VARIANCE-AWARE LEXICOGRAPHIC GENERALIZED LINEAR BANDITS: UNKNOWN VARIANCE AND TIGHTER BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates the Lexicographic Multi-Objective Generalized Linear Bandit problem, in which a learner optimizes m hierarchically ordered objectives by sequentially selecting arms and observing their corresponding reward vectors. While multi-objective bandits have been widely studied, existing work largely overlooks the exploitation of reward variance. Furthermore, current variance-aware approaches for generalized linear bandits typically require prior knowledge of the variance and suffer from a suboptimal variance-independent regret term of $\widetilde{O}(\sqrt{dT})$. To address these limitations, we propose a novel variance-aware algorithm that achieves a regret bound of $\widetilde{O}(d\sqrt{\sum_{t=1}^{T}(\sigma_t^i)^2} + d)$ for any objective $i \in [m]$, where $(\sigma_t^i)^2$, d, and T represent the reward variance, arm dimension, and time horizon, respectively. Crucially, our algorithm eliminates the need for prior variance knowledge and significantly reduces the variance-independent term to $\widetilde{O}(d)$. While this method relies on maximum likelihood estimation, which can be computationally intensive, we further propose a computationally efficient online alternative that achieves a regret bound of $\widetilde{O}(d\sqrt{T})$. Finally, we validate our theoretical findings through both numerical simulations and real-world experiments.

1 INTRODUCTION

Multi-armed bandits (MAB) have emerged as a fundamental framework for addressing the exploration-exploitation trade-off in sequential decision-making under uncertainty (Robbins, 1952; Lattimore & Szepesvári, 2020). In a bandit problem, a learner is presented with a set of actions or arms, each associated with a probability distribution of rewards. The learner's goal is to devise a strategy that maximizes the cumulative reward over time by dynamically selecting which arm to pull (Auer, 2002; Dani et al., 2008; Kleinberg et al., 2008). This decision-making process requires the learner to balance learning about the arms' reward distributions (exploration) with choosing the most promising arms (exploitation). MAB has found applications across various domains, including news recommendation (Li et al., 2010), clinical trials (Villar et al., 2015), and adaptive routing in networks (Bubeck & Cesa-Bianchi, 2012).

In many real-world scenarios, decision-making needs to consider multiple objectives, thus broadening the scope of bandit models from single-objective to multi-objective contexts. In these cases, after pulling an arm, the learner receives a reward vector containing multiple, often conflicting objectives. The learner's goal is to develop strategies that perform effectively across these diverse dimensions (Drugan & Nowe, 2013). Most research on multi-objective bandits assumes that all objectives hold equal importance (Lu et al., 2019; Cai et al., 2023; Xu & Klabjan, 2023). However, some real-world applications necessitate different levels of importance among objectives. For example, in radiation treatment for cancer patients, target coverage is prioritized over the proximity to organs at risk (Jee et al., 2007). To deal with such hierarchical or prioritized objectives, lexicographic ordering is proposed (Ehrgott, 2005), which ranks objectives based on their relative importance, with the most significant objective as the primary criterion. Only when arms are indistinguishable with respect to the primary criterion does the next objective come into consideration.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with representative variance-aware results for linear bandits.

Algorithm	Regret	Model	Multi-obj.	Online	Var.-free
Kim et al. (2022)	$\widetilde{O}\big(d^{1.5}\sqrt{V_T}+d^2\big)$	SLB	✗	✗	✓
Zhao et al. (2023a)	$\widetilde{O}\big(d\sqrt{V_T}+d\big)$	SLB	✗	✓	✓
VA-LGLB (linear)	$\widetilde{O}\Big(d\sqrt{V_T^i}+d\Big)$	SLB	✓	✓	✓
Zhao et al. (2023b)	$\widetilde{O}\Big(d\sqrt{V_T}+\boldsymbol{\sqrt{dT}}\Big)$	GLB	✗	✗	✗
VA-LGLB (general)	$\widetilde{O}\Big(d\sqrt{V_T^i}+\boldsymbol{d}\Big)$	GLB	✓	✗	✓
Online-LGLB	$\widetilde{O}(d\sqrt{T})$	GLB	✓	✓	–

$V_T=\sum_{t=1}^T\sigma_t^2$ and $V_T^i=\sum_{t=1}^T(\sigma_t^i)^2$; Multi-obj.: support for multiple objectives;
Online: support for online parameter updates; Var.-free: no prior knowledge of variance is required.

Bandit models have recently been extended to lexicographic settings, but existing studies focus mainly on worst-case regret guarantees (Tekin & Turgay, 2018; Hüyük & Tekin, 2021; Xue et al., 2024)[1]. Such guarantees fail to adapt to favorable problem instances, where the stochastic rewards may exhibit low variance and significantly smaller regret is achievable. In contrast, variance-aware regret bounds have been well studied for single-objective bandits. For generalized linear bandits (GLB), Zhao et al. (2023b) establish a regret bound of $\widetilde{O}(d\sqrt{\sum_{t=1}^T\sigma_t^2}+\sqrt{dT})$, but their algorithm requires prior knowledge of the reward variances. Moreover, compared with stochastic linear bandits (SLB), which achieve $\widetilde{O}(d\sqrt{\sum_{t=1}^T\sigma_t^2}+d)$ without prior variance knowledge (Zhao et al., 2023a), there remains a notable gap in the variance-independent term. This gap becomes more pronounced over long horizons T, suggesting that the variance-independent term in existing variance-aware GLB analysis may be improvable. These observations raise a natural question:

Can we achieve variance-aware regret guarantees for lexicographic GLB without prior variance knowledge, while attaining regret guarantees comparable to those for SLB?

To answer this question, we study the *lexicographic generalized linear bandit* (LGLB) problem with m objectives, where each arm is represented by a feature vector $\boldsymbol{x}\in\mathbb{R}^d$, and each objective follows a generalized linear model. Table 1 summarizes the comparison between our results and prior work. Our main contributions are summarized as follows:

- We develop VA-LGLB, a variance-aware algorithm that eliminates the need for prior knowledge of the reward variances. Unlike existing GLB methods (Zhao et al., 2023b), its confidence construction must estimate the unknown variances from nonlinear residuals while jointly controlling the resulting variance-estimation and parameter-estimation errors.
- We introduce AAE-LGLB, a lexicographic active-arm elimination procedure that aggregates uncertainty across higher-priority objectives. Unlike independent objective-wise elimination (Xue et al., 2024), AAE-LGLB aggregates uncertainty across preceding objectives through $\bar{\beta}_{t,s}^i=\max_{j\in[i]}\widehat{\beta}_{t,s}^j$ and priority-dependent elimination thresholds. We prove that it preserves the lexicographic optimal arm while controlling the reward gap.
- Combining the variance-adaptive confidence analysis with lexicographic elimination, we establish an objective-wise regret bound of $\widetilde{O}\big(\Lambda_i(d\sqrt{\sum_{t=1}^T(\sigma_t^i)^2}+d)\big)$, where $\Lambda_i=1+\lambda+\cdots+\lambda^{i-1}$. Compared with the method of Zhao et al. (2023b), our method requires no prior variance knowledge and improves the variance-independent term.
- To reduce the computational cost of VA-LGLB, we further develop ONLINE-LGLB, which replaces the offline maximum likelihood estimation (MLE) updates with recursive online Newton steps and achieves $\widetilde{\mathcal{O}}(\Lambda_i d\sqrt{T})$ regret.

[1]Due to space limitations, a more comprehensive discussion of related work can be found in Appendix A.

2 PROBLEM SETTING

Notation. In this paper, plain letters (e.g., x) represent scalars, lowercase bold letters (e.g., $\boldsymbol{x}$) represent vectors, and uppercase bold letters (e.g., $\mathbf{V}$) represent matrices. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, its Euclidean norm is denoted by $\|\boldsymbol{x}\|_2$. The norm induced by a positive-definite matrix $\mathbf{V} \in \mathbb{R}^{d\times d}$ is $\|\boldsymbol{x}\|_{\mathbf{V}} = \sqrt{\boldsymbol{x}^\top \mathbf{V}\boldsymbol{x}}$. The index i is used to distinguish between different objectives, e.g., y_t^i is the reward of the i-th objective at round t. For any positive integer N, $[N]$ denotes the set $\{1, 2, \ldots, N\}$.

This paper studies multi-objective GLB under lexicographic ordering, where the rewards are vectors with multiple elements, each generated by a generalized linear model (P. McCullagh, 1989).

In round t, the learner first observes the arm set $\mathcal{A}_t \subseteq \mathbb{R}^d$, then plays an arm $\boldsymbol{x}_t \in \mathcal{A}_t$ and receives a stochastic reward vector $\boldsymbol{y}_t = [y_t^1, y_t^2, \ldots, y_t^m] \in \mathbb{R}^m$. The i-th element of $\boldsymbol{y}_t$ is generated from a generalized linear model as follows,

$$y_t^i = \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) + \epsilon_t^i, i \in [m]$$

where $\mu^i(\cdot)$ is the link function of the i-th objective and $\boldsymbol{\theta}_*^i$ is the unknown parameter vector for the i-th objective. The term ϵ_t^i is a random noise term with zero mean and variance $(\sigma_t^i)^2$, i.e.,

$$\mathrm{E}[\epsilon_t^i|\mathcal{G}_{t-1}] = 0, \mathrm{Var}[\epsilon_t^i|\mathcal{G}_{t-1}] = (\sigma_t^i)^2, i \in [m] \tag{1}$$

where $\mathcal{G}_{t-1} \triangleq \{\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_{t-1}, y_{t-1}, \boldsymbol{x}_t\}$ denotes the σ-filtration and $\mathcal{G}_0 = \emptyset$. For a simplified statement of the main regret bound, we first impose the following variance ordering:

$$\sigma_t^m \geq \sigma_t^{m-1} \geq \cdots \geq \sigma_t^2 \geq \sigma_t^1. \tag{2}$$

This ordering is not required by the algorithm and does not imply any intrinsic relationship between objective priority and observation noise. Appendix E gives the general result using $\bar{\sigma}_t^i = \max_{j\in[i]} \sigma_t^j$.

Comparing different arms in a multi-objective setting is challenging due to the conflicts among objectives. For example, it is not immediately evident whether the reward vector $[1, 0]$ is preferable to $[0, 1]$. To resolve such ambiguities, lexicographic ordering is introduced (Hüyük & Tekin, 2021).

Definition 1 (Lexicographic Order) *Given two vectors $\boldsymbol{u} = [u^1, \ldots, u^m]$ and $\boldsymbol{v} = [v^1, \ldots, v^m]$, $\boldsymbol{u}$ dominates $\boldsymbol{v}$ if and only if there exists $i \in [m]$ such that $u^j = v^j$ for $j \in [i-1]$, and $u^i > v^i$.*

Definition 2 (Lexicographic Optimal Arm) *An arm $\boldsymbol{x}_t^* \in \mathcal{A}_t$ is lexicographic optimal if its expected reward vector is not lexicographically dominated by any other arm in $\mathcal{A}_t$.*

Accordingly, we extend the standard regret (Filippi et al., 2010) to the multi-objective setting:

$$R^i(T) = \sum_{t=1}^{T} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i), \quad \forall i \in [m].$$

Finally, we introduce some standard assumptions following existing work on GLBs (Filippi et al., 2010) and lexicographic bandits (Xue et al., 2024).

Assumption 1 *The coefficients $\{\boldsymbol{\theta}_*^i\}_{i\in[m]}$ and the contextual vectors $\boldsymbol{x}$ are bounded, such that $\|\boldsymbol{\theta}_*^i\|_2 \leq B$ and $\|\boldsymbol{x}\|_2 \leq 1$ for all $\boldsymbol{x} \in \mathcal{A}_t, t \in [T]$.*

Assumption 2 *The link functions $\{\mu^i(\cdot)\}_{i\in[m]}$ are continuously differentiable, and their derivatives satisfy $\kappa \leq \dot{\mu}^i(\cdot) \leq L$ for some constants $\kappa, L > 0$.*

Assumption 3 *The link functions $\{\mu^i(\cdot)\}_{i\in[m]}$ are bounded by U, and the noise is bounded by R, such that $|\mu^i(\cdot)| \leq U$ and $|y_t^i - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)| \leq R$.*

Assumption 4 *The trade-off among different objectives is characterized by some $\lambda \geq 0$, such that for any $i \in \{2, 3, \ldots, m\}$,*

$$\mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) \leq \lambda \cdot \max_{j\in[i-1]} \{\mu^j(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^j) - \mu^j(\boldsymbol{x}^\top \boldsymbol{\theta}_*^j)\}.$$

The parameter λ quantifies the trade-off among conflicting objectives: moving away from the optimal arm $\boldsymbol{x}_t^*$, the gain in the i-th objective is at most λ times the largest loss incurred in the higher-priority objectives. For example, $\lambda = 1$ means that a unit loss in the first objective yields at most a unit gain in the second, as in moving from $[1, 0]$ to $[0, 1]$.

3 ALGORITHMS

In this section, we first develop a variance-aware algorithm, which can adapt to the variance of rewards but incurs high computational costs. Then, we propose an alternative online algorithm to reduce the computational expenses.

3.1 VARIANCE-AWARE ALGORITHM: VA-LGLB

In this part, we present our first algorithm called Variance-Aware Lexicographic Generalized Linear Bandits (VA-LGLB), whose details are presented in Algorithm 1.

Structural Overview. VA-LGLB shares a similar structure to that of existing SupLin+ algorithms (Auer, 2002; Chu et al., 2011; Bengs et al., 2022; Zhao et al., 2023a;b; Di et al., 2024), which divides the decision-making process into multiple stages. Each stage employs historical data with varying levels of uncertainty, thus leading to a delicate analysis of variance during decision-making. A critical step in SupLin+ algorithms is to eliminate suboptimal arms between stages, which is particularly challenging for the lexicographic bandit problem due to the risk of discarding lexicographic optimal arms. To address this issue, we introduce Active Arm Elimination for LGLB (AAE-LGLB), a method tailored to the structure of the LGLB problem.

VA-LGLB initializes the number of stages as $S = \lceil 3/2 \cdot \log_2 T \rceil$. For each stage $s \in [S]$, it sets the regular matrices to $\mathbf{V}_{1,s} = 2^{-2s}\mathbf{I}$, and initializes the historical trial set $\Psi_{1,s}$ as empty. For each $i \in [m]$, the estimator $\hat{\boldsymbol{\theta}}^i_{1,s}$ and confidence radius $\hat{\beta}^i_{1,s}$ are set as $\mathbf{0}$ and $2^{-s}(B + \frac{1}{\kappa})$, respectively.

At the start of each round t, VA-LGLB sets the stage $s = 1$, initializes the available arm set $\mathcal{A}_{t,1} = \mathcal{A}_t$ and sets an update flag to False. It then enters a loop that continues until an arm $\boldsymbol{x}_t$ is selected. Within this loop, VA-LGLB considers three cases:

Case 1 (Step 7): VA-LGLB first checks whether all arms in $\mathcal{A}_{t,s}$ satisfy $\|\boldsymbol{x}\|_{\mathbf{V}^{-1}_{t,s}} \leq T^{-3/2}$, where $\|\boldsymbol{x}\|_{\mathbf{V}^{-1}_{t,s}}$ quantifies the uncertainty in the reward estimation of arm $\boldsymbol{x}$. If this condition is met, it implies that VA-LGLB has sufficiently explored all arms in $\mathcal{A}_{t,s}$. Thus, VA-LGLB invokes AAE-LGLB (Algorithm 2) to eliminate suboptimal arms, resulting in a refined arm set $\mathcal{A}_{t,T}$.

AAE-LGLB takes the estimated terms $\{\hat{\beta}^i_{t,s}, \hat{\boldsymbol{\theta}}^i_{t,s}\}_{i\in[m]}$ and an exploration parameter W as input to filter promising arms from the candidate set $\mathcal{A}_{t,s}$. It initializes the arm set $\mathcal{A}^0_{t,s} = \mathcal{A}_{t,s}$ and defines the surrogate confidence radius for each objective $i \in [m]$ as $\bar{\beta}^i_{t,s} = \max_{j\in[i]} \hat{\beta}^j_{t,s}$. After this, AAE-LGLB sequentially refines this set based on the importance of each objective, from the first to the last. Precisely, for each objective $i \in [m]$, it first finds the arm $\hat{\boldsymbol{x}}^i_{t,s}$ in $\mathcal{A}^{i-1}_{t,s}$ that maximizes the estimated reward, i.e., $\hat{\boldsymbol{x}}^i_{t,s} = \operatorname{argmax}_{\boldsymbol{x}\in\mathcal{A}^{i-1}_{t,s}} \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}^i_{t,s})$.

Next, the arm set is updated from $\mathcal{A}^{i-1}_{t,s}$ to $\mathcal{A}^i_{t,s}$ by retaining only those arms whose estimated rewards are close to that of $\hat{\boldsymbol{x}}^i_{t,s}$, i.e.,

$$\mathcal{A}^i_{t,s} = \{\boldsymbol{x} \in \mathcal{A}^{i-1}_{t,s} \mid \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}^i_{t,s}) \geq \mu^i(\hat{\boldsymbol{x}}^{i\top}_{t,s} \hat{\boldsymbol{\theta}}^i_{t,s}) - (2 + 4\lambda + \cdots + 4\lambda^{i-1}) \cdot \bar{\beta}^i_{t,s} \cdot L \cdot W\}.$$

This procedure is repeated for all m objectives, and returns the final arm set $\mathcal{A}^m_{t,s}$, which contains arms most likely to yield high rewards across all objectives. After obtaining the filtered set $\mathcal{A}_{t,T}$, VA-LGLB randomly plays an arm $\boldsymbol{x}_t$ from the refined set and terminates the inner loop.

The surrogate radius $\bar{\beta}^i_{t,s}$ captures the largest estimation uncertainty among objectives $1, \ldots, i$, and the factor $1 + \lambda + \cdots + \lambda^{i-1}$ in the threshold arises from recursively propagating the uncertainty of higher-priority objectives through Assumption 4.

Case 2 (Step 10): If there exists an arm $\boldsymbol{x}_t$ in $\mathcal{A}_{t,s}$ such that $\|\boldsymbol{x}_t\|_{\mathbf{V}^{-1}_{t,s}} > 2^{-s}$, this indicates $\boldsymbol{x}_t$ has high uncertainty, necessitating further exploration. In this case, VA-LGLB selects $\boldsymbol{x}_t$ and plays it directly. The associated observations $\boldsymbol{y}_t = [y^1_t, \ldots, y^m_t]$ are recorded, and the flag is set to True, signaling that the historical trial set and variables of stage s require updating.

Case 3 (Step 13): If neither the first nor the second condition is met, VA-LGLB invokes the AAE-LGLB to refine the arm set $\mathcal{A}_{t,s}$ using the threshold 2^{-s}. After that, the stage index s is incremented, and the process repeats, gradually narrowing down the set of arms until an arm $\boldsymbol{x}_t$ is selected.

Algorithm 1 VA-LGLB

Input: $\delta, T, B, \kappa, L, R, \lambda$
Set $S = \lceil 3/2 \cdot \log_2 T \rceil$
Initialize $\mathbf{V}_{1,s} = 2^{-2s}\mathbf{I}$, $\Psi_{1,s} = \emptyset$, $\hat{\boldsymbol{\theta}}^i_{1,s} = \mathbf{0}$, $\hat{\beta}^i_{1,s} = 2^{-s}(B + \frac{1}{\kappa})$ for all $s \in [S]$ and $i \in [m]$
for $t = 1, 2, \ldots, T$ **do**
  Set $s = 1, \mathcal{A}_{t,1} = \mathcal{A}_t$, flag = False
  repeat
    if $\forall\, \boldsymbol{x} \in \mathcal{A}_{t,s}, \|\boldsymbol{x}\|_{\mathbf{V}^{-1}_{t,s}} \leq T^{-3/2}$ **then**
      Run Algorithm 2 to refine arms: $\mathcal{A}_{t,T} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s}, \{\hat{\beta}^i_{t,s}, \hat{\boldsymbol{\theta}}^i_{t,s}\}_{i\in[m]}, T^{-\frac{3}{2}}\right)$
      Randomly play an arm $\boldsymbol{x}_t \in \mathcal{A}_{t,T}$
    else if $\exists\, \boldsymbol{x}_t \in \mathcal{A}_{t,s}, \|\boldsymbol{x}_t\|_{\mathbf{V}^{-1}_{t,s}} > 2^{-s}$ **then**
      Play the arm $\boldsymbol{x}_t$ and observe $\boldsymbol{y}_t = [y^1_t, \ldots, y^m_t]$
      Set flag = True
    else
      Run Algorithm 2 to refine arms: $\mathcal{A}_{t,s+1} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s}, \{\hat{\beta}^i_{t,s}, \hat{\boldsymbol{\theta}}^i_{t,s}\}_{i\in[m]}, 2^{-s}\right)$
      Update $s = s + 1$
    end if
  until an arm $\boldsymbol{x}_t$ is played
  if flag = True **then**
    Update the index set $\Psi_{t+1,s} = \Psi_{t,s} \cup \{t\}$
    Compute the weight $w_t = 2^{-s}/\|\boldsymbol{x}_t\|_{\mathbf{V}^{-1}_{t,s}}$
    Update the estimators $\{\hat{\boldsymbol{\theta}}^i_{t+1,s}\}_{i\in[m]}$ by Eq. (3)
    Update the radii $\{\hat{\beta}^i_{t+1,s}\}_{i\in[m]}$ by Eq. (4)
    Update the matrix $\mathbf{V}_{t+1,s} = \mathbf{V}_{t,s} + w_t^2 \cdot \boldsymbol{x}_t\boldsymbol{x}_t^\top$
  else
    Set $\mathbf{V}_{t+1,s} = \mathbf{V}_{t,s}$, $\Psi_{t+1,s} = \Psi_{t,s}$ and $\hat{\boldsymbol{\theta}}^i_{t+1,s} = \hat{\boldsymbol{\theta}}^i_{t,s}$, $\hat{\beta}^i_{t+1,s} = \hat{\beta}^i_{t,s}$ for all $i \in [m]$
  end if
  For $s' \in [S] - \{s\}$ and $i \in [m]$, $\mathbf{V}_{t+1,s'} = \mathbf{V}_{t,s'}$, $\Psi_{t+1,s'} = \Psi_{t,s'}$, $\hat{\beta}^i_{t+1,s'} = \hat{\beta}^i_{t,s'}$ and $\hat{\boldsymbol{\theta}}^i_{t+1,s'} = \hat{\boldsymbol{\theta}}^i_{t,s'}$
end for

Algorithm 2 AAE-LGLB

Input: $\mathcal{A}_{t,s}, \{\hat{\beta}^i_{t,s}, \hat{\boldsymbol{\theta}}^i_{t,s}\}_{i\in[m]}, W$
Set $\mathcal{A}^0_{t,s} = \mathcal{A}_{t,s}$ and $\bar{\beta}^i_{t,s} = \max_{j\in[i]} \hat{\beta}^j_{t,s}, \forall i \in [m]$
for $i = 1, 2, \ldots, m$ **do**
  $\hat{\boldsymbol{x}}^i_{t,s} = \text{argmax}_{\boldsymbol{x}\in\mathcal{A}^{i-1}_{t,s}}\, \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}^i_{t,s})$
  $\mathcal{A}^i_{t,s} = \{\boldsymbol{x} \in \mathcal{A}^{i-1}_{t,s} | \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}^i_{t,s}) \geq \mu^i(\hat{\boldsymbol{x}}^{i\top}_{t,s} \hat{\boldsymbol{\theta}}^i_{t,s}) - (2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1}) \cdot \bar{\beta}^i_{t,s} \cdot L \cdot W\}$
end for
Return the filtered arm set $\mathcal{A}^m_{t,s}$

Model Update. When an arm is played and the flag is True, VA-LGLB updates the index set to $\Psi_{t+1,s} = \Psi_{t,s} \cup \{t\}$ and computes a weight $w_t = 2^{-s}/\|\boldsymbol{x}_t\|_{\mathbf{V}^{-1}_{t,s}}$. Using the newly obtained data in stage s, VA-LGLB estimates $\hat{\boldsymbol{\theta}}^i_{t+1,s}$ for each objective $i \in [m]$ via maximum likelihood estimation (MLE) as follows:

$$2^{-2s}\kappa\boldsymbol{\theta} + \sum_{\tau\in\Psi_{t+1,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}) - y^i_\tau)\boldsymbol{x}_\tau = \mathbf{0}. \tag{3}$$

It is worth noting that when μ^i is the identity function (i.e., the SLB setting), Eq. (3) reduces to the solution of a weighted least squares problem. In this case, $\hat{\boldsymbol{\theta}}^i_{t+1,s}$ admits a closed-form solution and **can be updated online** efficiently (e.g., via the Sherman-Morrison formula). This estimation procedure allows us to construct a confidence region that contains the unknown parameter $\boldsymbol{\theta}^i_*$ with

high probability. As detailed in Theorem 1, the center of the confidence region is the estimator $\hat{\boldsymbol{\theta}}^i_{t+1,s}$, and its radius is computed by Eq. (4), which depends on the estimated variance at stage s. Finally, VA-LGLB updates the weighted covariance matrix to prepare for the next round, i.e.,

$$\mathbf{V}_{t+1,s} = \mathbf{V}_{t,s} + w_t^2 \cdot \boldsymbol{x}_t \boldsymbol{x}_t^\top.$$

If the flag remains False, VA-LGLB retains the values of $\mathbf{V}_{t+1,s}$, $\Psi_{t+1,s}$, as well as the estimators and radii from the current round t. For any other stages $s' \neq s$, VA-LGLB preserves all stored variables from the current round t. We conclude by providing theoretical guarantees for VA-LGLB.

Theorem 1 *Let $\gamma_t = \ln(4t^2 mS/\delta)$ for any $t \geq 1$ and $\bar{s}_t = \lceil \log_2(64(L/\kappa)\sqrt{\gamma_t}) \rceil$. With probability at least $1-\delta$, for any objective $i \in [m]$,*

$$\|\boldsymbol{\theta}^i_* - \hat{\boldsymbol{\theta}}^i_{t,s}\|_{\mathbf{V}_{t,s}} \leq \hat{\beta}^i_{t,s}, \forall t \geq 1$$

where

$$\hat{\beta}^i_{t,s} = \frac{16 \cdot 2^{-s}}{\kappa} \sqrt{8\widehat{\mathrm{Var}}^i_{t,s} \cdot \gamma_t + 18R^2 \cdot \gamma_t^2} + \frac{6 \cdot 2^{-s}}{\kappa} R \cdot \gamma_t + 2^{-s} B \tag{4}$$

and

$$\widehat{\mathrm{Var}}^i_{t,s} = \begin{cases} \sum_{\tau \in \Psi_{t,s}} w_\tau^2 (y^i_\tau - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}^i_{t,s}))^2 & s \geq \bar{s}_t, \\ |\Psi_{t,s}| \cdot R^2 & \textit{otherwise.} \end{cases}$$

Based on the above confidence region, the regret bound of VA-LGLB is provided as follows.

Theorem 2 *Let $\Lambda_i = 1 + \lambda + \cdots + \lambda^{i-1}$ for $i \in [m]$. With probability at least $1-\delta$, for any objective $i \in [m]$, the regret of VA-LGLB satisfies*

$$R^i(T) \leq \widetilde{O}\left(\Lambda_i \cdot d\left(\frac{L}{\kappa}\sqrt{\sum_{t=1}^T (\sigma_t^i)^2} + \frac{L^2}{\kappa^2} + L\right)\right).$$

Remark 1 Theorem 2 shows that VA-LGLB achieves a regret bound of $\widetilde{O}(\Lambda_i \cdot (d\sqrt{\sum_{t=1}^T (\sigma_t^i)^2} + d))$ for any objective $i \in [m]$, which *improves the dependence on T* relative to the single-objective GLB regret bound of $\widetilde{O}(d\sqrt{\sum_{t=1}^T \sigma_t^2} + \sqrt{dT})$ established by Zhao et al. (2023b). Since bandit problems often involve large T, this improvement is practically significant. For instance, the news recommendation system (Li et al., 2010) involves approximately $T = 4.7 \times 10^6$ interactions per day. Moreover, unlike the existing single-objective algorithm (Zhao et al., 2023b), VA-LGLB *does not require prior knowledge of the variance*, which further enhances its practicality.

Remark 2 Although an additional term Λ_i is included in our regret bound, this is the cost of optimizing multiple objectives simultaneously. Specifically, $\Lambda_1 = 1$ implies that the first objective incurs no additional regret compared to single-objective algorithms (Zhao et al., 2023b), while the subsequent objectives suffer higher regret as their priority decreases. VA-LGLB is the first variance-aware algorithm for multi-objective bandits. If $\sigma_t^i = 0$ for all $t \in [T]$, its regret bound is $\widetilde{O}(\Lambda_i \cdot d)$, which outperforms the worst-case regret bound of $\widetilde{O}(d\sqrt{T})$ in terms of T (Lu et al., 2019).

Practical Implications of λ and Λ_i. Exploiting trade-off information is a standard way to handle conflicts between objectives in multi-objective optimization (Kaliszewski, 2000; Keeney, 2002; Nowak & Trzaskalik, 2022). In practice, domain-specific requirements often constrain the trade-off intensity λ to moderate ranges. For instance, in smart grid energy management (Alzahrani et al., 2023), a 12% cost reduction may result in only a 6% emission increase, implying a sensitivity coefficient of $\lambda \approx 0.5$. Similarly, traffic signal control (Akyol et al., 2024) may achieve a 30% reduction in pedestrian delay at the cost of a 10% increase in vehicle emissions, yielding $\lambda \approx 0.3$. Such empirical evidence suggests that moderate values of λ are prevalent in real-world scenarios. Thus, Λ_i remains well-behaved, ensuring that our regret bounds do not suffer from excessive degradation.

The factor $\Lambda_i = 1 + \lambda + \cdots + \lambda^{i-1}$ arises in our analysis from recursively propagating uncertainty across higher-priority objectives. It is bounded uniformly in i when $0 \leq \lambda < 1$, grows linearly when $\lambda = 1$, and can grow geometrically when $\lambda > 1$. We do not establish a matching lower bound for this dependence, and whether the precise dependence on λ and i is unavoidable remains open.

Algorithm 3 Online-LGLB

Input: $\delta, T, B, \kappa, U, R, \lambda, \alpha = \max\{1, \kappa/2\}$
Initialize $\mathbf{V}_1 = \alpha\mathbf{I}, \hat{\boldsymbol{\theta}}_1^1 = \cdots = \hat{\boldsymbol{\theta}}_1^m = \mathbf{0}, \beta_1 = B\sqrt{d}$
for $t = 1, 2, \ldots, T$ **do**
  Set $s = 1$ and $\mathcal{A}_{t,1} = \mathcal{A}_t$
  repeat
    if $\forall\, \boldsymbol{x} \in \mathcal{A}_{t,s}, \|\boldsymbol{x}\|_{\mathbf{V}_t^{-1}} \leq 1/\sqrt{T}$ **then**
      Run Algorithm 2 to refine arms: $\mathcal{A}_{t,T} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s}, \beta_t, \{\hat{\boldsymbol{\theta}}_t^i\}_{i\in[m]}, \frac{1}{\sqrt{T}}\right)$
      Randomly choose an arm $\boldsymbol{x}_t \in \mathcal{A}_{t,T}$
    else if $\exists\, \boldsymbol{x}_t \in \mathcal{A}_{t,s}, \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}} > 2^{-s}$ **then**
      Choose the arm $\boldsymbol{x}_t$
    else
      Run Algorithm 2 to refine arms: $\mathcal{A}_{t,s+1} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s}, \beta_t, \{\hat{\boldsymbol{\theta}}_t^i\}_{i\in[m]}, 2^{-s}\right)$
      Update $s = s + 1$
    end if
  until an arm $\boldsymbol{x}_t$ is chosen
  Play the arm $\boldsymbol{x}_t$ and observe $\boldsymbol{y}_t = [y_t^1, \ldots, y_t^m]$
  Update the matrix $\mathbf{V}_{t+1} = \mathbf{V}_t + \frac{\kappa}{2}\boldsymbol{x}_t\boldsymbol{x}_t^\top$
  Compute the gradients $\{\nabla\ell_t^i(\hat{\boldsymbol{\theta}}_t^i)\}_{i\in[m]}$ by Eq. (5)
  Update the estimators $\{\hat{\boldsymbol{\theta}}_{t+1}^i\}_{i\in[m]}$ by Eq. (6)
  Compute the confidence radius β_{t+1} by Eq. (7)
end for

3.2 Online Algorithm: Online-LGLB

Although VA-LGLB can adjust to reward variability, it is expensive in both storage and computation due to the MLE performed in each round. To address this, we propose an alternative online algorithm for LGLB, termed Online-LGLB. Details of Online-LGLB are provided in Algorithm 3.

Online-LGLB adopts a decision-making framework similar to that of VA-LGLB, dividing each round into multiple stages and progressively eliminating less promising arms. At the beginning of each round t, Online-LGLB initializes the stage index $s = 1$ and sets the available arm set to $\mathcal{A}_{t,s} = \mathcal{A}_t$. It then enters a loop to select an arm. At each iteration within the loop, Online-LGLB first checks whether all arms in the current subset $\mathcal{A}_{t,s}$ satisfy $\|\boldsymbol{x}\|_{\mathbf{V}_t^{-1}} \leq 1/\sqrt{T}$. If they do, Online-LGLB invokes AAE-LGLB to filter arms based on the current estimators and an exploration width $1/\sqrt{T}$. An arm $\boldsymbol{x}_t$ is then randomly selected from the filtered set $\mathcal{A}_{t,T}$. If there exists an arm $\boldsymbol{x}_t$ whose confidence term exceeds 2^{-s}, i.e., $\|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}} > 2^{-s}$, Online-LGLB chooses $\boldsymbol{x}_t$ directly for exploration. If all arms in $\mathcal{A}_{t,s}$ have confidence terms smaller than 2^{-s}, the arm set $\mathcal{A}_{t,s}$ is further refined by the AAE-LGLB subroutine with an exploration width 2^{-s}, and the index s is incremented to apply more stringent criteria.

Model Update. Upon playing the arm $\boldsymbol{x}_t$, Online-LGLB observes the corresponding rewards $\boldsymbol{y}_t$. Using the action–reward pair $(\boldsymbol{x}_t, \boldsymbol{y}_t)$, Online-LGLB updates the estimators via the Online Newton Step (ONS) method. Precisely, for each objective $i \in [m]$, Online-LGLB first computes the gradient of the loss function with respect to the current estimator, i.e.,

$$\nabla\ell_t^i(\hat{\boldsymbol{\theta}}_t^i) = (\mu^i(\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}_t^i) - y_t^i)\cdot\boldsymbol{x}_t. \tag{5}$$

Subsequently, the estimator is updated by solving the following constrained minimization problem:

$$\hat{\boldsymbol{\theta}}_{t+1}^i = \operatorname*{argmin}_{\|\boldsymbol{\theta}\|_2\leq B} \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t^i\|_{\mathbf{V}_{t+1}}^2}{2} + \langle\boldsymbol{\theta}, \nabla\ell_t^i(\hat{\boldsymbol{\theta}}_t^i)\rangle \tag{6}$$

where $\mathbf{V}_{t+1} = \mathbf{V}_t + \frac{\kappa}{2}\boldsymbol{x}_t\boldsymbol{x}_t^\top$. Finally, Online-LGLB computes the confidence radius β_{t+1} according to Eq. (7). Theoretical guarantees for the estimator are given below.

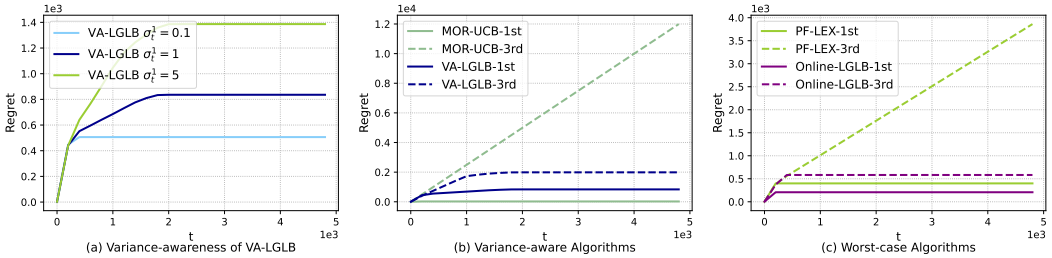


Figure 1: Comparison of our algorithms versus MOR-UCB and PF-LEX: Synthetic dataset.

Theorem 3 *With probability at least* $1-\delta$*, for any* $i \in [m]$*,*

$$\|\boldsymbol{\theta}_*^i - \hat{\boldsymbol{\theta}}_{t+1}^i\|_{\mathbf{V}_{t+1}} \leq \beta_{t+1}, \forall t \geq 0$$

where

$$\beta_{t+1} = \left(\alpha B^2 + \frac{16(U+R)^2}{\kappa} d \ln\left(1 + \frac{\kappa t}{2\alpha d}\right) + \frac{\kappa}{2} + \frac{16R^2}{\kappa} \ln\left(\frac{m\sqrt{1+4B^2t}}{\delta}\right)\right)^{1/2}. \quad (7)$$

Given this confidence region, we derive the regret bound for Online-LGLB as follows.

Theorem 4 *With probability at least* $1-\delta$*, for any* $i \in [m]$*, the regret of Online-LGLB satisfies*

$$R^i(T) \leq \widetilde{O}\left(\Lambda_i \cdot L \cdot \kappa^{-1} \cdot d\sqrt{T}\right).$$

Remark 3 Theorem 4 shows that Online-LGLB achieves a regret bound of $\widetilde{O}(\Lambda_i \cdot d\sqrt{T})$ for any objective $i \in [m]$, matching the optimal dependencies on d and T observed in single-objective GLB settings (Jun et al., 2017). The additional factor Λ_i, which is independent of d and T, characterizes the overhead of managing multiple objectives. Specifically, $\Lambda_1 = 1$ implies that for the primary objective, Online-LGLB recovers the performance of single-objective algorithms (Jun et al., 2017).

There are three main differences between VA-LGLB and Online-LGLB. First, VA-LGLB maintains S estimators, each computed from a subset of historical trial data, while Online-LGLB uses only one estimator computed from all the encountered action-reward pairs. Second, VA-LGLB updates its estimators by the MLE method, as shown in Eq. (3), which requires storing the historical trial data and is computationally expensive. Online-LGLB uses the ONS method described in Eq. (6), which is an online method and thus computationally efficient. Third, VA-LGLB adapts to the variance of the rewards through the variance-aware radius in Eq. (4), a feature that Online-LGLB lacks.

4 Experiments

We evaluate our algorithms on both synthetic and real-world datasets. As baselines, we consider MOR-UCB (Zhao et al., 2023b), a variance-aware algorithm for single-objective GLB, and PF-LEX (Hüyük & Tekin, 2021), a worst-case algorithm for lexicographic MAB. Each experiment is repeated 10 times, and we report the average regret. Detailed experimental settings are provided in Appendix B.

Results on Synthetic Data. We consider a controlled $K = d = 10$ setting with three objectives, where arms correspond to standard basis vectors and rewards follow a generalized linear model with Gaussian noise. The instance is constructed such that identifying the lexicographic optimal arm requires resolving all objectives. Figure 1(a) demonstrates the variance-awareness of VA-LGLB. We present the regret curves for the first objective, with the variance parameters σ_t^i set to 0.1, 1, and 5 for the three curves, respectively. As expected, higher variance leads to larger regret, aligning with our theoretical guarantees in Theorem 2 and thereby demonstrating VA-LGLB's sensitivity to variance. Notably, all three curves eventually plateau, confirming the effectiveness of VA-LGLB across different variance levels. Figures 1(b) and (c) run all algorithms with variance $\sigma_t^i = 1$, showing the regret curves for the first and third objectives. In Figure 1(b), the curves for the first and third objectives of

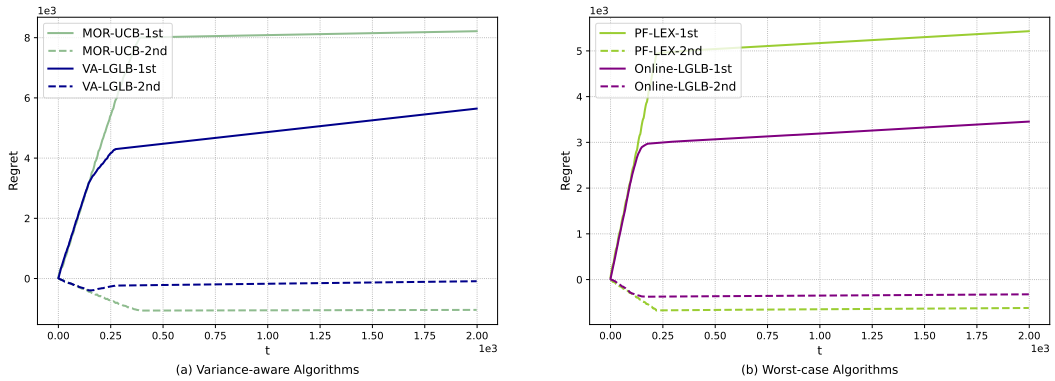


Figure 2: Comparison of our algorithms versus MOR-UCB and PF-LEX: Real-world dataset.

VA-LGLB stabilize, indicating that VA-LGLB has successfully identified the lexicographic optimal arm. In contrast, MOR-UCB exhibits non-convergence in its third objective, indicating its failure to simultaneously optimize both the first and third objectives. Figure 1(c) compares two algorithms whose theoretical regret bounds are worst-case. Our algorithm, Online-LGLB, outperforms PF-LEX in both the first and third objectives. Although PF-LEX is a multi-objective bandit algorithm, it struggles to optimize the third objective due to its chain-relation-based approach, which may lead to the loss of the optimal arm during the arm elimination process.

Results on Real-world Data. We evaluate our algorithms on the SW-LLVM dataset (Zuluaga et al., 2016), which consists of compiler optimization configurations characterized by binary features, with performance and memory footprint as the two objectives. A subset of 16 operating points is used to construct a stochastic bandit instance based on empirical statistics. Figure 2(a) compares variance-aware algorithms. VA-LGLB achieves lower regret than MOR-UCB on the first objective, consistent with our theoretical improvement of the variance-independent term from $\widetilde{O}(\sqrt{dT})$ to $\widetilde{O}(d)$. Notably, although MOR-UCB is a single-objective algorithm, it successfully identifies the lexicographic optimal arm on the SW-LLVM dataset. This is because the expected rewards on the first objective differ across arms, which is sufficient for identifying the lexicographic optimum. In Figure 2(b), Online-LGLB incurs lower regret on the first objective, consistent with its regret bound of $\widetilde{O}(\sqrt{T})$, compared to PF-LEX's regret bound of $\widetilde{O}(T^{2/3})$. These experimental results support our theoretical analysis. Lastly, we observe that the regret on the second objective is negative. This arises because many suboptimal arms in SW-LLVM have higher expected rewards than the lexicographic optimal arm on the second objective, a consequence of the dataset's structure.

5 Conclusion and Future Work

In this paper, we developed two algorithms for the LGLB problem: VA-LGLB, a variance-aware method, and Online-LGLB, an efficient online method. VA-LGLB achieves a regret bound of $\widetilde{O}(d\sqrt{\sum_{t=1}^{T}(\sigma_t^i)^2} + d)$ for each objective $i \in [m]$ and, to the best of our knowledge, is the first variance-aware algorithm for multi-objective bandits. Beyond handling multiple objectives, it has two advantages over the single-objective method of Zhao et al. (2023b): it requires no prior knowledge of the variances, and it improves the variance-independent term from $\widetilde{O}(\sqrt{dT})$ to $\widetilde{O}(d)$. Because VA-LGLB is computationally expensive, we also designed Online-LGLB, which attains $\widetilde{O}(d\sqrt{T})$ regret for all objectives and matches the dependence on d and T of state-of-the-art single-objective algorithms (Jun et al., 2017; Zhang et al., 2016). In addition to the theoretical guarantees, experiments on both synthetic and real-world datasets support our theoretical results.

Our analysis focuses on unknown reward variances and tighter regret bounds, while recent work has improved the dependence on curvature parameters (Sawarni et al., 2024; Zhang et al., 2025). It would therefore be interesting to combine our framework with recent self-concordant GLB analysis to further improve the dependence on L and κ. Another important challenge is to develop an LGLB algorithm that is both variance-aware and computationally efficient.

AI USE STATEMENT

We used generative AI tools for language polishing. The authors reviewed the AI-assisted edits for accuracy and consistency with the scientific content and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 2 specifies the lexicographic generalized linear bandit model and its assumptions, while Section 3 gives the VA-LGLB and Online-LGLB algorithms, including the active arm elimination procedure. The appendix provides derivations of the main regret bounds and an extension that removes the variance-ordering assumption. Appendix B describes the synthetic instance, reward generation, repeated-run protocol, and construction of the 16-arm SW-LLVM instance; Table 2 lists the empirical means and variances used for the latter experiment.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 1–9, 2012.

Görkem Akyol, Sadullah Göncü, and Mehmet Ali Silgu. Multi-objective optimization framework for trade-off among pedestrian delays and vehicular emissions at signal-controlled intersections. *Arabian Journal for Science and Engineering*, 49:14117 – 14130, 2024.

Ahmad Alzahrani, Ghulam Hafeez, Sajjad Ali, Sadia Murawwat, Muhammad Iftikhar Khan, Khalid Rehman, and Azher M. Abed. Multi-objective energy optimization with load and distributed energy source scheduling in the smart power grid. *Sustainability*, 15(13), 2023.

Cagin Ararat and Cem Tekin. Vector optimization with stochastic bandit feedback. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 2165–2190, 2023.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(11):397–422, 2002.

Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 939–947, 2016.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 1764–1786, 2022.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Xin-Qiang Cai, Pushi Zhang, Li Zhao, Bian Jiang, Masashi Sugiyama, and Ashley J. Llorens. Distributional pareto-optimal multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems 36*, pp. 15593–15613, 2023.

Sayak Ray Chowdhury and Aditya Gopalan. No-regret algorithms for multi-task Bayesian optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 1873–1881, 2021.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Yan Dai, Ruosong Wang, and Simon Shaolei Du. Variance-aware sparse linear bandits. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning*, pp. 355–366, 2008.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7111–7123, 2022.

Madalina M. Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks*, pp. 1–8, 2013.

Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin, Heidelberg, 2005.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pp. 586–594, 2010.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Raphaël Féraud, Robin Allesiardo, Tanguy Urvoy, and Fabrice Clérot. Random forest for the contextual bandit problem. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 93–101, 2016.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Alihan Hüyük and Cem Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning*, 110(6):1233–1266, 2021.

Kyung-Wook Jee, Daniel L. McShan, and Benedick A. Fraass. Lexicographic ordering: intuitive multicriteria optimization for imrt. *Physics in Medicine & Biology*, 52:1845–1861, 2007.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pp. 99–109, 2017.

Ignacy Kaliszewski. Using trade-off information in decision-making algorithms. *Computers & Operations Research*, 27(2):161–182, 2000.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. In *Advances in Neural Information Processing Systems 36*, pp. 10897–10908, 2023.

Ralph L. Keeney. Common mistakes in making value trade-offs. *Operations Research*, 50(6): 935–945, 2002.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. In *Advances in Neural Information Processing Systems 35*, pp. 1060–1072, 2022.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. In *Advances in Neural Information Processing Systems 36*, 2023.

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

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Xiang Li and Qiang Sun. Variance-aware decision making with linear function approximation under heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3080–3086, 2019.

Subhojyoti Mukherjee, K. P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Efficient-ucbv: An almost optimal algorithm using variance estimates. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6417–6424, 2018.

Maciej Nowak and Tadeusz Trzaskalik. A trade-off multiobjective dynamic programming procedure and its application to project portfolio selection. *Annals of Operations Research*, 311(2):1155–1181, 2022.

John A. Nelder P. McCullagh. *Generalized Linear Models (2nd ed.)*. Routledge, 1989.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58 (5):527–535, 1952.

Ayush Sawarni, Nirjhar Das, Siddharth Barman, and Gaurav Sinha. Generalized linear bandits with limited adaptivity. In *Advances in Neural Information Processing Systems 37*, pp. 8329–8369, 2024.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems 31*, pp. 8430–8439, 2018.

Aleksandrs Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(73):2533–2568, 2014.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 1015–1022, 2010.

Cem Tekin and Eralp Turgay. Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing*, 66(14):3799–3813, 2018.

Eralp Turgay, Doruk Oner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pp. 1673–1681, 2018.

Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science*, 30(2):199 – 215, 2015.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *Proceedings of the 40th International Conference on International Conference on Machine Learning*, pp. 38499–38517, 2023.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36*, pp. 70880 – 70891, 2023.

Bo Xue, Ji Cheng, Fei Liu, Yimu Wang, and Qingfu Zhang. Multiobjective lipschitz bandits under lexicographic ordering. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 16238–16246, 2024.

Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre Gaillard. Efficient kernelized ucb for contextual bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 5689–5720, 2022.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 392–401, 2016.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. In *Advances in Neural Information Processing Systems 38*, pp. 69244–69277, 2025.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. In *Advances in Neural Information Processing Systems 34*, pp. 4342–4355, 2021.

Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *Proceedings of 36th Conference on Learning Theory*, pp. 4977–5020, 2023a.

Heyang Zhao, Dongruo Zhou, Jiafan He, and Quanquan Gu. Optimal online generalized linear regression with stochastic noise and its application to heteroscedastic bandits. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 42259–42279, 2023b.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 746–755, 2020.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11492–11502, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 4532–4576, 2021.

Marcela Zuluaga, Andreas Krause, and Markus Püschel. e-pal: An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research*, 17(104):1–32, 2016.

A Related Work

Stochastic Contextual Bandits. The literature on stochastic contextual bandits can be categorized into non-parametric and parametric models. The most commonly studied non-parametric contextual bandit model is the Lipschitz bandit model, which assumes that the reward function satisfies the Lipschitz condition (Kleinberg et al., 2008; Slivkins, 2014; Kang et al., 2023). The seminal work of Kleinberg et al. (2008) proposed a zooming algorithm for Lipschitz bandits, which achieves a regret bound of $\widetilde{O}(T^{(d_z+1)/(d_z+2)})$, where d_z is the zooming dimension of the reward function. Additionally, researchers have explored various techniques for non-parametric contextual bandits, including random forests (Féraud et al., 2016), Gaussian processes (Srinivas et al., 2010), neural networks (Zhou et al., 2020), and kernel-based methods (Zenati et al., 2022).

Parametric contextual bandits rely on predefined models to estimate rewards, often assuming a linear relationship between the context and the reward (Chu et al., 2011; Shao et al., 2018). The SupLinRel algorithm, introduced by Auer (2002), is a notable example that uses a linear model to compute upper confidence bounds in the finite-armed bandit setting. Dani et al. (2008) extended this work to the infinite-armed setting by refining the confidence region technique, resulting in a regret bound of $\widetilde{O}(d\sqrt{T})$. Filippi et al. (2010) were the first to formalize the GLB model and proposed an algorithm achieving a regret bound of $\widetilde{O}(d\sqrt{T})$. However, their algorithm required storing all previously encountered arms and rewards for parameter estimation, which is computationally expensive. Zhang et al. (2016) addressed this issue for logistic reward functions using the Online Newton Step method. Jun et al. (2017) further extended Zhang et al.'s algorithm to accommodate a general link function. Additionally, parametric contextual bandits have been studied in various contexts, including dynamic environments (Zhao et al., 2020) and robust bandits (Ding et al., 2022).

Variance-aware Bandits. Audibert et al. (2009) pioneered variance-aware bandits by introducing an algorithm that incorporated reward means and variances to handle the exploration-exploitation trade-off. Building on this, Mukherjee et al. (2018) refined this framework through arm elimination, achieving modest improvements in variance-dependent regret bounds. A notable advancement came from Zhou et al. (2021), who proposed a weighted ridge regression scheme for SLB, achieving a regret bound of $\widetilde{O}(d\sqrt{\sum_{t=1}^{T}\sigma_t^2}+\sqrt{dT})$. However, their algorithm required knowledge of the reward variance. Later, Zhang et al. (2021) removed this requirement at the cost of a higher regret bound of $\widetilde{O}(d^{4.5}\sqrt{\sum_{t=1}^{T}\sigma_t^2}+d^5)$. Kim et al. (2022) further improved this regret bound to $\widetilde{O}(d^{1.5}\sqrt{\sum_{t=1}^{T}\sigma_t^2}+d^2)$, yet these algorithms, which do not require variance knowledge, remain computationally intensive (Zhang et al., 2021; Kim et al., 2022). Recent work (Zhao et al., 2023a) bridges this gap for SLB, achieving an efficient $\widetilde{O}(d\sqrt{\sum_{t=1}^{T}\sigma_t^2}+d)$ regret bound without requiring variance knowledge. In contrast, the existing variance-aware algorithm for GLB (Zhao et al., 2023b) is computationally inefficient, requires prior knowledge of the reward variance, and achieves a regret bound of $\widetilde{O}(d\sqrt{\sum_{t=1}^{T}\sigma_t^2}+\sqrt{dT})$. Additionally, variance-aware methods have been extended to sparse linear bandits (Dai et al., 2023), dueling bandits (Di et al., 2024), and heavy-tailed reward environments (Li & Sun, 2024).

Multi-objective Bandits. Drugan & Nowe (2013) introduced two basic approaches for multi-objective multi-armed bandits (MOMAB): encoding multiple objectives into a single one using a scalarized function, and constructing an approximated Pareto front. Turgay et al. (2018) focused on multi-objective contextual bandits, achieving a Pareto regret bound of $\widetilde{O}(T^{(d_p+1)/(d_p+2)})$, where d_p is the Pareto zooming dimension. Lu et al. (2019) improved this Pareto regret bound to $\widetilde{O}(d\sqrt{T})$ for multi-objective generalized linear bandits. Additionally, Chowdhury & Gopalan (2021) explored multi-objective bandits within the context of non-parametric Bayesian optimization. Another line of research involves Pareto front identification, which examines the cost of identifying all Pareto optimal arms (Auer et al., 2016; Ararat & Tekin, 2023; Kone et al., 2023). In the domain of lexicographic bandits, Tekin & Turgay (2018) studied contextual bandits with two objectives and established a regret bound of $\widetilde{O}(T^{(d_c+2)/(d_c+3)})$, where d_c is the dimension of contextual vectors. This work was generalized by Hüyük & Tekin (2021) to consider more than two objectives, leading to a priority-based regret bound of $\widetilde{O}((KT)^{\frac{2}{3}})$ for lexicographic MOMAB. Recently, Xue et al. (2024) investigated the

lexicographic Lipschitz bandit problem, proposing a regret bound of $\widetilde{O}(T^{(d_z^i+1)/(d_z^i+2)})$ for the i-th objective, where d_z^i is the zooming dimension.

Existing algorithms for multi-objective bandits (Lu et al., 2019; Hüyük & Tekin, 2021) fail to adapt to the reward variance and therefore incur high regret even for simple bandit problems. Furthermore, current regret bounds for non-parametric contextual bandits (Tekin & Turgay, 2018; Xue et al., 2024) grow nearly linearly when d_c and d_z^i are large. To address these limitations, we propose variance-aware algorithms for the parametric lexicographic GLB.

B Experimental Setting

In this section, we present the implementation details of our experiments. All experiments were conducted on a Windows 10 laptop equipped with an Intel(R) Core(TM) i7-11700 CPU and 32 GB of memory.

Synthetic Dataset. We set $K = d = 10$. For each round $t \geq 1$, the arm set is fixed as $\mathcal{A}_t = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_K\} \subseteq \mathbb{R}^d$. For any $k \in [K]$, the arm vector $\tilde{\boldsymbol{x}}_k$ is set to the k-th standard basis vector in $\mathbb{R}^d$, whose k-th element is 1 and whose remaining elements are 0. The number of objectives is set to $m = 3$, with the parameter vectors given by $\boldsymbol{\theta}_*^i = [\theta_*^i(1), \theta_*^i(2), \ldots, \theta_*^i(10)], i \in [3]$. The elements of $\boldsymbol{\theta}_*^1$, $\boldsymbol{\theta}_*^2$ and $\boldsymbol{\theta}_*^3$ are defined as $\theta_*^1(k) = 5 - 5 \times \min_{p \in \{1,4,8\}} |k-p|$, $\theta_*^2(k) = 5 - 10 \times \min_{p \in \{3,7\}} |k-p|$ and $\theta_*^3(k) = 5 - 10 \times |k - 7.5|$, respectively, for $k \in [10]$. The expected reward for arm $\tilde{\boldsymbol{x}}_k$ and objective $i \in [m]$ is

$$\mu^i(\tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta}_*^i) = 5 \times \frac{1}{1 + e^{-\tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta}_*^i}}.$$

The design of $\boldsymbol{\theta}_*^1$ makes $\{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_4, \tilde{\boldsymbol{x}}_8\}$ the optimal arms for the first objective. Among these, $\{\tilde{\boldsymbol{x}}_4, \tilde{\boldsymbol{x}}_8\}$ are optimal for both the first and second objectives. Therefore, all three objectives must be considered to find the lexicographic optimal arm $\{\tilde{\boldsymbol{x}}_8\}$. After playing $\boldsymbol{x}_t$, the reward of the i-th objective is drawn from a Gaussian distribution with mean $\mu^i(\tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta}_*^i)$ and variance σ_t^i. To reduce the randomness, each algorithm is run 10 times, and the average regrets for the first and third objectives are reported. The regret curves for the second objective are omitted, as Theorems 2 and 4 show that its regret falls between those of the first and third objectives.

Real-world Dataset. Following prior work on multi-objective bandits (Auer et al., 2016), we evaluate the performance of our algorithms on the real-world SW-LLVM dataset (Zuluaga et al., 2016). This dataset comprises 1,024 operating points, each characterized by 11 binary features representing different compiler settings. The corresponding objectives are the performance and memory footprint of software compiled with these settings. As in Auer et al. (2016), we select the four most informative binary features (features 7–10) to define a subset of 16 operating points for our experiments, enabling the generation of stochastic data. Each operating point comprises 64 samples, from which we compute the empirical means and variances. These statistics are then used as the parameters of Gaussian distributions to simulate stochastic rewards. Each algorithm is executed 10 times, and we report the average regrets with respect to both performance and memory footprint. The (expected reward, variance) pairs for all arms are summarized in Table 2, where Arm 16 represents the lexicographic optimal arm.

Table 2: Expected Rewards and Variances of Each Arm

Index	First Objective	Second Objective	Index	First Objective	Second Objective
1	(210.17, 43.66)	(27.41, 1.87)	9	(230.70, 26.78)	(28.42, 0.31)
2	(210.72, 44.63)	(28.61, 0.36)	10	(231.69, 26.52)	(28.20, 0.32)
3	(235.71, 535.88)	(13.31, 1.71)	11	(246.49, 26.80)	(28.13, 0.11)
4	(238.35, 582.37)	(13.20, 1.57)	12	(247.59, 25.07)	(28.00, 0.00)
5	(227.38, 46.70)	(28.52, 0.41)	13	(243.60, 51.89)	(15.31, 0.21)
6	(239.27, 104.00)	(17.78, 1.05)	14	(257.72, 50.27)	(19.19, 0.34)
7	(228.08, 49.54)	(28.61, 0.24)	15	(244.85, 50.46)	(15.22, 0.17)
8	(239.89, 102.39)	(18.17, 0.27)	**16**	(259.02, 46.71)	(19.33, 0.41)

C Proof of Theorem 1

To construct the confidence intervals, we adopt a proof framework inspired by those used in single-objective settings (Zhao et al., 2023a; Di et al., 2024). However, our model (e.g., GLB) differs from those in previous single-objective studies (e.g., SLB and dueling bandits), and we address the problem in a multi-objective setting. Therefore, we present a complete proof to ensure rigor and clarity. We begin by introducing some necessary notations. For any $s \in [S], t \in [T]$, and $\tau \in \Psi_{t,s}$, we define the following terms:

$$\begin{aligned} g_\tau^i(\boldsymbol{\theta}) &= \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}) - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i), \\ G_{t,s}^i(\boldsymbol{\theta}) &= 2^{-2s}\kappa\boldsymbol{\theta} + \sum_{\tau \in \Psi_{t,s}} w_\tau^2 \cdot g_\tau^i(\boldsymbol{\theta}) \cdot \boldsymbol{x}_\tau, \\ \epsilon_\tau^i &= y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i), \\ \boldsymbol{z}_{t,s}^i &= \sum_{\tau \in \Psi_{t,s}} w_\tau^2 \epsilon_\tau^i \boldsymbol{x}_\tau. \end{aligned} \tag{8}$$

Next, we provide the following lemma to establish the confidence region based on prior knowledge of the variance.

Lemma 1 *With probability at least $1-\delta$, for any $i \in [m]$ and $s \in [S]$,*

$$\|\boldsymbol{\theta}_*^i - \hat{\boldsymbol{\theta}}_{t,s}^i\|_{\mathbf{V}_{t,s}} \le \beta_{t,s}^i, \forall t \ge 1$$

where

$$\beta_{t,s}^i = \frac{16 \cdot 2^{-s}}{\kappa} \sqrt{\sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 \ln(4t^2 m S/\delta)} + \frac{6 \cdot 2^{-s}}{\kappa} R \ln(4t^2 m S/\delta) + 2^{-s} B.$$

Proof: Using the notation from Eq. (8), we have $G_{t,s}^i(\boldsymbol{\theta}_*^i) = 2^{-2s}\kappa\boldsymbol{\theta}_*^i$, which implies

$$G_{t,s}^i(\hat{\boldsymbol{\theta}}_{t,s}^i) - G_{t,s}^i(\boldsymbol{\theta}_*^i) = 2^{-2s}\kappa(\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i) + \sum_{\tau \in \Psi_{t,s}} w_\tau^2 \cdot g_\tau^i(\hat{\boldsymbol{\theta}}_{t,s}^i) \cdot \boldsymbol{x}_\tau. \tag{9}$$

Applying the mean value theorem, there exists some $\eta \in [0,1]$ such that, with $\bar{\boldsymbol{\theta}}_{t,s}^i = \eta\hat{\boldsymbol{\theta}}_{t,s}^i + (1-\eta)\boldsymbol{\theta}_*^i$, we have the following identity:

$$g_\tau^i(\hat{\boldsymbol{\theta}}_{t,s}^i) = \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i) - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i) = \dot{\mu}^i(\boldsymbol{x}_\tau^\top \bar{\boldsymbol{\theta}}_{t,s}^i) \cdot \boldsymbol{x}_\tau^\top (\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i).$$

Substituting this into Eq. (9), we obtain

$$G_{t,s}^i(\hat{\boldsymbol{\theta}}_{t,s}^i) - G_{t,s}^i(\boldsymbol{\theta}_*^i) = 2^{-2s}\kappa(\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i) + \sum_{\tau \in \Psi_{t,s}} w_\tau^2 \cdot \dot{\mu}^i(\boldsymbol{x}_\tau^\top \bar{\boldsymbol{\theta}}_{t,s}^i) \cdot \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\top \cdot (\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i).$$

Letting $\mathbf{F}(\bar{\boldsymbol{\theta}}_{t,s}^i) = 2^{-2s}\kappa\mathbf{I} + \sum_{\tau \in \Psi_{t,s}} w_\tau^2 \cdot \dot{\mu}^i(\boldsymbol{x}_\tau^\top \bar{\boldsymbol{\theta}}_{t,s}^i) \cdot \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\top$, the above equation simplifies to

$$G_{t,s}^i(\hat{\boldsymbol{\theta}}_{t,s}^i) - G_{t,s}^i(\boldsymbol{\theta}_*^i) = \mathbf{F}(\bar{\boldsymbol{\theta}}_{t,s}^i)(\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i).$$

Thus, we have

$$\begin{aligned} \|G_{t,s}^i(\hat{\boldsymbol{\theta}}_{t,s}^i) - G_{t,s}^i(\boldsymbol{\theta}_*^i)\|_{\mathbf{V}_{t,s}^{-1}}^2 &= (\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i)^\top \mathbf{F}(\bar{\boldsymbol{\theta}}_{t,s}^i) \mathbf{V}_{t,s}^{-1} \mathbf{F}(\bar{\boldsymbol{\theta}}_{t,s}^i)(\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i) \\ &\ge \kappa^2 (\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i)^\top \mathbf{V}_{t,s} (\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i) = \kappa^2 \|\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t,s}}^2 \end{aligned} \tag{10}$$

where the inequality follows from the fact that $\mathbf{F}(\bar{\boldsymbol{\theta}}_{t,s}^i) \succeq \kappa \mathbf{V}_{t,s}$.

Recall from Eq. (3) that $\hat{\boldsymbol{\theta}}_{t,s}^i$ satisfies the following equation:

$$2^{-2s}\kappa\hat{\boldsymbol{\theta}}_{t,s}^i + \sum_{\tau \in \Psi_{t,s}} w_\tau^2 (\mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i) - y_\tau^i)\boldsymbol{x}_\tau = \mathbf{0}.$$

Using the definition $g_\tau^i(\boldsymbol{\theta}) = \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}) - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)$ from Eq. (8), we obtain

$$G_{t,s}^i(\hat{\boldsymbol{\theta}}_{t,s}^i) = 2^{-2s}\kappa\hat{\boldsymbol{\theta}}_{t,s}^i + \sum_{\tau \in \Psi_{t,s}} w_\tau^2 \cdot g_\tau^i(\hat{\boldsymbol{\theta}}_{t,s}^i) \cdot \boldsymbol{x}_\tau = \sum_{\tau \in \Psi_{t,s}} w_\tau^2 (y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i))\boldsymbol{x}_\tau = \boldsymbol{z}_{t,s}^i.$$

Given that $G_{t,s}^i(\boldsymbol{\theta}_*^i) = 2^{-2s}\kappa\boldsymbol{\theta}_*^i$, it follows that

$$G_{t,s}^i(\hat{\boldsymbol{\theta}}_{t,s}^i) - G_{t,s}^i(\boldsymbol{\theta}_*^i) = \boldsymbol{z}_{t,s}^i - 2^{-2s}\kappa\boldsymbol{\theta}_*^i. \tag{11}$$

Substituting Eq. (11) into Eq. (10), we derive

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t,s}} &\leq \|\frac{1}{\kappa}\boldsymbol{z}_{t,s}^i - 2^{-2s}\boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t,s}^{-1}} \leq 2^{-2s}\|\boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t,s}^{-1}} + \frac{1}{\kappa}\|\boldsymbol{z}_{t,s}^i\|_{\mathbf{V}_{t,s}^{-1}} \\ &\leq 2^{-s}B + \frac{1}{\kappa}\|\boldsymbol{z}_{t,s}^i\|_{\mathbf{V}_{t,s}^{-1}}.\end{aligned} \tag{12}$$

Next, we bound the term $\|\boldsymbol{z}_{t,s}^i\|_{\mathbf{V}_{t,s}^{-1}}$, where $\boldsymbol{z}_{t,s}^i = \sum_{\tau\in\Psi_{t,s}} w_\tau^2\epsilon_\tau^i\boldsymbol{x}_\tau$ as defined in Eq. (8). Given that the variance of ϵ_τ^i is $(\sigma_\tau^i)^2$ and $w_\tau \leq 1$ for all $\tau \in \Psi_{t,s}$, we have

$$\mathrm{E}[(w_\tau\epsilon_\tau^i)^2|\mathcal{G}_{\tau-1}] \leq (w_\tau\sigma_\tau^i)^2, |w_\tau\epsilon_\tau^i| \leq |\epsilon_\tau^i| \leq R.$$

For a fixed objective $i \in [m]$ and stage $s \in [S]$, since $\|w_t\boldsymbol{x}_t\|_{\mathbf{V}_{t,s}^{-1}} = 2^{-s}$, we can apply Theorem 1 of Zhao et al. (2023a) to conclude that with probability at least $1-\delta$, for any $t \geq 1$,

$$\|\boldsymbol{z}_{t,s}^i\|_{\mathbf{V}_{t,s}^{-1}} \leq 16\cdot 2^{-s}\sqrt{\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\ln(4t^2/\delta)} + 6\cdot 2^{-s}R\ln(4t^2/\delta).$$

By applying a union bound over the m objectives and S stages, we obtain that with probability at least $1-\delta$, for any $i \in [m]$, $s \in [S]$ and $t \geq 1$,

$$\|\boldsymbol{z}_{t,s}^i\|_{\mathbf{V}_{t,s}^{-1}} \leq 16\cdot 2^{-s}\sqrt{\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\ln(4t^2mS/\delta)} + 6\cdot 2^{-s}R\ln(4t^2mS/\delta).$$

Substituting this result into Eq. (12), we get

$$\|\hat{\boldsymbol{\theta}}_{t,s}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t,s}} \leq \frac{16\cdot 2^{-s}}{\kappa}\sqrt{\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\ln(4t^2mS/\delta)} + \frac{6\cdot 2^{-s}}{\kappa}R\ln(4t^2mS/\delta) + 2^{-s}B.$$

This completes the proof of Lemma 1. □

Although Lemma 1 establishes a confidence region for the estimator in Algorithm 1, its radius depends on the variance term $(\sigma_\tau^i)^2$, which is unknown. To eliminate this dependence, we first bound the variance term using $(\epsilon_\tau^i)^2$, as shown in the following lemma.

Lemma 2 *With probability at least $1-\delta$, for any $i \in [m]$, $s \in [S]$ and $t \geq 1$,*

$$\begin{aligned}\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2 &\leq 2\sum_{\tau\in\Psi_{t,s}}(w_\tau\epsilon_\tau^i)^2 + \frac{14}{3}R^2\ln(4t^2mS/\delta), \\ \sum_{\tau\in\Psi_{t,s}}(w_\tau\epsilon_\tau^i)^2 &\leq \frac{3}{2}\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2 + \frac{7}{3}R^2\ln(4t^2mS/\delta).\end{aligned}$$

Proof: Given that for all $\tau \in \Psi_{t,s}$, the variance of ϵ_τ^i is $(\sigma_\tau^i)^2$ and $|\epsilon_\tau^i| \leq R$, we have

$$\begin{aligned}&\mathrm{E}[(\epsilon_\tau^i)^2 - (\sigma_\tau^i)^2|\mathcal{G}_{\tau-1}] = 0, \\ \sum_{\tau\in\Psi_{t,s}}\mathrm{E}[w_\tau^2((\epsilon_\tau^i)^2 - (\sigma_\tau^i)^2)^2|\mathcal{G}_{\tau-1}] &\leq \sum_{\tau\in\Psi_{t,s}}\mathrm{E}[w_\tau^2(\epsilon_\tau^i)^4|\mathcal{G}_{\tau-1}] \leq R^2\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2.\end{aligned}$$

For a fixed $i \in [m]$ and $s \in [S]$, applying Freedman's inequality (Freedman, 1975) to ϵ_τ^i for all $\tau \in \Psi_{t,s}$, we have that with probability at least $1-\delta$, for all $t \geq 1$, the following inequality holds:

$$\left|\sum_{\tau\in\Psi_{t,s}} w_\tau^2((\epsilon_\tau^i)^2 - (\sigma_\tau^i)^2)\right| \leq \sqrt{2R^2\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\ln(4t^2/\delta)} + \frac{4}{3}R^2\ln(4t^2/\delta).$$

Taking a union bound over all $i \in [m]$ and $s \in [S]$, and substituting δ/mS for δ, we have that with probability at least $1-\delta$, for all $i \in [m]$, $s \in [S]$ and $t \geq 1$, the following inequality holds:

$$\left| \sum_{\tau \in \Psi_{t,s}} w_\tau^2((\epsilon_\tau^i)^2 - (\sigma_\tau^i)^2) \right| \leq \sqrt{2R^2 \sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 \ln(4t^2 mS/\delta)} + \frac{4}{3} R^2 \ln(t^2 mS/\delta). \tag{13}$$

Using Young's inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$, we obtain

$$\sum_{\tau \in \Psi_{t,s}} w_\tau^2((\epsilon_\tau^i)^2 - (\sigma_\tau^i)^2) \leq \frac{1}{2} \sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 + \frac{7}{3} R^2 \ln(4t^2 mS/\delta).$$

Thus, we obtain

$$\sum_{\tau \in \Psi_{t,s}} (w_\tau \epsilon_\tau^i)^2 \leq \frac{3}{2} \sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 + \frac{7}{3} R^2 \ln(4t^2 mS/\delta).$$

Reapplying Young's inequality to Eq. (13), we derive

$$\sum_{\tau \in \Psi_{t,s}} w_\tau^2((\sigma_\tau^i)^2 - (\epsilon_\tau^i)^2) \leq \frac{1}{2} \sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 + \frac{7}{3} R^2 \ln(4t^2 mS/\delta).$$

Hence, we conclude:

$$\sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 \leq 2 \sum_{\tau \in \Psi_{t,s}} (w_\tau \epsilon_\tau^i)^2 + \frac{14}{3} R^2 \ln(4t^2 mS/\delta).$$

This completes the proof of Lemma 2. □

Lemma 2 establishes an upper bound for the sum of variances. However, this bound depends on $\epsilon_\tau^i = y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)$, where $\boldsymbol{\theta}_*^i$ is unknown. Thus, we further bound this unobservable noise term using computable residuals and derive a computable radius for the confidence region, as detailed in the following lemma.

Lemma 3 *Let* $\bar{s}_t = \lceil \log_2(64(L/\kappa)\sqrt{\ln(4t^2 mS/\delta)}) \rceil$. *With probability at least* $1-\delta$, *for* $i \in [m]$, $t \geq 1$, *and* $s \geq \bar{s}_t$,

$$\sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 \leq 8 \sum_{\tau \in \Psi_{t,s}} w_\tau^2 (y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i))^2 + 18 R^2 \ln(4t^2 mS/\delta),$$

$$\sum_{\tau \in \Psi_{t,s}} w_\tau^2 (y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i))^2 \leq 4 \sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 + 7 R^2 \ln(4t^2 mS/\delta).$$

Proof: According to Lemma 2, for a fixed $s \in [S]$,

$$\sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 \leq 2 \sum_{\tau \in \Psi_{t,s}} (w_\tau \epsilon_\tau^i)^2 + \frac{14}{3} R^2 \ln(4t^2 mS/\delta).$$

Given that $\epsilon_\tau^i = y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)$ and applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain:

$$\begin{aligned} \sum_{\tau \in \Psi_{t,s}} (w_\tau \sigma_\tau^i)^2 \leq\ & 4 \sum_{\tau \in \Psi_{t,s}} w_\tau^2 (y_\tau^i - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i))^2 + 4 \sum_{\tau \in \Psi_{t,s}} w_\tau^2 (\mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i))^2 \\ & + \frac{14}{3} R^2 \ln(4t^2 mS/\delta). \end{aligned} \tag{14}$$

Now, we bound the term $\sum_{\tau \in \Psi_{t,s}} w_\tau^2 (\mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i))^2$. By Assumption 2, since $\dot{\mu}^i(\cdot) \leq L$, we have

$$\sum_{\tau \in \Psi_{t,s}} w_\tau^2 (\mu^i(\boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_{t,s}^i))^2 \leq \sum_{\tau \in \Psi_{t,s}} w_\tau^2 L^2 (\boldsymbol{\theta}_*^i - \hat{\boldsymbol{\theta}}_{t,s}^i)^\top \cdot \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\top \cdot (\boldsymbol{\theta}_*^i - \hat{\boldsymbol{\theta}}_{t,s}^i).$$

Rearranging this inequality yields:

$$\sum_{\tau\in\Psi_{t,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 \le L^2(\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i)^\top\cdot\sum_{\tau\in\Psi_{t,s}} w_\tau^2\boldsymbol{x}_\tau\boldsymbol{x}_\tau^\top\cdot(\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i).$$

Since $\mathbf{V}_{t,s}=2^{-2s}\mathbf{I}+\sum_{\tau\in\Psi_{t,s}} w_\tau^2\boldsymbol{x}_\tau\boldsymbol{x}_\tau^\top$, the preceding inequality further yields

$$\sum_{\tau\in\Psi_{t,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 \le L^2\|\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i\|_{\mathbf{V}_{t,s}}^2. \tag{15}$$

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

$$\|\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i\|_{\mathbf{V}_{t,s}} \le \frac{16\cdot 2^{-s}}{\kappa}\sqrt{\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\ln(4t^2mS/\delta)}+\frac{6\cdot 2^{-s}}{\kappa}R\ln(4t^2mS/\delta)+2^{-s}B.$$

Using the inequality $(a+b)^2\le 2a^2+2b^2$, we further bound Eq. (15) as

$$\begin{aligned}&\sum_{\tau\in\Psi_{t,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2\\ &\le \frac{2^{-2s}L^2}{\kappa^2}\left(512\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\cdot\ln(4t^2mS/\delta)+2\left(6R\ln(4t^2mS/\delta)+\kappa B\right)^2\right).\end{aligned}$$

Given that $s\ge\bar{s}_t=\lceil\log_2(64(L/\kappa)\sqrt{\ln(4t^2mS/\delta)})\rceil$, we obtain

$$\sum_{\tau\in\Psi_{t,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 \le \frac{1}{8}\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2+R^2\ln(4t^2mS/\delta). \tag{16}$$

Substituting Eq. (16) into Eq. (14), we have

$$\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2 \le 4\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2+\frac{1}{2}\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2+9R^2\ln(4t^2mS/\delta).$$

Rearranging the above inequality, we have

$$\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2 \le 8\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2+18R^2\ln(4t^2mS/\delta).$$

For the second inequality in Lemma 3, by applying the inequality $(a+b)^2\le 2a^2+2b^2$ to the term on the left-hand side, we obtain

$$\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 \le 2\sum_{\tau\in\Psi_{t,s}}(w_\tau\epsilon_\tau^i)^2+2\sum_{\tau\in\Psi_{t,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2. \tag{17}$$

According to Lemma 2, we have

$$\sum_{\tau\in\Psi_{t,s}}(w_\tau\epsilon_\tau^i)^2 \le \frac{3}{2}\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2+\frac{7}{3}R^2\ln(4t^2mS/\delta).$$

Substituting this result into Eq. (17), we obtain

$$\begin{aligned}\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 \le 3\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2+\frac{14}{3}R^2\ln(4t^2mS/\delta)\\ +2\sum_{\tau\in\Psi_{t,s}} w_\tau^2(\mu^i(\boldsymbol{x}_\tau^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2.\end{aligned} \tag{18}$$

By reusing Eq. (16) to bound the last term on the right-hand side of Eq. (18), we find

$$\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 \le 4\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2+7R^2\ln(4t^2mS/\delta).$$

This completes the proof of Lemma 3. □

Proof of Theorem 1. Equipped with Lemmas 1 and 3, we are ready to prove Theorem 1. Lemma 1 shows that

$$\|\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i\|_{\mathbf{V}_{t,s}} \leq \frac{16\cdot 2^{-s}}{\kappa}\sqrt{\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2\ln(4t^2mS/\delta)}+\frac{6\cdot 2^{-s}}{\kappa}R\ln(4t^2mS/\delta)+2^{-s}B. \quad (19)$$

Meanwhile, Lemma 3 implies that for any $s \geq \bar{s}_t$,

$$\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2 \leq 8\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2+18R^2\ln(4t^2mS/\delta). \quad (20)$$

For any $s < \bar{s}_t$, since $w_\tau \leq 1$ and $\sigma_\tau^i \leq R$, we can easily derive

$$\sum_{\tau\in\Psi_{t,s}}(w_\tau\sigma_\tau^i)^2 \leq |\Psi_{t,s}|\cdot R^2. \quad (21)$$

Substituting Eq. (20) and Eq. (21) into Eq. (19), we derive the confidence region radius that is independent of prior variance knowledge, given by

$$\|\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i\|_{\mathbf{V}_{t,s}} \leq \hat{\beta}_{t,s}^i$$

where

$$\hat{\beta}_{t,s}^i=\frac{16\cdot 2^{-s}}{\kappa}\sqrt{8\widehat{\mathrm{Var}}_{t,s}^i\cdot\gamma_t+18R^2\cdot\gamma_t^2}+\frac{6\cdot 2^{-s}}{\kappa}R\gamma_t+2^{-s}B, \gamma_t=\ln(4t^2mS/\delta)$$

and

$$\widehat{\mathrm{Var}}_{t,s}^i=\begin{cases}\sum_{\tau\in\Psi_{t,s}} w_\tau^2(y_\tau^i-\mu^i(\boldsymbol{x}_\tau^\top\hat{\boldsymbol{\theta}}_{t,s}^i))^2 & s\geq\bar{s}_t,\\ |\Psi_{t,s}|\cdot R^2 & \text{otherwise.}\end{cases}$$

This completes the proof of Theorem 1. □

D Proof of Theorem 2

Based on Theorem 1, we deduce the following confidence intervals for the expected rewards of each arm.

Lemma 4 *With probability at least $1-\delta$, for any $i\in[m]$ and $s\in[S]$,*

$$|\mu^i(\boldsymbol{x}^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^i)| \leq \hat{\beta}_{t,s}^i\cdot L\cdot\|\boldsymbol{x}\|_{\mathbf{V}_{t,s}^{-1}}, t\geq 1.$$

Proof: Given that $\dot{\mu}^i(\cdot)\leq L$ as stated in Assumption 2, we have

$$|\mu^i(\boldsymbol{x}^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^i)| \leq L\cdot|\boldsymbol{x}^\top\boldsymbol{\theta}_*^i-\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^i|.$$

Applying the Cauchy–Schwarz inequality, this inequality can be rewritten as

$$|\mu^i(\boldsymbol{x}^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^i)| \leq L\cdot\|\boldsymbol{x}\|_{\mathbf{V}_{t,s}^{-1}}\cdot\|\boldsymbol{\theta}_*^i-\hat{\boldsymbol{\theta}}_{t,s}^i\|_{\mathbf{V}_{t,s}}.$$

According to the confidence region specified in Lemma 1 and Theorem 1, with probability at least $1-\delta$, for any $i\in[m]$ and $s\in[S]$,

$$|\mu^i(\boldsymbol{x}^\top\boldsymbol{\theta}_*^i)-\mu^i(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^i)| \leq \hat{\beta}_{t,s}^i\cdot L\cdot\|\boldsymbol{x}\|_{\mathbf{V}_{t,s}^{-1}}, t\geq 1.$$

Thus, the proof of Lemma 4 is complete. □

Next, we present a lemma concerning the AAE-LGLB algorithm. This lemma establishes that sequentially eliminating arms from the first to the last objective does not result in the loss of the optimal arm, and that the remaining arms are promising.

Lemma 5 *Consider the AAE-LGLB algorithm run in VA-LGLB. Suppose that the input arm set $\mathcal{A}_{t,s}$ contains $\boldsymbol{x}_t^*$ and satisfies $\|\boldsymbol{x}\|_{\mathbf{V}_{t,s}^{-1}} \leq W$ for all $\boldsymbol{x} \in \mathcal{A}_{t,s}$. With probability at least $1-\delta$, $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^m$ and for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^m$,*

$$\mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top\boldsymbol{\theta}_*^i) \leq 4(1+\lambda+\lambda^2+\cdots+\lambda^{i-1})\cdot\bar{\beta}_{t,s}^i \cdot L \cdot W, i \in [m].$$

Proof: We use induction on the objective index $i \in [m]$ to prove this lemma. For the base case $i=1$, consider the candidate arm set $\mathcal{A}_{t,s}^0 = \mathcal{A}_{t,s}$ and the filtered arm set $\mathcal{A}_{t,s}^1$ defined as follows:

$$\mathcal{A}_{t,s}^1 = \{\boldsymbol{x} \in \mathcal{A}_{t,s}^0 | \mu^1(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^1) \geq \mu^1(\hat{\boldsymbol{x}}_{t,s}^{1\top}\hat{\boldsymbol{\theta}}_{t,s}^1) - 2\bar{\beta}_{t,s}^1 LW\}. \tag{22}$$

Since $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^0$ and $\hat{\boldsymbol{x}}_{t,s}^1 \in \mathcal{A}_{t,s}^0$, Lemma 4 implies that

$$\mu^1(\boldsymbol{x}_t^{*\top}\hat{\boldsymbol{\theta}}_{t,s}^1) \geq \mu^1(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^1) - \hat{\beta}_{t,s}^1 LW. \tag{23}$$

Since $\boldsymbol{x}_t^*$ is the optimal arm in $\mathcal{A}_{t,s}^0$, it follows that $\mu^1(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^1) \geq \mu^1(\hat{\boldsymbol{x}}_{t,s}^{1\top}\boldsymbol{\theta}_*^1)$. Therefore, Eq. (23) implies

$$\mu^1(\boldsymbol{x}_t^{*\top}\hat{\boldsymbol{\theta}}_{t,s}^1) \geq \mu^1(\hat{\boldsymbol{x}}_{t,s}^{1\top}\boldsymbol{\theta}_*^1) - \hat{\beta}_{t,s}^1 LW.$$

Applying Lemma 4 again, we further obtain

$$\mu^1(\boldsymbol{x}_t^{*\top}\hat{\boldsymbol{\theta}}_{t,s}^1) \geq \mu^1(\hat{\boldsymbol{x}}_{t,s}^{1\top}\hat{\boldsymbol{\theta}}_{t,s}^1) - 2\hat{\beta}_{t,s}^1 LW. \tag{24}$$

Combining Eq. (22) and Eq. (24) confirms that $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^1$ since $\hat{\beta}_{t,s}^1 = \bar{\beta}_{t,s}^1$.

Next, we need to show that $\mu^1(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^1) - \mu^1(\boldsymbol{x}^\top\boldsymbol{\theta}_*^1) \leq 4\bar{\beta}_{t,s}^1 LW, \forall \boldsymbol{x} \in \mathcal{A}_{t,s}^1$. Using the confidence bounds and elimination step in Eq. (22), for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^1$ we obtain

$$\mu^1(\boldsymbol{x}^\top\boldsymbol{\theta}_*^1) \geq \mu^1(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^1) - \hat{\beta}_{t,s}^1 LW \geq \mu^1(\hat{\boldsymbol{x}}_{t,s}^{1\top}\hat{\boldsymbol{\theta}}_{t,s}^1) - 3\hat{\beta}_{t,s}^1 LW.$$

Since $\hat{\boldsymbol{x}}_{t,s}^1 = \mathrm{argmax}_{\boldsymbol{x}\in\mathcal{A}_{t,s}^0} \mu^1(\boldsymbol{x}^\top\hat{\boldsymbol{\theta}}_{t,s}^1)$ and $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^0$, it follows that

$$\mu^1(\boldsymbol{x}^\top\boldsymbol{\theta}_*^1) \geq \mu^1(\hat{\boldsymbol{x}}_{t,s}^{1\top}\hat{\boldsymbol{\theta}}_{t,s}^1) - 3\hat{\beta}_{t,s}^1 LW \geq \mu^1(\boldsymbol{x}_t^{*\top}\hat{\boldsymbol{\theta}}_{t,s}^1) - 3\hat{\beta}_{t,s}^1 LW \geq \mu^1(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^1) - 4\hat{\beta}_{t,s}^1 LW.$$

This shows that $\mu^1(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^1) - \mu^1(\boldsymbol{x}^\top\boldsymbol{\theta}_*^1) \leq 4\hat{\beta}_{t,s}^1 LW$, completing the proof for the case $i=1$ since $\hat{\beta}_{t,s}^1 = \bar{\beta}_{t,s}^1$.

Following the standard framework of the induction method, we now prove that if $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^{i-1}$ and for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^{i-1}$,

$$\mu^j(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^j) - \mu^j(\boldsymbol{x}^\top\boldsymbol{\theta}_*^j) \leq 4(1+\lambda+\cdots+\lambda^{j-1})\bar{\beta}_{t,s}^j LW, j \in [i-1], \tag{25}$$

then $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^i$ and for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^i$,

$$\mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top\boldsymbol{\theta}_*^i) \leq 4(1+\lambda+\cdots+\lambda^{i-1})\bar{\beta}_{t,s}^i LW.$$

Assumption 4 states that

$$\mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) \geq \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top}\boldsymbol{\theta}_*^i) - \lambda \cdot \max_{j\in[i-1]}\{\mu^j(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^j) - \mu^j(\hat{\boldsymbol{x}}_{t,s}^{i\top}\boldsymbol{\theta}_*^j)\}.$$

Substituting Eq. (25) into the right-hand side of the above inequality yields

$$\mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) \geq \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top}\boldsymbol{\theta}_*^i) - 4\lambda\cdot(1+\lambda+\cdots+\lambda^{i-2})\bar{\beta}_{t,s}^{i-1} LW. \tag{26}$$

Using the confidence intervals given in Lemma 4, we have

$$\mu^i(\boldsymbol{x}_t^{*\top}\hat{\boldsymbol{\theta}}_{t,s}^i) + \hat{\beta}_{t,s}^i LW \geq \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i), \quad \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top}\boldsymbol{\theta}_*^i) \geq \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top}\hat{\boldsymbol{\theta}}_{t,s}^i) - \hat{\beta}_{t,s}^i LW.$$

Thus, substituting the preceding two inequalities into Eq. (26) and using $\bar{\beta}_{t,s}^i = \max_{j\in[i]}\hat{\beta}_{t,s}^j$, we obtain

$$\mu^i(\boldsymbol{x}_t^{*\top}\hat{\boldsymbol{\theta}}_{t,s}^i) \geq \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top}\hat{\boldsymbol{\theta}}_{t,s}^i) - 2\bar{\beta}_{t,s}^i LW - 4\lambda\cdot(1+\lambda+\cdots+\lambda^{i-2})\bar{\beta}_{t,s}^i LW.$$

This implies $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^i$, because $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^{i-1}$ and $\mathcal{A}_{t,s}^i$ is defined as follows:

$$\mathcal{A}_{t,s}^i = \{\boldsymbol{x} \in \mathcal{A}_{t,s}^{i-1} | \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_{t,s}^i) \geq \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top} \hat{\boldsymbol{\theta}}_{t,s}^i) - (2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1})\bar{\beta}_{t,s}^i LW\}. \quad (27)$$

Next, we prove that for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^i$,

$$\mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \leq 4(1 + \lambda + \cdots + \lambda^{i-1})\bar{\beta}_{t,s}^i LW.$$

Using the confidence interval in Lemma 4, we get that for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^{i-1}$,

$$\mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \geq \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_{t,s}^i) - \hat{\beta}_{t,s}^i LW, \quad \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_{t,s}^i) \geq \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) - \hat{\beta}_{t,s}^i LW. \quad (28)$$

Combining the first inequality in Eq. (28) with the filtering operation in Eq. (27), we get that for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^i$,

$$\mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \geq \mu^i(\hat{\boldsymbol{x}}_{t,s}^{i\top} \hat{\boldsymbol{\theta}}_{t,s}^i) - \hat{\beta}_{t,s}^i LW - (2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1})\bar{\beta}_{t,s}^i LW.$$

Since $\hat{\boldsymbol{x}}_{t,s}^i = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{A}_{t,s}^{i-1}} \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_{t,s}^i)$ and $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^{i-1}$, this inequality implies

$$\mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \geq \mu^i(\boldsymbol{x}_t^{*\top} \hat{\boldsymbol{\theta}}_{t,s}^i) - \hat{\beta}_{t,s}^i LW - (2 + 4\lambda + 4\lambda^2 + \cdots + 4\lambda^{i-1})\bar{\beta}_{t,s}^i LW.$$

Finally, using the second inequality in Eq. (28) and $\bar{\beta}_{t,s}^i = \max_{j \in [i]} \hat{\beta}_{t,s}^j$ to further bound the right-hand side of the preceding inequality, we obtain

$$\mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \geq \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - 4(1 + \lambda + \lambda^2 + \cdots + \lambda^{i-1})\bar{\beta}_{t,s}^i LW.$$

Thus, the proof of Lemma 5 is complete. □

Lemma 6 *With probability at least $1 - \delta$, for any $t \geq 1$ and $s \geq 1$, $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}$.*

Proof. We use induction on the stage index s to prove this lemma. For $s = 1$, $\boldsymbol{x}_t^* \in \mathcal{A}_{t,1}$ clearly holds since $\mathcal{A}_{t,1} = \mathcal{A}_t$. For $s \geq 1$, if $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}$, then by Lemma 5, $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s+1}$. Therefore, $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}$ for all $s \geq 1$. □

Lemma 7 *For the VA-LGLB algorithm, let s_t be the stage index in the **if** case at round t. With probability at least $1 - \delta$, for any $t \geq 1$ and $\boldsymbol{x} \in \mathcal{A}_{t,T}$,*

$$\mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \leq 4(1 + \lambda + \cdots + \lambda^{i-1}) \cdot \bar{\beta}_{t,s_t}^i \cdot L \cdot T^{-\frac{3}{2}}, i \in [m].$$

Proof. By Lemma 6, we have $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s_t}$. Therefore, Lemma 5 can be applied to the operation $\mathcal{A}_{t,T} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s_t}, \{\hat{\beta}_{t,s_t}^i, \hat{\boldsymbol{\theta}}_{t,s_t}^i\}_{i \in [m]}, T^{-\frac{3}{2}}\right)$ in the **if** case. Setting $W = T^{-\frac{3}{2}}$ in Lemma 5 completes the proof of Lemma 7. □

Lemma 8 *For the VA-LGLB algorithm, with probability at least $1 - \delta$, for any $t \geq 1$, $s \geq 2$ and $\boldsymbol{x} \in \mathcal{A}_{t,s}$,*

$$\mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \leq 8(1 + \lambda + \cdots + \lambda^{i-1}) \cdot \bar{\beta}_{t,s-1}^i \cdot L \cdot 2^{-s}, i \in [m].$$

Proof. For $s \geq 2$, Lemma 6 implies that $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s-1}$. Therefore, we can apply Lemma 5 to the set $\mathcal{A}_{t,s} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s-1}, \{\hat{\beta}_{t,s-1}^i, \hat{\boldsymbol{\theta}}_{t,s-1}^i\}_{i \in [m]}, 2^{-s+1}\right)$ in the **else** case. Setting $W = 2^{-s+1}$ in Lemma 5 completes the proof of Lemma 8. □

Proof of Theorem 2. Define $\Psi_{T+1,0} = [T] \setminus \bigcup_{s \in [S]} \Psi_{T+1,s}$ to represent the trials conducted in the **if** case. According to Lemma 7, we have

$$\sum_{t \in \Psi_{T+1,0}} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \leq \sum_{t \in \Psi_{T+1,0}} 4\Lambda_i L \bar{\beta}_{t,s_t}^i \cdot T^{-\frac{3}{2}} \leq 4\Lambda_i L \cdot \max_{s \in [S]} \bar{\beta}_{T,s}^i \cdot T^{-\frac{1}{2}}.$$

Using the definition of $\hat{\beta}^i_{T,s}$ in Theorem 1 and the identity $\bar{\beta}^i_{T,s} = \max_{j\in[i]} \hat{\beta}^j_{T,s}$, we obtain

$$\begin{aligned}
&\sum_{t\in\Psi_{T+1,0}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \\
&\leq 4\Lambda_i L\left(\frac{16}{\kappa}\sqrt{16(U+R)^2\gamma_T + 18R^2\gamma_T^2} + \frac{6}{\kappa}R\gamma_T + B\right) \\
&\leq 4\Lambda_i L\left(\frac{64}{\kappa}\sqrt{(U+R)^2\gamma_T} + \frac{86}{\kappa}R\gamma_T + B\right).
\end{aligned} \tag{29}$$

Next, we analyze the trials in the set $\Psi_{T+1,s}$. Based on the radius of the confidence region presented in Theorem 1, we consider two cases.

Case 1: $s \leq \lceil \log_2(64(L/\kappa)\sqrt{\gamma_T})\rceil$. Given that $w_t = 2^{-s}/\|\boldsymbol{x}_t\|_{\mathbf{V}^{-1}_{t,s}}$, we have

$$\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \leq 2U|\Psi_{T+1,s}| = 2U\cdot 2^{2s}\sum_{t\in\Psi_{T+1,s}} \|w_t\boldsymbol{x}_t\|^2_{\mathbf{V}^{-1}_{t,s}}.$$

Using $\mathbf{V}_{t,s} = 2^{-2s}\mathbf{I} + \sum_{\tau\in\Psi_{t,s}} w_\tau^2\cdot\boldsymbol{x}_\tau\boldsymbol{x}_\tau^\top$ and applying Lemma 11 from Abbasi-yadkori et al. (2011), we obtain

$$\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \leq 4U\cdot 2^{2s}\cdot d\ln(1+2^{2s+2}T/d).$$

Since $s \leq \lceil \log_2(64(L/\kappa)\sqrt{\gamma_T})\rceil$, we can further simplify this inequality to

$$\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \leq 16U\cdot(64L/\kappa)^2\gamma_T\cdot d\ln(1+2^{2s+2}T/d). \tag{30}$$

Case 2: $s > \lceil \log_2(64(L/\kappa)\sqrt{\gamma_T})\rceil$. According to Lemma 8, we have

$$\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \leq 8\sum_{t\in\Psi_{T+1,s}} \Lambda_i\bar{\beta}^i_{t,s-1}L\cdot 2^{-s}.$$

Using $w_t = 2^{-s}/\|\boldsymbol{x}_t\|_{\mathbf{V}^{-1}_{t,s}}$ and Lemma 11 of Abbasi-yadkori et al. (2011), we obtain

$$\begin{aligned}
\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) &\leq 8\sum_{t\in\Psi_{T+1,s}} \Lambda_i\bar{\beta}^i_{t,s-1}L\cdot 2^s\cdot\|w_t\boldsymbol{x}_t\|^2_{\mathbf{V}^{-1}_{t,s}} \\
&\leq 8\Lambda_i\bar{\beta}^i_{T,s-1}L\cdot 2^s\cdot 2d\ln(1+2^{2s+2}T/d).
\end{aligned}$$

Substituting the confidence radius from Theorem 1 for $\hat{\beta}^i_{T,s-1}$ and using $\bar{\beta}^i_{T,s-1} = \max_{j\in[i]}\hat{\beta}^j_{T,s-1}$, we obtain

$$\begin{aligned}
&\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \\
&\leq 16\Lambda_i Ld\ln(1+2^{2s+2}T/d)\cdot\left(\frac{16}{\kappa}\sqrt{8\max_{j\in[i]}\widehat{\mathrm{Var}}^j_{T,s-1}\cdot\gamma_T + 18R^2\gamma_T^2} + \frac{6}{\kappa}R\gamma_T + B\right).
\end{aligned}$$

Replacing $\widehat{\mathrm{Var}}^j_{T,s-1}$ with its upper bound from Lemma 3, we obtain

$$\begin{aligned}
&\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}^i_*) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}^i_*) \\
&\leq 16\Lambda_i Ld\ln(1+2^{2s+2}T/d)\left(\frac{16}{\kappa}\sqrt{32\max_{j\in[i]}\sum_{\tau\in\Psi_{T+1,s-1}}(w_\tau\sigma^j_\tau)^2\cdot\gamma_T + 74R^2\gamma_T^2} + \frac{6}{\kappa}R\gamma_T + B\right) \\
&\leq 16\Lambda_i Ld\ln(1+2^{2s+2}T/d)\left(\frac{64}{\kappa}\sqrt{2\sum_{\tau\in\Psi_{T+1,s-1}}(w_\tau\sigma^i_\tau)^2\cdot\gamma_T} + \frac{150}{\kappa}R\gamma_T + B\right).
\end{aligned} \tag{31}$$

To finish the proof of Theorem 2, we decompose the regret as follows:

$$\begin{aligned}R^i(T) &= \sum_{t\in\Psi_{T+1,0}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}_*^i) + \sum_{s=1}^{S}\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}_*^i)\\ &= \sum_{t\in\Psi_{T+1,0}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}_*^i) + \sum_{s=1}^{\bar{s}_T}\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}_*^i)\\ &\quad + \sum_{s=\bar{s}_T+1}^{S}\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}_*^i).\end{aligned}$$

Combining the results from Eq. (29), Eq. (30) and Eq. (31), we have

$$\begin{aligned}R^i(T) \le\ & 4\Lambda_i L\left(\frac{64}{\kappa}\sqrt{(U+R)^2\gamma_T} + \frac{86}{\kappa}R\gamma_T + B\right)\\ &+ 16SU\cdot(64L/\kappa)^2\gamma_T\cdot d\ln(1+2^{2S+2}T/d)\\ &+ \sum_{s=1}^{S} 16\Lambda_i L d\ln(1+2^{2s+2}T/d)\left(\frac{64}{\kappa}\sqrt{2\sum_{\tau\in\Psi_{T+1,s-1}}(w_\tau\sigma_\tau^i)^2\cdot\gamma_T} + \frac{150}{\kappa}R\gamma_T + B\right).\end{aligned}$$

By simplifying the second and third terms, we get

$$\begin{aligned}R^i(T) \le\ & 4\Lambda_i L\left(\frac{64}{\kappa}\sqrt{(U+R)^2\gamma_T} + \frac{86}{\kappa}R\gamma_T + B\right) + 32S^2U\cdot(64L/\kappa)^2\gamma_T\cdot d\ln(1+4T/d)\\ &+ 32S\Lambda_i L d\ln(1+4T/d)\sum_{s=1}^{S}\left(\frac{64}{\kappa}\sqrt{2\sum_{\tau\in\Psi_{T+1,s-1}}(w_\tau\sigma_\tau^i)^2\cdot\gamma_T} + \frac{150}{\kappa}R\gamma_T + B\right).\end{aligned}$$

Applying the Cauchy-Schwarz inequality to the third term, we obtain

$$\begin{aligned}R^i(T) &\le 4\Lambda_i\left(\frac{64L}{\kappa}\sqrt{(U+R)^2\gamma_T} + \frac{86L}{\kappa}R\gamma_T + LB\right)\\ &\quad + 32S^2U\cdot(64L/\kappa)^2\gamma_T\cdot d\ln(1+4T/d)\\ &\quad + 32S\Lambda_i L d\ln(1+4T/d)\sqrt{S}\left(\frac{64}{\kappa}\sqrt{2\sum_{t\in[T]}(w_t\sigma_t^i)^2\cdot\gamma_T} + \frac{150}{\kappa}R\gamma_T + B\right)\\ &= \widetilde{O}\left(\Lambda_i\cdot d\left(\frac{L}{\kappa}\sqrt{\sum_{t=1}^{T}(\sigma_t^i)^2} + \frac{L^2}{\kappa^2} + L\right)\right).\end{aligned}$$

This completes the proof of Theorem 2. □

E Regret Analysis Without Eq. (2)

In the main text, for simplicity in algorithm design and discussion, we assume that lower-priority objectives exhibit greater variance, as formalized in Eq. (2). In this section, we provide a corollary that removes this assumption.

Corollary 1 *Let $\Lambda_i = 1+\lambda+\cdots+\lambda^{i-1}$ and $\bar{\sigma}_t^i = \max_{j\in[i]}\sigma_t^j$ for $i\in[m]$. If Eq. (2) does not hold, then with probability at least $1-\delta$, for any objective $i\in[m]$, the regret of VA-LGLB satisfies*

$$R^i(T) \le \widetilde{O}\left(\Lambda_i\cdot d\left(\frac{L}{\kappa}\sqrt{\sum_{t=1}^{T}(\bar{\sigma}_t^i)^2} + \frac{L^2}{\kappa^2} + L\right)\right).$$

Proof: Recall Eq. (31) from the proof of Theorem 2, which states that for all $s \in [S]$ and $i \in [m]$,

$$\begin{aligned}
&\sum_{t \in \Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \\
&\leq 16\Lambda_i L d \ln(1 + 2^{2s+2}T/d) \left(\frac{16}{\kappa} \sqrt{32 \max_{j \in [i]} \sum_{\tau \in \Psi_{T+1,s}} (w_\tau \sigma_\tau^j)^2 \cdot \gamma_T + 74R^2\gamma_T^2} + \frac{6}{\kappa} R\gamma_T + B \right) \\
&\leq 16\Lambda_i L d \ln(1 + 2^{2s+2}T/d) \left(\frac{64}{\kappa} \sqrt{2 \sum_{\tau \in \Psi_{T+1,s}} (w_\tau \bar{\sigma}_\tau^i)^2 \cdot \gamma_T} + \frac{150}{\kappa} R\gamma_T + B \right).
\end{aligned} \tag{32}$$

Following the same procedure as in the proof of Theorem 2, we derive

$$\begin{aligned}
R^i(T) &\leq 4\Lambda_i \left(\frac{64L}{\kappa} \sqrt{(U+R)^2 \gamma_T} + \frac{86L}{\kappa} R\gamma_T + LB \right) \\
&\quad + 32S^2 U \cdot (64L/\kappa)^2 \gamma_T \cdot d \ln(1 + 4T/d) \\
&\quad + 32S\Lambda_i L d \ln(1 + 4T/d) \sqrt{S} \left(\frac{64}{\kappa} \sqrt{2 \sum_{t \in [T]} (w_t \bar{\sigma}_t^i)^2 \cdot \gamma_T} + \frac{150}{\kappa} R\gamma_T + B \right) \\
&= \widetilde{O} \left(\Lambda_i \cdot d \left(\frac{L}{\kappa} \sqrt{\sum_{t=1}^T (\bar{\sigma}_t^i)^2} + \frac{L^2}{\kappa^2} + L \right) \right).
\end{aligned}$$

This completes the proof of Corollary 1. □

F Proof of Theorem 3

To construct the confidence intervals, we adopt a proof framework inspired by that used in the single-objective setting (Xue et al., 2023). However, since our paper assumes bounded rewards, whereas the single-objective work considers heavy-tailed rewards (Xue et al., 2023), we provide a complete proof to ensure the rigor and clarity of our exposition. Specifically, for any $i \in [m]$ and $t \in [T]$, we define the loss function of the action-reward pair $(\boldsymbol{x}_t, y_t^i)$ and its conditional expectation as follows:

$$\ell_t^i(\boldsymbol{\theta}) = -y_t^i \cdot \boldsymbol{x}_t^\top \boldsymbol{\theta} + \int_0^{\boldsymbol{x}_t^\top \boldsymbol{\theta}} \mu^i(z) dz, f_t^i(\boldsymbol{\theta}) = \mathrm{E}[\ell_t^i(\boldsymbol{\theta}) | \mathcal{G}_{t-1}], \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{33}$$

Next, we propose the following lemma to establish that the loss function exhibits strong convexity.

Lemma 9 *Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$. For any $t > 0$ and $i \in [m]$,*

$$\ell_t^i(\boldsymbol{\theta}_1) - \ell_t^i(\boldsymbol{\theta}_2) \leq \nabla \ell_t^i(\boldsymbol{\theta}_1)^\top (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) - \frac{\kappa}{2} (\boldsymbol{x}_t^\top \boldsymbol{\theta}_1 - \boldsymbol{x}_t^\top \boldsymbol{\theta}_2)^2.$$

Proof. Define $L_t^i(x) = -y_t^i x + \int_0^x \mu^i(z) dz, x \in \mathbb{R}$. By Assumption 2, we have $\ddot{L}_t^i(x) = \dot{\mu}^i(x) \geq \kappa$ for all $i \in [m]$. Hence, $L_t^i(x)$ is a κ-strongly convex function, which indicates that for any $x_1, x_2 \in \mathbb{R}$,

$$L_t^i(x_1) - L_t^i(x_2) \leq \dot{L}_t^i(x_1)(x_1 - x_2) - \frac{\kappa}{2} (x_1 - x_2)^2.$$

By setting $x_1 = \boldsymbol{x}_t^\top \boldsymbol{\theta}_1$ and $x_2 = \boldsymbol{x}_t^\top \boldsymbol{\theta}_2$, the above inequality becomes

$$L_t^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_1) - L_t^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_2) \leq \dot{L}_t^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_1) \cdot (\boldsymbol{x}_t^\top \boldsymbol{\theta}_1 - \boldsymbol{x}_t^\top \boldsymbol{\theta}_2) - \frac{\kappa}{2} (\boldsymbol{x}_t^\top \boldsymbol{\theta}_1 - \boldsymbol{x}_t^\top \boldsymbol{\theta}_2)^2.$$

Since $L_t^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}) = \ell_t^i(\boldsymbol{\theta})$ and $\dot{L}_t^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}) \cdot \boldsymbol{x}_t = \nabla \ell_t^i(\boldsymbol{\theta})$, we conclude the proof of Lemma 9. □

We now present Lemma 10, which shows that for each $i \in [m]$, $\boldsymbol{\theta}_*^i$ is the minimizer of the expected loss function $f_t^i(\boldsymbol{\theta})$.

Lemma 10 *Let $\boldsymbol{\theta} \in \mathbb{R}^d$. For any $i \in [m]$ and $t \geq 1$, $f_t^i(\boldsymbol{\theta}) - f_t^i(\boldsymbol{\theta}_*^i) \geq 0$.*

Proof. By the definition of $f_t^i(\cdot)$ and the fact that $\mathrm{E}[y_t^i|\mathcal{G}_{t-1}] = \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)$, we have

$$\begin{aligned} f_t^i(\boldsymbol{\theta}) - f_t^i(\boldsymbol{\theta}_*^i) &= \mathrm{E}[\ell_t^i(\boldsymbol{\theta}) - \ell_t^i(\boldsymbol{\theta}_*^i)|\mathcal{G}_{t-1}] \\ &= \int_0^{\boldsymbol{x}_t^\top \boldsymbol{\theta}} \mu^i(z)dz - \int_0^{\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i} \mu^i(z)dz - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \cdot (\boldsymbol{x}_t^\top \boldsymbol{\theta} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i). \end{aligned}$$

As shown in the proof of Lemma 9, $\int_0^x \mu^i(z)dz$ is κ-strongly convex. Therefore, we obtain the lower bound

$$\begin{aligned} f_t^i(\boldsymbol{\theta}) - f_t^i(\boldsymbol{\theta}_*^i) &= \int_0^{\boldsymbol{x}_t^\top \boldsymbol{\theta}} \mu^i(z)dz - \int_0^{\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i} \mu^i(z)dz - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \cdot (\boldsymbol{x}_t^\top \boldsymbol{\theta} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \\ &\geq \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \cdot (\boldsymbol{x}_t^\top \boldsymbol{\theta} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \cdot (\boldsymbol{x}_t^\top \boldsymbol{\theta} - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \\ &= 0. \end{aligned}$$

This concludes the proof of Lemma 10. □

Next, we generalize Lemma 3 from Zhang et al. (2016) to a multi-objective setting by leveraging the properties of the Online Newton Step (ONS) method.

Lemma 11 *For any $i \in [m]$ and $t > 1$,*

$$\nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i)^\top (\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i) - \frac{1}{2}\|\nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i)\|^2_{\mathbf{V}_{t+1}^{-1}} \leq \frac{1}{2}\left(\|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}} - \|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}}\right).$$

Proof of Theorem 3. Equipped with Lemma 9, Lemma 10 and Lemma 11, we can now complete the proof of Theorem 3. Lemma 9 indicates that

$$\ell_t^i(\hat{\boldsymbol{\theta}}_t^i) - \ell_t^i(\boldsymbol{\theta}_*^i) \leq \nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i)^\top (\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i) - \frac{\kappa}{2}(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)^2.$$

Taking the conditional expectation of both sides given $\mathcal{G}_{t-1}$, we obtain

$$f_t^i(\hat{\boldsymbol{\theta}}_t^i) - f_t^i(\boldsymbol{\theta}_*^i) \leq \nabla f_t^i(\hat{\boldsymbol{\theta}}_t^i)^\top (\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i) - \frac{\kappa}{2}(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)^2.$$

Combining this with Lemma 10, we derive

$$0 \leq (\nabla f_t^i(\hat{\boldsymbol{\theta}}_t^i) - \nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i))^\top (\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i) - \frac{\kappa}{2}(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)^2 + \nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i)^\top (\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i). \tag{34}$$

Applying Lemma 11 to bound the last term on the right-hand side of Eq. (34), we obtain

$$\begin{aligned} 0 \leq (\nabla f_t^i(\hat{\boldsymbol{\theta}}_t^i) - \nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i))^\top (\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i) - \frac{\kappa}{2}(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)^2 + \frac{1}{2}\|\nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i)\|^2_{\mathbf{V}_{t+1}^{-1}} \\ + \frac{1}{2}\left(\|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}} - \|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}}\right). \end{aligned} \tag{35}$$

Recalling the definition of $\ell_t^i(\cdot)$ and $f_t^i(\cdot)$ in Eq. (33), we have

$$\nabla f_t^i(\hat{\boldsymbol{\theta}}_t^i) = (\mu^i(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i))\boldsymbol{x}_t, \nabla \ell_t^i(\hat{\boldsymbol{\theta}}_t^i) = (\mu^i(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i) - y_t^i)\boldsymbol{x}_t.$$

Substituting the preceding two expressions into Eq. (35), we obtain

$$\begin{aligned} 0 \leq \frac{1}{2}\left(\|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}} - \|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}}\right) - \frac{\kappa}{2}(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)^2 \\ + (y_t^i - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)) \cdot (\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) + \frac{1}{2}\|(\mu^i(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i) - y_t^i)\boldsymbol{x}_t\|^2_{\mathbf{V}_{t+1}^{-1}}. \end{aligned}$$

Given $\mathbf{V}_{t+1} = \mathbf{V}_t + \frac{\kappa}{2}\boldsymbol{x}_t\boldsymbol{x}_t^\top$ and $\epsilon_t^i = y_t^i - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)$, the preceding inequality yields

$$\begin{aligned} 0 \leq \frac{1}{2}\left(\|\hat{\boldsymbol{\theta}}_t^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_t} - \|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|^2_{\mathbf{V}_{t+1}}\right) - \frac{\kappa}{4}(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i)^2 \\ + \epsilon_t^i \cdot (\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) + 2(U+R)^2\|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t+1}^{-1}} \end{aligned}$$

Summing this inequality over all $\tau \in [t]$ and rearranging terms, we obtain

$$\|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t+1}}^2 \leq \|\hat{\boldsymbol{\theta}}_1^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_1}^2 - \frac{\kappa}{2}\sum_{\tau=1}^{t}(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)^2 + 8(U+R)^2 \sum_{\tau=1}^{t}\|\boldsymbol{x}_\tau\|_{\mathbf{V}_{\tau+1}^{-1}}^2 + \sum_{\tau=1}^{t} 2\epsilon_\tau^i \cdot (\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i). \tag{36}$$

According to Corollary 8 of Abbasi-Yadkori et al. (2012), with probability at least $1-\delta$, for any $i \in [m]$ and $t \geq 1$,

$$\sum_{\tau=1}^{t} \epsilon_\tau^i \cdot (\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i) \leq R\sqrt{2\left(1 + \sum_{\tau=1}^{t}(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)^2\right) \ln\left(\frac{m\sqrt{1 + \sum_{\tau=1}^{t}(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)^2}}{\delta}\right)}.$$

Applying the inequality $\sqrt{ab} \leq a/c + cb$ with $c = 8R/\kappa$, we obtain

$$\sum_{\tau=1}^{t} \epsilon_\tau^i \cdot (\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i) \leq \frac{\kappa}{4} + \frac{\kappa}{4}\sum_{\tau=1}^{t}(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)^2 + \frac{8R^2}{\kappa}\ln\left(\frac{m\sqrt{1 + \sum_{\tau=1}^{t}(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_\tau^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)^2}}{\delta}\right).$$

Substituting this result into Eq. (36), we obtain

$$\|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t+1}}^2 \leq \|\hat{\boldsymbol{\theta}}_1^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_1}^2 + 8(U+R)^2 \sum_{\tau=1}^{t}\|\boldsymbol{x}_\tau\|_{\mathbf{V}_{\tau+1}^{-1}}^2 + \frac{\kappa}{2} + \frac{16R^2}{\kappa}\ln\left(\frac{m\sqrt{1 + \sum_{\tau=1}^{t}(\boldsymbol{x}_\tau^\top \hat{\boldsymbol{\theta}}_t^i - \boldsymbol{x}_\tau^\top \boldsymbol{\theta}_*^i)^2}}{\delta}\right).$$

Given that $\mathbf{V}_1 = \alpha\mathbf{I}$, $\hat{\boldsymbol{\theta}}_1^i = \mathbf{0}$, $\|\boldsymbol{x}_t\|_2 \leq 1$ for all $t \geq 1$ and $\|\boldsymbol{\theta}_*^i\|_2 \leq B$ for all $i \in [m]$, we have

$$\|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t+1}}^2 \leq \alpha B^2 + 8(U+R)^2 \sum_{\tau=1}^{t}\|\boldsymbol{x}_\tau\|_{\mathbf{V}_{\tau+1}^{-1}}^2 + \frac{\kappa}{2} + \frac{16R^2}{\kappa}\ln\left(\frac{m\sqrt{1+4B^2t}}{\delta}\right).$$

By applying Lemma 12 from Hazan et al. (2007), which states that $\sum_{\tau=1}^{t}\|\boldsymbol{x}_\tau\|_{\mathbf{V}_{\tau+1}^{-1}}^2 \leq \frac{2}{\kappa}\ln\frac{\det(\mathbf{V}_{t+1})}{\det(\mathbf{V}_1)}$, we can simplify this further to

$$\|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t+1}}^2 \leq \alpha B^2 + \frac{16(U+R)^2}{\kappa}\ln\frac{\det(\mathbf{V}_{t+1})}{\det(\mathbf{V}_1)} + \frac{\kappa}{2} + \frac{16R^2}{\kappa}\ln\left(\frac{m\sqrt{1+4B^2t}}{\delta}\right).$$

Finally, applying Lemma 10 from Abbasi-yadkori et al. (2011), which states that $\det(\mathbf{V}_{t+1}) \leq (\alpha + \kappa t/(2d))^d$, we obtain

$$\|\hat{\boldsymbol{\theta}}_{t+1}^i - \boldsymbol{\theta}_*^i\|_{\mathbf{V}_{t+1}}^2 \leq \alpha B^2 + \frac{16(U+R)^2}{\kappa} d\ln\left(1 + \frac{\kappa t}{2\alpha d}\right) + \frac{\kappa}{2} + \frac{16R^2}{\kappa}\ln\left(\frac{m\sqrt{1+4B^2t}}{\delta}\right).$$

This completes the proof of Theorem 3. □

G Proof of Theorem 4

We first introduce two lemmas related to the confidence interval and AAE-LGLB used in Online-LGLB, which are the counterparts of Lemma 4 and Lemma 5, respectively. The proofs of Lemma 12 and Lemma 13 are omitted as they closely follow the proofs of Lemma 4 and Lemma 5.

Lemma 12 *With probability at least $1-\delta$, for any $i \in [m]$,*

$$|\mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_t^i)| \leq \beta_t L \|\boldsymbol{x}\|_{\mathbf{V}_t^{-1}}, t \geq 1.$$

Lemma 13 *Consider the AAE-LGLB algorithm run in Online-LGLB. Suppose that the input arm set $\mathcal{A}_{t,s}$ contains $\boldsymbol{x}_t^*$ and satisfies $\|\boldsymbol{x}\|_{\mathbf{V}_t^{-1}} \leq W$ for all $\boldsymbol{x} \in \mathcal{A}_{t,s}$. With probability at least $1-\delta$, $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s}^m$ and for any $\boldsymbol{x} \in \mathcal{A}_{t,s}^m$,*

$$\mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \leq 4(1+\lambda+\lambda^2+\cdots+\lambda^{i-1})\beta_t L W, i \in [m].$$

Next, we propose two lemmas related to the AAE-LGLB algorithm in the **if** and **else** cases of Online-LGLB. These lemmas are counterparts to Lemma 7 and Lemma 8 in the proof of Theorem 2, respectively.

Lemma 14 *Consider the Online-LGLB algorithm. Let s_t be the stage index in the **if** case at round t. With probability at least $1-\delta$, for any $\boldsymbol{x} \in \mathcal{A}_{t,T}$,*

$$\mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \leq 4(1+\lambda+\cdots+\lambda^{i-1})\beta_t L \cdot T^{-\frac{1}{2}}, i \in [m].$$

Proof. By Lemma 6, we have $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s_t}$. Therefore, Lemma 13 is applicable to the operation $\mathcal{A}_{t,T} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s_t}, \beta_t, \{\hat{\boldsymbol{\theta}}_t^i\}_{i\in[m]}, T^{-\frac{1}{2}}\right)$ in the **if** case. Setting $W = T^{-\frac{1}{2}}$ in Lemma 13 completes the proof of Lemma 14. □

Lemma 15 *For the Online-LGLB algorithm, with probability at least $1-\delta$, for any $s \geq 1$ and $\boldsymbol{x} \in \mathcal{A}_{t,s}$,*

$$\mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}^\top \boldsymbol{\theta}_*^i) \leq 8(1+\lambda+\cdots+\lambda^{i-1})\beta_t L \cdot 2^{-s}, i \in [m].$$

Proof. For $s=1$, Lemma 15 holds because $\|\boldsymbol{\theta}_*^i\| \leq B$ and $\|\boldsymbol{x}\| \leq 1$ for any $\boldsymbol{x} \in \mathcal{A}_{t,s}$. For $s \geq 2$, by Lemma 6, we have $\boldsymbol{x}_t^* \in \mathcal{A}_{t,s-1}$. Therefore, Lemma 13 can be applied to the operation $\mathcal{A}_{t,s} = \text{AAE-LGLB}\left(\mathcal{A}_{t,s-1}, \beta_t, \{\hat{\boldsymbol{\theta}}_t^i\}_{i\in[m]}, 2^{-s+1}\right)$ in the **else** case. Setting $W = 2^{-s+1}$ in Lemma 13 completes the proof of Lemma 15. □

Let $\Psi_{T+1,s}$ be the set of trials for which $\|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}} > 2^{-s}$, i.e., $\Psi_{T+1,s} = \{t \in [T] | \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}} > 2^{-s}\}$. Clearly, the number of trials in $\Psi_{T+1,s}$ satisfies

$$|\Psi_{T+1,s}| \leq 2^s \sum_{t\in\Psi_{T+1,s}} \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}}. \tag{37}$$

Proof of Theorem 4. Let $\Psi_{T+1,0} = [T] \setminus \bigcup_{s\in[S]} \Psi_{T+1,s}$ denote the trials conducted in the **if** case. For any $i \in [m]$, we first decompose the regret as

$$R^i(T) = \sum_{t\in\Psi_{T+1,0}} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) + \sum_{s=1}^{S} \sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i). \tag{38}$$

According to Lemma 14, the regret incurred over the trials in $\Psi_{T+1,0}$ can be bounded as

$$\begin{aligned}\sum_{t\in\Psi_{T+1,0}} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) &\leq \sum_{t\in\Psi_{T+1,0}} 4\Lambda_i \beta_t L \cdot T^{-\frac{1}{2}} \\ &\leq 4\Lambda_i \beta_T L |\Psi_{T+1,0}| \cdot T^{-\frac{1}{2}} \leq 4\Lambda_i \beta_T L \sqrt{T}.\end{aligned} \tag{39}$$

Similarly, applying Lemma 15, the regret incurred over the trials in $\Psi_{T+1,s}$ can be bounded by

$$\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top} \boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^\top \boldsymbol{\theta}_*^i) \leq \sum_{t\in\Psi_{T+1,s}} 8\Lambda_i \beta_t L \cdot 2^{-s} \leq 8\Lambda_i \beta_T L \cdot 2^{-s} \cdot |\Psi_{T+1,s}|.$$

Substituting Eq. (37) into the above inequality yields

$$\sum_{t\in\Psi_{T+1,s}} \mu^i(\boldsymbol{x}_t^{*\top}\boldsymbol{\theta}_*^i) - \mu^i(\boldsymbol{x}_t^{\top}\boldsymbol{\theta}_*^i) \le 8\Lambda_i\beta_T L \sum_{t\in\Psi_{T+1,s}} \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}}. \tag{40}$$

Substituting Eq. (39) and Eq. (40) into Eq. (38), we obtain

$$\begin{aligned} R^i(T) &\le 4\Lambda_i\beta_T L\sqrt{T} + 8\Lambda_i\beta_T L \sum_{s=1}^{S} \sum_{t\in\Psi_{T+1,s}} \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}} \\ &\le 4\Lambda_i\beta_T L\sqrt{T} + 8\Lambda_i\beta_T L \sum_{t=1}^{T} \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}}. \end{aligned}$$

Applying the Cauchy–Schwarz inequality to the last term on the right-hand side, we get

$$R^i(T) \le 4\Lambda_i\beta_T L\sqrt{T} + 8\Lambda_i\beta_T L\sqrt{T}\sqrt{\sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\mathbf{V}_t^{-1}}}. \tag{41}$$

According to Lemmas 10 and 11 in Abbasi-yadkori et al. (2011), we have

$$\sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\mathbf{V}_t^{-1}} \le \frac{4}{\kappa} \ln\left(\frac{\det(\mathbf{V}_{T+1})}{\det(\mathbf{V}_1)}\right) \le \frac{4d}{\kappa} \ln\left(1 + \frac{\kappa T}{2\alpha d}\right).$$

Thus, Eq. (41) can be further bounded by

$$R^i(T) \le 4\Lambda_i\beta_T L\sqrt{T} + 16\Lambda_i\beta_T L\sqrt{T}\sqrt{\frac{d}{\kappa} \ln\left(1 + \frac{\kappa T}{2\alpha d}\right)}. \tag{42}$$

Recalling that the radius of the confidence region in Theorem 3 is

$$\beta_T = \left(\alpha B^2 + \frac{16(U+R)^2}{\kappa} d \ln\left(1 + \frac{\kappa T}{2\alpha d}\right) + \frac{\kappa}{2} + \frac{16R^2}{\kappa} \ln\left(\frac{m\sqrt{1+4B^2T}}{\delta}\right)\right)^{\frac{1}{2}},$$

substituting this into Eq. (42) yields

$$R^i(T) = \widetilde{O}\left(\Lambda_i \cdot \frac{L}{\kappa} \cdot d\sqrt{T}\right).$$

This completes the proof of Theorem 4. □