log(K)^{3/8} B_T^{1/4} T^{3/4})$. Notably, if the refined analysis of Wang et al. (2023) is applied in place of that of Russac et al. (2019) as described in Section 2.3, these regrets can be improved to $\tilde{\mathcal{O}}(d^{3/4} B_T^{1/4} T^{3/4})$ for `D-RandLinUCB` and $\tilde{\mathcal{O}}(d^{3/4} \log(K)^{3/8} B_T^{1/4} T^{3/4})$ for `D-LinTS`.

[1] For simplicity we present the Gaussian case. More generally, one may sample from any distribution $\mathcal{D}(\delta, a)$ that satisfies suitable concentration and anti-concentration properties (Kim and Tewari, 2020; Kveton et al., 2020; Vaswani et al., 2020).

3 A BAYESIAN TREATMENT

While randomized algorithms such as `D-RandLinUCB` and `D-LinTS` introduce probabilistic exploration, they remain fundamentally rooted in frequentist point estimation. Because the WRLS objective only outputs point estimates $\hat{\theta}_t$, these algorithms construct a surrogate distribution by injecting artificial Gaussian noise scaled by the matrix V_t. This motivates us to develop a more principled tool that properly grounds these randomized algorithms.

In this section, we introduce Weighted Sequential Bayesian (WSB) inference that is able to handle the time-varying reward parameter θ_t. By maintaining the WSB posteriors, we establish the foundation of randomized exploration, and as a natural byproduct, we obtain a dynamic prior penalty (Π_t) in our concentration bounds that frequentist point estimates do not capture.

While previous work usually focuses on exponential weighting, we consider general weight sequences $\{w_{s,t} \in [0,1] : 1 \le s \le t\}$ that are non-decreasing in s (i.e., $w_{s-1,t} \le w_{s,t}$) and satisfy the multiplicative consistency condition $w_{s,t} \ge w_{t-1,t} w_{s,t-1}$. This includes exponential weights $w_{s,t} = \gamma^{t-s}$ for some $\gamma \in (0,1)$, which we adopt as a canonical example throughout the paper.

WSB posteriors. For analytical tractability, we assume Gaussian reward noise with known variance $\sigma^2 > 0$; this is not a restriction since prior algorithms also use σ when defining UCB confidence levels (see, e.g., Lemma 1). For any Gaussian prior $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$ with mean $\mu_0 \in \mathbb{R}^d$ and covariance $\Sigma_0 \in \mathbb{R}^{d\times d}$, the posterior distribution over θ at round t remains Gaussian. We denote this posterior by $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$, where the posterior mean and covariance are given as follows:

$$\mu_t = \Sigma_t \left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2} \sum_{s=1}^{t} w_{s,t} X_s r_s \right) \tag{2}$$

and

$$\Sigma_t^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{s=1}^{t} w_{s,t} X_s X_s^\top. \tag{3}$$

With exponential weighting, defined by $w_{s,t} = \gamma^{t-s}$ for some $\gamma \in (0,1)$, the updates in (2) and (3) simplify into recursive forms, requiring no storage of past data:

$$\begin{aligned} \mu_t &= \Sigma_t \left(\gamma \Sigma_{t-1}^{-1}\mu_{t-1} + \sigma^{-2} X_t r_t + (1-\gamma)\Sigma_0^{-1}\mu_0 \right) \\ \Sigma_t^{-1} &= \gamma \Sigma_{t-1}^{-1} + \sigma^{-2} X_t X_t^\top + (1-\gamma)\Sigma_0^{-1}. \end{aligned} \tag{4}$$

Remark 1. The WSB posterior corresponds to applying Bayes' rule with tempered likelihoods $\rho_t(\theta) \propto \rho_0(\theta) \prod_{s=1}^{t} p(r_s|X_s,\theta)^{w_{s,t}}$, which defines a generalized Bayesian posterior (Bissiri et al., 2016; Grünwald and van Ommen, 2017). This mirrors the design in non-parametric approaches such as weighted Gaussian Processes (Deng et al., 2022), which is necessary to handle non-stationarity.

Remark 2. Under an uninformative prior ($\mu_0 = \mathbf{0}$, $\Sigma_0^{-1} = \lambda\sigma^2 \mathbb{I}_d$), the WSB posterior mean μ_t recovers the WRLS estimate $\hat{\theta}_t$. Thus, WSB naturally generalizes WRLS-based approaches; while matching their point estimates, it natively provides a full posterior covariance for sampling. Furthermore, while the WRLS initialization penalty ($\lambda\|\theta\|_2^2$) is bounded by a static constant $\sqrt{\lambda}S$, WSB evaluates its prior penalty ($\|\theta - \mu_0\|^2_{\Sigma_0^{-1}}$) through the shrinking posterior covariance Σ_t. This yields a dynamic prior penalty term (Π_t) in our concentration bounds that tightens over time.

Concentration bounds for WSB posteriors. We also provide concentration bounds for the WSB posteriors. Our analysis mirrors the improved decomposition for WRLS from Wang et al. (2023) by separating parameter drift from stochastic noise. A key distinction, however, is that we isolate the prior penalty Π_t from the noise term β_t^{WSB}, allowing it to be controlled separately in Lemma 3.

Lemma 2. *For any horizon $T \in \mathbb{N}_+$, error probability $\delta \in (0,1)$, and prior $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, with probability at least $1-\delta$, the following inequalities hold for all posteriors $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ for all $t \in [T]$ simultaneously:*

$$\|\mu_{t-1} - \theta_t^*\|_{\Sigma_{t-1}^{-1}} \le \alpha_{t-1}^{\mathrm{WSB}} + \beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1}$$

and $\forall x \in \mathcal{X}_t$, $|\langle x, \mu_{t-1} - \theta_t^* \rangle| \le (\alpha_{t-1}^{\mathrm{WSB}} + \beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|x\|_{\Sigma_{t-1}}$, where

- $\alpha_t^{\mathrm{WSB}} = L\sqrt{d}\sum_{k=1}^{t} \sqrt{\sum_{s=1}^{k} w_{s,t}} \|\theta_k^* - \theta_{k+1}^*\|_2$ *is the drift-induced bias,*
- $\beta_t^{\mathrm{WSB}}(\delta') = \sqrt{2\log(\frac{1}{\delta'}) + d\log(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2 \sum_{s=1}^{t} w_{s,t}^2}{d\sigma^2})}$ *is the confidence level evaluated at a given failure allocation δ', and*
- $\Pi_t = \max_{\theta:\|\theta\|_2 \le S} \|\Sigma_0^{-1}(\mu_0 - \theta)\|_{\Sigma_t}$ *is the time-decaying prior term.*

The proof of Lemma 2 is given in Appendix C. The projection bound for arbitrary actions x follows by the dual norm inequality $|\langle x, z\rangle| \le \|x\|_{\Sigma_{t-1}} \|z\|_{\Sigma_{t-1}^{-1}}$, making this step explicit. The drift term α_t^{WSB}, with $w_{s,t} = \gamma^{t-s}$, essentially coincides with α_t^{WRLS} in WRLS that accumulates the effects of non-stationarity. The confidence level β_t^{WSB} captures the stochastic noise only. Unlike β_t^{WRLS}, which also absorbs the regularization term $\sqrt{\lambda}S$, our formulation isolates this regularization effect to the prior term Π_t.

Calculating Π_t exactly requires solving a maximization problem, which is computationally impractical. To address this, we provide two tractable upper bounds on Π_t that will be useful in both analysis and implementation.

Lemma 3. *For any prior $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, the following inequalities hold for all posteriors $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ and $t \in [T]$, simultaneously:*

$$\Pi_t \le \Pi_t^{\mathrm{cxv}} \le \Pi_t^{\Delta},$$

where

- $(\Pi_t^{\text{cxv}})^2 = \|\mu_0\|_{M_t}^2 + S^2\lambda_{\max}(M_t) + 2S\sqrt{\mu_0^\top M_t^2 \mu_0}$
- $\Pi_t^{\Delta} = \|\mu_0\|_{M_t} + \sqrt{\lambda_{\max}(M_t)}S$

with $M_t = \Sigma_0^{-1}\Sigma_t\Sigma_0^{-1}$.

The proof of Lemma 3 can be found in Appendix C. The bound Π_t^{cxv} handles the interaction across the full spectrum of M_t to resolve alignment issues, while the simpler upper bound Π_t^{Δ} remains easier to evaluate and suffices for most theoretical purposes. To explicitly see how this penalty dynamically improves upon the static WRLS bound, we note that as Σ_t contracts, $\Pi_t \to 0$, unlike the fixed $\sqrt{\lambda}S$ term.

3.1 PROOF SKETCH OF LEMMA 2

Lemma 2 provides a unified bound on the *estimation error* around the WSB posterior mean μ_{t-1} and the current reward parameter θ_t^*. To establish this bound, we follow a common approach in the literature (see e.g., Abbasi-Yadkori et al. (2011); Russac et al. (2019); Wang et al. (2023)), using a *noise-free surrogate* to separate the contributions of parameter drift and stochastic noise. Specifically, we define the *noise-free surrogate posterior mean*:

$$\bar{\mu}_t = \Sigma_{t-1}\left(\Sigma_0^{-1}\theta_t^* + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}X_s X_s^\top \theta_s^*\right), \quad (5)$$

which allows us to decompose the *estimation error* $(\mu_{t-1} - \theta_t^*)$ into two terms: a *drift part* $(\bar{\mu}_t - \theta_t^*)$ and a *noise part* $(\mu_{t-1} - \bar{\mu}_t)$, which we can control independently.

The *drift part* captures the deviation caused by non-stationarity. As this part is unaffected by noise, it can be bounded deterministically, mirroring the bias control in WRLS, but with general weighting $w_{s,t}$:

Lemma 4. *Given the prior $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, the following inequalities hold for all posteriors $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ and $t \in \mathbb{N}_+$:*

$$\|\bar{\mu}_t - \theta_t^*\|_{\Sigma_{t-1}^{-1}} \le \alpha_{t-1}^{\text{WSB}}$$

and $\forall x \in \mathcal{X}_t$, $|\langle x, \bar{\mu}_t - \theta_t^*\rangle| \le \alpha_{t-1}^{\text{WSB}}\|x\|_{\Sigma_{t-1}}$, where α_t^{WSB} is defined as in Lemma 2.

The *noise part* accounts for the randomness in the observations and the influence of the initial prior.

Lemma 5. *For any horizon $T \in \mathbb{N}_+$, $\delta \in (0,1)$, and prior $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, with probability at least $1-\delta$, the following inequalities hold for all posteriors and $t \in [T]$, simultaneously:*

$$\|\mu_{t-1} - \bar{\mu}_t\|_{\Sigma_{t-1}^{-1}} \le \beta_{t-1}^{\text{WSB}}(\delta/T) + \Pi_{t-1}$$

and for all $x \in \mathcal{X}_t$ we have

$$|\langle x, \mu_{t-1} - \bar{\mu}_t\rangle| \le (\beta_{t-1}^{\text{WSB}}(\delta/T) + \Pi_{t-1})\|x\|_{\Sigma_{t-1}},$$

where $\beta_t^{\text{WSB}}(\cdot)$ and Π_t are defined as in Lemma 2.

The proofs of Lemmas 4 and 5 are given in Appendix C, which together prove Lemma 2. We note that in addition to isolating the dynamic penalty Π_t, our technical analysis makes a second improvement. Inspired by modern martingale techniques (Howard et al., 2020), we significantly simplify the proof of the self-normalized concentration bound (in Appendix B). Instead of the complex stopping-time sequences traditionally required in the linear bandit literature (Abbasi-Yadkori et al., 2011; Russac et al., 2019), we directly leverage Ville's inequality (Ville, 1939) to establish a clean, self-normalized super-martingale structure. This allows us to elegantly isolate the noise process at any given step before taking a standard union bound over the finite horizon T, providing a streamlined and highly accessible alternative framework for non-stationary concentrations.

4 ALGORITHMS

Building on our WSB formulation in Section 3, we instantiate three algorithms corresponding to widely used exploration strategies: UCB, randomized UCB, and Thompson Sampling (TS). We show that their regrets closely match or improve upon those of their WRLS-based counterparts. Proofs are in Appendix D. Pseudo-code for `WSB-LinUCB`, `WSB-RandLinUCB`, and `WSB-LinTS` in their exponential-weighted form, $w_{s,t} = \gamma^{t-s}$ with $\gamma \in (0,1)$, are presented in Algorithms 1 to 3, respectively.

4.1 DETERMINISTIC EXPLORATION WITH WSB

4.1.1 Upper Confidence Bound

Although WSB natively supports probabilistic exploration, we first instantiate a deterministic UCB algorithm to provide a direct baseline comparison against frequentist methods. At each round t, the `WSB-LinUCB` algorithm selects

$$X_t = \arg\max_{x \in \mathcal{X}_t}\{\langle x, \mu_{t-1}\rangle + (\beta_{t-1}^{\text{WSB}}(\delta/T) + \Pi_{t-1})\|x\|_{\Sigma_{t-1}}\}.$$

As the Bayesian analogue to `LB-WeightUCB`, it simply replaces the WRLS estimate and fixed penalty with our posterior mean μ_{t-1} and dynamic prior term Π_{t-1}.

Theorem 1. *For a horizon $T \in \mathbb{N}_+$, any $\delta \in (0,1)$ and prior $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, with probability at least $1-2\delta$, the following inequality holds for all posteriors $\rho_t(\theta) =$*

Algorithm 1 `WSB-LinUCB` (Weighted Sequential Bayesian Upper Confidence Bound)

Require: Probability $\delta \in (0,1)$, discount factor $\gamma \in (0,1)$, prior dist. $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$

Initialize: $\upsilon = \mathrm{Tr}(\Sigma_0)$
for each round $t \geq 1$ **do**
  Receive action set $\mathcal{X}_t$
  Compute $\beta_{t-1}^{\mathrm{WSB}}(\delta/T)$ and Π_{t-1} according to Lemmas 2 and 3, respectively
  Play action $X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle \mu_{t-1}, x\rangle + (\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|x\|_{\Sigma_{t-1}}\}$ and **receive reward** r_t
  Update posterior (μ_t, Σ_t) according to (4)

$\mathcal{N}(\theta|\mu_t, \Sigma_t)$ *and all* $t \in [T]$ *simultaneously,*

$$R_T \leq 2L\sqrt{\lambda_{\max}(\Sigma_0)} \sum_{t=1}^{T} \alpha_{t-1}^{\mathrm{WSB}} + 2^{3/2}\sigma\sqrt{CdT\Lambda_T}(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0),$$

with $C = \max\{1, L^2\lambda_{\max}(\Sigma_0)/\sigma^2\}$, $\Pi_0 \leq \|\mu_0\|_{\Sigma_0^{-1}} + S\sqrt{\lambda_{\max}(\Sigma_0^{-1})}$ *and* $\Lambda_T = \sum_{t=1}^{T} \log(\frac{1}{w_{t-1,t}}) + \log(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2\sum_{t=1}^{T} w_{t,T}}{d\sigma^2})$.

Theorem 1 holds for weighting schemes $w_{s,t}$ that are non-decreasing and satisfy the multiplicative consistency condition. The following corollary exemplifies this result for the case of exponential weights.

Corollary 1. *Suppose* $w_{s,t} = \gamma^{t-s}$. *For a horizon* $T \in \mathbb{N}_+$, *any* $\gamma \in (1/T, 1)$ *and prior* $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, *with probability at least* $1 - 1/T$, *the following inequality holds for all posteriors* $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ *and all* $t \in [T]$ *simultaneously,*

$$R_T \leq \tilde{\mathcal{O}}(\sqrt{d\lambda_{\max}(\Sigma_0)}B_T(1-\gamma)^{-3/2} + dT\sqrt{1-\gamma}).$$

In particular, setting $\lambda_{\max}(\Sigma_0) = 1/d$ *and* $\gamma = 1 - \max\{1/T, \sqrt{B_T/dT}\}$ *yields* $R_T \leq \tilde{\mathcal{O}}(d\sqrt{T})$ *when the drift is mild* ($B_T < d/T$), *and* $R_T \leq \tilde{\mathcal{O}}(d^{3/4}B_T^{1/4}T^{3/4})$ *when the drift is large* ($B_T \geq d/T$).

The regret of `WSB-LinUCB` matches that of `LB-WeightUCB` while improving upon `D-LinUCB`.

4.2 RANDOMIZED EXPLORATION WITH WSB

As discussed in Section 3, while frequentist algorithms must construct surrogate distributions to achieve randomized exploration, our WSB framework natively provides a full posterior covariance. The following result extends Kim and Tewari (2020, Theorem 7) to the WSB setting, and incorporates the refined weighted analysis of Wang et al. (2023) to handle the drift term, yielding improved regret guarantees. The events $\mathcal{E}^{\mathrm{WSB}}$, $\mathcal{E}_t^{\mathrm{Conc.}}$, and $\mathcal{E}_t^{\mathrm{Anti\text{-}Conc.}}$ are defined in Appendix D.2.

Theorem 2. *Let* $p_1, p_2, p_3 \in (0,1)$, *a horizon* $T \in \mathbb{N}_+$, *and suppose there exists constants* $c_1, c_2 \geq 1$, *which may depend on* T *such that* $\mathbb{P}(\mathcal{E}^{WSB}) \geq 1 - p_1$, $\mathbb{P}(\mathcal{E}_t^{Conc.}) \geq 1 - p_2$, *and* $\mathbb{P}(\mathcal{E}_t^{Anti\text{-}Conc.}) \geq p_3$. *Then, for any prior* $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, *the following inequality holds for all posteriors* $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ *and all* $t \in [T]$ *simultaneously,*

$$R_T \leq 2L\sqrt{\lambda_{\max}(\Sigma_0)}\sum_{t=1}^{T}\alpha_{t-1}^{\mathrm{WSB}} + (c_1+c_2)\left(1 + \frac{2}{p_3 - p_2}\right)\sqrt{dT\Lambda_T} + T(p_1+p_2),$$

where Λ_T *is as in Theorem 1.*

4.2.1 Randomized Upper Confidence Bound

`WSB-RandLinUCB` adapts the randomized UCB principle to the Bayesian WSB setting. At each round t, a sample $\eta_t \sim \mathcal{N}(0, a^2)$ is drawn, and the action is chosen as

$$X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle x, \mu_{t-1}\rangle + \eta_t\|x\|_{\Sigma_{t-1}}\}.$$

Corollary 2. *Suppose* $w_{s,t} = \gamma^{t-s}$, $c_1 = \beta_T^{WSB}(1/T^2) + \Pi_0$, $c_2 = a\sqrt{2\log(T/2)}$, *and* $a^2 = 14c_1^2$. *For any* $\gamma \in (1/T, 1)$ *and prior* $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, *with probability at least* $1 - 1/T$, *the following inequality holds for all posteriors* $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ *and all* $t \in [T]$ *simultaneously,*

$$R_T \leq \tilde{\mathcal{O}}(\sqrt{d\lambda_{\max}(\Sigma_0)}B_T(1-\gamma)^{-3/2} + dT\sqrt{1-\gamma}).$$

In particular, setting $\lambda_{\max}(\Sigma_0) = 1/d$ *and* $\gamma = 1 - \max\{1/T, \sqrt{B_T/dT}\}$ *yields* $R_T \leq \tilde{\mathcal{O}}(d\sqrt{T})$ *when the drift is mild* ($B_T < d/T$), *and* $R_T \leq \tilde{\mathcal{O}}(d^{3/4}B_T^{1/4}T^{3/4})$ *when the drift is large* ($B_T \geq d/T$).

`WSB-RandLinUCB` improves upon the regret of the randomized UCB algorithm `D-RandLinUCB` by a factor of $d^{1/8}$, and removes the reliance on complex *local norms* by incorporating the refined analysis of Wang et al. (2023). With this improvement, `WSB-RandLinUCB` achieves the same order of regret as `WSB-LinUCB` while reducing over-conservatism in the arm-selection criterion, often resulting in better empirical performance.

4.2.2 Thompson Sampling

`WSB-LinTS` implements TS by drawing a parameter vector directly from the WSB posterior and acting greedily with respect to it. At each round t, one draws $\tilde{\mu}_{t-1} = \mu_{t-1} + \Sigma_{t-1}^{1/2}\eta_t$ with $\eta_t \sim \mathcal{N}(0, a^2\mathbb{I}_d)$, and chooses

$$X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle x, \tilde{\mu}_{t-1}\rangle\}.$$

Algorithm 2 `WSB-RandLinUCB` (Weighted Sequential Bayesian Randomized Upper Confidence Bound)

Require: Probability $\delta \in (0,1)$, discount factor $\gamma \in (0,1)$, prior dist. $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, confidence level $a > 0$

1: **for** each round $t \geq 1$ **do**
2: Receive action set $\mathcal{X}_t$
3: Randomly sample $\eta_t \sim \mathcal{N}(0, a^2)$
4: **Play action** $X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle \mu_{t-1}, x \rangle + \eta_t \|x\|_{\Sigma_{t-1}}\}$ and **receive reward** r_t
5: **Update posterior** (μ_t, Σ_t) according to (4)

Algorithm 3 `WSB-LinTS` (Weighted Sequential Bayesian Thompson Sampling)

Require: Probability $\delta \in (0,1)$, discount factor $\gamma \in (0,1)$, prior dist. $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, confidence level $a > 0$

1: **for** each round $t \geq 1$ **do**
2: Receive action set $\mathcal{X}_t$
3: Randomly sample $\tilde{\mu}_{t-1} = \mu_{t-1} + \Sigma_{t-1}^{1/2} \eta_t$ with $\eta_t \sim \mathcal{N}(0, a^2 \mathbb{I}_d)$
4: **Play action** $X_t = \arg\max_{x \in \mathcal{X}_t} \{\langle \tilde{\mu}_{t-1}, x \rangle\}$ and **receive reward** r_t
5: **Update posterior** (μ_t, Σ_t) according to (4)

Corollary 3. *Suppose* $w_{s,t} = \gamma^{t-s}$, $c_1 = \beta_T^{WSB}(1/T^2) + \Pi_0$, $c_2 = a\sqrt{2\log(KT/2)}$, *and* $a^2 = 14c_1^2$. *For any* $\gamma \in (1/T, 1)$ *and prior* $\pi(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, *with probability at least* $1 - 1/T$, *the following inequality holds for all posteriors* $\rho_t(\theta) = \mathcal{N}(\theta|\mu_t, \Sigma_t)$ *and all* $t \in [T]$ *simultaneously,*

$$R_T \leq \tilde{\mathcal{O}}\left(\sqrt{d\lambda_{\max}(\Sigma_0)} B_T (1-\gamma)^{-3/2} + dT\sqrt{\log(K)(1-\gamma)}\right).$$

In particular, setting $\lambda_{\max}(\Sigma_0) = 1/d$ *and* $\gamma = 1 - \max\{1/T, \sqrt{B_T/d\sqrt{\log(K)}T}\}$ *yields* $R_T \leq \tilde{\mathcal{O}}(d\sqrt{T\log(K)})$ *when the drift is mild* ($B_T < d\sqrt{\log(K)}/T$), *and* $R_T \leq \tilde{\mathcal{O}}(d^{3/4}\log(K)^{3/8} B_T^{1/4} T^{3/4})$ *when the drift is large* ($B_T \geq d\sqrt{\log(K)}/T$).

`WSB-LinTS` achieves improved regret guarantees compared to `D-LinTS`. Similarly, `WSB-LinTS` improves upon the regret of `D-LinTS` by a factor of $d^{1/8}$. Again, the theoretical improvement is mainly due to the technique by Wang et al. (2023), our WSB formulation provides a native, principled posterior for Thompson Sampling.

Our derived frequentist regret of $\tilde{O}(d^{3/4} B_T^{1/4} T^{3/4})$ matches the current state-of-the-art for smooth and weighted strategies in non-stationary linear bandits. While this bound is a factor of $(dT/B_T)^{1/12}$ above the theoretical lower bound of $\Omega(d^{2/3} B_T^{1/3} T^{2/3})$ established for this problem class, this gap reflects open challenge in the literature: achieving minimax optimal rates while relying solely on smooth, continuous adaptation. We note that Wei and Luo (2021) successfully match this lower bound; however, their approach relies on a complex, black-box master-base algorithmic structure that requires periodic resets. In contrast, weighted strategies like WSB are highly appealing in practice exactly because they offer continuous adaptation without the need for forced restarts or fixed memory budgets. Ultimately, WSB achieves highly competitive frequentist performance for continuous methods while offering a more principled foundation for uncertainty quantification.

5 EXPERIMENTS

Setting. We study two synthetic non-stationary scenarios designed to capture both abrupt and gradual changes. In both scenarios, the time horizon $T = 4000$ and the reward noise $\sigma^2 = 0.15$. In both scenarios, the unknown parameter sequence $\{\theta_t^*\}$ evolves on the unit sphere in $\mathbb{R}^d$, with variation across all coordinates. Arms are sampled from the same unit sphere with $K = |\mathcal{X}_t| = 48$. This ensures that Assumption 1 holds with $L = S = 1$. In the *abruptly changing* scenario, $\{\theta_t^*\}$ is piecewise constant and switches between four spherical anchors at every $T/4$ rounds. In the *slowly varying* scenario, $\{\theta_t^*\}$ moves gradually between these anchors along shortest paths on the sphere. The discount parameter $\gamma = 1 - \max\{1/T, \sqrt{B_T/dT}\}$, except for `D-LinTS` and `WSB-LinTS`, where $\gamma = 1 - \max\{1/T, \sqrt{B_T/d\sqrt{\log(K)}T}\}$. Since the sequence $\{\theta_t^*\}$ is known, we use the exact variation budget B_T.

All WSB-based algorithms use $w_{s,t} = \gamma^{t-s}$ and the same Gaussian prior $\mathcal{N}(0, \mathbb{I}_d)$. Following Kim and Tewari (2020), we use a truncated Gaussian with zero mean and standard deviation σ for both `D-RandLinUCB` and `WSB-RandLinUCB`. This ensures that the randomly sampled confidence level lies within the upper confidence bounds used by `D-LinUCB`, `LB-WeightUCB`, and `WSB-LinUCB` with high probability. As in Kim and Tewari (2020), we adopt the non-inflated variant by setting the scaling factor $a = 1$ for `D-RandLinUCB`, `WSB-RandLinUCB`, `D-LinTS`, and `WSB-LinTS`.

Results. The cumulative regrets across all evaluated dimensions are reported in Table 2, and the corresponding round-by-round cumulative regret trajectories are provided in Appendix E. Across both *abruptly changing* and *slowly drifting* scenarios, randomized exploration substantially outperforms deterministic UCB-based exploration. Among the deterministic methods, `WSB-LinUCB` improves over `LB-WeightUCB` in lower dimensions, but becomes slightly worse as the dimension increases. However, neither `WSB-LinUCB` nor `LB-WeightUCB` outperforms `D-LinUCB`, which relies on the more complex *local norm*. In contrast, the randomized WSB variants consistently improve upon their WRLS-based coun-

Table 2: Cumulative regret at horizon (T) (mean $\pm$ std) across 100 independent trials for varying dimensions (d), under both abruptly changing and slowly drifting scenarios.

Abruptly Changing

ALGORITHM	$d = 2$	$d = 4$	$d = 6$	$d = 16$	$d = 32$
`D-LinUCB` (Russac et al., 2019)	455.8 ± 21.4	713.0 ± 73.2	867.6 ± 90.5	1277.9 ± 99.8	1290.0 ± 125.9
`LB-WeightUCB` (Wang et al., 2023)	522.5 ± 19.4	822.7 ± 88.7	993.4 ± 102.1	1360.0 ± 106.8	1324.6 ± 128.5
`WSB-LinUCB` (**Ours**, Algorithm 1)	470.4 ± 18.8	780.6 ± 83.2	990.2 ± 100.8	1477.9 ± 116.0	1427.5 ± 136.6
`D-RandLinUCB` (Kim and Tewari, 2020)	293.8 ± 20.6	381.7 ± 42.9	402.4 ± 51.7	476.4 ± 41.8	503.6 ± 49.9
`WSB-RandLinUCB` (**Ours**, Algorithm 2)	267.5 ± 20.7	355.0 ± 44.9	366.5 ± 53.9	433.7 ± 44.1	474.2 ± 50.7
`D-LinTS` (Kim and Tewari, 2020)	375.8 ± 19.6	598.8 ± 49.2	720.3 ± 59.9	1077.0 ± 74.5	1138.6 ± 101.5
`WSB-LinTS` (**Ours**, Algorithm 3)	328.9 ± 21.2	458.9 ± 50.4	501.3 ± 51.5	719.5 ± 47.8	845.5 ± 61.0

Slowly Drifting

ALGORITHM	$d = 2$	$d = 4$	$d = 6$	$d = 16$	$d = 32$
`D-LinUCB` (Russac et al., 2019)	440.4 ± 23.3	674.9 ± 167.8	857.8 ± 167.1	1346.1 ± 221.7	1284.0 ± 263.5
`LB-WeightUCB` (Wang et al., 2023)	504.6 ± 24.3	776.5 ± 194.3	962.7 ± 191.6	1405.7 ± 234.3	1301.7 ± 267.0
`WSB-LinUCB` (**Ours**, Algorithm 1)	446.3 ± 23.5	737.9 ± 185.4	969.0 ± 194.8	1531.2 ± 257.7	1405.3 ± 288.4
`D-RandLinUCB` (Kim and Tewari, 2020)	169.3 ± 14.0	236.1 ± 36.2	289.5 ± 29.7	403.5 ± 35.3	435.5 ± 37.1
`WSB-RandLinUCB` (**Ours**, Algorithm 2)	130.8 ± 14.5	194.1 ± 28.2	241.4 ± 24.3	367.3 ± 23.9	405.5 ± 33.6
`D-LinTS` (Kim and Tewari, 2020)	210.9 ± 16.6	437.9 ± 92.0	608.7 ± 94.9	1101.1 ± 146.8	1128.9 ± 210.1
`WSB-LinTS` (**Ours**, Algorithm 3)	136.6 ± 13.9	268.3 ± 49.4	368.2 ± 45.2	702.7 ± 63.5	840.6 ± 107.7

terparts. `WSB-RandLinUCB` achieves lower regret than `D-RandLinUCB` across all dimensions and both scenarios, while `WSB-LinTS` substantially outperforms `D-LinTS`. These results indicate that the practical benefits of the WSB posterior are most pronounced for randomized exploration. In this case, the posterior covariance can be used directly for uncertainty-driven action selection, while the prior influence is captured through the dynamic prior term Π_t rather than through a fixed worst-case initialization penalty. This yields less conservative randomized exploration without relying on surrogate posterior distributions or *local norms*.

Ablation. An ablation study is presented in Appendix F, where we examine the sensitivity of the WSB-based algorithms to prior misspecification. We vary the norm of the prior mean as $\|\mu_0\|_2 \in \{1, 10, 100\}$, while the true parameter sequence satisfies the bound $S = 1$. The results show that moderate misspecification, such as $\|\mu_0\|_2 = 10$, leads to a noticeable but controlled increase in regret across both abruptly changing and slowly drifting environments. In contrast, severe misspecification with $\|\mu_0\|_2 = 100$ can substantially degrade performance, particularly for the randomized exploration methods. This highlights that the dynamic prior term Π_t is not merely a technical artifact: when the prior mean is far outside the true parameter scale, its influence can dominate the early posterior updates and delay adaptation. Overall, the ablation confirms that WSB can accommodate biased priors, but also emphasizes the practical importance of choosing a prior scale compatible with the assumed parameter bound S.

Lastly, as discussed in the introduction, GP-based methods provide a Bayesian treatment of uncertainty, but they do so through kernel representations whose size grows with the accumulated history. In long-horizon bandit problems, this growing representation can make posterior updates and action selection computationally prohibitive. WSB takes a complementary parametric route: by maintaining a Gaussian posterior over the finite-dimensional reward parameter, it retains Bayesian uncertainty quantification for both deterministic and randomized exploration while preserving efficient recursive updates.

6 CONCLUSION

We introduce a new concentration inequality for WSB posteriors that incorporates prior information. Building on this result, we instantiate the WSB framework through three exploration algorithms, `WSB-LinUCB`, `WSB-RandLinUCB`, and `WSB-LinTS`, which achieve improved theoretical and empirical regrets. To support this framework, we provide a simplified proof for the time-uniform concentration of vector-valued martingales using Ville's inequality, offering a clean, alternative subroutine of independent interest for future linear bandit literature. Together, these results demonstrate that Bayesian principles, when paired with weighted updates, yield practical and theoretically sound algorithms for non-stationary sequential decision-making. While our current analysis relies on a known total variation budget B_T to optimize the weighting parameter, the WSB framework natively accommodates master-base meta-tuning paradigms or online restart heuristics, making fully automated, data-driven drift adaptation a promising direction for future work.

Author Contributions

All authors contributed to the conceptualization and design of the research. N. W. and Y.-S. W. developed and refined the theoretical analysis. N. W., .A. A. and M. K. designed and executed the empirical evaluations. All authors participated in the drafting and critical revision of the manuscript.

Acknowledgements

This work was supported by grants from the Novo Nordisk Foundation (NNF) under grant number NNF21OC0070621 and the Carlsberg Foundation (CF) under grant number CF21-0250. Y.-S. W. was also supported by the Academia Sinica Postdoctoral Scholar Program, grant number AS-PD-1151-M15-2. The authors thank the anonymous reviewers for their insightful feedback, which helped improve the clarity and technical precision of the paper.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2016.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 2012.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under nonstationarity. *Management Science*, 2022.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, 2008.

Yuntian Deng, Xingyu Zhou, Baekjin Kim, Ambuj Tewari, Abhishek Gupta, and Ness Shroff. Weighted Gaussian process bandits for non-stationary environments. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. Regret bounds for generalized linear bandits under parameter drift. *arXiv preprint arXiv:2103.05750*, 2021a.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. A technical note on non-stationary parametric bandits: Existing mistakes and preliminary solutions. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2021b.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2011.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 2017.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Baekjin Kim and Ambuj Tewari. Randomized exploration for non-stationary stochastic linear bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Andreas Krause and Cheng Ong. Contextual gaussian process bandit optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2010.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 2010.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized UCB for bandit problems. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Jean Ville. Étude critique de la notion de collectif. *Bulletin of the American Mathematical Society*, 1939.

Jing Wang, Peng Zhao, and Zhi-Hua Zhou. Revisiting weighted strategy for non-stationary parametric bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Peng Zhao and Lijun Zhang. Non-stationary linear bandits revisited. *arXiv preprint arXiv:2103.05324*, 2021.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Supplementary Material: Weighted Sequential Bayesian Inference for Non-Stationary Linear Contextual Bandits

Nicklas Werge[1] **Yi-Shan Wu**[1,2] **Abdullah Akgül**[1] **Melih Kandemir**[1]

[1]Department of Mathematics and Computer Science, University of Southern Denmark
[2]Research Center for Information Technology Innovation, Academia Sinica

A SOME TECHNICAL PROPOSITIONS

Proposition A.1 (Determinant inequalities)**.** *For some $p \in \mathbb{N}_+$, define $\overline{\Sigma}_t^{-1} = \overline{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^t w_{s,t}^p X_s X_s^\top$, where $\overline{\Sigma}_0 \succ 0$. Under Assumption 1, we have*

$$\det\left(\overline{\Sigma}_t^{-1}\right) \leq \left(\frac{\mathrm{Tr}(\overline{\Sigma}_0^{-1}) + \frac{L^2}{\sigma^2}\sum_{s=1}^t w_{s,t}^p}{d}\right)^d \quad \textit{and} \quad \frac{\det\left(\overline{\Sigma}_0\right)}{\det\left(\overline{\Sigma}_t\right)} \leq \left(1 + \frac{\mathrm{Tr}(\overline{\Sigma}_0)L^2\sum_{s=1}^t w_{s,t}^p}{d\sigma^2}\right)^d.$$

Proof of Proposition A.1. Let $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of $\overline{\Sigma}_t$. Recall that $\overline{\Sigma}_t$ is positive definite, thus, the eigenvalues $\lambda_1, \ldots, \lambda_d$ are positive. Also, note that $\det(\overline{\Sigma}_t^{-1}) = \prod_{i=1}^d \lambda_i^{-1}$ and $\mathrm{Tr}(\overline{\Sigma}_t^{-1}) = \sum_{i=1}^d \lambda_i^{-1}$. Hence, $\det(\overline{\Sigma}_t^{-1}) \leq (\mathrm{Tr}(\overline{\Sigma}_t^{-1})/d)^d$ by the inequality of arithmetic and geometric means. Next, since $\|X_t\|_2 \leq L$ for any t (Assumption 1), we obtain that

$$\begin{aligned}
\mathrm{Tr}(\overline{\Sigma}_t^{-1}) &= \mathrm{Tr}(\overline{\Sigma}_0^{-1}) + \frac{1}{\sigma^2}\sum_{s=1}^t w_{s,t}^p \,\mathrm{Tr}(X_s X_s^\top) \\
&= \mathrm{Tr}(\overline{\Sigma}_0^{-1}) + \frac{1}{\sigma^2}\sum_{s=1}^t w_{s,t}^p \|X_s\|_2^2 \\
&\leq \mathrm{Tr}(\overline{\Sigma}_0^{-1}) + \frac{L^2}{\sigma^2}\sum_{s=1}^t w_{s,t}^p,
\end{aligned}$$

which shows the first inequality of the proposition. For the second inequality, we use that

$$\begin{aligned}
\frac{\det\left(\overline{\Sigma}_0\right)}{\det\left(\overline{\Sigma}_t\right)} &= \det\left(\overline{\Sigma}_0\overline{\Sigma}_t^{-1}\right) \\
&= \det\left(\overline{\Sigma}_0\left(\overline{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^t w_{s,t}^p X_s X_s^\top\right)\right) \\
&= \det\left(\mathbb{I}_d + \frac{1}{\sigma^2}\overline{\Sigma}_0\sum_{s=1}^t w_{s,t}^p X_s X_s^\top\right).
\end{aligned}$$

Next, as $\mathrm{Tr}(\mathbb{I}_d + \frac{1}{\sigma^2}\overline{\Sigma}_0\sum_{s=1}^t w_{s,t}^p X_s X_s^\top) = \mathrm{Tr}(\mathbb{I}_d) + \frac{1}{\sigma^2}\mathrm{Tr}(\overline{\Sigma}_0\sum_{s=1}^t w_{s,t}^p X_s X_s^\top)$, we only need to notice that $\mathrm{Tr}(\overline{\Sigma}_0\sum_{s=1}^t w_{s,t}^p X_s X_s^\top) \leq \mathrm{Tr}(\overline{\Sigma}_0)L^2\sum_{s=1}^t w_{s,t}^p$ by Cauchy-Schwarz inequality. □

Proposition A.2 (Mahalanobis convexity bound)**.** *For some $p \in \mathbb{N}_+$, define $\overline{\Sigma}_t^{-1} = \overline{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t} w_{s,t}^p X_s X_s^\top$, where $\overline{\Sigma}_0 \succ 0$. Under Assumption 1, we have*

$$\max_{\theta:\|\theta\|_2 \le S} \|\overline{\Sigma}_0^{-1}(\mu-\theta)\|_{\overline{\Sigma}_t} \le \sqrt{\|\mu\|_{\overline{M}_t}^2 + S^2\lambda_{\max}(\overline{M}_t) + 2S\sqrt{\mu^\top \overline{M}_t^2 \mu}},$$

for any $\mu \in \mathbb{R}^d$, where $\overline{M}_t = \overline{\Sigma}_0^{-1}\overline{\Sigma}_t\overline{\Sigma}_0^{-1}$.

Proof of Proposition A.2. Let $\overline{M}_t = U\Lambda U^\top$ be an eigendecomposition with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ ordered so that $\lambda_{\max}(\overline{M}_t) = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{d-1} \ge \lambda_d = \lambda_{\min}(\overline{M}_t) > 0$. Since U is orthogonal ($U^\top U = UU^\top = \mathbb{I}_d$), the Euclidean norm is invariant: $\|x\|_2 = \|U^\top x\|_2$. Thus, for any $\theta \in \mathbb{R}^d$,

$$\|\mu-\theta\|_{\overline{M}_t} = \|\overline{M}_t^{1/2}(\mu-\theta)\|_2 = \|U^\top \overline{M}_t^{1/2}(\mu-\theta)\|_2 = \|U^\top U\Lambda^{1/2}U^\top(\mu-\theta)\|_2 = \|\Lambda^{1/2}U^\top\mu - \Lambda^{1/2}U^\top\theta\|_2.$$

Equivalently, with $b = \overline{M}_t^{1/2}\mu$ (such that $U^\top b = \Lambda^{1/2}U^\top\mu$) and $z = U^\top\theta$ (such that $\|z\|_2 = \|\theta\|_2 \le S$), we have

$$\|\mu-\theta\|_{\overline{M}_t} = \|U^\top b - \Lambda^{1/2} z\|_2.$$

The map $z \mapsto \|U^\top b - \Lambda^{1/2}z\|_2$ is convex; hence a maximizer over the convex set $\{z : \|z\|_2 \le S\}$ lies on the boundary $\|z\|_2 = S$. Hence, we can write $z = Se$ with $\|e\|_2 = 1$ and consider

$$\phi(e) = \|U^\top b - S\Lambda^{1/2}e\|_2^2 = \|b\|_2^2 + S^2 e^\top \Lambda e - 2Sb^\top\Lambda^{1/2}e.$$

Since $\Lambda \preceq \lambda_{\max}(\overline{M}_t)\mathbb{I}_d$ and $\|e\|_2 = 1$, we have $e^\top\Lambda e \le \lambda_{\max}(\overline{M}_t)$. For the cross term, applying the Cauchy-Schwarz inequality directly yields:

$$-2Sb^\top\Lambda^{1/2}e \le 2S\|\Lambda^{1/2}b\|_2\|e\|_2 = 2S\|\Lambda^{1/2}b\|_2.$$

Substituting these bounds back into $\phi(e)$ gives:

$$\phi(e) \le \|b\|_2^2 + S^2\lambda_{\max}(\overline{M}_t) + 2S\|\Lambda^{1/2}b\|_2.$$

Notice that $\|b\|_2^2 = \mu^\top\overline{M}_t\mu = \|\mu\|_{\overline{M}_t}^2$, and for the cross-term argument, we can expand $\|\Lambda^{1/2}b\|_2 = \sqrt{b^\top\Lambda b} = \sqrt{\mu^\top\overline{M}_t^{1/2}(U\Lambda U^\top)\overline{M}_t^{1/2}\mu} = \sqrt{\mu^\top\overline{M}_t^2\mu}$. Taking the square root on both sides of our inequality yields:

$$\|\mu-\theta\|_{\overline{M}_t} \le \sqrt{\|\mu\|_{\overline{M}_t}^2 + S^2\lambda_{\max}(\overline{M}_t) + 2S\sqrt{\mu^\top\overline{M}_t^2\mu}},$$

for every θ with $\|\theta\|_2 \le S$. Maximizing over the boundary completes the proof. □

B SELF-NORMALIZED CONCENTRATION INEQUALITY FOR VECTOR-VALUED MARTINGALES

In this appendix, we provide a simplified proof of the self-normalized concentration inequality for vector-valued martingales (Abbasi-Yadkori et al., 2011) using Ville's inequality (Ville, 1939), and provide its weighted corollary, which we rely on to bound the stochastic noise in Appendix C.

We begin by restating the classical theorem. Let $\{\mathcal{F}_t\}_{t\ge0}$ be a filtration. Let $\{\epsilon_t\}_{t\ge1}$ be a real-valued stochastic process such that ϵ_t is $\mathcal{F}_t$-measurable and conditionally σ-sub-Gaussian, meaning that for all $\nu \in \mathbb{R}$, $\mathbb{E}[\exp(\nu\epsilon_t) \mid \mathcal{F}_{t-1}] \le \exp(\nu^2\sigma^2/2)$ almost surely. Let $\{X_t\}_{t\ge1}$ be an $\mathbb{R}^d$-valued stochastic process where X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \ge 0$, define

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \quad \text{and} \quad S_t = \sum_{s=1}^{t}\epsilon_s X_s.$$

Theorem B.1 (Abbasi-Yadkori et al. (2011), Theorem 1). *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following inequality holds for all $t \geq 0$:*

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

For our non-stationary analysis in Appendix C, we need a bound that applies to a predictable sequence of weights $\{\kappa_t\}_{t \geq 1}$ (i.e., κ_t is $\mathcal{F}_{t-1}$-measurable), which can be obtained as a direct corollary of Theorem B.1.

Corollary B.1 (Weighted Extension). *Let Λ be a predictable positive-definite prior covariance matrix. Define $\overline{\Lambda}_t = \Lambda + \sum_{s=1}^{t} \kappa_s^2 X_s X_s^\top$ and $S_t = \sum_{s=1}^{t} \kappa_s X_s \epsilon_s$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $t \geq 0$,*

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

Proof. Because the scalar weight κ_t is $\mathcal{F}_{t-1}$-measurable, the product $\tilde{X}_t = \kappa_t X_t$ remains an $\mathcal{F}_{t-1}$-measurable, predictable random vector. Directly applying Theorem B.1 by substituting X_t with $\tilde{X}_t$ exactly yields the stated result. □

B.1 A SIMPLIFIED PROOF VIA VILLE'S INEQUALITY

We now provide the modernized proof for the bound in Theorem B.1. We rely on Ville's inequality, a time-uniform extension of Markov's inequality for non-negative super-martingales. A friendly proof and further discussions can be found in Howard et al. (2020).

Lemma B.1 (Ville's Inequality). *Let $(M_t)_{t \geq 0}$ be a non-negative super-martingale with respect to a filtration $(\mathcal{F}_t)_{t \geq 0}$. For any starting time $t_0 \in \mathbb{N}$ and threshold $u > 0$,*

$$\mathbb{P}(\exists t \geq t_0 : M_t \geq u) \leq \frac{\mathbb{E}[M_{t_0}]}{u}.$$

Modernized Proof of Theorem B.1. We aim to show that with probability at least $1-\delta$, for all $t \geq 0$:

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

This is equivalent to showing that with probability at most δ, there exists $t \geq 0$ such that:

$$L_t \equiv \left(\frac{\det(V)}{\det(\overline{V}_t)} \right)^{1/2} \exp \left(\frac{\|S_t\|^2_{\overline{V}_t^{-1}}}{2\sigma^2} \right) \geq \frac{1}{\delta}.$$

If we can show that the process $(L_t)_{t \geq 0}$ is a non-negative super-martingale with initial expected value $\mathbb{E}[L_0] = 1$, we can apply Ville's Inequality (Lemma B.1) with the threshold $u = 1/\delta$:

$$\mathbb{P}(\exists t \geq 0 : L_t \geq 1/\delta) \leq \frac{\mathbb{E}[L_0]}{1/\delta} = \delta,$$

which would immediately imply the theorem.

We construct this super-martingale by the method of mixtures. Consider the exponential process for any fixed direction $x \in \mathbb{R}^d$:

$$M_t(x) = \exp \left(\frac{1}{\sigma} x^\top S_t - \frac{1}{2} x^\top (\overline{V}_t - V) x \right). \tag{B.1}$$

Since X_t and $M_{t-1}(x)$ are $\mathcal{F}_{t-1}$-measurable, we have:

$$\begin{aligned}
\mathbb{E}[M_t(x) \mid \mathcal{F}_{t-1}] &= M_{t-1}(x) \exp \left(-\frac{1}{2} (x^\top X_t)^2 \right) \mathbb{E} \left[\exp \left(\frac{\epsilon_t x^\top X_t}{\sigma} \right) \,\middle|\, \mathcal{F}_{t-1} \right] \\
&\leq M_{t-1}(x) \exp \left(-\frac{1}{2} (x^\top X_t)^2 \right) \exp \left(\frac{(x^\top X_t / \sigma)^2 \sigma^2}{2} \right) \\
&= M_{t-1}(x),
\end{aligned}$$

where the inequality uses the conditionally σ-sub-Gaussian property of ϵ_t. Thus, $(M_t(x))_{t\geq 0}$ is a non-negative super-martingale with $M_0(x) = \exp(0) = 1$.

To control the deviation in all directions simultaneously, we integrate the fixed super-martingales over the probability density $h(x)$ of the Gaussian prior $\mathcal{N}(0, V^{-1})$. The mixture process $M_t = \int_{\mathbb{R}^d} M_t(x)h(x)dx$ is also a non-negative super-martingale with initial value $\mathbb{E}[M_0] = 1$ due to Fubini's theorem:

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = \int_{\mathbb{R}^d} \mathbb{E}[M_t(x) \mid \mathcal{F}_{t-1}]h(x)dx \leq \int_{\mathbb{R}^d} M_{t-1}(x)h(x)dx = M_{t-1}.$$

Substituting the definitions of $M_t(x)$ and $h(x)$ into the integral:

$$\begin{aligned} M_t &= \int_{\mathbb{R}^d} \exp\left(\frac{x^\top S_t}{\sigma} - \frac{1}{2}x^\top(\overline{V}_t - V)x\right) \frac{\det(V)^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}x^\top V x\right) dx \\ &= \frac{\det(V)^{1/2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\left(\frac{x^\top S_t}{\sigma} - \frac{1}{2}x^\top \overline{V}_t x\right) dx. \end{aligned}$$

Applying the standard Gaussian integral identity $\int \exp(b^\top x - \frac{1}{2}x^\top Ax)dx = \frac{(2\pi)^{d/2}}{\det(A)^{1/2}} \exp\left(\frac{1}{2}b^\top A^{-1}b\right)$ with $A = \overline{V}_t$ and $b = S_t/\sigma$, we obtain:

$$M_t = \frac{\det(V)^{1/2}}{(2\pi)^{d/2}} \cdot \frac{(2\pi)^{d/2}}{\det(\overline{V}_t)^{1/2}} \exp\left(\frac{1}{2}\left(\frac{S_t}{\sigma}\right)^\top \overline{V}_t^{-1}\left(\frac{S_t}{\sigma}\right)\right) = \left(\frac{\det(V)}{\det(\overline{V}_t)}\right)^{1/2} \exp\left(\frac{\|S_t\|^2_{\overline{V}_t^{-1}}}{2\sigma^2}\right).$$

Therefore, $M_t = L_t$. Applying Ville's inequality to this mixture process completes the proof. □

C A BAYESIAN TREATMENT OF NON-STATIONARY LINEAR CONTEXTUAL BANDITS

This appendix provides the detailed proofs for the results stated in Section 3, which establish high-probability confidence bounds for the WSB posteriors. In particular, Lemma 2 establishes a uniform deviation inequality for Bayesian posteriors over a finite horizon T. Unlike prior work, our analysis avoids the introduction of an auxiliary covariance matrix by building on the refined framework of Wang et al. (2023).

We structure the appendix as follows: We begin by analyzing the two components of the estimation error separately — namely, the effect of parameter drift (Lemma 4) and stochastic noise including prior influence (Lemma 5). We then combine these results via a finite-horizon union bound to obtain the full posterior confidence bound stated in Lemma 2. Finally, we prove the upper bounds on the prior term (Lemma 3).

Proof of Lemma 4. We aim to bound the drift-induced deviation between the surrogate posterior mean $\bar{\mu}_t$ in (5) and the current reward parameter θ_t^*:

$$\bar{\mu}_t - \theta_t^* = \Sigma_{t-1}\left(\Sigma_0^{-1}\theta_t^* + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}X_s X_s^\top \theta_s^*\right) - \theta_t^*.$$

Using the definition of the posterior covariance Σ_{t-1}^{-1} from (3) gives

$$\begin{aligned} \bar{\mu}_t - \theta_t^* &= \Sigma_{t-1}\left(\Sigma_0^{-1}\theta_t^* + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}X_s X_s^\top \theta_s^* - \Sigma_{t-1}^{-1}\theta_t^*\right) \\ &= \Sigma_{t-1}\left(\Sigma_0^{-1}\theta_t^* + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}X_s X_s^\top \theta_s^* - \Sigma_0^{-1}\theta_t^* - \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}X_s X_s^\top \theta_t^*\right) \\ &= \Sigma_{t-1}\left(\frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}X_s X_s^\top (\theta_s^* - \theta_t^*)\right). \end{aligned}$$

Now, for any arbitrary vector $x \in \mathbb{R}^d$, we have

$$\langle x, \bar{\mu}_t - \theta_t^* \rangle = \left\langle x, \Sigma_{t-1} \left(\frac{1}{\sigma^2} \sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top (\theta_s^* - \theta_t^*) \right) \right\rangle,$$

which by Cauchy-Schwarz inequality yields the following inequality:

$$|\langle x, \bar{\mu}_t - \theta_t^* \rangle| \leq \|x\|_{\Sigma_{t-1}} \left\| \frac{1}{\sigma^2} \sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top (\theta_s^* - \theta_t^*) \right\|_{\Sigma_{t-1}}.$$

We now focus on bounding the right-hand norm. We first expand the inner sum over variations (i.e., $\theta_s^* - \theta_t^* = \sum_{k=s}^{t-1} (\theta_k^* - \theta_{k+1}^*)$), swap the order of summation, apply the triangle inequality, use the Cauchy-Schwarz inequality, invoke Assumption 1, and then apply the triangle inequality once more:

$$\begin{aligned}
\left\| \frac{1}{\sigma^2} \sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top (\theta_s^* - \theta_t^*) \right\|_{\Sigma_{t-1}} &= \left\| \frac{1}{\sigma^2} \sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top \left(\sum_{k=s}^{t-1} (\theta_k^* - \theta_{k+1}^*) \right) \right\|_{\Sigma_{t-1}} \\
&= \left\| \sum_{k=1}^{t-1} \left(\frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} X_s X_s^\top (\theta_k^* - \theta_{k+1}^*) \right) \right\|_{\Sigma_{t-1}} \\
&\leq \sum_{k=1}^{t-1} \left\| \frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} X_s X_s^\top (\theta_k^* - \theta_{k+1}^*) \right\|_{\Sigma_{t-1}} \\
&\leq \sum_{k=1}^{t-1} \left\| \frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} X_s \|X_s\|_2 \|\theta_k^* - \theta_{k+1}^*\|_2 \right\|_{\Sigma_{t-1}} \\
&\leq L \sum_{k=1}^{t-1} \left\| \frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} X_s \|\theta_k^* - \theta_{k+1}^*\|_2 \right\|_{\Sigma_{t-1}} \\
&\leq L \sum_{k=1}^{t-1} \frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} \|X_s\|_{\Sigma_{t-1}} \|\theta_k^* - \theta_{k+1}^*\|_2.
\end{aligned}$$

Applying Cauchy-Schwarz to the inner sum yields

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} \|X_s\|_{\Sigma_{t-1}} &= \frac{1}{\sigma^2} \sum_{s=1}^{k} \sqrt{w_{s,t-1}} \sqrt{w_{s,t-1}} \|X_s\|_{\Sigma_{t-1}} \\
&\leq \sqrt{\sum_{s=1}^{k} w_{s,t-1}} \sqrt{\frac{1}{\sigma^2} \sum_{s=1}^{k} w_{s,t-1} \|X_s\|_{\Sigma_{t-1}}^2} \\
&\leq \sqrt{d} \sqrt{\sum_{s=1}^{k} w_{s,t-1}},
\end{aligned}$$

where we bounded the second factor as follows:

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{s=1}^{k} w_{s,t-1}\|X_s\|^2_{\Sigma_{t-1}} &= \frac{1}{\sigma^2}\sum_{s=1}^{k} w_{s,t-1}\operatorname{Tr}\left(X_s^\top \Sigma_{t-1} X_s\right) \\
&= \operatorname{Tr}\left(\Sigma_{t-1}\left(\frac{1}{\sigma^2}\sum_{s=1}^{k} w_{s,t-1} X_s X_s^\top\right)\right) \\
&\le \operatorname{Tr}\left(\Sigma_{t-1}\left(\frac{1}{\sigma^2}\sum_{s=1}^{k} w_{s,t-1} X_s X_s^\top\right)\right) + \operatorname{Tr}\left(\Sigma_{t-1}\left(\frac{1}{\sigma^2}\sum_{s=k+1}^{t-1} w_{s,t-1} X_s X_s^\top\right)\right) + \operatorname{Tr}\left(\Sigma_{t-1}\Sigma_0^{-1}\right) \\
&= \operatorname{Tr}\left(\Sigma_{t-1}\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top\right)\right) \\
&= \operatorname{Tr}\left(\Sigma_{t-1}\Sigma_{t-1}^{-1}\right) = \operatorname{Tr}(\mathbb{I}_d) = d.
\end{aligned}$$

By putting everything together, we obtain:

$$\left\|\frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top(\theta_s^* - \theta_t^*)\right\|_{\Sigma_{t-1}} \le \sqrt{d}L\sum_{k=1}^{t-1}\sqrt{\sum_{s=1}^{k} w_{s,t-1}}\|\theta_k^* - \theta_{k+1}^*\|_2.$$

Combining all bounds proves that $|\langle x, \bar{\mu}_t - \theta_t^*\rangle| \le \alpha_{t-1}^{\mathrm{WSB}}\|x\|_{\Sigma_{t-1}}$, and taking $x = \Sigma_{t-1}^{-1}(\bar{\mu}_t - \theta_t^*)$ yields $\|\bar{\mu}_t - \theta_t^*\|_{\Sigma_{t-1}^{-1}} \le \alpha_{t-1}^{\mathrm{WSB}}$. □

Proof of Lemma 5. We aim to bound the deviation between the posterior mean μ_{t-1} in (2) and the surrogate posterior mean $\bar{\mu}_t$ in (5) with high probability:

$$\begin{aligned}
\mu_{t-1} - \bar{\mu}_t &= \Sigma_{t-1}\left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s r_s\right) - \Sigma_{t-1}\left(\Sigma_0^{-1}\theta_t^* + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top \theta_s^*\right) \\
&= \Sigma_{t-1}\left(\Sigma_0^{-1}(\mu_0 - \theta_t^*) + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right),
\end{aligned}$$

by the definition of the rewards r_t. For any arbitrary vector $x \in \mathbb{R}^d$, we can express the equality above as

$$\begin{aligned}
\langle x, \mu_{t-1} - \bar{\mu}_t\rangle &= \left\langle x, \Sigma_{t-1}\Sigma_0^{-1}(\mu_0 - \theta_t^*)\right\rangle + \left\langle x, \Sigma_{t-1}\left(\frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right)\right\rangle \\
&= \left\langle x, \Sigma_0^{-1}(\mu_0 - \theta_t^*)\right\rangle_{\Sigma_{t-1}} + \left\langle x, \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right\rangle_{\Sigma_{t-1}},
\end{aligned}$$

which, by the Cauchy-Schwarz inequality, yields

$$|\langle x, \mu_{t-1} - \bar{\mu}_t\rangle| \le \|x\|_{\Sigma_{t-1}}\left(\left\|\Sigma_0^{-1}(\mu_0 - \theta_t^*)\right\|_{\Sigma_{t-1}} + \left\|\frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right\|_{\Sigma_{t-1}}\right).$$

Now we can bound each term inside the parenthesis. For the first term, we simply use Assumption 1 to obtain our prior term:

$$\left\|\Sigma_0^{-1}(\mu_0 - \theta_t^*)\right\|_{\Sigma_{t-1}} \le \max_{\theta:\|\theta\|_2 \le S}\|\Sigma_0^{-1}(\mu_0 - \theta)\|_{\Sigma_{t-1}} \equiv \Pi_{t-1},$$

which we further bound in Lemma 3. Next, the second term is a martingale term, which can be bounded with high probability using a weighted variant of self-normalized concentration inequality for vector-valued martingales (Appendix B). For this analysis, we introduce the auxiliary covariance matrix $\widetilde{\Sigma}_t$, defined as $\widetilde{\Sigma}_t^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t} w_{s,t}^2 X_s X_s^\top$, which is also positive semi-definite. Since $\{w_{s,t} \in [0,1] : 1 \le s \le t\}$, and hence, $w_{s,t} \ge w_{s,t}^2$, we have $\Sigma_t^{-1} =$

$\Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t} w_{s,t} X_s X_s^\top \succeq \Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t} w_{s,t}^2 X_s X_s^\top = \widetilde{\Sigma}_t^{-1}$. Importantly, $\widetilde{\Sigma}_t$ is purely used for analysis and is not required in our algorithm. Because $A \succeq B \Rightarrow A^{-1} \preceq B^{-1}$ for positive definite matrices, we get $\Sigma_t \preceq \widetilde{\Sigma}_t$. Hence,

$$\left\|\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right\|_{\Sigma_{t-1}} \leq \left\|\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right\|_{\widetilde{\Sigma}_{t-1}}.$$

Thus, we can apply Corollary B.1 alongside a union bound over $t \in [T]$ with a localized failure probability $\delta_t = \delta/T$ for each time step. This guarantees that with probability at least $1-\delta$, the following inequality holds simultaneously for all $t \in [T]$:

$$\left\|\sum_{s=1}^{t-1} w_{s,t-1} X_s \varepsilon_s\right\|_{\widetilde{\Sigma}_{t-1}} \leq \sigma\sqrt{2\log\left(\frac{T}{\delta}\sqrt{\frac{\det(\widetilde{\Sigma}_{t-1}^{-1})}{\det(\Sigma_0^{-1})}}\right)} = \sigma\sqrt{2\log\left(\frac{T}{\delta}\sqrt{\frac{\det(\Sigma_0)}{\det(\widetilde{\Sigma}_{t-1})}}\right)}.$$

Applying Proposition A.1 with $p = 2$, gives

$$\frac{\det(\Sigma_0)}{\det(\widetilde{\Sigma}_{t-1})} \leq \left(1 + \frac{\mathrm{Tr}(\Sigma_0) L^2 \sum_{s=1}^{t-1} w_{s,t-1}^2}{d\sigma^2}\right)^d.$$

Summing everything together gives that with probability at least $1-\delta$, for all $t \in [T]$ simultaneously:

$$|\langle x, \mu_{t-1} - \bar{\mu}_t\rangle| \leq \|x\|_{\Sigma_{t-1}} \left(\Pi_{t-1} + \beta_{t-1}^{\mathrm{WSB}}(\delta/T)\right).$$

At last, taking $x = \Sigma_{t-1}^{-1}(\mu_{t-1} - \bar{\mu}_t)$ and applying the dual norm inequality yields that with probability at least $1-\delta$, for all $t \in [T]$ simultaneously:

$$\|\mu_{t-1} - \bar{\mu}_t\|_{\Sigma_{t-1}^{-1}} \leq \Pi_{t-1} + \beta_{t-1}^{\mathrm{WSB}}(\delta/T).$$

□

Proof of Lemma 2. The result follows directly by combining the decomposition

$$\mu_{t-1} - \theta_t^* = (\mu_{t-1} - \bar{\mu}_t) + (\bar{\mu}_t - \theta_t^*),$$

with the uniform bounds from Lemma 5 and Lemma 4, respectively. Specifically, for any $x \in \mathbb{R}^d$, the projection bound follows explicitly by applying the dual norm inequality $|\langle x, z\rangle| \leq \|x\|_{\Sigma_{t-1}} \|z\|_{\Sigma_{t-1}^{-1}}$ to the self-normalized vector error. Thus, with probability at least $1-\delta$, we have for all $t \in [T]$ simultaneously:

$$|\langle x, \mu_{t-1} - \theta_t^*\rangle| \leq |\langle x, \mu_{t-1} - \bar{\mu}_t\rangle| + |\langle x, \bar{\mu}_t - \theta_t^*\rangle| \leq \|x\|_{\Sigma_{t-1}} \left(\Pi_{t-1} + \beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \alpha_{t-1}^{\mathrm{WSB}}\right),$$

where Π_{t-1} is the prior term, $\beta_{t-1}^{\mathrm{WSB}}(\delta/T)$ is the noise concentration term, and $\alpha_{t-1}^{\mathrm{WSB}}$ is the drift bound. Finally, taking $x = \Sigma_{t-1}^{-1}(\mu_{t-1} - \theta_t^*)$ yields

$$\|\mu_{t-1} - \theta_t^*\|_{\Sigma_{t-1}^{-1}} \leq \Pi_{t-1} + \beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \alpha_{t-1}^{\mathrm{WSB}},$$

with probability at least $1-\delta$, concluding the proof. □

Proof of Lemma 3. The bound Π_t^{cxv} follows directly from applying Proposition A.2 under $p = 1$ and replacing variables with the corresponding prior metrics:

$$\Pi_t \leq \sqrt{\|\mu_0\|_{M_t}^2 + S^2\lambda_{\max}(M_t) + 2S\sqrt{\mu_0^\top M_t^2 \mu_0}} \equiv \Pi_t^{\mathrm{cxv}}.$$

The Π_t^{Δ}-bound comes from applying the triangle inequality:

$$\Pi_t \leq \|\Sigma_0^{-1}\mu_0\|_{\Sigma_t} + S\|\Sigma_0^{-1}\|_{\Sigma_t} \leq \|\Sigma_0^{-1}\mu_0\|_{\Sigma_t} + S\sqrt{\lambda_{\max}(\Sigma_0^{-1}\Sigma_t\Sigma_0^{-1})} = \|\mu_0\|_{M_t} + S\sqrt{\lambda_{\max}(M_t)} \equiv \Pi_t^{\Delta},$$

with $M_t = \Sigma_0^{-1}\Sigma_t\Sigma_0^{-1}$. Finally, to see that $\Pi_t^{\mathrm{cxv}} \leq \Pi_t^{\Delta}$, note that by the definition of the spectral norm, $\sqrt{\mu_0^\top M_t^2 \mu_0} \leq \sqrt{\lambda_{\max}(M_t)\mu_0^\top M_t \mu_0} = \sqrt{\lambda_{\max}(M_t)}\|\mu_0\|_{M_t}$. Substituting this into $(\Pi_t^{\mathrm{cxv}})^2$ reveals a perfect square expanding inequality:

$$(\Pi_t^{\mathrm{cxv}})^2 \leq \|\mu_0\|_{M_t}^2 + S^2\lambda_{\max}(M_t) + 2S\sqrt{\lambda_{\max}(M_t)}\|\mu_0\|_{M_t} = \left(\|\mu_0\|_{M_t} + S\sqrt{\lambda_{\max}(M_t)}\right)^2 = (\Pi_t^{\Delta})^2.$$

Taking the square root on both sides confirms $\Pi_t^{\mathrm{cxv}} \leq \Pi_t^{\Delta}$, completing the proof. □

D REGRET GUARANTEES OF ALGORITHMS

This appendix establishes regret guarantees for our WSB algorithms, treating each exploration paradigm separately. In the case of *deterministic exploration* (Appendix D.1), we prove the regret bound for `WSB-LinUCB` (Appendix D.1.1). For *randomized exploration* (Appendix D.2), we first develop the auxiliary events and instantaneous-regret bounds needed for perturbation-based policies, before deriving algorithm-specific guarantees for `WSB-RandLinUCB` (Appendix D.2.1) and `WSB-LinTS` (Appendix D.2.2).

We begin with a technical lemma that controls the cumulative variance terms, $\sum_{t=1}^{T}\|X_t\|^2_{\overline{\Sigma}_{t-1}}$. In WRLS-based approaches, this term is typically achieved using auxiliary results, such as the *weighted potential lemmas* presented in Faury et al. (2021a, Lemma 8) and/or Wang et al. (2023, Lemma 11). Here, we extend this type of control to the WSB posterior for any weighted scheme $\{w_{s,t}\}$ that are non-decreasing in s (i.e., $w_{s-1,t} \le w_{s,t}$) and satisfy the multiplicative consistency condition (i.e., $w_{s,t} \ge w_{t-1,t}w_{s,t-1}$) in the unit interval.

Lemma D.1 (Weighted potential lemma)**.** *For some $p \in \mathbb{N}_+$, define $\overline{\Sigma}_t^{-1} = \overline{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t} w_{s,t}^p X_s X_s^\top$, where $\overline{\Sigma}_0 \succ 0$. Under Assumption 1, we have*

$$\sum_{t=1}^{T}\|X_t\|^2_{\overline{\Sigma}_{t-1}} \le 2\sigma^2 \max\{1, L^2\lambda_{\max}(\overline{\Sigma}_0)/\sigma^2\}\left(d\sum_{t=1}^{T}\log\left(\frac{1}{w_{t-1,t}^p}\right) + \log\left(\frac{\det(\overline{\Sigma}_0)}{\det(\overline{\Sigma}_T)}\right)\right).$$

Proof of Lemma D.1. First, since $\{w_{s,t}\}$ is non-decreasing and satisfies the multiplicative consistency condition, we have $w_{s,t}^p \ge w_{t-1,t}^p\, w_{s,t-1}^p$ for all $s \le t-1$ (equality for exponential weights) and $w_{t,t}^p \ge w_{t-1,t}^p$. Hence,

$$\begin{aligned}
\overline{\Sigma}_t^{-1} &= \overline{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\left(\sum_{s=1}^{t-1} w_{s,t}^p X_s X_s^\top + w_{t,t}^p X_t X_t^\top\right)\\
&\succeq w_{t-1,t}^p\left(\overline{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1}^p X_s X_s^\top + \frac{1}{\sigma^2}X_t X_t^\top\right)\\
&= w_{t-1,t}^p\left(\overline{\Sigma}_{t-1}^{-1} + \frac{1}{\sigma^2}X_t X_t^\top\right).
\end{aligned}$$

Moreover, for any positive semi-definite matrices A, B, $A + B = A^{1/2}\big(I + A^{-1/2}BA^{-1/2}\big)A^{1/2}$. Applying this with $A = \overline{\Sigma}_{t-1}^{-1}$ and $B = \frac{1}{\sigma^2}X_t X_t^\top$ yields

$$\overline{\Sigma}_{t-1}^{-1} + \frac{1}{\sigma^2}X_t X_t^\top = \overline{\Sigma}_{t-1}^{-1/2}\left(\mathbb{I}_d + \frac{1}{\sigma^2}\overline{\Sigma}_{t-1}^{1/2}X_t X_t^\top \overline{\Sigma}_{t-1}^{1/2}\right)\overline{\Sigma}_{t-1}^{-1/2}.$$

Next, taking the determinant on both sides gives us

$$\begin{aligned}
\det\left(\overline{\Sigma}_t^{-1}\right) &\ge \det\left(w_{t-1,t}^p\overline{\Sigma}_{t-1}^{-1}\right)\det\left(\mathbb{I}_d + \frac{1}{\sigma^2}\left(\overline{\Sigma}_{t-1}^{1/2}X_t\right)\left(\overline{\Sigma}_{t-1}^{1/2}X_t\right)^\top\right)\\
&= (w_{t-1,t}^p)^d\det\left(\overline{\Sigma}_{t-1}^{-1}\right)\det\left(\mathbb{I}_d + \frac{1}{\sigma^2}\left(\overline{\Sigma}_{t-1}^{1/2}X_t\right)\left(\overline{\Sigma}_{t-1}^{1/2}X_t\right)^\top\right)\\
&= (w_{t-1,t}^p)^d\det\left(\overline{\Sigma}_{t-1}^{-1}\right)\left(1 + \frac{1}{\sigma^2}\|X_t\|^2_{\overline{\Sigma}_{t-1}}\right)\\
&\ge (w_{t-1,t}^p)^d\det\left(\overline{\Sigma}_{t-1}^{-1}\right)\left(1 + \frac{1}{\sigma^2\max\{1, L^2\lambda_{\max}(\overline{\Sigma}_0)/\sigma^2\}}\|X_t\|^2_{\overline{\Sigma}_{t-1}}\right)\\
&\ge (w_{t-1,t}^p)^d\det\left(\overline{\Sigma}_{t-1}^{-1}\right)\exp\left(\frac{1}{2\sigma^2\max\{1, L^2\lambda_{\max}(\overline{\Sigma}_0)/\sigma^2\}}\|X_t\|^2_{\overline{\Sigma}_{t-1}}\right),
\end{aligned}$$

using $\det(1 + xx^\top) = 1 + \|x\|_2^2$ for any $x \in \mathbb{R}^d$, $\max\{1, L^2\lambda_{\max}(\overline{\Sigma}_0)/\sigma^2\} \ge 1$, and $1 + z \ge \exp(z/2)$ for any $z \in [0,1]$. Note that the term $C = \max\{1, L^2\lambda_{\max}(\overline{\Sigma}_0)/\sigma^2\}$ is used in the denominator so that $\frac{1}{\sigma^2 C}\|X_t\|^2_{\overline{\Sigma}_{t-1}} \in [0,1]$. This is due to the fact that $\overline{\Sigma}_{t-1} \preceq \overline{\Sigma}_0$ and $\|X_t\|_2 \le L$ together imply $\|X_t\|^2_{\overline{\Sigma}_{t-1}} \le \lambda_{\max}(\overline{\Sigma}_{t-1})\|X_t\|_2^2 \le L^2\lambda_{\max}(\overline{\Sigma}_0)$. Finally, by exploiting the telescoping structure of the inequality above, we can derive the desired result. □

D.1 DETERMINISTIC EXPLORATION WITH WSB CONFIDENCE BOUNDS

D.1.1 Upper Confidence Bound (`WSB-LinUCB`)

Proof of Theorem 1. For any $x \in \mathbb{R}^d$, by positive definiteness, we have $\|x\|_{\Sigma_{t-1}} = \sqrt{x^\top \Sigma_{t-1} x} \leq \sqrt{\lambda_{\max}(\Sigma_{t-1})}\|x\|_2$. From the update $\Sigma_{t-1}^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{s=1}^{t-1} w_{s,t-1} X_s X_s^\top \succeq \Sigma_0^{-1}$, we have $\Sigma_{t-1} \preceq \Sigma_0$, hence $\lambda_{\max}(\Sigma_{t-1}) \leq \lambda_{\max}(\Sigma_0)$. Therefore, using Assumption 1 (which gives $\|x\|_2 \leq L$), $\|x\|_{\Sigma_{t-1}} \leq \sqrt{\lambda_{\max}(\Sigma_0)}\|x\|_2 \leq L\sqrt{\lambda_{\max}(\Sigma_0)}$. By Lemma 2, together with the fact that $\alpha_{t-1}^{\mathrm{WSB}}\|x\|_{\Sigma_{t-1}} \leq L\sqrt{\lambda_{\max}(\Sigma_0)}\alpha_{t-1}^{\mathrm{WSB}}$, we have for any $\delta \in (0,1)$, with probability at least $1-\delta$, that

$$\langle X_t^*, \theta_t^* \rangle \leq \langle X_t^*, \mu_{t-1} \rangle + L\sqrt{\lambda_{\max}(\Sigma_0)}\alpha_{t-1}^{\mathrm{WSB}} + (\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t^*\|_{\Sigma_{t-1}},$$

and with probability at least $1-\delta$ that

$$\langle X_t, \theta_t^* \rangle \geq \langle X_t, \mu_{t-1} \rangle - L\sqrt{\lambda_{\max}(\Sigma_0)}\alpha_{t-1}^{\mathrm{WSB}} - (\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t\|_{\Sigma_{t-1}}.$$

Hence, for any $\delta \in (0,1)$, we obtain, with probability at least $1-2\delta$, that

$$\begin{aligned}\langle X_t^* - X_t, \theta_t^* \rangle \leq & \langle X_t^* - X_t, \mu_{t-1} \rangle + 2L\sqrt{\lambda_{\max}(\Sigma_0)}\alpha_{t-1}^{\mathrm{WSB}} + (\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\left(\|X_t^*\|_{\Sigma_{t-1}} + \|X_t\|_{\Sigma_{t-1}}\right) \\ \leq & 2L\sqrt{\lambda_{\max}(\Sigma_0)}\alpha_{t-1}^{\mathrm{WSB}} + 2(\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t\|_{\Sigma_{t-1}},\end{aligned}$$

where the second inequality comes from the upper confidence bound selection criteria; $\langle X_t^*, \mu_{t-1} \rangle + (\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t^*\|_{\Sigma_{t-1}} \leq \langle X_t, \mu_{t-1} \rangle + (\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t\|_{\Sigma_{t-1}}$. Thus, the regret R_T can be upper bounded as follows:

$$R_T \leq 2L\sqrt{\lambda_{\max}(\Sigma_0)}\sum_{t=1}^{T}\alpha_{t-1}^{\mathrm{WSB}} + 2\sum_{t=1}^{T}(\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t\|_{\Sigma_{t-1}},$$

where the first term corresponds to the bias component, while the second term accounts for the variance component. The variance term can be directly bounded using the fact that $\beta_t^{\mathrm{WSB}}(\delta)$ is non-decreasing in t and Π_t is non-increasing in t (see its definition in Lemma 2), along with the Cauchy-Schwarz inequality, Lemma D.1 and Proposition A.1 with $p=1$:

$$\begin{aligned}&2\sum_{t=1}^{T}(\beta_{t-1}^{\mathrm{WSB}}(\delta/T) + \Pi_{t-1})\|X_t\|_{\Sigma_{t-1}} \leq 2(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0)\sum_{t=1}^{T}\|X_t\|_{\Sigma_{t-1}} \leq 2(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0)\sqrt{T}\sqrt{\sum_{t=1}^{T}\|X_t\|_{\Sigma_{t-1}}^2} \\ &\leq 2\sigma\sqrt{2\max\{1, L^2\lambda_{\max}(\Sigma_0)/\sigma^2\}}(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0)\sqrt{T}\sqrt{d\sum_{t=1}^{T}\log\left(\frac{1}{w_{t-1,t}}\right) + \log\left(\frac{\det(\Sigma_0)}{\det(\Sigma_T)}\right)} \\ &\leq 2\sigma\sqrt{2\max\{1, L^2\lambda_{\max}(\Sigma_0)/\sigma^2\}}(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0)\sqrt{dT}\sqrt{\sum_{t=1}^{T}\log\left(\frac{1}{w_{t-1,t}}\right) + \log\left(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2\sum_{t=1}^{T} w_{t,T}}{d\sigma^2}\right)},\end{aligned}$$

which yields the desired bound:

$$R_T \leq 2L\sqrt{\lambda_{\max}(\Sigma_0)}\sum_{t=1}^{T}\alpha_{t-1}^{\mathrm{WSB}} + 2^{3/2}\sigma\sqrt{\max\{1, L^2\lambda_{\max}(\Sigma_0)/\sigma^2\}}(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0)\sqrt{dT\Lambda_T},$$

with $\Lambda_T = \sum_{t=1}^{T}\log\left(\frac{1}{w_{t-1,t}}\right) + \log\left(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2\sum_{t=1}^{T} w_{t,T}}{d\sigma^2}\right)$. □

Proof of Corollary 1. By substituting $w_{s,t} = \gamma^{t-s}$ into the regret bound of R_T (Theorem 1), we obtain the following expression:

$$R_T \leq 4L^2\sqrt{d\lambda_{\max}(\Sigma_0)}\frac{1}{(1-\gamma)^{3/2}}B_T + 2\sigma\sqrt{2\max\{1, L^2\lambda_{\max}(\Sigma_0)/\sigma^2\}}(\beta_T^{\mathrm{WSB}}(\delta/T) + \Pi_0)\sqrt{dT\Lambda_T'},$$

with $\Lambda_T < \Lambda'_T = T \log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2}{d\sigma^2(1-\gamma)}\right)$, where we used that $\sum_{t=1}^{T} w_{t,T} = \sum_{t=1}^{T} \gamma^{T-t} = \frac{1-\gamma^T}{1-\gamma} < \frac{1}{1-\gamma}$ and that

$$
\begin{aligned}
\sum_{t=1}^{T}\sum_{k=1}^{t-1}\sqrt{\sum_{s=1}^{k} w_{s,t-1}}\|\theta_k^* - \theta_{k+1}^*\|_2 &= \sum_{t=1}^{T}\sum_{k=1}^{t-1}\sqrt{\sum_{s=1}^{k} \gamma^{t-s-1}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= \sum_{t=1}^{T}\sum_{k=1}^{t-1}\sqrt{\frac{\gamma^{t-1}(\gamma^{-k}-1)}{1-\gamma}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= \sum_{k=1}^{T-1}\sum_{t=k+1}^{T}\sqrt{\frac{\gamma^{t-1}(\gamma^{-k}-1)}{1-\gamma}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= \sum_{k=1}^{T-1}\sum_{t=k+1}^{T}\sqrt{\gamma^{t-1}}\frac{\sqrt{\gamma^{-k}-1}}{\sqrt{1-\gamma}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= \sum_{k=1}^{T-1}\frac{\sqrt{\gamma^k}-\sqrt{\gamma^T}}{1-\sqrt{\gamma}}\frac{\sqrt{\gamma^{-k}-1}}{\sqrt{1-\gamma}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&\le \sum_{k=1}^{T-1}\frac{\sqrt{\gamma^k}-\sqrt{\gamma^T}}{(1-\sqrt{\gamma})(\frac{1+\sqrt{\gamma}}{2})}\frac{\sqrt{\gamma^{-k}-1}}{\sqrt{1-\gamma}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= 2\sum_{k=1}^{T-1}\frac{\sqrt{\gamma^k}-\sqrt{\gamma^T}}{1-\gamma}\frac{\sqrt{\gamma^{-k}-1}}{\sqrt{1-\gamma}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= 2\sum_{k=1}^{T-1}\frac{(\sqrt{\gamma^k}-\sqrt{\gamma^T})\sqrt{\gamma^{-k}-1}}{(1-\gamma)^{3/2}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&\le 2\sum_{k=1}^{T-1}\frac{\sqrt{\gamma^k}\sqrt{\gamma^{-k}}}{(1-\gamma)^{3/2}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= 2\sum_{k=1}^{T-1}\frac{1}{(1-\gamma)^{3/2}}\|\theta_k^* - \theta_{k+1}^*\|_2 \\
&= \frac{2}{(1-\gamma)^{3/2}}B_T.
\end{aligned}
$$

Since the regret bound contains a term of the form $T\sqrt{\log(1/\gamma)}$, the discount factor γ cannot become smaller than $1/T$, i.e., $\gamma \ge 1/T$. Consequently, this leads to the inequality $\log(1/\gamma) \le C(1-\gamma)$ where $C = \log(T)/(1-1/T)$. By disregarding logarithmic dependencies on T, we can bound the regret as follows:

$$
R_T \le \tilde{\mathcal{O}}\left(\frac{\sqrt{d\lambda_{\max}(\Sigma_0)}B_T}{(1-\gamma)^{3/2}} + dT\sqrt{1-\gamma}\right).
$$

Setting $\lambda_{\max}(\Sigma_0) = 1/d$, this simplifies to

$$
R_T \le \tilde{\mathcal{O}}\left(\frac{B_T}{(1-\gamma)^{3/2}} + dT\sqrt{1-\gamma}\right).
$$

When B_T is small, specifically $B_T < d/T$, choosing $\gamma = 1 - 1/T$ yields the regret bound $R_T \le \tilde{\mathcal{O}}(d\sqrt{T})$. For cases where B_T is larger, i.e., $B_T \ge d/T$, setting $\gamma = 1 - \sqrt{B_T/dT}$ results in $R_T \le \tilde{\mathcal{O}}(B_T^{1/4}(dT)^{3/4})$, which completes the proof. □

D.2 RANDOMIZED EXPLORATION THROUGH WSB PERTURBATION

Our analysis builds on Kim and Tewari (2020), who introduced perturbation-based methods for non-stationary linear contextual bandits. But we adopt the refined weighted analysis of Wang et al. (2023), which allows us to extend their results to the WSB framework and obtain improved regret bounds.

Let $\tilde{f}_t(x)$ denote the algorithm-specific selection criterion. For instance:

- `WSB-RandLinUCB`: $\tilde{f}_t(x) = \langle x, \mu_{t-1} \rangle + \eta_t \|x\|_{\Sigma_{t-1}}$, with $\eta_t \sim \mathcal{N}(0, a^2)$ a scalar perturbation.
- `WSB-LinTS`: $\tilde{f}_t(x) = \langle x, \mu_{t-1} \rangle + x^\top \Sigma_{t-1}^{1/2} \eta_t$, with $\eta_t \sim \mathcal{N}(0, a^2 \mathbb{I}_d)$ a d-dimensional Gaussian vector. Equivalently, one may write $\tilde{f}_t(x) = \langle x, \mu_{t-1} \rangle + \eta_{t,x} \|x\|_{\Sigma_{t-1}}$ with $\eta_{t,x} \sim \mathcal{N}(0, a^2)$.

The central idea is that randomized algorithms can be analyzed by controlling the effect of (in this case) Gaussian perturbations. Concentration bounds ensure that perturbations remain bounded, while anti-concentration guarantees that the optimal action is not systematically discarded. Following Kim and Tewari (2020), we define auxiliary events that combine these perturbation properties with our WSB confidence bounds (Lemma 2). With these events in place, the refined analysis of Wang et al. (2023) applies directly, yielding improved regret guarantees for randomized exploration without the need for local norms.

We introduce the following events, which control both estimation error and random perturbations:

$$\begin{aligned}
\mathcal{E}^{\text{WSB}} &\equiv \{\forall t \geq 1, \forall x \in \mathcal{X}_t : |\langle x, \mu_t - \bar{\mu}_t \rangle| \leq c_1 \|x\|_{\Sigma_{t-1}}\}, \\
\mathcal{E}_t^{\text{Conc.}} &\equiv \{\forall x \in \mathcal{X}_t : |\tilde{f}_t(x) - \langle x, \mu_t \rangle| \leq c_2 \|x\|_{\Sigma_{t-1}}\}, \\
\mathcal{E}_t^{\text{Anti-Conc.}} &\equiv \{\tilde{f}_t(X_t^*) - \langle X_t^*, \mu_t \rangle > c_1 \|X_t^*\|_{\Sigma_{t-1}}\}.
\end{aligned}$$

The next lemma adapts Kim and Tewari (2020, Theorem 3) to the WSB framework.

Lemma D.2. *Let $p_1, p_2, p_3 \in (0,1)$, a time horizon $T \in \mathbb{N}$, and suppose there exists constants $c_1, c_2 \geq 1$ which may depend on T such that $\mathbb{P}(\mathcal{E}^{WSB}) \geq 1 - p_1$, $\mathbb{P}(\mathcal{E}_t^{Conc.}) \geq 1 - p_2$, and $\mathbb{P}(\mathcal{E}_t^{Anti\text{-}Conc.}) \geq p_3$. Then, for any prior $\pi(\theta) = \mathcal{N}(\theta | \mu_0, \Sigma_0)$, the following inequality holds for all posteriors $\rho_t(\theta) = \mathcal{N}(\theta | \mu_t, \Sigma_t)$ and $t \in [T]$ simultaneously:*

$$\langle X_t^* - X_t, \bar{\mu}_t \rangle \leq p_2 + (c_1 + c_2) \left(1 + \frac{2}{p_3 - p_2}\right) \|X_t\|_{\Sigma_{t-1}},$$

where $X_t = \arg\max_{x \in \mathcal{X}_t} \tilde{f}_t(x)$ with $\tilde{f}_t(x)$ denoting the algorithm-specific selection criterion.

Proof of Lemma D.2. The proof follows directly from analogous steps as in Kim and Tewari (2020, Theorem 3), with all arguments carrying over after replacing their WRLS confidence events and local norms with our WSB concentration events and the norm $\|\cdot\|_{\Sigma_{t-1}}$, and using our notation. □

Proof of Theorem 2. We begin with the standard decomposition of the regret:

$$R_T \leq 2L\sqrt{\lambda_{\max}(\Sigma_0)} \sum_{t=1}^{T} \alpha_{t-1}^{\text{WSB}} + \sum_{t=1}^{T} \langle X_t^* - X_t, \bar{\mu}_t \rangle.$$

Applying Lemma D.2 to the last term yields

$$\sum_{t=1}^{T} \langle X_t^* - X_t, \bar{\mu}_t \rangle = \sum_{t=1}^{T} \langle X_t^* - X_t, \bar{\mu}_t \rangle \, \mathbb{1}_{\{\mathcal{E}^{\text{WSB}}\}} + T\mathbb{P}(\overline{\mathcal{E}^{\text{WSB}}}) \leq (c_1 + c_2) \left(1 + \frac{2}{p_3 - p_2}\right) \sum_{t=1}^{T} \|X_t\|_{\Sigma_{t-1}} + T(p_1 + p_2).$$

By Cauchy–Schwarz inequality,

$$\begin{aligned}
\sum_{t=1}^{T} \|X_t\|_{\Sigma_{t-1}} &\leq \sqrt{T} \sqrt{\sum_{t=1}^{T} \|X_t\|_{\Sigma_{t-1}}^2} \leq \sqrt{T} \sqrt{d \sum_{t=1}^{T} \log\left(\frac{1}{w_{t-1,t}}\right) + \log\left(\frac{\det(\Sigma_0)}{\det(\Sigma_T)}\right)} \\
&\leq \sqrt{dT} \sqrt{\sum_{t=1}^{T} \log\left(\frac{1}{w_{t-1,t}}\right) + \log\left(1 + \frac{\operatorname{Tr}(\Sigma_0) L^2 \sum_{t=1}^{T} w_{t,T}}{d\sigma^2}\right)},
\end{aligned}$$

where the last two inequalities follows from Lemma D.1 and Proposition A.1 with $p = 1$. Substituting this bound completes the proof. □

The crucial step in the above proof is the control of the drift-related term $2L\sqrt{\lambda_{\max}(\Sigma_0)}\sum_{t=1}^{T}\alpha_{t-1}^{\mathrm{WSB}}$. This is precisely where the refined weighted analysis of Wang et al. (2023) enters, which avoids the virtual windowing arguments used in prior work and eliminates the need for *local norms*. It is this refinement in the treatment of drift that yields the improved regret guarantees in our WSB analysis.

D.2.1 Randomized Upper Confidence Bound (`WSB-RandLinUCB`)

The following lemma, which follows from Kim and Tewari (2020, Lemmas 4, 5, and 6), establishes high-probability control of the events $\mathcal{E}^{\mathrm{WSB}}$, $\mathcal{E}_t^{\text{Conc.}}$, and $\mathcal{E}_t^{\text{Anti-Conc.}}$ for `WSB-RandLinUCB` with suitable choices of c_1 and c_2.

Lemma D.3. *For `WSB-RandLinUCB`, the arm-selection criterion is $\tilde{f}_t(x) = \langle x, \mu_t\rangle + \eta_t\|x\|_{\Sigma_{t-1}}$, with $\eta_t \sim \mathcal{N}(0, a^2)$. Then, the following hold with appropriate constant choices:*

- *($\mathcal{E}^{\mathrm{WSB}}$). If $c_1 = \sqrt{4\log(T) + d\log(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2\sum_{t=1}^{T} w_{t,T}^2}{d\sigma^2})} + \Pi_0$, then $\mathbb{P}(\mathcal{E}^{\mathrm{WSB}}) \geq 1 - 1/T$.*
- *($\mathcal{E}_t^{\text{Conc.}}$). If $c_2 = a\sqrt{2\log(T/2)}$, then $\mathbb{P}(\overline{\mathcal{E}_t^{\text{Conc.}}}) \leq 1/T$.*
- *($\mathcal{E}_t^{\text{Anti-Conc.}}$). If $a^2 = 14c_1^2$, then $\mathbb{P}(\mathcal{E}_t^{\text{Anti-Conc.}}) \geq \exp(-1/4)/(8\sqrt{\pi})$.*

Proof of Lemma D.3. The three statements follow directly from Kim and Tewari (2020, Lemmas 4, 5, and 6), respectively. The arguments extend directly to our WSB setting after replacing the WRLS confidence and *local norms* with our WSB concentration and the norm $\|\cdot\|_{\Sigma_{t-1}}$. □

Proof of Corollary 2. The argument follows by analogous reasoning to the proof of Corollary 1. We have

$$R_T \leq 4L^2\sqrt{d\lambda_{\max}(\Sigma_0)}\frac{1}{(1-\gamma)^{3/2}}B_T + (c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right)\sqrt{dT\Lambda_T'} + T(p_1 + p_2),$$

where Λ_T' is given as in the proof of Corollary 1. With the choices of c_1 and c_2, this simplifies in $\tilde{\mathcal{O}}$-notation to

$$R_T \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{d\lambda_{\max}(\Sigma_0)}B_T}{(1-\gamma)^{3/2}} + dT\sqrt{1-\gamma}\right).$$

Setting $\lambda_{\max}(\Sigma_0) = 1/d$, we obtain

$$R_T \leq \tilde{\mathcal{O}}\left(\frac{B_T}{(1-\gamma)^{3/2}} + dT\sqrt{1-\gamma}\right).$$

If B_T is small, specifically $B_T < d/T$, choosing $\gamma = 1 - 1/T$ yields $R_T \leq \tilde{\mathcal{O}}(d\sqrt{T})$. On the other hand, if $B_T \geq d/T$, setting $\gamma = 1 - \sqrt{B_T/dT}$ gives $R_T \leq \tilde{\mathcal{O}}(B_T^{1/4}(dT)^{3/4})$. This completes the proof. □

D.2.2 Thompson Sampling (`WSB-LinTS`)

The following lemma, which follows from Kim and Tewari (2020, Lemmas 4, 5, and 6), establishes high-probability control of the events $\mathcal{E}^{\mathrm{WSB}}$, $\mathcal{E}_t^{\text{Conc.}}$, and $\mathcal{E}_t^{\text{Anti-Conc.}}$ for `WSB-LinTS` with suitable choices of c_1 and c_2.

Lemma D.4. *For `WSB-LinTS`, the arm-selection criterion is $\tilde{f}_t(x) = \langle x, \mu_t\rangle + x^\top\Sigma_{t-1}^{1/2}\eta_t$ with $\eta_t \sim \mathcal{N}(0, a^2\mathbb{I}_d)$. Equivalently, $\tilde{f}_t(x) = \langle x, \mu_t\rangle + \eta_{t,x}\|x\|_{\Sigma_{t-1}}$, with $\eta_{t,x} \sim \mathcal{N}(0, a^2)$. Then, the following hold with appropriate constant choices:*

- *($\mathcal{E}^{\mathrm{WSB}}$). If $c_1 = \sqrt{4\log(T) + d\log(1 + \frac{\mathrm{Tr}(\Sigma_0)L^2\sum_{t=1}^{T} w_{t,T}^2}{d\sigma^2})} + \Pi_0$, then $\mathbb{P}(\mathcal{E}^{\mathrm{WSB}}) \geq 1 - 1/T$.*
- *($\mathcal{E}_t^{\text{Conc.}}$). If $c_2 = a\sqrt{\log(KT/2)}$, then $\mathbb{P}(\overline{\mathcal{E}_t^{\text{Conc.}}}) \leq 1/T$.*
- *($\mathcal{E}_t^{\text{Anti-Conc.}}$). If $a^2 = 14c_1^2$, then $\mathbb{P}(\mathcal{E}_t^{\text{Anti-Conc.}}) \geq \exp(-1/4)/(8\sqrt{\pi})$.*

Proof of Lemma D.4. The three statements follow directly from Kim and Tewari (2020, Lemmas 4, 5, and 6), respectively. The arguments extend to our WSB setting after replacing the WRLS confidence and *local norms* with the WSB concentration and the norm $\|\cdot\|_{\Sigma_{t-1}}$. □

Proof of Corollary 3. The argument follows by analogous reasoning to the proof of Corollaries 1 and 2;

$$R_T \leq 4L^2\sqrt{d\lambda_{\max}(\Sigma_0)}\frac{1}{(1-\gamma)^{3/2}}B_T + (c_1+c_2)\left(1+\frac{2}{p_3-p_2}\right)\sqrt{dT\Lambda'_T} + T(p_1+p_2),$$

where Λ'_T is given as in the proof of Corollary 1. With the choices of c_1 and c_2, this simplifies in $\tilde{\mathcal{O}}$-notation to

$$R_T \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{d\lambda_{\max}(\Sigma_0)}B_T}{(1-\gamma)^{3/2}} + dT\sqrt{\log(K)}\sqrt{1-\gamma}\right).$$

Setting $\lambda_{\max}(\Sigma_0) = 1/d$, we obtain

$$R_T \leq \tilde{\mathcal{O}}\left(\frac{B_T}{(1-\gamma)^{3/2}} + dT\sqrt{\log(K)}\sqrt{1-\gamma}\right).$$

If B_T is small, specifically $B_T < d/T$, choosing $\gamma = 1 - 1/T$ yields $R_T \leq \tilde{\mathcal{O}}(d\sqrt{\log(K)T})$. On the other hand, if $B_T \geq d/T$, setting $\gamma = 1 - \sqrt{B_T/dT\sqrt{\log(K)}}$ gives $R_T \leq \tilde{\mathcal{O}}(B_T^{1/4}\log(K)^{3/8}(dT)^{3/4})$. □

E REGRET CURVES

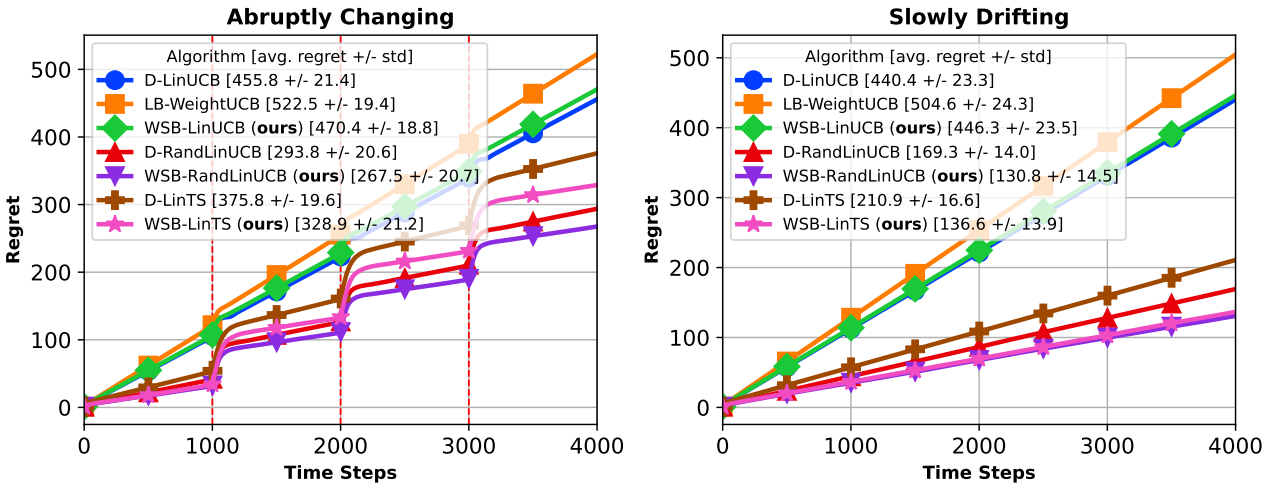

Figure 1: Cumulative regret comparison with parameter dimension $d = 2$ for the abruptly changing scenario (left column) and the slowly varying scenario (right column), averaged over 100 independent trials. Vertical red dashed lines denote the environment change-points in the abruptly changing scenario, which remain unknown to the algorithms.

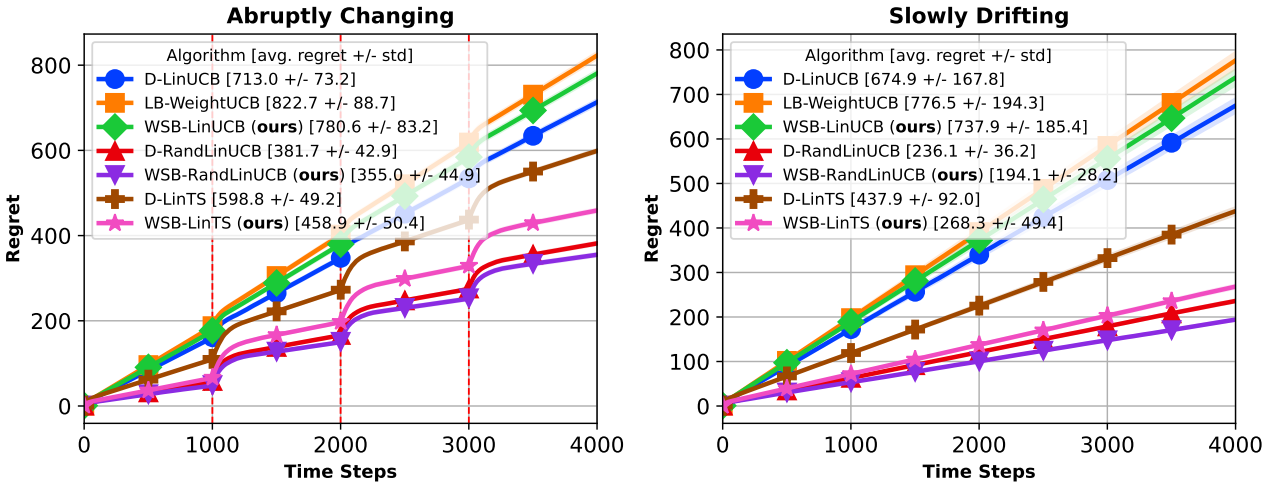

Figure 2: Cumulative regret comparison with parameter dimension $d = 4$ for the abruptly changing scenario (left column) and the slowly varying scenario (right column), averaged over 100 independent trials. Vertical red dashed lines denote the environment change-points in the abruptly changing scenario, which remain unknown to the algorithms.

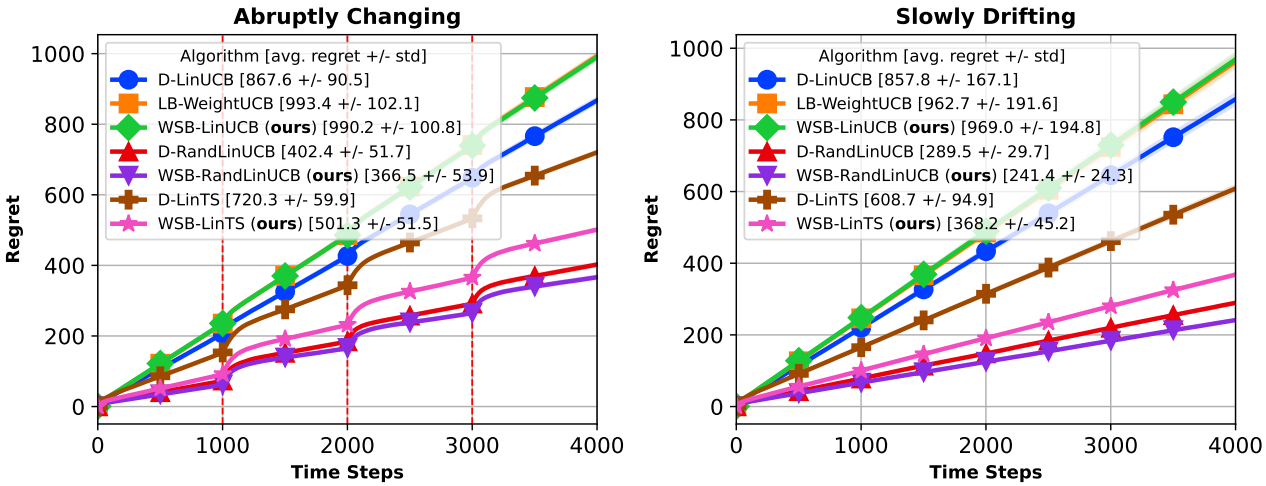

Figure 3: Cumulative regret comparison with parameter dimension $d = 6$ for the abruptly changing scenario (left column) and the slowly varying scenario (right column), averaged over 100 independent trials. Vertical red dashed lines denote the environment change-points in the abruptly changing scenario, which remain unknown to the algorithms.

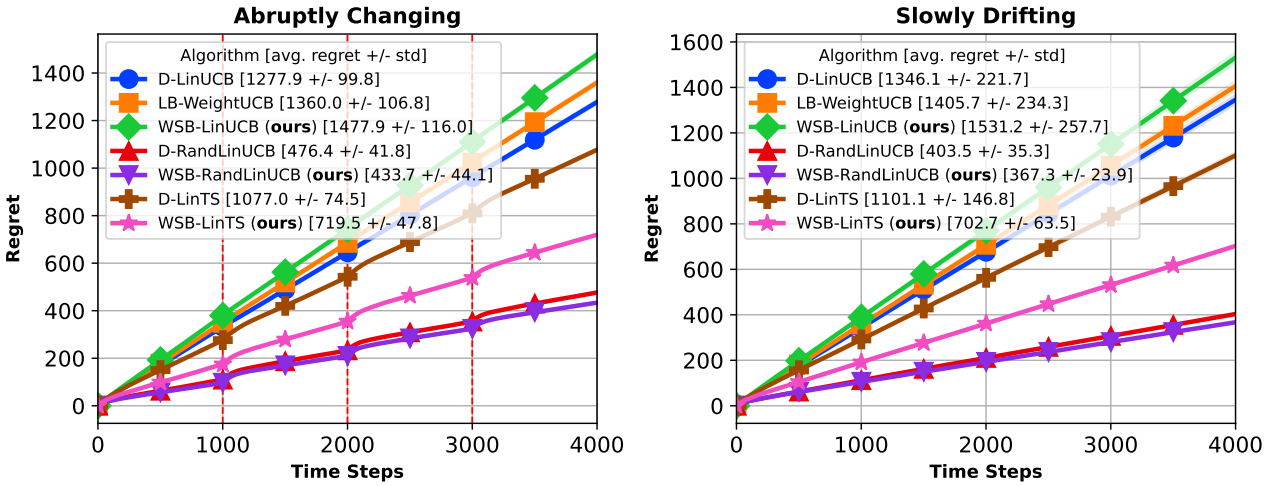

Figure 4: Cumulative regret comparison with parameter dimension $d = 16$ for the abruptly changing scenario (left column) and the slowly varying scenario (right column), averaged over 100 independent trials. Vertical red dashed lines denote the environment change-points in the abruptly changing scenario, which remain unknown to the algorithms.

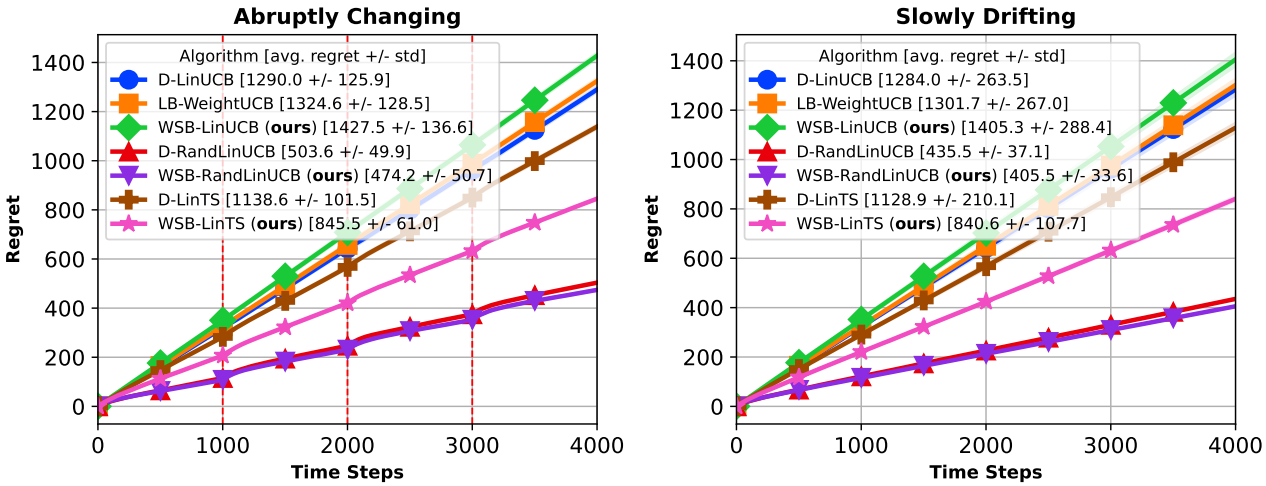

Figure 5: Cumulative regret comparison with parameter dimension $d = 32$ for the abruptly changing scenario (left column) and the slowly varying scenario (right column), averaged over 100 independent trials. Vertical red dashed lines denote the environment change-points in the abruptly changing scenario, which remain unknown to the algorithms.

F ABLATION STUDY

Table 3: Ablation study reporting cumulative regret (mean $\pm$ std) across context dimensions $d \in \{2, 4, 6, 16, 32\}$ in both abruptly changing and slowly drifting environments. The study evaluates sensitivity to prior misspecification, $\|\mu_0\|_2 \in \{1, 10, 100\}$, relative to the true parameter bound $S = 1$.

Abruptly Changing

ALGORITHM	$d = 2$	$d = 4$	$d = 6$	$d = 16$	$d = 32$
`WSB-LinUCB` ($\|\mu_0\|_2 = 1$)	474.5 ± 17.9	788.3 ± 84.0	996.6 ± 107.3	1480.6 ± 115.4	1429.9 ± 135.6
`WSB-LinUCB` ($\|\mu_0\|_2 = 10$)	528.5 ± 22.7	878.5 ± 91.9	1092.2 ± 113.7	1536.1 ± 121.8	1446.7 ± 137.4
`WSB-LinUCB` ($\|\mu_0\|_2 = 100$)	1642.6 ± 408.9	2228.5 ± 401.9	2258.4 ± 329.5	1928.6 ± 170.4	1516.2 ± 140.9
`WSB-RandLinUCB` ($\|\mu_0\|_2 = 1$)	273.7 ± 21.6	352.6 ± 44.2	368.5 ± 58.4	443.5 ± 45.7	486.5 ± 53.9
`WSB-RandLinUCB` ($\|\mu_0\|_2 = 10$)	627.2 ± 134.5	593.0 ± 120.0	562.7 ± 101.1	628.6 ± 102.9	606.2 ± 99.4
`WSB-RandLinUCB` ($\|\mu_0\|_2 = 100$)	2515.1 ± 781.0	2399.7 ± 592.5	2066.5 ± 432.2	1733.9 ± 195.3	1321.3 ± 170.4
`WSB-LinTS` ($\|\mu_0\|_2 = 1$)	332.2 ± 19.8	453.0 ± 47.5	505.2 ± 50.6	714.2 ± 44.1	851.5 ± 60.2
`WSB-LinTS` ($\|\mu_0\|_2 = 10$)	472.6 ± 68.4	552.6 ± 71.7	582.6 ± 58.6	792.1 ± 59.2	892.7 ± 80.9
`WSB-LinTS` ($\|\mu_0\|_2 = 100$)	2204.6 ± 664.1	2020.5 ± 481.9	1844.1 ± 378.5	1672.4 ± 181.3	1315.2 ± 157.6

Slowly Drifting

ALGORITHM	$d = 2$	$d = 4$	$d = 6$	$d = 16$	$d = 32$
`WSB-LinUCB` ($\|\mu_0\|_2 = 1$)	451.1 ± 23.2	743.6 ± 186.9	978.6 ± 196.6	1538.7 ± 258.3	1407.5 ± 288.4
`WSB-LinUCB` ($\|\mu_0\|_2 = 10$)	520.1 ± 40.8	862.9 ± 215.4	1091.4 ± 215.5	1596.8 ± 268.3	1423.8 ± 290.9
`WSB-LinUCB` ($\|\mu_0\|_2 = 100$)	1943.9 ± 991.6	2538.4 ± 910.5	2394.9 ± 690.9	1972.5 ± 338.5	1487.3 ± 300.2
`WSB-RandLinUCB` ($\|\mu_0\|_2 = 1$)	138.7 ± 16.8	199.5 ± 32.1	247.8 ± 28.3	366.3 ± 27.1	407.4 ± 34.2
`WSB-RandLinUCB` ($\|\mu_0\|_2 = 10$)	629.3 ± 308.5	539.9 ± 205.9	484.5 ± 159.0	569.3 ± 99.7	564.2 ± 183.8
`WSB-RandLinUCB` ($\|\mu_0\|_2 = 100$)	2809.5 ± 1742.7	2671.7 ± 1244.4	2167.5 ± 889.9	1831.0 ± 357.5	1352.1 ± 318.8
`WSB-LinTS` ($\|\mu_0\|_2 = 1$)	137.9 ± 13.8	268.6 ± 48.7	371.3 ± 46.0	706.2 ± 63.0	838.8 ± 108.1
`WSB-LinTS` ($\|\mu_0\|_2 = 10$)	354.1 ± 147.4	413.6 ± 100.0	471.2 ± 73.3	791.1 ± 81.4	872.5 ± 131.9
`WSB-LinTS` ($\|\mu_0\|_2 = 100$)	2371.1 ± 1414.0	2257.5 ± 996.0	1926.6 ± 764.9	1765.3 ± 331.5	1330.4 ± 314.6

We evaluate empirical robustness by testing the sensitivity of the WSB-based algorithms to prior misspecification. Specifically, we vary the norm of the prior mean as $\|\mu_0\|_2 \in \{1, 10, 100\}$, while the true parameter sequence satisfies the bound $S = 1$. The results show that WSB is reasonably stable when the prior scale is comparable to, or moderately larger than, the true parameter scale. Moving from $\|\mu_0\|_2 = 1$ to $\|\mu_0\|_2 = 10$ increases regret, but the algorithms generally retain the same qualitative behavior across both abruptly changing and slowly drifting environments. In contrast, the extreme setting $\|\mu_0\|_2 = 100$ leads to a substantial degradation in performance, especially for the randomized exploration methods. This confirms that the prior influence captured by the dynamic term Π_t is practically relevant: although sequential posterior updates can reduce the effect of the prior over time, a prior mean far outside the true parameter scale can dominate early decisions and delay adaptation. Overall, the ablation highlights both the robustness and the limitations of WSB under prior misspecification.