

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 BREAKING THE TOTAL VARIANCE BARRIER: SHARP SAMPLE COMPLEXITY FOR LINEAR HETEROSCEDAS- TIC BANDITS WITH FIXED ACTION SET

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent years have witnessed increasing interests in tackling heteroscedastic noise in bandits and reinforcement learning (e.g. Zhou et al., 2021; Zhao et al., 2023a; Jia et al., 2024; Pacchiano, 2025). In these works, the cumulative variance of the noise $\Lambda = \sum_{t=1}^T \sigma_t^2$, where σ_t^2 is the variance of the noise at round t , is used to characterize the statistical complexity of the problem, yielding *simple regret* bounds of order $\tilde{\mathcal{O}}(d\sqrt{\Lambda/T^2})$ for d -dimensional linear bandits with heteroscedastic noise (Zhou et al., 2021; Zhao et al., 2023a). However, with a closer look, Λ remains the same order even if the noise is close to zero at half of the rounds, which indicates that the Λ -dependence is not optimal.

In this paper, we revisit the linear bandit problem with heteroscedastic noise. We consider the setting where the action set is fixed throughout the learning process. We propose a novel variance-adaptive algorithm VAE (Variance-Aware Exploration with Elimination) for large action set, which actively explores actions that maximizes the information gain among a candidate set of actions that are not eliminated. With the active-exploration strategy, we show that VAE achieves a *simple regret* with a nearly *harmonic-mean* dependent rate, i.e., $\tilde{\mathcal{O}}\left(d\left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{\mathcal{O}}(d)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}\right)$ ¹ where $\sigma^{(i)}$ is the i -th smallest variance among $\{\sigma_t\}_{t=1}^T$. For finitely many actions, we propose a variance-aware variant of G-optimal design based exploration, which achieves a simple regret of $\tilde{\mathcal{O}}\left(\sqrt{d \log |\mathcal{A}|} \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{\mathcal{O}}(d)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}\right)$. We also establish a nearly matching lower bound for the fixed action set setting indicating that *harmonic-mean* dependent rate is unavoidable. To the best of our knowledge, this is the first work that breaks the $\sqrt{\Lambda}$ barrier for linear bandits with heteroscedastic noise.

1 INTRODUCTION

The stochastic multi-armed bandit (MAB) problem is a fundamental framework for studying the exploration-exploitation trade-off in sequential decision-making (Auer et al., 2002). In the classic stochastic bandit setting, an agent repeatedly selects an arm from a set of arms and receives a stochastic reward associated with the chosen arm. The goal of the agent is to maximize the cumulative reward over a series of rounds by balancing exploration (trying out different arms to gather information) and exploitation (choosing the best-known arm based on past observations). Over the past few decades, various algorithms have been proposed to tackle the stochastic bandit problem from the perspectives of minimax optimal sample complexity (Audibert & Bubeck, 2009; Ménard & Garivier, 2017; Jin et al., 2021; 2023).

To further leverage the heteroscedastic nature of the noise in real world applications, recent works have extended the classic bandit framework to account for heteroscedastic noise, where the variance

¹The formal notation is given by $\tilde{\mathcal{O}}\left(d\left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\iota(d, T)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}\right)$, where $\iota(d, T) = \tilde{\mathcal{O}}(d)$ is a function of d and T . For simplicity, we use $\tilde{\mathcal{O}}(d)$ to denote $\iota(d, T)$ throughout the paper.

of the noise can vary across different arms and time steps (Zhou et al., 2021; Zhao et al., 2023a; Jia et al., 2024; Pacchiano, 2025). These works have shown that by taking into account the varying variance of the noise, it is possible to design more efficient algorithms that achieve better performance in terms of regret bounds. In detail, Zhou et al. (2021) first considered the linear bandit problem with heteroscedastic noise and proposed a variance-aware algorithm that achieves a regret bound of order $\tilde{\mathcal{O}}(d\sqrt{\Lambda} + \sqrt{dT})$ where $\Lambda = \sum_{t=1}^T \sigma_t^2$ is the cumulative variance of the noise along T time steps and d is the dimension of the feature space. Later, Zhou & Gu (2022) improved the cumulative regret bound to $\tilde{\mathcal{O}}(d\sqrt{\Lambda} + d)$, yielding a simple regret² bound of order $\tilde{\mathcal{O}}(d\sqrt{\Lambda/T^2} + d/T)$. More recently, Jia et al. (2024) proposed VarCB, which achieves tighter $\tilde{\mathcal{O}}(\sqrt{|\mathcal{A}|\Lambda d} + d^2)$ cumulative regret for bandits with few actions, where $|\mathcal{A}|$ is the size of the action set and further extend the results to general function classes. Notably, they also proved a lower bound of order $\Omega(\sqrt{\min(|\mathcal{A}|, d)\Lambda} + d)$ when $d \leq \sqrt{|\mathcal{A}|T}$, indicating that the dependence on $\sqrt{\Lambda}$ is unavoidable in the worst case.

Despite this progress, existing regret bounds depend on the total variance term Λ , which overlooks the heterogeneity of information gain across actions and time steps with different noise levels. Consider an extreme case: if $\sigma_t^2 \approx 0$ for all $t \leq t_0$ with $t_0 = \tilde{\mathcal{O}}(d) \ll T$, the d -dimensional weighting parameter in the linear bandit problem could be recovered almost exactly. In such a case, the regret bound should be essentially independent of the noise variance after time step t_0 . This motivates an important open question in heteroscedastic linear bandits:

Can we improve upon the $\sqrt{\Lambda}$ dependence in the regret bounds in heteroscedastic linear bandits?

In this paper, we revisit the problem of best-arm identification in linear bandits under heteroscedastic noise, where the variances of the reward distribution may vary significantly across actions. The primary performance metric we focus on is the simple regret, which measures the suboptimality of the action recommended after a fixed budget of exploration. Our results highlight the fundamental role of the harmonic mean of the variances in characterizing the attainable regret rate.

Our main contributions are summarized as follows:

- **Variance-adaptive exploration for large action sets.** We propose a novel algorithm, VAEEL (Variance-Aware Exploration with Elimination), designed to handle large (potentially infinite) action sets. The key idea is to maintain a candidate set of promising actions and actively explore those that maximize the information gain subject to elimination rules. We prove that VAEEL achieves a simple regret bound of

$$\tilde{\mathcal{O}}\left(d\left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{\mathcal{O}}(d)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}\right),$$

where d is the feature dimension, and $\{[\sigma^{(i)}]^2\}$ are the ordered list of the variance sequence $\{\sigma_t^2\}$. This establishes a nearly harmonic-mean dependent rate for the simple regret.

- **Variance-aware G-optimal design for finite action sets.** For the case of a finite action set \mathcal{A} , we propose a variance-adaptive variant of G-optimal design based exploration. We show that this strategy achieves a simple regret bound with improved dependence on the dimension d as follows

$$\tilde{\mathcal{O}}\left(\sqrt{d \log |\mathcal{A}|} \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{\mathcal{O}}(d)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}\right).$$

- **Lower bound matching the harmonic-mean rate.** We establish a nearly matching lower bound for the fixed-action setting, showing that the harmonic-mean dependence is intrinsic to the problem. This demonstrates that our algorithms are essentially optimal in their variance dependence.
- **Breaking the $\sqrt{\Lambda}$ barrier.** To the best of our knowledge, this is the first work that surpasses the classical $\sqrt{\Lambda}$ -type dependence in simple regret bounds for linear bandits with heteroscedastic noise, where Λ denotes the variance proxy commonly used in prior analyses. A comprehensive comparison on the simple regret bounds is provided in Table 1 for the reader's reference.

²Simple regret quantifies the expected gap between the optimal reward and the reward of the arm proposed by the algorithm.

108

109 Table 1: Comparison between different algorithms for stochastic (linear) contextual bandits. Here
110 d is the feature dimension, T is the number of rounds, $\{\sigma_t\}_{t \in [T]}$ is the variance of noise at round
111 $t \in [T]$, $|\mathcal{A}|$ is the size of the arm set, and $\Lambda = \sum_{t=1}^T \sigma_t^2$ is the cumulative variance of the noise.
112 The lower bounds from Jia et al. (2024); He & Gu (2025) are derived for the cumulative regret.
113 We convert them to be comparable to our simple regret by dividing them by T . Note that their
114 lower bounds are derived for the worst case sequence of noise variance, while our lower bound has
115 a refined dependence on the noise variance sequence.

Algorithm	Simple Regret Upper Bound	Simple Regret Lower Bound
Weighted OFUL (Zhou et al., 2021)	$d\sqrt{\Lambda/T^2}$	-
Weighted OFUL+ (Zhou & Gu, 2022)	$d\sqrt{\Lambda/T^2}$	-
VOFUL (Zhang et al., 2021)	$d^{9/2}\sqrt{\Lambda/T^2}$	-
VOFUL2 (Kim et al., 2021)	$d^{3/2}\sqrt{\Lambda/T^2}$	-
SAVE (Zhao et al., 2023a)	$d\sqrt{\Lambda/T^2}$	-
LinNATS (Xu et al., 2023)	$d^{3/2}\sqrt{\Lambda/T^2}$	-
VarCB (Jia et al., 2024)	$\sqrt{ \mathcal{A} \Lambda d/T^2}$	$\Omega(\sqrt{\min(\mathcal{A} , d)\Lambda/T^2})$ $\tilde{\Omega}(d\sqrt{\Lambda/T^2})$
He & Gu (2025)	-	-
VAEE (Ours)	$d\left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{O}(d)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}$	$\Omega\left(d\left(\sum_{i=1}^t \frac{1}{\sigma_i^2}\right)^{-\frac{1}{2}}\right)$
VAGD (Ours)	$\sqrt{d \log \mathcal{A} \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{O}(d)} \frac{1}{[\sigma^{(i)}]^2}\right]^{-\frac{1}{2}}}$	-

128

129

130 **Notations.** We use bold lowercase letters (e.g., \mathbf{a}) to denote vectors and bold uppercase letters (e.g.,
131 \mathbf{A}) to denote matrices. For a vector $\mathbf{a} \in \mathbb{R}^d$, we use $\|\mathbf{a}\|_2$ to denote its Euclidean norm. For a
132 positive definite matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we define the elliptical norm of a vector \mathbf{a} as $\|\mathbf{a}\|_{\mathbf{A}} = \sqrt{\mathbf{a}^\top \mathbf{A} \mathbf{a}}$.
133 We use \mathbf{I}_d to denote the $d \times d$ identity matrix. For a set \mathcal{A} , we use $|\mathcal{A}|$ to denote its cardinality. We
134 use $\tilde{O}(\cdot)$ to hide logarithmic factors in $d, T, 1/\delta, 1/\sigma_{\min}, 1/\sigma_{\max}$. For a sequence $\{a_t\}_{t=1}^T$, we use
135 $a^{(i)}$ to denote the i -th smallest element in the sequence.

136

137

2 RELATED WORK

138

139

140 **Variance-Aware Regret for Linear Bandits with Heteroscedastic Noise.** The incorporation of
141 variance information in linear bandit algorithms has garnered significant attention in recent years,
142 leading to substantial improvements in regret bounds. Early work by Kirschner & Krause (2018)
143 introduced the concept of information-directed sampling for bandits with heteroscedastic noise,
144 demonstrating that leveraging variance information can lead to more efficient exploration strategies.
145 Later, Zhou et al. (2021) proposed a variance-aware algorithm for linear bandits that achieves
146 a regret bound of order $\tilde{O}(d\sqrt{\Lambda} + \sqrt{dT})$, where $\Lambda = \sum_{t=1}^T \sigma_t^2$ is the cumulative variance of the
147 noise. This result was further improved by Zhou & Gu (2022) to a tighter bound of $\tilde{O}(d\sqrt{\Lambda} + d)$.
148 Zhao et al. (2023b) later proposed a peeling-based algorithm that achieves a similar regret bound.

149

150

151 The challenge of unknown conditional variances has been addressed by several researchers. Zhang
152 et al. (2021) and Kim et al. (2021) developed algorithms that operates without prior knowledge of the
153 variance, achieving regret bounds that adapt to the observed noise levels. However, these approaches
154 are not tractable for large action sets and incur sub-optimal dependence on d . Zhao et al. (2023a)
155 proposed a computationally efficient algorithm that achieves a regret bound of order $\tilde{O}(d\sqrt{\Lambda} + d)$
156 without requiring prior knowledge of the variances.

157

158

159

160

161

162 More recently, Pacchiano (2025) extended the variance-aware framework of Zhao et al.
163 (2023a) to the general function approximation setting, achieving a regret bound of order
164 $\tilde{O}(d_{\text{eluder}}\sqrt{\log(\mathcal{F})\Lambda} + d_{\text{eluder}}\log(\mathcal{F}))$, where d_{eluder} is the eluder dimension and $\log(\mathcal{F})$ is the
165 log-covering number of the function class \mathcal{F} . Concurrently, Jia et al. (2024) introduced VarCB, an
166 algorithm that attains a regret bound of $\tilde{O}(\sqrt{|\mathcal{A}|\Lambda d} + d^2)$ for bandits with few actions, and extended
167 their results to general function classes. They also established a worst-case lower bound of order
168 $\Omega(\sqrt{\min(|\mathcal{A}|, d)\Lambda} + d)$ when $d \leq \sqrt{|\mathcal{A}|T}$. He & Gu (2025) further studied the setting where the
169 action set can change arbitrarily over time and proved an instance-dependent lower bound of order

162 $\Omega(d\sqrt{\Lambda}/\log T)$ for the expected cumulative regret. These results indicate that the $\sqrt{\Lambda}$ dependence
 163 is unavoidable in such settings.
 164

165 **Bandits with Heavy-Tailed Noise.** The topic of robustness to heavy-tailed rewards has received
 166 considerable attention in recent years, addressing the limitations of classical bandit algorithms that
 167 assume sub-Gaussian or bounded noise. Bubeck et al. (2013) pioneered this research direction by
 168 studying heavy-tailed rewards in multi-armed bandits, establishing that standard concentration in-
 169 equalities fail in such environments. For linear bandits, Medina & Yang (2016) proposed truncation-
 170 based methods and median-of-means estimators to handle heavy-tailed noise, achieving sublinear
 171 regret bounds. Shao et al. (2018) adopted median-of-means techniques with a well-designed allo-
 172 cation of decisions to achieve nearly optimal regret bounds. Later, Xue et al. (2020) introduced a
 173 SupLin-based algorithm (Chu et al., 2011) which further improved the dimension dependence in the
 174 regret bounds. More related works include Li & Sun (2024), Huang et al. (2023), which proposed
 175 Huber regression based algorithms to handle heteroscedastic heavy-tailed noise. Recently, Ye et al.
 176 (2025) proposed a Catoni's estimator based algorithm that achieves adaptive regret bounds in bandits
 177 with general function approximation.
 178

179 **Variance-Dependent Bounds in MDPs.** As a natural extension of bandits, Markov Decision Pro-
 180 cesses (MDPs) have also been studied under the lens of variance-dependent regret bounds. In tabular
 181 MDPs, Zanette & Brunskill (2019) first established a variance-dependent regret bound which scales
 182 with the square root of the maximum variance of the value function. Afterwards, Zhou et al. (2023)
 183 proposed MVP-V, an algorithm that achieves a regret bound scaling with the square root of the total
 184 variance of the value function, achieving worst-case optimal regret bound. In MDPs with linear
 185 function approximation, Zhao et al. (2023a) proposed a variance-aware algorithm that achieves a
 186 second-order and horizon-free regret bound. More recently, there have been several works (Wang
 187 et al., 2024; Zhao et al., 2024; Wang et al., 2025; Zhao et al., 2025) presenting variance-dependent
 188 regret bounds in MDPs with general function approximation.
 189

3 PRELIMINARIES

191 We consider a heteroscedastic variant of the classic linear bandit problem. Let T be the total number
 192 of rounds. The action set \mathcal{A} is fixed. At each round $t \in [T]$, the interaction between the agent and
 193 the environment is as follows:

1. The agent selects $\mathbf{a}_t \in \mathcal{A}$ based on the past observations $\mathcal{F}_{t-1} = (\mathbf{a}_1, r_1, \dots, \mathbf{a}_{t-1}, r_{t-1})$ up to
 time $t-1$.
2. The environment generates the stochastic noise η_t at round t and reveals the stochastic reward
 $r_t = \langle \theta^*, \mathbf{a}_t \rangle + \eta_t$ to the agent.

194 WLOG, we assume that for all $\mathbf{a} \in \mathcal{A}$, it holds that $\|\mathbf{a}\|_2 \leq 1$ and $\|\theta^*\|_2 \leq 1$.
 195

200 **Remark 3.1.** Our assumption that the action set is fixed is necessary for achieving the harmonic-
 201 mean dependent rate. In scenarios where the action set can change arbitrarily over time, it is possible
 202 to construct instances where the cumulative variance $\Lambda = \sum_{t=1}^T \sigma_t^2$ remains the most appropriate
 203 measure of statistical complexity. This is because an adversarially chosen action set can force the
 204 algorithm to repeatedly explore less informative actions when the noise level is low, thereby negating
 205 the benefits of a harmonic-mean based approach. A detailed study of this phenomenon is provided
 206 by He & Gu (2025), which demonstrates that in the case of adversarially changing contexts, there
 207 exists a lower bound of order $\Omega(d\sqrt{\Lambda}/\log T)$ for the expected cumulative regret, indicating that the
 208 $\sqrt{\Lambda}$ dependence is unavoidable in such settings.
 209

210 Therefore, to fully leverage the advantages of our proposed variance-adaptive algorithms and
 211 achieve the improved regret bounds, we focus on the standard stochastic linear bandit setting (Latti-
 212 more & Szepesvári, 2020) where the action set is fixed throughout the learning process.
 213

214 We introduce the following assumption on the noise η_t .
 215

Assumption 3.2. The noise η_t is conditionally σ_t -sub-Gaussian, i.e., for all $\lambda \in \mathbb{R}$, it holds that
 $\mathbb{E}[\exp(\lambda\eta_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda^2\sigma_t^2/2)$, where \mathcal{F}_{t-1} is the filtration up to round $t-1$. We assume
 that there exist known constants $\sigma_{\min}, \sigma_{\max} > 0$ such that $\sigma_{\min} \leq \sigma_t \leq \sigma_{\max}$ for all $t \in [T]$.
 216

216 **Remark 3.3.** This assumption follows from the original formulation of heteroscedastic bandits by
 217 Kirschner & Krause (2018). Later works (Zhou et al., 2021; Zhao et al., 2023a; Jia et al., 2024) have
 218 slightly generalized this assumption to only require the variance of η_t to be bounded by σ_t^2 and the
 219 magnitude of η_t to be bounded by a constant. However, this generalization does not significantly
 220 affect our analysis or results, as we will discuss in Appendix E that our algorithms can be extended
 221 to handle heavy-tailed noise by replacing the least-squares estimator with a robust estimator.

222 In this paper, we focus on the best-arm identification problem in linear bandits with heteroscedastic
 223 noise. The performance of an algorithm is measured by the simple regret defined as follows:
 224

$$225 \quad \text{SR}(T) = \mathbb{E} \left[\max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\theta}^*, \mathbf{a} \rangle - \langle \boldsymbol{\theta}^*, \hat{\mathbf{a}}_T \rangle \right], \quad (3.1)$$

226 where $\hat{\mathbf{a}}_T$ is the action recommended by the algorithm after T rounds of exploration.
 227

228 **Remark 3.4.** In the stochastic linear bandit literature, simple regret is closely connected to cumu-
 229 lative regret. In particular, if an algorithm achieves cumulative regret of order $\tilde{\mathcal{O}}(\sqrt{dT})$, then its
 230 simple regret can be shown to be of order $\tilde{\mathcal{O}}(\sqrt{d/T})$ (Lattimore & Szepesvári, 2020). In the het-
 231 eroscedastic setting, however, the varying and unpredictable noise levels make this relationship more
 232 subtle. For example, a harmonic-mean dependence of the simple regret on the variances does not
 233 necessarily translate to the same dependence for cumulative regret. In this work, we therefore focus
 234 on directly analyzing the simple regret of our proposed algorithms.
 235

236 4 VARIANCE-AWARE EXPLORATION WITH ELIMINATION

237 In this section, we propose Variance-Aware Exploration with Elimination (VAEE), a variance-
 238 adaptive approach designed for linear bandits operating in environments with heteroscedastic noise
 239 and potentially large action spaces. The algorithm is displayed in Algorithm 1, which builds upon
 240 the Optimism in the Face of Uncertainty for Linear bandits (OFUL) framework while incorporating
 241 variance information to improve exploration efficiency and regret bounds.
 242

243 **Variance Adaptation.** The algorithm explicitly incorporates variance information σ_t observed at
 244 each time step and uses variance-weighted updates for both the covariance matrix and parameter
 245 estimation (Zhou et al., 2021). This allows the algorithm to adaptively adjust its confidence sets
 246 based on the observed noise levels, leading to more accurate estimates of the underlying parameters.
 247

248 **Active Exploration.** Algorithm 1 employs an active exploration strategy that selects actions based
 249 on their potential to maximize information gain. Specifically, at each round, the algorithm chooses
 250 the action that maximizes the uncertainty in the parameter estimate, as measured by the Mahalanobis
 251 distance with respect to the inverse covariance matrix. This encourages exploration of actions that
 252 are expected to provide the most informative feedback.
 253

Algorithm 1 Variance-Aware Exploration with Elimination (VAEE)

254 **Require:** $\mathcal{A} \subset \mathbb{R}^d, \delta$.

255 1: Initialize $V_0 \leftarrow \lambda I_d$, $\hat{\boldsymbol{\theta}}_0 \leftarrow 0$, $\mathcal{A}_1 \leftarrow \mathcal{A}$.
 256 2: **for** $t = 1, \dots, T$ **do**
 257 3: Pull the action $\mathbf{a}_t \leftarrow \max_{\mathbf{e} \in \mathcal{A}_t} \|\mathbf{e}\|_{V_{t-1}^{-1}}$.
 258 4: The agent receives the reward r_t and the variance σ_t .
 259 5: Calculate $V_t \leftarrow V_{t-1} + \sigma_t^{-2} \mathbf{a}_t \mathbf{a}_t^\top$.
 260 6: Calculate $\hat{\boldsymbol{\theta}}_t \leftarrow V_t^{-1} \sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s r_s$.
 261 7: Set confidence set as follows $\mathcal{C}_t \leftarrow \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\|_{V_t^{-1}}^2 \leq \beta_t\}$.
 262 8: Eliminate low rewarding arms: $\mathcal{A}_{t+1} \leftarrow \{\mathbf{a} \in \mathcal{A}_t : \max_{\mathbf{e} \in \mathcal{A}_t} \min_{\boldsymbol{\theta} \in \mathcal{C}_t} \langle \boldsymbol{\theta}, \mathbf{e} \rangle \leq \max_{\boldsymbol{\theta} \in \mathcal{C}_t} \langle \boldsymbol{\theta}, \mathbf{a} \rangle\}$.
 263 9: **end for**
 264

265 4.1 CASE STUDY ON WHY WEIGHTED OFUL FAILS

266 Now we present a case study to demonstrate that Weighted OFUL fails for structural reason rather
 267 than the analysis, especially when the variance sequence includes low-variance windows and in-

270
271 Table 2: Simple and Cumulative Regrets of Algorithm 1 and Weighted-OFUL (Zhou & Gu, 2022).
272 We use $R(T) \leq \sum_{t=1}^T \text{SR}(t)$ to upper bound the regret of Algorithm 1.

Scenario	σ_t^2	Simple Regret		Cumulative Regret	
		Algorithm 1	Weighted OFUL	Algorithm 1	Weighted OFUL
Fast-Decaying Noise	$\sigma_t^2 = 1/t^2$	$\tilde{O}(\frac{d}{T^{3/2}})$	$\tilde{O}(\frac{d}{T})$	$\tilde{O}(d)$	$\tilde{O}(d)$
Flat Noise ($1/d$)	$\sigma_t^2 \equiv 1/d$	$\tilde{O}(\sqrt{\frac{d}{T}})$	$\tilde{O}(\sqrt{\frac{d}{T}})$	$\tilde{O}(\sqrt{dT})$	$\tilde{O}(\sqrt{dT})$
Many Moderate Spike	$\alpha \in (0, 1)$, $\sigma_t^2 = \begin{cases} x, & t \leq \alpha T, \\ 1, & t > \alpha T, \end{cases}$ with $x = T^{-1/3}$	$\tilde{O}(\frac{d}{T^{2/3}})$	$\tilde{O}(\frac{d}{\sqrt{T}})$	$\tilde{O}(dT^{1/3})$	$\tilde{O}(d\sqrt{T})$
Front-Loaded Super-Precision	$\sigma_t^2 = \begin{cases} \min\{1/2, t^{-2}\}, & t \leq T^{4/5}, \\ 1/2, & t > T^{4/5}, \end{cases}$	$\tilde{O}(\frac{d}{T^{6/5}})$	$\tilde{O}(\frac{d}{\sqrt{T}})$	$\tilde{O}(d)$	$\tilde{O}(d\sqrt{T})$

283 formation for learning the d -dimensional vector grows anisotropically across coordinates. Our
284 variance-sequence-aware design reallocates exploration toward weak coordinates and achieves an
285 instance-variance-sequence dependent bound. In the example below, our algorithm attains simple
286 regret in order $\varepsilon e^{-\Theta(\log T/\varepsilon^2)}$, while Weighted OFUL yields $\varepsilon e^{-\Theta(\log T)}$. When $\varepsilon = T^{-1/4}$, the
287 simple regret of VAE is sharper than that of Weighted OFUL by a factor of $T^{\Theta(\sqrt{T})}$.

288 **Setup.** Consider $e_1 = (1, 0)$, $e_2 = (0, 1)$, and $x = (1 - \varepsilon, \varepsilon)$ with $\theta^* = e_1$, confidence radii
289 $\beta_t \leq \beta = \Theta(\log T)$, and $\varepsilon = T^{-1/4}$. There is a global window W of length L where the noise
290 variance is $\sigma_t^2 = T^{-\alpha}$ for $t \in W$ with $\alpha \in (0, 1)$; outside W the variance is constant.

291 For simplicity, assume the first coordinate is already well explored: pulls of e_1 and x both supply
292 substantial information about the first coordinate, and both our method and weighted OFUL sample
293 x (or e_1) frequently. Thus, the first-coordinate information dominates that of the second, and we
294 focus on the error caused by limited information in the second coordinate.

295 **Weighted OFUL in the low-variance window.** *Case 1 (start with e_2).* Each pull of e_2 increases
296 the second-coordinate information by T^α . Once about $\log T/\varepsilon^2$ information has been accumulated
297 via e_2 , the policy switches to x . Intuitively, with high probability $\text{UCB}_x(t) \approx (1 - \varepsilon) + \varepsilon = 1$ and
298 $\text{UCB}_2(t) \approx \varepsilon$, so $\text{UCB}_2(t) \leq \text{UCB}_x(t)$ and x is selected.

299 *Case 2 (keep pulling x).* Each pull of x contributes only $\varepsilon^2 T^\alpha$ to the e_2 direction, so after L pulls
300 the total second-coordinate information is at most $L \varepsilon^2 T^\alpha$.

301 **Simple Regret of Weighted OFUL.** Choose $L = c_L T^{-\alpha} \frac{\log T}{\varepsilon^4} \Rightarrow L \varepsilon^2 T^\alpha = c_L \log T$. By a
302 standard Chernoff/Hoeffding concentration, the failure probability of recommending $\hat{a}_T = x$ decays
303 as $\exp(-\Theta(\log T))$, hence the simple regret satisfies $\text{SR}(T) \leq \varepsilon \cdot \exp(-\Theta(\log T))$.

304 **Simple Regret of Algorithm 1.** Since our algorithm pulls the arm with the largest exploration
305 bonus (Line 2 of Algorithm 1), it allocates the window W to e_2 and gains $L T^\alpha \asymp c_L \frac{\log T}{\varepsilon^4}$ units
306 of second-coordinate information within W . By the standard concentration inequality, the failure
307 probability of recommending $\hat{a}_T = x$ decays as $\exp(-\Theta(\log T/\varepsilon^2))$, hence the simple regret
308 satisfies $\text{SR}(T) \leq \varepsilon \cdot \exp(-\Theta(\log T/\varepsilon^2))$. Since $\varepsilon = T^{-1/4}$, our simple regret is significantly
309 lower than that of Weighted OFUL by a factor of $T^{\Theta(\sqrt{T})}$.

310 4.2 THEORETICAL RESULTS FOR VAE

311 We now present the main theoretical results for VAE. The following theorem establishes a simple
312 regret bound with harmonic-mean dependence.

313 **Theorem 4.1** (Simple Regret of VAE). Set $\beta_t = 2\sqrt{\lambda} + 16\sqrt{\log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}}$ and
314 $\lambda = 1$ in Algorithm 1. Let $\sigma_T^{(i)}$ be the i -th smallest element in $\{\sigma_\tau^2\}_{\tau=1}^T$. With probability at least
315 $1 - \delta$, the simple regret of Algorithm 1 satisfies

$$316 \text{SR}(T) = \tilde{O}(\sqrt{d}) \min_{1 \leq k \leq T+1} \left\{ x = \sqrt{\frac{\iota(T) - k + 1}{\sum_{i=k}^T \frac{1}{[\sigma_T^{(i)}]^2}}} \mid x \in [\sigma_T^{(k-1)}, \sigma_T^{(k)}] \right\},$$

324 where $\iota(T) = 2d \log(1 + \sum_{\tau \in [T]} \sigma_{\tau}^{-2}/d)$.
 325

326 **Remark 4.2.** Theorem 4.1 provides a simple regret bound for VAE that depends on the harmonic
 327 mean of the variances σ_t^2 . To see this, we can simplify the bound by substituting $k = \tilde{O}(d)$ and
 328 $\iota(T) = \tilde{O}(d)$, which yields the following simplified expression:

$$329 \quad \text{SR}(T) = \tilde{O}\left(d \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{\tilde{O}(d)} \frac{1}{[\sigma^{(i)}]^2} \right]^{-\frac{1}{2}}\right). \quad (4.1)$$

333 This bound in (4.1) highlights that the simple regret decreases as the harmonic mean of the variances
 334 increases, effectively capturing the influence of low-variance actions on the overall performance.
 335 Notably, this result breaks the traditional $\sqrt{\Lambda}$ barrier, where $\Lambda = \sum_{t=1}^T \sigma_t^2$, demonstrating that
 336 our variance-adaptive approach can achieve significantly better performance in environments with
 337 heteroscedastic noise.

338 **Remark 4.3.** It is worth noting that our harmonic-mean dependence is subtracted by the contribu-
 339 tion of the $\tilde{O}(d)$ smallest variances. This subtraction is unavoidable due to the inherent difficulty of
 340 estimating a d -dimensional parameter, which requires at least d well-explored actions. In the worst-
 341 case scenario, these d actions may correspond to the smallest variances, and then it is impossible to
 342 achieve a near zero simple regret with only $T = O(1)$ rounds of noise-free exploration. Therefore,
 343 the subtraction term in our bound is necessary to account for this fundamental limitation.

344 Nonetheless, we can still show that our simple regret bound is strictly sharper than the $\sqrt{\Lambda}$ -type
 345 bounds in prior works (Zhou et al., 2021; Zhao et al., 2023a; Jia et al., 2024). To see this, we
 346 observe that $\min_{1 \leq k \leq T+1} \left\{ x = \sqrt{\frac{\iota(T)-k+1}{\sum_{i=k}^T \frac{1}{[\sigma_T^{(i)}]^2}}} \mid x \in [\sigma_T^{(k-1)}, \sigma_T^{(k)}] \right\}$ is the solution to the following
 347 equation: $x^2 = \frac{\iota(t)}{\sum_{i=1}^t \frac{1}{\max(\sigma_i^2, x^2)}}$. We further have

$$351 \quad x^2 \leq \frac{\iota(t)}{\sum_{i=1}^t \frac{1}{\sigma_i^2 + x^2}} \leq \frac{\iota(t)}{\frac{t}{x^2 + t^{-1} \sum_{i=1}^t \sigma_i^2}} = \frac{\iota(t)(x^2 + t^{-1} \sum_{i=1}^t \sigma_i^2)}{t}, \quad (4.2)$$

354 where the second inequality follows from mean inequality. Rearranging the terms yields $x^2 =$
 355 $\tilde{O}(d\Lambda/t^2)$ when $t = \Omega(d)$. Please refer to Appendix B for detailed derivations.

356 4.3 COMPARISON WITH WEIGHTED OFUL

358 In this subsection, we present a case study to illustrate the limitations of using the cumulative vari-
 359 ance $\Lambda = \sum_{t=1}^T \sigma_t^2$ as a measure of statistical complexity in linear bandits with heteroscedastic
 360 noise and the potential weakness of existing algorithms that rely on this measure. For simplicity, we
 361 assume $\sum_{t=1}^T 1/\sigma_t^2 \gg \sum_{i=1}^{\tilde{O}(d)} 1/[\sigma^{(i)}]^2$. Therefore, according to (4.1) and Zhou & Gu (2022),

$$363 \quad \text{SR}_{\text{Alg 1}} \asymp d \left(\sum_{t=1}^T \frac{1}{\sigma_t^2} \right)^{-1/2}, \quad \text{SR}_{\text{Weighted-OFUL}} \asymp d \frac{\sqrt{\sum_{t=1}^T \sigma_t^2}}{T} \quad (4.3)$$

366 First, by the HM-AM inequality, we have $T/(\sum_t \frac{1}{\sigma_t^2}) \leq \sum_t \sigma_t^2/T$, which leads to a general rela-
 367 tionship between the regret bounds of our methods: $\text{SR}_{\text{Alg 1}} \leq \text{SR}_{\text{Weighted-OFUL}}$ for any sequence
 368 σ_t^2 . Therefore, our regret bound is always sharper. In the Table 2, we demonstrate the specific rate
 369 of improvement for some special variance sequences. We note that an improvement in simple regret
 370 does not necessarily lead to an improvement in cumulative regret. For example, this is evident in
 371 fast-decaying noise, as discussed in Remark 3.4.

373 5 STOCHASTIC LINEAR BANDIT WITH FINITE ACTION SPACE

376 In this section, we consider the special case where the action set \mathcal{A} is finite. We propose a variance-
 377 adaptive G-optimal design based exploration strategy and establish a simple regret bound with
 harmonic-mean dependence which improved over Theorem 4.1 by a factor of \sqrt{d} .

378 **Algorithm 2** Variance Adaptive G-Optimal Design (VAGD)

379 **Require:** $\mathcal{A} \subset \mathbb{R}^d, \delta$.

380 1: Find nearly G -optimal design $\pi \in \Delta(\mathcal{A})$ with $|\text{supp}(\pi)| \leq 4d \log \log d + 16$ as described in

381 Theorem 5.2 that minimizes

382

383
$$\max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\|_{V(\pi)^{-1}} \text{ subject to } \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a}) = 1.$$

384

385 2: Let $\mathcal{T}_0(\mathbf{a}) \leftarrow \emptyset$ for all $\mathbf{a} \in \mathcal{A}$.

386 3: **for** $t = 1, \dots, T$ **do**

387 4: Pull the action $\mathbf{a}_t := \arg\min_{\mathbf{a} \in \mathcal{A}} \sum_{\tau \in \mathcal{T}(\mathbf{a})} \frac{1}{\sigma_\tau^2 \cdot \pi(\mathbf{a})}$

388 5: Observe reward r_t and variance σ_t .

389 6: Update the set $\mathcal{T}_t(\mathbf{a}_t) \leftarrow \mathcal{T}_{t-1}(\mathbf{a}_t) \cup \{t\}$ and $\mathcal{T}_t(\mathbf{a}) \leftarrow \mathcal{T}_{t-1}(\mathbf{a})$ for all $\mathbf{a} \neq \mathbf{a}_t$.

390 7: **end for**

391 8: Outout $\mathbf{a}_{T+1} = \arg\max_{\mathbf{a} \in \mathcal{A}} \langle \hat{\theta}_T, \mathbf{a} \rangle$ where $\hat{\theta}_T = V_T^{-1} \sum_{t=1}^T \sigma_t^{-2} r_t \mathbf{a}_t$ and $V_T = I + \sum_{t=1}^T \sigma_t^{-2} \mathbf{a}_t \mathbf{a}_t^\top$.

394

395 5.1 VARIANCE-ADAPTIVE G-OPTIMAL DESIGN BASED EXPLORATION

396

397

398 **G -optimal design.** In Algorithm 2, we need to find a nearly G -optimal design $\pi \in \Delta(\mathcal{A})$ that maximizes $\log \det V(\pi)$. We first introduce some necessary notations and definitions regarding D -optimal and G -optimal designs. Let $\pi : \mathcal{A} \rightarrow [0, 1]$ be a distribution on \mathcal{A} so that $\sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a}) = 1$. Based on $\pi \in \mathcal{P}(\mathcal{A}_\ell)$, define $V(\pi) \in \mathbb{R}^{d \times d}$ and $g(\pi) \in \mathbb{R}$ as follows $V(\pi) = \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a}) \mathbf{a} \mathbf{a}^\top$, $g(\pi) = \max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\|_{V(\pi)^{-1}}^2$. A design π is defined as a G -optimal design if it minimises g . And a design π is defined as a D -optimal design if it maximises $f(\pi) = \log \det V(\pi)$. The set $\text{Supp}(\pi)$ is sometimes called the core set. The following theorem characterizes the size of the core set and the minimum of g and establishes the equivalence of G -optimal and D -optimal designs.

400

401

402

403

404

405

406

407 **Theorem 5.1.** (Lattimore & Szepesvári, 2020, Kiefer-Wolfowitz) Assume that $\mathcal{A} \subset \mathbb{R}^d$ is compact and $\text{span}(\mathcal{A}) = \mathbb{R}^d$. The following are equivalent: (a) π^* is a minimiser of g ; (b) π^* is a maximiser of $f(\pi) = \log \det V(\pi)$; (c) $g(\pi^*) = d$.

408 Furthermore, there exists a minimiser π^* of g such that $|\text{Supp}(\pi^*)| \leq d(d+1)/2$.

409

410 However, the core set size of the G -optimal design given by Theorem 5.1 is at most $d(d+1)/2$, which may cause additional overhead in our variance-adaptive algorithm. To address this issue, we can find an approximate G -optimal design with a smaller core set size using the following theorem.

411

412 **Theorem 5.2** (Lattimore et al. 2020). Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is compact and $\text{span}(\mathcal{A}) = \mathbb{R}^d$. There exists a probability distribution $\pi \in \Delta(\mathcal{A})$ such that $g(\pi) \leq 2d$ and the cardinality of the core set of π is at most $4d \log \log d + 16$.

413

414

415

416

417

418 **Adaptive arm selection.** After obtaining the approximate G -optimal design π , we use it to guide the arm selection process. Unlike traditional G -optimal design-based algorithms, which pull arms according to the fixed distribution π , our algorithm adaptively selects arms based on the observed variances σ_t . Specifically, at each round t , we choose the arm \mathbf{a}_t that has been pulled the fewest times relative to its probability under π , weighted by the inverse of the observed variance. This adaptive strategy prevents over-exploration caused by the unpredictable and heteroscedastic nature of noise and ensures that we collect sufficient information from all arms in the core set of π .

418

419

420

421

422

423

424

425

426 **Weighted least-squares estimator.** Inspired by Zhou et al. (2021), we use a variance-weighted least-squares estimator to estimate the unknown parameter θ^* . Specifically, after T rounds of exploration, we compute the estimator $\hat{\theta}_T$ as follows:

427

428

429

430

431

$$\hat{\theta}_T = V_T^{-1} \sum_{t=1}^T \sigma_t^{-2} r_t \mathbf{a}_t, \quad V_T = I + \sum_{t=1}^T \sigma_t^{-2} \mathbf{a}_t \mathbf{a}_t^\top.$$

432 Under the finite action space regime, we show that this estimator achieves a tighter confidence bound
 433 compared to the general case, replacing the \sqrt{d} factor with $\sqrt{\log(|\mathcal{A}|)}$ in the confidence radius.
 434

435 Finally, we recommend the action \mathbf{a}_{T+1} that maximizes the estimated reward based on $\hat{\theta}_T$.
 436

437 5.2 SIMPLE REGRET BOUND FOR VAGD

439 **Theorem 5.3** (Simple Regret of Algorithm 2). Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is compact and $\text{span}(\mathcal{A}) =$
 440 \mathbb{R}^d . If we follow Algorithm 2, then it holds that with probability at least $1 - \delta$,

$$441 \quad \langle \boldsymbol{\theta}^*, \mathbf{a}^* \rangle - \langle \boldsymbol{\theta}^*, \mathbf{a}_{T+1} \rangle \leq 2 \sqrt{d \log(|\mathcal{A}|/\delta) / \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{4d \log \log d + 16} \frac{1}{[\sigma_T^{(i)}]^2} \right]}.$$

444 **Remark 5.4.** Theorem 5.3 establishes a simple regret bound for Algorithm 2 that depends on the
 445 harmonic mean of the noise variances σ_t^2 , while improving the dependence on the dimension d com-
 446 pared to Theorem 4.1. In particular, the \sqrt{d} factor in the numerator is replaced by $\sqrt{\log(|\mathcal{A}|)}$, which
 447 can be substantially smaller when the action set \mathcal{A} is finite and of moderate size. This improvement is
 448 obtained by exploiting the finite action space structure and employing a variance-adaptive G-optimal
 449 design exploration strategy. Consequently, Algorithm 2 is especially effective in settings with lim-
 450 ited action sets, enabling more efficient exploration and improved simple regret performance.
 451

452 6 LOWER BOUND

454 In this section, we establish a lower bound for the simple regret in linear bandits with heteroscedastic
 455 noise. Our lower bound nearly matches the upper bound in Theorem 4.1 up to logarithmic factors,
 456 demonstrating the optimality of our proposed algorithm.

457 **Theorem 6.1** (Instance-dependent lower bound.). For any $d \geq 2$ and $T \geq 1$, and any algorithm \mathcal{A} ,
 458 there exists a linear bandit instance with heteroscedastic Gaussian noise satisfying our assumptions
 459 such that the simple regret is lower bounded as follows:

$$460 \quad \mathbb{E}[\text{SR}(T)] \geq \frac{3}{16} d \cdot \left(\sum_{t=1}^T \frac{1}{\sigma_t^2} \right)^{-1/2}.$$

463 **Remark 6.2.** Theorem 6.1 establishes an variance-sequence-dependent lower bound for the simple
 464 regret in linear bandits with heteroscedastic noise. This lower bound matches the upper bound in
 465 Theorem 4.1 up to logarithmic factors, indicating that our proposed algorithm is nearly optimal in
 466 terms of its dependence on the harmonic mean of the variances σ_t^2 . This result highlights the funda-
 467 mental difficulty of the best-arm identification problem in linear bandits with heteroscedastic noise
 468 and underscores the effectiveness of our variance-adaptive approach. In contrast, in Table 1, the
 469 worst-case lower bound is derived by constructing an instance where all variances are equal, which
 470 does not capture the complexity of heteroscedastic linear bandits, especially when the variances vary
 471 significantly across actions and time steps.

472 7 CONCLUSION AND FUTURE WORK

475 In this paper, we study the sample complexity of stochastic linear bandits with heteroscedastic noise
 476 under a fixed action set. We propose a variance-adaptive algorithm that achieves a nearly instance-
 477 optimal simple regret bound, characterized by the harmonic mean of the noise variances. We further
 478 establish a nearly matching lower bound, demonstrating the optimality of our algorithm. Together,
 479 these results provide a comprehensive characterization of the statistical complexity of linear bandits
 480 with heteroscedastic noise.

481 There are several promising directions for future work. First, one could consider settings where the
 482 context is not fixed but instead sampled from an unknown distribution. Second, it would be natural
 483 to extend our results to the case where the noise variances are unknown and must be estimated
 484 from data. Third, in reinforcement learning, the variance of the noise is governed by the transition
 485 dynamics, which can themselves be estimated from historical data. Extending our results to Markov
 decision processes using such estimated variances would be an interesting direction to explore.

486 REFERENCES
487

488 Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic
489 bandits. *Advances in neural information processing systems*, 24, 2011.

490 Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits.
491 In *COLT*, pp. 217–226, 2009.

492 Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit
493 problem. *Machine learning*, 47(2):235–256, 2002.

495 Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans-
496 actions on Information Theory*, 59(11):7711–7717, 2013.

497 Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff func-
498 tions. In *Proceedings of the fourteenth international conference on artificial intelligence and
499 statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

500 David A Harville. Matrix algebra from a statistician’s perspective, 1998.

501 Jiafan He and Quanquan Gu. Variance-dependent regret lower bounds for contextual bandits. *arXiv
502 preprint arXiv:2503.12020*, 2025.

503 Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforce-
504 ment learning with function approximation: Minimax optimal and instance-dependent regret
505 bounds. *Advances in Neural Information Processing Systems*, 36:56576–56588, 2023.

506 Zeyu Jia, Jian Qian, Alexander Rakhlin, and Chen-Yu Wei. How does variance shape the regret
507 in contextual bandits? *Advances in Neural Information Processing Systems*, 37:83730–83785,
508 2024.

509 Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal
510 thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR,
511 2021.

512 Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration
513 is fast and optimal. In *International Conference on Machine Learning*, pp. 15239–15261. PMLR,
514 2023.

515 Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive
516 linear bandits and horizon-free linear mixture mdps. *arXiv preprint arXiv:2111.03289*, 2021.

517 Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with het-
518 eroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

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

520 Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in
521 bandits and in rl with a generative model. In *International conference on machine learning*, pp.
522 5662–5670. PMLR, 2020.

523 Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, New York, 1986.
524 ISBN 978-0-387-96317-4.

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

527 Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In
528 *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

529 Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochas-
530 tic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237, 2017.

531 Aldo Pacchiano. Second order bounds for contextual bandits with function approximation. In *The
532 Thirteenth International Conference on Learning Representations*, 2025.

540 Mark S. Pinsker. *Information and Information Stability of Random Variables and Processes*.
 541 Holden-Day, San Francisco, 1964. English translation of the 1960 Russian edition.
 542

543 David Ruppert. The elements of statistical learning: data mining, inference, and prediction, 2004.
 544

545 Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear
 546 stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*,
 547 31, 2018.

548 Qiang Sun. Do we need to estimate the variance in robust mean estimation? *arXiv preprint*
 549 *arXiv:2107.00118*, 2021.

550 Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being
 551 distributional: Second-order bounds for reinforcement learning. In *International Conference on*
 552 *Machine Learning*, pp. 51192–51213. PMLR, 2024.

553

554 Zhiyong Wang, Dongruo Zhou, John CS Lui, and Wen Sun. Model-based rl as a minimalist approach
 555 to horizon-free and second-order bounds. In *The Thirteenth International Conference on Learning*
 556 *Representations*, 2025.

557 Ruitu Xu, Yifei Min, and Tianhao Wang. Noise-adaptive thompson sampling for linear contextual
 558 bandits. *Advances in Neural Information Processing Systems*, 36:23630–23657, 2023.

559

560 Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear
 561 bandits with heavy-tailed payoffs. *arXiv preprint arXiv:2004.13465*, 2020.

562 Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In
 563 *Proceedings of the 18th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*,
 564 pp. 222–227. IEEE, 1977.

565 Chenlu Ye, Yujia Jin, Alekh Agarwal, and Tong Zhang. Catoni contextual bandits are robust to
 566 heavy-tailed rewards. In *Forty-second International Conference on Machine Learning*, 2025.

567

568 Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement
 569 learning without domain knowledge using value function bounds. In *International Conference on*
 570 *Machine Learning*, pp. 7304–7312. PMLR, 2019.

571

572 Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets
 573 for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*,
 574 34:4342–4355, 2021.

575

576 Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret
 577 bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In
 578 *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4977–5020. PMLR, 2023a.

579

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

582

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for
 583 reinforcement learning with general function approximation. *Advances in Neural Information*
 584 *Processing Systems*, 37:94684–94735, 2024.

585

Runze Zhao, Yue Yu, Ruhan Wang, Chunfeng Huang, and Dongruo Zhou. Instance-dependent
 586 continuous-time reinforcement learning via maximum likelihood estimation. *arXiv preprint*
 587 *arXiv:2508.02103*, 2025.

588

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning
 589 for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349,
 590 2022.

591

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning
 592 for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–
 593 4576. PMLR, 2021.

594 Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforce-
595 ment learning: Best of both worlds in stochastic and deterministic environments. In *International*
596 *Conference on Machine Learning*, pp. 42878–42914. PMLR, 2023.
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

648 **A LLM USAGE**
649650 We used an LLM only for grammatical and stylistic polishing of the manuscript. No research ideas
651 or results were generated by the LLM. The authors wrote and verified all technical content.
652653 **B PROOF OF THEOREM 4.1**
654655 **Lemma B.1** (Matrix Inversion Lemma, Harville (1998)). For any invertible matrix $A \in \mathbb{R}^{d \times d}$,
656 vector $u, v \in \mathbb{R}^d$, it holds that
657

658
$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.$$

659

660 **Lemma B.2** (Elliptical Potential Lemma, Abbasi-Yadkori et al. (2011)). For any sequence of vec-
661 tors $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$, let $V_0 = \lambda \mathbf{I}$ for some $\lambda > 0$ and $V_t = V_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$ for $t \geq 1$. If $\|\mathbf{x}_t\|_2 \leq L$ for
662 all t , then we have

663
$$\sum_{t=1}^T \min\{1, \|\mathbf{x}_t\|_{V_{t-1}}^2\} \leq 2d \log \frac{\lambda + TL^2}{d\lambda}.$$

664
665

666 **Lemma B.3.** With probability at least $1 - \delta$, it holds for all $t \in [T]$ that
667

668
$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \beta_t := 2\sqrt{\lambda} + 16\sqrt{\log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}}.$$

669
670

671 *Proof.* The proof follows from the standard analysis of OFUL (Abbasi-Yadkori et al., 2011) with
672 variance-weighted updates.
673

674 We have

675
$$\begin{aligned} \|\hat{\theta}_t - \theta^*\|_{V_t}^2 &= (\hat{\theta}_t - \theta^*)^\top V_t (\hat{\theta}_t - \theta^*) \\ 676 &= \left(\sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s r_s - \sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \mathbf{a}_s^\top \theta^* - \lambda \theta^* \right)^\top V_t^{-1} \cdot V_t \cdot V_t^{-1} \cdot \left(\sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s r_s - \sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \mathbf{a}_s^\top \theta^* - \lambda \theta^* \right) \\ 677 &= \left(\sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \eta_s - \lambda \theta^* \right) V_t^{-1} \left(\sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \eta_s - \lambda \theta^* \right) \\ 678 &\leq 2\lambda^2 \|\theta^*\|_{V_t^{-1}}^2 + 2 \underbrace{\left(\sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \eta_s \right)^\top V_t^{-1} \left(\sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \eta_s \right)}_{I_{0,t}}, \end{aligned} \tag{B.1}$$

679
680
681
682
683
684
685
686

687 where the second equality follows from the definition of $\hat{\theta}_t$, the third equality follows from the
688 definition of r_s and η_s , the inequality follows from Young's inequality. To further bound $I_{0,t}$, we
689 introduce the following notation:

690
$$\mathbf{d}_0 = 0, \quad \mathbf{d}_t = \sum_{s=1}^t \sigma_s^{-2} \mathbf{a}_s \eta_s, \quad \mathcal{I}_t = \mathbf{1}(0 \leq s \leq t, I_{0,s} \leq \gamma_s), \gamma_s := 64 \log(4s^2/\delta) \cdot d \log \frac{d\lambda + s\sigma_{\min}^{-2}}{d\lambda}.$$

691
692

693 Decomposing $I_{0,t}$ into a martingale difference sequence, we have

694
$$\begin{aligned} I_{0,t} &= \mathbf{d}_{t-1}^\top V_t^{-1} \mathbf{d}_{t-1} + 2\sigma_t^{-2} \eta_t \mathbf{a}_t^\top V_t^{-1} \mathbf{d}_{t-1} + \sigma_t^{-4} \eta_t^2 \mathbf{a}_t^\top V_t^{-1} \mathbf{a}_t \\ 695 &\leq I_{0,t-1} + 2 \underbrace{\sigma_t^{-2} \eta_t \mathbf{a}_t^\top V_t^{-1} \mathbf{d}_{t-1}}_{I_{1,t}} + \underbrace{\sigma_t^{-4} \eta_t^2 \mathbf{a}_t^\top V_t^{-1} \mathbf{a}_t}_{I_{2,t}}. \end{aligned} \tag{B.2}$$

696
697
698

699 From the matrix inversion lemma (Lemma B.1), we have
700

701
$$I_{1,t} = \sigma_t^{-2} \eta_t \mathbf{a}_t^\top \left(V_{t-1}^{-1} - \frac{\sigma_t^{-2} V_{t-1}^{-1} \mathbf{a}_t \mathbf{a}_t^\top V_{t-1}^{-1}}{1 + \sigma_t^{-2} \mathbf{a}_t^\top V_{t-1}^{-1} \mathbf{a}_t} \right) \mathbf{d}_{t-1}$$

$$\begin{aligned}
&= \sigma_t^{-2} \eta_t \left(\mathbf{a}_t V_{t-1}^{-1} \mathbf{d}_{t-1} - \frac{\sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2 \mathbf{a}_t V_{t-1}^{-1} \mathbf{d}_{t-1}}{1 + \sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2} \right) \\
&= \sigma_t^{-2} \eta_t \frac{\mathbf{a}_t V_{t-1}^{-1} \mathbf{d}_{t-1}}{1 + \sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2}.
\end{aligned}$$

Based on our assumption on the noise η_t , $I_{1,t} \cdot \mathcal{I}_t$ is also a sub-Gaussian random variable with variance proxy bounded by

$$\sigma_t^{-2} \frac{\|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2 \|\mathbf{d}_{t-1}\|_{V_{t-1}^{-1}}^2}{(1 + \sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2)^2} \cdot \mathcal{I}_t.$$

Adding $I_{1,s} \cdot \mathcal{I}_s$ up to t and using Lemma D.3, we have with probability at least $1 - \delta/2$,

$$\begin{aligned}
\sum_{s=1}^t I_{1,s} \cdot \mathcal{I}_s &\leq \sqrt{2 \log(2/\delta) \sum_{s=1}^t \sigma_s^{-2} \frac{\|\mathbf{a}_s\|_{V_{s-1}^{-1}}^2 \|\mathbf{d}_{s-1}\|_{V_{s-1}^{-1}}^2}{(1 + \sigma_s^{-2} \|\mathbf{a}_s\|_{V_{s-1}^{-1}}^2)^2} \cdot \mathcal{I}_s} \\
&\leq \sqrt{2 \log(2/\delta) \sum_{s=1}^t \min\{1, \|\sigma_s^{-1} \mathbf{a}_s\|_{V_{s-1}^{-1}}^2\} \cdot \gamma_t} \\
&\leq \sqrt{2 \gamma_t \log(2/\delta) \cdot 2d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}} \\
&\leq \frac{1}{4} \gamma_t + 8 \log(2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda},
\end{aligned} \tag{B.3}$$

where the second inequality follows from the definition of \mathcal{I}_s and the fact that $\frac{\sigma_s^{-2} \|\mathbf{a}_s\|_{V_{s-1}^{-1}}^2}{(1 + \sigma_s^{-2} \|\mathbf{a}_s\|_{V_{s-1}^{-1}}^2)^2} \leq 1$, the third inequality follows from Lemma B.2, and the last inequality follows from Young's inequality.

Using union bound over all $t \geq 1$, we have with probability at least $1 - \delta/2$,

$$\sum_{s=1}^t I_{1,s} \cdot \mathcal{I}_s \leq \frac{1}{4} \gamma_t + 8 \log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}$$

for all $t \geq 1$.

For the second term $I_{2,t}$, it follows from the matrix inversion lemma (Lemma B.1) that

$$\begin{aligned}
I_{2,t} &= \sigma_t^{-4} \eta_t^2 \mathbf{a}_t^\top \left(V_{t-1}^{-1} - \frac{\sigma_t^{-2} V_{t-1}^{-1} \mathbf{a}_t \mathbf{a}_t^\top V_{t-1}^{-1}}{1 + \sigma_t^{-2} \mathbf{a}_t^\top V_{t-1}^{-1} \mathbf{a}_t} \right) \mathbf{a}_t \\
&= \sigma_t^{-4} \eta_t^2 \left(\|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2 - \frac{\sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^4}{1 + \sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2} \right) \\
&= \sigma_t^{-4} \eta_t^2 \frac{\|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2}{1 + \sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2},
\end{aligned}$$

Using union bound over $t \geq 1$, we have with probability at least $1 - \delta/2$,

$$I_{2,t} \leq \sigma_t^{-4} \sigma_t^2 \log(4t^2/\delta) \frac{\|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2}{1 + \sigma_t^{-2} \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2}$$

756 for all $t \geq 1$.

757 Thus, for any $t \geq 1$,

$$\begin{aligned}
 759 \quad \sum_{s=1}^t I_{2,s} &\leq \sum_{s=1}^t \sigma_s^{-2} \log(4s^2/\delta) \frac{\|\mathbf{a}_s\|_{V_{s-1}^{-1}}^2}{1 + \sigma_s^{-2} \|\mathbf{a}_s\|_{V_{s-1}^{-1}}^2} \\
 760 \quad &\leq \log(4t^2/\delta) \sum_{s=1}^t \min\{1, \|\sigma_s^{-1} \mathbf{a}_s\|_{V_{s-1}^{-1}}^2\} \\
 761 \quad &\leq 2 \log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}, \tag{B.4}
 \end{aligned}$$

762 where the last inequality follows from Lemma B.2.

763 Substituting (B.3) and (B.4) into (B.2), and using induction on t , we have with probability at least
764 $1 - \delta$,

$$765 \quad I_{0,t} \leq \gamma_t := 64 \log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda},$$

766 which further implies that

$$767 \quad \|\hat{\theta}_t - \theta^*\|_{V_t}^2 \leq 2\lambda^2 \|\theta^*\|_{V_t^{-1}}^2 + 2I_{0,t} \leq 2\lambda + 256 \log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}.$$

768 \square

769 **Lemma B.4.** If we follow Algorithm 1 to choose the action \mathbf{a}_t , then it holds for any $t \in [T]$ that

$$\begin{aligned}
 770 \quad \|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2 &\leq \min_{1 \leq k \leq t+1} \left\{ x^2 = \frac{\iota(t) - k + 1}{\sum_{i=k}^t \frac{1}{[\sigma_t^{(i)}]^2}} \middle| \sigma_t^{(i)} \text{ is the } i\text{-th smallest element in } \{\sigma_\tau^2\}_{\tau=1}^t, \right. \\
 771 \quad &\quad \left. x \in [\sigma_t^{(k-1)}, \sigma_t^{(k)}] \right\},
 \end{aligned}$$

772 where $\iota(t) = 2d \log \left(\frac{d + \sum_{\tau \in [t]} \sigma_\tau^{-2}}{d} \right)$.

773 *Proof.* When $x \in [0, 1]$, $x \leq 2 \log(1 + x)$, which further indicates that

$$\begin{aligned}
 774 \quad \sum_{\tau \in [t]} \min\left\{1, \frac{1}{\sigma_\tau^2} \|\mathbf{a}_\tau\|_{V_{\tau-1}^{-1}}^2\right\} &\leq 2 \sum_{\tau \in [t]} \log\left(1 + \frac{1}{\sigma_\tau^2} \|\mathbf{a}_\tau\|_{V_{\tau-1}^{-1}}^2\right) \\
 775 \quad &\leq 2 \log \frac{\det(V_t)}{\det(V_0)} \\
 776 \quad &\leq 2d \log \left(\frac{d + \sum_{\tau \in [t]} \sigma_\tau^{-2}}{d} \right). \tag{B.5}
 \end{aligned}$$

777 Let

$$778 \quad \iota(t) := 2d \log \left(\frac{d + \sum_{\tau \in [t]} \sigma_\tau^{-2}}{d} \right).$$

779 Note that for $i \leq t$,

$$780 \quad \|\mathbf{a}_t\|_{V_{t-1}^{-1}} \leq \|\mathbf{a}_t\|_{V_{i-1}^{-1}} \leq \|\mathbf{a}_i\|_{V_{i-1}^{-1}}$$

801 due to the fact that $V_{t-1} \succeq V_{i-1}$ and the definition of \mathbf{a}_i in Algorithm 1. Therefore, following
802 inequality (B.5), we obtain that

$$803 \quad \sum_{\tau \in [t]} \min\left\{1, \frac{1}{\sigma_\tau^2} \|\mathbf{a}_\tau\|_{V_{\tau-1}^{-1}}^2\right\} \leq \sum_{\tau \in [t]} \min\left\{1, \frac{1}{\sigma_\tau^2} \|\mathbf{a}_t\|_{V_{\tau-1}^{-1}}^2\right\} \leq \iota(t). \tag{B.6}$$

As LHS of (B.6) is strictly increasing with respect to $\|\mathbf{a}_t\|_{V_{t-1}^{-1}}$ as long as $\iota(t) < t$, we can derive a bound for $\|\mathbf{a}_t\|_{V_{t-1}^{-1}}$ by solving

$$\sum_{\tau \in [t]} \min \left\{ 1, \frac{1}{\sigma_\tau^2} x^2 \right\} = \iota(t). \quad (\text{B.7})$$

To solve (B.7), we define a sorted sequence of $\{\sigma_i\}_{i=1}^t$ in increasing order, denoted as

$$\sigma_t^{(1)} \leq \sigma_t^{(2)} \leq \dots \leq \sigma_t^{(t)}.$$

Let $\sigma_t^{(0)} := 0$. Suppose that $x \in [\sigma_t^{(k-1)}, \sigma_t^{(k)}]$ for some $k \in [t]$. Then for $i < k$, we have

$$\min \left\{ 1, \frac{1}{[\sigma_t^{(i)}]^2} x^2 \right\} = 1;$$

for $i \geq k$, we have

$$\min \left\{ 1, \frac{1}{[\sigma_t^{(i)}]^2} x^2 \right\} = \frac{x^2}{[\sigma_t^{(i)}]^2}.$$

After rewriting the LHS of the above equality, we have

$$k - 1 + \sum_{i=k}^t \frac{x^2}{[\sigma_t^{(i)}]^2} = \iota(t).$$

We can then rearrange the above inequality to obtain

$$\|\mathbf{a}_t\|_{V_{t-1}^{-1}}^2 \leq x^2,$$

where

$$x^2 := \min_{1 \leq k \leq t+1} \left\{ \frac{\iota(t) - k + 1}{\sum_{i=k}^t \frac{1}{[\sigma_t^{(i)}]^2}} \middle| [\sigma_t^{(k-1)}]^2 \leq \frac{\iota(t) - k + 1}{\sum_{i=k}^t \frac{1}{[\sigma_t^{(i)}]^2}} \leq [\sigma_t^{(k)}]^2 \right\}.$$

This completes the proof. \square

Remark B.5. In the previous proof, x^2 is the solution for the implicit equation (B.7). We can also rewrite it as

$$sx^2 \sum_{i=1}^t \frac{1}{\max(\sigma_i^2, x^2)} = \iota(t),$$

which implies that

$$x^2 = \frac{\iota(t)}{\sum_{i=1}^t \frac{1}{\max(\sigma_i^2, x^2)}}.$$

We further have

$$x^2 \leq \frac{\iota(t)}{\sum_{i=1}^t \frac{1}{\sigma_i^2 + x^2}} \leq \frac{\iota(t)}{\frac{t}{x^2 + t^{-1} \sum_{i=1}^t \sigma_i^2}} = \frac{\iota(t)(x^2 + t^{-1} \sum_{i=1}^t \sigma_i^2)}{t}.$$

Rearranging the above inequality, we obtain

$$x^2 \leq \frac{\iota(t) \sum_{i=1}^t \sigma_i^2}{t[t - \iota(t)]} = \tilde{O}\left(\frac{d \sum_{i=1}^t \sigma_i^2}{t^2}\right)$$

when $t = \Omega(d)$.

864 **Theorem B.6** (Simple Regret of VAE, restatement of Theorem 4.1). If we set $\lambda = 1$ and $\beta_t =$
 865 $2\sqrt{\lambda} + 16\sqrt{\log(4t^2/\delta) \cdot d \log \frac{d\lambda + t\sigma_{\min}^{-2}}{d\lambda}}$ in Algorithm 1, then with probability at least $1 - \delta$, the
 866 simple regret of Algorithm 1 is bounded as
 867

$$868 \text{SR}(T) = \tilde{O}(\sqrt{d}) \min_{1 \leq k \leq T+1} \left\{ x = \sqrt{\frac{\iota(T) - k + 1}{\sum_{i=k}^T \frac{1}{[\sigma_T^{(i)}]^2}}} \left[\sigma_T^{(i)} \right]^2 \text{ is the } i\text{-th smallest element in } \{\sigma_\tau^2\}_{\tau=1}^T, \right. \\ 869 \left. x \in [\sigma_T^{(k-1)}, \sigma_T^{(k)}] \right\}, \\ 870$$

$$871 \text{where } \iota(T) = 2d \log \left(\frac{d + \sum_{\tau \in [T]} \sigma_\tau^{-2}}{d} \right). \\ 872 \\ 873$$

874 *Proof.* By Lemma B.3, with probability at least $1 - \delta$, it holds for all $t \in [T]$ that $\theta^* \in \mathcal{C}_t := \{\theta \in$
 875 $\mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{V_t} \leq \beta_t\}$. In the following, we condition on the event that $\theta^* \in \mathcal{C}_t$ for all $t \in [T]$.
 876

877 With the conditioned event, we can show by induction that for any $t \in [T]$, $\mathbf{a}^* \in \mathcal{A}_t$:

$$878 \max_{\theta \in \mathcal{C}_{t-1}} \langle \theta, \mathbf{a}^* \rangle \geq \langle \theta^*, \mathbf{a}^* \rangle \geq \max_{\mathbf{a} \in \mathcal{A}_{t-1}} \langle \theta^*, \mathbf{a} \rangle \geq \max_{\mathbf{a} \in \mathcal{A}_{t-1}} \min_{\theta \in \mathcal{C}_{t-1}} \langle \theta, \mathbf{a} \rangle, \\ 879$$

880 where the first inequality follows from the fact that $\theta^* \in \mathcal{C}_{t-1}$, the second inequality follows from
 881 the definition of \mathbf{a}^* , and the last inequality follows from the definition of \mathcal{A}_{t-1} .
 882

883 Let $\mathbf{a}^* = \operatorname{argmax}_{\mathbf{a} \in \mathcal{A}} \langle \theta^*, \mathbf{a} \rangle$ be the optimal action. By the definition of simple regret, we have
 884

$$885 \text{SR}(T) = \langle \theta^*, \mathbf{a}^* - \mathbf{a}_T \rangle \\ 886 \leq \max_{\mathbf{a} \in \mathcal{A}_T} \max_{\theta \in \mathcal{C}_{T-1}} \langle \theta, \mathbf{a} \rangle - \langle \theta^*, \mathbf{a}_T \rangle \\ 887 \leq \max_{\mathbf{a} \in \mathcal{A}_T} \langle \hat{\theta}_{T-1}, \mathbf{a} \rangle + \beta_T \|\mathbf{a}_T\|_{V_{T-1}^{-1}} - \langle \hat{\theta}_{T-1}, \mathbf{a}_T \rangle + \beta_T \|\mathbf{a}_T\|_{V_{T-1}^{-1}}, \\ 888$$

889 where the first inequality follows from the event that $\theta^* \in \mathcal{C}_t$ for all $t \in [T]$, and the fact that
 890 $\mathbf{a}^* \in \mathcal{A}_T$, the second inequality follows from the definition of \mathcal{C}_{T-1} and the definition of \mathbf{a}_T in
 891 Algorithm 1.

892 Then it suffices to bound $\max_{\mathbf{a} \in \mathcal{A}_T} \langle \hat{\theta}_{T-1}, \mathbf{a} \rangle - \langle \hat{\theta}_{T-1}, \mathbf{a}_T \rangle$. Since $\mathbf{a}_T \in \mathcal{A}_T$, it is guaranteed that
 893 there exists $\theta'_T \in \mathcal{C}_{T-1}$ such that
 894

$$895 \langle \theta'_T, \mathbf{a}_T \rangle - \max_{\mathbf{a} \in \mathcal{A}_T} \min_{\theta \in \mathcal{C}_{T-1}} \langle \theta, \mathbf{a} \rangle \geq 0, \\ 896$$

897 which further implies that
 898

$$899 \max_{\mathbf{a} \in \mathcal{A}_T} \langle \hat{\theta}_{T-1}, \mathbf{a} \rangle - \langle \hat{\theta}_{T-1}, \mathbf{a}_T \rangle = \max_{\mathbf{a} \in \mathcal{A}_T} \langle \hat{\theta}_{T-1}, \mathbf{a} \rangle - \langle \theta'_T, \mathbf{a}_T \rangle + \langle \theta'_T - \hat{\theta}_{T-1}, \mathbf{a}_T \rangle \\ 900 \leq \max_{\mathbf{a} \in \mathcal{A}_T} \langle \hat{\theta}_{T-1}, \mathbf{a} \rangle - \langle \theta'_T, \mathbf{a}_T \rangle + \beta_T \|\mathbf{a}_T\|_{V_{T-1}^{-1}} \\ 901 \leq 3\beta_T \max_{\mathbf{a} \in \mathcal{A}_T} \|\mathbf{a}\|_{V_{T-1}^{-1}} = 3\beta_T \|\mathbf{a}_T\|_{V_{T-1}^{-1}}, \\ 902$$

903 where the last inequality holds because $\max_{\mathbf{a} \in \mathcal{A}_T} \langle \hat{\theta}_{T-1}, \mathbf{a} \rangle - \max_{\mathbf{a} \in \mathcal{A}_T} \min_{\theta \in \mathcal{C}_{T-1}} \langle \theta, \mathbf{a} \rangle \leq$
 904 $2\beta_T \max_{\mathbf{a} \in \mathcal{A}_T} \|\mathbf{a}\|_{V_{T-1}^{-1}}$ based on the definition of \mathcal{C}_{T-1} and the last equality follows from the action
 905 selection rule in Algorithm 1.

906 Hence, the simple regret of Algorithm 1 is bounded as
 907

$$908 \text{SR}(T) \leq 5\beta_T \|\mathbf{a}_T\|_{V_{T-1}^{-1}} \\ 909 \leq \tilde{O}(\sqrt{d}) \min_{1 \leq k \leq T+1} \left\{ x = \sqrt{\frac{\iota(T) - k + 1}{\sum_{i=k}^T \frac{1}{[\sigma_T^{(i)}]^2}}} \left[\sigma_T^{(i)} \right]^2 \text{ is the } i\text{-th smallest element in } \{\sigma_\tau^2\}_{\tau=1}^T, \right. \\ 910 \left. x \in [\sigma_T^{(k-1)}, \sigma_T^{(k)}] \right\}, \\ 911$$

912 where the last inequality follows from the choice of β_T and the result in the previous lemma. \square
 913

918 **C PROOF OF THEOREM 6.1**
 919

920 The proofs of our lower bound require the following Lemmas, which is standard technique used for
 921 lower bound.

922 **Lemma C.1** (Le Cam two-point method (Le Cam, 1986)). Let P and Q be probability measures on
 923 the same measurable space, and let $\phi : \Omega \rightarrow \{0, 1\}$ be any (possibly randomized) test. Then
 924

$$925 \frac{1}{2} \left(P\{\phi = 1\} + Q\{\phi = 0\} \right) \geq \frac{1}{2} \left(1 - \delta_{\text{TV}}(P, Q) \right), \\ 926$$

927 where $\delta_{\text{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$ is total variation distance. Moreover, by Pinsker's in-
 928 equality,

$$929 \delta_{\text{TV}}(P, Q) \leq \sqrt{\frac{1}{2} \text{KL}(P\|Q)}. \\ 930$$

931 Hence the average error of any test is bounded below by

$$932 \frac{1}{2} \left(P\{\phi = 1\} + Q\{\phi = 0\} \right) \geq \frac{1}{2} \left(1 - \sqrt{\frac{1}{2} \text{KL}(P\|Q)} \right). \\ 933$$

934 **Lemma C.2** (Pinsker's inequality (Pinsker, 1964)). For any probability measures P, Q ,

$$936 \delta_{\text{TV}}(P, Q) \leq \sqrt{\frac{1}{2} \text{KL}(P\|Q)}. \\ 937$$

938 **Lemma C.3** (Yao's minimax principle (Yao, 1977)). Let Π be the set of deterministic algorithms
 939 (measurable decision rules), \mathcal{P} a family of instances with loss $L(\pi, \theta)$ and let \mathcal{D} be distributions
 940 over \mathcal{P} . Then

$$941 \inf_{\pi \in \Pi} \sup_{\theta \in \mathcal{P}} \mathbb{E}_{\theta} [L(\pi, \theta)] \geq \sup_{\mu \in \mathcal{D}} \inf_{\pi \in \Pi} \mathbb{E}_{\theta \sim \mu} \mathbb{E}_{\theta} [L(\pi, \theta)].$$

942 In other words, the worst-case risk of the best deterministic algorithm is at least the Bayes risk under
 943 any prior μ .

944 Now, we are ready to prove our lower bound.

945 *Proof of Theorem 6.1. Step 1 (per-coordinate two-point divergence).* Let action set $\mathcal{A} =$
 946 $\{-1, 1\}^d$. The unknown parameter belongs to

$$947 \Theta = \{-c, +c\}^d \quad \text{for some } c > 0. \\ 948$$

949 The optimal arm for θ is $\mathbf{a}^*(\theta) = \text{sign}(\theta)$. Since each coordinate mistake costs $2c$, the *simple regret*
 950 is

$$951 \text{SR}(T) = \mathbb{E}_{\theta} [\langle \theta, \mathbf{a}^*(\theta) \rangle - \langle \theta, \hat{\mathbf{a}}_T \rangle] = 2c \cdot \mathbb{E}_{\theta} [\text{Ham}(\hat{\mathbf{a}}_T, \text{sign}(\theta))], \\ 952$$

953 where

$$954 \text{Ham}(\mathbf{u}, \mathbf{v}) \triangleq \sum_{j=1}^d \mathbb{1}\{u_j \neq v_j\}. \\ 955$$

956 Let $S = \sum_{t=1}^T \sigma_t^{-2}$ denote the total precision. Fix $j \in [d]$ and consider neighboring parameters that
 957 differ only on coordinate j :

$$958 \theta_j^{(j,+)} = +c, \quad \theta_j^{(j,-)} = -c, \quad \theta_k^{(j,+)} = \theta_k^{(j,-)} \in \{\pm c\} \quad (k \neq j). \\ 959$$

960 Let $\mathbb{P}_+^{(j)} \equiv \mathbb{P}_{\theta^{(j,+)}}, \mathbb{P}_-^{(j)} \equiv \mathbb{P}_{\theta^{(j,-)}}$ be the laws of the full transcript under these two instances.
 961 By the chain rule for KL and the fact that $\mathbf{a}_t = \pi_t(H_{t-1})$ contributes to KL,
 962

$$963 \text{KL}(\mathbb{P}_+^{(j)} \| \mathbb{P}_-^{(j)}) = \sum_{t=1}^T \mathbb{E} [\text{KL}(\mathcal{N}(\mu_t^{(j,+)}, \sigma_t^2) \| \mathcal{N}(\mu_t^{(j,-)}, \sigma_t^2))], \\ 964$$

965 where $\mu_t^{(j,\pm)} = \langle \theta^{(j,\pm)}, \mathbf{a}_t \rangle$. The means differ only on coordinate j , so $\mu_t^{(j,+)} - \mu_t^{(j,-)} = (+c -$
 966 $-c)a_{t,j} = 2c a_{t,j}$ and thus
 967

$$968 \text{KL}(\mathcal{N}(\mu_t^{(j,+)}, \sigma_t^2) \| \mathcal{N}(\mu_t^{(j,-)}, \sigma_t^2)) = \frac{(\mu_t^{(j,+)} - \mu_t^{(j,-)})^2}{2\sigma_t^2} = \frac{(2c a_{t,j})^2}{2\sigma_t^2} = \frac{2c^2}{\sigma_t^2}, \\ 969$$

972 since $a_{t,j}^2 = 1$. Summing over t yields the instance divergence
 973

$$974 \quad 975 \quad 976 \quad \text{KL}(\mathbb{P}_+^{(j)} \parallel \mathbb{P}_-^{(j)}) = 2c^2 \sum_{t=1}^T \frac{1}{\sigma_t^2} = 2c^2 S.$$

977 **Step 2 (Le Cam + Pinsker \Rightarrow per-coordinate average error).** Apply Lemma C.1 with the test
 978 $\phi = \mathbb{1}\{\hat{s}_j = +1\}$ between $P_+^{(j)}$ and $P_-^{(j)}$, and bound the TV distance by Lemma C.2. We get
 979

$$980 \quad 981 \quad \frac{1}{2}(P_+^{(j)}\{\hat{s}_j \neq +1\} + P_-^{(j)}\{\hat{s}_j \neq -1\}) \geq \frac{1}{2}\left(1 - \sqrt{\frac{1}{2} \text{KL}(P_+^{(j)} \parallel P_-^{(j)})}\right) = \frac{1}{2}(1 - c\sqrt{S}).$$

982 Choose $c = \frac{1}{4}S^{-1/2}$ to obtain the uniform per-coordinate bound
 983

$$984 \quad 985 \quad \frac{1}{2}(P_+^{(j)}\{\hat{s}_j \neq +1\} + P_-^{(j)}\{\hat{s}_j \neq -1\}) \geq \frac{3}{8} \quad \text{for all } j \in [d].$$

987 **Step 3 (aggregate to d coordinates under Hamming loss).** This part requires the following lemma.
 988 The proof of this Lemma is defer to Appendix C.1.

989 **Lemma C.4** (From two-point bounds to a d -dimensional Hamming-risk lower bound). Let $\Theta = \{\pm c\}^d$ and put the uniform prior on Θ . Let H_T denote the full transcript and let $\hat{s} = \hat{s}(H_T) \in \{\pm 1\}^d$ be any estimator of $\text{sign}(\theta)$. For each coordinate $j \in [d]$, fix two instances $\theta^{(j,+)}, \theta^{(j,-)} \in \Theta$ that differ only in coordinate j (i.e., $\theta_j^{(j,+)} = +c$, $\theta_j^{(j,-)} = -c$ and $\theta_k^{(j,+)} = \theta_k^{(j,-)}$ for all $k \neq j$). Denote by $P_+^{(j)}$ and $P_-^{(j)}$ the corresponding laws of H_T under these two instances. Assume that for every j ,

$$996 \quad \frac{1}{2}(P_+^{(j)}\{\hat{s}_j \neq +1\} + P_-^{(j)}\{\hat{s}_j \neq -1\}) \geq \eta, \quad \text{for some } \eta \in (0, 1/2]. \quad (\text{C.1})$$

997 Then the Bayes Hamming risk under the uniform prior satisfies
 998

$$999 \quad \mathbb{E}_\theta \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] \geq \eta d, \quad (\text{C.2})$$

1000 and, consequently, by Yao's minimax principle,
 1001

$$1002 \quad \inf_{\pi} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] \geq \eta d. \quad (\text{C.3})$$

1004 Invoke Lemma C.4 with $\eta = \frac{3}{8}$ to conclude that the Bayes Hamming risk under the uniform prior
 1005 on Θ satisfies
 1006

$$1007 \quad \mathbb{E}_\theta \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] \geq \frac{3}{8} d.$$

1008 By Lemma C.3 (Yao's principle), this also lower-bounds the minimax Hamming risk:
 1009

$$1010 \quad \inf_{\pi} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] \geq \frac{3}{8} d.$$

1013 **Step 4 (convert Hamming loss to simple regret).** In $\Theta = \{\pm c\}^d$, each coordinate mistake costs
 1014 exactly $2c$ in value. Therefore
 1015

$$1016 \quad \inf_{\pi} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\text{SR}(\pi, \theta, T)] \geq 2c \cdot \frac{3}{8} d = \frac{3}{4} c d = \frac{3}{16} d S^{-1/2},$$

1018 where we used $c = \frac{1}{4}S^{-1/2}$. □
 1019

1020 C.1 PROOF OF LEMMA C.4

1022 *Proof of Lemma C.4.* By definition of Hamming distance,
 1023

$$1024 \quad 1025 \quad \text{Ham}(\hat{s}, \text{sign}(\theta)) = \sum_{j=1}^d \mathbb{1}\{\hat{s}_j \neq \text{sign}(\theta_j)\}.$$

1026 Taking expectation under the model instance θ and then averaging over the uniform prior on Θ , the
 1027 Bayes Hamming risk equals
 1028

$$1029 \quad \mathbb{E}_\theta \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] = \sum_{j=1}^d \mathbb{E}_\theta \mathbb{P}_\theta \{ \hat{s}_j \neq \text{sign}(\theta_j) \}. \quad (\text{C.4})$$

1032 Fix a coordinate $j \in [d]$. Write $\theta = (\theta_j, \theta_{-j})$, and condition on θ_{-j} . Under the *uniform* prior on
 1033 Θ , we have $\mathbb{P}\{\theta_j = +c \mid \theta_{-j}\} = \mathbb{P}\{\theta_j = -c \mid \theta_{-j}\} = \frac{1}{2}$. Hence, conditionally on θ_{-j} , the law of
 1034 the transcript H_T is the equal mixture
 1035

$$1036 \quad \mathcal{M}_{\theta_{-j}}^{(j)} = \frac{1}{2} P_+^{(j)} + \frac{1}{2} P_-^{(j)}, \quad (\text{C.5})$$

1038 where $P_+^{(j)}$ and $P_-^{(j)}$ are the endpoint laws that differ only in coordinate j (with θ_{-j} held fixed).
 1039

1040 **Identify the conditional Bayes error for coordinate j with the average two-point error.** Consider
 1041 the indicator loss for estimating $\text{sign}(\theta_j)$ by the (measurable) decision rule $\hat{s}_j(H_T) \in \{\pm 1\}$. The
 1042 conditional Bayes error probability for coordinate j , given θ_{-j} , is
 1043

$$1043 \quad \text{Err}_j(\theta_{-j}) \triangleq \mathbb{E}_{H_T \sim \mathcal{M}_{\theta_{-j}}^{(j)}} \left[\frac{1}{2} \mathbb{1}\{\hat{s}_j(H_T) \neq +1\} + \frac{1}{2} \mathbb{1}\{\hat{s}_j(H_T) \neq -1\} \right].$$

1045 Using (C.5), this equals the *average* of the two endpoint errors:
 1046

$$1047 \quad \text{Err}_j(\theta_{-j}) = \frac{1}{2} P_+^{(j)} \{ \hat{s}_j \neq +1 \} + \frac{1}{2} P_-^{(j)} \{ \hat{s}_j \neq -1 \}. \quad (\text{C.6})$$

1049 By the assumption (C.1), we have, for every θ_{-j} ,

$$1050 \quad \text{Err}_j(\theta_{-j}) \geq \eta. \quad (\text{C.7})$$

1053 **Average over θ_{-j} and sum over j .** By the tower property (law of total expectation),
 1054

$$1055 \quad \mathbb{E}_\theta \mathbb{P}_\theta \{ \hat{s}_j \neq \text{sign}(\theta_j) \} = \mathbb{E}_{\theta_{-j}} [\text{Err}_j(\theta_{-j})].$$

1056 Combining with (C.7) yields
 1057

$$1058 \quad \mathbb{E}_\theta \mathbb{P}_\theta \{ \hat{s}_j \neq \text{sign}(\theta_j) \} \geq \eta \quad \text{for every } j \in [d]. \quad (\text{C.8})$$

1060 Summing (C.8) over $j = 1, \dots, d$ and using (C.4) gives
 1061

$$1062 \quad \mathbb{E}_\theta \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] = \sum_{j=1}^d \mathbb{E}_\theta \mathbb{P}_\theta \{ \hat{s}_j \neq \text{sign}(\theta_j) \} \geq \eta d,$$

1064 which is (C.2).
 1065

1066 **From Bayes to Minimax.** By Yao’s minimax principle (Lemma C.3), the Bayes risk under the
 1067 uniform prior lower-bounds the minimax (worst-case) risk over Θ of any deterministic policy:
 1068

$$1069 \quad \inf_{\pi} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] \geq \mathbb{E}_\theta \mathbb{E}_\theta [\text{Ham}(\hat{s}, \text{sign}(\theta))] \geq \eta d,$$

1071 which is (C.3). This completes the proof. \square
 1072

1073 D PROOF OF THEOREM E.3

1075 **Lemma D.1.** If we follow Algorithm 2 to choose the action $\mathbf{a}_{1:T}$, and compute $\hat{\theta}_T$ and V_T , then it
 1076 holds that
 1077

$$1078 \quad V_T \succeq \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{|\text{supp}(\pi)|} \frac{1}{[\sigma_T^{(i)}]^2} \right] V(\pi).$$

1080
1081 *Proof.* Consider the action $\mathbf{a}_m \in \mathcal{A}$ such that $\pi(\mathbf{a}_m) > 0$ and $\sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)}$ is minimized. It is
1082 straightforward to see that
1083

$$V_T \succeq \sum_{\mathbf{a} \in \mathcal{A}} \left[\sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} \right] \pi(\mathbf{a}) \mathbf{a} \mathbf{a}^T = \sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} V(\pi). \quad (\text{D.1})$$

1086 Then it suffices to lower bound $\sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)}$.
1087

1088 We consider the arms in $\text{supp}(\pi) \setminus \{\mathbf{a}_m\}$. For the round $t(\mathbf{a})$ when the arm \mathbf{a} is pulled and
1089 $\sum_{\tau \in \mathcal{T}_t(\mathbf{a})} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a})} \geq \sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)}$, it is guaranteed by the selection rule that \mathbf{a} will not
1090 be pulled in the subsequent rounds. We denote by $\sigma(\mathbf{a})$ the last variance observed when pulling the
1091 arm \mathbf{a} . Then we have
1092

$$\begin{aligned} \sum_{\mathbf{a} \in \text{supp}(\pi) \setminus \{\mathbf{a}_m\}} \pi(\mathbf{a}) \sum_{\tau \in \mathcal{T}_T(\mathbf{a})} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a})} &\leq \sum_{\mathbf{a} \in \text{supp}(\pi) \setminus \{\mathbf{a}_m\}} \pi(\mathbf{a}) \left[\sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} + \frac{1}{[\sigma(\mathbf{a})]^2 \cdot \pi(\mathbf{a})} \right] \\ &\leq [1 - \pi(\mathbf{a}_m)] \sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} + \sum_{\mathbf{a} \in \text{supp}(\pi) \setminus \{\mathbf{a}_m\}} \frac{1}{[\sigma(\mathbf{a})]^2} \\ &\leq [1 - \pi(\mathbf{a}_m)] \sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} + \sum_{i=1}^{|\text{supp}(\pi)|-1} \frac{1}{[\sigma_T^{(i)}]^2}, \end{aligned} \quad (\text{D.2})$$

1102 where the first inequality is due to the definition of $\sigma(\mathbf{a})$ and the last inequality follows from the fact
1103 that $\{\sigma(\mathbf{a})\}_{\mathbf{a} \in \text{supp}(\pi) \setminus \{\mathbf{a}_m\}}$ are the variance signals observed at distinct rounds.
1104

1105 Rearranging (D.2), we obtain that
1106

$$\sum_{\mathbf{a} \in \text{supp}(\pi)} \pi(\mathbf{a}) \sum_{\tau \in \mathcal{T}_T(\mathbf{a})} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a})} \leq \sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} + \sum_{i=1}^{|\text{supp}(\pi)|-1} \frac{1}{[\sigma_T^{(i)}]^2}. \quad (\text{D.3})$$

1110 Note that LHS of (D.3) is exactly $\sum_{t=1}^T \frac{1}{\sigma_t^2}$, which further indicates that
1111

$$\sum_{\tau \in \mathcal{T}_T(\mathbf{a}_m)} \frac{1}{\sigma_{\tau}^2 \cdot \pi(\mathbf{a}_m)} \geq \sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{|\text{supp}(\pi)|-1} \frac{1}{[\sigma_T^{(i)}]^2}.$$

1115 As a result, we have
1116

$$V_T \succeq \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{|\text{supp}(\pi)|-1} \frac{1}{[\sigma_T^{(i)}]^2} \right] V(\pi)$$

1119 following (D.1). □

1120 **Lemma D.2.** In Algorithm 2, for any arm $\mathbf{a} \in \mathcal{A}$, with probability at least $1 - \delta$, we have
1121

$$\begin{aligned} |\langle \hat{\theta}_T - \theta^*, \mathbf{a} \rangle| &\leq \|\mathbf{a}\|_{V_T^{-1}} \sqrt{2 \log(2|\mathcal{A}|/\delta)} \\ &\leq \|\mathbf{a}\|_{V(\pi)^{-1}} \sqrt{2 \log(2|\mathcal{A}|/\delta) / \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{|\text{supp}(\pi)|-1} \frac{1}{[\sigma_T^{(i)}]^2} \right]}. \end{aligned}$$

1127 *Proof.* From the definition of $\hat{\theta}_T$ and V_T , it is straightforward to obtain the following inequality:
1128

$$\begin{aligned} |\langle \hat{\theta}_T - \theta^*, \mathbf{a} \rangle| &= |\langle V_T^{-1} \sum_{t=1}^T \sigma_t^{-2} r_t \mathbf{a}_t - V_T^{-1} V_T \cdot \theta^*, \mathbf{a} \rangle| \\ &= |\langle V_T^{-1} \sum_{t=1}^T \sigma_t^{-2} \mathbf{a}_t \eta_t, \mathbf{a} \rangle| \end{aligned}$$

$$1134 \quad 1135 \quad = \left| \sum_{t=1}^T \sigma_t^{-2} \eta_t \langle \mathbf{a}_t, V_T^{-1} \mathbf{a} \rangle \right|. \quad (D.4)$$

1137

1138 Applying Lemma D.3 to (D.4), we have that with probability at least $1 - \delta$,

$$1139 \quad 1140 \quad |\langle \hat{\theta}_T - \theta^*, \mathbf{a} \rangle| \leq \sqrt{2 \sum_{t=1}^T \sigma_t^{-2} \langle \mathbf{a}_t, V_T^{-1} \mathbf{a} \rangle^2 \log(2/\delta)} \\ 1141 \quad 1142 \quad \leq \sqrt{2 \sum_{t=1}^T [\sigma_t^{-1} \mathbf{a}_t^\top V_T^{-1} \mathbf{a}]^2 \log(2/\delta)} \\ 1143 \quad 1144 \quad = \sqrt{2 \sum_{t=1}^T \mathbf{a}^\top V_T^{-1} \sigma_t^{-1} \mathbf{a}_t \sigma_t^{-1} \mathbf{a}_t^\top V_T^{-1} \mathbf{a} \log(2/\delta)} \\ 1145 \quad 1146 \quad \leq \sqrt{2 \mathbf{a}^\top V_T^{-1} \mathbf{a} \log(2/\delta)},$$

1151 where the last inequality is due to the fact that $\sum_{t=1}^T \sigma_t^{-2} \mathbf{a}_t \mathbf{a}_t^\top = V_T$.

1152 By Lemma D.1, we further have

$$1153 \quad 1154 \quad |\langle \hat{\theta}_T - \theta^*, \mathbf{a} \rangle| \leq \|\mathbf{a}\|_{V_T^{-1}} \sqrt{2 \log(2/\delta)} \\ 1155 \quad 1156 \quad \leq \|\mathbf{a}\|_{V(\pi)^{-1}} \sqrt{2 \log(2/\delta) / \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{|\text{supp}(\pi)|} \frac{1}{[\sigma_T^{(i)}]^2} \right]},$$

1157 from which the desired result follows by taking a union bound over all $\mathbf{a} \in \mathcal{A}$. \square

1158 **Lemma D.3** (Hoeffding's inequality). Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a filtration so that for all $i \in [n]$, x_i is \mathcal{G}_i -measurable, while $\mathbb{E}[x_i | \mathcal{G}_{i-1}] = 0$ and $x_i | \mathcal{G}_{i-1}$ is a σ_i -sub-Gaussian random variable. Then, for any $t > 0$, with probability at least $1 - \delta$, it holds that

$$1159 \quad 1160 \quad \sum_{i=1}^n x_i \leq \sqrt{2 \sum_{i=1}^n \sigma_i^2 \log(1/\delta)}.$$

1161 **Theorem D.4** (Simple Regret of Algorithm 2, restatement of Theorem 5.3). Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is compact and $\text{span}(\mathcal{A}) = \mathbb{R}^d$. If we follow Algorithm 2, then it holds that with probability at least $1 - \delta$,

$$1162 \quad 1163 \quad \langle \theta^*, \mathbf{a}^* \rangle - \langle \theta^*, \mathbf{a}_{T+1} \rangle \leq 2 \sqrt{d \log(|\mathcal{A}|/\delta) / \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{4d \log \log d + 16} \frac{1}{[\sigma_T^{(i)}]^2} \right]}.$$

1164 *Proof.* From the definition of \mathbf{a}_{T+1} , we have

$$1165 \quad 1166 \quad \langle \theta^*, \mathbf{a}^* \rangle - \langle \theta^*, \mathbf{a}_{T+1} \rangle = \langle \hat{\theta}_T - \theta^*, \mathbf{a}_{T+1} \rangle + \langle \theta^* - \hat{\theta}_T, \mathbf{a}^* \rangle + \langle \hat{\theta}_T, \mathbf{a}^* - \mathbf{a}_{T+1} \rangle \\ 1167 \quad 1168 \quad \leq |\langle \hat{\theta}_T - \theta^*, \mathbf{a}_{T+1} \rangle| + |\langle \hat{\theta}_T - \theta^*, \mathbf{a}^* \rangle| \\ 1169 \quad 1170 \quad \leq (\|\mathbf{a}_{T+1}\|_{V(\pi)^{-1}} + \|\mathbf{a}^*\|_{V(\pi)^{-1}}) \sqrt{2 \log(2|\mathcal{A}|/\delta) / \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{|\text{supp}(\pi)|} \frac{1}{[\sigma_T^{(i)}]^2} \right]} \\ 1171 \quad 1172 \quad \leq 2 \sqrt{d \log(2|\mathcal{A}|/\delta) / \left[\sum_{t=1}^T \frac{1}{\sigma_t^2} - \sum_{i=1}^{4d \log \log d + 16} \frac{1}{[\sigma_T^{(i)}]^2} \right]},$$

1173 where the first inequality follows from the definition of \mathbf{a}_{T+1} , the second inequality is due to Lemma D.2 and the last inequality is due to Theorem 5.2. \square

1188 **Algorithm 3** Variance-Aware Exploration with Elimination (VAEE for heavy-tailed noise)

1189 **Require:** $\mathcal{A} \subset \mathbb{R}^d, \delta$.

1190 1: Initialize $V_0 \leftarrow \lambda I_d, \hat{\theta}_0 \leftarrow 0, \mathcal{A}_1 \leftarrow \mathcal{A}$.

1191 2: **for** $t = 1, \dots, T$ **do**

1192 3: Pull the action $\mathbf{a}_t \leftarrow \max_{\mathbf{e} \in \mathcal{A}_t} \|\mathbf{e}\|_{V_{t-1}^{-1}}$.

1193 4: The agent receives the reward r_t and the variance σ_t .

1194 5: Calculate $V_t \leftarrow V_{t-1} + \sigma_t^{-2} \mathbf{a}_t \mathbf{a}_t^\top$.

1195 6: Calculate

1196
$$\hat{\theta}_t \leftarrow \underset{\|\theta\|_2 \leq 1}{\operatorname{argmin}} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{i=1}^t \ell_{\tau_i} \left(\frac{r_i - \langle \mathbf{a}_i, \theta \rangle}{\sigma_i} \right),$$

1197 where $\tau_i = \tau_0 \cdot \frac{\sqrt{1 + \sigma_i^{-2} \|\mathbf{a}_i\|_{V_{i-1}^{-1}}^2}}{\sigma_i^{-1} \|\mathbf{a}_i\|_{V_{i-1}^{-1}}}$.

1198 7: Set confidence set as follows

1199
$$\mathcal{C}_t \leftarrow \{\theta \mid \|\theta - \hat{\theta}_t\|_{V_t^{-1}}^2 \leq \beta_t\}.$$

1200 8: Eliminate low rewarding arms: $\mathcal{A}_{t+1} \leftarrow \{\mathbf{a} \in \mathcal{A}_t : \max_{\mathbf{e} \in \mathcal{A}_t} \min_{\theta \in \mathcal{C}_t} \langle \theta, \mathbf{e} \rangle \leq \max_{\theta \in \mathcal{C}_t} \langle \theta, \mathbf{a} \rangle\}$.

1201 9: **end for**

E EXTENSION TO HEAVY-TAILED NOISE

In this section, we extend our results to the setting where the noise is heavy-tailed. Specifically, we consider the following assumption on the noise.

Assumption E.1. For any round t ($t \geq 1$), the noise η_t satisfies that

$$\mathbb{E}[\eta_t | \mathbf{a}_{1:t}, \eta_{1:t-1}] = 0, \quad \mathbb{E}[\eta_t^2 | \mathbf{a}_{1:t}, \eta_{1:t-1}] \leq \sigma_t^2.$$

This assumption is more general than the sub-Gaussian assumption on the noise, which only requires the second moment of the noise to be bounded.

To handle the heavy-tailed noise, we consider the following adaptive pseudo-Huber regression estimator (Ruppert, 2004; Sun, 2021; Li & Sun, 2024):

$$\ell_\tau(x) := \tau \cdot (\sqrt{\tau^2 + x^2} - \tau), \quad (\text{E.1})$$

where $\tau > 0$ is a robustification parameter. The pseudo-Huber loss behaves like the squared loss when $|x|$ is small, and behaves like the absolute loss when $|x|$ is large, which is firstly applied by Li & Sun (2024) into the heteroscedastic linear bandit setting.

Lemma E.2 (Theorem 2.1, Li & Sun 2024). Let $\kappa = d \cdot \log(1 + T/d\sigma_{\min}^2)$. If we set $\tau_0 \geq \max\{\sqrt{2\kappa}, 2\sqrt{d}\}/\sqrt{\log(2T^2/\delta)}$, then with probability at least $1 - 4\delta$, it holds for all $t \in [T]$ that

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \beta_t := 32 \left(\frac{\kappa}{\tau_0} + \sqrt{\kappa \log \frac{2t^2}{\delta}} + \tau_0 \log \frac{2t^2}{\delta} \right) + 5\sqrt{\lambda}.$$

Theorem E.3 (Simple Regret of Algorithm 3). Consider the linear bandit problem with heavy-tailed noise satisfying Assumption E.1. If we set $\tau_0 = \Theta(\sqrt{d})$ and $\lambda = 1$ in Algorithm 3, then with probability at least $1 - 4\delta$, it holds that

$$\text{SR}(T) = \tilde{O}(\sqrt{d}) \cdot \min_{1 \leq k \leq T+1} \left\{ x = \sqrt{\frac{\iota(T) - k + 1}{\sum_{i=k}^T \frac{1}{[\sigma_T^{(i)}]^2}}} \mid x \in [\sigma_T^{(k-1)}, \sigma_T^{(k)}] \right\},$$

where $\iota(T) = 2d \log \left(\frac{d + \sum_{\tau \in [T]} \sigma_\tau^{-2}}{d} \right)$. Recall that $\{\sigma_T^{(i)}\}_{i=1}^T$ is the sorted sequence of $\{\sigma_t\}_{t=1}^T$ in the ascending order.

1242 *Proof.* The proof follows the same line as the proof of Theorem 4.1, with the only difference being
1243 the confidence radius β_t . By setting $\tau_0 = \Theta(\sqrt{d})$, we have $\beta_t = \tilde{O}(\sqrt{d})$. Following the same
1244 analysis as in Theorem 4.1, we can obtain the desired result. \square
1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295