Variance-Dependent Regret Bounds for Non-stationary Linear Bandits

Zhiyong Wang* **Jize Xie†** **Yi Chen‡** **John C.S. Lui§** **Dongruo Zhou¶**

Abstract

We investigate the non-stationary stochastic linear bandit problem where the reward distribution evolves each round. Existing algorithms characterize the non-stationarity by the *total variation budget* B_K, which is the summation of the change of the consecutive feature vectors of the linear bandits over K rounds. However, such a quantity only measures the non-stationarity with respect to the expectation of the reward distribution, which makes existing algorithms sub-optimal under the general non-stationary distribution setting. In this work, we propose algorithms that utilize the *variance* of the reward distribution as well as the B_K, and show that they can achieve tighter regret upper bounds. Specifically, we introduce two novel algorithms: Restarted WeightedOFUL$^+$ and Restarted SAVE$^+$. These algorithms address cases where the variance information of the rewards is known and unknown, respectively. Notably, when the total variance V_K is much smaller than K, our algorithms outperform previous state-of-the-art results on non-stationary stochastic linear bandits under different settings. Experimental evaluations further validate the superior performance of our proposed algorithms over existing works.

1 Introduction

In this work, we study non-stationary stochastic bandits, which is a generalization of the classical stationary stochastic bandits, where the reward distribution is non-stationary. The intuition about the non-stationary setting comes from real-world applications such as dynamic pricing and ads allocation, where the environment changes rapidly and deviates significantly from stationarity [4, 10]. Most of the existing works in stochastic bandits consider a stationary setting where the goal of the agent is to minimize the *static regret*, *i.e.*, the summation of suboptimality gaps between the agent's selected arm and the fixed, time-independent best arm that maximizes the expectation of the reward distribution [3, 2, 26, 35, 36, 37, 47]. In contrast, for the non-stationary setting, the emphasis shifts to minimizing the *dynamic regret*, which represents the gap between the cumulative reward of selecting the time-dependent optimal arm at each time and that of the learner [10, 11, 42, 43]. As we can always treat a stationary bandit instance as a special case of the non-stationary bandit instance, designing algorithms that work well under the non-stationary setting is significantly more challenging.

There have been a series of works aiming to minimize the *dynamic regret* for non-stationary stochastic bandits, such as Multi-Armed Bandits (MAB) [4, 20, 7, 38], linear bandits [10, 11, 43, 39, 34], general function approximation [17, 29, 30], and the even more challenging reinforcement learning (RL) setting [28, 33, 19, 12, 39]. In this work, we mainly consider the linear bandit setting, where each arm is a contextual vector, and the expected reward of each arm is assumed to be the linear product

*The Chinese University of Hong Kong; zhiyongwangwzy@gmail.com
†Hong Kong University of Science and Technology; jxiebj@connect.ust.hk
‡Hong Kong University of Science and Technology; yichen@ust.hk
§The Chinese University of Hong Kong; cslui@cse.cuhk.edu.hk
¶Indiana University Bloomington; dz13@iu.edu

ICML 2024 Workshop on Foundations of Reinforcement Learning and Control.

of the arm with an unknown feature vector. Most existing *dynamic regret* results for non-stationary linear bandits depend on both the *non-stationarity measurement* and the number of interaction rounds. Specifically, assume K is the total number of rounds in bandits, and for each $k \in [K]$, $\mathbf{x}$ is one of the arms, $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k+1}$ are the feature vectors at k and $k+1$ rounds, satisfying $\|\mathbf{x}\|_2 \leq 1$. Then, the non-stationarity measurement is often defined as the summation of the changes in the mean of the reward distribution, which is

$$B_K := \sum_{k=1}^{K} \max_{\mathbf{x} \in \mathbb{R}^d} |\langle \mathbf{x}, \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1} \rangle| = \sum_{k=1}^{K} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_2 \,. \tag{1.1}$$

Existing works for non-stationary linear bandits [31, 22, 44, 33, 10, 43] achieved a regret upper bound of $\widetilde{O}(d^{7/8} B_K^{\frac{1}{4}} K^{\frac{3}{4}})$, where d is the problem dimension. A recent work [39] proposed a black-box reduction method that can achieve a regret upper bound of $\widetilde{O}(d B_K^{\frac{1}{3}} K^{\frac{2}{3}})$, yet in a slightly limited setting with *a fixed arm set* across all rounds. Such regret bounds clearly demonstrate that the regret grows as long as the non-stationarity grows, which is aligned with intuition.

Although existing works clearly demonstrate the relationship between the B_K and the regret, we claim that it is not sufficient for us to fully characterize the non-stationary level of the reward distributions. Consider applications such as hyperparameter tuning in physical systems, the noise distribution may highly depends on the evaluation point since the measurement noise often largely varies with the chosen parameter settings [25]. For linear bandits, such examples suggest that the non-stationarity not only consists of the change of the mean of the distribution, but also the variance of the distribution. However, none of the previous works on non-stationary linear bandits considered how to leverage the variance information to improve regret bounds in the above heteroscedastic noise setting. Therefore, an open question arises:

Can we design even better algorithms for non-stationary linear bandits by considering its variance information?

In this paper, we answer this question affirmatively. We assume that at the k-th round, the reward distribution of an arm $\mathbf{x}$ satisfies $r_k \sim \langle \boldsymbol{\theta}_k, \mathbf{x} \rangle + \epsilon_k$, where ϵ_k is a zero-mean noise variable with variance σ_k^2. Our contributions are:

- For the case where the reward variance σ_k^2 at round k can be observed and the *total variation budget* B_K is known, we propose the Restarted-WeightedOFUL$^+$ algorithm, which uses variance-based weighted linear regression to deal with heteroscedastic noises [46, 45] and a restarted scheme to forget some historical data to hedge against the non-stationarity. We prove that the regret upper bound of Restarted-WeightedOFUL$^+$ is $\widetilde{O}(d^{7/8}(B_K V_K)^{1/4}\sqrt{K} + d^{5/6} B_K^{1/3} K^{2/3})$. Notably, our regret surpasses the best result for non-stationary linear bandits $\widetilde{O}(d B_K^{1/3} K^{2/3})$ [39] when the total variance $V_K = \widetilde{O}(1)$ is small, which indicates that additional variance information benefits non-stationary linear bandit algorithms. It is worth noting that our algorithms could also work in the more general setting with the arm sets vary through rounds, unlike [39], which only addresses a fixed arm set.
- For the case where the reward variance σ_k^2 is unknown but the total variance V_K and variation budget B_K are known, we propose the Restarted-SAVE$^+$ algorithm. It maintains a multi-layer weighted linear regression structure with carefully-designed weight within each layer to handle the unknown variances. We prove that Restarted-SAVE$^+$ can achieve a regret upper bound of $\widetilde{O}(d^{\frac{4}{5}} V_K^{\frac{2}{5}} B_K^{\frac{1}{5}} K^{\frac{2}{5}} + d^{\frac{2}{3}} B_K^{\frac{1}{3}} K^{\frac{2}{3}})$. Specifically, when $V_K = \widetilde{O}(1)$, our regret is also better than the existing best result $\widetilde{O}(d B_K^{1/3} K^{2/3})$ [39], which again verifies the effect of the variance information.
- Lastly, we propose Restarted-SAVE$^+$-BOB for the case where both the reward variance σ_k^2 and B_K are unknown. Restarted-SAVE$^+$-BOB equips a *bandit-over-bandit* (BOB) framework to handle the unknown B_K, and also maintains a multi-layer structure as Restarted-SAVE$^+$. We show that Restarted-SAVE$^+$-BOB achieves a regret upper bound of $\widetilde{O}(d^{\frac{4}{5}} V_K^{\frac{2}{5}} B_K^{\frac{1}{5}} K^{\frac{2}{5}} + d^{\frac{2}{3}} B_K^{\frac{1}{3}} K^{\frac{2}{3}} + d^{\frac{1}{5}} K^{\frac{7}{10}})$, and it behaves the same as Restarted-SAVE$^+$ when $V_K = \widetilde{O}(1)$ and $B_K = \Omega(d^{-14} K^{1/10})$.
- We also conduct experimental evaluations to validate the outperformance of our proposed algorithms over existing works.

Model	Algorithm	Regret	Variance -Dependent	Varying Arm Set	Require B_K
Linear Bandit	SW-UCB [10]	$\widetilde{O}(d^{\frac{7}{8}}B_K^{\frac{1}{4}}K^{\frac{3}{4}})$	No	Yes	Yes
	BOB [10]	$\widetilde{O}\Big(d^{\frac{7}{8}}B_K^{\frac{1}{4}}K^{\frac{3}{4}}\Big)$	No	Yes	No
	RestartUCB [43]	$\widetilde{O}\Big(d^{\frac{7}{8}}B_K^{\frac{1}{4}}K^{\frac{3}{4}}\Big)$	No	Yes	Yes
	RestartUCB-BOB [43]	$\widetilde{O}\Big(d^{\frac{7}{8}}B_K^{\frac{1}{4}}K^{\frac{3}{4}}\Big)$	No	Yes	No
	LB-WeightUCB [34]	$\widetilde{O}\Big(d^{\frac{3}{4}}B_K^{\frac{1}{4}}K^{\frac{3}{4}}\Big)$	No	Yes	Yes
	MASTER + OFUL [39]	$\widetilde{O}\Big(dB_K^{\frac{1}{3}}K^{\frac{2}{3}}\Big)$	No	No	No
	Restarted-WeightedOFUL$^+$ **(Ours)**	$\widetilde{O}\Big(d^{\frac{7}{8}}(B_KV_K)^{\frac{1}{4}}K^{\frac{1}{2}} + d^{\frac{5}{6}}B_K^{\frac{1}{3}}K^{\frac{2}{3}}\Big)$	Yes	Yes	Yes
	Restarted SAVE$^+$ **(Ours)**	$\widetilde{O}\Big(d^{\frac{4}{5}}V_K^{\frac{2}{5}}B_K^{\frac{1}{5}}K^{\frac{2}{5}} + d^{\frac{2}{3}}B_K^{\frac{1}{3}}K^{\frac{2}{3}}\Big)$	Yes	Yes	Yes
	Restarted SAVE$^+$-BOB **(Ours)**	$\widetilde{O}\Big(d^{\frac{4}{5}}V_K^{\frac{2}{5}}B_K^{\frac{1}{5}}K^{\frac{2}{5}} + d^{\frac{2}{3}}B_K^{\frac{1}{3}}K^{\frac{2}{3}} + d^{\frac{1}{5}}K^{\frac{7}{10}}\Big)$	Yes	Yes	No
MAB	Rerun-UCB-V [38]	$\widetilde{O}\Big(\lvert\mathcal{A}\rvert^{\frac{2}{3}}B_K^{\frac{1}{3}}V_K^{\frac{1}{3}}K^{\frac{1}{3}} + \lvert\mathcal{A}\rvert^{\frac{1}{2}}B_K^{\frac{1}{2}}K^{\frac{1}{2}}\Big)$	Yes	No	Yes
	Lower Bound [38]	$\widetilde{\Omega}\Big(B_K^{\frac{1}{3}}V_K^{\frac{1}{3}}K^{\frac{1}{3}} + B_K^{\frac{1}{2}}K^{\frac{1}{2}}\Big)$	Yes	No	-

Table 1: Comparison of non-stationary bandits in terms of regret guarantee. K is the total rounds, d is the problem dimension for linear bandits, B_K is the *total variation budget* defined in Section 3 (for the MAB setting, $B_K = \sum_{k=1}^{K} \|\mu_k - \mu_{k+1}\|_\infty$, where μ_k is the mean of the reward distribution at round k), V_K is the *total variance* defined in Section 3, $|\mathcal{A}|$ is the number of arms for MAB.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We denote by $[n]$ the set $\{1, \ldots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and a positive semi-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we denote by $\|\mathbf{x}\|_2$ the vector's Euclidean norm and define $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n = 1, 2, \ldots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq Cb_n$ holds for all $n \geq 1$ and write $a_n = \Omega(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \geq Cb_n$ holds for all $n \geq 1$. We use $\widetilde{O}(\cdot)$ to further hide the polylogarithmic factors.

2 Related Work

Non-stationary (Linear) Bandits There have been a series of works about non-stationary bandits [4, 20, 7, 38, 11, 31, 5, 9, 29, 43, 23, 39, 30, 8, 16, 32, 27, 1, 14]. In non-stationary linear bandits, the unknown feature vector $\boldsymbol{\theta}_k$ can be dynamically and adversarially adjusted, with the total change upper bounded by the *total variation budget* B_K over K rounds, *i.e.*, $\sum_{k=1}^{K-1} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2 \leq B_K$. To address this problem, some works proposed forgetting strategies such as sliding window, restart, and weighted regression [11, 31, 43]. Kim and Tewari [23] also introduced the randomized exploration with weighting strategy. The regret upper bounds in these works are all of $\widetilde{O}(B_K^{\frac{1}{4}}K^{\frac{3}{4}})$. A recent work by [39] proposed the MASTER-OFUL algorithm based on a black-box approach, which can achieve a regret bound of $\widetilde{O}(B_K^{\frac{1}{3}}K^{\frac{2}{3}})$ in the case where *the arm set is fixed over K rounds*. To the best of our knowledge, none of the existing works consider how to utilize the variance information to

improve the regret bound in the case with time-dependent variances. The only exception of utilizing the variance information in the non-stationary bandit setting is Wei et al. [38], which proposed the Rerun-UCB-V algorithm for the non-stationary MAB setting with a regret dependent on the action set size $|\mathcal{A}|$. To compare with, the regret upper bounds of our algorithms are independent of the action set size, thus our algorithms are more efficient for the case where the number of actions is large.

Linear Bandits with Heteroscedastic Noises Some recent works study the heteroscedastic linear bandit problem, where the noise distribution is assumed to vary over time. Kirschner and Krause [25] first proposed the linear bandit model with heteroscedastic noise. In this model, the noise at round $k \in [K]$ is assumed to be σ_k-sub-Gaussian. Some follow-up works relaxed the σ_k-sub-Gaussian assumption by assuming the noise at the k-th round to be of variance σ_k^2 [46, 40, 24, 45, 15, 41]. Specifically, Zhou et al. [46] and Zhou and Gu [45] considered the case where σ_k is observed by the learner after the k-th round. [40] and [24] proposed statistically efficient but computationally inefficient algorithms for the unknown-variance case. A recent work by [41] proposed an algorithm that achieves both statistical and computational efficiency in the unknown-variance setting. Dai et al. [15] also considered a specific heteroscedastic linear bandit problem where the linear model is sparse.

3 Problem Setting

We consider a heteroscedastic variant of the classic non-stationary linear contextual bandit problem. Let K be the total number of rounds. At each round $k \in [K]$, the learner interacts with the environment as follows: (1) the environment generates an arbitrary arm set $\mathcal{D}_k \subseteq \mathbb{R}^d$ where each element represents a feasible arm for the learner to choose, and also generates an *unknown* feature vector $\boldsymbol{\theta}_k$; (2) the leaner observes $\mathcal{D}_k$ and selects $\mathbf{a}_k \in \mathcal{D}_k$; (3) the environment generates the stochastic noise ϵ_k and reveals the stochastic reward $r_k = \langle \boldsymbol{\theta}_k, \mathbf{a}_k \rangle + \epsilon_k$ to the leaner. We assume the ℓ_2 norm of the feasible actions is upper bounded by A, *i.e.*, for all $k \in [K]$, $\mathbf{a} \in \mathcal{D}_k$: $\|\mathbf{a}\|_2 \leq A$.

Following Zhou et al. [46], Zhao et al. [41], we assume the following condition on the random noise ϵ_k at each round k:

$$\mathbb{P}\left(|\epsilon_k| \leq R\right) = 1, \quad \mathbb{E}[\epsilon_k | \mathbf{a}_{1:k}, \epsilon_{1:k-1}] = 0, \quad \mathbb{E}[\epsilon_k^2 | \mathbf{a}_{1:k}, \epsilon_{1:k-1}] = \sigma_k^2. \tag{3.1}$$

For the case with known variance, we assume that at each round k, the variance σ_k is also revealed to the learner together with the reward r_k; in the unknown variance case, the σ_k can not be observed.

Following [10, 11, 31, 43], we assume the sum of ℓ_2 differences of consecutive θ_k's is upper bounded by the *total variation budget* B_K, *i.e.*, $\sum_{k=1}^{K-1} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2 \leq B_K$, where the $\boldsymbol{\theta}_k$'s can be adversarially chosen by an oblivious adversary. We also assume that the *total variance* is upper bounded by V_K, which is $\sum_{k=1}^{K} \sigma_k^2 \leq V_K$. The goal of the agent is to minimize the *dynamic regret* defined as follows: $\text{Regret}(K) = \sum_{k \in [K]} \left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k \rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k \rangle \right)$, where $\mathbf{a}_k^* = \text{argmax}_{\mathbf{a} \in \mathcal{D}_k} \langle \mathbf{a}, \boldsymbol{\theta}_k \rangle$ is the optimal arm at round k which gives the highest expected reward.

4 Non-stationary Linear Contextual Bandit with Known Variance

In this section, we introduce our Algorithm 1 under the setting where the variance σ_k^2 at k-th iteration is known to the agent in prior. We start from WeightedOFUL$^+$ [45], an *weighted ridge regression*-based algorithm for heteroscedastic linear bandits under the stationary reward assumption. For our non-stationary linear bandit setting where $\boldsymbol{\theta}_k$ is changing over the round k, WeightedOFUL$^+$ aims to build an $\widehat{\boldsymbol{\theta}}_k$ which estimates the feature vector $\boldsymbol{\theta}_k$ by using the solution to the following regression:

$$\widehat{\boldsymbol{\theta}}_k \leftarrow \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{k-1} \bar{\sigma}_t^{-2} (\langle \boldsymbol{\theta}, \mathbf{a}_t \rangle - r_t)^2 + \lambda \|\boldsymbol{\theta}\|_2^2, \tag{4.2}$$

where the weight is defined as in (4.1). After obtaining $\widehat{\boldsymbol{\theta}}_k$, WeightedOFUL$^+$ chooses arm $\mathbf{a}_k$ by maximizing the upper confidence bound (UCB) of $\langle \mathbf{a}, \widehat{\boldsymbol{\theta}} \rangle$, with an exploration bonus $\widehat{\beta}_k \|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}$, where $\widehat{\boldsymbol{\Sigma}}_k$ is the covariance matrix over $\mathbf{a}_k$. The weight $\bar{\sigma}_k^2$ is introduced to balance the different past examples based on their reward variance σ_k^2, and such a strategy has been proved as a state-of-the-art algorithm for the stationary heteroscedastic linear bandits [45]. However, the non-stationary nature of our setting prevents us from directly using $\widehat{\boldsymbol{\theta}}_k$ defined in (4.2) as an estimate to $\boldsymbol{\theta}$. Therefore, inspired by the *restarting* strategy which has been adopted by previous algorithms for non-stationary

Algorithm 1 Restarted-WeightedOFUL$^+$

Require: Regularization parameter $\lambda > 0$; B, an upper bound on the ℓ_2-norm of $\boldsymbol{\theta}_k$ for all $k \in [K]$; confidence radius $\widehat{\beta}_k$, variance parameters α, γ; restart window size w.

$\widehat{\boldsymbol{\Sigma}}_1 \leftarrow \lambda \mathbf{I}$, $\widehat{\mathbf{b}}_1 \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_1 \leftarrow \mathbf{0}$, $\widehat{\beta}_1 = \sqrt{\lambda} B$

for $k = 1, \ldots, K$ **do**

  if $k \% w == 0$ **then**

    $\widehat{\boldsymbol{\Sigma}}_k \leftarrow \lambda \mathbf{I}$, $\widehat{\mathbf{b}}_k \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_k \leftarrow \mathbf{0}$, $\widehat{\beta}_k = \sqrt{\lambda} B$

  end if

  Observe $\mathcal{D}_k$ and choose $\mathbf{a}_k \leftarrow \operatorname{argmax}_{\mathbf{a} \in \mathcal{D}_k} \langle \mathbf{a}, \boldsymbol{\theta}_k \rangle + \widehat{\beta}_k \|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}$

  Observe (r_k, σ_k), set $\bar{\sigma}_k$ as

$$\bar{\sigma}_k \leftarrow \max\{\sigma_k, \alpha, \gamma \|\mathbf{a}_k\|^{1/2}_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}\} \tag{4.1}$$

  $\widehat{\boldsymbol{\Sigma}}_{k+1} \leftarrow \widehat{\boldsymbol{\Sigma}}_k + \mathbf{a}_k \mathbf{a}_k^\top / \bar{\sigma}_k^2$, $\widehat{\mathbf{b}}_{k+1} \leftarrow \widehat{\mathbf{b}}_k + r_k \mathbf{a}_k / \bar{\sigma}_k^2$, $\widehat{\boldsymbol{\theta}}_{k+1} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k+1}^{-1} \widehat{\mathbf{b}}_{k+1}$

end for

linear bandits [43], we propose Restarted-WeightedOFUL$^+$, which periodically restarts itself and runs WeightedOFUL$^+$ as its submodule. The restart window size is set as w, which is used to balance the nonstationarity and the total regret and will be fine-tuned in the next steps. Combined with the restart window size w, we set $\{\widehat{\beta}_k\}_{k \geq 1}$ to

$$\widehat{\beta}_k = 12\sqrt{d \log(1 + \frac{(k\%w)A^2}{\alpha^2 d\lambda}) \log(32(\log(\frac{\gamma^2}{\alpha} + 1)\frac{(k\%w)^2}{\delta})} + 30 \log(32(\log(\frac{\gamma^2}{\alpha}) + 1)\frac{(k\%w)^2}{\delta})\frac{R}{\gamma^2} + \sqrt{\lambda} B. \tag{4.3}$$

We now propose the theoretical guarantee for Algorithm 8. The following key lemma shows how nonstationarity affects our estimation of the reward of each arm.

Lemma 4.1. Let $0 < \delta < 1$. Then with probability at least $1 - \delta$, for any action $\mathbf{a} \in \mathbb{R}^d$, we have

$$|\mathbf{a}^\top (\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)| \leq \underbrace{\frac{A^2}{\alpha}\sqrt{\frac{dw}{\lambda}} \sum_{t=w \cdot \lfloor k/w \rfloor + 1}^{k-1} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2}_{\text{Drifting term}} + \underbrace{\widehat{\beta}_k \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}}_{\text{Stochastic term}}.$$

Here we provide a proof sketch of Lemma 4.1 to show the technical challenge we need to overcome. Without loss of generality, we prove the lemma for $k \in [1, w]$. We have

$$|\mathbf{a}^\top (\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)| \leq \left| \mathbf{a}^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \sum_{t=1}^{k-1} \frac{\mathbf{a}_t \mathbf{a}_t^\top}{\bar{\sigma}_t^2} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k) \right| + \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}} \| \sum_{t=1}^{k-1} \frac{\mathbf{a}_t \epsilon_t}{\bar{\sigma}_t^2} \|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}} + \sqrt{\lambda} B \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}, \tag{4.4}$$

For the first term, it gets involved by the nonstationarity of $\boldsymbol{\theta}_k$. By rearranging the summation orders and several calculation steps, we have

$$\left| \mathbf{a}^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \sum_{t=1}^{k} \frac{\mathbf{a}_t \mathbf{a}_t^\top}{\bar{\sigma}_t^2} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_k) \right| \leq \sum_{t=1}^{k-1} |\mathbf{a}^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}| \cdot \| \frac{\mathbf{a}_t}{\bar{\sigma}_t} \|_2 \cdot \| \sum_{s=t}^{k-1} (\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}) \|_2 \leq \frac{A^2}{\alpha} \sqrt{\frac{dw}{\lambda}} \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 ,$$

We would like to highlight the subtleties in both our algorithm design and analysis to get the desired improvement. First, from here, we can see the necessity of introducing α in the design of $\bar{\sigma}_k$ in Eq.(4.1), which makes it possible to upper bound $\bar{\sigma}_k^{-1}$ and get a tunable α in the drifting term, which can subsequently be used to optimize the regret bound. Second, we show that it is essential to split the term $\bar{\sigma}_t^{-2}$ as how we did. Only by doing that can we bound the $\sum_{t=1}^{s} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}$ term by d with the elliptical potential lemma. Otherwise, we can a $1/\alpha^2$ term rather than the A/α term, which will hurt the final regret bound. For the second term in Eq.(4.4), a vanilla way to control it is adopting a self-normalized concentration inequality from [2]. However, it can not utilize variance information, but just the magnitude of the noise, which fails to get a tight bound with the variance information. Inspired by [45, 46, 41], we adapt a variance-adaptive concentration inequality in Theorem F.1 to get a tighter bound. Similar arguments also hold for the proof of Theorem 5.1 for the unknown variance case. We refer to Appendix B for the full proof.

Lemma 4.1 suggests that under the non-stationary setting, the difference between the true expected reward and our estimated reward will be upper bounded by two separate terms. The first drifting term charcterizes the error caused by the non-stationary environment, and the second stochastic term charcterizes the error caused by the estimation of the stochastic environment. Note that similar bound has also been discovered in Touati and Vincent [33]. We want to emphasize that our bound differs from existing ones in 1) an additional variance parameter α in the drifting term, and 2) a weighted convariance matrix $\widehat{\mathbf{\Sigma}}$ rather than a vanilla convariance matrix.

Next, we present our first main theorem.

Theorem 4.2. Let $0 < \delta < 1$. Suppose that for all $k \geq 1$ and all $\mathbf{a} \in \mathcal{D}_k$, $\langle \mathbf{a}, \boldsymbol{\theta}_k \rangle \in [-1, 1]$, $\|\boldsymbol{\theta}^*\|_2 \leq B$, $\|\mathbf{a}\|_2 \leq A$. With probability at least $1 - \delta$, the regret of Restarted-WeightedOFUL$^+$ is bounded by

$$\text{Regret}(K) \leq \frac{2A^2 B_K w^{\frac{3}{2}}}{\alpha}\sqrt{\frac{d}{\lambda}} + 4\widehat{\beta}\sqrt{V_K + K\alpha^2}\sqrt{\frac{Kd\iota}{w}} + \frac{4d\iota K\widehat{\beta}\gamma^2}{w} + \frac{4d\iota K}{w}, \tag{4.5}$$

where $\iota = \log(1 + \frac{wA^2}{d\lambda\alpha^2})$, and $\widehat{\beta} = \widetilde{O}(\sqrt{d} + R/\gamma^2 + \sqrt{\lambda}B)$. Specifically, by treating A, λ, B, R as constants and setting $\gamma^2 = R/\sqrt{d}$, we have

$$\text{Regret}(K) = \widetilde{O}(B_K w^{3/2} d^{1/2}\alpha^{-1} + dK\alpha/\sqrt{w} + d\sqrt{KV_K/w} + dK/w). \tag{4.6}$$

Proof. See Appendix C. □

Remark 4.3. For the stationary linear bandit case where $B_K = 0$, we can set the restart window size $w = K$ and the variance parameter $\alpha = 1/\sqrt{K}$, then we obtain an $\widetilde{O}(d\sqrt{V_K} + d)$ regret for Algorithm 8, which is identical to the one in Zhou and Gu [45].

Next, we aim to select parameters α and w in order to optimize (4.6).

Corollary 4.4. Assume that $B_K, V_K \in [\Omega(1), O(K)]$. Then by selecting

$$\begin{aligned} w &= d^{1/4}\sqrt{V_K/B_K}, && dV_K^6 \geq K^4 B_K^2, \\ w &= d^{1/6}(K/B_K)^{1/3} && \text{otherwise.} \end{aligned}$$

and $\alpha = d^{-1/4}B_K^{1/2} w K^{-1/2}$, the final regret is in the order

$$\text{Regret}(K) = \widetilde{O}(d^{7/8}(B_K V_K)^{1/4}\sqrt{K} + d^{5/6}B_K^{1/3}K^{2/3}). \tag{4.7}$$

Remark 4.5. We compare the regret of Algo.8 in Corollary 4.4 with previous results in the following special cases.

- In the worst case where $V_K = O(K)$, our result becomes $\widetilde{O}(d^{7/8}B_K^{1/4}K^{3/4})$, matching the state-of-the-art results for restarting and sliding window strategies [10, 43].
- In the case where the *total variance* is small, *i.e.*, $V_K = \widetilde{O}(1)$, assuming that $K^4 > d$, our result becomes $\widetilde{O}(d^{5/6}B_K^{1/3}K^{2/3})$, better than all the previous results [10, 43, 34, 39].

Remark 4.6. Wei et al. [38] has studied non-stationary MAB with dynamic variance. With the knowledge of V_K and B_K, Wei et al. [38] proposed a restart-based Rerun-UCB-V algorithm with a $\widetilde{O}(|\mathcal{A}|^{\frac{2}{3}} B_K^{\frac{1}{3}} V_K^{\frac{1}{3}} K^{\frac{1}{3}} + |\mathcal{A}|^{\frac{1}{2}} B_K^{\frac{1}{2}} K^{\frac{1}{2}})$ regret, where $\mathcal{A}$ is the action set. Reduced to the MAB setting, our Restarted-WeightedOFUL$^+$ achieves an $\widetilde{O}(|\mathcal{A}|^{7/8}(B_K V_K)^{1/4}\sqrt{K} + |\mathcal{A}|^{5/6}B_K^{1/3}K^{2/3})$ regret, which is worse than Wei et al. [38]. We claim that this is due to the generality of the linear bandits, which brings us a looser bound to the drifting term in Lemma 4.1. When restricting to the MAB setting, our drifting term enjoys a tighter bound, which could further tighten our final regret. To develop an algorithm achieving the same regret as Wei et al. [38] is beyond the scope of this work.

Remark 4.7. Wei et al. [38] has established a lower bound $\widetilde{\Omega}(B_K^{\frac{1}{3}} V_K^{\frac{1}{3}} K^{\frac{1}{3}} + B_K^{\frac{1}{2}} K^{\frac{1}{2}})$ for MAB with total variance V_K and total variation budget B_K. There still exist gaps between our regret and their lower bound regarding the dependence of K, V_K, B_K, and we leave to fix the gaps as future work.

Algorithm 2 Restarted SAVE$^+$

Require: $\alpha > 0$; the upper bound on the ℓ_2-norm of $\mathbf{a}$ in $\mathcal{D}_k (k \geq 1)$, i.e., A; the upper bound on the ℓ_2-norm of $\boldsymbol{\theta}_k$ $(k \geq 1)$, i.e., B; restart window size w.

1: Initialize $L \leftarrow \lceil \log_2(1/\alpha) \rceil$.

2: Initialize the estimators for all layers: $\widehat{\boldsymbol{\Sigma}}_{1,\ell} \leftarrow 2^{-2\ell} \cdot \mathbf{I}$, $\widehat{\mathbf{b}}_{1,\ell} \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_{1,\ell} \leftarrow \mathbf{0}$, $\widehat{\beta}_{1,\ell} \leftarrow 2^{-\ell+1}$, $\widehat{\Psi}_{1,\ell} \leftarrow \emptyset$ for all $\ell \in [L]$.

3: **for** $k = 1, \ldots, K$ **do**

4: **if** $k \% w == 0$ **then**

5: Set $\widehat{\boldsymbol{\Sigma}}_{k,\ell} \leftarrow 2^{-2\ell} \cdot \mathbf{I}$, $\widehat{\mathbf{b}}_{k,\ell} \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_{k,\ell} \leftarrow \mathbf{0}$, $\widehat{\beta}_{1,\ell} \leftarrow 2^{-\ell+1}$, $\widehat{\Psi}_{k,\ell} \leftarrow \emptyset$ for all $\ell \in [L]$.

6: **end if**

7: Observe $\mathcal{D}_k$, choose $\mathbf{a}_k \leftarrow \operatorname{argmax}_{\mathbf{a} \in \mathcal{D}_k} \min_{\ell \in [L]} \langle \mathbf{a}, \widehat{\boldsymbol{\theta}}_{k,\ell} \rangle + \widehat{\beta}_{k,\ell} \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}}$ and observe r_k.

8: Set $\ell_k \leftarrow L + 1$

9: Let $\mathcal{L}_k \leftarrow \{\ell \in [L] : \|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} \geq 2^{-\ell}\}$, set $\ell_k \leftarrow \min(\mathcal{L}_k)$ if $\mathcal{L}_k \neq \emptyset$

10: $\widehat{\Psi}_{k,\ell_k} \leftarrow \widehat{\Psi}_{k,\ell_k} \cup \{k\}$

11: **if** $\mathcal{L}_k \neq \emptyset$ **then**

12: Set $w_k \leftarrow \frac{2^{-\ell_k}}{\|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_k}^{-1}}}$ and update

$$\widehat{\boldsymbol{\Sigma}}_{k+1,\ell_k} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,\ell_k} + w_k^2 \mathbf{a}_k \mathbf{a}_k^\top, \widehat{\mathbf{b}}_{k+1,\ell} \leftarrow \widehat{\mathbf{b}}_{k,\ell_k} + w_k^2 \cdot r_k \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k+1,\ell_k} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k+1,\ell_k}^{-1} \widehat{\mathbf{b}}_{k+1,\ell_k}.$$

13: Compute the adaptive confidence radius $\widehat{\beta}_{k+1,l}$for the next round according to (5.1).

14: **end if**

15: For $\ell \neq \ell_k$ let $\widehat{\boldsymbol{\Sigma}}_{k+1,\ell} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,\ell}$, $\widehat{\mathbf{b}}_{k+1,\ell} \leftarrow \widehat{\mathbf{b}}_{k,\ell}$, $\widehat{\boldsymbol{\theta}}_{k+1,\ell} \leftarrow \widehat{\boldsymbol{\theta}}_{k,\ell}$, $\widehat{\beta}_{k+1,\ell} \leftarrow \widehat{\beta}_{k,\ell}$.

16: **end for**

5 Non-stationary Linear Contextual Bandit with Unknown Variance and Total Variation Budget

By Theorem 4.2, we know that Algorithm 8 is able to utilize the total variance V_K and obtain a better regret result compared with existing algorithms which do not utilize V_K. However, the success of Algorithm 8 depends on the knowledge of the per-round variance σ_k, and it also depends on a good selection of restart window size w, whose optimal selection depends on both V_K and B_K. In this section, we aim to relax these two requirements with still better regret results.

5.1 Unknown Per-round Variance, Known V_K and B_K

We first aim to relax the requirement that each σ_k^2 is known to the agent at the beginning of k-th round. We follow the SAVE algorithm [41] which introduces a multi-layer structure [13, 21] to deal with unknown σ_k^2. In detail, SAVE maintains multiple estimates to the current feature vector θ_k, which we denote them as $\widehat{\boldsymbol{\theta}}_{k,1}, ..., \widehat{\boldsymbol{\theta}}_{k,L}$ in line 2. Each $\widehat{\boldsymbol{\theta}}_{k,\ell}$ is calculated based on a subset $\widehat{\Psi}_{k,\ell} \subseteq [k-1]$ of samples $\{(\mathbf{a}_t, r_t)\}$. The rule that whether to add the current k to some $\widehat{\Psi}_{k,\ell}$ is based on the uncertainty of $\mathbf{a}_k$ with the sample set $\{(\mathbf{a}_t, r_t)\}_{t \in \widehat{\Psi}_{k,\ell}}$. As long as $\mathbf{a}_k$ is too uncertain w.r.t. some level ℓ_k (line 9), we add k to $\widehat{\Psi}_{k,\ell}$ and update the estimate $\widehat{\boldsymbol{\theta}}_{k,\ell_k}$ accordingly (line 12). Each $\widehat{\boldsymbol{\theta}}_{k,\ell_k}$ is calculated as the solution of a weighted regression problem, where the weight w_k is selected as the inverse of the uncertainty of the arm $\mathbf{a}_k$ w.r.t. the samples in the ℓ-th layer. Maintaining L different $\widehat{\boldsymbol{\theta}}_{k,\ell}, \ell \in [L]$, Algorithm 2 then calculates L number of UCB for each arm $\mathbf{a}$ w.r.t. L different $\widehat{\boldsymbol{\theta}}_{k,\ell}$, and selects the arm which maximizes the minimization of L UCBs (line 7). It has been shown in Zhao et al. [41] that such a multilayer structure is able to utilize the V_K information without knowing the per-round variance σ_k^2. Similar to Algorithm 8, in order to deal with the nonstationarity issue, we introduce a restarting scheme that Algorithm 2 restarts itself by a restart window size w (line 5).

Next we show the theoretical guarantee of Algorithm 2. We call the restart time rounds *grids* and denote them by $g_1, g_2, \ldots g_{\lceil \frac{K}{w} \rceil - 1}$, where $g_i \% w = 0$ for all $i \in [\lceil \frac{K}{w} \rceil - 1]$. Let i_k be the grid index of time round k, *i.e.*, $g_{i_k} \leq k < g_{i_k+1}$. We denote $\widehat{\Psi}_{k,\ell} := \{t : t \in [g_{i_k}, k-1], \ell_t = \ell\}$. We define

the confidence radius $\widehat{\beta}_{k,\ell}$ at round k and layer ℓ as

$$\widehat{\beta}_{k,\ell} := 16 \cdot 2^{-\ell} \sqrt{\Big(8\widehat{\mathrm{Var}}_{k,\ell} + 6R^2 \log(\frac{4(w+1)^2 L}{\delta}) + 2^{-2\ell+4}\Big)} \times \sqrt{\log(\frac{4w^2 L}{\delta})} + 6 \cdot 2^{-\ell} R \log(\frac{4w^2 L}{\delta}) + 2^{-\ell} B, \tag{5.1}$$

$$\text{where } \widehat{\mathrm{Var}}_{k,\ell} := \begin{cases} \sum_{i \in \widehat{\Psi}_{k,\ell}} w_i^2 \big(r_i - \langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \mathbf{a}_i \rangle\big)^2, & \text{If} \quad 2^\ell \geq 64 \sqrt{\log\Big(\frac{4(w+1)^2 L}{\delta}\Big)} \\ R^2 \Big|\widehat{\Psi}_{k,\ell}\Big|, & \text{otherwise.} \end{cases}$$

Note that our selection of the confidence radius $\widehat{\beta}_{k,\ell}$ only depends on $\widehat{\mathrm{Var}}_{k,\ell}$, which serves as an estimate of the total variance of samples at ℓ-th layer without knowing σ_k^2.

We build the theoretical guarantee of Algorithm 2 as follows.

Theorem 5.1. Let $0 < \delta < 1$. Suppose that for all $k \geq 1$ and all $\mathbf{a} \in \mathcal{D}_k$, $\langle \mathbf{a}, \boldsymbol{\theta}_k \rangle \in [-1, 1]$, $\|\boldsymbol{\theta}^*\|_2 \leq B$, $\|\mathbf{a}\|_2 \leq A$. If $\{\beta_{k,\ell}\}_{k \geq 1, \ell \in [L]}$ is defined in (5.1), then the cumulative regret of Algorithm 2 is bounded as follows with probability at least $1 - 3\delta$:

$$\text{Regret}(K) = \widetilde{O}\Big(\frac{A^2 \sqrt{d} w^{\frac{3}{2}} B_K}{\alpha} + \big(w\alpha^2 + d\big) \cdot \sqrt{\frac{K}{w} V_K} + \big(1 + R\big) \cdot \big(K\alpha^2 + \frac{Kd}{w}\big)\Big) \tag{5.2}$$

Specifically, regarding A, R as constants, we have

$$\text{Regret}(K) = \widetilde{O}(\sqrt{d} w^{1.5} B_K / \alpha + \alpha^2 (K + \sqrt{wKV_K}) + d\sqrt{KV_K / w} + dK/w).$$

Proof. See Appendix D for the full proof. □

Remark 5.2. Like Remark 4.3, we consider the case where $B_K = 0$. We set $w = K$ and $\alpha^2 = 1/K\sqrt{V_K}$, then we obtain a regret $\widetilde{O}(d\sqrt{V_K} + d)$, which matches the regret of the SAVE algorithm in Zhao et al. [41].

Corollary 5.3. Assume that $B_K, V_K \in [\Omega(1), O(K)]$, then by selecting

$$\begin{aligned} &w = d^{1/3}(K/B_K)^{1/3}, && K^2 \geq V_K^3 d / B_K, \\ &w = d^{2/5}(KV_K)^{1/5} / B_K^{2/5} && \text{otherwise.} \end{aligned}$$

and $\alpha = d^{1/6}\sqrt{w} B_K^{1/3} / (K^{1/3} + (V_K K w)^{1/6})$, we have

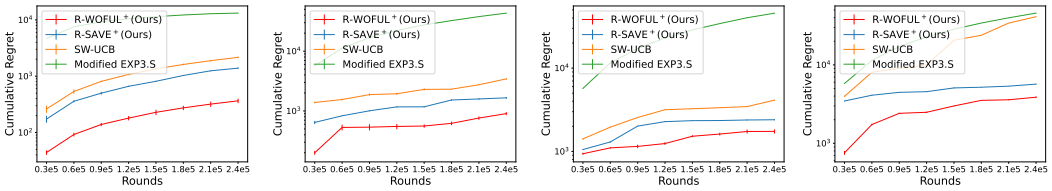
$$\text{Regret}(K) = \widetilde{O}(d^{4/5} V_K^{2/5} B_K^{1/5} K^{2/5} + d^{2/3} B_K^{1/3} K^{2/3}). \tag{5.3}$$

Remark 5.4. We discuss the regret of Algo.2 in Corollary 5.3 in the following special cases. In the case where the *total variance* is small, *i.e.*, $V_K = \widetilde{O}(1)$, assuming that $K^2 > d$, our result becomes $\widetilde{O}(d^{2/3} B_K^{1/3} K^{2/3})$, better than all the previous results [10, 43, 34, 39]. In the worst case where $V_K = O(K)$, our result becomes $\widetilde{O}(d^{4/5} B_K^{1/5} K^{4/5})$.

Unknown Per-round Variance, Unknown V_K and B_K In Corollary 5.3, we need to know the *total variance* V_K and *total variation budget* B_K to select the optimal w and α. To deal with the more general case where V_K and B_K are unknown, we can employ the *Bandits-over-Bandits* (BOB) mechanism ([11, 34, 43]). We name the Restarted SAVE$^+$ algorithm with BOB mechanism as "Restarted SAVE$^+$-BOB". Due to the space limit, we put the algorithm design, descriptions, and theoretical analysis of Restarted SAVE$^+$-BOB (Algo.3) in Appendix A.

6 Experiments

To validate the effectiveness of our methods, we conduct a series of experiments on the synthetic data. All the experiments are run on an AMD Ryzen5 7640H CPU. About 60 hours are needed to implement all the experiments.



(a) $B_K = 1$ (b) $B_K = 10$ (c) $B_K = 20$ (d) $B_K = K^{1/3}$

Figure 1: The regret of Restarted-WeightedOFUL$^+$, Restarted SAVE$^+$, SW-UCB and Modified EXP3.S under different total rounds.

Problem Setting Following the experimental set up in [11], we consider the 2-armed bandits setting, where the action set $\mathcal{D}_k = \{(1, 0), (0, 1)\}$, and

$$\boldsymbol{\theta}_k = \begin{pmatrix} 0.5 + \frac{3}{10}\sin(5B_K\pi k/K) \\ 0.5 + \frac{3}{10}\sin(\pi + 5B_K\pi k/K) \end{pmatrix}.$$

It is easy to see that the total variation budget can be bounded as B_K. At each round k, the ϵ_k satisfies $\epsilon_k \sim \text{Bernoulli}(0.5/k) - 0.5/k$. We can verify that under such a distribution for ϵ_k, the variance of the reward distribution at k-th round is $(1 - 0.5/k) \cdot 0.5/k$, and the total variance $V_K \sim \log K$.

Baseline algorithms We compare the proposed Restarted-WeightedOFUL$^+$ and Restarted SAVE$^+$ with SW-UCB [11] and Modified EXP3.S [6]. For Restarted-WeightedOFUL$^+$, we set $\lambda = 1$, $\widehat{\beta}_k = 10$, $w = 1000$, and we grid search the variance parameters α and γ, both among values {1, 1.5, 2, 2.5, 3}. Finally we set $\alpha = 1$, and $\gamma = 2$. For Restarted SAVE$^+$ we set $w = 1000$, $\widehat{\beta}_{k,\ell} = 2^{-\ell+1}$, and grid search L from 1 to 10 with stepsize of 1 and finally choose $L = 6$. For SW-UCB, we set $\lambda = 1$, $w = 1000$, $\beta_k = 10$. The Modified EXP3.S requires two parameters $\bar{\alpha}$ and $\bar{\gamma}$, and we set $\bar{\gamma} = 0.01$ and $\bar{\alpha} = \frac{1}{K}$.

To test the algorithms' performance under different total time horizons, we let K vary from 3×10^4 to 2.4×10^5, with a stepsize of 3×10^4, and plot the cumulative regret $\text{Regret}(K)$ for these different *total time step* K. We set $B_K = 1, 10, 20$, and $K^{1/3}$ to observe their performance with different B_K.

Result We plot the results in Figure.1, where all the empirical results are averaged over ten independent trials and the error bar is the standard error divided by $\sqrt{10}$. The results are consistent with our theoretical findings. It is evident that our algorithms significantly outperform both SW-UCB and Modified EXP3.S. Among our proposed algorithms, Restarted-WeightedOFUL$^+$ achieves the best performance. This can be attributed to the fact that it knows the variance and can make more informed decisions. Although Restarted SAVE$^+$ performed slightly worse than Restarted-WeightedOFUL$^+$, it still outperforms the baseline algorithms, particularly when $B_K = K^{1/3}$. These results highlight the superiority of our methods.

7 Conclusion and Future Work

We study non-stationary stochastic linear bandits in this work. We propose Restarted-WeightedOFUL$^+$ and Restarted SAVE$^+$, two novel algorithms that utilize the dynamic variance information of the dynamic reward distribution. We show that both of our algorithms are able to achieve better dynamic regret compared with best existing results [39] under several parameter regimes, *e.g.*, when the total variance V_K is small. Experiment results backup our theoretical claim. It is worth noting there still exist gaps between our current obtained regret and the lower bound [38], and to fix such a gap is left as our future work.

References

[1] ABBASI-YADKORI, Y., GYÖRGY, A. and LAZIĆ, N. (2023). A new look at dynamic regret for non-stationary stochastic bandits. *Journal of Machine Learning Research* **24** 1–37.

[2] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

[3] AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning* **47** 235–256.

[4] AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing* **32** 48–77.

[5] AUER, P., GAJANE, P. and ORTNER, R. (2019). Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory* (A. Beygelzimer and D. Hsu, eds.), vol. 99 of *Proceedings of Machine Learning Research*. PMLR.

[6] BESBES, O., GUR, Y. and ZEEVI, A. (2014). Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *SSRN Electronic Journal* .

[7] BESBES, O., GUR, Y. and ZEEVI, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems* **27**.

[8] CHEN, W., WANG, L., ZHAO, H. and ZHENG, K. (2021). Combinatorial semi-bandit in the non-stationary environment. In *Uncertainty in Artificial Intelligence*. PMLR.

[9] CHEN, Y., LEE, C.-W., LUO, H. and WEI, C.-Y. (2019). A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Thirty-Second Conference on Learning Theory* (A. Beygelzimer and D. Hsu, eds.), vol. 99 of *Proceedings of Machine Learning Research*. PMLR.

[10] CHEUNG, W. C., SIMCHI-LEVI, D. and ZHU, R. (2018). Hedging the drift: Learning to optimize under non-stationarity. *Available at SSRN 3261050* .

[11] CHEUNG, W. C., SIMCHI-LEVI, D. and ZHU, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

[12] CHEUNG, W. C., SIMCHI-LEVI, D. and ZHU, R. (2020). Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*. PMLR.

[13] CHU, W., LI, L., REYZIN, L. and SCHAPIRE, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings.

[14] CLERICI, G., LAFORGUE, P. and CESA-BIANCHI, N. (2023). Linear bandits with memory: from rotting to rising.

[15] DAI, Y., WANG, R. and DU, S. S. (2022). Variance-aware sparse linear bandits. *arXiv preprint arXiv:2205.13450* .

[16] DENG, Y., ZHOU, X., KIM, B., TEWARI, A., GUPTA, A. and SHROFF, N. (2022). Weighted gaussian process bandits for non-stationary environments. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

[17] FAURY, L., RUSSAC, Y., ABEILLE, M. and CALAUZÈNES, C. (2021). Regret bounds for generalized linear bandits under parameter drift. *arXiv preprint arXiv:2103.05750* .

[18] FREEDMAN, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability* 100–118.

[19] GAJANE, P., ORTNER, R. and AUER, P. (2018). A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066* .

[20] GARIVIER, A. and MOULINES, E. (2011). On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory* (J. Kivinen, C. Szepesvári, E. Ukkonen and T. Zeugmann, eds.). Springer Berlin Heidelberg, Berlin, Heidelberg.

[21] HE, J., ZHOU, D. and GU, Q. (2021). Uniform-pac bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems* **34** 14188–14199.

[22] KIM, B. and TEWARI, A. (2020). Randomized exploration for non-stationary stochastic linear bandits. In *Uncertainty in Artificial Intelligence*.

[23] KIM, B. and TEWARI, A. (2020). Randomized exploration for non-stationary stochastic linear bandits. In *Conference on Uncertainty in Artificial Intelligence*. PMLR.

[24] KIM, Y., YANG, I. and JUN, K.-S. (2022). Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems* **35** 1060–1072.

[25] KIRSCHNER, J. and KRAUSE, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*. PMLR.

[26] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

[27] LIU, Y., VAN ROY, B. and XU, K. (2023). A definition of non-stationary bandits. *arXiv preprint arXiv:2302.12202* .

[28] MAO, W., ZHANG, K., ZHU, R., SIMCHI-LEVI, D. and BASAR, T. (2021). Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*. PMLR.

[29] RUSSAC, Y., CAPPÉ, O. and GARIVIER, A. (2020). Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113* .

[30] RUSSAC, Y., FAURY, L., CAPPÉ, O. and GARIVIER, A. (2021). Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

[31] RUSSAC, Y., VERNADE, C. and CAPPÉ, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems* .

[32] SUK, J. and KPOTUFE, S. (2022). Tracking most significant arm switches in bandits. In *Conference on Learning Theory*. PMLR.

[33] TOUATI, A. and VINCENT, P. (2020). Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870* .

[34] WANG, J., ZHAO, P. and ZHOU, Z.-H. (2023). Revisiting weighted strategy for non-stationary parametric bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

[35] WANG, Z., LIU, X., LI, S. and LUI, J. C. (2023). Efficient explorative key-term selection strategies for conversational contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37.

[36] WANG, Z., XIE, J., LIU, X., LI, S. and LUI, J. (2024). Online clustering of bandits with misspecified user models. *Advances in Neural Information Processing Systems* **36**.

[37] WANG, Z., XIE, J., YU, T., LI, S. and LUI, J. (2024). Online corrupted user detection and regret minimization. *Advances in Neural Information Processing Systems* **36**.

[38] WEI, C.-Y., HONG, Y.-T. and LU, C.-J. (2016). Tracking the best expert in non-stationary stochastic environments. *Advances in neural information processing systems* **29**.

[39] WEI, C.-Y. and LUO, H. (2021). Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*. PMLR.

[40] ZHANG, Z., YANG, J., JI, X. and DU, S. S. (2021). Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems* **34** 4342–4355.

[41] ZHAO, H., HE, J., ZHOU, D., ZHANG, T. and GU, Q. (2023). Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. *arXiv preprint arXiv:2302.10371* .

[42] ZHAO, P. and ZHANG, L. (2021). Non-stationary linear bandits revisited. *arXiv preprint arXiv:2103.05324* .

[43] ZHAO, P., ZHANG, L., JIANG, Y. and ZHOU, Z.-H. (2020). A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

[44] ZHAO, P., ZHANG, L., JIANG, Y. and ZHOU, Z.-H. (2020). A simple approach for non-stationary linear bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.), vol. 108 of *Proceedings of Machine Learning Research*. PMLR.

[45] ZHOU, D. and GU, Q. (2022). Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems* **35** 36337–36349.

[46] ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

[47] ZUO, J., ZHANG, Z., WANG, Z., LI, S., HAJIESMAILI, M. and WIERMAN, A. (2024). Adversarial attacks on online learning to rank with click feedback. *Advances in Neural Information Processing Systems* **36**.

A Restarted SAVE$^+$-BOB

In this section, we provide the details of our proposed Restarted SAVE$^+$-BOB algorithm. The Restarted SAVE$^+$-BOB algorithm is summarized in Algo.3. We divide the K rounds into $\lceil \frac{K}{H} \rceil$ blocks, with each block having H rounds (except the last one may have less than H). Within each block i, we use a fixed (α_i, w_i) pair to run the Restarted SAVE$^+$ algorithm. To adaptively learn the optimal (α, w) pair without the knowledge of V_K and B_K, we employ an adversarial bandit algorithm (Exp3 in [4]) as the meta-learner to select α_i, w_i over time for $i \in \lceil \frac{K}{H} \rceil$ blocks. Specifically, in each block, the meta learner selects a (α, w) pair from the candidate pool to feed to Restarted SAVE$^+$, and the cumulative reward received by Restarted SAVE$^+$ within the block is fed to the meta-learner as the reward feedback to select a better pair for the next block.

We set H to be $\lceil d^{\frac{2}{5}} K^{\frac{2}{5}} \rceil$, and set the candidate pool of (α, w) pairs for the Exp3 algorithm as:

$$\mathcal{P} = \{(w, \alpha) : w \in \mathcal{W}, \alpha \in \mathcal{J}\}, \tag{A.1}$$

where

$$\mathcal{W} = \{w_i = d^{\frac{1}{3}} 2^{i-1} | i \in \lceil \frac{1}{3} \log_2 K \rceil + 1\} \cup \{w_i = d^{\frac{2}{5}} 2^{i-1} | i \in \lceil \frac{2}{5} \log_2 K \rceil + 1\}, \tag{A.2}$$

and

$$\mathcal{J} = \{\alpha_i = d^{\frac{1}{3}} 2^{-i+1} | i \in \lceil \frac{1}{3} \log_2 K \rceil + 1\} \cup \{\alpha_i = d^{\frac{11}{30}} 2^{-i+1} | i \in \lceil \frac{11}{30} \log_2 K \rceil + 1\}. \tag{A.3}$$

The algorithm also labels all the $|\mathcal{P}| = \big(\lceil \frac{1}{3} \log_2 K \rceil + \lceil \frac{2}{5} \log_2 K \rceil + 2\big) \cdot \big(\lceil \frac{1}{3} \log_2 K \rceil + \lceil \frac{11}{30} \log_2 K \rceil + 2\big)$ candidate pairs of parameters in $\mathcal{P}$,*i.e.*, $\mathcal{P} = \{(w_i, \alpha_i)\}_{i=1}^{|\mathcal{P}|}$. The algorithm initializes $\{s_{j,1}\}_{j=1}^{|\mathcal{P}|}$ to be $s_{j,1} = 1, \quad \forall j = 0, 1, \ldots, |\mathcal{P}|$, which means that at the beginning, the algorithm selects a pair from $\mathcal{P}$ uniformly at random. At the beginning of each block $i \in [\lceil K/H \rceil]$, the meta-learner (Exp3) calculates the distribution $(p_{j,i})_{j=1}^{|\mathcal{P}|}$ over the candidate set $\mathcal{P}$ by

$$p_{j,i} = (1 - \gamma) \frac{s_{j,i}}{\sum_{u=1}^{|\mathcal{P}|} s_{u,i}} + \frac{\gamma}{|\mathcal{P}| + 1}, \quad \forall j = 1, \ldots, |\mathcal{P}|, \tag{A.4}$$

where γ is defined as

$$\gamma = \min \left\{ 1, \sqrt{\frac{(|\mathcal{P}| + 1) \ln(|\mathcal{P}| + 1)}{(e - 1) \lceil K/H \rceil}} \right\}. \tag{A.5}$$

Then, the meta-learner draws a j_i from the distribution $(p_{j,i})_{j=1}^{|\mathcal{P}|}$, and sets the pair of parameters in block i to be (w_{j_i}, α_{j_i}), and runs the base algorithm Algo.2 from scratch in this block with (w_{j_i}, α_{j_i}), then feeds the cumulative reward in the block $\sum_{k=(i-1)H+1}^{\min\{i \cdot H, K\}} r_k$ to the meta-learner. The meta-learner rescales $\sum_{k=(i-1)H+1}^{\min\{i \cdot H, K\}} r_k$ to $\frac{\sum_{k=(i-1)H+1}^{\min\{i \cdot H, K\}} r_k}{H + R\sqrt{\frac{H}{2} \log\left(K(\frac{K}{H} + 1)\right) + \frac{2}{3} \cdot R \log\left(K(\frac{K}{H} + 1)\right)}}$ to make it in the range $[0, 1]$ with high probability (supported by Lemma F.7). The meta-learner updates the parameter $s_{j_i, i+1}$ to be

$$s_{j_i, i+1} = s_{j_i, i} \cdot \exp\left(\frac{\gamma}{(|\mathcal{P}| + 1) p_{j_i, i}} \left(\frac{1}{2} + \frac{\sum_{k=(i-1)H+1}^{\min\{i \cdot H, K\}} r_k}{H + R\sqrt{\frac{H}{2} \log\left(K(\frac{K}{H} + 1)\right) + \frac{2}{3} \cdot R \log\left(K(\frac{K}{H} + 1)\right)}} \right) \right), \tag{A.6}$$

and keep others unchanged, *i.e.*, $s_{u,i+1} = s_{u,i}, \ \forall u \neq j_i$. After that, the algorithm will go to the next block, and repeat the same process in block $i + 1$.

We have the following theorem to bound the regret of Restarted SAVE$^+$-BOB.

Theorem A.1. By using the BOB framework with Exp3 as the meta-algorithm and Restarted SAVE$^+$ as the base algorithm, with the candidate pool $\mathcal{P}$ for Exp3 specified as in Eq.(A.1), Eq.(A.2), Eq.(A.3), and $H = \lceil d^{\frac{2}{5}} K^{\frac{2}{5}} \rceil$, then the regret of Restarted SAVE$^+$-BOB (Algo.3) satisfies

$$\text{Regret}(K) = \widetilde{O}(d^{4/5} V_K^{2/5} B_K^{1/5} K^{2/5} + d^{2/3} B_K^{1/3} K^{2/3} + d^{2/5} K^{7/10}). \tag{A.7}$$

Algorithm 3 Restarted SAVE$^+$-BOB

Require: total time rounds K; problem dimension d; noise upper bound R; $\alpha > 0$; the upper bound on the ℓ_2-norm of $\mathbf{a}$ in $\mathcal{D}_k (k \geq 1)$, i.e., A; the upper bound on the ℓ_2-norm of $\boldsymbol{\theta}_k$ $(k \geq 1)$, i.e., B.

1: Initialize $H = \lceil d^{\frac{2}{5}} K^{\frac{2}{5}} \rceil$; $\mathcal{P}$ as defined in Eq.(A.1), and index the $|\mathcal{P}| = \left(\lceil \frac{1}{3}\log_2 K \rceil + \lceil \frac{2}{5}\log_2 K \rceil + 2\right) \cdot \left(\lceil \frac{1}{3}\log_2 K \rceil + \lceil \frac{11}{30}\log_2 K \rceil + 2\right)$ items in $\mathcal{P}$, *i.e.*, $\mathcal{P} = \{(w_i, \alpha_i)\}_{i=1}^{|\mathcal{P}|}$; $\gamma = \min\left\{1, \sqrt{\frac{(|\mathcal{P}|+1)\ln(|\mathcal{P}|+1)}{(e-1)\lceil K/H \rceil}}\right\}$; $\{s_{j,1}\}_{j=1}^{|\mathcal{P}|}$ is set to $s_{j,1} = 1, \quad \forall j = 0, 1, \ldots, |\mathcal{P}|$.

2: **for** $i = 1, 2, \ldots, \lceil K/H \rceil$ **do**

3: Calculate the distribution $(p_{j,i})_{j=1}^{|\mathcal{P}|}$ by $p_{j,i} = (1-\gamma)\frac{s_{j,i}}{\sum_{u=1}^{|\mathcal{P}|} s_{u,i}} + \frac{\gamma}{|\mathcal{P}|+1}, \quad \forall j = 1, \ldots, |\mathcal{P}|$.

4: Set $j_i \leftarrow j$ with probability $p_{j,i}$, and $(w_i, \alpha_i) \leftarrow (w_{i_i}, \alpha_{j_i})$.

5: Run Algo.2 from scratch in block i (*i.e.*, in rounds $k = (i-1)H+1, \ldots, \min\{i \cdot H, K\}$) with $(w, \alpha) = (w_i, \alpha_i)$.

6: Update $s_{j_i, i+1} = s_{j_i,i} \cdot \exp\left(\frac{\gamma}{(|\mathcal{P}|+1)p_{j_i,i}}\left(\frac{1}{2} + \frac{\sum_{k=(i-1)H+1}^{\min\{i\cdot H, K\}} r_k}{H + R\sqrt{\frac{H}{2}\log\left(K(\frac{K}{H}+1)\right) + \frac{2}{3}\cdot R\log\left(K(\frac{K}{H}+1)\right)}}\right)\right)$, and keep all the others unchanged, *i.e.*, $s_{u,i+1} = s_{u,i}$, $\forall u \neq j_i$.

7: **end for**

Proof. See Appendix E for the full proof. □

Remark A.2. We discuss the regret of Algo.3 in Corollary 5.3 in the following special cases. In the case where the *total variance* is small, *i.e.*, $V_K = \widetilde{O}(1)$, assuming $K^2 > d$, our result becomes $\widetilde{O}(d^{2/3}B_K^{1/3}K^{2/3} + d^{1/5}K^{7/10})$, when $d^{14}B_K^{10} > K$, it becomes $\widetilde{O}(d^{2/3}B_K^{1/3}K^{2/3})$, better than all the previous results [10, 43, 34, 39]. In the worst case where $V_K = O(K)$, our result becomes $\widetilde{O}(d^{4/5}B_K^{1/5}K^{4/5})$.

B Proof of Lemma 4.1

For simplicity, we denote

$$\widehat{\beta} := 12\sqrt{d\log(1+\frac{wA^2}{\alpha^2 d\lambda})\log(32(\log(\frac{\gamma^2}{\alpha}+1)\frac{w^2}{\delta})} + 30\log(32(\log(\frac{\gamma^2}{\alpha})+1)\frac{w^2}{\delta})\frac{R}{\gamma^2} + \sqrt{\lambda}B. \tag{B.1}$$

It is obvious that $\widehat{\beta} \geq \widehat{\beta}_k$ for all $k \in [K]$. We call the restart time rounds *grids* and denote them by $g_1, g_2, \ldots g_{\lceil \frac{K}{w} \rceil - 1}$, where $g_i \% w = 0$ for all $i \in [\lceil \frac{K}{w} \rceil - 1]$. Let i_k be the grid index of time round k, *i.e.*, $g_{i_k} \leq k < g_{i_k+1}$.

For ease of exposition and without loss of generality, we prove the lemma for $k \in [1, w]$. We calculate the estimation difference $|\boldsymbol{a}^\top(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)|$ for any $\mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}\|_2 \leq A$, $k \in [1, w]$. By definition:

$$\widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{b}_k = \widehat{\boldsymbol{\Sigma}}_k^{-1}(\sum_{t=1}^{k-1}\frac{r_t\mathbf{a}_t}{\bar{\sigma}_t^2}) = \widehat{\boldsymbol{\Sigma}}_k^{-1}(\sum_{t=1}^{k-1}\frac{\mathbf{a}_t\mathbf{a}_t^\top\boldsymbol{\theta}_t}{\bar{\sigma}_t^2} + \sum_{t=1}^{k-1}\frac{\mathbf{a}_t\epsilon_t}{\bar{\sigma}_t^2}), \tag{B.2}$$

where $\widehat{\Sigma}_k = \lambda\boldsymbol{I} + \sum_{t=g_{i_k}}^{k-1}\frac{\mathbf{a}_t\mathbf{a}_t^\top}{\bar{\sigma}_t^2}$.

Then we have

$$\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k = \widehat{\boldsymbol{\Sigma}}_k^{-1}(\sum_{t=1}^{k-1}\frac{\mathbf{a}_t\mathbf{a}_t^\top}{\bar{\sigma}_t^2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_k) + \sum_{t=1}^{k-1}\frac{\mathbf{a}_t\epsilon_t}{\bar{\sigma}_t^2}) - \lambda\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{\theta}_k\,. \tag{B.3}$$

Therefore

$$|\mathbf{a}^\top(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)| \leq \left|\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_k^{-1}\sum_{t=1}^{k-1}\frac{\mathbf{a}_t\mathbf{a}_t^\top}{\bar{\sigma}_t^2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_k)\right| + \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}\|\sum_{t=1}^{k-1}\frac{\mathbf{a}_t\epsilon_t}{\bar{\sigma}_t^2}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}} + \lambda\|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}}\|\widehat{\boldsymbol{\Sigma}}_k^{-\frac{1}{2}}\boldsymbol{\theta}_k\|_2\,, \tag{B.4}$$

where we use the Cauchy-Schwarz inequality.

For the first term, we have that for any $k \in [1, w]$

$$\begin{aligned}
\left|\mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \sum_{t=1}^{k} \frac{\mathbf{a}_t \mathbf{a}_t^\top}{\bar{\sigma}_t^2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_k)\right| &\leq \sum_{t=1}^{k-1} |\mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}| \cdot |\frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top (\sum_{s=t}^{k-1}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}))| \quad \text{(triangle inequality)} \\
&\leq \sum_{t=1}^{k-1} |\mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}| \cdot \|\frac{\mathbf{a}_t}{\bar{\sigma}_t}\|_2 \cdot \|\sum_{s=t}^{k-1}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1})\|_2 \\
&\qquad \text{(Cauchy-Schwarz)} \\
&\leq \frac{A}{\alpha} \sum_{t=1}^{k-1} |\mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}| \cdot \|\sum_{s=t}^{k-1}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1})\|_2 \quad (\|\mathbf{a}_t\| \leq A,\, \bar{\sigma}_t \geq \alpha) \\
&\leq \frac{A}{\alpha} \sum_{s=1}^{k-1} \sum_{t=1}^{s} |\mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}| \cdot \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|_2 \\
&\qquad (\textstyle\sum_{t=1}^{k-1}\sum_{s=t}^{k-1} = \sum_{s=1}^{k-1}\sum_{t=1}^{s}) \\
&\leq \frac{A}{\alpha} \sum_{s=1}^{k-1} \sqrt{\left[\sum_{t=1}^{s} \mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \mathbf{a}\right] \cdot \left[\sum_{t=1}^{s} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}\right]} \cdot ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 \\
&\qquad \text{(Cauchy-Schwarz)} \\
&\leq \frac{A}{\alpha} \sum_{s=1}^{k-1} \sqrt{\left[\sum_{t=1}^{s} \mathbf{a}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \mathbf{a}\right] \cdot d} \cdot ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 \qquad ((\star)) \\
&\leq \frac{A\|\mathbf{a}\|_2}{\alpha} \sqrt{d} \sum_{s=1}^{k-1} \sqrt{\frac{\sum_{t=1}^{k-1} 1}{\lambda}} \cdot ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 \quad (\lambda_{\max}(\widehat{\mathbf{\Sigma}}_k^{-1}) \leq \tfrac{1}{\lambda}) \\
&\leq \frac{A^2}{\alpha} \sqrt{\frac{dw}{\lambda}} \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2\,, \qquad \text{(B.5)}
\end{aligned}$$

where the inequality $(\star)$ follows from the fact that $\sum_{t=1}^{s} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t} \leq d$ that can be proved as follows. We have $\sum_{t=1}^{k-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t} = \sum_{t=1}^{k-1} \mathrm{tr}\left(\frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top \widehat{\mathbf{\Sigma}}_k^{-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t}\right) = \mathrm{tr}\left(\widehat{\mathbf{\Sigma}}_k^{-1} \sum_{t=1}^{k-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top\right)$. Given the eigenvalue decomposition $\sum_{t=1}^{k-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)^\top$, we have $\widehat{\mathbf{\Sigma}}_k = \mathrm{diag}(\lambda_1 + \lambda, \ldots, \lambda_d + \lambda)^\top$, and $\mathrm{tr}\left(\widehat{\mathbf{\Sigma}}_k^{-1} \sum_{t=1}^{k-1} \frac{\mathbf{a}_t}{\bar{\sigma}_t} \frac{\mathbf{a}_t}{\bar{\sigma}_t}^\top\right) = \sum_{i=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda} \leq d$.

For the second term, by the assumption on ϵ_k, we know that

$$\begin{aligned}
&|\epsilon_k/\bar{\sigma}_k| \leq R/\alpha, \\
&|\epsilon_k/\bar{\sigma}_k| \cdot \min\{1, \|\mathbf{a}_k/\bar{\sigma}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\} \leq R\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}/\bar{\sigma}_k^2 \leq R/\gamma^2, \\
&\mathbb{E}[\epsilon_k|\mathbf{a}_{1:k}, \epsilon_{1:k-1}] = 0,\ \mathbb{E}[(\epsilon_k/\bar{\sigma}_k)^2|\mathbf{a}_{1:k}, \epsilon_{1:k-1}] \leq 1,\ \|\mathbf{a}_k/\bar{\sigma}_k\|_2 \leq A/\alpha,
\end{aligned}$$

Therefore, setting $\mathcal{G}_k = \sigma(\mathbf{a}_{1:k}, \epsilon_{1:k-1})$, and using that σ_k is $\mathcal{G}_k$-measurable, applying Theorem F.1 to $(\boldsymbol{x}_k, \eta_k) = (\boldsymbol{a}_k/\bar{\sigma}_k, \epsilon_k/\bar{\sigma}_k)$ with $\epsilon = R/\gamma^2$, we get that with probability at least $1 - \delta$, for all $k \in [1, w]$,

$$\|\sum_{t=1}^{k-1} \frac{\mathbf{a}_t \epsilon_t}{\bar{\sigma}_t^2}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \leq 12\sqrt{d\log(1 + \frac{(k\%w)A^2}{\alpha^2 d\lambda})\log(32(\log(\frac{\gamma^2}{\alpha}+1)\frac{(k\%w)^2}{\delta})} + 30\log(32(\log(\frac{\gamma^2}{\alpha})+1)\frac{(k\%w)^2}{\delta})\frac{R}{\gamma^2}. \qquad \text{(B.6)}$$

For the last term

$$\lambda\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \|\widehat{\mathbf{\Sigma}}_k^{-\frac{1}{2}} \boldsymbol{\theta}_k\|_2 \leq \lambda\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \|\widehat{\mathbf{\Sigma}}_k^{-\frac{1}{2}}\|_2 \|\boldsymbol{\theta}_k\|_2 \leq \lambda\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \frac{1}{\sqrt{\lambda_{\min}(\widehat{\mathbf{\Sigma}}_k)}} \|\boldsymbol{\theta}_k\|_2 \leq \sqrt{\lambda} B \|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\,, \qquad \text{(B.7)}$$

where we use the fact that $\lambda_{\min}(\widehat{\mathbf{\Sigma}}_k) \geq \lambda$.

Therefore, with probabilty at least $1-\delta$, we have

$$\begin{aligned}
|\mathbf{a}^\top(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)| \leq & \frac{A^2}{\alpha}\sqrt{\frac{dw}{\lambda}}\sum_{t=1}^{k-1}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 \\
& + \|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\Bigg(12\sqrt{d\log(1+\frac{(k\%w)A^2}{\alpha^2 d\lambda})\log(32(\log(\frac{\gamma^2}{\alpha}+1)\frac{(k\%w)^2}{\delta})} \\
& + 30\log(32(\log(\frac{\gamma^2}{\alpha})+1)\frac{(k\%w)^2}{\delta})\frac{R}{\gamma^2} + \sqrt{\lambda}B\Bigg) \\
= & \frac{A^2}{\alpha}\sqrt{\frac{dw}{\lambda}}\sum_{t=1}^{k-1}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 + \widehat{\beta}_k\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\,, \qquad \text{(B.8)}
\end{aligned}$$

where $\widehat{\beta}_k$ is defined in Eq.(4.3).

C Proof for Theorem 4.2

For simplicity of analysis, we only analyze the regret over the first grid, *i.e.*, we try to analyze Regret($\widetilde{K}$) for $\widetilde{K} \in [1, w]$. Denote $\mathcal{E}_1$ as the event when Lemma 4.1 holds. Therefore, under event $\mathcal{E}_1$, for any $\widetilde{K} \in [1, w]$, the regret can be bounded by

$$\begin{aligned}
\text{Regret}(\widetilde{K}) = & \sum_{k=1}^{\widetilde{K}}\left[\langle \mathbf{a}_k^* - \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right] \\
= & \sum_{k=1}^{\widetilde{K}}\big[\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\rangle + (\langle \mathbf{a}_k^*, \widehat{\boldsymbol{\theta}}_k\rangle + \widehat{\beta}_k\|\mathbf{a}_k^*\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}) - (\langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_k\rangle + \widehat{\beta}_k\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}) + \langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\rangle \\
& + \widehat{\beta}_k\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} - \widehat{\beta}_k\|\mathbf{a}_k^*\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\big] \\
\leq & \frac{2A^2}{\alpha}\sqrt{\frac{dw}{\lambda}}\sum_{k=1}^{\widetilde{K}}\sum_{t=1}^{k-1}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 + 2\sum_{k=1}^{\widetilde{K}}\min\left\{1, \widehat{\beta}_k\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\}, \qquad \text{(C.1)}
\end{aligned}$$

where in the last inequality we use the definition of event $\mathcal{E}_1$, the arm selection rule in Line 7 of Algo.8, and $0 \leq \langle \mathbf{a}_k^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}^*\rangle \leq 2$.

Then we will bound the two terms in Eq.(C.1).

For the first term, we have

$$\begin{aligned}
& \frac{2A^2}{\alpha}\sqrt{\frac{dw}{\lambda}}\sum_{k=1}^{\widetilde{K}}\sum_{t=1}^{k-1}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 \\
= & \frac{2A^2}{\alpha}\sqrt{\frac{dw}{\lambda}}\sum_{t=1}^{\widetilde{K}-1}\sum_{k=t}^{\widetilde{K}}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 \\
\leq & \frac{2A^2}{\alpha}\sqrt{\frac{dw}{\lambda}}w\sum_{t=1}^{\widetilde{K}-1}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2\,. \qquad \text{(C.2)}
\end{aligned}$$

To bound the second term in Eq.(C.1), we decompose the set $[\widetilde{K}]$ into a union of two disjoint subsets $[K] = \mathcal{I}_1 \cup \mathcal{I}_2$.

$$\mathcal{I}_1 = \left\{k \in [\widetilde{K}] : \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \geq 1\right\},\ \mathcal{I}_2 = \left\{k \in [\widetilde{K}] : \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} < 1\right\}. \qquad \text{(C.3)}$$

Then the following upper bound of $|\mathcal{I}_1|$ holds:

$$|\mathcal{I}_1| = \sum_{k\in\mathcal{I}_1}\min\left\{1, \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|^2_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\}$$

$$\leq \sum_{k=1}^{\widetilde{K}} \min\left\{1, \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|^2_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\}$$
$$\leq 2d\iota, \tag{C.4}$$

where $\iota = \log(1 + \frac{wA^2}{d\lambda\alpha^2})$, the first equality holds since $\|\frac{\mathbf{x}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \geq 1$ for $k \in \mathcal{I}_1$, the last inequality holds due to Lemma F.2 together with the fact $\|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_2 \leq \frac{A}{\alpha}$ since $\bar{\sigma}_k \geq \alpha$ and $\|\mathbf{a}_k\|_2 \leq A$.

Then, we have

$$\begin{aligned}
&\sum_{k=1}^{\widetilde{K}} \min\left\{1, \widehat{\beta}_k \|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\} \\
&= \sum_{k\in\mathcal{I}_1} \min\left\{1, \bar{\sigma}_k\widehat{\beta}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\} + \sum_{k\in\mathcal{I}_2} \min\left\{1, \bar{\sigma}_k\widehat{\beta}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\} \\
&\leq \left[\sum_{k\in\mathcal{I}_1} 1\right] + \sum_{k\in\mathcal{I}_2} \bar{\sigma}_k\widehat{\beta}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \\
&\leq 2d\iota + \widehat{\beta} \sum_{k\in\mathcal{I}_2} \bar{\sigma}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}},
\end{aligned} \tag{C.5}$$

where the first inequality holds since $\min\{1, x\} \leq 1$ and also $\min\{1, x\} \leq x$, the second inequality holds by Eq.(C.4), and the fact the $\widehat{\beta} \geq \widehat{\beta}_k$ for all $k \in [K]$ ($\widehat{\beta}$ is defined in Eq.(B.1)). Next we further bound the second summation term in (C.5). We decompose $\mathcal{I}_2 = \mathcal{J}_1 \cup \mathcal{J}_2$, where

$$\mathcal{J}_1 = \left\{k \in \mathcal{I}_2 : \bar{\sigma}_k = \sigma_k \cup \bar{\sigma}_k = \alpha\right\},\ \mathcal{J}_2 = \left\{k \in \mathcal{I}_2 : \bar{\sigma}_k = \gamma\sqrt{\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}}\right\}.$$

Then $\sum_{k\in\mathcal{I}_2} \bar{\sigma}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} = \sum_{k\in\mathcal{J}_1} \bar{\sigma}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} + \sum_{k\in\mathcal{J}_2} \bar{\sigma}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}$. First, for $k \in \mathcal{J}_1$, we have

$$\begin{aligned}
\sum_{k\in\mathcal{J}_1} \bar{\sigma}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} &\leq \sum_{k\in\mathcal{J}_1} (\sigma_k + \alpha) \min\left\{1, \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\} \\
&\leq \sqrt{\sum_{k=1}^{\widetilde{K}} (\sigma_k + \alpha)^2} \sqrt{\sum_{k=1}^{\widetilde{K}} \min\left\{1, \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\}^2} \\
&\leq \sqrt{2\sum_{k=1}^{\widetilde{K}} (\sigma_k^2 + \alpha^2)} \sqrt{\sum_{k=1}^{\widetilde{K}} \min\left\{1, \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|^2_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\}} \\
&\leq 2\sqrt{\sum_{k=1}^{\widetilde{K}} \sigma_k^2 + \widetilde{K}\alpha^2\sqrt{d\iota}}\,,
\end{aligned} \tag{C.6}$$

where the first inequality holds since $\bar{\sigma}_k \leq \sigma_k + \alpha$ for $k \in \mathcal{J}_1$ and $\|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \leq 1$ since $k \in \mathcal{J}_1 \subseteq \mathcal{I}_2$, the second inequality holds by Cauchy-Schwarz inequality, the third inequality holds due to $(a+b)^2 \leq 2(a^2+b^2)$, and the last inequality holds due to Lemma F.2.

Finally we bound the summation for $k \in \mathcal{J}_2$. When $k \in \mathcal{J}_2$, we have $\bar{\sigma}_k = \gamma^2 \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\Sigma}_k^{-1}}$. Therefore we have

$$\begin{aligned}
\sum_{k\in\mathcal{J}_2} \bar{\sigma}_k \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} &= \sum_{k\in\mathcal{J}_2} \gamma^2 \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|^2_{\widehat{\mathbf{\Sigma}}_k^{-1}} \\
&\leq \sum_{k=1}^{\widetilde{K}} \gamma^2 \min\left\{1, \|\frac{\mathbf{a}_k}{\bar{\sigma}_k}\|^2_{\widehat{\mathbf{\Sigma}}_k^{-1}}\right\} \\
&\leq 2\gamma^2 d\iota\,,
\end{aligned} \tag{C.7}$$

where in the first inequality we use the fact that $\|\frac{\mathbf{a}_k}{\widetilde{\sigma}_k}\|_{\widehat{\mathbf{\Sigma}}_k^{-1}} \leq 1$ since $k \in \mathcal{J}_2 \subseteq \mathcal{I}_2$, and in the last inequality we use Lemma F.2.

Therefore, with Eq.(C.1), Eq.(C.2), Eq.(C.5), Eq.(C.6), Eq.(C.7), we can get the regret upper bound for $\widetilde{K} \in [1, w]$

$$\text{Regret}(\widetilde{K}) \leq \frac{2A^2 w^{\frac{3}{2}}}{\alpha}\sqrt{\frac{d}{\lambda}}\sum_{k=1}^{\widetilde{K}-1}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_2 + 4\widehat{\beta}\sqrt{d\iota}\sqrt{\sum_{k\in[\widetilde{K}]}\sigma_k^2 + w\alpha^2} + 4d\iota\gamma^2\widehat{\beta} + 4d\iota\,. \tag{C.8}$$

Therefore, by the same deduction, we can get that

$$\text{Regret}([g_i, g_{i+1}]) \leq \frac{2A^2 w^{\frac{3}{2}}}{\alpha}\sqrt{\frac{d}{\lambda}}\sum_{k=g_i}^{g_{i+1}-1}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_2 + 4\widehat{\beta}\sqrt{d\iota}\sqrt{\sum_{k=g_i}^{g_{i+1}}\sigma_k^2 + w\alpha^2} + 4d\iota\gamma^2\widehat{\beta} + 4d\iota\,, \tag{C.9}$$

where we use $\text{Regret}([g_i, g_{i+1}])$ to denote the regret accumulated in the time period $[g_i, g_{i+1}]$.

Finally, without loss of generality, we assume $K\%w = 0$. Then we have

$$\begin{aligned}
\text{Regret}(\widetilde{K}) &= \sum_{i=0}^{\frac{K}{w}-1}\text{Regret}([g_i, g_{i+1}]) \\
&\leq \frac{2A^2 w^{\frac{3}{2}}}{\alpha}\sqrt{\frac{d}{\lambda}}\sum_{i=0}^{\frac{K}{w}-1}\sum_{k=g_i}^{g_{i+1}-1}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_2 + 4\widehat{\beta}\sqrt{d\iota}\sum_{i=0}^{\frac{K}{w}-1}\sqrt{\sum_{k=g_i}^{g_{i+1}}\sigma_k^2 + w\alpha^2} + \frac{4d\iota\gamma^2\widehat{\beta}K}{w} + \frac{4dK\iota}{w} \\
&\leq \frac{2A^2 w^{\frac{3}{2}}}{\alpha}\sqrt{\frac{d}{\lambda}}\sum_{k=1}^{K-1}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_2 + 4\widehat{\beta}\sqrt{d\iota}\sqrt{\frac{K}{w}\sum_{i=0}^{\frac{K}{w}-1}(\sum_{k=g_i}^{g_{i+1}}\sigma_k^2 + w\alpha^2)} + \frac{4d\iota\gamma^2\widehat{\beta}K}{w} + \frac{4dK\iota}{w} \\
&\leq \frac{2A^2 w^{\frac{3}{2}}B_K}{\alpha}\sqrt{\frac{d}{\lambda}} + 4\widehat{\beta}\sqrt{\frac{Kd\iota}{w}}\sqrt{\sum_{k=1}^{K}\sigma_k^2 + K\alpha^2} + \frac{4d\iota\gamma^2\widehat{\beta}K}{w} + \frac{4dK\iota}{w}\,,
\end{aligned}$$

where in the second inequality we use Cauchy-Schwarz inequality, and the last inequality holds due to $\sum_{k\in[K-1]}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_2 \leq B_K$.

D Proof for Theorem 5.1

Recall that we call the restart time rounds *grids* and denote them by $g_1, g_2, \ldots g_{\lceil\frac{K}{w}\rceil-1}$, where $g_i\%w = 0$ for all $i \in [\lceil\frac{K}{w}\rceil - 1]$. Let i_k be the grid index of time round k, *i.e.*, $g_{i_k} \leq k < g_{i_k+1}$. We denote $\widehat{\Psi}_{k,\ell} := \{t : t \in [g_{i_k}, k-1], \ell_t = \ell\}$.

For simplicity of analysis, we first try to bound the regret over the first grid, *i.e.*, we try to analyze $\text{Regret}(\widetilde{K})$ for $\widetilde{K} \in [1, w]$. Note that in this case, for any $k \in [\widetilde{K}]$ with $\widetilde{K} \in [1, w]$, we have $g_{i_k} = 1$, so $\widehat{\Psi}_{k,\ell} := \{t : t \in [1, k-1], \ell_t = \ell\}$.

First, we calculate the estimation difference $|\boldsymbol{a}^\top(\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}_k)|$ for any $\mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}\|_2 \leq A$. Recall that by definition, $\widehat{\Sigma}_{k,\ell} = 2^{-2\ell}\boldsymbol{I} + \sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t^\top$, $\widehat{\boldsymbol{b}}_{k,\ell} = \sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2 r_t\mathbf{a}_t$, and

$$\widehat{\boldsymbol{\theta}}_{k,\ell} = \widehat{\Sigma}_{k,\ell}^{-1}\widehat{\boldsymbol{b}}_{k,\ell} = \widehat{\Sigma}_{k,\ell}^{-1}(\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2 r_t\mathbf{a}_t) = \widehat{\Sigma}_{k,\ell}^{-1}(\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t^\top\boldsymbol{\theta}_t + \sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\epsilon_t)\,.$$

Then we have

$$\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}_k = \widehat{\Sigma}_{k,\ell}^{-1}(\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_k) + \sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\epsilon_t) - 2^{-2\ell}\widehat{\Sigma}_{k,\ell}^{-1}\boldsymbol{\theta}_k\,. \tag{D.1}$$

Therefore, we can get

$$|\mathbf{a}^\top(\widehat{\boldsymbol{\theta}}_{k,\ell}-\boldsymbol{\theta}_k)| \le \left|\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}_k)\right| + \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}}\|\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\epsilon_t\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} + 2^{-2\ell}\|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}}\|\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-\frac{1}{2}}\boldsymbol{\theta}_k\|_2\,, \tag{D.2}$$

where we use the Cauchy-Schwarz inequality.

For the first term, we have that for any $k \in [1, w]$

$$\begin{aligned}
\left|\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}_k)\right| &\le \sum_{t\in\widehat{\Psi}_{k,\ell}} |\mathbf{a}^\top\boldsymbol{\Sigma}_{k,\ell}^{-1}w_t\mathbf{a}_t|\cdot|w_t\mathbf{a}_t^\top(\sum_{s=t}^{k-1}(\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1}))| \\
&\qquad \text{(triangle inequality)} \\
&\le \sum_{t\in\widehat{\Psi}_{k,\ell}} |\mathbf{a}^\top\boldsymbol{\Sigma}_{k,\ell}^{-1}w_t\mathbf{a}_t|\cdot\|w_t\mathbf{a}_t\|_2\cdot\|\sum_{s=t}^{k-1}(\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1})\|_2 \\
&\qquad \text{(Cauchy-Schwarz)} \\
&\le A\sum_{t\in\widehat{\Psi}_{k,\ell}} |\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}w_t\mathbf{a}_t|\cdot\|\sum_{s=t}^{k-1}(\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1})\|_2 \\
&\qquad (\|\mathbf{a}_t\|\le A,\, w_t=\tfrac{2^{-\ell_t}}{\|\mathbf{a}_t\|_{\widehat{\Sigma}_{t,\ell_t}^{-1}}}\le 1) \\
&\le A\sum_{s=1}^{k-1}\sum_{t\in\widehat{\Psi}_{k,\ell}} |\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}w_t\mathbf{a}_t|\cdot\|\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1}\|_2 \\
&\le A\sum_{s=1}^{k-1}\sqrt{\left[\sum_{t\in\widehat{\Psi}_{k,\ell}}\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}\mathbf{a}\right]\cdot\left[\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t\mathbf{a}_t{}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}w_t\mathbf{a}_t\right]}\cdot||\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1}||_2 \\
&\qquad \text{(Cauchy-Schwarz)} \\
&\le A\sum_{s=1}^{k-1}\sqrt{\left[\sum_{t\in\widehat{\Psi}_{k,\ell}}\mathbf{a}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}\mathbf{a}\right]\cdot d}\cdot||\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1}||_2 \qquad ((\star)) \\
&\le A\|\mathbf{a}\|_2\sqrt{d}\sum_{s=1}^{k-1}\sqrt{2^{2\ell}\sum_{t\in\widehat{\Psi}_{k,\ell}}1}\cdot||\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1}||_2 \\
&\qquad (\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1})\le\tfrac{1}{2^{-2\ell}}=2^{2\ell}) \\
&\le A^2 2^\ell\sqrt{dw}\sum_{s=1}^{k-1}||\boldsymbol{\theta}_s-\boldsymbol{\theta}_{s+1}||_2\,, \qquad (D.3)
\end{aligned}$$

where the inequality $(\star)$ follows from the fact that $\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t\mathbf{a}_t{}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}w_t\mathbf{a}_t \le d$ that can be proved as follows. We have $\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t\mathbf{a}_t{}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}w_t\mathbf{a}_t = \sum_{t\in\widehat{\Psi}_{k,\ell}}\text{tr}\left(w_t\mathbf{a}_t{}^\top\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}w_t\mathbf{a}_t\right) = \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t{}^\top\right)$. Given the eigenvalue decomposition $\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t{}^\top = \text{diag}(\lambda_1,\ldots,\lambda_d)^\top$, we have $\widehat{\boldsymbol{\Sigma}}_{k,\ell} = \text{diag}(\lambda_1+\lambda,\ldots,\lambda_d+\lambda)^\top$, and $\text{tr}\left(\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\mathbf{a}_t\mathbf{a}_t{}^\top\right) = \sum_{i=1}^d\frac{\lambda_j}{\lambda_j+\lambda}\le d$.

For the second term in Eq.(D.2), we can apply Theorem F.3 for the layer ℓ. In detail, for any $k\in[K]$, for each $t\in\widehat{\Psi}_{k,\ell}$, we have

$$\|w_t\mathbf{a}_t\|_{\widehat{\boldsymbol{\Sigma}}_{t,\ell}^{-1}} = 2^{-\ell},\quad \mathbb{E}[w_t^2\epsilon_t^2|\mathcal{F}_t]\le w_t^2\mathbb{E}[\epsilon_t^2|\mathcal{F}_t]\le w_t^2\sigma_t^2,\quad |w_t\epsilon_t|\le|\epsilon_t|\le R,$$

where the last inequality holds due to the fact that $w_t = \frac{2^{-\ell_t}}{\|\mathbf{a}_t\|_{\widehat{\mathbf{\Sigma}}_{t,\ell_t}^{-1}}} \leq 1$. According to Theorem F.3, and taking a union bound, we can deduce that with probability at least $1-\delta$, for all $\ell \in [L]$, for all round $k \in \Psi_{K+1,\ell}$,

$$\|\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2 \mathbf{a}_t \epsilon_t\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}} \leq 16\cdot 2^{-\ell}\sqrt{\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\sigma_t^2 \log(\frac{4w^2L}{\delta})} + 6\cdot 2^{-\ell} R\log(\frac{4w^2L}{\delta})\,. \tag{D.4}$$

For simplicity, we denote $\mathcal{E}_{\text{conf}}$ as the event such that Eq.(D.4) holds.

For the third term in Eq.(D.2), we have

$$2^{-2\ell}\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}}\|\widehat{\mathbf{\Sigma}}_{k,\ell}^{-\frac{1}{2}}\boldsymbol{\theta}_k\|_2 \leq 2^{-2\ell}\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}}\|\widehat{\mathbf{\Sigma}}_{k}^{-\frac{1}{2}}\|_2\|\boldsymbol{\theta}_k\|_2 \leq 2^{-2\ell}\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}}\frac{1}{\sqrt{\lambda_{\min}(\widehat{\mathbf{\Sigma}}_{k,\ell})}}\|\boldsymbol{\theta}_k\|_2 \leq 2^{-\ell}B\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k}^{-1}}\,, \tag{D.5}$$

where we use the fact that $\lambda_{\min}(\widehat{\mathbf{\Sigma}}_{k,\ell}) \geq 2^{-2\ell}$.

For simplicity, we denote $\ell^* = \lceil \frac{1}{2}\log_2\log\left(4(w+1)^2L/\delta\right)\rceil + 8$. Then, under $\mathcal{E}_{\text{conf}}$, by the definition of $\widehat{\beta}_{k,\ell}$ in Eq.(5.1), Lemma F.4 and Lemma F.5, with probability at least $1-\delta$, we have for all $\ell^*+1 \leq \ell \leq L$,

$$\widehat{\beta}_{k,\ell} \geq 16\cdot 2^{-\ell}\sqrt{\sum_{t\in\widehat{\Psi}_{k,\ell}} w_t^2\sigma_t^2\log(\frac{4w^2L}{\delta})} + 6\cdot 2^{-\ell}R\log(\frac{4w^2L}{\delta}) + 2^{-\ell}B. \tag{D.6}$$

Therefore, with Eq.(D.2), Eq.(D.3), Eq.(D.4), Eq.(D.5), Eq.(D.6), with probability at least $1-3\delta$, for all $\ell^*+1 \leq \ell \leq L$ we have

$$|\mathbf{a}^\top(\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}_k)| \leq A^2 2^\ell\sqrt{dw}\sum_{s=1}^{k-1}||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 + \widehat{\beta}_{k,\ell}\|\mathbf{a}\|_{\widehat{\Sigma}_{k,\ell}^{-1}}\,. \tag{D.7}$$

Then for all $k \in [K]$ such that $\ell^*+1 \leq \ell_k \leq L$, with probability at least $1-3\delta$ we have

$$\begin{aligned}
\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle &\leq \min_{\ell\in[L]}\langle \mathbf{a}_k^*, \widehat{\boldsymbol{\theta}}_{k,\ell}\rangle + A^2 2^\ell\sqrt{dw}\sum_{s=1}^{k-1}||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 + \widehat{\beta}_{k,\ell}\|\mathbf{a}_k^*\|_{\widehat{\Sigma}_{k,\ell}^{-1}} \\
&\leq A^2 2^L\sqrt{dw}\sum_{s=1}^{k-1}||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 + \min_{\ell\in[L]}\langle \mathbf{a}_k^*, \widehat{\boldsymbol{\theta}}_{k,\ell}\rangle + \widehat{\beta}_{k,\ell}\|\mathbf{a}_k^*\|_{\widehat{\Sigma}_{k,\ell}^{-1}} \\
&\leq A^2 2^L\sqrt{dw}\sum_{s=1}^{k-1}||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 + \min_{\ell\in[L]}\langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k,\ell}\rangle + \widehat{\beta}_{k,\ell}\|\mathbf{a}_k\|_{\widehat{\Sigma}_{k,\ell}^{-1}} \\
&\leq A^2 2^L\sqrt{dw}\sum_{s=1}^{k-1}||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 + \langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k,\ell_k-1}\rangle + \widehat{\beta}_{k,\ell_k-1}\|\mathbf{a}_k\|_{\widehat{\Sigma}_{k,\ell_k-1}^{-1}}\,,
\end{aligned} \tag{D.8}$$

where the first inequality holds because of Eq.(D.7), the third inequality holds because of the arm selection rule in Line 8 of Algo.2.

We decompose the regret for $\widetilde{K} \in [1,w]$ as follows

$$\begin{aligned}
\text{Regret}(\widetilde{K}) &= \sum_{k\in[\widetilde{K}]}\left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right) \\
&= \sum_{\ell\in[\ell^*]}\sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}}\left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right) + \sum_{\ell\in[L]\setminus[\ell^*]}\sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}}\left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right) \\
&\quad + \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,L+1}}\left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right).
\end{aligned} \tag{D.9}$$

We will bound the three terms separately. For the first term, we have for layer $\ell \in [\ell^*]$ and round $k \in \widehat{\Psi}_{\widetilde{K}+1,\ell}$, we have

$$\begin{aligned}
\sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}^*\rangle\right) &\le 2\left|\Psi_{K+1,\ell}\right| \\
&= 2^{2\ell+1} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \|w_k \mathbf{a}_k\|^2_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,\ell}} \\
&\le 2\cdot 128^2 \log(\frac{4(w+1)^2 L}{\delta}) \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \|w_k \mathbf{a}_k\|^2_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,\ell}} \\
&\le 2\cdot 128^2 \log(\frac{4(w+1)^2 L}{\delta}) \cdot 2d\log(1+\frac{2^{2\ell} w A^2}{d}) \\
&= \widetilde{O}(d)\,, \qquad \text{(D.10)}
\end{aligned}$$

where the first inequality holds because the reward is in $[-1,1]$, the equation follows from the fact that $\|w_k \mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,\ell}} = 2^{-\ell}$ holds for all $k \in \Psi_{K+1,\ell}$, the second inequality holds due to the fact that $2^{\ell^*} \le 128\sqrt{\log(4(w+1)^2 L/\delta)}$, and the last inequality holds due to Lemma F.2.

Therefore

$$\sum_{\ell\in[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right) = \widetilde{O}(d)\,. \qquad \text{(D.11)}$$

For the second part in Eq.(D.9), we have

$$\begin{aligned}
&\sum_{\ell\in[L]\setminus[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle\right) \\
&\le \sum_{\ell\in[L]\setminus[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \Bigg(\langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k,\ell-1}\rangle + \widehat{\beta}_{k,\ell-1}\|\mathbf{a}_k\|_{\widehat{\Sigma}^{-1}_{k,\ell-1}} \\
&\quad + A^2 2^L \sqrt{dw} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 - \langle \mathbf{a}_k, \boldsymbol{\theta}_k\rangle \Bigg) \\
&\le 2 \sum_{\ell\in[L]\setminus[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \widehat{\beta}_{k,\ell-1}\|\mathbf{a}_k\|_{\widehat{\Sigma}^{-1}_{k,\ell-1}} + A^2\sqrt{dw} \sum_{\ell\in[L]\setminus[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} 2^L \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2\,, \qquad \text{(D.12)}
\end{aligned}$$

where the inequality holds due to Eq.(D.8), the second inequality holds due to Eq.(D.7). We then try to bound the two terms.

For the first term in Eq.(D.12), we have

$$\begin{aligned}
\sum_{\ell\in[L]\setminus[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \widehat{\beta}_{k,\ell-1}\|\mathbf{a}_k\|_{\widehat{\Sigma}^{-1}_{k,\ell-1}} &\le \sum_{\ell\in[L]\setminus[\ell^*]} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \widehat{\beta}_{k,\ell-1}\cdot 2^{-\ell} \\
&\le \sum_{\ell\in[L]\setminus[\ell^*]} \widehat{\beta}_{\widetilde{K},\ell-1}\cdot 2^{-\ell} \left|\widehat{\Psi}_{\widetilde{K}+1,\ell}\right| \\
&= \sum_{\ell\in[L]\setminus[\ell^*]} \widehat{\beta}_{\widetilde{K},\ell-1}\cdot 2^{\ell} \sum_{k\in\widehat{\Psi}_{\widetilde{K}+1,\ell}} \|w_k \mathbf{a}_k\|^2_{\Sigma^{-1}_{k,\ell}} \\
&\le \sum_{\ell\in[L]\setminus[\ell^*]} \widehat{\beta}_{\widetilde{K},\ell-1}\cdot 2^{\ell}\cdot 2d\log(1+\frac{2^{2\ell}\widetilde{K}A^2}{d}) \\
&= \widetilde{O}(d\cdot 2^{\ell}\cdot \widehat{\beta}_{\widetilde{K},\ell-1})
\end{aligned}$$

$$= \widetilde{O}\Big(d(\sqrt{\sum_{k=1}^{\widetilde{K}} \sigma_k^2} + R + 1)\Big), \tag{D.13}$$

where the first inequality holds because by the algorithm design, we have for all $k \in \widehat{\Psi}_{\widetilde{K}+1,\ell}$: $\|\mathbf{a}_k\|_{\widehat{\Sigma}_{k,\ell-1}^{-1}} \leq 2^{-\ell}$; the second inequality holds because for all $k \in \widehat{\Psi}_{\widetilde{K}+1,\ell}$, $\widehat{\beta}_{k,\ell-1} \leq \widehat{\beta}_{\widetilde{K},\ell-1}$; the first equality holds because for all $k \in \widehat{\Psi}_{\widetilde{K}+1,\ell}$, $\|w_k \mathbf{a}_k\|^2_{\Sigma_{k,\ell}^{-1}} = 2^{-2\ell}$; the third inequality holds by Lemma F.2; the last two equalities hold because by Lemma F.4 and Lemma F.5, we have $\widehat{\beta}_{\widetilde{K},\ell-1} = \widetilde{O}\Big(2^{-\ell}(\sqrt{\sum_{k=1}^{\widetilde{K}} \sigma_k^2} + R + 1)\Big)$.

For the second term in Eq.(D.12), we have

$$\begin{aligned} A^2\sqrt{dw} \sum_{\ell \in [L]\setminus[\ell^*]} \sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,\ell}} 2^L \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 &\leq A^2 2^L \sqrt{dw} \sum_{k \in [\widetilde{K}-1]} \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 \\ &\leq \frac{A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha} \sum_{k=1}^{\widetilde{K}-1} ||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 \end{aligned} \tag{D.14}$$

Therefore, with this, Eq.(D.12), and Eq.(D.13), we have

$$\sum_{\ell \in [L]\setminus[\ell^*]} \sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,\ell}} (\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k \rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k \rangle) \leq \frac{A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha} \sum_{k=1}^{\widetilde{K}-1} ||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \widetilde{O}\Big(d(\sqrt{\sum_{k=1}^{\widetilde{K}} \sigma_k^2} + R + 1)\Big). \tag{D.15}$$

Finally, for the last term in Eq.(D.9), we have

$$\begin{aligned} &\sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,L+1}} (\langle \mathbf{a}_k^*, \boldsymbol{\theta}_k \rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}_k \rangle) \\ &\leq \sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,L+1}} \Big(\langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k,L} \rangle + \widehat{\beta}_{k,L} \|\mathbf{a}_k\|_{\widehat{\Sigma}_{k,L}^{-1}} \\ &\quad + A^2 2^L \sqrt{dw} \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 - \langle \mathbf{a}_k, \boldsymbol{\theta}_k \rangle \Big) \\ &\leq \sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,L+1}} \Big(2\widehat{\beta}_{k,L} \|\mathbf{a}_k\|_{\widehat{\Sigma}_{k,L}^{-1}} + A^2 2^{L+1} \sqrt{dw} \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 \Big) \\ &\leq \sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,L+1}} \Big(2^{-L+1} \widehat{\beta}_{k,L} + A^2 2^{L+1} \sqrt{dw} \sum_{s=1}^{k-1} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}||_2 \Big) \\ &\leq \frac{2A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha} \sum_{k=1}^{\widetilde{K}-1} ||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \sum_{k \in \widehat{\Psi}_{\widetilde{K}+1,L+1}} 2^{-L+1} \widehat{\beta}_{\widetilde{K},L} \\ &\leq \frac{2A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha} \sum_{k=1}^{\widetilde{K}-1} ||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + w \cdot 2\alpha \cdot \widehat{\beta}_{\widetilde{K},L} \\ &= \frac{2A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha} \sum_{k=1}^{\widetilde{K}-1} ||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \widetilde{O}\Big(w\alpha^2 \cdot (\sqrt{\sum_{k=1}^{\widetilde{K}} \sigma_k^2} + R + 1) \Big), \end{aligned} \tag{D.16}$$

where the first inequality holds due to Eq.(D.8), the second inequality holds due to Eq.(D.7), the third inequality holds because by the algorithm design, we have for all $k \in \widehat{\Psi}_{\widetilde{K}+1,L+1}$: $\|\mathbf{a}_k\|_{\widehat{\Sigma}_{k,L}^{-1}} \leq 2^{-L}$,

the fourth inequality holds due to the same reasons as before, and the fact that $\widehat{\beta}_{\widetilde{K},L} \geq \widehat{\beta}_{k,L}$ for all $k \in \widehat{\beta}_{\widetilde{K},L}$; the last inequality holds due to $\widehat{\beta}_{\widetilde{K},\ell-1} = \widetilde{O}\Big(\alpha(\sqrt{\sum_{k=1}^{\widetilde{K}} \sigma_k^2} + R + 1)\Big)$.

Plugging Eq.(D.15), Eq.(D.16), and Eq.(D.11) into Eq.(D.9), we can get that for $\widetilde{K} \in [1, w]$

$$\text{Regret}(\widetilde{K}) = \widetilde{O}\Big(\frac{A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha}\sum_{k=1}^{\widetilde{K}-1}||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \big(w\alpha^2 + d\big)\cdot\big(\sqrt{\sum_{k=1}^{\widetilde{K}}\sigma_k^2} + R + 1\big)\Big). \quad \text{(D.17)}$$

By the same deduction we can get

$$\text{Regret}([g_i, g_{i+1}]) = \widetilde{O}\Big(\frac{A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha}\sum_{k=g_i}^{g_{i+1}}||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \big(w\alpha^2 + d\big)\cdot\big(\sqrt{\sum_{k=g_i}^{g_{i+1}}\sigma_k^2} + R + 1\big)\Big). \quad \text{(D.18)}$$

Finally, without loss of generality, we assume $K\%w = 0$. Then we have

$$\begin{aligned}
\text{Regret}(K) &= \sum_{i=0}^{\frac{K}{w}-1}\text{Regret}([g_i, g_{i+1}]) \\
&= \widetilde{O}\Big(\frac{A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha}\sum_{i=0}^{\frac{K}{w}-1}\sum_{k=g_i}^{g_{i+1}}||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \big(w\alpha^2 + d\big)\cdot\sum_{i=0}^{\frac{K}{w}-1}\big(\sqrt{\sum_{k=g_i}^{g_{i+1}}\sigma_k^2} + R + 1\big)\Big) \\
&\leq \widetilde{O}\Big(\frac{A^2\sqrt{d}w^{\frac{3}{2}}}{\alpha}\sum_{k=1}^{K-1}||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 + \big(w\alpha^2 + d\big)\cdot\big(\sqrt{\frac{K}{w}\sum_{i=0}^{\frac{K}{w}-1}\sum_{k=g_i}^{g_{i+1}}\sigma_k^2} + \frac{KR}{w} + \frac{K}{w}\big)\Big) \\
&\leq \widetilde{O}\Big(\frac{A^2\sqrt{d}w^{\frac{3}{2}}B_K}{\alpha} + \big(w\alpha^2 + d\big)\cdot\sqrt{\frac{K}{w}\sum_{k=1}^{K}\sigma_k^2} + \big(1 + R\big)\cdot\big(K\alpha^2 + \frac{Kd}{w}\big)\Big),
\end{aligned}$$

where the first inequality holds due to the Cauchy-Schwarz inequality, the last inequality holds because $\sum_{k=1}^{K-1}||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}||_2 \leq B_K$.

E Proof of Theorem A.1

With the candidate pool set $\mathcal{P}$ designed as in Eq.(A.1), Eq.(A.2), Eq.(A.3), and $H = \lceil d^{\frac{2}{5}}K^{\frac{2}{5}}\rceil$, we have $|\mathcal{P}| = O(\log K)$, and for any $w \in \mathcal{W}$, $w \leq H$.

We denote the optimal (w, α) with the knowledge of V_K and B_K in Corollary 5.3 as (w^*, α^*). We denote the best approximation of (w^*, α^*) in the candidate set $\mathcal{P}$ as (w^+, α^+). Then we can decompose the regret as follows

$$\begin{aligned}
\text{Regret}(K) = \sum_{k=1}^{K}\langle\mathbf{a}_t^*, \boldsymbol{\theta}_k\rangle - \langle\mathbf{a}_t, \boldsymbol{\theta}_k\rangle &= \underbrace{\sum_{k=1}^{K}\langle\mathbf{a}_t^*, \boldsymbol{\theta}_k\rangle - \sum_{i=1}^{\lceil\frac{K}{H}\rceil}\sum_{k=(i-1)H+1}^{iH}\langle\mathbf{a}_t(w^+, \alpha^+), \boldsymbol{\theta}_k\rangle}_{(1)} \\
&\quad + \underbrace{\sum_{i=1}^{\lceil\frac{K}{H}\rceil}\sum_{k=(i-1)H+1}^{iH}\langle\mathbf{a}_t(w^+, \alpha^+), \boldsymbol{\theta}_k\rangle - \langle\mathbf{a}_t(w_i, \alpha_i), \boldsymbol{\theta}_k\rangle}_{(2)}.
\end{aligned} \quad \text{(E.1)}$$

The first term (1) is the dynamic regret of Restarted SAVE$^+$ with the best parameters in the candidate pool $\mathcal{P}$. The second term (2) is the regret overhead of meta-algorithm due to adaptive exploration of unknown optimal parameters.

By the design of the candidate pool set $\mathcal{P}$ in Eq.(A.1), Eq.(A.2), Eq.(A.3), we have that there exists a pair $(w^+, \alpha^+) \in \mathcal{P}$ such that $w^+ < w^* < 2w^+$, and $\alpha^+ < \alpha^* < 2\alpha^+$. Therefore, employing the regret bound in Theorem 5.1, we can get

$$
\begin{aligned}
(1) &\leq \sum_{i=1}^{\lceil \frac{K}{H} \rceil} \widetilde{O}(\sqrt{d} w^{+1.5} B_i/\alpha^+ + \alpha^{+2}(H + \sqrt{w^+ H V_i}) + d\sqrt{H V_i / w^+} + dH/w^+) \\
&\leq \widetilde{O}(\sqrt{d} w^{+1.5} B_K/\alpha^+ + \alpha^{+2}(K + \sqrt{w^+ H \frac{K}{H} \sum_{i=1}^{\lceil \frac{K}{H} \rceil} V_i}) + d\sqrt{H \frac{K}{H} \sum_{i=1}^{\lceil \frac{K}{H} \rceil} V_i / w^+} + dK/w^+) \\
&= \widetilde{O}(\sqrt{d} w^{+1.5} B_K/\alpha^+ + \alpha^{+2}(K + \sqrt{w^+ K V_K}) + d\sqrt{K V_K / w^+} + dK/w^+) \\
&= \widetilde{O}(\sqrt{d} w^{*1.5} B_K/\alpha^* + \alpha^{*2}(K + \sqrt{w^* K V_K}) + d\sqrt{K V_K / w^*} + dK/w^*) \\
&= \widetilde{O}(d^{4/5} V_K^{2/5} B_K^{1/5} K^{2/5} + d^{2/3} B_K^{1/3} K^{2/3}) \,,
\end{aligned} \tag{E.2}
$$

where we denote B_i as the total variation budget in block i, V_i is the total variance in block i, the second inequality is by Cauchy–Schwarz inequality, the first equality holds due to $\sum_{i=1}^{\lceil \frac{K}{H} \rceil} B_i = B_K$, $\sum_{i=1}^{\lceil \frac{K}{H} \rceil} V_i = V_K$, the second equality holds due to $w^+ < w^* < 2w^+$ and $\alpha^+ < \alpha^* < 2\alpha^+$, the last equality holds by Corollary 5.3.

We then try to bound the second term (2). We denote by $\mathcal{E}$ the event such that Lemma F.7 holds, and denote by $R_i := \sum_{k=(i-1)H+1}^{iH} \langle \mathbf{a}_t(w^+, \alpha^+), \boldsymbol{\theta}_k \rangle - \langle \mathbf{a}_t(w_i, \alpha_i), \boldsymbol{\theta}_k \rangle$ the instantaneous regret of the meta learner in the block i. Then we have

$$
\begin{aligned}
(2) &= \mathbb{E}\left[\sum_{i=1}^{\lceil \frac{K}{H} \rceil} R_i \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\lceil \frac{K}{H} \rceil} R_i | \mathcal{E} \right] P(\mathcal{E}) + \mathbb{E}\left[\sum_{i=1}^{\lceil \frac{K}{H} \rceil} R_i | \overline{\mathcal{E}} \right] P(\overline{\mathcal{E}}) \\
&\leq \widetilde{O}\left(L_{\max} \sqrt{\frac{K}{H} |\mathcal{P}|} \right) \cdot (1 - \frac{2}{K}) + \widetilde{O}(K) \cdot \frac{2}{K} \\
&= \widetilde{O}(\sqrt{H |\mathcal{P}| K}) \\
&= \widetilde{O}(d^{\frac{1}{5}} K^{\frac{7}{10}}) \,,
\end{aligned} \tag{E.3}
$$

where $L_{\max} := \max_{i \in [\lceil \frac{K}{H} \rceil]} L_i$, the first inequality holds due to the standard regret upper bound result for Exp3 [4], the third equality holds due to Lemma F.7, the last equality holds since $H = \lceil d^{\frac{2}{5}} K^{\frac{2}{5}} \rceil$, and $|\mathcal{P}| = O(\log K)$.

Finally, combining the above results for term (1) and term (2), we have

$$
\text{Regret}(K) = \widetilde{O}(d^{4/5} V_K^{2/5} B_K^{1/5} K^{2/5} + d^{2/3} B_K^{1/3} K^{2/3} + d^{\frac{1}{5}} K^{\frac{7}{10}}). \tag{E.4}
$$

F Technical Lemmas

Theorem F.1 (Theorem 4.3, [45])**.** Let $\{\mathcal{G}_k\}_{k=1}^{\infty}$ be a filtration, and $\{\mathbf{x}_k, \eta_k\}_{k \geq 1}$ be a stochastic process such that $\mathbf{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \lambda, \epsilon > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \geq 1$, let $y_k = \langle \boldsymbol{\mu}^*, \mathbf{x}_k \rangle + \eta_k$ and suppose that $\eta_k, \mathbf{x}_k$ also satisfy

$$
\mathbb{E}[\eta_k | \mathcal{G}_k] = 0, \ \mathbb{E}[\eta_k^2 | \mathcal{G}_k] \leq \sigma^2, \ |\eta_k| \leq R, \ \|\mathbf{x}_k\|_2 \leq L. \tag{F.1}
$$

For $k \geq 1$, let $\mathbf{Z}_k = \lambda \mathbf{I} + \sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^\top$, $\mathbf{b}_k = \sum_{i=1}^{k} y_i \mathbf{x}_i$, $\boldsymbol{\mu}_k = \mathbf{Z}_k^{-1} \mathbf{b}_k$, and

$\beta_k = 12\sqrt{\sigma^2 d \log(1 + kL^2/(d\lambda)) \log(32(\log(R/\epsilon) + 1)k^2/\delta)}$

$$+ 24\log(32(\log(R/\epsilon)+1)k^2/\delta)\max_{1\le i\le k}\{|\eta_i|\min\{1,\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\}\} + 6\log(32(\log(R/\epsilon)+1)k^2/\delta)\epsilon.$$

Then, for any $0<\delta<1$, we have with probability at least $1-\delta$ that,

$$\forall k\ge 1,\ \big\|\textstyle\sum_{i=1}^k \mathbf{x}_i\eta_i\big\|_{\mathbf{Z}_k^{-1}} \le \beta_k,\ \|\boldsymbol{\mu}_k-\boldsymbol{\mu}^*\|_{\mathbf{Z}_k} \le \beta_k + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2.$$

Lemma F.2 (Lemma 11, [2])**.** For any $\lambda>0$ and sequence $\{\mathbf{x}_k\}_{k=1}^K \subset \mathbb{R}^d$ for $k\in[K]$, define $\mathbf{Z}_k = \lambda\mathbf{I}+\sum_{i=1}^{k-1}\mathbf{x}_i\mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_k\|_2\le L$ holds for all $k\in[K]$, we have

$$\sum_{k=1}^K \min\big\{1,\|\mathbf{x}_k\|^2_{\mathbf{Z}_k^{-1}}\big\} \le 2d\log\big(1+KL^2/(d\lambda)\big).$$

Theorem F.3 (Theorem 2.1, [41])**.** Let $\{\mathcal{G}_k\}_{k=1}^\infty$ be a filtration, and $\{\mathbf{x}_k,\eta_k\}_{k\ge1}$ be a stochastic process such that $\mathbf{x}_k\in\mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k\in\mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L,\sigma,\lambda,\epsilon>0$, $\boldsymbol{\mu}^*\in\mathbb{R}^d$. For $k\ge1$, let $y_k=\langle\boldsymbol{\mu}^*,\mathbf{x}_k\rangle+\eta_k$, where $\eta_k,\mathbf{x}_k$ satisfy

$$\mathbb{E}[\eta_k|\mathcal{G}_k]=0,\ |\eta_k|\le R,\ \sum_{i=1}^k\mathbb{E}[\eta_i^2|\mathcal{G}_i]\le v_k,\ \text{ for }\forall\ k\ge1$$

For $k\ge1$, let $\mathbf{Z}_k=\lambda\mathbf{I}+\sum_{i=1}^k\mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_k=\sum_{i=1}^k y_i\mathbf{x}_i$, $\boldsymbol{\mu}_k=\mathbf{Z}_k^{-1}\mathbf{b}_k$, and

$$\beta_k = 16\rho\sqrt{v_k\log(4w^2/\delta)}+6\rho R\log(4w^2/\delta),$$

where $\rho\ge\sup_{k\ge1}\|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}$. Then, for any $0<\delta<1$, we have with probability at least $1-\delta$ that,

$$\forall k\ge1,\ \big\|\textstyle\sum_{i=1}^k\mathbf{x}_i\eta_i\big\|_{\mathbf{Z}_k^{-1}}\le\beta_k, \|\boldsymbol{\mu}_k-\boldsymbol{\mu}^*\|_{\mathbf{Z}_k}\le\beta_k+\sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2.$$

Lemma F.4 (Adopted from Lemma B.4, [41])**.** Let weight w_i be defined in Algorithm 2. With probability at least $1-2\delta$, for all $k\ge1$, $\ell\in[L]$, the following two inequalities hold simultaneously:

$$\sum_{i\in\widehat{\Psi}_{k+1,\ell}} w_i^2\sigma_i^2 \le 2\sum_{i\in\widehat{\Psi}_{k+1,\ell}} w_i^2\epsilon_i^2 + \frac{14}{3}R^2\log(4w^2L/\delta),$$
$$\sum_{i\in\widehat{\Psi}_{k+1,\ell}} w_i^2\epsilon_i^2 \le \frac{3}{2}\sum_{i\in\widehat{\Psi}_{k+1,\ell}} w_i^2\sigma_i^2 + \frac{7}{3}R^2\log(4w^2L/\delta).$$

For simplicity, we denote $\mathcal{E}_{\mathbb{V}}$ as the event such that the two inequalities in Lemma F.4 holds.

Lemma F.5 (Adopted from Lemma B.5, [41])**.** Suppose that $\|\boldsymbol{\theta}^*\|_2\le B$. Let weight w_i be defined in Algorithm 2. On the event $\mathcal{E}_{\text{conf}}$ and $\mathcal{E}_{\mathbb{V}}$ (defined in Eq.(D.4), Lemma F.4), for all $k\ge1$, $\ell\in[L]$ such that $2^\ell\ge64\sqrt{\log(4(w+1)^2L/\delta)}$, we have the following inequalities:

$$\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 \le 8\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i-\langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 + 6R^2\log(4(w+1)^2L/\delta) + 2^{-2\ell+2}B^2,$$
$$\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i-\langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 \le \frac{3}{2}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 + \frac{7}{3}R^2\log(4w^2L/\delta) + 2^{-2\ell}B^2.$$

Lemma F.6 ([18])**.** Let $M,v>0$ be fixed constants. Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a filtration so that for all $i\in[n]$, x_i is $\mathcal{G}_i$-measurable, while almost surely

$$\mathbb{E}\left[x_i|\mathcal{G}_{i-1}\right]=0,\quad |x_i|\le M,\quad \sum_{i=1}^n\mathbb{E}[x_i^2|\mathcal{G}_{i-1}]\le v.$$

Then for any $\delta>0$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^n x_i \le \sqrt{2v\log(1/\delta)}+2/3\cdot M\log(1/\delta).$$

Lemma F.7. Let $N = \lceil \frac{K}{H} \rceil$. Denote by L_i the absolute value of cumulative rewards for episode i, i.e., $L_i = \sum_{k=(i-1)H+1}^{iH} r_k$, then

$$\mathbb{P}\left[\forall i \in [N], L_i \leq H + R\sqrt{\frac{H}{2}\log\left(K(\frac{K}{H}+1)\right)} + \frac{2}{3}\cdot R\log\left(K(\frac{K}{H}+1)\right)\right] \geq 1 - \frac{1}{K}. \quad \text{(F.2)}$$

Proof. By Lemma F.6, we have that with probability at least $1 - 1/K$

$$\begin{aligned}
\sum_{k=(i-1)\cdot H+1}^{i\cdot H} \epsilon_i &\leq \sqrt{2\sum_{k=(i-1)\cdot H+1}^{i\cdot H} \sigma_k^2 \log(NK)} + 2/3 \cdot R\log(NK) \\
&\leq \sqrt{2H\frac{R^2}{4}\log(NK)} + 2/3\cdot R\log(NK) \\
&\leq R\sqrt{\frac{H}{2}\log\left(K\cdot(\frac{K}{H}+1)\right)} + \frac{2}{3}\cdot R\log\left(K\cdot(\frac{K}{H}+1)\right), \qquad \text{(F.3)}
\end{aligned}$$

where we use union bound, and in the second inequality we use the fact that since $|\epsilon_k| \leq R$, we have $\sigma_k^2 \leq \frac{R^2}{4}$. Finally, together with the assumption that $r_k \leq 1$ for all $k \in [K]$, we complete the proof. □