Noise-Adaptive Confidence Sets for Linear Bandits and Application to Bayesian Optimization

Kwang-Sung Jun [1] **Jungtaek Kim** [2]

Abstract

Adapting to a priori unknown noise level is a very important but challenging problem in sequential decision-making as efficient exploration typically requires knowledge of the noise level, which is often loosely specified. We report significant progress in addressing this issue for linear bandits in two respects. First, we propose a novel confidence set that is 'semi-adaptive' to the unknown sub-Gaussian parameter σ_*^2 in the sense that the (normalized) confidence width scales with $\sqrt{d\sigma_*^2 + \sigma_0^2}$ where d is the dimension and σ_0^2 is the specified sub-Gaussian parameter (known) that can be much larger than σ_*^2. This is a significant improvement over $\sqrt{d\sigma_0^2}$ of the standard confidence set of Abbasi-Yadkori et al. (2011), especially when d is large. We show that this leads to an improved regret bound in linear bandits. Second, for bounded rewards, we propose a novel variance-adaptive confidence set that has much improved numerical performance upon prior art. We then apply this confidence set to develop, as we claim, the first practical variance-adaptive linear bandit algorithm via an optimistic approach, which is enabled by our novel regret analysis technique. Both of our confidence sets rely critically on 'regret equality' from online learning. Our empirical evaluation in diverse Bayesian optimization tasks shows that our proposed algorithms demonstrate better or comparable performance compared to existing methods.

1. Introduction

In linear bandits (Abe & Long, 1999; Auer, 2002; Dani et al., 2008; Abbasi-Yadkori et al., 2011), the learner faces the challenge of making judicious sequential decisions with observed features so that the rewards obtained from those decisions would be maximized.[1] Specifically, at each time step $t = 1, 2, \ldots, n$, the learner is given an arm set $\mathcal{X}_t \subset \mathbb{R}^d$ and chooses an arm $x_t \in \mathcal{X}_t$ to observe a reward

$$y_t = \langle x_t, \theta^* \rangle + \eta_t, \tag{1}$$

where $\theta^* \in \mathbb{R}^d$ is unknown, $\eta_t \mid \mathcal{F}_{t-1}$ is a zero-mean stochastic noise, and $\mathcal{F}_{t-1} := \sigma(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$. The goal is to maximize the cumulative reward $\sum_{t=1}^n y_t$. The standard theoretical performance measure is cumulative pseudo-regret (or simply 'regret') defined as

$$\mathrm{Reg}_n := \sum_{t=1}^n \langle x_{t,*}, \theta^* \rangle - \langle x_t, \theta^* \rangle, \tag{2}$$

where $x_{t,*} = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle$ is the arm with the highest mean reward. The significance of this problem lies in its wide applicability, ranging from recommendation systems (Li et al., 2010) to online advertising (Avadhanula et al., 2021) where the learner must balance between exploration (acquiring information about the model) and exploitation (pulling the arm estimated to give high reward).

A critical aspect of the linear bandit problem is the fact that an algorithm requires prior knowledge of the noise level (or an upper bound of it) and that its performance critically depends on the *specified* noise level rather than the *actual* noise level. In this paper, we make significant progress in addressing this issue by proposing two algorithms that can semi or fully adapt to the actual noise level in two different setups.

The first setup is when the noise $\eta_t \mid \mathcal{F}_{t-1}$ is σ_*^2-sub-Gaussian where the sub-Gaussian parameter specified to the algorithm is σ_0^2 that can be much larger than σ_*^2. We propose a novel linear bandit algorithm called **LOSAN** (Linear Optimism with Semi-Adaptivity to Noise). The critical ingredient for this algorithm is a novel confidence set whose (normalized) confidence width contains online variance estimators and is no larger than $\tilde{\mathcal{O}}(\sqrt{d\sigma_*^2 + \sigma_0^2})$ with high probability where $\tilde{\mathcal{O}}$ hides polylogarithmic factors. This is no worse than $\tilde{\mathcal{O}}(\sqrt{d\sigma_0^2 + \sigma_0^2})$ of the standard confidence

[1]University of Arizona [2]University of Pittsburgh. Correspondence to: Kwang-Sung Jun <kjun@cs.arizona.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]The implementation of our proposed methods is available at https://github.com/jungtaekkim/LOSAN-LOFAV.

set of Abbasi-Yadkori et al. (2011), and in fact our confidence set can be significantly smaller when σ_0^2 is largely over-specified. LOSAN leverages our novel confidence set to perform optimistic exploration, which we prove to have a regret bound of $\tilde{\mathcal{O}}(\sigma_0\sqrt{dn} + \sigma_* d\sqrt{n})$. This bound, compared to the state-of-the-art bound of $\tilde{\mathcal{O}}(\sigma_0 d\sqrt{n})$, effectively achieves a factor of $\sqrt{d}$ improvement when $\sigma_*^2 \ll \sigma_0^2$. We present our confidence set and LOSAN in Section 2.

The second setup is the case of bounded noise; i.e., $\forall t, \eta_t \in [-R, R]$ almost surely for some $R \geq 0$. Among the recent studies reporting regret bounds that adapt to the unknown variance σ_t^2 of the noise η_t (Zhang et al., 2021; Kim et al., 2022; Xu et al., 2023; Zhao et al., 2023a), the seminal work of Zhao et al. (2023a) proposed an algorithm called SAVE and proved a regret bound of $\tilde{\mathcal{O}}(d\sqrt{R^2 + \sum_{t=1}^n \sigma_t^2})$, which is unimprovable in general as it matches the optimal worst-case regret (Lattimore & Szepesvári, 2020). Note that this rate is orderwise never worse than existing regret bounds with the sub-Gaussian assumption like OFUL (Abbasi-Yadkori et al., 2011) that achieves $d\sqrt{R^2 n}$.

However, SAVE follows the SupLinRel (Auer, 2002) style whose numerical performance is bad compared to other approaches such as optimistic (Dani et al., 2008) or posterior sampling (Agrawal & Goyal, 2013) strategies since it builds confidence bounds based on a small subset of samples only. Furthermore, their algorithm adapts to the variance that is unknown yet *fixed deterministically* before the bandit game starts. This severely limits its applicability. For example, such a setting cannot incorporate the environment where the variance at time t changes as a function of the chosen arm at time t or the past behavior of the learner.

To overcome these limitations, we propose a novel confidence set that not only removes the limiting assumption on the noise discussed above but also significantly improves the numerical performance. We then propose a novel optimistic-style linear bandit algorithm called **LOFAV** (Linear Optimism with Full Adaptivity to Variance) that computes the confidence set with all the available samples rather than a small subset of them. Our analysis shows that LOFAV enjoys the same order of regret as SAVE, which is optimal up to logarithmic factors. LOFAV can be implemented with the computational complexity of $\tilde{\mathcal{O}}((d^2 \max_{t=1}^n |\mathcal{X}_t|)n)$, which is the same as SAVE and is within a logarithmic factor of that of the standard algorithm OFUL (Abbasi-Yadkori et al., 2011). We present our confidence set, LOFAV, and their analysis in Section 3.

Finally, we empirically validate the performance of both of our algorithms in synthetic experiments and the benchmarks widely used in the Bayesian optimization community including NATS-Bench (Dong et al., 2021). In this empirical analysis, our algorithms yield better or comparable performance compared to OFUL and potentially the simple discrete Bayesian optimization strategy (Garrido-Merchán & Hernández-Lobato, 2020). We present our empirical results in Section 4, discuss related work in Section 5, and conclude our paper with exciting future research directions in Section 6.

Preliminaries. Throughout, we assume that both the confidence set and bandit algorithms have prior knowledge of S such that $\|\theta^*\| \leq S$. Note that techniques employed in Gales et al. (2022) can be readily applied to remove this assumption with only a constant factor inflation in the leading term of $\sqrt{n}$ in the regret bound and a polynomial factor in the lower order term. We assume that $\mathcal{X}_t \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. For a vector x, we define $\|x\| := \|x\|_2$ as the euclidean norm. We define $\|x\|_V := \sqrt{x^\top V x}$ where V is a positive semi-definite matrix. We use $a \lesssim b$ to denote that there exists an absolute constant $c > 0$ such that $a \leq cb$. We use $\tilde{\mathcal{O}}$ as the big-O notation that omits polylogarithmic factors. Let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. Define $\overline{\ln\ln}(x) := \ln\ln(e \vee x)$ and $[a..b] := \{a, a+1, \ldots, b\}$.

2. Semi-Adaptation for Sub-Gaussian Noise

In this section, we assume the standard linear bandit setup with the reward model of (1) where the conditional noise $\eta_t \mid \mathcal{F}_{t-1}$ satisfies the following assumption, which is standard in linear bandits.

Assumption A1. The noise $\eta_t \mid \mathcal{F}_{t-1}$ is σ_*^2-sub-Gaussian, $\forall t$. The algorithm has a prior knowledge of σ_0^2 such that $\sigma_*^2 \leq \sigma_0^2$.

We first introduce our confidence set. Our intuition comes from the standard FTRL (Follow-The-Regularized-Leader) regret equality (Orabona, 2023, Lemma 7.1) (restated in Lemma C.7). For the case of the squared loss $\ell(\theta) = \frac{1}{2}(x_t^\top \theta - y_t)^2$ with the standard online ridge regression estimator $\theta_t = \arg\min_\theta \sum_{s=1}^t \ell(\theta) + \frac{\lambda}{2}\|\theta\|^2$, the regret equality implies

$$\sum_{s=1}^t \ell_s(\theta_{s-1}) - \ell_s(\theta^*) = \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^t \ell_s(\theta_{s-1})\|x_s\|_{V_s^{-1}}^2 - \frac{1}{2}\|\theta_t - \theta^*\|_{V_t}^2,$$

where

$$V_t := \lambda I + \sum_{s=1}^t x_s x_s^\top.$$

The negative term on the RHS is often not useful and thus ignored in online learning. However, this term exactly appears in confidence sets for linear bandits such as the standard self-normalized confidence set (SNCS) (Abbasi-Yadkori

et al., 2011). Using the fact that $\sum_{s=1}^t \ell_s(\theta_{s-1}) - \ell_s(\theta^*) \geq -\sigma_*^2 \ln(1/\delta)$ with high probability (see Lemma C.8), which we call the 'negative regret bound', we obtain

$$\frac{1}{2}\|\theta_t - \theta^*\|_{V_t}^2 \leq \frac{\lambda}{2}\|\theta^*\|^2 + \underbrace{\sum_{s=1}^t \ell_s(\theta_{s-1})\|x_s\|_{V_s^{-1}}^2}_{=:A_t} + \sigma_*^2 \ln(\frac{1}{\delta}) \ .$$

One can bound $\|\theta^*\| \leq S$ and use Assumption A1 to bound $\sigma_*^2 \leq \sigma_0^2$. One can further try to upper bound A_t to construct a confidence set. Indeed, this is the approach taken by Dekel et al. (2012) for online selective sampling (with minor differences) where they assume bounded noise; i.e., $\eta_t \mid \mathcal{F}_{t-1} \in [0,1]$ almost surely. This means that we have $\sigma_0^2 = 1/4$. This allows them to control $\ell_s(\theta_{s-1})$ loosely such that $A_t \leq \bar{A}_t := O(\sigma_0^2 d \ln(t))$ and construct a confidence set

$$\left\{\theta \in \mathbb{R}^d : \frac{1}{2}\|\theta_t - \theta\|_{V_t}^2 \leq \frac{\lambda}{2}S^2 + \bar{A}_t + \sigma_0^2 \ln(\frac{1}{\delta})\right\} \ .$$

While details differ, this is the essence of Dekel et al. (2012). However, this does not provide an improvement over SNCS. In another work of Jun et al. (2017), a slightly different technique of online-to-confidence-set conversion (see Section 5 for details) also introduces a term like the second term above; they use a union bound over the time steps to control $A_t \leq O(\sigma_0^2 S d \ln^2(t))$, which leads to an even looser bound than SNCS.

Departing from prior work, we propose to keep A_t as is, which motivates the following confidence set:

$$\left\{\theta \in \mathbb{R}^d : \frac{1}{2}\|\theta_t - \theta\|_{V_t}^2 \leq \frac{\lambda}{2}S^2 + A_t + \sigma_0^2 \ln(\frac{1}{\delta})\right\} \ . \quad (3)$$

The benefit is the following observation: If $\ell_s(\theta_{s-1}) \approx \ell_s(\theta^*)$, then $A_t = O(\sigma_*^2 d \ln(t))$ with high probability. Then, we can have a confidence set whose (normalized) confidence width (i.e., the upper bound on $\|\theta_t - \theta\|_{V_t}$) is of order $\mathcal{O}\left(\sqrt{d\sigma_*^2 + \sigma_0^2}\right)$ rather than $\mathcal{O}(\sqrt{d\sigma_0^2})$ of SNCS. Note that the assumption of $\ell_s(\theta_{s-1}) \approx \ell_s(\theta^*)$ is sensible since for large enough s, the loss of θ_{s-1} should be sufficiently similar to the true parameter θ^*.

However, our intuition does not easily lead to a confidence set whose radius is $\sqrt{d\sigma_*^2 + \sigma_0^2}$ for technical reasons that we omit here. Still, one can see that the proposed confidence set above cannot be strictly better than SNCS. The reason is that A_t contains $\ell_1(\hat{\theta}_0) = \ell_1(0) = y_1^2/2$, which can be $\Omega(B^2)$ where $B = \max_t \max_{x \in \mathcal{X}_1} |\langle x, \theta^* \rangle|$ that can be as large as S. Such a dependence does not appear in SNCS (note the factor S can be canceled out by setting a large λ), and thus (3) cannot be strictly better than SNCS. Instead, we found that an added assumption of $\|x_s\|_{V_s^{-1}}^2 \leq \frac{1}{2}$ helps attain the desired inequality (Lemma C.10):

$$A_t \lesssim \sum_{s=1}^t \ell_s(\theta^*) D_s^2 + \lambda S^2 + \sigma_0^2 \ln(1/\delta) \ .$$

While the tools proposed by Zhao et al. (2023a) can be used to satisfy $\|x_s\|_{V_s^{-1}}^2 \leq \frac{1}{2}$. However, this requires using weighted ridge regression, which changes the left-hand side $\frac{1}{2}\|\theta_t - \theta\|_{V_t}^2$ of the constraint now involves V_t that consists of the weighted versions of $\{x_s\}_{s=1}^t$, which will block us from applying the standard elliptical potential lemma (e.g., Abbasi-Yadkori et al. (2011, Lemma 11)) from the existing analysis technique. That is, even if we can achieve the target confidence width, the regret analysis must be done differently to accommodate the change. Therefore, both our confidence set and the regret analysis are our novelty.

Proposed confidence set. Departing from the standard ridge regression, we use a weighted ridge regression estimator inspired by Zhao et al. (2023a). Define the weight w_t:

$$w_t = 1 \wedge \frac{1}{\|x_t\|_{\Sigma_{t-1}^{-1}}} \in (0,1],$$

where $\Sigma_t = \lambda I + \sum_{s=1}^t w_s^2 x_s x_s^\top$. Since $\|w_t x_t\|_{\Sigma_{t-1}^{-1}}^2 \leq 1$, using Woodbury matrix identity, we obtain

$$D_t^2 := \|w_t x_t\|_{\Sigma_t^{-1}}^2 \leq \frac{1}{2} \ ,$$

which will be the key property that enables our semi-adaptive bound. The weighted ridge regression estimator we use is then defined as

$$\hat{\theta}_t = \arg\min_\theta \sum_{s=1}^t w_s^2 \ell_s(\theta) + \frac{\lambda}{2}\|\theta\|_2^2,$$

where $\ell_s(\theta) := \frac{1}{2}(x_s^\top \theta - y_s)^2$ is the squared loss. Equivalently, one can assume the noise model of

$$w_t y_t = \langle w_t x_t, \theta^* \rangle + w_t \eta_t,$$

and write down the weighted estimator as

$$\hat{\theta}_t = \arg\min_\theta \sum_{s=1}^t \ell_s^{\mathsf{w}}(\theta) + \frac{\lambda}{2}\|\theta\|_2^2,$$

where $\ell_s^{\mathsf{w}}(\theta) := \frac{1}{2}(w_s x_s^\top \theta - w_s y_s)^2 = w_s^2 \ell(\theta)$.

We then construct the following confidence set

$$\mathcal{C}_t^{\text{semi}} = \{\theta \in \mathbb{R}^d : \frac{1}{2}\|\hat{\theta}_t - \theta\|_{\Sigma_t}^2 \leq \frac{\lambda}{2}S^2 + \sum_{s=1}^t \ell_s^{\mathsf{w}}(\hat{\theta}_{s-1}) D_s^2 + \sigma_0^2 \ln(1/\delta) =: \gamma_t\}, \quad (4)$$

Algorithm 1 LOSAN (Linear Optimism with Semi-Adaptivity to Noise)

Input: norm bound S, sub-Gaussian parameter σ_0^2
for $t = 1, 2, \ldots$ **do**
  Observe the arm set $\mathcal{X}_t$.
  Pull $x_t = \arg\max_{x \in \mathcal{X}_t} \max_{\theta \in \mathcal{C}_{t-1}^{\text{semi}}} \langle x, \theta \rangle$ where $\mathcal{C}_{t-1}^{\text{semi}}$ is defined in (4) and

$$\max_{\theta \in \mathcal{C}_{t-1}^{\text{semi}}} \langle x, \theta \rangle = \langle x, \hat{\theta}_{t-1} \rangle + \sqrt{2\gamma_{t-1}} \|x\|_{\Sigma_{t-1}^{-1}} .$$

  Receive reward y_t.
end for

and show that the confidence set is time-uniformly valid as follows. Throughout, all the proofs are deferred to the appendix.

Theorem 2.1. *Take Assumption A1. Then,*

$$\mathbb{P}(\forall t \geq 1, \theta^* \in \mathcal{C}_t^{\text{semi}}) \geq 1 - \delta.$$

We show later (Appendix A) that, with high probability,

$$\gamma_t = O\left(\lambda S^2 + \sigma_*^2 d \ln(1 + \frac{t}{d\lambda}) + \sigma_0^2 \ln(1/\delta)\right) ,$$

which shows an orderwise improvement upon SNCS that has $O\left(\lambda S^2 + \sigma_0^2 d \ln(1 + \frac{t}{d\lambda}) + \sigma_0^2 \ln(1/\delta)\right)$. The gap becomes much larger when $\sigma_*^2 \ll \sigma_0^2$ or even $\sigma_*^2 = 0$.

Proposed bandit algorithm. We are now ready to present our algorithm, which follows the optimistic approach (Dani et al., 2008; Abbasi-Yadkori et al., 2011) with our confidence set $\mathcal{C}_t^{\text{semi}}$, which pulls the arm with the largest upper confidence bound. The full pseudo-code is presented in Algorithm 1.

Theorem 2.2. *Let $B = \max_{t=1}^{\infty} \max_{x \in \mathcal{X}_t} |\langle x, \theta^* \rangle|$. Under Assumption A1, Algorithm 1 with $\lambda = \sigma_0^2 / S^2$ satisfies that, with probability at least $1 - O(\delta)$, $\forall n \geq 1$,*

$$\text{Reg}_n \lesssim \sigma_* d \sqrt{n} + \sigma_0 \sqrt{dn \ln(1/\delta)} + dB,$$

where we omit $\text{polylog}(d, n, S, \sigma_0^2)$ *factors.*

Theorem 2.2 shows an improvement upon the state-of-the-art regret bound of $\sigma_0 d \sqrt{n \ln(1/\delta)} + dB$ which is reported in Lattimore & Szepesvári (2020, Exercise 19.3). When the sub-Gaussian parameter is largely over-specified (i.e., $\sigma_0^2 \gg \sigma_*^2$), our algorithm achieves a factor of $\sqrt{d}$ improvement, which is significant when d is not too small. To our knowledge, our confidence set and LOSAN are the first ones to achieve the semi-adaptivity under the sub-Gaussian noise assumption.

3. Full Adaptation to Bounded Noise

In this section, we turn to the bounded noise case. Specifically, we assume the model of (1) with the following noise assumption.

Assumption A2. The reward noise η_t satisfy $\eta_t \in [-R, R]$ for some $R \geq 0$ with probability 1 for every t. The algorithm has a priori knowledge of R.

Of recently proposed studies on variance-adaptive linear bandits (see Section 5 for more discussion), Zhao et al. (2023a) for the first time proposed an algorithm called SAVE that enjoys $\tilde{\mathcal{O}}(d\sqrt{(R^2 + \sum_{t=1}^n \sigma_t^2)})$ with polynomial time and space complexity. In fact, its space and time complexity is of the same order (up to logarithmic factors) as the standard linear bandits such as OFUL (Abbasi-Yadkori et al., 2011), which is $O(d^2 \max_{t=1}^n |\mathcal{X}_t| n)$ time complexity and $O(d^2)$ space complexity. SAVE has improved both the regret bound (a factor of $\sqrt{d}$) and the time complexity (exponential to polynomial) upon the previous state-of-the-art Kim et al. (2022).

However, SAVE is inherently based on SupLinRel (Auer, 2002) or SupLinUCB (Chu et al., 2011), which maintains L disjoint buckets of the observed samples. Since each estimator $\hat{\theta}_\ell$ is computed from the samples from ℓ-th bucket only, SupLinRel-style algorithms are usually an order of magnitude worse than the standard algorithmic frameworks such as optimistic approach (Abbasi-Yadkori et al., 2011) or posterior sampling approach (Agrawal & Goyal, 2013).

Is it possible to achieve the same order of regret bound and computational complexity without wasting samples? We answer this question in the affirmative by developing an optimistic-style algorithm. As is usual, we first need to construct a confidence set for the unknown parameter θ^*. While one can leverage the existing confidence set used for SAVE, it works under the assumption where the conditional variance of $\eta_t \mid \mathcal{F}_{t-1}$ at time t is fixed ahead of time before the bandit game starts. That is, they assume that there exists a sequence of deterministic values $\bar{\sigma}_1^2, \ldots, \bar{\sigma}_n^2$ such that

$$\forall t \in [n], (\sigma_t^2 \mid \mathcal{F}_{t-1}) = \bar{\sigma}_t^2 . \tag{5}$$

We found this unrealistic as it can only deal with the noise that varies only as a function of the time step t (e.g., a seasonal effect of the customers' behavior in recommendation systems). For example, such an assumption cannot effectively capture the case where the noise variance changes as a function of the specific arm being pulled. Furthermore, we have found that the variance-adaptive confidence set used in SAVE is quite loose, which requires a lot of samples until outperforms even SNCS; see Figure 1.

Motivated by the limitations of the sample-inefficiency of the prior art, we propose a novel confidence set and bandit algorithm that remove the limited assumption on the variance and exhibit much improved numerical performance. Our proposed confidence set computes L estimators and builds a confidence set as an intersection of L base confidence sets.

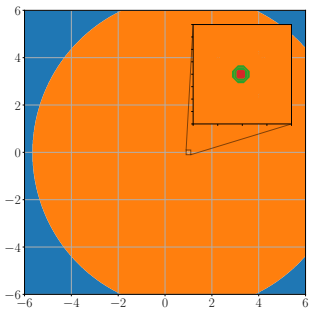


Figure 1. The green line (very small) represents our confidence set $\mathcal{C}_t^{\text{full}}$ and the orange area represents the confidence set of Zhao et al. (2023a), which is also implemented as an intersection of L confidence sets like ours. We use $n = 500{,}000$ samples, $d = 2$, $L = 9$, and $\sigma_t^2 = 0.1, \forall t$. With $\theta^* = (1, 0)$, the upper confidence bound on the mean reward of the arm $x = (1, 0)$ is 1.01 with our method while it is 7.66 with their method and 1.05 with SNCS.

Inspired by the base confidence set of SAVE (Zhao et al., 2023a), we construct our base confidence sets by leveraging weighted ridge regression estimators, but with a critical difference that (i) we leverage the regret equality (Lemma C.7), which results in a significantly tightened confidence set (see Figure 1 for numerical results) and (ii) we use an exponential cover to remove the restrictive assumption (5) and adapt to any conditional variance, and (iii) compute a set of secondary estimator that will be the center of our confidence ellipsoid.

Proposed confidence set. Fix $L \in \mathbb{N}_+$ and let $\ell \in [L]$. We let $\rho_\ell = 2^{-\ell}$ and set the regularization parameter $\lambda_\ell = \frac{R^2}{S^2}\rho_\ell^2$. We set the weights

$$w_{s,\ell} = 1 \wedge \frac{\rho_\ell}{\|x_s\|_{\Sigma_{s-1,\ell}^{-1}}} ,$$

where

$$\Sigma_{t,\ell} = \lambda_\ell I + \sum_{s=1}^{t} w_{s,\ell}^2 x_s x_s^\top . \tag{6}$$

The weight ensures that $\|w_{t,\ell} x_t\|_{\Sigma_{t-1,\ell}^{-1}} \leq \rho_\ell$. Define $\ell_{s,\ell}^{\mathsf{w}}(\theta) := w_{s,\ell}^2 \ell_s(\theta)$ and $\ell_{s,\ell}^{\mathsf{w}}(\theta; \theta') = \frac{w_{s,\ell}^2}{2}(x_s^\top(\theta - \theta'))^2$. We then compute the weighted ridge regression estimator:

$$\hat{\theta}_{t,\ell} = \arg\min_\theta L_{t,\ell}(\theta),$$

where

$$L_{t,\ell}(\theta) = \sum_{s=1}^{t} \ell_{s,\ell}^{\mathsf{w}}(\theta) + \frac{\lambda_\ell}{2}\|\theta\|_2^2 .$$

We then compute our secondary estimators as follows:

$$K_{t,\ell}(\theta) := \sum_{s=1}^{t} \ell_{s,\ell}^{\mathsf{w}}(\theta) + \sum_{s=1}^{t} \ell_{s,\ell}^{\mathsf{w}}(\theta; \hat{\theta}_{s-1}) + \frac{\lambda_\ell}{2}\|\theta\|^2,$$

$$\bar{\theta}_{t,\ell} := \arg\min_\theta K_{t,\ell}(\theta), \tag{7}$$

With $\overline{\Sigma}_{t,\ell} := \lambda_\ell I + 2\sum_{s=1}^{t} w_{s,\ell}^2 x_s x_s^\top$, we define our confidence set as an intersection of confidence ellipsoids centered at $\bar{\theta}_{t,\ell}$ as follows:

$$\mathcal{C}_t^{\text{full}} := \cap_{\ell=1}^{L} \mathcal{C}_{t,\ell}^{\text{full}}, \tag{8}$$

where

$$\forall \ell \in [L], \mathcal{C}_{t,\ell}^{\text{full}} = \left\{ \theta \in \mathbb{R}^d : \frac{1}{2}\|\theta - \bar{\theta}_{t,\ell}\|_{\overline{\Sigma}_{t,\ell}}^2 \leq \beta_{t,\ell} \right\},$$

and $\beta_{t,\ell}$ is defined based on its previous version $\beta_{t-1,\ell}$. Specifically, with $D_{t,\ell}^2 := \|w_{t,\ell} x_t\|_{\Sigma_{t,\ell}^{-1}}^2$,

$$\begin{aligned}\beta_{t,\ell} := & L_{t,\ell}(\hat{\theta}_{t,\ell}) - K_{t,\ell}(\bar{\theta}_{t,\ell}) + \frac{\lambda_\ell}{2} S^2 + \sum_{s=1}^{t} \ell_{s,\ell}^{\mathsf{w}}(\hat{\theta}_{s-1,\ell}) D_{s,\ell}^2 \\ & + \sqrt{8\rho_\ell^2 \bar{\beta}_{t-1,\ell} \left(\textstyle\sum_{s=1}^{t} \ell_{s,\ell}^{\mathsf{w}}(\hat{\theta}_{s-1}) + R^2 \ln(2L/\delta) \right) \xi_{t,\ell}} \\ & + 2^{k_{t,\ell}} \rho_\ell R \sqrt{2\beta_{0,\ell}} \xi_{t,\ell},\end{aligned} \tag{9}$$

where $\bar{\beta}_{t-1,\ell} := \max_{s=0}^{t-1} \beta_{s,\ell}$, $\beta_{0,\ell} := \frac{\lambda_\ell}{2} S^2$, $\xi_{t,\ell} := \ln(\sqrt{\pi(t+1)} \cdot \frac{6.8L \cdot k_{t,\ell} \ln^2(1+k_{t,\ell})}{\delta})$, and $k_{t,\ell} := 1 \vee \lceil \log_2(\sqrt{\bar{\beta}_{t-1,\ell}/\beta_{0,\ell}}) \rceil$.

Our confidence set enjoys the following correctness guarantee.

Theorem 3.1. *Under Assumption A2, we have*

$$\mathbb{P}(\forall t \geq 1, \theta^* \in \mathcal{C}_t^{\text{full}}) \geq 1 - \delta.$$

As one will see from the proof, the introduction of the secondary estimator $\bar{\theta}_{t,\ell}$ is a nonessential part of the theoretical guarantee as it only helps the numerical tightness. The secondary estimator only appears during the attempt to write down the confidence set as a canonical quadratic form – such an estimator also appears for the same reason in Abbasi-Yadkori et al. (2012). One can easily see that it is possible to use $\hat{\theta}_{t,\ell}$ directly with a slightly looser confidence set.

We later show that (Lemma B.2)

$$\beta_{t,\ell} \leq \bar{\beta}_{t,\ell} = \tilde{\mathcal{O}}(\rho_\ell^2 (R^2 + \sum_{s=1}^{t} \sigma_s^2) \ln^2(t/\delta)) ,$$

which scales with the conditional variances as desired and matches the order of the confidence set of Zhao et al. (2023a). Note that due to the particular weights being used, each confidence set ℓ is tight in certain regimes only – this is the reason why we take an intersection over L of them.

Algorithm 2 LOFAV (Linear Optimism with Full Adaptivity to Variance)

Input: norm bound S, time horizon n, the number of levels L
for $t = 1, 2, \ldots, n$ **do**
  Observe the arm set $\mathcal{X}_t$.
  Pull $x_t = \arg\max_{x \in \mathcal{X}_t} \max_{\theta \in \mathcal{C}_{t-1}^{\text{full}}} \langle x, \theta \rangle$ where

$$\max_{\theta \in \mathcal{C}_{t-1}^{\text{full}}} \langle x, \theta \rangle = \min_{\ell \in [L]} \langle x, \bar{\theta}_{t-1,\ell} \rangle + \sqrt{2\beta_{t-1,\ell}} \|x\|_{\overline{\Sigma}_{t-1,\ell}^{-1}}.$$

  and $\mathcal{C}_{t-1}^{\text{full}}$, $\bar{\theta}_{t-1,\ell}$, $\beta_{t-1,\ell}$, and $\overline{\Sigma}_{t-1,\ell}$ are defined in (8), (6), (7), and (9), respectively.
  Receive reward y_t.
end for

Proposed bandit algorithm. Equipped with our improved confidence set, we construct an OFUL-style algorithm, which we call **LOFAV** (Linear Optimism with Full Adaptivity to Variance); see Algorithm 2. The time and space complexity is the same as OFUL, i.e., $O(d^2 n \max_{t=1}^n |\mathcal{X}_t|)$, up to logarithmic factors when setting $L = O(\ln(n/d))$ as suggested in Theorem 3.2 below. For efficient implementation, one needs to maintain sufficient statistics for the estimators $\hat{\theta}_{t,\ell}$ and $\bar{\theta}_{t,\ell}$, and update the inverse of the covariance matrices $\Sigma_{t,\ell}^{-1}$ and $\overline{\Sigma}_{t,\ell}^{-1}$ using the matrix inversion lemma. Evaluating the loss functions such as $L_{t,\ell}(\hat{\theta}_{t,\ell})$ can be done incrementally as well by expanding the square and extracting sufficient statistics.

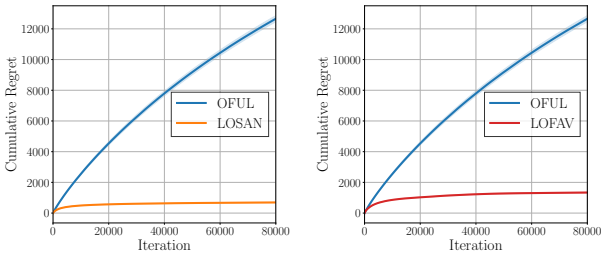
We report our regret analysis result in Theorem 3.2 below, which attains the optimal variance-adaptive regret bound up to logarithmic factors.

Theorem 3.2. *Under Assumption A2, Algorithm 2 with* $\delta \in (0, 1/2]$ *and* $L = 1 \vee \lceil \frac{1}{2} \log_2(n/d) \rceil$ *achieves, with probability at least* $1 - \delta$,

$$\text{Reg}_n \lesssim d \sqrt{\left(R^2 + \sum_{t=1}^n \sigma_t^2\right) \ln^2\left(\frac{1}{\delta}\right)} + dB,$$

where we omit $\text{polylog}(d, n, S, R)$ *factors.*

The regret analysis is challenging since, unlike SAVE and other variance-adaptive algorithms such as LinNATS (Xu et al., 2023), LOFAV does not perform the stratification of the arms that provides an easy control of how many samples fall in each bucket, which is directly related to the regret in a systematic manner. That is, each bucket ℓ is explicitly associated with a particular instantaneous regret bound of order $2^{-\ell}$). We overcome this difficulty by turning to the peeling-based regret analysis (Kim et al., 2022; He et al., 2021), which provides a strong control on the instantaneous regret when used with the elliptical potential 'count' lemma (Kim et al., 2022) that we restate in Lemma C.2.



(a) Sub-Gaussian noise (b) Bounded noise

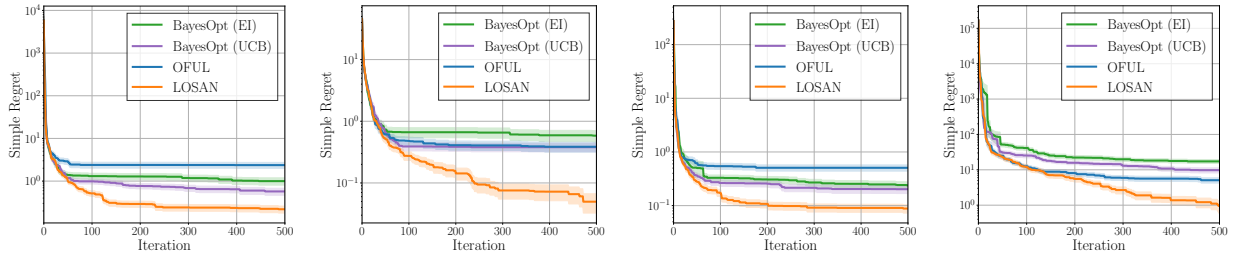
Figure 2. Results of synthetic experiments with LOSAN and LOFAV. To fairly compare our algorithms to OFUL, we perform each experiment over 50 rounds where $S = 1.0$, $d = 32$, $|\mathcal{X}_t| = 128$, and σ_0 or $R = 1.0$.

Practical version. The performance of LOFAV can be further improved by adding extra L confidence sets without further splitting the target failure rate δ. This is because there is an event being assumed in the proof of confidence set $\mathcal{C}_t^{\text{full}}$ that automatically implies a semi-adaptive style confidence set (one set for each $\ell \in [L]$). The reason why it helps is that there are extra logarithmic factors and constants in $\mathcal{C}_t^{\text{full}}$, which can only be overcome after observing a large number of samples. Equipping LOFAV with semi-adaptive style confidence sets helps avoid excessive exploration for the small-sample regime. We provide a precise description of this modification in Appendix D.

Anytime version. The current version of LOFAV requires knowledge of the time horizon n. This can be easily lifted by changing the union bound over L confidence sets into a union bound over $\mathbb{N}_+$. Specifically, one can divide the failure rate δ into $\{\delta_\ell\}_{\ell=1}^\infty$ such that $\sum_{\ell=1}^\infty \delta_\ell = \delta$; e.g., $\delta_\ell = \frac{\delta}{\ell^2} \cdot \frac{6}{\pi^2}$. The algorithm needs to maintain only $L_t = \Theta(\ln(t/d))$ confidence sets up to time step t.

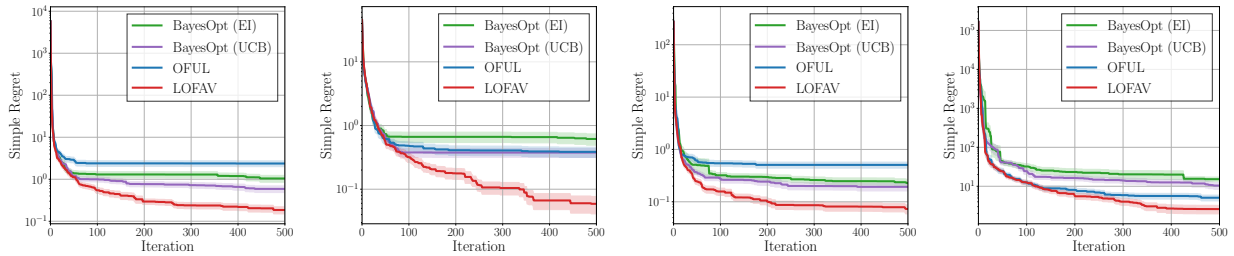
4. Experiments

We conduct several experiments with two noise types, i.e., the sub-Gaussian and bounded noises, to verify the theoretical analysis of our algorithms. First, synthetic experiments are tested to show empirical results of our methods and OFUL in terms of cumulative regrets. Second, the experiments of Bayesian optimization, i.e., simple benchmark functions and neural architecture search benchmarks, are carried out to compare our algorithms to OFUL and simple Bayesian optimization algorithms. Unless noted otherwise, we perform each experiment with 50 random trials. Moreover, the sample mean and the standard error of the sample are depicted in Figures 2 to 5. For sub-Gaussian and bounded noises, we assume that the true noises are sampled from the distributions with $\sigma_*^2 = 0.01^2$ (Guassian) and $\sigma_t^2 = 0.01^2$ ($\eta \sim \text{Uniform}\{-0.01, 0.01\}$) respectively. For LOFAV, we implement a practical version discussed in Section 3.



(a) Beale (b) Branin (c) Three-Hump Camel (d) Zakharov 4D

Figure 3. Bayesian optimization results of LOSAN with random Fourier features and sub-Gaussian noises for four benchmark functions. We perform each experiment over 50 rounds where $S = 1.0$, $d = 128$, $|\mathcal{X}_t| = 512$, and $\sigma_0 = 1.0$.



(a) Beale (b) Branin (c) Three-Hump Camel (d) Zakharov 4D

Figure 4. Bayesian optimization results of LOFAV with random Fourier features and bounded noises for four benchmark functions. We perform each experiment over 50 rounds where $S = 1.0$, $d = 128$, $|\mathcal{X}_t| = 512$, and $R = 1.0$.

4.1. Synthetic Experiments

To generate synthetic experiments, we sample an unknown parameter θ^* and an arm set $\mathcal{X}_t$ from multivariate normal distributions with zero mean and unit variance and then normalize them in order to locate them on a hypersphere of radius S. As presented in Figure 2, our methods LOSAN and LOFAV exhibit better performance in terms of cumulative regrets than OFUL. We perform a more comprehensive experiments in Appendix E.

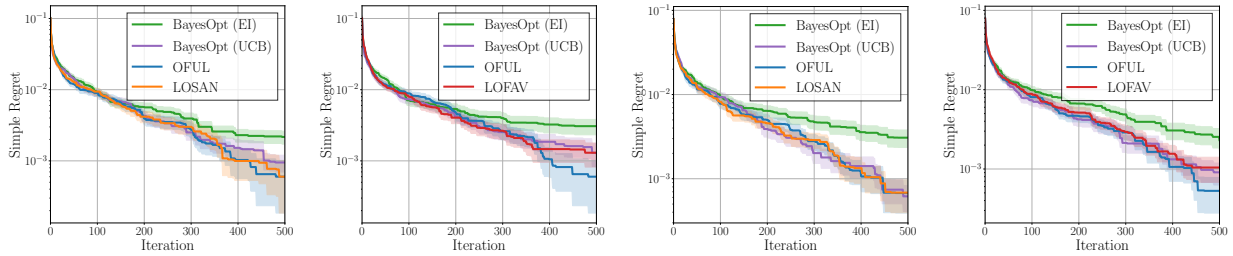
4.2. Application to Bayesian Optimization

We expand the application of our linear bandit models to Bayesian optimization (Garnett, 2023). Along with OFUL, we compare our methods to simple Bayesian optimization algorithms with Gaussian process regression (Rasmussen & Williams, 2006) and either expected improvement acquisition function (Jones et al., 1998) or Gaussian process upper confidence bound (Srinivas et al., 2010). For the Gaussian process regression, we use a linear kernel where kernel parameters are sought by marginal likelihood maximization. We select this linear kernel for fair comparison to linear bandit models. Multi-start L-BFGS-B (Byrd et al., 1995) is utilized in the process of acquisition function optimization. To deal with a fixed number of arms with Bayesian optimization, we modify the standard Bayesian optimization algorithm to choose the nearest arm after determining the next point through Bayesian optimization (Garrido-Merchán & Hernández-Lobato, 2020).

For LOSAN, LOFAV, OFUL, and Bayesian optimization approaches, we make use of random Fourier features (Rahimi & Recht, 2007) in order to solve Bayesian optimization problems using linear models or Gaussian process regression with the linear kernel. Each original point is transformed into a 128-dimensional random feature following the work by Rahimi & Recht (2007). Other configurations for these algorithms are the same as the configurations used in Section 4.1. For arm selection, we uniformly sample a fixed number of arms from a specific search space depending on benchmarks. Moreover, instead of cumulative regret, we use simple regret as the performance measure in the experiments of Bayesian optimization tasks where the simple regret at time t is set to be the best instantaneous regret until t.

Benchmark functions. We test four benchmark functions: Beale, Branin, Three-Hump Camel, and Zakharov 4D functions. As illustrated in Figures 3 and 4, our LOSAN and LOFAV are better than OFUL and Bayesian optimization.

NATS-Bench. We utilize NATS-Bench (Dong et al., 2021), which is a set of benchmarks on neural architecture search (Zoph & Le, 2017), in order to examine our algo-



(a) CIFAR-100 (Sub-Gaussian) (b) CIFAR-100 (Bounded) (c) ImageNet (Sub-Gaussian) (d) ImageNet (Bounded)

Figure 5. Bayesian optimization results of LOSAN and LOFAV with random Fourier features and sub-Gaussian or bounded noises for NATS-Bench. We perform each experiment over 50 rounds where $S = 1.0$, $d = 128$, $|\mathcal{X}_t| = 512$, and σ_0 or $R = 1.0$.

rithms and baseline methods. We employ the size search space of NATS-Bench for CIFAR-100 and ImageNet-16-120. Since the number of possible architecture candidates, i.e., 32,768, is significantly greater than the number of arms often used in the linear bandits, we sample 512 arms from the continuous space of the size search space. Figure 5 demonstrates that there is no single method that dominates the rest. LOSAN shows comparable results for CIFAR-100 (sub-Gaussian) and ImagenNet (sub-Gaussian).

5. Related Work

Heteroscedastic linear bandits. For the sub-Gaussian noise, Kirschner & Krause (2018) first considered the linear bandit problem where the noise at time t is assumed to be σ_t^2-sub-Gaussian where σ_t^2 is known to the learner. Zhou et al. (2021) and Zhou & Gu (2022) consider the setup where the noise σ_t^2 is known only after the learner chooses the arm at time t.

For the bounded *noise* setup $\eta_t \in [-1, 1]$ (often appear as *rewards* being bounded), the seminal work of Zhang et al. (2021) proposes a linear bandit algorithm called VOFUL whose regret bound was shown to be adaptive to the unknown variances $\{\sigma_t^2\}_{t=1}^n$. They achieved a regret upper bound of order $d^{4.5}\sqrt{1 + \sum_{t=1}^n \sigma_t^2} + d^5$. Kim et al. (2022) then improved the regret bound of VOFUL to $d^{1.5}\sqrt{1 + \sum_{t=1}^n \sigma_t^2} + d^2$. However, VOFUL has an exponential time complexity. Zhao et al. (2023a) has made a breakthrough by achieving the optimal variance-adaptive worst-case regret bound of $d\sqrt{1 + \sum_{t=1}^n \sigma_t^2} + d$ with a computationally efficient algorithm called SAVE. However, as detailed in Section 3, SAVE makes a limiting assumption on the variance, and the algorithm is not practical. Our algorithm LOFAV overcomes these two limitations, which we claim to be the first practical variance-adaptive algorithm. Xu et al. (2023) further extends SAVE to a Thompson sampling style algorithm, which adds an extra factor of $\sqrt{d}$ in the regret bound – this is known to be unavoidable for linear Thompson sampling style algorithms (Hamidi & Bayati, 2020). Unfortunately, they also use the inefficient SupLinRel-style algorithm. This also means that their time complexity per round w.r.t. the arm set size $|\mathcal{X}_t|$ scales with $d^2|\mathcal{X}_t|$ rather than $d|\mathcal{X}_t|$ of the standard linear Thompson sampling Agrawal & Goyal (2013), which is one of the main benefits of linear Thompson sampling.

Improved confidence sets. Recently, there have been quite a few studies that improve confidence sets for linear models upon the standard self-normalized confidence set (SNCS) (Abbasi-Yadkori et al., 2011, Theorem 2). Emmenegger et al. (2023) propose a novel confidence set construction based on weighted sequential likelihood ratio tests, which is empirically shown to be tighter than SNCS. However, the analysis does not show a tighter bound than SNCS due to an extra factor of S (the bound on $\|\theta^*\|_2$), which requires further investigation. Flynn et al. (2023) propose a novel adaptive martingale mixture to construct an improved confidence set that is both numerically and provably tighter than SNCS. However, the degree of tightness was not precisely quantified as an orderwise improvement. In stark contrast, our confidence set for LOSAN enjoys an improvement that is precisely quantified in terms of the true noise level σ_*^2 and the specified noise level σ_0^2. We are not aware of any comparable results in prior work.

Confidence sets via online learning regret bounds. In one way or another, regret bounds of online learning (OL) algorithms play an important role in constructing or analyzing novel confidence bounds or sets. The seminal work of Rakhlin & Sridharan (2017) makes a strong case by showing that the existence of an OL regret bound implies a confidence bound. For (generalized) linear models, we have found that there have been three types of results that leverage OL regret bounds to construct confidence sets. The first is to construct a confidence set by running a specific OL algorithm (e.g., online Newton step (Hazan et al., 2007)) and leverage its regret bound to quantify the confidence width, which relies critically on a negative term in the regret bound (Dekel et al., 2012; Crammer & Gentile, 2013; Gen-

tile & Orabona, 2014; Zhang et al., 2016). The second is the so-called online-to-confidence-set conversion (Abbasi-Yadkori et al., 2012; Jun et al., 2017), which constructs a confidence set by *regressing* on the prediction made by running an OL algorithm $\mathcal{A}$ whose confidence width becomes a function of the regret bound of $\mathcal{A}$. The advantage of this method is that we are not married to those OL algorithms that have a particular negative term in the regret bound, which provides more flexibility. The third is the so-called regret-to-confidence-set conversion (Lee et al., 2024), which constructs a confidence set with the maximum likelihood estimator but characterizes the confidence width with an achievable OL regret bound. Similar to Rakhlin & Sridharan (2017), this technique only requires the *existence* of a regret bound achieved by an online learner, and thus one can even use the regret bounds of computationally intractable OL algorithms such as those in Mayo et al. (2022). Interestingly, the role of OL regret bounds here is solely an analysis tool, perhaps providing a shortcut to otherwise complicated analysis. Our confidence set for LOSAN belongs to the first technique above since we leverage the negative term in the regret bound directly. In this regard, we remark that Emmenegger et al. (2023) also use OL regret bounds for the analysis of their confidence set algorithm.

6. Conclusion

Our advances in noise-adaptive confidence sets and applications to linear bandits and Bayesian optimization open up numerous exciting future directions. First, investigating whether similar variance-adaptive worst-case regret bounds are possible in more generic hypothesis classes and various noise models is an open question. Second, the weighted ridge regression considered in this paper computes the weights in a sequential manner. It would be interesting to study if there exists a batch counterpart of the weights, which could be more useful for the fixed design case. Finally, it would be interesting to develop a Thompson sampling version of the variance-adaptive algorithm whose time complexity w.r.t. the arm set size $|\mathcal{X}_t|$ is $O(d|\mathcal{X}_t|)$ per iteration.

Acknowledgements

This work used, in part, the H2P cluster of the University of Pittsburgh Center for Research Computing, which is supported by National Science Foundation award number OAC-2117681. Kwang-Sung Jun was supported in part by the National Science Foundation under grant CCF-2327013.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–19, 2011.

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3–11, 1999.

Agrawal, S. and Goyal, N. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 127–135, 2013.

Auer, P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Avadhanula, V., Colini Baldeschi, R., Leonardi, S., Sankararaman, K. A., and Schrijvers, O. Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference*, pp. 2805–2817, 2021.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, 2013.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Chowdhury, S. R., Saux, P., Maillard, O., and Gopalan, A. Bregman deviations of generic exponential families. In Neu, G. and Rosasco, L. (eds.), *Proceedings of the Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pp. 394–449. PMLR, 12–15 Jul 2023.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pp. 208–214, 2011.

Cover, T. M. and Ordentlich, E. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, 1996.

Crammer, K. and Gentile, C. Multiclass Classification with Bandit Feedback Using Adaptive Regularization. *Mach. Learn.*, 90(3):347–383, 2013.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Dekel, O., Gentile, C., and Sridharan, K. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13:2655–2697, 2012.

Dong, X., Liu, L., Musial, K., and Gabrys, B. NATS-Bench: Benchmarking NAS algorithms for architecture topology and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3634–3646, 2021.

Emmenegger, N., Mutny, M., and Krause, A. Likelihood ratio confidence sets for sequential decision making. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Fan, X., Grama, I., and Liu, Q. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1–22, 2015.

Flynn, H., Reeb, D., Kandemir, M., and Peters, J. Improved algorithms for stochastic linear bandits using tail bounds for martingale mixtures. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Gales, S. B., Sethuraman, S., and Jun, K.-S. Norm-Agnostic Linear Bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Garnett, R. *Bayesian Optimization*. Cambridge University Press, 2023.

Garrido-Merchán, E. C. and Hernández-Lobato, D. Dealing with categorical and integer-valued variables in Bayesian optimization with Gaussian processes. *Neurocomputing*, 380:20–35, 2020.

Gentile, C. and Orabona, F. On Multilabel Classification and Ranking with Bandit Feedback. *Journal of Machine Learning Research*, 15:2451–2487, 2014.

Hamidi, N. and Bayati, M. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic Regret Algorithms for Online Convex Optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 99–109, 2017.

Kim, Y., Yang, I., and Jun, K.-S. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 358–384. PMLR, 2018.

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

Lee, J., Yun, S.-Y., and Jun, K.-S. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the International Conference on World Wide Web*, pp. 661–670, 2010.

Mayo, J. J., Hadiji, H., and van Erven, T. Scale-free unconstrained online learning for curved losses. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 4464–4497. PMLR, 2022.

Orabona, F. A modern introduction to online learning, 2023.

Orabona, F. and Jun, K.-S. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 2023.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, pp. 1177–1184, 2007.

Rakhlin, A. and Sridharan, K. On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities. In *Proceedings of the Conference On Learning Theory (COLT)*, pp. 1704–1722, 2017.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1015–1022, 2010.

Ville, J. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

Xu, R., Min, Y., and Wang, T. Noise-adaptive thompson sampling for linear contextual bandits. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-h. Online Stochastic Linear Optimization under One-bit Feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pp. 392–401, 2016.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Variance-Aware Confidence Set: Variance-Dependent Bound for Linear Bandits and Horizon-Free Bound for Linear Mixture MDP. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4977–5020. PMLR, 12–15 Jul 2023a.

Zhao, Y., Stephens, C., Szepesvári, C., and Jun, K.-S. Revisiting simple regret: Fast rates for returning a good arm. *Proceedings of the International Conference on Machine Learning (ICML)*, 2023b.

Zhou, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture MDPs. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36337–36349, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 4532–4576. PMLR, 2021.

Zoph, B. and Le, Q. V. Neural architecture search with reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Appendix

Table of Contents

A Details for Semi-Adaptation **12**

A.1 Proof of the Confidence Set (Theorem 2.1) . . . 12

A.2 Analysis of the (Normalized) Confidence Width $\sqrt{\gamma_t}$ . . . 13

A.3 Regret Analysis (Proof of Theorem 2.2) . . . 13

B Details for Full Adaptation **14**

B.1 Proof of Theorem 3.1 . . . 14

B.2 Proof of Theorem 3.2 . . . 18

C Utility Lemmas **20**

C.1 Online Ridge Regression . . . 23

D Implementation Details of LOFAV **26**

E Additional Experiments for LOSAN **27**

A. Details for Semi-Adaptation

For convenience, we introduce the following model assumption, which helps us use simpler notations for both Section A and B.

Assumption A3. Let $\{(z_t \in \mathbb{R}^d, u_t \in \mathbb{R})\}_{t=1}^{\infty}$ be a sequence of feature vectors and observed labels. Let $\mathcal{F}_t = \sigma(z_1, u_1, \ldots, z_t, u_t, z_{t+1})$. Assume that $u_t = z_t^\top \theta^* + \nu_t$ where $\nu_t \mid \mathcal{F}_{t-1}$ is σ_*^2-sub-Gaussian (i.e., $\forall \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda \nu_t) \mid \mathcal{F}_{t-1}] \le e^{\lambda^2 \sigma^2/2}$). Let $\sigma_t^2 = \mathbb{E}[\nu_t^2 \mid \mathcal{F}_{t-1}]$. Let $L_t(\theta) = \sum_{s=1}^t f_s(\theta) + \frac{\lambda}{2}\|\theta\|^2$ where $f_s(\theta) = \frac{1}{2}(z_s^\top \theta - u_s)^2$ and define

$$\hat{\theta}_t = \arg\min_\theta L_t(\theta) \quad \text{and} \quad \Sigma_t = \lambda I + \sum_{s=1}^t z_s z_s^\top .$$

Specifically, the model assumption in Section 2 can be reduced to Assumption A3 by setting $z_t = w_t x_t$, $u_t = w_t y_t$, $\nu_t = w_t \eta_t$, $f_s(\theta) = \ell_s^{\mathsf{w}}(\theta)$ with the same sub-Gaussian parameter σ_*^2. With this, other notations like $\hat{\theta}_{t-1}$, $\Sigma_t = \lambda I + \sum_{s=1}^t w_s^2 x_s x_s^\top = \lambda I + \sum_{s=1}^t z_s z_s^\top$, and $D_s^2 = \|w_t x_t\|_{\Sigma_t^{-1}}^2 = \|z_t\|_{\Sigma_t^{-1}}^2$ remain the same. Throughout this section, we use this set of notations to avoid clutter.

A.1. Proof of the Confidence Set (Theorem 2.1)

Let $r_s := z_s^\top(\hat{\theta}_{s-1} - \theta^*)$. The regret equality of FTRL (Lemma C.7) states that

$$\frac{1}{2}\|\hat{\theta}_t - \theta^*\|_{\Sigma_t}^2 = \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^t f_s(\hat{\theta}_{s-1}) D_s^2 + \sum_{s=1}^t f_s(\theta^*) - f_s(\hat{\theta}_{s-1}) .$$

The negative regret bound (Lemma C.8) implies that, with probability at least $1-\delta$,

$$\forall t \ge 1, \sum_{s=1}^t (f_s(\theta^*) - f_s(\hat{\theta}_{s-1})) \le \sigma_*^2 \ln(1/\delta) .$$

Thus, we have

$$\begin{aligned}\forall t \geq 1, \frac{1}{2}\|\hat{\theta}_t - \theta^*\|_{\Sigma_t}^2 &\leq \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 + \sigma_*^2 \ln(1/\delta) \\ &\leq \frac{\lambda}{2} S^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 + \sigma_0^2 \ln(1/\delta) \,. \qquad \text{(Assumption A1)}\end{aligned}$$

Therefore, it is easy to see that with probability at least $1-\delta$,

$$\forall t \geq 1, \theta^* \in \mathcal{C}_t^{\text{semi}} \,.$$

This concludes the proof of Theorem 2.1.

A.2. Analysis of the (Normalized) Confidence Width $\sqrt{\gamma}_t$

Proposition A.1. *With probability at least* $1-3\delta$*,*

$$\gamma_t \lesssim \lambda S^2 + \sigma_*^2 d \ln\left(1 + \frac{t}{d\lambda}\right) + \sigma_0^2 \ln(1/\delta) \,.$$

Proof. By Lemma C.10, we have, with probability at least $1-2\delta$,

$$\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \leq \frac{3}{2}\lambda\|\theta^*\|_2^2 + 3\sum_{s=1}^{t} f_s(\theta^*) D_s^2 + 10\sigma_*^2 \ln(1/\delta) \,.$$

Thus,

$$\gamma_t \lesssim \lambda S^2 + 3\sum_{s=1}^{t} f_s(\theta^*) D_s^2 + \sigma_0^2 \ln(1/\delta) \,. \qquad (\sigma_*^2 \leq \sigma_0^2)$$

It remains to study how $\sum_{s=1}^{t} f_s(\theta^*) D_s^2$ scales.

Since ν_s is σ_*^2-sub-Gaussian, ν_s^2 is ($v = 2\sigma_*^4$,$c = 2\sigma_*^2$)-subgamma Boucheron et al. (2013, Section 2.4). Then, we can use the concentration of subgamma random variables Boucheron et al. (2013, Theorem 2.3) to obtain

$$\begin{aligned}\sum_{s=1}^{t} f_s(\theta^*) D_s^2 &= \sum_{s=1}^{t} \nu_s^2 D_s^2 \\ &\leq \sum_{s=1}^{t} \sigma_*^2 D_s^2 + \sqrt{4\sigma_*^4 \sum_{s=1}^{t} D_s^4 \ln(1/\delta)} + 2\sigma_*^2 \ln(1/\delta) \\ &\overset{(a)}{\leq} \sigma_*^2 \cdot d\ln(1 + \frac{t}{d\lambda}) + \sigma_*^2 \sqrt{4d\ln(1+\frac{t}{d\lambda})\ln(1/\delta)} + 2\sigma_*^2 \ln(1/\delta) \\ &\leq 2\sigma_*^2 \cdot d\ln(1 + \frac{t}{d\lambda}) + 3\sigma_*^2 \ln(1/\delta)\end{aligned}$$

where (a) is by $D_s^4 \leq D_s^2$ (since $D_s^2 \leq 1$) and the standard elliptical potential lemma (see Lemma C.1). This concludes the proof. □

A.3. Regret Analysis (Proof of Theorem 2.2)

Define the event

$$\mathcal{E}_1 := \{\forall t \geq 1, \theta^* \in \mathcal{C}_t\} \,.$$

Let $\text{reg}_t = \langle x_t^*, \theta^* \rangle - \langle x_t, \theta^* \rangle$ where $x_t^* = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle$. The following lemma is standard in linear bandit analysis.

Lemma A.2. *Let $\tilde{\theta}_t = \arg\max_{\theta \in \mathcal{C}_t} \langle x_t, \theta \rangle$. Under the event $\mathcal{E}_1$,*

$$\textit{reg}_t \le \|x_t\|_{\Sigma_{t-1}^{-1}} \sqrt{8\gamma_n}\,.$$

Proof. We have

$$\begin{aligned}
\mathrm{reg}_t &\le \langle x_t, \tilde{\theta}_t - \theta^* \rangle && \text{(def'n of } x_t\text{)}\\
&\le \|x_t\|_{\Sigma_{t-1}^{-1}} \|\tilde{\theta}_t - \theta^*\|_{\Sigma_{t-1}} && \text{(Cauchy-Schwarz)}\\
&\le \|x_t\|_{\Sigma_{t-1}^{-1}} (\|\tilde{\theta}_t - \hat{\theta}_{t-1}\|_{\Sigma_{t-1}} + \|\hat{\theta}_{t-1} - \theta^*\|_{\Sigma_{t-1}}) && \text{(triangle inequality)}\\
&\le \|x_t\|_{\Sigma_{t-1}^{-1}} \sqrt{8\gamma_{t-1}} && \text{(by } \mathcal{E}_1\text{)}\\
&\le \|x_t\|_{\Sigma_{t-1}^{-1}} \sqrt{8\gamma_n}\,. && \text{(monotonicity)}
\end{aligned}$$

□

For the regret analysis, using the fact that $\mathrm{reg}_t \le 2B$,

$$\sum_t^n \mathrm{reg}_t \le \sum_{t=1}^n \mathbb{1}\{w_t \ne 1\}\, 2B + \sum_{t=1}^n \mathbb{1}\{w_t = 1\}\, \mathrm{reg}_t\,.$$

For the first term,

$$\begin{aligned}
\sum_{t=1}^n \mathbb{1}\{w_t \ne 1\}\, 2B &= 2B \sum_{t=1}^n \mathbb{1}\left\{w_t \ne 1, \|w_t x_t\|^2_{\Sigma_{t-1}^{-1}} = 1\right\} && \text{(def'n of } w_t\text{)}\\
&\le 2B \sum_{t=1}^n \mathbb{1}\left\{\|w_t x_t\|^2_{\Sigma_{t-1}^{-1}} \ge 1\right\}\\
&\le 6B \cdot d \ln\left(1 + \frac{2}{\lambda}\right)\,. && \text{(by EPC (Lemma C.2))}
\end{aligned}$$

For the second term,

$$\begin{aligned}
\sum_{t=1}^n \mathbb{1}\{w_t = 1\}\, \mathrm{reg}_t &\le \sum_{t=1}^n \mathbb{1}\{w_t = 1\}\, \|x_t\|_{\Sigma_{t-1}^{-1}} \sqrt{8\gamma_n}\\
&= \sum_{t=1}^n \mathbb{1}\{w_t = 1\}\, \|w_t x_t\|_{\Sigma_{t-1}^{-1}} \sqrt{8\gamma_n}\\
&\le \sqrt{8\gamma_n} \sqrt{n \sum_{t=1}^n \|w_t x_t\|^2_{\Sigma_{t-1}^{-1}}}\\
&\le \sqrt{8\gamma_n} \sqrt{n 2d \ln(1 + \frac{n}{d\lambda})}
\end{aligned}$$

where the last line is due to the elliptical potential lemma (Abbasi-Yadkori et al., 2011, Lemma 11) and the fact that $\|w_t x_t\|^2_{\Sigma_{t-1}^{-1}} \le 1$ by the definition of w_t. Applying Proposition A.1 and noting that the event $\mathcal{E}_1$ happens with probability at least $1 - \delta$ conclude the proof.

B. Details for Full Adaptation

B.1. Proof of Theorem 3.1

The confidence set is an intersection of $\mathcal{C}_{t,\ell}$'s. Therefore, it suffices to prove that the confidence set $\mathcal{C}_{t,\ell}$ contains θ^* with probability at least $1 - \frac{\delta}{L}$. This is a direct consequence of Theorem B.1 below where we set $z_t = w_{t,\ell} x_t$, $u_t = w_{t,\ell} y_t$, and $\lambda = \lambda_\ell$, and replace δ with $\delta/(2L)$. Then, $f_s(\theta) = \ell^{\mathsf{w}}_{s,\ell}(\theta)$ and other symbols becomes just a matter of adding the extra subscript ℓ (e.g., $D_s^2 = D^2_{s,\ell}$, $\xi_t = \xi_{t,\ell}$, $k_t = k_{t,\ell}$, $\beta_t = \beta_{t,\ell}$, $\hat{\theta}_t = \hat{\theta}_{t,\ell}$, etc.)

Theorem B.1. *Take Assumption A3 with the added assumption of $\nu_t \in [-R, R]$. Define*

$$K_t(\theta) = \sum_{s=1}^{t} f_s(\theta) + \sum_{s=1}^{t} \frac{1}{2}(z_s^\top(\theta - \hat{\theta}_{s-1}))^2 + \frac{\lambda}{2}\|\theta\|^2$$
$$\textit{and} \quad \bar{\theta}_t = \arg\min_{\theta} K_t(\theta) \ .$$

Furthermore, define $\overline{\Sigma}_t = 2\sum_{s=1}^{t} z_s z_s^\top + \lambda I$.

Let $D_s^2 := \|z_s\|^2_{\Sigma_s^{-1}}$ and define β_t recursively as follows:

$$\beta_t = L_t(\hat{\theta}_t) - K_t(\bar{\theta}_t) + \frac{\lambda}{2}S^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1})D_s^2 + \rho\sqrt{\bar{\beta}_{t-1}}\sqrt{8\left(\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) + R^2\ln(1/\delta)\right)\xi_t} + 2^{k_t}\rho R\sqrt{2\beta_0}\xi_t$$

where $\bar{\beta}_{t-1} = \max_{s=1}^{t-1}\beta_s$, $\beta_0 = \frac{\lambda}{2}S^2$, $\xi_t := \ln(\sqrt{\pi(t+1)} \cdot \frac{3.4 \cdot k_t \ln^2(1+k_t)}{\delta})$ with $k_t = 1 \vee \lceil \log_2(\sqrt{\bar{\beta}_{t-1}/\beta_0})\rceil$. Define the confidence set

$$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \frac{1}{2}\|\theta - \bar{\theta}_t\|^2_{\overline{\Sigma}_t} \le \beta_t\right\} \ .$$

Assume $\forall s \ge 1, D_s^2 \le \rho^2$ for some $\rho^2 > 0$. Then,

$$\mathbb{P}(\forall t \ge 1, \theta^* \in \mathcal{C}_t) \ge 1 - 2\delta$$

Proof. This proof is inspired by Zhao et al. (2023b), but details differ since we leverage the regret equality (Lemma C.7), which helps shorten the proof and provides a numerically tight derivation.

To describe the plan, we will show $\bar{M}_t^2 := \frac{1}{2}\|\theta^* - \bar{\theta}_t\|^2_{\overline{\Sigma}_t} \le \beta_t, \forall t \ge 0$ under an event $\mathcal{E}_1$ (defined below) that holds with probability at least $1 - \delta$. For this, we use induction. Define $I_s = \mathbb{1}\left\{\bar{M}_s^2 \le \beta_s\right\}$ and let $\bar{I}_s = 1 - I_s$.

First, we show the base case of $I_0 = 1$. This is trivial since $\bar{M}_0^2 = \frac{1}{2}\|\theta^*\|^2_{\lambda I} \le \frac{\lambda}{2}S^2 = \beta_0$. It remains to prove that, assuming $I_0 = \cdots = I_{t-1} = 1$, we have $I_t = 1$.

So, let us assume the inductive hypothesis $I_0 = \cdots = I_{t-1} = 1$. Define $r_t = z_t^\top(\hat{\theta}_{t-1} - \theta^*)$, which implies that $f_s(\hat{\theta}_{s-1}) - f_s(\theta^*) = \frac{1}{2}r_s^2 - r_s\nu_s$. Recall the regret equality (Lemma C.7):

$$\begin{aligned}
M_t^2 := \frac{1}{2}\|\theta^* - \hat{\theta}_t\|^2_{\Sigma_t} &= \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1})D_s^2 - \sum_{s=1}^{t}\left(f_s(\hat{\theta}_{s-1}) - f_s(\theta^*)\right) \\
&= \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1})D_s^2 + \sum_{s=1}^{t}(-\frac{1}{2}r_s^2) + \sum_{s=1}^{t} r_s\nu_s \\
\implies M_t^2 + \sum_{s=1}^{t}\frac{1}{2}r_s^2 &\le \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1})D_s^2 + \sum_{s=1}^{t} r_s\nu_s \ .
\end{aligned}$$

Note that

$$\sum_{s=1}^{t} r_s\nu_s = \sum_{s=1}^{t} I_{s-1}r_s\nu_s + \sum_{s=1}^{t}\bar{I}_{s-1}r_s\nu_s = \sum_{s=1}^{t} I_{s-1}r_s\nu_s \qquad \text{(inductive hypothesis)}$$

We assume the event in Corollary C.5 with

- $b_k = 2^k\rho R\sqrt{2\beta_0}$.
- $X_s = I_{s-1}r_s\nu_s$,

which holds with probability at least $1-\delta$. Recall that $k_t = 1 \vee \lceil \log_2(\sqrt{\bar{\beta}_{t-1}/\beta_0}) \rceil$, which means that $k_t = \min\{k \in \mathbb{N}_+ : b_k \geq \rho R\sqrt{2\bar{\beta}_{t-1}}\}$. Then, since

$$\begin{aligned}
|X_s| \leq RI_{s-1}|r_s| &\leq R\|z_s\|_{\overline{\Sigma}_{s-1}^{-1}} \|\bar{\theta}_{s-1} - \theta^*\|_{\overline{\Sigma}_{s-1}} && \text{(Cauchy-Schwarz)} \\
&\leq R\|z_s\|_{\overline{\Sigma}_{s-1}^{-1}} \sqrt{2\beta_{s-1}} && (I_{s-1}=1) \\
&\leq R\rho\sqrt{2\beta_{s-1}} && \text{(assumption in the theorem)} \\
&\leq R\rho\sqrt{2\bar{\beta}_{t-1}} \ \leq b_{k_t} \ .
\end{aligned}$$

Thus, we have $\overline{(X_s)}_{b_k} = X_s, \forall s \in [t]$, so we have

$$\forall t \geq 1, \sum_{s=1}^{t} I_{s-1} r_s \nu_s \leq \sqrt{2\sum_{s=1}^{t} I_{s-1} r_s^2 \nu_s^2 \xi_t} + 2^{k_t}\rho R\sqrt{2\beta_0}\xi_t \ . \tag{10}$$

Note that, with a similar derivation as the bound on $|X_s|$, we have

$$I_{s-1} r_s^2 \leq 2\rho^2 \bar{\beta}_{t-1} \ .$$

Then, using Lemma C.8, with probability at least $1-\delta$, we have

$$\forall t \geq 1, \sum_{s=1}^{t} \nu_s^2 = \sum_{s=1}^{t} 2f_s(\theta^*) = 2\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) + 2\sum_{s=1}^{t} f_s(\theta^*) - f_s(\hat{\theta}_{s-1}) \leq 2\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) + 2R^2\ln(1/\delta) \ .$$

Thus,

$$M_t^2 + \sum_{s=1}^{t} \frac{1}{2} r_s^2 \leq \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 + \sqrt{8\rho^2\bar{\beta}_{t-1}\left(\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) + R^2\ln(1/\delta)\right)\xi_t} + 2^{k_t}\rho R\sqrt{2\beta_0}\xi_t =: \beta_t \ .$$

Recall that $L_t(\theta) = \sum_{s=1}^{t} f_s(\theta) + \frac{\lambda}{2}\|\theta\|^2$. Since $M_t^2 = L_t(\theta^*) - L_t(\hat{\theta}_t)$ (verify this with Taylor's theorem), the LHS above can be rewritten as

$$\begin{aligned}
&\sum_{s=1}^{t} f_s(\theta^*) + \frac{\lambda}{2}\|\theta^*\|^2 - L_t(\hat{\theta}_t) + \sum_{s=1}^{t} \frac{1}{2}(z_s^\top(\hat{\theta}_{s-1} - \theta^*))^2 \\
&= K_t(\theta^*) - L_t(\hat{\theta}_t) \\
&= K_t(\theta^*) - K_t(\bar{\theta}_t) + K_t(\bar{\theta}_t) - L_t(\hat{\theta}_t) \\
&= \frac{1}{2}\|\theta^* - \bar{\theta}_t\|_{\overline{\Sigma}_t}^2 + K_t(\bar{\theta}_t) - L_t(\hat{\theta}_t)
\end{aligned}$$

Rearranging the terms and using the bound $\|\theta^*\|^2 \leq S^2$ prove $I_t = 1$, which completes the inductive proof. □

Next, we find a nondecreasing upper bound on β_t that directly depends on the true variances $\{\sigma_s^2\}_{s=1}^t$.

Lemma B.2. *Take the assumptions of Theorem B.1. Suppose $\lambda = \frac{R^2}{S^2}\rho^2\lambda_0$ for some absolute constant $\lambda_0 > 0$. Then, with probability at least $1 - O(\delta)$,*

$$\forall t \geq 1, \bar{\beta}_t \leq \beta_t^* := c\rho^2(R^2 + \sum_{s=1}^{t} \sigma_s^2)\ln^2(t/\delta)$$

for some absolute constant $c > 0$.

Proof. Using the definition of k_t, we have

$$\begin{aligned}\beta_t &\le L_t(\hat{\theta}_t) - K_t(\bar{\theta}_t) \\ &\quad + \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^t f_s(\hat{\theta}_{s-1})D_s^2 + \sqrt{8\rho^2\bar{\beta}_{t-1}\left(\sum_{s=1}^t f_s(\hat{\theta}_{s-1}) + R^2\ln(1/\delta)\right)}\,\xi_t + \rho R\sqrt{8\bar{\beta}_{t-1}}\xi_t\,.\end{aligned}$$

First, we notice that we can easily bound the first two terms in β_t:

$$L_t(\hat{\theta}_t) - K_t(\bar{\theta}_t) \le L_t(\bar{\theta}_t) - K_t(\bar{\theta}_t) \le 0\,.$$

Furthermore,

$$2^{k_t}\rho\sqrt{2\beta_0}\xi_t \le \rho\sqrt{8\bar{\beta}_{t-1}}\xi_t\,.$$

Let $A_t = \sum_{s=1}^t f_s(\hat{\theta}_{s-1})$. We use $D_s^2 \le \rho^2$ and the Fenchel-Young inequality $xy \le \frac{1}{2a}x^2 + \frac{a}{2}y^2, \forall a > 0$ to obtain

$$\begin{aligned}\beta_t &\le \frac{\lambda}{2}S^2 + \rho^2 A_t + 8\rho^2(A_t + R^2\ln(1/\delta))\xi_t + \frac{1}{4}\bar{\beta}_{t-1} + 8\rho^2R^2\xi_t^2 + \frac{1}{4}\bar{\beta}_{t-1} \\ &\le c\lambda S^2 + c\rho^2\xi_t(A_t + R^2\xi_t) + \frac{1}{2}\bar{\beta}_{t-1}\end{aligned}$$

for some absolute constant $c > 0$.

Let $\xi = \ln(1/\delta)$. We bound A_t with probability at least $1 - O(\delta)$ as follows:

$$\begin{aligned}A_t &= \sum_{s=1}^t f_s(\theta^*) + \sum_{s=1}^t (f_s(\hat{\theta}_{s-1}) - f_s(\theta^*)) \\ &\lesssim \sum_{s=1}^t \sigma_s^2 + \sum_{s=1}^t r_s^2 + R^2\xi && \text{(Lemma C.12 and Lemma C.9)} \\ &\lesssim \sum_{s=1}^t \sigma_s^2 + \lambda S^2 + \rho^2\sum_s \nu_s^2 + R^2\xi && \text{(Lemma C.11; } D_s^2 \le \rho^2\text{)} \\ &\lesssim \sum_{s=1}^t \sigma_s^2 + \lambda S^2 + R^2\xi\,. && \text{(Lemma C.12; } \rho^2 \le 1\text{)}\end{aligned}$$

Thus, there exists an absolute constant $c_1, c_2 > 0$ such that, letting $Q_t = \lambda S^2 + R^2 + \sum_{s=1}^t \sigma_s^2$,

$$\begin{aligned}\beta_t &\le c_1\lambda S^2 + c_1\rho^2\xi_t^2 Q_t + \frac{1}{2}\bar{\beta}_{t-1} \\ &\le \underbrace{c_2\lambda S^2 + c_2\rho^2 Q_t\left(\ln^2(t/\delta) + \overline{\ln\ln}^2(\sqrt{\bar{\beta}_{t-1}/\beta_0})\right)}_{=:\,\beta_t'} + \frac{1}{2}\bar{\beta}_{t-1}\end{aligned}$$

for some absolute constant $c_2 > e$. We define $\beta_0' = \frac{\lambda}{2}S^2$. Note that β_t' can be compactly written as

$$\beta_t' = a_t + b_t\,\overline{\ln\ln}^2(\sqrt{\bar{\beta}_{t-1}/\beta_0})$$

where

$$\begin{aligned}a_t &:= c_2(\lambda S^2 + \rho^2 Q_t\ln^2(t/\delta)) \\ b_t &:= c_2\rho^2 Q_t\,.\end{aligned}$$

Let

$$\beta_t^\dagger = a_t + 2b_t(e + \overline{\ln\ln}^2(2a_t/\beta_0) + \overline{\ln\ln}^2(2b_t/\beta_0)) \ .$$

Note that $\beta_t' \le \beta_t^\dagger$ and that $\beta_t^\dagger$ is a nondecreasing function of t.

We claim that $\beta_t \le 2\beta_t^\dagger, \forall t \ge 0$. We use induction. First, we trivially have $\beta_0 = \frac{\lambda}{2}S^2 \le \beta_0^\dagger$ using $c_2 \ge e$.

For the general case, assume $\beta_{t-1} \le 2\beta_{t-1}^*$. If $\bar{\beta}_{t-1} \le e^2\beta_0$, then the $\overline{\ln\ln}^2(\cdot)$ term is 0, so we trivially have that $\beta_t' \le \beta_t^\dagger$ and thus

$$\beta_t \le \beta_t^\dagger + \frac{1}{2}\bar{\beta}_{t-1} \le \beta_t^\dagger + \beta_{t-1}^\dagger \le 2\beta_t^\dagger \ .$$

If $\bar{\beta}_{t-1} > e^2\beta_0$, then we use Lemma C.6 with $x = \bar{\beta}_{t-1}/\beta_0$, $a = a_t/\beta_0$, and $b = b_t/\beta_0$ to derive

$$\begin{aligned} a + b\ln^2\ln(\sqrt{\bar{\beta}_{t-1}/\beta_0}) &\le a + b\ln^2\ln(\bar{\beta}_{t-1}/\beta_0) && (\bar{\beta}_{t-1}/\beta_0 \ge e^2 \ge 1) \\ &\le \bar{\beta}_{t-1}/\beta_0 \ . && (\text{Lemma C.6; } \bar{\beta}_{t-1}/\beta_0 \ge e^2 \ge e) \end{aligned}$$

Thus,

$$\begin{aligned} \beta_t' &= a_t + b_t\,\overline{\ln\ln}^2(\sqrt{\bar{\beta}_{t-1}/\beta_0}) \\ &\le \bar{\beta}_{t-1} \le \beta_{t-1}^\dagger \le \beta_t^\dagger \ , \end{aligned}$$

which implies $\beta_t \le 2\beta_t^\dagger$. Observing $\forall s \in [t], \beta_s \le 2\beta_s^\dagger \implies \bar{\beta}_t \le 2\beta_t^\dagger$ concludes the claim.

Finally, we need to show that $\beta_t^\dagger \le \beta_t^*$. First, note that $b_t \lesssim a_t$, so $\overline{\ln\ln}^2(2b_t/\beta_0) \le \overline{\ln\ln}^2(2a_t/\beta_0)$. Furthermore, using the fact that $\sigma_s^2 \le R^2$, one can show that $a_t/\beta_0 \lesssim 1 + \frac{t}{\lambda_0} \lesssim t\ln^2(t/\delta)$ since λ_0 is an absolute constant. Thus, $b_t(e + \overline{\ln\ln}^2(2a_t/\beta_0)) \lesssim \rho^2 Q_t\,\overline{\ln\ln}^2(t\ln^2(t/\delta)) \lesssim \lambda S^2 + \rho^2 Q_t \ln^2(t/\delta) \lesssim a_t$. This implies that $\beta_t^\dagger \lesssim a_t$, and one can easily show that $a_t \lesssim \beta_t^*$. □

B.2. Proof of Theorem 3.2

By Lemma B.2, we have that $\beta_{t,\ell} \le \beta_{t,\ell}^* = c\rho_\ell^2(R^2 + \sum_{s=1}^t \sigma_s^2)\psi_t^2$ where $\psi_t^2 = \ln^2(t/\delta)$. Defining $\beta_{t,0}^* = c\psi_t^2(R^2 + \sum_{s=1}^t \sigma_s^2)$, we can write down

$$\beta_{t,\ell}^* = 2^{-2\ell}\beta_{t,0}^* \ .$$

Define $t' = t-1$ and $\tilde{\theta}_{t,\ell} = \arg\max_{\theta \in C_{t',\ell}} \langle x_t, \theta\rangle$. Define $\text{reg}_t := \langle x_{t,*} - x_t, \theta^*\rangle$, which is the instantaneous regret at time t. We first show the elementary bound that is useful throughout:

$$\begin{aligned} \text{reg}_t &:= \langle x_{t,*} - x_t, \theta^*\rangle && (x_{t,*} := \max_{x\in\mathcal{X}_t}\langle x, \theta^*\rangle) \\ &\le \min_{\ell \in [\ell^*+1..L]} \langle x_t, \tilde{\theta}_{t,\ell} - \theta^*\rangle \\ &= \min_{\ell \in [\ell^*+1..L]} \langle x_t, \tilde{\theta}_{t,\ell} - \bar{\theta}_{t',\ell} + \bar{\theta}_{t',\ell} - \theta^*\rangle \\ &\le \min_{\ell \in [\ell^*+1..L]} \sqrt{8}\|x_t\|_{\overline{\Sigma}_{t',\ell}^{-1}}\sqrt{\beta_{t',\ell}^*} \\ &\le \min_{\ell \in [\ell^*+1..L]} \sqrt{8}\|x_t\|_{\Sigma_{t',\ell}^{-1}} 2^{-\ell}\sqrt{\beta_{n,0}^*} \ . && (\overline{\Sigma}_{t',\ell} \succeq \Sigma_{t',\ell}; \text{ monotonicity of } \beta_{t,0}^*) \end{aligned}$$

Also note the trivial upper bound: $\text{reg}_t \le 2B$.

We classify the time steps $[n]$ into the following three:

- $\mathcal{T}_1 := \{t \in [n] : \text{reg}_t > \sqrt{8\beta_{n,0}^*}\}$

- $\mathcal{T}_2 := \{t \in [n] : \mathrm{reg}_t \le 2^{-2L}\sqrt{8\beta^*_{n,0}}\}$
- $\mathcal{T}_3 := \left\{t \in [n] : \exists \ell \in [1..L] \text{ s.t. } \mathrm{reg}_t \in \left(2^{-2\ell}\sqrt{8\beta^*_{n,0}},\ \cdot 2^{-2(\ell-1)}\sqrt{8\beta^*_{n,0}}\right]\right\}$.

Hereafter, we define $\Sigma_{t,\ell}[\mathcal{T}] := \lambda_\ell I + \sum_{s\in[t]\cap\mathcal{T}} w^2_{s,\ell} x_s x_s^\top$.

Case 1. $t \in \mathcal{T}_1$
We have

$$\sqrt{8\beta^*_{n,0}} < \mathrm{reg}_t \le \|x_t\|_{\Sigma^{-1}_{t',1}} 2^{-1}\sqrt{8\beta^*_{n,0}} \implies \|x_t\|_{\Sigma^{-1}_{t',1}} \ge 2 .$$

This means that $\frac{2^{-1}}{\|x_t\|_{\Sigma^{-1}_{t',1}}} \le \frac{1}{4}$, so $w_{t,1} < 1$. Then, we have $\|w_{t,1}x_t\|_{\Sigma^{-1}_{t',1}} = \frac{1}{2}$. Thus, using $\mathrm{reg}_t \le 2B$,

$$\begin{aligned}
\sum_{t\in\mathcal{T}_1} \mathrm{reg}_t &\le \sum_{t\in\mathcal{T}_1} \mathbb{1}\left\{\|w_{t,1}x_t\|_{\Sigma^{-1}_{t',1}} = \frac{1}{2}\right\} 2B \\
&\le \sum_{t\in\mathcal{T}_1} \mathbb{1}\left\{\|w_{t,1}x_t\|_{\Sigma^{-1}_{t',1}[\mathcal{T}_1]} \ge \frac{1}{2}\right\} 2B \\
&\lesssim Bd\ln(1+\frac{S^2}{R^2}) . && \text{(EPC (Lemma C.2))}
\end{aligned}$$

Case 2. $t \in \mathcal{T}_2$
Since $\mathrm{reg}_t \le 2^{-2L}\sqrt{8\beta^*_{n,0}}$,

$$\sum_{t\in\mathcal{T}_2} \mathrm{reg}_t \le \sum_{t\in\mathcal{T}_2} 2^{-2L}\sqrt{8\beta^*_{n,0}} \le n\cdot 2^{-2L}\sqrt{8\beta^*_{n,0}} .$$

Case 3. $t \in \mathcal{T}_3$
Define $\mathcal{T}_{3,\ell} = \{t \in \mathcal{T}_3 : \ell_t = \ell\}$. Note that

$$\begin{aligned}
2^{-2\ell}\sqrt{8\beta^*_{n,0}} < \mathrm{reg}_t \le \|x_t\|_{\Sigma^{-1}_{t',\ell}} 2^{-\ell}\sqrt{8\beta^*_{n,0}} \quad &\implies \quad 2^{-\ell} \le \|x_t\|^2_{\Sigma^{-1}_{t',\ell}} \\
&\implies \quad w_{t,\ell} = \frac{2^{-\ell}}{\|x_t\|_{\Sigma^{-1}_{t',\ell}}} . \qquad (11)
\end{aligned}$$

That is, $w_{t,\ell}$ can be written without '$\wedge 1$' from its definition. Using this,

$$\begin{aligned}
\sum_{t\in\mathcal{T}_3} \mathrm{reg}_t &= \sum_{\ell=1}^{L}\sum_{t\in\mathcal{T}_{3,\ell}} \mathrm{reg}_t \\
&= \sum_{\ell=1}^{L}\sum_{t\in\mathcal{T}_{3,\ell}} \mathbb{1}\left\{\|w_{t,\ell}x_t\|_{\Sigma^{-1}_{t',\ell}} = 2^{-\ell}\right\} \mathrm{reg}_t && \text{(by (11))} \\
&\le \sum_{\ell=1}^{L}\sum_{t\in\mathcal{T}_{3,\ell}} \mathbb{1}\left\{\|w_{t,\ell}x_t\|_{\Sigma^{-1}_{t',\ell}} = 2^{-\ell}\right\} 2^{-2(\ell-1)}\sqrt{8\beta^*_{n,0}} && (t \in \mathcal{T}_{3,\ell}) \\
&\le \sum_{\ell=1}^{L}\sum_{t\in\mathcal{T}_{3,\ell}} \mathbb{1}\left\{\|w_{t,\ell}x_t\|_{\Sigma^{-1}_{t',\ell}[\mathcal{T}_{3,\ell}]} \ge 2^{-\ell}\right\} 2^{-2(\ell-1)}\sqrt{8\beta^*_{n,0}} \\
&= \sum_{\ell=1}^{L} 2^{2\ell} d\ln\left(1+\frac{2^{2\ell}S^2}{R^2}\right) 2^{-2(\ell-1)}\sqrt{8\beta^*_{n,0}} && \text{(EPC (Lemma C.2))} \\
&\lesssim L\cdot d\ln\left(1+\frac{4^L S^2}{R^2}\right)\sqrt{\beta^*_{n,0}} .
\end{aligned}$$

Altogether, we have

$$\sum_{t=1}^{n} \mathrm{reg}_t \lesssim Bd \ln\left(1+\frac{S^2}{R^2}\right) + \left(n2^{-2L} + L \cdot d \ln\left(1+\frac{4^L S^2}{R^2}\right)\right)\sqrt{\beta^*_{n,0}}\,.$$

Using the definition $L = 1 \vee \lceil \frac{1}{2}\log_2(n/d)\rceil$,

$$\mathrm{Reg}_n \lesssim \ln(n/d)\cdot d\sqrt{\left(R^2+\sum_{t=1}^{n}\sigma_t^2\right)\ln^2(n/\delta)} + Bd\ln(1+\frac{S^2}{R^2})\,.$$

C. Utility Lemmas

The following lemma is the well known elliptical potential lemma from linear algebra.

Lemma C.1. *(Elliptical potential lemma; e.g., Hazan et al. (2007, Lemma 11)) Let $x_1,\dots,x_t \in \mathbb{R}^d$ be a sequence of vectors with $\|x_s\|_2 \le X, \forall s \in [t]$ for some $X > 0$. Let $V_t = \lambda I + \sum_{s=1}^t \sum_{s=1}^t x_s x_s^\top$ for some $\tau > 0$. Let $|A|$ be the determinant of the matrix A. Then,*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \le \ln\left(\frac{|V_t|}{|\lambda I|}\right)$$

and $\ln\left(\frac{|V_t|}{|\lambda I|}\right) \le d\ln\left(1+\frac{X^2 t}{d\lambda}\right)$.

Proof. The following is a well-known identity (e.g., see the proof of Abbasi-Yadkori et al. (2011, Lemma 11)) :

$$\sum_{s=1}^{t} \ln(1+\|x_t\|^2_{V_{t-1}^{-1}}) = \ln\left(\frac{|V_t|}{|\lambda I|}\right)\,.$$

We now lower bound the left-hand side above. Letting $D_s^2 = \|x_s\|^2_{V_s^{-1}}$, we have

$$\begin{aligned}
\sum_{s=1}^{t} \ln(1+\|x_s\|^2_{V_{s-1}^{-1}}) &= \sum_{s=1}^{t} \ln\left(1+\frac{D_s^2}{1-D_s^2}\right) && \text{(Woodbury matrix identity)}\\
&= \sum_{s=1}^{t} \ln\left(\frac{1}{1-D_s^2}\right)\\
&\ge \sum_{s=1}^{t} D_s^2\,. && (\ln(1+x)\le x, \forall x)
\end{aligned}$$

□

The following lemma is a simplified version of the elliptical potential count lemma (Kim et al., 2022, Lemma 4) by using $\ln(1+x) \ge \ln(2)x, \forall x \in [0,1]$, which is a generalization of Lattimore & Szepesvári (2020, Exercise 19.3).

Lemma C.2. *(Elliptical potential count; Kim et al. (2022, Lemma 4) and Lattimore & Szepesvári (2020, Exercise 19.3)) Let $x_1,\dots,x_t \in \mathbb{R}^d$ be a sequence of vectors with $\|x_s\|_2 \le X, \forall s \in [t]$ for some $X > 0$. Let $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$ for some $\lambda > 0$. Let $J = \{s \in [t] : \|x_s\|^2_{V_{s-1}^{-1}} \ge L^2\}$ for some $L^2 \le 1$. Then,*

$$|J| \le 3\frac{d}{L^2}\ln\left(1+\frac{2X^2}{L^2\lambda}\right)\,.$$

Lemma C.3. *Let $(X_t)_{t=1}^\infty$ be a sequence of random variables adapted to filtration $(\mathcal{G}_t)_{t=0}^\infty$. Assume $X_t \mid \mathcal{G}_{t-1}$ is σ_t^2-sub-Gaussian. Then, for any $a > 0$, we have*

$$1-\delta \le \mathbb{P}\left(\forall t \ge 1, \sum_{s=1}^{t} X_s \le \frac{a}{2}\sum_{s=1}^{t}\sigma_s^2 + \frac{1}{a}\ln(1/\delta)\right)\,.$$

Proof. Define $M_t = \exp(a \sum_{s=1}^t X_s - \frac{a^2}{2} \sum_{s=1}^t \sigma_s^2)$ and verify that M_t is a supermartingale. Applying Ville's inequality (Ville, 1939) concludes the proof. □

Lemma C.4. *Let $(X_t)_{t=1}^\infty$ be a sequence of random variables adapted to a filtration $(\mathcal{G}_t)_{t=1}^\infty$ such that $X_t \mid \mathcal{G}_{t-1} \in [-R, R]$ and $\mathbb{E}[X_t \mid \mathcal{G}_{t-1}] = 0$ almost surely. Then,*

$$1 - \delta \le \mathbb{P}\left(\forall t \ge 1, \forall \alpha \in [-1/R, 1/R], \sum_{s=1}^t \ln(1 + \alpha X_s) \le \ln\left(\frac{\sqrt{\pi(t+1)}}{\delta}\right)\right) . \tag{12}$$

Furthermore, let $q(x) := \frac{-\ln(1-x)-x}{x^2}$ and $\psi_t := \ln(\sqrt{\pi(t+1)}/\delta)$. Then,

$$1 - \delta \le \mathbb{P}\left(\forall t \in \mathbb{N}_+, \forall \alpha \in [0, 1/R), \left|\sum_{s=1}^t X_s\right| \le q(R\alpha)\alpha \sum_{s=1}^t X_s^2 + \frac{1}{\alpha}\psi_t\right) .$$

Finally, tuning α implies

$$1 - \delta \le \mathbb{P}\left(\forall t \ge 1, \left|\sum_{s=1}^t X_s\right| \le \sqrt{2 \sum_{s=1}^t X_s^2 \psi_t} + R\psi_t\right) .$$

Proof. To show the third statement, we choose $\alpha = \left(R + \sqrt{\frac{\sum_{s=1}^t X_s^2}{\psi_t}}\right)^{-1}$. One can similarly obtain an upper bound on $\sum_{s=1}^t -X_s$.

The proof of the first statement requires modifying Orabona & Jun (2023, Theorem 1) in two ways: we deal with (i) a sequence of random variables that are not necessarily i.i.d. and (ii) the range of X_t that is $[-R, R]$ rather than $[0, 1]$.

The rest of the proof requires the background on the coin betting problem; with the full details, refer to Orabona & Jun (2023). Suppose we have an algorithm $\mathcal{A}$ with the initial wealth of $W_0 = 1$. At each time step t, the algorithm $\mathcal{A}$ commits to a betting fraction $\alpha_t \in [-1/R, 1/R]$ based on the past observations $c_1, \dots, c_{t-1} \in [-R, R]$. Once the coin outcome $c_t \in [-R, R]$ is revealed, the wealth of $\mathcal{A}$ denoted by W_t becomes $W_t = (1 + \alpha_t c_t) W_{t-1}$. Suppose we set $c_t = X_t$. Since (W_t) forms a nonnegative martingale, we can apply Ville's inequality (Ville, 1939) to have

$$1 - \delta \le \mathbb{P}\left(\forall t \ge 1, \ln(W_t) < \ln(1/\delta)\right) . \tag{13}$$

To maximize the wealth, it is natural to choose an algorithm that has a small log-wealth regret w.r.t. the best betting fraction $\alpha \in [-1/R, 1/R]$ in hindsight, which is defined as

$$\text{Regret}_t := \max_{\alpha \in [-1/R, 1/R]} \sum_{s=1}^t \ln(1 + \alpha X_t) - \ln(W_t) . \tag{14}$$

One can construct an efficient $\mathcal{A}$ via a reduction of coin betting problem to the two-stock online portfolio problem. In this problem, an algorithm $\mathcal{B}$ starts from wealth $W_0^{\text{portfolio}} = 1$. At each time step t, the algorithm $\mathcal{B}$ determines the fraction $(b_t, 1 - b_t)$, which means that it will invest b_t fraction of the current wealth $W_{t-1}^{\text{portfolio}}$ to the first stock and the rest to the second stock. Then, the price change ratios $(w_{t,1}, w_{t,2})$ are revealed, and the wealth of $\mathcal{B}$ changes: $W_t^{\text{portfolio}} = b_t w_{t,1} W_{t-1}^{\text{portfolio}} + (1 - b_t) w_{t,2} W_{t-1}^{\text{portfolio}}$.

Similar to Orabona & Jun (2023, Lemma 1), one can easily construct a reduction of coin betting to online portfolio: given a coin outcome $c_t \in [-\ell, u]$, we can set the price change ratio of the first stock and the second stock as $w_{t,1} = 1 + \frac{c_t}{\ell}$ and $w_{t,2} = 1 - \frac{c_t}{u}$ respectively. The online portfolio algorithm $\mathcal{B}$ will then produce the next investment fraction $(b_{t+1}, 1 - b_{t+1})$. We can then set α_{t+1} in the coin betting problem as $\alpha_{t+1} = -\frac{1}{u} + b_{t+1}\left(\frac{1}{\ell} + \frac{1}{u}\right)$. Performing this conversion at every iteration t with $\ell = u = R$ satisfies that $W_t = W_t^{\text{portfolio}}$. This implies that coin betting is a special case of the online portfolio problem, and the log-wealth regret in online portfolio coincides with the log-wealth regret in coin betting. Our choice of online portfolio algorithm is universal portfolio equipped with the Dirichlet$(\frac{1}{2}, \frac{1}{2})$ prior (Cover & Ordentlich, 1996). This gives the regret bound of $\text{Regret}_t \le \ln \frac{\sqrt{\pi}\Gamma(t+1)}{\Gamma(t+\frac{1}{2})} \le \ln(\sqrt{\pi(t+1)})$.

Then, the event in (13) implies

$$\ln(1/\delta) > \ln(W_t) = \max_{\alpha\in[-1/R,1/R]} \sum_{s=1}^{t} \ln(1+\alpha X_t) - \text{Regret}_t \geq \max_{\alpha\in[-1/R,1/R]} \sum_{s=1}^{t} \ln(1+\alpha X_t) - \ln(\sqrt{\pi(t+1)})$$

This completes the proof of (12).

To prove the second statement, assume the event inside the probability probability statement inside (12). We use the following inequality from Fan et al. (2015, Eq. 4.11):

$$\forall \lambda \in [0,1), \xi \geq -1, \ln(1+\lambda\xi) \geq \lambda\xi + \xi^2(\lambda + \ln(1-\lambda)) \ .$$

For $\alpha \in [0, 1/R)$, we apply this inequality with $\lambda = R\alpha$ and $\xi = X_s/R$ to obtain

$$\sum_{s=1}^{t} X_s \leq q(R\alpha)\alpha \sum_{s=1}^{t} X_s^2 + \frac{1}{\alpha}\psi_t \ .$$

Obtaining an upper bound on $\sum_{s=1}^{t} X_s$ is symmetric to the proof above.

□

Corollary C.5. *Define the clipping operator* $\overline{(x)}_y = (\frac{y}{|x|} \wedge 1)x$ *if* $x \neq 0$ *and* $\overline{(x)}_y = 0$ *otherwise. Let* $(b_k)_{k=1}^{\infty}$ *be a sequence of positive integers. Then, under the same setting of Lemma C.4, with probability at least* $1-\delta$*,*

$$\forall k \in \mathbb{N}_+, \forall t \geq 1, \sum_{s=1}^{t} \overline{(X_s)}_{b_k} - \mathbb{E}[\overline{(X_s)}_{b_k} \mid \mathcal{G}_{s-1}] \leq \sqrt{2\sum_{s=1}^{t}(\overline{(X_s)}_{b_k} - \mathbb{E}[\overline{(X_s)}_{b_k} \mid \mathcal{G}_{s-1}])^2 \xi_{t,k} + b_k \xi_{t,k}}$$

where $\xi_{t,k} = \ln(\sqrt{\pi(t+1)} \cdot \frac{3.39 \cdot k \ln^2(1+k)}{\delta})$.

Proof. The statement follows by a simple union bound argument applied to Lemma C.4. □

Lemma C.6. *Define* $\overline{\ln\ln}(x) := \ln\ln(e \vee x)$ *and let* $\overline{\ln\ln}^2(x) = (\overline{\ln\ln}(x))^2$*. Let* $x \geq e$*. Throughout, we take* $\ln\ln(x) = 0$ *for* $x \leq e$*. Then, for* $a, b > 0$*,*

$$x \geq a + 2b(e + \overline{\ln\ln}^2(2a) + \overline{\ln\ln}^2(2b)) \implies x \geq a + b\ln^2\ln(x) \ .$$

Proof. We prove the contraposition:

$$x < a + b\ln^2\ln(x) \implies x < a + 2b(e + \overline{\ln\ln}^2(2a) + \overline{\ln\ln}^2(2b)) \ .$$

If $x < a + be$, then the statement follows trivially.

If $x \geq a + be$, then $z := \frac{x-a}{b} \geq e$. Then,

$$\begin{aligned} z &< \ln^2\ln(bz + a) \\ &\leq \ln^2\ln(2bz \vee 2a)) \\ &\leq \overline{\ln\ln}^2(2bz) \vee \overline{\ln\ln}^2(2a) \end{aligned}$$

where the last line can be shown by the case-by-case reasoning on whether $2bz \geq e$ or not and whether $2a \geq e$ or not.

If $2bz \geq e$, then

$$\begin{aligned} \overline{\ln\ln}^2(2bz) &\leq \ln^2(2\ln(z) \vee 2\ln(2b)) \\ &\overset{(a)}{\leq} \ln^2(2\ln(z)) + 2\ln^2(2) + 2\overline{\ln\ln}^2(2b) \qquad (z \geq e) \end{aligned}$$

$$\leq \frac{z}{2} + 1 + \overline{\ln\ln}^2(2b) \ . \qquad (z \geq e;\, 2\ln^2(2) \leq 1)$$

If $2bz < e$, then the inequality above is also true, trivially.

The two displays above imply

$$z < 2(1 + \overline{\ln\ln}^2(2b) + \overline{\ln\ln}^2(2a)) \ .$$

Using definition of z concludes the proof. □

C.1. Online Ridge Regression

The following regret equality provides an essential tool for analyses in this paper.

Lemma C.7. *(Regret Equality; Orabona (2023, Lemma 7.1)) Take Assumption A3 except for the stochastic modeling of u_t. Let $\Sigma_t = \lambda I + \sum_{s=1}^t z_s z_s^\top$. Then,*

$$\sum_{s=1}^t f_s(\hat{\theta}_{s-1}) - f_s(\theta^*) = \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^t f_s(\hat{\theta}_{s-1})\|z_s\|^2_{\Sigma_t^{-1}} - \frac{1}{2}\|\theta^* - \hat{\theta}_t\|^2_{\Sigma_t} \ .$$

Lemma C.8. *Take Assumption A3. Then,*

$$1 - \delta \leq \mathbb{P}\left(\sum_{s=1}^t (f_s(\theta^*) - f_s(\hat{\theta}_{s-1})) \leq \sigma_*^2 \ln(1/\delta)\right) \ .$$

Proof. Let $r_s := z_s^\top(\hat{\theta}_{s-1} - \theta^*)$. We have the following identity.

$$f_s(\hat{\theta}_{s-1}) - f_s(\theta^*) = \frac{1}{2}r_s^2 - r_s\nu_s \ . \tag{15}$$

Using (15) and the sub-Gaussianity of $(\nu_s)_{s=1}^t$, we can use Lemma C.3 to have that, with probability at least $1-\delta$,

$$\forall t \geq 1, \sum_{s=1}^t f_s(\theta^*) - f_s(\hat{\theta}_{s-1}) \leq \frac{a}{2}\sum_{s=1}^t r_s^2\sigma_*^2 + \frac{1}{a}\ln(1/\delta) - \frac{1}{2}\sum_{s=1}^t r_s^2 \ .$$

Choosing $a = \frac{1}{\sigma_*^2}$ concludes the proof. □

Lemma C.9. *Take Assumption A3. Then, with probability at least $1-\delta$,*

$$\forall t \geq 1, \sum_{s=1}^t (f_s(\hat{\theta}_{s-1}) - f_s(\theta^*)) \leq \sum_{s=1}^t r_s^2 + \sigma_*^2\ln(1/\delta) \ .$$

Proof. Note that, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{s=1}^t (f_s(\hat{\theta}_{s-1}) - f_s(\theta^*)) &= \sum_{s=1}^t \frac{1}{2}r_s^2 - r_s\nu_s \\
&\leq \sum_{s=1}^t \frac{1}{2}r_s^2 + \frac{a}{2}\sum_{s=1}^t r_s^2\sigma_*^2 + \frac{1}{a}\ln(1/\delta) \\
&= \sum_{s=1}^t r_s^2 + \sigma_*^2\ln(1/\delta) \ . \qquad \text{(choose } a = 1/\sigma_*^2\text{)}
\end{aligned}$$

□

Lemma C.10. *Take Assumption A3. If $D_t := \|z_t\|^2_{\Sigma_t^{-1}} \le \frac{1}{2}, \forall t \ge 1$, then with probability at least $1-2\delta$,*

$$\forall t, \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \le 10\sigma_*^2 \ln(1/\delta) + \frac{3}{2}\lambda\|\theta^*\|_2^2 + 4\sum_{s=1}^{t} f_s(\theta^*) D_s^2 \,.$$

Proof. Let $r_s := z_s^\top(\hat{\theta}_{s-1} - \theta^*)$.

By the regret equality (Lemma C.7), we have

$$\sum_{s=1}^{t} \frac{1}{2} r_s^2 - r_s\nu_s = \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) - f_s(\theta^*) \le \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \,. \tag{16}$$

Then, using Lemma C.3, with probability at least $1-\delta$, we have, $\forall t \ge 1$,

$$\begin{aligned}
&\sum_{s=1}^{t} \frac{1}{2} r_s^2 \le \frac{1}{4}\sum_{s=1}^{t} r_s^2 + 2\sigma_*^2 \ln(1/\delta) + \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \,, \\
\implies &\frac{1}{4}\sum_{s=1}^{t} r_s^2 \le 2\sigma_*^2 \ln(1/\delta) + \frac{\lambda}{2}\|\theta^*\|^2 + \sum_{s}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{s=1}^{t} (f_s(\hat{\theta}_{s-1}) - f_s(\theta^*)) D_s^2 &= \sum_{s=1}^{t} \frac{1}{2} r_s^2 D_s^2 - \sum_{s=1}^{t} r_s\nu_s D_s^2 \\
&\le \frac{1}{4}\sum_{s=1}^{t} r_s^2 - \sum_{s=1}^{t} r_s\nu_s D_s^2 && (D_s^2 \le \tfrac{1}{2}) \\
&\le \sum_{s=1}^{t} r_s\nu_s(\frac{1}{2} - D_s^2) + \frac{\lambda}{4}\|\theta^*\|^2 + \frac{1}{2}\sum_{s}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \,. && \text{(by (16))}
\end{aligned}$$

Using Lemma C.3, with probability at least $1-\delta$, we have, $\forall t \ge 1$,

$$\sum_{s=1}^{t} r_s\nu_s(\frac{1}{2} - D_s^2) \le \frac{\xi}{2}\sum_{s=1}^{t} r_s^2 \underbrace{(\frac{1}{2} - D_s^2)^2}_{\le \frac{1}{4}} + \frac{\sigma_*^2}{\xi}\ln(1/\delta)$$

for some $\xi > 0$. Thus,

$$\begin{aligned}
\sum_{s=1}^{t} (f_s(\hat{\theta}_{s-1}) - f_s(\theta^*)) D_s^2 &\le \frac{\xi}{8}\sum_{s=1}^{t} r_s^2 + \frac{\sigma_*^2}{\xi}\ln(1/\delta) + \frac{\lambda}{4}\|\theta^*\|^2 + \frac{1}{2}\sum_{s}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \\
&\le (\xi + \frac{1}{\xi})\sigma_*^2 \ln(1/\delta) + (\xi+1)\frac{\lambda}{4}\|\theta^*\|^2 + (\frac{\xi}{2} + \frac{1}{2})\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 \\
&\qquad\qquad \text{(use the bound on } \tfrac{1}{4}\textstyle\sum_s r_s^2) \\
&= \frac{5}{2}\sigma_*^2 \ln(1/\delta) + \frac{3}{8}\lambda\|\theta^*\|^2 + \frac{3}{4}\sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 && (\text{choose } \xi = \tfrac{1}{2}) \\
\implies \sum_{s=1}^{t} f_s(\hat{\theta}_{s-1}) D_s^2 &\le 10\sigma_*^2 \ln(1/\delta) + \frac{3}{2}\lambda\|\theta^*\|^2 + 4\sum_{s=1}^{t} f_s(\theta^*) D_s^2 \,.
\end{aligned}$$

□

Lemma C.11. *Take Assumption A3. Assume* $\|z_s\|_{\Sigma_{s-1}^{-1}} \leq 1, \forall s$*. Then, with probability at least* $1-\delta$

$$\forall t \geq 1, \sum_{s=1}^{t} (z_s^\top (\hat{\theta}_{s-1} - \theta^*))^2 \leq 8 \left(\frac{\lambda}{2} \|\theta^*\|^2 - \frac{1}{2} \|\theta^* - \hat{\theta}_t\|_{\Sigma_t}^2 + \frac{1}{2} \sum_{s=1}^{t} \nu_s^2 D_s^2 \right) + 16\sigma_*^2 \ln(1/\delta) \ .$$

Proof. Let $r_s = z_s^\top (\hat{\theta}_{s-1} - \theta^*)$ From the regret equality (Lemma C.7)

$$\begin{aligned}
\frac{1}{2} \sum_{s=1}^{t} r_s^2 - \sum_{s=1}^{t} \nu_s r_s &= \frac{\lambda}{2} \|\theta^*\|^2 - \frac{1}{2} \|\hat{\theta}_t - \theta^*\|_{\Sigma_t}^2 + \frac{1}{2} \sum_{s=1}^{t} r_s^2 D_s^2 - \sum_{s=1}^{t} \nu_s r_s D_s^2 + \frac{1}{2} \sum_{s=1}^{t} \nu_s^2 D_s^2 \\
\implies \frac{1}{2} \sum_{s=1}^{t} r_s^2 (1 - D_s^2) &= \underbrace{\frac{\lambda}{2} \|\theta^*\|^2 - \frac{1}{2} \|\hat{\theta}_t - \theta^*\|_{\Sigma_t}^2 + \frac{1}{2} \sum_{s=1}^{t} \nu_s^2 D_s^2}_{=:\, A} + \sum_{s=1}^{t} \nu_s r_s (1 - D_s^2)
\end{aligned}$$

where (a) is obtained by tuning the scalar in the exponential martingale carefully.

With Lemma C.3, one can obtain, with probability at least $1-\delta$,

$$\forall t \geq 1, \sum_{s=1}^{t} \nu_s r_s (1 - D_s^2) \leq \frac{1}{4} \sum_{s=1}^{t} r_s^2 (1 - D_s^2)^2 + 2\sigma_*^2 \ln(1/\delta) \ .$$

Thus,

$$\begin{aligned}
\sum_{s=1}^{t} r_s^2 (1 - D_s^2) &\leq 4A + 8\sigma_*^2 \ln(1/\delta) \\
\implies \sum_{s=1}^{t} r_s^2 &\leq 8A + 16\sigma_*^2 \ln(1/\delta)
\end{aligned}$$

where the last line is due to the fact that $\|x_s\|_{\Sigma_{s-1}^{-1}} \leq 1 \implies D_s^2 \leq \frac{1}{2}$ using the Woodbury matrix identity. □

Lemma C.12. *Take Assumption A3 with the added assumption of* $\nu_t \in [-R, R], \forall t$*, with probability 1.*

$$1 - \delta \leq \mathbb{P} \left(\forall t \geq 1, \sum_{s=1}^{t} \nu_s^2 \leq \frac{3}{2} \sum_{s=1}^{t} \sigma_s^2 + R^2 \ln(1/\delta) \right) \ .$$

Proof. Using the standard sub-Gaussian inequality (Lemma C.3) with the fact that $\nu_s^2 - \sigma_s^2$ is R^2-sub-Gaussian, we have, with probability at least $1-\delta$,

$$\begin{aligned}
\forall t \geq 1, \sum_{s=1}^{t} \nu_s^2 &= \sum_{s=1}^{t} \sigma_s^2 + \sum_{s=1}^{t} (\nu_s^2 - \sigma_s^2) \\
&\leq \sum_{s=1}^{t} \sigma_s^2 + \frac{a}{2} \sum_{s=1}^{t} \mathbb{E}[(\nu_s^2 - \sigma_s^2)^2 \mid \mathcal{F}_{s-1}] + \frac{R^2}{a} \ln(1/\delta) \\
&\leq \sum_{s=1}^{t} \sigma_s^2 + \frac{a}{2} \sum_{s=1}^{t} \sigma_s^2 + \frac{R^2}{a} \ln(1/\delta) \\
&\leq \frac{3}{2} \sum_{s=1}^{t} \sigma_s^2 + R^2 \ln(1/\delta) \ . && \text{(choose } a = 1\text{)}
\end{aligned}$$

Note that the above holds true for every t simultaneously, with probability at least $1-\delta$. □

D. Implementation Details of LOFAV

For LOFAV, we can define an extra confidence set:

$$\forall \ell \in [L], \mathcal{C}^{\text{full}}_{t,L+\ell} := \left\{ \theta : \frac{1}{2} \|\hat{\theta}_{t,\ell} - \theta\|^2_{\Sigma_{t,\ell}} \leq \frac{\lambda_\ell}{2} S^2 + \sum_{s=1}^{t} \ell^{\mathsf{w}}_{s,\ell}(\hat{\theta}_{s-1,\ell}) D^2_{s,\ell} + R^2 \ln(2L/\delta) =: \gamma_{t,\ell} \right\} .$$

Then, change the algorithm so we pull $x_t = \arg\max_{x \in \mathcal{X}_t} \max_{\theta \in \overline{\mathcal{C}}^{\text{full}}_{t-1}} \langle x, \theta \rangle$ where $\overline{\mathcal{C}}^{\text{full}}_{t-1} := \cap_{\ell=1}^{2L} \mathcal{C}^{\text{full}}_{t-1,\ell}$. Specifically,

$$\max_{\theta \in \overline{\mathcal{C}}^{\text{full}}_{t-1}} \langle x, \theta \rangle = \min \left\{ \min_{\ell \in [L]} \langle x, \bar{\theta}_{t-1,\ell} \rangle + \sqrt{2\beta_{t-1,\ell}} \|x\|_{\overline{\Sigma}^{-1}_{t-1,\ell}}, \min_{\ell \in [L]} \langle x, \hat{\theta}_{t-1,\ell} \rangle + \sqrt{2\gamma_{t-1,\ell}} \|x\|_{\Sigma^{-1}_{t-1,\ell}} \right\} .$$

One can verify easily that

$$1 - \delta \leq \mathbb{P}(\forall t \in [n], \theta^* \in \mathcal{C}^{\text{full}}_t)$$

since the event we need for $\{\mathcal{C}^{\text{full}}_{t,\ell}\}_{\ell=L+1}^{2L}$ to be true (i.e., the upper deviation of the negative regret) is already assumed in the proof of the correctness of $\{\mathcal{C}^{\text{full}}_{t,\ell}\}_{\ell=1}^{L}$

E. Additional Experiments for LOSAN

In this section, we provide additional experiments on toy dataset to verify our theory and identify strengths and weaknesses of LOSAN.

In addition to OFUL, we accommodate recent advances of confidence sets by using OFUL with the improved confidence set proposed in Chowdhury et al. (2023, Appendix F), which coincides with the version of AMM-UCB in Flynn et al. (2023, Appendix C.2). Specifically, instead of the confidence set of OFUL (Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020), i.e.,

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t}^2 \le \left(\sqrt{\lambda} S + \sqrt{\sigma_0^2 \ln\left(\frac{|V_t|}{|\lambda I|}\right) + 2\sigma_0^2 \ln(1/\delta)} \right)^2 \right\}$$

where $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$, $\hat{\theta}_t = V_t^{-1} X_t^\top y_t$, $X_t \in \mathbb{R}^{t \times d}$ is the design matrix, and $y_t \in \mathbb{R}^t$ is the vector of rewards from time step 1 to t, we use

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t}^2 \le \lambda S^2 + \sigma_0^2 \ln\left(\frac{|V_t|}{|\lambda I|}\right) + 2\sigma_0^2 \ln(1/\delta) \right\} .$$

We call this variation OFUL-C. We exclude any methods that are computationally more demanding (i.e., orderwise more computations) than OFUL such as CMM-UCB of Flynn et al. (2023). For all methods, we set $\lambda = 10 \cdot \frac{\sigma_0^2}{S^2}$ where σ_0^2 is the specified noise and S is the known upper bound of the norm of the unknown parameter $\|\theta^*\|_2$.

More experiments for LOSAN. We consider a set of hard instances where the suboptimality gap of all the suboptimal arms is $\Delta := 4\sqrt{\sigma_0^2 d^2/n}$ as we describe below, which is meant to achieve the worst-case regret bound for OFUL, roughly. We set $d = 20$, $n = 50{,}000$, $\sigma_0^2 = 1.0$, and $\theta^* = (S, 0, \ldots, 0) \in \mathbb{R}^d$. We take a fixed arm set $\mathcal{X}$ and use it throughout the time steps (i.e., $\mathcal{X}_t = \mathcal{X}, \forall t$) with $|\mathcal{X}| = 400$. The arm set consists of a single best arm $x^* = \frac{1}{S}\theta^*$ and the rest of the arms whose suboptimality gap is all equal to Δ. To ensure this, each suboptimal arm x has its first coordinate $x(1)$ as $1 - \frac{\Delta}{S}$ and the other coordinates as a vector $(x(2), \ldots, x(d)) \in \mathbb{R}^{d-1}$ uniformly drawn from a sphere of radius $\sqrt{1 - (1 - \frac{\Delta}{S})^2}$. This ensures that every arm has a unit Euclidean norm and the suboptimality gap is equal to Δ.

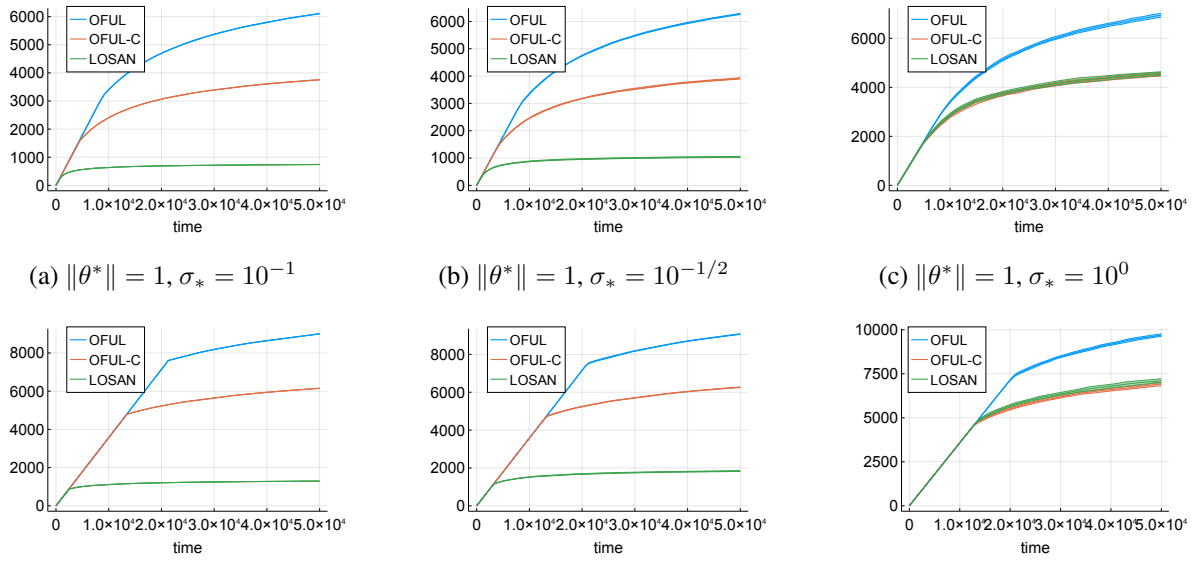
We set the parameter S of all the baseline algorithms as $\|\theta^*\|$ and the specified noise as $\sigma_0^2 = 1$. We try $\|\theta\|_2 \in \{1, 10\}$ and $\sigma_* \in \{10^{-1}, 10^{-1/2}, 10^0\}$ and draw the reward noise by $\eta_t \sim \mathcal{N}(0, \sigma_*^2)$, which results in total 6 experiments. We repeat each experiment 20 times and report their average regret along with their twice standard error as the error band in Figure 6 As one can see, LOSAN significantly outperforms all the other methods when the noise is over-specified (see (a-b) and (d-e)). Interestingly, in the just-specified setting of (c) and (f), LOSAN outperforms OFUL and is on par with LOSAN, which shows that LOSAN is not only adaptive to the true noise but also numerically tight enough to be competitive with the state-of-the-art confidence set. Note that, when $\|\theta^*\|$ is large ($\|\theta^*\| = 10$), LOSAN is slightly worse than OFUL-C.

A just-specified setting where LOSAN is worse than OFUL and OFUL-C. Note that LOSAN is not without a weakness in the just-specified setting. We here report a case where LOSAN performs worse than OFUL and OFUL-C and explain why.

The problem instance considered here is an easy problem case in the sense that not many samples are required to have small regret. This instance has a high signal-noise ratio in the sense that the smallest suboptimality gap $\Delta_{\min} := \min_{x \in \mathcal{X}, x \neq x^*} \langle x^* - x, \theta^* \rangle$, where x^* is the best arm, is large compared to the noise level σ_*^2. Specifically, we draw θ^* uniformly at random from a ($d = 20$)-dimensional sphere of radius 15 (thus $\|\theta^*\| = 15$), and each arm is drawn uniformly at random from the unit sphere. We sample 800 such arms ($|\mathcal{X}| = 800$).

When we created such an instance over 20 trials, we obtained that the average and the standard deviation of $\Delta_{\min}$ are approximately 5.38 and 0.84. When considering Gaussian noise with variance $\sigma_*^2 = 1$, the value of $\Delta_{\min}$ is much larger than σ_*, and thus one can expect that an algorithm could identify the best arm roughly after pulling 20 arms that are sufficiently linearly independent. Furthermore, $\Delta_{\max} := \min_{x \in \mathcal{X}, x \neq x^*} \langle x^* - x, \theta^* \rangle$ has the average and the standard deviation of approximately 24.79 and 0.68. The large value of $\Delta_{\max}$ implies that the cost of pulling a suboptimal arm can be very large.

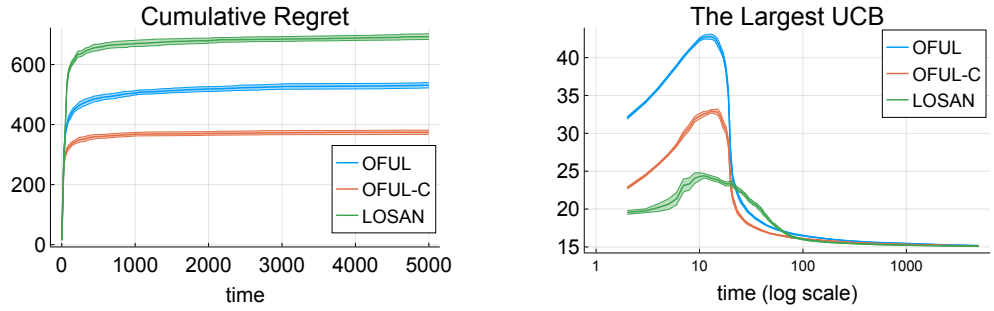
We run OFUL, OFUL-C, and LOSAN with 20 trials where each trial samples a fresh θ^* and an arm set. We report the resulting regret in Figure 7(a) where the error band is twice the standard error. As one can see, LOSAN is outperformed by



(a) $\|\theta^*\| = 1, \sigma_* = 10^{-1}$ (b) $\|\theta^*\| = 1, \sigma_* = 10^{-1/2}$ (c) $\|\theta^*\| = 1, \sigma_* = 10^{0}$

(d) $\|\theta^*\| = 10, \sigma_* = 10^{-1}$ (e) $\|\theta^*\| = 10, \sigma_* = 10^{-1/2}$ (f) $\|\theta^*\| = 10, \sigma_* = 10^{0}$

Figure 6. Toy Experiments with Gaussian noise.



(a) (b)

Figure 7. A case where LOSAN performs worse than baseline method.

both OFUL and OFUL-C. The problem is easy enough that once an algorithm finds the best arm, then it rarely increases the regret, which means that they have arrived at a regime where the regret is polylogarithmic. While the slope of the regret after time 1000 is mostly the same for all methods, LOSAN incurs more regret early on.

To inspect the reason, we plot the largest upper confidence bound (UCB), i.e., $\max_{x\in\mathcal{X},\theta\in\mathcal{C}_{t-1}}\langle x,\theta\rangle$), in Figure 7(b). Both OFUL and OFUL-C drop the value of the largest UCB significantly around $t = 20$. While LOSAN initially has a tighter confidence bound at the beginning, its drop of confidence bound happens at a much slower rate during which LOSAN pulls suboptimal arms more frequently, resulting in high regret. We speculate that such a difference is within a constant factor at best; we would be very surprised if one can show that LOSAN is orderwise worse than OFUL/OFUL-C (i.e., exist a series of instances where the difference can be arbitrarily large). Further examination and potential improvement for LOSAN is left as future work.

As a comparison, the instance used in Figure 6 has $\Delta_{\min} = \Delta_{\max}$, which means that not knowing the best arm has the same cost, and the bottleneck of the regret is mainly to precisely locate θ^*. On the other hand, the bottleneck in the spherical instance used here is to quickly figuring out θ^* very roughly (since that is enough to find the best arm), and any delay in doing so results in large cost in regret (due to large $\Delta_{\max}$).