NESTED MODEL SELECTION IN LINEAR CONTEXTUAL BANDITS VIA EXPLORATION AND COMPARISON

Anonymous authors
Paper under double-blind review

ABSTRACT

We study nested model selection in stochastic linear contextual bandits with an unknown intercept for each arm and an unknown contextual parameter. The contextual parameter may be zero, may lie in a low-dimensional subspace, or may require the full d-dimensional space, and the goal is to achieve regret comparable to an oracle that knows the smallest subspace containing this parameter. The main challenge is that arm selection depends on the observed contexts, which can bias the data used for model selection. We first show that the assumptions used in prior work do not prevent this issue, and construct an instance on which the prior method incurs linear regret. To address this challenge, we propose two algorithms, `SST-UCB` and `SSC-UCB`, that use uniform exploration to obtain reliable observations for estimating the contextual parameter. `SST-UCB` selects the smallest subspace that is compatible with the observations, while `SSC-UCB` further compares the rewards obtained by learners at different levels and can control regret without identifying the true subspace exactly. For K arms and horizon n, both algorithms retain the standard $\widetilde{\mathcal{O}}(\sqrt{Kn})$ regret when contexts do not affect rewards, while adapting their leading regret term to the dimension of the smallest valid subspace when contexts matter. We further prove lower bounds showing that the guarantee of `SSC-UCB` is optimal up to logarithmic factors when the dimension of the smallest valid subspace does not exceed the number of arms.

1 INTRODUCTION

Many sequential decision problems reveal contextual information before each action is taken. In applications such as recommendation, pricing, and treatment allocation, this information can determine which action is best. Contextual bandits (Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011) provide a widely studied framework for learning in such settings. However, not all contextual information is useful: contexts may be irrelevant, or only a small part of the feature space may affect rewards. Therefore, a learner that always uses the full contextual model may waste samples estimating unnecessary parameters, while a learner that ignores contexts may incur large regret when the optimal action depends on the observed contexts. This creates a *nested model selection* problem: the learner should adapt to the smallest valid contextual subspace while minimizing regret.

We study this problem in a stochastic linear contextual bandit with K arms and an *intercept* for each arm. At each round t, the learner observes a context vector $\boldsymbol{\alpha}_{i,t} \in \mathbb{R}^d$ for every arm i, chooses an arm A_t, and observes the reward only for the chosen arm. The reward of arm i is:

$$g_{i,t} = \mu_i + \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t} \rangle + \eta_{i,t}, \qquad i \in [K], \tag{1}$$

where the intercept μ_i is the baseline reward of arm i, $\boldsymbol{\theta}^*$ is the unknown contextual parameter, and $\eta_{i,t}$ is noise. This formulation naturally captures models with different amounts of contextual information. If $\boldsymbol{\theta}^* = 0$, contexts do not affect rewards, and the problem reduces to a standard stochastic multi-armed bandit (Auer et al., 2002). We refer to this case as the *simple model*. At the other extreme, the *full model* allows $\boldsymbol{\theta}^*$ to be any vector in $\mathbb{R}^d$. More generally, $\boldsymbol{\theta}^*$ may lie in one of several nested subspaces:

$$\{0\} = V_0 \subsetneq V_1 \subsetneq \cdots \subsetneq V_L = \mathbb{R}^d, \qquad d_\ell = \dim(V_\ell).$$

Throughout this paper, the term *model* refers to a restriction of $\boldsymbol{\theta}^*$ to one of these subspaces. We define the *true level* as the smallest index ℓ^* such that $\boldsymbol{\theta}^* \in V_{\ell^*}$. Importantly, neither ℓ^* nor the

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.

corresponding true subspace is known to the learner. The learner must decide how much contextual information is sufficient while simultaneously minimizing regret. Our goal is to learn and exploit rewards nearly as well as a learner that knows ℓ^* from the beginning.

Motivating Example: Consider a food delivery platform that recommends one of K restaurants to a user at each visit. Each restaurant i has a baseline appeal to the user, represented by μ_i, which may capture factors such as brand preference or location. At round t, the platform also observes a context vector $\boldsymbol{\alpha}_{i,t} \in \mathbb{R}^d$ describing the current menu of restaurant i, including features such as cuisine type, price, ingredients, and available dishes. The expected reward from recommending restaurant i can be modeled as $\mu_i + \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t} \rangle$, where $\boldsymbol{\theta}^*$ captures how much the user cares about different features. Different users may require different levels of contextual information. Some users consistently prefer the same restaurants regardless of the current menu, corresponding to $\boldsymbol{\theta}^* = 0$. For others, a small set of features, such as cuisine and price, may be sufficient to explain their choices, so that $\boldsymbol{\theta}^*$ lies in a low dimensional subspace. Other users may care about more detailed properties, such as ingredients or specific dishes, and therefore require a richer feature representation. This naturally gives a sequence of nested models, from using no contextual information, to using a small set of features, and finally to using the full set of features. The platform does not know in advance which level is sufficient for a given user and must learn it while making recommendations.

The binary version of this problem, which distinguishes the simple model from the full model, was first studied by Chatterji et al. (2020), who proposed the OSOM algorithm. This problem was later extended to model selection among linear bandits of different complexity by Ghosh et al. (2021a). These methods estimate the contextual parameter $\boldsymbol{\theta}^*$ to determine whether the simple model is sufficient, and otherwise use a richer contextual model. However, OSOM (Chatterji et al., 2020) has two distinct issues caused by the dependence between arm selection and the observed contexts.

- First, the design matrix argument in Lemma 7 of Chatterji et al. (2020) treats the selected contexts as retaining the marginal context distribution of each arm. This does not hold in general because the learner chooses an arm after observing the current contexts. Even when every arm has a well-conditioned context distribution, adaptive arm selection may favor contexts that provide little information in some directions. Proposition 1 gives a counterexample to this argument.
- Second, OSOM estimates each intercept μ_i using the average reward observed from arm i. The analysis in Lemma 6 of Chatterji et al. (2020) requires this average to concentration around μ_i. However, although the contextual term $\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t} \rangle$ has mean zero for each fixed arm i, it need not have mean zero conditional on that arm being selected. Consequently, the resulting intercept estimates can be systematically biased. Proposition 2 shows that this bias can cause an actual failure of OSOM: there exists a problem instance on which OSOM incurs linear regret.

These counterexamples show that observations collected by an adaptive policy cannot in general be used directly for reliable model selection. To address this issue, we use uniform exploration to collect observations so that arm selection is independent of the current contexts, and propose two algorithms, `SST-UCB` and `SSC-UCB`.

Our first algorithm, `SST-UCB`, uses the uniform observations to estimate $\boldsymbol{\theta}^*$ and selects the smallest subspace compatible with the data. It then exploits using a bandit algorithm at the selected level. When the selected model is too small, confidence information from the full model is also used to remove arms that cannot be optimal. As more uniform observations are collected, incorrect smaller subspaces can be ruled out, while the cost of uniform exploration remains controlled.

Accurately identifying the true subspace, however, is not always necessary for achieving low regret. For example, a smaller model may make nearly the same decisions as the true model and therefore incur little regret even if it is not exactly correct. This motivates our second algorithm, `SSC-UCB`, which maintains learners at several candidate levels and compares the rewards they obtain. A smaller model is eliminated only when its decisions are detectably worse, and may remain active as long as it performs competitively. In this way, the algorithm can control regret without waiting to identify the true subspace exactly.

Both algorithms achieve $\widetilde{\mathcal{O}}(\sqrt{Kn})$ regret when contexts do not affect rewards and have a leading term $\widetilde{\mathcal{O}}(\sqrt{(K + d_{\ell^*})n})$ when contexts matter. We prove matching lower bounds of $\Omega(\sqrt{Kn})$ and $\Omega(\sqrt{(K + d_{\ell^*})n})$, respectively. Moreover, when $d_{\ell^*} \leq K$, `SSC-UCB` achieves $\widetilde{\mathcal{O}}(\sqrt{(K + d_{\ell^*})n})$

regret, matching the lower bound up to logarithmic factors. Thus, SSC−UCB is near minimax optimal in this regime while adapting to the smallest valid contextual subspace.

Contributions. Our contributions are summarized as follows.

- We identify two flaws in the design and analysis of OSOM (Chatterji et al., 2020) and give two counterexamples. The first shows that well-conditioned marginal context distributions do not imply that the selected contexts remain well conditioned under adaptive arm selection. The second shows that the biased intercept estimates cause OSOM to incur linear regret.
- We extend OSOM's setting to nested model selection, and propose two algorithms, SST−UCB and SSC−UCB. Both algorithms retain the standard $\widetilde{\mathcal{O}}(\sqrt{Kn})$ regret when contexts are irrelevant and adapt their leading regret term to the dimension d_{ℓ^*} of the true subspace. SSC−UCB further compares the rewards obtained at different levels and provides a regret guarantee that does not depend on the subspace gap.
- We prove regret lower bounds of $\Omega(\sqrt{Kn})$ when contexts are irrelevant and $\Omega(\sqrt{(K+d_{\ell^*})n})$ when contexts matter. When $d_{\ell^*} \leq K$, the latter bound matches the corresponding guarantee of SSC−UCB up to logarithmic factors.
- We conduct synthetic experiments across different true model levels. The results are consistent with our theoretical guarantees: SST−UCB and SSC−UCB retain low regret when contexts are irrelevant and adapt when contextual information is needed, while SSC−UCB consistently improves over SST−UCB in the contextual settings.

2 Related Work

Stochastic and linear contextual bandits. Classical stochastic multi-armed bandits study exploration and exploitation when each arm has an unknown fixed mean (Lai & Robbins, 1985; Auer et al., 2002; Lattimore & Szepesvári, 2020). Linear contextual bandits extend this objective to reward models with covariates, and algorithms such as SupLinUCB (Chu et al., 2011) and OFUL (Dani et al., 2008; Abbasi-Yadkori et al., 2011) control regret when the feature dimension is fixed. Our setting combines these two structures: the reward contains arm-specific intercepts, so the simple model is an ordinary stochastic bandit, while the full model introduces an unknown contextual parameter, and the learner needs to infer the underlying model while minimizing regret under adaptive data collection. This perspective is related to recent work on the tension between regret minimization and inference. For example, Banerjee et al. (2023) study spectral exploration in low-regret linear bandit trajectories with rich action sets. Simchi-Levi & Wang (2023) quantify the regret-inference trade-off in multi-armed bandit experimental design. Collaborative and personalized linear bandits (Ghosh et al., 2021b) use similarity, clustering, or shared structure across multiple users or agents to transfer information and improve individual learning. Safety-constrained linear bandits (Amani et al., 2019) impose constraints on the actions that can be selected and aim to minimize regret while ensuring that chosen actions remain safe with high probability.

Model selection in contextual and linear bandits. Classical stochastic multi-armed bandits and linear contextual bandits provide regret guarantees when the correct model is known in advance (Auer et al., 2002; Chu et al., 2011; Abbasi-Yadkori et al., 2011). Model selection asks whether similar guarantees can be achieved without knowing the appropriate underlying model. OSOM (Chatterji et al., 2020) is the closest predecessor of our work. It studies model selection between a stochastic multi-armed bandit and a linear contextual bandit with an intercept for each arm. Ghosh et al. (2021a) further extend this setting to models of different complexity using a related analysis. Our work revisits this line of work and shows that the analysis of OSOM does not follow from its stated assumptions because arm selection depends on the current contexts. More general forms of adaptation have also been studied. Foster et al. (2019) consider nested linear classes in stochastic contextual bandits and obtain regret that adapts to the dimension of the smallest realizable class. Muthukumar & Krishnamurthy (2022) revisit the problem of distinguishing a stochastic multi-armed bandit from a linear contextual bandit. They develop algorithms with adaptive exploration that do not require feature diversity in their general guarantees, and also give an extension to nested linear contextual models under additional assumptions. Ghosh et al. (2024) study model selection for general stochastic contextual bandits using successive refinement, with specialized results for linear

contextual bandits. Moradipari et al. (2022) considers model selection among multiple feature maps or parameter models on stochastic linear bandits.

Exploration and statistical inference. Exploration is useful not only for regret minimization but also for obtaining observations that support reliable inference. Forced exploration has been used in stochastic and contextual bandits (Auer et al., 2002; Langford & Zhang, 2007; Goldenshluger & Zeevi, 2013; Bastani & Bayati, 2020), and Muthukumar & Krishnamurthy (2022) explicitly design exploration for model selection in linear contextual bandits. Related work studies how low-regret actions affect statistical inference in linear bandits (Banerjee et al., 2023) and the trade-off between regret and inference in bandit experiments (Simchi-Levi & Wang, 2023). In our setting, by using uniform exploration, the selected arm is independent of the current contexts, which removes the selection bias that invalidates the analysis of prior work. These exploration rounds are interleaved with exploitation, allowing the learner to obtain reliable observations for model selection while continuing to minimize regret.

3 Problem Setup and Goals

In this section, we introduce the problem setup, which is largely adapted from Chatterji et al. (2020). There are K arms, and n denotes a time horizon unknown to the learner. At each round t, the learner observes context vectors $\boldsymbol{\alpha}_{i,t} \in \mathbb{R}^d$ for every arm $i \in [K]$. The learner then selects an arm $A_t \in [K]$ and observes the reward $g_{A_t,t}$, which follows Equation (1). We assume that $\mu_i \in [-1, 1]$, $\|\boldsymbol{\theta}^*\| \leq 1$, $\|\boldsymbol{\alpha}_{i,t}\| \leq 1$, and reward noises $\eta_{i,t}$ are independent zero-mean σ-sub-Gaussian random variables, independent of the context vectors. As in prior work, we also assume that the marginal context distribution is well conditioned.

Assumption 1 (Well-conditioned marginal context distribution)**.** For every round t, the context vectors $\{\boldsymbol{\alpha}_{i,t}\}_{i=1}^K$ are independent of the past and mutually independent across arms. Moreover, $\mathbb{E}[\boldsymbol{\alpha}_{i,t}] = 0$ and $\mathbb{E}[\boldsymbol{\alpha}_{i,t}\boldsymbol{\alpha}_{i,t}^\top] = \boldsymbol{\Sigma}_c \succeq \rho_{\min} \boldsymbol{I}_d$ for some $\rho_{\min} > 0$.

Note that since one goal of this work is to address the issues in Chatterji et al. (2020) under its original setting, we deliberately retain its assumptions above for a fair comparison. The next definition extends the binary model selection problem of Chatterji et al. (2020) to multiple nested levels.

Definition 1 (Model levels)**.** Let

$$\{0\} = V_0 \subsetneq V_1 \subsetneq \cdots \subsetneq V_L = \mathbb{R}^d$$

be known strictly nested subspaces with $d_\ell = \dim(V_\ell)$. Model level ℓ consists of all problem instances satisfying Equation (1) whose contextual parameter lies in V_ℓ. Level 0 is the simple model because $V_0 = \{0\}$ and contexts do not affect rewards. Level L is the full model because $V_L = \mathbb{R}^d$.

For the nested model selection, let $\boldsymbol{U}_\ell \in \mathbb{R}^{d \times d_\ell}$ be an orthonormal basis for V_ℓ and let $\boldsymbol{\Pi}_\ell = \boldsymbol{U}_\ell \boldsymbol{U}_\ell^\top$. For level 0, we use the convention $\boldsymbol{\Pi}_0 = 0$. The *true level* and *subspace gap* are defined as

$$\ell^* = \min\{\ell : \boldsymbol{\theta}^* \in V_\ell\}, \qquad \gamma_\ell = \|(\boldsymbol{I} - \boldsymbol{\Pi}_\ell)\boldsymbol{\theta}^*\|.$$

Thus the true level $\ell^* = 0$ iff $\boldsymbol{\theta}^* = 0$, and $\gamma_{\ell^*} = 0$. The true level ℓ^* is *unknown* to the learner.

For any true level ℓ^*, we define the regret R_n with respect to the optimal arm a_t^* as

$$R_n = \sum_{t=1}^{n} \left[\mu_{a_t^*} + \left\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{a_t^*,t} \right\rangle - \mu_{A_t} - \left\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{A_t,t} \right\rangle\right], \qquad a_t^* \in \arg\max_i \left\{\mu_i + \left\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t} \right\rangle\right\}.$$

When $\ell^* = 0$, we have $\boldsymbol{\theta}^* = 0$, so this reduces to the standard stochastic bandit regret

$$R_n = n\mu^* - \sum_{t=1}^{n} \mu_{A_t}, \qquad \mu^* = \max_i \mu_i.$$

For this case, we also define the arm gap $\Delta_i = \mu^* - \mu_i$.

The goal is to achieve regret nearly matching that of an oracle learner that knows the true level ℓ^*. Specifically, at level $\ell^* = 0$, this oracle reduces to a standard multi-armed bandit learner, such as UCB (Auer et al., 2002). When $\ell^* \geq 1$, it can run a linear contextual bandit algorithm, such as SupLinUCB (Chu et al., 2011), on the true subspace V_{ℓ^*}. Accordingly, our goal is to adapt to the unknown ℓ^* so that the regret depends on the intrinsic dimension d_{ℓ^*}, rather than the full dimension d, while retaining the standard stochastic bandit rate when $\ell^* = 0$.

4 WHY A NEW SOLUTION IS NEEDED

In this section, we explain why a new algorithmic design is needed by providing two counterexamples to OSOM (Chatterji et al., 2020). The first (Proposition 1) shows that its design matrix argument does not follow from the stated assumptions, while the second (Proposition 2) shows that OSOM's intercept estimation can fail under adaptive arm selection.

To identify the true level ℓ^*, the learner must estimate the contextual parameter $\boldsymbol{\theta}^*$. To this end, OSOM's analysis uses the identity (Lemma 7 of Chatterji et al. (2020))

$$\mathbb{E}\left[\boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top \mid \mathcal{F}_{t-1}\right] = \boldsymbol{\Sigma}_c, \quad \text{where } \mathcal{F}_{t-1} \text{ is the history before round } t. \tag{2}$$

based on the fact that every arm has marginal covariance $\boldsymbol{\Sigma}_c$ (Assumption 1). This step is invalid because A_t is chosen after observing the contexts, so the event $A_t = i$ can depend on the value of $\boldsymbol{\alpha}_{i,t}$. Consequently, the conditional distribution of $\boldsymbol{\alpha}_{i,t}$ given $A_t = i$ need not equal its marginal distribution, and the selected context need not retain covariance $\boldsymbol{\Sigma}_c$. Proposition 1 gives a counterexample to this argument and the proof is given in Appendix A.

Proposition 1. For every $K \geq 3$, there is a stochastic context distribution satisfying Assumption 1 with $\rho_{\min} = 1/2$ and an adaptive arm selection rule, such that

$$\lambda_{\min}\big(\mathbb{E}[\boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top]\big) = 2^{-K} < \rho_{\min}/2.$$

This contradicts the identity (Equation 2) used in the proof of OSOM's Lemma 7 (Chatterji et al., 2020). Consequently, the argument underlying OSOM is invalid under its stated assumptions.

The second issue of OSOM is in estimating the arm intercepts μ_i. It estimates μ_i by the average reward observed when arm i is selected, and its analysis treats the contextual term $\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle$ as zero mean because $\mathbb{E}[\boldsymbol{\alpha}_{i,t}] = 0$. However, the event $A_t = i$ depends on the current contexts, so $\mathbb{E}\left[\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle \mid A_t = i\right]$ need not be zero. Consequently, the empirical reward average of arm i can be systematically biased and need not concentrate around μ_i. Proposition 2 shows that this bias can cause OSOM to incur linear regret with constant probability. The proof is in Appendix B.

Proposition 2. There exists a problem instance satisfying OSOM's original assumptions such that for every time horizon $n \geq 2$, OSOM incurs linear regret $R_n = \Omega(n)$ with probability at least 1/2.

In this work, we address these issues through controlled uniform exploration. The key idea is to collect a set of samples in which the selected arm is independent of the current contexts, thereby removing the bias described above and providing a well-conditioned design for model testing.

5 ALGORITHMS

In this section, we introduce our algorithm design. We first rewrite the reward model in Equation (1) as a standard linear contextual bandit (Chu et al., 2011; Abbasi-Yadkori et al., 2011). Define the augmented feature $\widetilde{\boldsymbol{x}}_{i,t} = (\boldsymbol{e}_i^\top, \boldsymbol{\alpha}_{i,t}^\top)^\top \in \mathbb{R}^{K+d}$, where $\boldsymbol{e}_i \in \mathbb{R}^K$ is the i-th standard basis vector. Then Equation (1) can be written as $g_{i,t} = \widetilde{\boldsymbol{x}}_{i,t}^\top \widetilde{\boldsymbol{\theta}}^* + \eta_{i,t}$, with $\widetilde{\boldsymbol{\theta}}^* = (\mu_1, \ldots, \mu_K, \boldsymbol{\theta}^{*\top})^\top \in \mathbb{R}^{K+d}$. Running SupLinUCB (Chu et al., 2011) on these $(K+d)$-dimensional features gives the full-dimensional contextual regret bound $R_n = \widetilde{\mathcal{O}}(\sqrt{(K+d)n})$. This regret bound is valid, but it does not exploit the possibility that $\boldsymbol{\theta}^*$ may lie in a lower-dimensional subspace V_{ℓ^*} or even $\boldsymbol{\theta}^* = 0$.

For each level $\ell > 0$, define the projected augmented feature $\widetilde{\boldsymbol{x}}_{i,t}^{(\ell)} = (\boldsymbol{e}_i^\top, (\boldsymbol{U}_\ell^\top \boldsymbol{\alpha}_{i,t})^\top)^\top \in \mathbb{R}^{K+d_\ell}$ and set $\widetilde{\boldsymbol{x}}_{i,t}^{(0)} = \boldsymbol{e}_i$. At the true level ℓ^*, the reward can therefore be written exactly using the $(K + d_{\ell^*})$-dimensional projected features.

5.1 SST-UCB: SUBSPACE SELECTION VIA TESTING

Our first algorithm, SST-UCB (Subspace Selection via Testing), selects the smallest subspace that is consistent with a confidence set for the contextual parameter. As shown in Algorithm 1, SST-UCB proceeds in phases, each consisting of a uniform exploration block (Lines 5 to 8) followed by an exploitation block (Lines 14 to 23). It first collects E_s uniform observations, then freezes the estimate

and constructs its confidence set (Lines 9 to 10). It then selects the smallest subspace consistent with the confidence set (Line 11). At each exploitation round, it removes arms that cannot be optimal (Line 16). Among the remaining arms, the algorithm uses UCB (Auer et al., 2002) when the selected level is zero and SupLinUCB (Chu et al., 2011) at a positive level (Lines 18 and 20). All observed rewards are used to update the estimators for subsequent phases (Lines 7 and 22).

Algorithm 1: `SST-UCB`: Subspace Selection via Testing

Input: number of arms K, subspaces $V_0, \dots, V_L$, noise parameter σ, confidence δ

1 Initialize design matrix $\widetilde{\boldsymbol{V}} \leftarrow \lambda \boldsymbol{I}_{K+d}$ and vector $\boldsymbol{b} \leftarrow 0$, where $\lambda \leftarrow 1/(K+1)$
2 Initialize $T_i \leftarrow 0$, $S_i \leftarrow 0$ for all arms $i \in [K]$, and round counter $t \leftarrow 1$
3 **for** *phases* $s = 1, 2, \dots$ **do**
4 $\quad E_s \leftarrow K2^{s-1}, \;\; H_s \leftarrow K4^{s-1}, \;\; \varepsilon_s \leftarrow \delta/(4s(s+1))$
5 $\quad$ **for** $\tau = 1, \dots, E_s$ **do**
6 $\quad\quad$ Draw A_t uniformly from $[K]$, and observe $g_{A_t,t}$
7 $\quad\quad \widetilde{\boldsymbol{V}} \leftarrow \widetilde{\boldsymbol{V}} + \widetilde{\boldsymbol{x}}_{A_t,t}\widetilde{\boldsymbol{x}}_{A_t,t}^\top, \;\; \boldsymbol{b} \leftarrow \boldsymbol{b} + \widetilde{\boldsymbol{x}}_{A_t,t} g_{A_t,t}, \;\; T_{A_t} \leftarrow T_{A_t} + 1, \;\; S_{A_t} \leftarrow S_{A_t} + g_{A_t,t}$
8 $\quad\quad t \leftarrow t + 1$
9 $\quad \widetilde{\boldsymbol{V}}_s \leftarrow \widetilde{\boldsymbol{V}}, \;\; \widehat{\widetilde{\boldsymbol{\theta}}}_s \leftarrow \widetilde{\boldsymbol{V}}_s^{-1}\boldsymbol{b}, \;\; \beta_s \leftarrow 1 + \sigma\sqrt{2\log(4/\delta) + \log\frac{\det(\widetilde{\boldsymbol{V}}_s)}{\lambda^{K+d}}}$
10 $\quad \mathcal{C}_s \leftarrow \left\{\boldsymbol{v} \in \mathbb{R}^{K+d} : \|\boldsymbol{v} - \widehat{\widetilde{\boldsymbol{\theta}}}_s\|_{\widetilde{\boldsymbol{V}}_s} \le \beta_s\right\}$
11 $\quad \widehat{\ell}_s \leftarrow \min\{\ell \in \{0, \dots, L\} : \mathcal{C}_s \cap (\mathbb{R}^K \times V_\ell) \neq \varnothing\}$
12 $\quad$ **if** $\widehat{\ell}_s > 0$ **then**
13 $\quad\quad$ Start a fresh SupLinUCB with horizon H_s, confidence ε_s, regularization λ, and noise parameter σ
14 $\quad$ **for** $\tau = 1, \dots, H_s$ **do**
15 $\quad\quad$ Observe contexts
16 $\quad\quad \mathcal{A}_{s,t} \leftarrow \left\{i \in [K] : \widetilde{\boldsymbol{x}}_{i,t}^\top \widehat{\widetilde{\boldsymbol{\theta}}}_s + \beta_s \|\widetilde{\boldsymbol{x}}_{i,t}\|_{\widetilde{\boldsymbol{V}}_s^{-1}} \ge \max_{j\in[K]}\left(\widetilde{\boldsymbol{x}}_{j,t}^\top \widehat{\widetilde{\boldsymbol{\theta}}}_s - \beta_s \|\widetilde{\boldsymbol{x}}_{j,t}\|_{\widetilde{\boldsymbol{V}}_s^{-1}}\right)\right\}$
17 $\quad\quad$ **if** $\widehat{\ell}_s = 0$ **then**
18 $\quad\quad\quad$ Select $A_t \in \arg\max_{i\in\mathcal{A}_{s,t}} U_i$, where $U_i = +\infty$ if $T_i = 0$, and otherwise $U_i = S_i/T_i + \sigma\sqrt{2\log(8K(T_i+1)^2/\delta)/T_i}$
19 $\quad\quad$ **else**
20 $\quad\quad\quad$ Select $A_t \in \mathcal{A}_{s,t}$ by continuing SupLinUCB on the $(K + d_{\widehat{\ell}_s})$-dimensional features
21 $\quad\quad$ Observe $g_{A_t,t}$ and update SupLinUCB if $\widehat{\ell}_s > 0$
22 $\quad\quad \widetilde{\boldsymbol{V}} \leftarrow \widetilde{\boldsymbol{V}} + \widetilde{\boldsymbol{x}}_{A_t,t}\widetilde{\boldsymbol{x}}_{A_t,t}^\top, \boldsymbol{b} \leftarrow \boldsymbol{b} + \widetilde{\boldsymbol{x}}_{A_t,t} g_{A_t,t}, T_{A_t} \leftarrow T_{A_t} + 1, S_{A_t} \leftarrow S_{A_t} + g_{A_t,t}$
23 $\quad\quad t \leftarrow t + 1$

Theorem 3 gives the regret guarantee for `SST-UCB`. Its proof is given in Appendix C and the notations A_n and Λ_n are defined in Equation (9).

Theorem 3 (Regret guarantee of `SST-UCB`)**.** *Under Assumption 1, `SST-UCB` satisfies the following bounds for all time horizon $n \ge 1$ with probability at least $1 - \delta$.*

If $\ell^* = 0$, then

$$R_n \le \left(10 + 4\sigma\sqrt{2\log\frac{8K(n+1)^2}{\delta}}\right)\sqrt{Kn} = \widetilde{\mathcal{O}}\left(\sqrt{Kn}\right). \tag{3}$$

If $\ell^* \ge 1$, then

$$R_n \le 16\sqrt{Kn} + C(1+\sigma)\Lambda_n^2\sqrt{(K+d_{\ell^*})n} + \frac{9A_n^2}{K\gamma_{\ell^*-1}^3} = \widetilde{\mathcal{O}}\left(\sqrt{(K+d_{\ell^*})n} + \frac{(K+d)^2}{K\widetilde{\rho}^2\gamma_{\ell^*-1}^3}\right), \tag{4}$$

where C is a universal constant.

Remark 1. For $\ell^* = 0$, Theorem 3 recovers the standard near-optimal UCB regret $\widetilde{\mathcal{O}}(\sqrt{Kn})$. For $\ell^* \ge 1$, its leading term $\widetilde{\mathcal{O}}(\sqrt{(K+d_{\ell^*})n})$ matches the near-optimal SupLinUCB regret.

5.2 SSC−UCB: SUBSPACE SELECTION VIA COMPARISON

SST−UCB (Algorithm 1) chooses the smallest model consistent with the confidence set. This mechanism only checks whether a model is consistent with the observations, but does not directly measure the quality of the decisions made using that model. A smaller model that passes the test may therefore be used throughout a long exploitation block, even when it leads to poor arm choices. Conversely, when the subspace gap is small, identifying the true model may require many samples even though using a smaller model incurs little regret.

These observations motivate SSC−UCB (Subspace Selection via Comparison). Rather than relying solely on parameter estimation to identify the true subspace, SSC−UCB compares the rewards obtained by learners operating at different candidate levels. A candidate is eliminated when its average reward is sufficiently lower than that of another candidate, after accounting for reward noise and the regret of the learners themselves. Thus, an underfitted model can be removed directly when it leads to worse decisions, but it may remain active as long as its decisions are competitive. This allows SSC−UCB to control regret without first identifying the true subspace exactly.

Algorithm 2: SSC−UCB: Subspace Selection via Comparison

Input: number of arms K, subspaces $V_0, \dots, V_L$, noise parameter $\sigma \geq 0$, confidence δ

1 Set $\lambda \leftarrow 1/(K+1)$ and round counter $t \leftarrow 1$
2 **for** *phases* $s = 1, 2, \dots$ **do**
3 $E_s \leftarrow K2^{s-1}$, $H_s \leftarrow K4^{s-1}$, $\zeta_s \leftarrow \delta/(16H_s^2)$
4 $\mathcal{J}_s \leftarrow \{\max\{j \in \{0, \dots, L\} : K + d_j \leq 2^q K\} : q = 0, \dots, s-1\}$
5 $\boldsymbol{V} \leftarrow \lambda \boldsymbol{I}_{K+d}$, $\boldsymbol{b} \leftarrow \boldsymbol{0}$
6 **for** $\tau = 1, \dots, E_s$ **do**
7 Draw A_t uniformly from $[K]$ and observe $g_{A_t,t}$
8 $\boldsymbol{V} \leftarrow \boldsymbol{V} + \widetilde{\boldsymbol{x}}_{A_t,t}\widetilde{\boldsymbol{x}}_{A_t,t}^\top$, $\boldsymbol{b} \leftarrow \boldsymbol{b} + \widetilde{\boldsymbol{x}}_{A_t,t} g_{A_t,t}$, $t \leftarrow t+1$
9 Let $\widehat{\boldsymbol{\theta}}_s$ be the last d coordinates of $\boldsymbol{V}^{-1}\boldsymbol{b}$
10 $r_s \leftarrow \left(1 + \sigma\sqrt{2\log(8E_s/\delta) + (K+d)\log\left(1 + \frac{2E_s(K+1)}{K+d}\right)}\right) / \sqrt{\lambda_{\min}(\boldsymbol{V})}$
11 $\mathcal{I} \leftarrow \{\max \mathcal{J}_s\} \cup \left\{j \in \mathcal{J}_s : \|(\boldsymbol{I} - \boldsymbol{\Pi}_j)\widehat{\boldsymbol{\theta}}_s\| \leq r_s\right\}$
12 For every $j \in \mathcal{I}$, start a fresh SupLinUCB learner at level j, horizon H_s, confidence ζ_s, regularization λ, and noise parameter σ
13 Set $N_j \leftarrow 0$, $S_j \leftarrow 0$ for every $j \in \mathcal{I}$
14 **for** $\tau = 1, \dots, H_s$ **do**
15 Independently draw $j \in \mathcal{I}$ with probability $(2K + d_j)^{-1} / \sum_{i \in \mathcal{I}} (2K + d_i)^{-1}$
16 Observe contexts and use learner j on features $\widetilde{\boldsymbol{x}}_{i,t}^{(j)}$ to select $A_t \in [K]$
17 Observe $g_{A_t,t}$ and update learner j
18 $N_j \leftarrow N_j + 1$, $S_j \leftarrow S_j + g_{A_t,t}$, $t \leftarrow t+1$
19 Set $w_i \leftarrow 24(1+\sigma)\log(1/\zeta_s)\sqrt{(K+d_i)/N_i}$ and $c_i \leftarrow 4(1+\sigma)\sqrt{\log(1/\zeta_s)/N_i}$ for every $i \in \mathcal{I}$ with $N_i > 0$
20 $\mathcal{D} \leftarrow \{j \in \mathcal{I} : N_j > 0,\ S_j/N_j + w_j < \max_{i \in \mathcal{I}: N_i > 0}(S_i/N_i - c_i)\}$
21 $\mathcal{I} \leftarrow \mathcal{I} \setminus \mathcal{D}$
22 **if** $\mathcal{D} \neq \varnothing$ **then**
23 Reset the SupLinUCB learner at level j for every $j \in \mathcal{I}$ with the same parameters
24 $N_j \leftarrow 0$, $S_j \leftarrow 0$ for every $j \in \mathcal{I}$

As shown in Algorithm 2, SSC−UCB first forms a set of candidate levels (Line 4), then uses the current uniform exploration samples to estimate the parameter and removes candidates that are inconsistent with this estimate (Lines 7 to 11). The parameter test always retains the largest candidate, so the active set $\mathcal{I}$ is nonempty. During exploitation, SSC−UCB draws a candidate learner before observing contexts, with probabilities proportional to $(2K + d_j)^{-1}$, and gives feedback only to selected learner (Lines 15 to 17). Selecting the learner before observing the contexts avoids introducing additional selection bias, while assigning smaller probabilities to larger models limits their

cost. The algorithm then compares the average rewards obtained by the active learners and eliminates a candidate j when its upper comparison bound falls below another candidate's lower bound (Lines 19 to 21). After an elimination, the surviving learners and comparison statistics are restarted within the same exploitation block (Line 24).

The following theorem gives the regret guarantee for SSC−UCB. Its proof is given in Appendix D. The notations $A_n, \Lambda_n, m_n, \Phi_n$ are defined in Equations (9) and (16).

Theorem 4 (SSC−UCB)**.** *Under Assumption 1, there is a universal constant C such that SSC−UCB satisfies the following bounds for all time horizon $n \geq 1$ with probability at least $1-\delta$.*

$$R_n \leq C\Phi_n\left[\sqrt{m_n}(K+d_{\ell^*})\sqrt{n/K} + m_n^{3/2}\sqrt{(K+d_{\ell^*})n}\right] = \widetilde{\mathcal{O}}\left(\frac{K+d_{\ell^*}}{\sqrt{K}}\sqrt{n}\right). \tag{5}$$

If $\ell^* \geq 1$, it also satisfies

$$\begin{aligned} R_n &\leq Cm_n\Phi_n\sqrt{(K+d_{\ell^*})n} + C\Phi_n\left[\frac{\sqrt{m_n}(K+d_{\ell^*})}{K} + m_n^{3/2}\sqrt{\frac{K+d_{\ell^*}}{K}}\right]\frac{A_n}{\gamma^2_{\ell^*-1}} \\ &= \widetilde{\mathcal{O}}\left(\sqrt{(K+d_{\ell^*})n} + \frac{(K+d_{\ell^*})(K+d)}{K\widetilde{\rho}\gamma^2_{\ell^*-1}}\right). \end{aligned} \tag{6}$$

If $\ell^* = 0$, it also satisfies

$$R_n \leq Cm_n\Phi_n\sqrt{Kn} = \widetilde{\mathcal{O}}\left(\sqrt{Kn}\right) \tag{7}$$

Remark 2 (Adaptivity and optimality)**.** Theorem 4 shows that SSC−UCB retains the standard $\widetilde{\mathcal{O}}(\sqrt{Kn})$ regret when $\ell^* = 0$. For $\ell^* \geq 1$, the leading term $\widetilde{\mathcal{O}}\left(\sqrt{(K+d_{\ell^*})n}\right)$ in Equation (6) matches the SupLinUCB regret and adapts to the true dimension d_{ℓ^*}. Compared with SST−UCB, SSC−UCB improves the dependence on the subspace gap from $\gamma^{-3}_{\ell^*-1}$ to $\gamma^{-2}_{\ell^*-1}$. More importantly, SSC−UCB also provides a bound independent of the subspace gap in Equation (5). When $d_{\ell^*} \leq K$, this becomes $\widetilde{\mathcal{O}}(\sqrt{(K+d_{\ell^*})n})$, matching the lower bound in Theorem 5 up to logarithmic factors.

6 Lower Bounds

In this section, we establish regret lower bounds that characterize the unavoidable dependence on K and d_{ℓ^*}. In particular, we prove an $\Omega(\sqrt{Kn})$ lower bound for $\ell^* = 0$ and an $\Omega(\sqrt{(K+d_{\ell^*})n})$ lower bound for $\ell^* \geq 1$, showing that the guarantee of SSC−UCB in Equation (5) is optimal up to logarithmic factors when $d_{\ell^*} \leq K$. The proof is given in Appendix E.

Theorem 5 (Lower bounds)**.** *There exists a universal constant $c_{\mathrm{lb}} > 0$ such that the following statements hold. If $K \geq 2$ and $n \geq K$, then for every policy π, there exists an instance satisfying Assumption 1 with $\boldsymbol{\theta}^* = 0$ such that*

$$\mathbb{E}[R_n] \geq c_{\mathrm{lb}}\sqrt{Kn}.$$

For any $\ell^* \geq 1$ with $d_{\ell^*} \geq 2$, $K \geq 2$, and $n \geq \max\{K, d^2_{\ell^*}\}$, for every policy π, there exists an instance satisfying Assumption 1 with $\boldsymbol{\theta}^* \in V_{\ell^*} \setminus V_{\ell^*-1}$, such that

$$\mathbb{E}[R_n] \geq c_{\mathrm{lb}}\sqrt{(K+d_{\ell^*})n}.$$

Remark 3 (Interpretation of the lower bound)**.** The two parts of the lower bound capture different sources of difficulty. Learning the K intercepts $\mu_1, \ldots, \mu_K$ requires $\Omega(\sqrt{Kn})$ regret, while learning the contextual parameter $\boldsymbol{\theta}^*$ within the true subspace requires $\Omega(\sqrt{d_{\ell^*}n})$ regret. Together, they show that the dependence on $K + d_{\ell^*}$ is intrinsic to the problem rather than an artifact of the augmented feature representation. Combined with Equation (5), this establishes the optimality of SSC−UCB up to logarithmic factors when $d_{\ell^*} \leq K$.

7 Experiments

We conduct synthetic experiments to evaluate the performance of our algorithms and whether they adapt to the true model level. Figure 1 compares SST−UCB and SSC−UCB with OSOM (Chatterji

et al., 2020), UCB (Auer et al., 2002), and full-dimensional SupLinUCB (Chu et al., 2011). We use $K = 100$, $d = 20$, horizon $n = 16,000$, confidence $\delta = 0.05$, and consider three true dimensions $d_{\ell^*} \in \{0, 5, 20\}$. Contexts are generated from independent Rademacher coordinates and normalized to unit norm. Thus, they are zero mean and satisfy Assumption 1. Arm intercepts are equally spaced from -0.1 to 0.1, and the reward noise is Gaussian with standard deviation 0.05. For $d_{\ell^*} = 0$, we set $\boldsymbol{\theta}^* = 0$. For $d_{\ell^*} = 5$, we draw a standard Gaussian vector in $\mathbb{R}^5$, normalize it, and embed it in the first five coordinates of $\mathbb{R}^{20}$. For $d_{\ell^*} = 20$, we independently draw and normalize a standard Gaussian vector in $\mathbb{R}^{20}$. These choices correspond to the nested subspaces $V_0 = \{0\}$, $V_1 = \mathrm{span}\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_5\}$, $V_2 = \mathbb{R}^{20}$. We report mean cumulative regret over 10 independent seeds in Figure 1, with shaded regions showing one standard error. For each seed, all algorithms are evaluated on the same sequence of contexts and reward noises to reduce variation across comparisons.

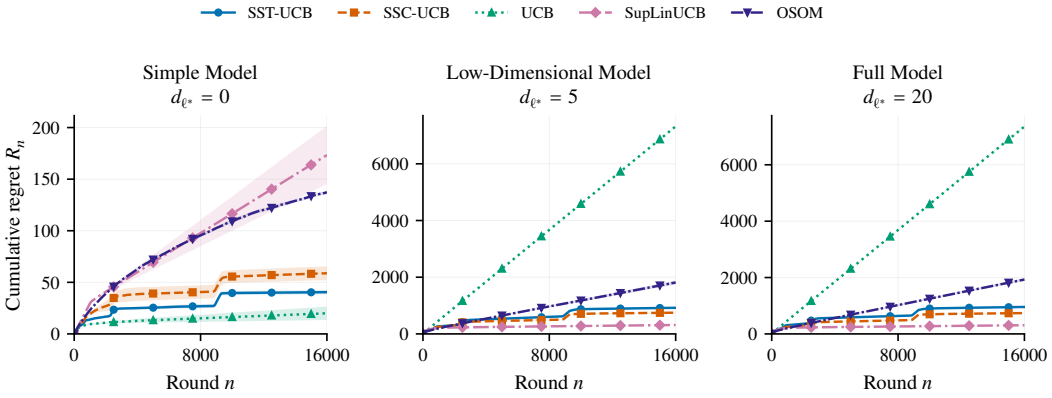


Figure 1: Comparison of cumulative regret over ten seeds.

The results are consistent with our theoretical guarantees. When $d_{\ell^*} = 0$, UCB achieves the lowest mean regret, while `SST-UCB` and `SSC-UCB` perform similarly and substantially better than OSOM and full-dimensional SupLinUCB. When $d_{\ell^*} = 5$ or 20, SupLinUCB achieves the lowest mean regret, followed by `SSC-UCB` and `SST-UCB`, whereas UCB incurs much larger regret because it ignores the contextual component. In particular, the results illustrate the intended adaptivity of our methods, i.e., they retain performance close to UCB when contexts are irrelevant, while automatically switching to contextual learning when contexts matter, without knowing the true level in advance. Across the contextual settings, `SSC-UCB` also consistently improves over `SST-UCB`, in line with its stronger regret guarantees.

8 Limitations

First, to isolate the issues in OSOM and enable a direct comparison, we retain its assumptions that contexts are independent across rounds and have well-conditioned marginal distributions. Extending our results beyond these assumptions remains open. Second, we assume that the nested subspaces $V_0, \ldots, V_L$ are known in advance and do not study how to learn such a hierarchy from data. Third, the guarantee of `SSC-UCB` in Equation (5) matches our lower bound up to logarithmic factors only when $d_{\ell^*} \leq K$. Closing the remaining gap when $d_{\ell^*} > K$ is left for future work.

9 Conclusion

In this paper, we studied nested model selection in linear contextual bandits, where the learner seeks to adapt to the smallest subspace needed to model the rewards while controlling regret. We first identified two issues in existing algorithm OSOM by giving two counterexamples. To address these issues, we proposed `SST-UCB` and `SSC-UCB`, which use uniform exploration to obtain reliable samples for inference. Both algorithms adapt their leading regret term to the unknown true dimension d_{ℓ^*}, and `SSC-UCB` further provides a guarantee independent of the subspace gap that matches our lower bound up to logarithmic factors when $d_{\ell^*} \leq K$. We also conducted experiments which validated these theoretical guarantees.

AI USE STATEMENT

In this work, we used generative AI tools for literature retrieval and analysis, language editing and polishing, drafting evaluation source code, and assistance in checking mathematical proofs. We have not used generative AI tools for generating research ideas, methodologies, or algorithms. We have reviewed all AI-assisted work. Literature retrieved or summarized with AI assistance was checked against the original sources, all mathematical arguments and proofs were independently examined by the authors for correctness, and all AI-assisted source code was manually reviewed by the authors. All research ideas and technical contributions in this work were developed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 2312–2320, 2011.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, May 2002.

Debangshu Banerjee, Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Exploration in linear bandits with rich action sets and its implications for inference. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8233–8262. PMLR, 25–27 Apr 2023.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

J. Bretagnolle and C. Huber. Estimation des densités : Risque minimax. In C. Dellacherie, P. A. Meyer, and M. Weil (eds.), *Séminaire de Probabilités XII*, pp. 342–363, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. ISBN 978-3-540-35856-5.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1844–1854. PMLR, 26–28 Aug 2020.

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

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

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Avishek Ghosh, Abishek Sankararaman, and Ramchandran Kannan. Problem-complexity adaptive model selection for stochastic linear bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1396–1404. PMLR, 13–15 Apr 2021a.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Collaborative learning and personalization in multi-agent stochastic linear bandits. *stat*, 1050:15, 2021b.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Model selection for generic contextual bandits. *IEEE Transactions on Information Theory*, 70(1):656–675, 2024.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

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*, WWW '10, pp. 661–670, 2010.

Ahmadreza Moradipari, Berkay Turan, Yasin Abbasi-Yadkori, Mahnoosh Alizadeh, and Mohammad Ghavamzadeh. Feature and parameter selection in stochastic linear bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15927–15958. PMLR, 17–23 Jul 2022.

Vidya K Muthukumar and Akshay Krishnamurthy. Universal and data-adaptive algorithms for model selection in linear contextual bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16197–16222. PMLR, 17–23 Jul 2022.

David Simchi-Levi and Chonghuan Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3086–3097. PMLR, 25–27 Apr 2023.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2011. ISSN 1615-3383.

A Proof of Proposition 1

Proposition 1. For every $K \geq 3$, there is a stochastic context distribution satisfying Assumption 1 with $\rho_{\min} = 1/2$ and an adaptive arm selection rule, such that

$$\lambda_{\min}\big(\mathbb{E}[\boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top]\big) = 2^{-K} < \rho_{\min}/2.$$

This contradicts the identity (Equation 2) used in the proof of OSOM's Lemma 7 (Chatterji et al., 2020). Consequently, the argument underlying OSOM is invalid under its stated assumptions.

Proof. Let $d = 2$. For every arm and round, draw $\boldsymbol{\alpha}_{i,t}$ uniformly from $\{\boldsymbol{e}_1, -\boldsymbol{e}_1, \boldsymbol{e}_2, -\boldsymbol{e}_2\}$, independently across arms and rounds. Then $\mathbb{E}[\boldsymbol{\alpha}_{i,t}] = 0$ and

$$\mathbb{E}[\boldsymbol{\alpha}_{i,t}\boldsymbol{\alpha}_{i,t}^\top] = \frac{1}{2}\boldsymbol{e}_1\boldsymbol{e}_1^\top + \frac{1}{2}\boldsymbol{e}_2\boldsymbol{e}_2^\top = \frac{1}{2}\boldsymbol{I}_2,$$

so Assumption 1 holds with $\rho_{\min} = 1/2$.

Consider the following context-dependent rule. At round t, if at least one arm has context on the first coordinate axis, namely $\boldsymbol{\alpha}_{i,t} \in \{\boldsymbol{e}_1, -\boldsymbol{e}_1\}$, select such an arm with the smallest index. If no such arm exists, select arm 1. Thus the selected context lies on the second coordinate axis if and only if all K arms lie on that axis. This event has probability 2^{-K}. Because this rule depends only on the current contexts and the current contexts are independent of the past, therefore, conditioned on the past history $\mathcal{F}_{t-1}$

$$\mathbb{E}[\boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top \mid \mathcal{F}_{t-1}] = (1 - 2^{-K})\boldsymbol{e}_1\boldsymbol{e}_1^\top + 2^{-K}\boldsymbol{e}_2\boldsymbol{e}_2^\top.$$

Averaging over the past gives the unconditional expectation. Its minimum eigenvalue is 2^{-K}, which is strictly less than $\rho_{\min}/2 = 1/4$ for every $K \geq 3$. Thus the identity used in the proof of OSOM's Lemma 7 does not follow from the marginal covariance assumption. □

B Proof of Proposition 2

Proposition 2. There exists a problem instance satisfying OSOM's original assumptions such that for every time horizon $n \geq 2$, OSOM incurs linear regret $R_n = \Omega(n)$ with probability at least 1/2.

Proof. Set $K = 2$, $d = 1$, $\mu_1 = 1/4$, and $\mu_2 = 0$. Let $\boldsymbol{e}_1$ be the unit vector in $\mathbb{R}^1$, and set $\boldsymbol{\theta}^* = \boldsymbol{e}_1$ and $\boldsymbol{\alpha}_{i,t} = x_{i,t}\boldsymbol{e}_1$, where the signs $x_{i,t}$ are independent across arms and rounds with $\Pr(x_{i,t} = 1) = \Pr(x_{i,t} = -1) = 1/2$. Take $\eta_{i,t} = 0$ and the sub-Gaussian parameter $\sigma = 1$. It is easy to verify that this setting satisfies all of OSOM's original assumptions. Then the rewards are

$$g_{1,t} = \frac{1}{4} + x_{1,t}, \qquad g_{2,t} = x_{2,t}.$$

We have $g_{1,t} \in \{-3/4, 5/4\}$ and $g_{2,t} \in \{-1, +1\}$. The optimal arm a_t^* is arm 2 when $(x_{1,t}, x_{2,t}) = (-1, +1)$, and is arm 1 otherwise. In particular, every suboptimal choice incurs regret at least $1/4$.

Consider Algorithm 1 (OSOM) of Chatterji et al. (2020) with time horizon $n \geq 2$ and confidence parameter $\delta \in (0, 1)$. Let $T_i(t) = \sum_{s=1}^t \mathbf{1}\{A_s = i\}$ and $\bar{g}_{i,t} = (\sum_{s=1}^t \mathbf{1}\{A_s = i\}g_{i,s})/T_i(t)$ be the arm count and reward average. For both the "simple" branch of OSOM and the "complex" branch when $x_{1,t} = x_{2,t}$, the arm selection rule (as in Equations (5) and (6) of Chatterji et al. (2020)) is:

$$A_t \in \arg\max_{i \in \{1,2\}} \widetilde{\mu}_{i,t-1}, \text{ where } \widetilde{\mu}_{i,t} = \bar{g}_{i,t} + b(T_i(t)),\ b(q) = \sqrt{\frac{1+q}{q^2}\left[1 + 2\log\left(\frac{2\sqrt{1+q}}{\delta}\right)\right]}. \quad (8)$$

The bonus term $b(q)$ is decreasing for $q \geq 1$, because

$$\frac{\mathrm{d}}{\mathrm{d}q}b(q)^2 = \frac{q - (q+2)\left[1 + 2\log\left(\frac{2\sqrt{1+q}}{\delta}\right)\right]}{q^3} < 0.$$

By the above setting, the optimal arm and reward satisfy

$$\Pr(a_t^* = 1) = \frac{3}{4}, \qquad \Pr(a_t^* = 2) = \frac{1}{4}, \qquad \mathbb{E}[g_{a_t^*,t}] = \frac{5/4 + 5/4 + 1 - 3/4}{4} = \frac{11}{16}.$$

OSOM initializes by selecting arm 1 in round 1 and arm 2 in round 2. These choices are both optimal with probability $(3/4)(1/4) = 3/16$. Hence, for time horizon $2 \le n < 256$,

$$\Pr(R_n > n/1024) \ge \Pr(R_2 \ge 1/4) = 1 - \frac{3}{16} > \frac{1}{2}.$$

It remains to consider $n \ge 256$. Set $m = \lceil 3n/4 \rceil$, so $n - m = \lfloor n/4 \rfloor \ge n/8$. We first show that the following four conditions hold simultaneously with probability at least $1/2$:

$$\sum_{s=1}^{n} \mathbb{1}\{a_s^* = 2\} \le \frac{n}{3}, \quad \sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2\} \ge \frac{3m}{16},$$
$$\sum_{s=1}^{m} g_{a_s^*,s} \le \frac{13m}{16}, \quad \sum_{t=m+1}^{n} \mathbb{1}\{x_{1,t} = x_{2,t}\} \ge \frac{n-m}{4}.$$

For the first and second conditions, since the variables $\mathbb{1}\{a_s^* = 2\}$ have mean $\frac{1}{4}$ and lie in $[0, 1]$, Hoeffding's inequality gives

$$\Pr\left(\sum_{s=1}^{n} \mathbb{1}\{a_s^* = 2\} > \frac{n}{3}\right) = \Pr\left(\sum_{s=1}^{n}\left(\mathbb{1}\{a_s^* = 2\} - \frac{1}{4}\right) > \frac{n}{12}\right)$$
$$\le \exp\left(-\frac{2(n/12)^2}{n}\right) = \exp\left(-\frac{n}{72}\right).$$

$$\Pr\left(\sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2\} < \frac{3m}{16}\right) = \Pr\left(\sum_{s=1}^{m}\left(\mathbb{1}\{a_s^* = 2\} - \frac{1}{4}\right) < -\frac{m}{16}\right)$$
$$\le \exp\left(-\frac{2(m/16)^2}{m}\right) = \exp\left(-\frac{m}{128}\right).$$

For the third condition, since the optimal rewards $g_{a_s^*,s}$ have mean $11/16$ and lie in $[-3/4, 5/4]$, Hoeffding's inequality yields

$$\Pr\left(\sum_{s=1}^{m} g_{a_s^*,s} > \frac{13m}{16}\right) = \Pr\left(\sum_{s=1}^{m}\left(g_{a_s^*,s} - \frac{11}{16}\right) > \frac{m}{8}\right)$$
$$\le \exp\left(-\frac{2(m/8)^2}{m \cdot 2^2}\right) = \exp\left(-\frac{m}{128}\right).$$

For the fourth condition, $x_{1,t} = x_{2,t}$ occurs for the two sign pairs $(+1, +1)$ and $(-1, -1)$, each of probability $1/4$. Thus the variables $\mathbb{1}\{x_{1,t} = x_{2,t}\}$ take values in $[0, 1]$ and have mean $1/2$. For the $n - m$ rounds after m, Hoeffding's inequality gives

$$\Pr\left(\sum_{t=m+1}^{n} \mathbb{1}\{x_{1,t} = x_{2,t}\} < \frac{n-m}{4}\right) = \Pr\left(\sum_{t=m+1}^{n}\left(\mathbb{1}\{x_{1,t} = x_{2,t}\} - \frac{1}{2}\right) < -\frac{n-m}{4}\right)$$
$$\le \exp\left(-\frac{2((n-m)/4)^2}{n-m}\right) = \exp\left(-\frac{n-m}{8}\right).$$

By the union bound, the probability of the four conditions fail to hold simultaneously is at most

$$\begin{aligned}
&\Pr(\text{at least one of the four conditions fails}) \\
&\le 2\exp(-m/128) + \exp(-n/72) + \exp(-(n-m)/8) \\
&\le 2\exp(-3n/512) + \exp(-n/72) + \exp(-n/64) \\
&\le 2e^{-3/2} + e^{-32/9} + e^{-4} < \frac{1}{2}.
\end{aligned}$$

Next, we prove that on the event that all the four conditions hold, we have $R_n > n/1024$. Suppose for a contradiction that $R_n \le n/1024$. Using the first two conditions and the monotonicity of $T_2(t)$, for $m \le t \le n$, we have

$$T_2(t) \ge T_2(m) = \sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2, A_s = 2\} + \sum_{s=1}^{m} \mathbb{1}\{a_s^* = 1, A_s = 2\}$$

$$
\begin{aligned}
&\geq \sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2, A_s = 2\} \\
&= \sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2\} - \sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2, A_s = 1\} \\
&\geq \frac{3m}{16} - 4R_n \geq \frac{3m}{16} - \frac{n}{256} \geq \frac{35n}{256} \geq \frac{n}{8},
\end{aligned}
$$

$$
\begin{aligned}
T_2(t) \leq T_2(n) &= \sum_{s=1}^{n} \mathbb{1}\{a_s^* = 2, A_s = 2\} + \sum_{s=1}^{n} \mathbb{1}\{a_s^* = 1, A_s = 2\} \\
&\leq \sum_{s=1}^{n} \mathbb{1}\{a_s^* = 2\} + \sum_{s=1}^{n} \mathbb{1}\{a_s^* = 1, A_s = 2\} \\
&\leq \frac{n}{3} + 4R_n \leq \frac{n}{3} + \frac{n}{256} < \frac{3n}{8} \leq \frac{m}{2} \leq \frac{t}{2},
\end{aligned}
$$

where we use the inequalities $R_n \geq \frac{1}{4}\sum_{s=1}^{m} \mathbb{1}\{a_s^* = 2,\ A_s = 1\}$ and $R_n \geq \frac{1}{4}\sum_{s=1}^{n} \mathbb{1}\{a_s^* = 1,\ A_s = 2\}$ since every suboptimal choice incurs at least $1/4$ regret.

On the other hand, arm 2 has reward $+1$ or -1, so $g_{2,s} = 1 - 2\mathbb{1}\{g_{2,s} = -1\}$, we have

$$
\begin{aligned}
\bar{g}_{2,t} &= \frac{1}{T_2(t)} \sum_{s=1}^{t} \mathbb{1}\{A_s = 2\}\, g_{2,s} \\
&= \frac{1}{T_2(t)} \sum_{s=1}^{t} \mathbb{1}\{A_s = 2\}\big(1 - 2\mathbb{1}\{g_{2,s} = -1\}\big) \\
&= \frac{1}{T_2(t)} \left(\sum_{s=1}^{t} \mathbb{1}\{A_s = 2\} - 2\sum_{s=1}^{t} \mathbb{1}\{A_s = 2,\ g_{2,s} = -1\} \right) \\
&= 1 - \frac{2}{T_2(t)} \sum_{s=1}^{t} \mathbb{1}\{A_s = 2,\ g_{2,s} = -1\}. \\
&\geq 1 - \frac{8R_n}{T_2(t)} \geq 1 - \frac{8(n/1024)}{n/8} = \frac{15}{16},
\end{aligned}
$$

where we use the fact that every choice with reward -1 is suboptimal.

Since the total reward of both arms cannot exceed the optimal total reward, every reward is at most $5/4$, and $m \geq 3n/4 \geq 3t/4$, we have

$$
\begin{aligned}
T_1(t)\bar{g}_{1,t} + T_2(t)\bar{g}_{2,t} = \sum_{s=1}^{t} g_{A_s,s} &\leq \sum_{s=1}^{m} g_{a_s^*,s} + \frac{5}{4}(t - m) \\
&\leq \frac{13m}{16} + \frac{5}{4}(t - m) = \frac{5t}{4} - \frac{7m}{16} \leq \frac{59t}{64} < \frac{15t}{16}.
\end{aligned}
$$

Since $T_1(t) + T_2(t) = t$, this forces $\bar{g}_{1,t} < 15/16$ because $\bar{g}_{2,t} \geq 15/16$ and both $T_1(t)$ and $T_2(t)$ are positive. Also, since $T_2(t) < \frac{t}{2}$, we have $T_1(t) > T_2(t)$. So the following holds since b is nonincreasing,

$$
\widetilde{\mu}_{1,t} < \frac{15}{16} + b(T_1(t)) \leq \frac{15}{16} + b(T_2(t)) \leq \widetilde{\mu}_{2,t}.
$$

So for every round in $\{m+1, \ldots, n\}$, if $x_{1,t} = x_{2,t}$, Equation (8) forces arm 2 to be selected, and each of these choices incurs regret $1/4$. We have shown that there are at least $(n-m)/4$ such rounds, giving $R_n \geq (n-m)/16 = \lfloor n/4 \rfloor /16 > n/1024$, a contradiction. Thus, for every $n \geq 2$,

$$
\Pr(R_n > n/1024) \geq \frac{1}{2}. \qquad \square
$$

C Proof of Theorem 3

We first define the following quantities used in the analysis:

$$\overline{\beta}_n = 1 + \sigma\sqrt{2\log(4/\delta) + (K+d)\log\left(1 + \frac{2n(K+1)}{K+d}\right)}, \qquad \Lambda_n = \log\frac{64K(n+1)^2}{\delta}$$
$$A_n = \frac{16}{\widetilde{\rho}}\max\left\{\overline{\beta}_n^2, \log\frac{4(K+d)n(n+1)}{\delta}\right\} = \widetilde{\mathcal{O}}((K+d)/\widetilde{\rho}). \tag{9}$$

Lemma 1 (Uniform exploration design)**.** *Suppose Assumption 1 holds. Consider any block of T exploration rounds in which, at each round t, the learner selects $A_t \sim \mathrm{Unif}([K])$ independently of the past and of the current contexts $\{\boldsymbol{\alpha}_{i,t}\}_{i=1}^K$. Let*

$$\widetilde{\boldsymbol{x}}_{i,t} = (\boldsymbol{e}_i^\top, \boldsymbol{\alpha}_{i,t}^\top)^\top \in \mathbb{R}^{K+d}, \qquad \widetilde{\rho} = \min\{1/K, \rho_{\min}\}.$$

If $T \geq \frac{4}{\widetilde{\rho}\log(e/2)}\log\frac{K+d}{\delta}$, then with probability at least $1-\delta$, we have

$$\lambda_{\min}\left(\sum_{t=1}^T \widetilde{\boldsymbol{x}}_{A_t,t}\widetilde{\boldsymbol{x}}_{A_t,t}^\top\right) \geq \frac{T\widetilde{\rho}}{2}.$$

Proof. Let $\boldsymbol{X}_t = \widetilde{\boldsymbol{x}}_{A_t,t}\widetilde{\boldsymbol{x}}_{A_t,t}^\top$. Uniform exploration makes A_t independent of the current contexts, and the contexts are independent across rounds, so $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ are independent positive semidefinite matrices. Since $\widetilde{\boldsymbol{x}}_{A_t,t} = (\boldsymbol{e}_{A_t}^\top, \boldsymbol{\alpha}_{A_t,t}^\top)^\top$, the rank-one matrix $\boldsymbol{X}_t$ has the block form

$$\boldsymbol{X}_t = \begin{pmatrix} \boldsymbol{e}_{A_t}\boldsymbol{e}_{A_t}^\top & \boldsymbol{e}_{A_t}\boldsymbol{\alpha}_{A_t,t}^\top \\ \boldsymbol{\alpha}_{A_t,t}\boldsymbol{e}_{A_t}^\top & \boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top \end{pmatrix}.$$

Taking expectations block by block, and conditioning on the uniformly selected arm A_t, gives

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}_t] &= \mathbb{E}\left[\begin{pmatrix} \boldsymbol{e}_{A_t}\boldsymbol{e}_{A_t}^\top & \boldsymbol{e}_{A_t}\boldsymbol{\alpha}_{A_t,t}^\top \\ \boldsymbol{\alpha}_{A_t,t}\boldsymbol{e}_{A_t}^\top & \boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}[\boldsymbol{e}_{A_t}\boldsymbol{e}_{A_t}^\top] & \mathbb{E}[\boldsymbol{e}_{A_t}\boldsymbol{\alpha}_{A_t,t}^\top] \\ \mathbb{E}[\boldsymbol{\alpha}_{A_t,t}\boldsymbol{e}_{A_t}^\top] & \mathbb{E}[\boldsymbol{\alpha}_{A_t,t}\boldsymbol{\alpha}_{A_t,t}^\top] \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{K}\sum_{i=1}^K \boldsymbol{e}_i\boldsymbol{e}_i^\top & \frac{1}{K}\sum_{i=1}^K \boldsymbol{e}_i\,\mathbb{E}[\boldsymbol{\alpha}_{i,t}]^\top \\ \frac{1}{K}\sum_{i=1}^K \mathbb{E}[\boldsymbol{\alpha}_{i,t}]\boldsymbol{e}_i^\top & \frac{1}{K}\sum_{i=1}^K \mathbb{E}[\boldsymbol{\alpha}_{i,t}\boldsymbol{\alpha}_{i,t}^\top] \end{pmatrix} \\
&= \begin{pmatrix} K^{-1}\boldsymbol{I}_K & 0 \\ 0 & \boldsymbol{\Sigma}_c \end{pmatrix}.
\end{aligned}$$

The two off-diagonal blocks are zero because Assumption 1 gives $\mathbb{E}[\boldsymbol{\alpha}_{i,t}] = 0$. The lower-right block equals $\boldsymbol{\Sigma}_c$ because Assumption 1 gives $\mathbb{E}[\boldsymbol{\alpha}_{i,t}\boldsymbol{\alpha}_{i,t}^\top] = \boldsymbol{\Sigma}_c$ for every arm i.

Take any $\boldsymbol{z} \in \mathbb{R}^{K+d}$ and write $\boldsymbol{z} = (\boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top$, where $\boldsymbol{u} \in \mathbb{R}^K$ and $\boldsymbol{v} \in \mathbb{R}^d$. Then

$$\begin{aligned}
\boldsymbol{z}^\top\begin{pmatrix} K^{-1}\boldsymbol{I}_K & 0 \\ 0 & \boldsymbol{\Sigma}_c \end{pmatrix}\boldsymbol{z} &= K^{-1}\|\boldsymbol{u}\|^2 + \boldsymbol{v}^\top\boldsymbol{\Sigma}_c\boldsymbol{v} \\
&\geq K^{-1}\|\boldsymbol{u}\|^2 + \rho_{\min}\|\boldsymbol{v}\|^2 \\
&\geq \widetilde{\rho}\Big(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2\Big) \\
&= \boldsymbol{z}^\top\widetilde{\rho}\boldsymbol{I}_{K+d}\boldsymbol{z}.
\end{aligned}$$

The first inequality uses $\boldsymbol{\Sigma}_c \succeq \rho_{\min}\boldsymbol{I}_d$, and the second uses $\widetilde{\rho} = \min\{1/K, \rho_{\min}\}$. By the definition of the Loewner order, we have $\mathbb{E}[\boldsymbol{X}_t] \succeq \widetilde{\rho}\boldsymbol{I}_{K+d}$. Thus, by Weyl's inequality, i.e., $\lambda_{\min}(\boldsymbol{A}+\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B})$, we have

$$\mu_{\min} := \lambda_{\min}\left(\sum_{t=1}^T \mathbb{E}[\boldsymbol{X}_t]\right) \geq \sum_{t=1}^T \lambda_{\min}(\mathbb{E}[\boldsymbol{X}_t]) \geq T\widetilde{\rho}.$$

Since the only non-zero eigenvalue of $\boldsymbol{x}\boldsymbol{x}^\top$ is $\|\boldsymbol{x}\|^2$, and by the assumption $\|\boldsymbol{\alpha}_{i,t}\| \leq 1$ in Section 3,

$$\lambda_{\max}(\boldsymbol{X}_t) = \|\widetilde{\boldsymbol{x}}_{A_t,t}\|^2 = \|\boldsymbol{e}_{A_t}\|^2 + \|\boldsymbol{\alpha}_{A_t,t}\|^2 = 1 + \|\boldsymbol{\alpha}_{A_t,t}\|^2 \leq 2.$$

Matrix Chernoff (Lemma 10) with $\varepsilon = 1/2$ gives

$$\Pr\left[\lambda_{\min}\left(\sum_{t=1}^{T} \boldsymbol{X}_t\right) \le \frac{T\widetilde{\rho}}{2}\right] \le (K+d)\left(\sqrt{\frac{2}{e}}\right)^{T\widetilde{\rho}/2}.$$

Therefore the RHS is at most δ whenever $T \ge \frac{4}{\widetilde{\rho}\log(e/2)} \log \frac{K+d}{\delta}$. □

Lemma 2. *For SST-UCB's confidence sets in Algorithm 1 (Lines 9 to 10), with probability at least $1-\delta/4$, the following bounds hold simultaneously at the end of every uniform exploration block:*

$$\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\right\|_{\widetilde{\boldsymbol{V}}_s} \le \beta_s, \qquad \left\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\right\| \le C_s := \frac{\beta_s}{\sqrt{\lambda_{\min}(\widetilde{\boldsymbol{V}}_s)}}.$$

If that endpoint has been reached by horizon n, then $\beta_s \le \overline{\beta}_n$.

Proof. Let $t_s = \sum_{r<s}(E_r + H_r) + E_s$ be the global round at the end of phase s's exploration. At this round, we have

$$\widetilde{\boldsymbol{V}}_s = \lambda \boldsymbol{I}_{K+d} + \sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t}\widetilde{\boldsymbol{x}}_{A_t,t}^\top, \qquad \boldsymbol{b} = \sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t} g_{A_t,t}.$$

Substituting the reward model (Equation 1) yields

$$\begin{aligned}
\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^* &= \widetilde{\boldsymbol{V}}_s^{-1} \sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t}\left(\widetilde{\boldsymbol{x}}_{A_t,t}^\top \widetilde{\boldsymbol{\theta}}^* + \eta_{A_t,t}\right) - \widetilde{\boldsymbol{\theta}}^* \\
&= \widetilde{\boldsymbol{V}}_s^{-1}\left[(\widetilde{\boldsymbol{V}}_s - \lambda \boldsymbol{I}_{K+d})\widetilde{\boldsymbol{\theta}}^* + \sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t}\eta_{A_t,t}\right] - \widetilde{\boldsymbol{\theta}}^* \\
&= \widetilde{\boldsymbol{V}}_s^{-1} \sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t}\eta_{A_t,t} - \lambda \widetilde{\boldsymbol{V}}_s^{-1}\widetilde{\boldsymbol{\theta}}^*.
\end{aligned}$$

By the Cauchy-Schwarz inequality and the triangle inequality,

$$\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\right\|_{\widetilde{\boldsymbol{V}}_s} \le \left\|\sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t}\eta_{A_t,t}\right\|_{\widetilde{\boldsymbol{V}}_s^{-1}} + \lambda\|\widetilde{\boldsymbol{\theta}}^*\|_{\widetilde{\boldsymbol{V}}_s^{-1}}.$$

Since $\eta_{A_t,t}$ is σ-sub-Gaussian conditional on the past, current contexts, and selected arm, by the self-normalized inequality (Theorem 1) of Abbasi-Yadkori et al. (2011), the following holds simultaneously at all rounds with probability at least $1-\delta/4$,

$$\left\|\sum_{t=1}^{t_s} \widetilde{\boldsymbol{x}}_{A_t,t}\eta_{A_t,t}\right\|_{\widetilde{\boldsymbol{V}}_s^{-1}} \le \sigma\sqrt{2\log(4/\delta) + \log\frac{\det(\widetilde{\boldsymbol{V}}_s)}{\lambda^{K+d}}}.$$

Since $\widetilde{\boldsymbol{V}}_s \succeq \lambda \boldsymbol{I}_{K+d}$,

$$\lambda\|\widetilde{\boldsymbol{\theta}}^*\|_{\widetilde{\boldsymbol{V}}_s^{-1}} \le \sqrt{\lambda}\|\widetilde{\boldsymbol{\theta}}^*\| = \sqrt{\frac{\sum_i \mu_i^2 + \|\boldsymbol{\theta}^*\|^2}{K+1}} \le 1.$$

This proves the first bound. For any vector $\boldsymbol{u}$, the Rayleigh quotient gives $\|\boldsymbol{u}\|_{\widetilde{\boldsymbol{V}}_s}^2 \ge \lambda_{\min}(\widetilde{\boldsymbol{V}}_s)\|\boldsymbol{u}\|^2$. Applying this inequality obtains the second bound.

Finally, if $t_s \le n$, then Lemma 11 and $\|\widetilde{\boldsymbol{x}}_{A_t,t}\|^2 \le 2$ give

$$\log\frac{\det(\widetilde{\boldsymbol{V}}_s)}{\lambda^{K+d}} \le (K+d)\log\left(1 + \frac{2t_s}{\lambda(K+d)}\right) \le (K+d)\log\left(1 + \frac{2n(K+1)}{K+d}\right).$$

Substituting into β_s proves $\beta_s \le \overline{\beta}_n$. □

Lemma 3. *Let $N_s = \sum_{r=1}^{s} E_r = K(2^s - 1)$ denote the cumulative number of uniform exploration rounds up to the end of phase s. With probability at least $1 - \delta/4$, every phase s satisfying*

$$N_s \geq \frac{4}{\widetilde{\rho}\log(e/2)} \log \frac{4(K+d)s(s+1)}{\delta} \tag{10}$$

has $\lambda_{\min}(\widetilde{\boldsymbol{V}}_s) \geq \lambda + N_s\widetilde{\rho}/2$. On this event, if the uniform exploration block in phase s is completed by horizon n and $N_s \geq A_n$, then

$$C_s \leq \sqrt{\frac{A_n}{8N_s}}. \tag{11}$$

Proof. Let $\boldsymbol{G}_{N_s}$ be the Gram matrix of the N_s uniform exploration samples collected through phase s. Applying Lemma 1 with $T = N_s$ and failure probability $\varepsilon_s = \delta/[4s(s+1)]$ gives exactly Equation (10), and hence $\Pr(\lambda_{\min}(\boldsymbol{G}_{N_s}) < N_s\widetilde{\rho}/2) \leq \varepsilon_s$. Since $\widetilde{\boldsymbol{V}}_s \succeq \lambda \boldsymbol{I}_{K+d} + \boldsymbol{G}_{N_s}$, we have $\lambda_{\min}(\widetilde{\boldsymbol{V}}_s) \geq \lambda + \lambda_{\min}(\boldsymbol{G}_{N_s})$. The union bound and the identity $\sum_{s\geq 1} 1/[s(s+1)] = 1$ then give $\lambda_{\min}(\widetilde{\boldsymbol{V}}_s) \geq \lambda + N_s\widetilde{\rho}/2$ with probability at least $1 - \delta/4$.

If $N_s \geq A_n$, by the definition of A_n and $16 \geq 4/\log(e/2)$, we have

$$\begin{aligned} N_s \geq A_n &\geq \frac{16}{\widetilde{\rho}} \log \frac{4(K+d)n(n+1)}{\delta} \\ &\geq \frac{16}{\widetilde{\rho}} \log \frac{4(K+d)s(s+1)}{\delta} \\ &\geq \frac{4}{\widetilde{\rho}\log(e/2)} \log \frac{4(K+d)s(s+1)}{\delta}. \end{aligned}$$

Since $\beta_s \leq \overline{\beta}_n$, we have

$$C_s := \frac{\beta_s}{\sqrt{\lambda_{\min}(\widetilde{\boldsymbol{V}}_s)}} \leq \frac{\overline{\beta}_n}{\sqrt{\lambda + N_s\widetilde{\rho}/2}} \leq \sqrt{\frac{2\overline{\beta}_n^2}{N_s\widetilde{\rho}}} \leq \sqrt{\frac{A_n}{8N_s}},$$

where the last step uses $A_n \geq 16\overline{\beta}_n^2/\widetilde{\rho}$. □

Lemma 4. *On the event of Lemma 2, we have $\widehat{\ell}_s \leq \ell^*$ for all phase s. Suppose now that $\ell^* \geq 1$. On the same event, $\widehat{\ell}_s = \ell^*$ whenever $C_s < \gamma_{\ell^*-1}/2$. Moreover, on the intersection of the events of Lemmas 2 and 3, every exploration endpoint reached by horizon n with $N_s \geq A_n/\gamma_{\ell^*-1}^2$ satisfies $\widehat{\ell}_s = \ell^*$, simultaneously for all $n \geq 1$.*

Proof. We work on the event in Lemma 2. The true parameter $\widetilde{\boldsymbol{\theta}}^*$ belongs both to $\mathcal{C}_s$ and to $\mathbb{R}^K \times V_{\ell^*}$. Hence level ℓ^* passes the test of Line 11 in Algorithm 1, and therefore $\widehat{\ell}_s \leq \ell^*$.

Suppose $\ell^* \geq 1$, and assume that a smaller level $\ell < \ell^*$ pass the test, then there is a vector $\boldsymbol{v} = (\boldsymbol{v}_\mu^\top, \boldsymbol{v}_\theta^\top)^\top \in \mathcal{C}_s$ with $\boldsymbol{v}_\theta \in V_\ell$. On the event in Lemma 2, $\widetilde{\boldsymbol{\theta}}^* \in \mathcal{C}_s$. By the definition of $\mathcal{C}_s$,

$$\|\widetilde{\boldsymbol{\theta}}^* - \widehat{\widetilde{\boldsymbol{\theta}}}_s\|_{\widetilde{\boldsymbol{V}}_s} \leq \beta_s, \qquad \|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \boldsymbol{v}\|_{\widetilde{\boldsymbol{V}}_s} \leq \beta_s.$$

For any vector $\boldsymbol{u} \in \mathbb{R}^{K+d}$, $\|\boldsymbol{u}\|_{\widetilde{\boldsymbol{V}}_s}^2 = \boldsymbol{u}^\top \widetilde{\boldsymbol{V}}_s \boldsymbol{u} \geq \lambda_{\min}(\widetilde{\boldsymbol{V}}_s)\|\boldsymbol{u}\|^2$, so $\|\boldsymbol{u}\| \leq \frac{\|\boldsymbol{u}\|_{\widetilde{\boldsymbol{V}}_s}}{\sqrt{\lambda_{\min}(\widetilde{\boldsymbol{V}}_s)}}$. Combining this inequality with the triangle inequality yields

$$\begin{aligned} \|\widetilde{\boldsymbol{\theta}}^* - \boldsymbol{v}\| &\leq \|\widetilde{\boldsymbol{\theta}}^* - \widehat{\widetilde{\boldsymbol{\theta}}}_s\| + \|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \boldsymbol{v}\| \\ &\leq \frac{\|\widetilde{\boldsymbol{\theta}}^* - \widehat{\widetilde{\boldsymbol{\theta}}}_s\|_{\widetilde{\boldsymbol{V}}_s} + \|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \boldsymbol{v}\|_{\widetilde{\boldsymbol{V}}_s}}{\sqrt{\lambda_{\min}(\widetilde{\boldsymbol{V}}_s)}} \leq \frac{2\beta_s}{\sqrt{\lambda_{\min}(\widetilde{\boldsymbol{V}}_s)}} = 2C_s. \end{aligned}$$

Since $\boldsymbol{v}_\theta \in V_\ell$, we have

$$\boldsymbol{\Pi}_\ell \boldsymbol{v}_\theta = \boldsymbol{v}_\theta, \qquad (\boldsymbol{I} - \boldsymbol{\Pi}_\ell)\boldsymbol{v}_\theta = 0.$$

Moreover, the components in V_ℓ and $V_\ell^\perp$ are orthogonal, so

$$\|\boldsymbol{\theta}^* - \boldsymbol{v}_\theta\|^2 = \|\boldsymbol{\Pi}_\ell(\boldsymbol{\theta}^* - \boldsymbol{v}_\theta)\|^2 + \|(\boldsymbol{I} - \boldsymbol{\Pi}_\ell)(\boldsymbol{\theta}^* - \boldsymbol{v}_\theta)\|^2 \geq \|(\boldsymbol{I} - \boldsymbol{\Pi}_\ell)(\boldsymbol{\theta}^* - \boldsymbol{v}_\theta)\|^2.$$

Consequently,

$$\gamma_\ell = \|(\boldsymbol{I} - \boldsymbol{\Pi}_\ell)\boldsymbol{\theta}^*\| = \|(\boldsymbol{I} - \boldsymbol{\Pi}_\ell)(\boldsymbol{\theta}^* - \boldsymbol{v}_\theta)\| \leq \|\boldsymbol{\theta}^* - \boldsymbol{v}_\theta\| \leq \|\widetilde{\boldsymbol{\theta}}^* - \boldsymbol{v}\| \leq 2C_s.$$

For $\ell < \ell^*$, we have $V_\ell \subseteq V_{\ell^*-1}$, so the distance from $\boldsymbol{\theta}^*$ to V_ℓ cannot be smaller than its distance to V_{ℓ^*-1}.

$$\gamma_\ell = \min_{\boldsymbol{v} \in V_\ell} \|\boldsymbol{\theta}^* - \boldsymbol{v}\| \geq \min_{\boldsymbol{v} \in V_{\ell^*-1}} \|\boldsymbol{\theta}^* - \boldsymbol{v}\| = \gamma_{\ell^*-1}.$$

Thus every passing smaller level requires $\gamma_{\ell^*-1} \leq 2C_s$. So if $C_s < \gamma_{\ell^*-1}/2$, no smaller level can pass, and therefore $\widehat{\ell}_s = \ell^*$.

On the joint event of Lemmas 2 and 3. Since $0 < \gamma_{\ell^*-1} \leq \|\boldsymbol{\theta}^*\| \leq 1$, the $N_s \geq A_n/\gamma_{\ell^*-1}^2$ implies $N_s \geq A_n$. Equation (11) gives

$$C_s \leq \sqrt{\frac{A_n}{8N_s}} \leq \frac{\gamma_{\ell^*-1}}{\sqrt{8}} < \frac{\gamma_{\ell^*-1}}{2}.$$

Therefore we have $\widehat{\ell}_s = \ell^*$. □

Lemma 5. *On the event of Lemma 2, for every phase s and every exploitation round t in that phase, optimal arms are retained in $\mathcal{A}_{s,t}$, i.e., $\arg\max_{i\in[K]}\{\mu_i + \langle\boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle\} \subseteq \mathcal{A}_{s,t}$. For every retained arm $i \in \mathcal{A}_{s,t}$, we have*

$$\max_{j\in[K]}(\mu_j + \langle\boldsymbol{\theta}^*, \boldsymbol{\alpha}_{j,t}\rangle) - (\mu_i + \langle\boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle) \leq 4\sqrt{2}C_s.$$

This bound holds even when $\widehat{\ell}_s \neq \ell^*$.

Proof. For clarity, we write $m_{i,t} = \mu_i + \langle\boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle = \widetilde{\boldsymbol{x}}_{i,t}^\top\widetilde{\boldsymbol{\theta}}^*$ and denote

$$m_t^* := \max_i m_{i,t}, \qquad \widehat{m}_{i,t}^s := \widetilde{\boldsymbol{x}}_{i,t}^\top\widehat{\widetilde{\boldsymbol{\theta}}}_s, \qquad w_{i,t}^s := \beta_s\|\widetilde{\boldsymbol{x}}_{i,t}\|_{\widetilde{\boldsymbol{V}}_s^{-1}}.$$

Thus, on the event of Lemma 2, we have

$$\begin{aligned}
|\widehat{m}_{i,t}^s - m_{i,t}| &= \left|\widetilde{\boldsymbol{x}}_{i,t}^\top(\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*)\right| = \left|(\widetilde{\boldsymbol{V}}_s^{-1/2}\widetilde{\boldsymbol{x}}_{i,t})^\top\widetilde{\boldsymbol{V}}_s^{1/2}(\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*)\right| \\
&\leq \|\widetilde{\boldsymbol{V}}_s^{-1/2}\widetilde{\boldsymbol{x}}_{i,t}\|_2\|\widetilde{\boldsymbol{V}}_s^{1/2}(\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*)\|_2 \\
&= \|\widetilde{\boldsymbol{x}}_{i,t}\|_{\widetilde{\boldsymbol{V}}_s^{-1}}\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\|_{\widetilde{\boldsymbol{V}}_s} \\
&\leq \beta_s\|\widetilde{\boldsymbol{x}}_{i,t}\|_{\widetilde{\boldsymbol{V}}_s^{-1}} = w_{i,t}^s.
\end{aligned}$$

Therefore, for any optimal arm i^*, its upper bound is at least m_t^*, which is greater than every arm's lower bound, i.e.,

$$\widehat{m}_{i^*,t}^s + w_{i^*,t}^s \geq m_{i^*,t} = m_t^* \geq m_{i,t} \geq \widehat{m}_{i,t}^s - w_{i,t}^s, \quad \forall i \in [K].$$

Therefore $i^* \in \mathcal{A}_{s,t}$.

For a retained arm i, we have $\widehat{m}_{i,t}^s + w_{i,t}^s \geq \widehat{m}_{i^*,t}^s - w_{i^*,t}^s$. It follows that

$$m_t^* - m_{i,t} \leq \widehat{m}_{i^*,t}^s + w_{i^*,t}^s - \widehat{m}_{i,t}^s + w_{i,t}^s \leq 2(w_{i^*,t}^s + w_{i,t}^s) \leq 4\sqrt{2}\,\frac{\beta_s}{\sqrt{\lambda_{\min}(\widetilde{\boldsymbol{V}}_s)}}.$$

The last inequality uses $\|\widetilde{\boldsymbol{x}}_{i,t}\| \leq \sqrt{2}$ and $\widetilde{\boldsymbol{V}}_s^{-1} \preceq \lambda_{\min}(\widetilde{\boldsymbol{V}}_s)^{-1}\boldsymbol{I}_{K+d}$. □

Lemma 6 (SupLinUCB guarantee)**.** *Consider a fresh SupLinUCB learner used by* `SST-UCB` *or* `SSC-UCB` *in phase s at level $j \in \{0, \dots, L\}$ with $\boldsymbol{\theta}^* \in V_j$, using the projected features $\widetilde{\boldsymbol{x}}_{i,t}^{(j)}$ and original rewards. Let $H = H_s$ be its local horizon, corresponding to the input T in Chu et al. (2011, Algorithm 3), and hence an upper bound on the number of rounds assigned to this learner. For confidence $0 < \zeta < 1$, use regularization $\lambda = 1/(K+1)$ and width coefficient a_H, where*

$$S_H = 1 + \lceil \log_2(H+1) \rceil, \qquad a_H = 1 + \sigma\sqrt{2\log(2KHS_H/\zeta)}.$$

With probability at least $1-\zeta$ conditional on its starting history, its regret relative to the best offered arm satisfies, simultaneously for every prefix of $0 \le h \le H$ rounds assigned to it,

$$R_{\text{base}}(h) \le \min\left\{4h, 32(1+\sigma)\log\frac{32K(H+1)}{\zeta}\sqrt{(K+d_j)h}\right\}.$$

If $\zeta \le 1/(16H^2)$, the same event also gives

$$R_{\text{base}}(h) \le 22(1+\sigma)\log(1/\zeta)\sqrt{(K+d_j)h}. \tag{12}$$

Proof. In Line 14 of Chu et al. (2011, Algorithm 3), we choose the smallest-index arm satisfying the width condition. Condition on the learner's starting history, contexts, and assigned rounds, using the continuation without deletions for `SSC-UCB`. The conditions for the layer independence argument of Chu et al. (2011, Lemma 4) are verified in Appendices C and D. Write $p = K + d_j$ and abbreviate the projected features by $\boldsymbol{x}$. Their norms are at most $\sqrt{2}$, and the exact parameter has norm at most $\sqrt{K+1}$. For a layer matrix $\boldsymbol{V} = \lambda \boldsymbol{I}_p + \sum_\tau \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\top$, the noise in a ridge prediction is $\sigma\|\boldsymbol{x}\|_{\boldsymbol{V}^{-1}}$-sub-Gaussian because

$$\sum_\tau (\boldsymbol{x}^\top \boldsymbol{V}^{-1}\boldsymbol{x}_\tau)^2 = \boldsymbol{x}^\top \boldsymbol{V}^{-1}(\boldsymbol{V} - \lambda \boldsymbol{I}_p)\boldsymbol{V}^{-1}\boldsymbol{x} \le \boldsymbol{x}^\top \boldsymbol{V}^{-1}\boldsymbol{x}.$$

The ridge bias is at most $\sqrt{\lambda(K+1)}\|\boldsymbol{x}\|_{\boldsymbol{V}^{-1}} = \|\boldsymbol{x}\|_{\boldsymbol{V}^{-1}}$. The sub-Gaussian tail bound therefore places the prediction error within $a_H\|\boldsymbol{x}\|_{\boldsymbol{V}^{-1}}$ with failure at most $\zeta/(KHS_H)$. Since $2^{-S_H} \le H^{-1/2}$, at most S_H layers are visited. Averaging over the conditioning and taking a union bound over arms, calls, and layers gives simultaneous prediction bounds with conditional probability at least $1-\zeta$.

On this event, the best offered arm survives each layer elimination, as in Chu et al. (2011, Lemma 5). An arm retained after layer v has regret at most $2^{1-v} + 2\cdot 2^{-v} = 4\cdot 2^{-v}$. Thus an arm sampled in layer $v \ge 2$ has regret at most $8\cdot 2^{-v}$, and this also holds at layer 1 because mean rewards lie in $[-2,2]$. The stopping-width rule incurs regret at most $2/\sqrt{H}$.

Let N_v count updates to layer v in the first h calls, and let $u_\tau = \boldsymbol{x}_\tau^\top \boldsymbol{V}_{\tau-1}^{-1}\boldsymbol{x}_\tau$ at each such update. Sampling gives $u_\tau > 4^{-v}/a_H^2$, so $\log(1+u_\tau) \ge 4^{-v}/(a_H^2 + 1/4)$ by $\log(1+z) \ge z/(1+z)$ and $v \ge 1$. The determinant identity and trace bound imply

$$\frac{N_v 4^{-v}}{a_H^2 + 1/4} \le \sum_{\tau \text{ in layer } v} \log(1+u_\tau) = \log\frac{\det(\boldsymbol{V}_{v,h})}{\det(\lambda \boldsymbol{I}_p)} \le p\log\left(1 + \frac{2N_v}{\lambda p}\right) \le p\log(1+3h).$$

Here $\boldsymbol{V}_{v,h}$ is the layer matrix after these calls, and $\lambda p \ge K/(K+1) \ge 2/3$. Set $B = (a_H^2 + 1/4)p\log(1+3h)$, so $N_v \le B4^v$ and $\sum_v N_v \le h$. Consequently,

$$\begin{aligned}\left(\sum_{v=1}^{S_H} 2^{-v}N_v\right)^2 &= \sum_v 4^{-v}N_v^2 + 2\sum_v 2^{-v}N_v \sum_{w<v} 2^{-w}N_w \\ &\le B\sum_v N_v + 2B\sum_v 2^{-v}N_v(2^v - 2) \le 3Bh.\end{aligned}$$

The middle step uses $\sum_{w=1}^{v-1} 2^w = 2^v - 2$. Combining the sampling and stopping regrets yields

$$R_{\text{base}}(h) \le \frac{2h}{\sqrt{H}} + 8\sqrt{3(a_H^2 + 1/4)ph\log(1+3h)}.$$

For $\mathcal{L} = \log(32K(H+1)/\zeta) > 4$, we have $\log(1+3h) \le \mathcal{L}$, $\log(2KHS_H/\zeta) \le 2\mathcal{L}$, and $\sqrt{a_H^2 + 1/4} \le 3/2 + 2\sigma\sqrt{\mathcal{L}}$. For $0 < h \le H$, division by $\sqrt{ph}$ gives

$$\frac{R_{\text{base}}(h)}{\sqrt{ph}} \le \sqrt{2} + 8\sqrt{3}\left(\frac{3}{2} + 2\sigma\sqrt{\mathcal{L}}\right)\sqrt{\mathcal{L}} \le 32(1+\sigma)\mathcal{L}.$$

If $\zeta \le 1/(16H^2)$, write $u = \log(1/\zeta) > 4$. Since $K \le H$ and $S_H \le 2H$, we instead have $\log(1+3h) \le u/2$ and $\log(2KHS_H/\zeta) \le 5u/2$. Using $\sqrt{a_H^2 + 1/4} \le \sqrt{5}/2 + \sigma\sqrt{5u}$ gives

$$\frac{R_{\text{base}}(h)}{\sqrt{ph}} \le \sqrt{2} + 4\sqrt{15u/2} + 8\sqrt{15/2}\,\sigma u \le 22(1+\sigma)u.$$

The last inequality uses $\sqrt{2}/u + 4\sqrt{15/(2u)} < 6$ for $u > 4$ and $8\sqrt{15/2} < 22$. Both conclusions include $h = 0$, and $R_{\text{base}}(h) \le 4h$ follows from mean rewards in $[-2, 2]$. □

Lemma 7. *In the simple model, with probability at least $1 - \delta/4$, every arm with a positive observation count satisfies, at every round,*

$$\left|\frac{S_i}{T_i} - \mu_i\right| \le \sigma\sqrt{\frac{2\log(8K(T_i+1)^2/\delta)}{T_i}}.$$

Proof. Let $\xi_1, \dots, \xi_m$ denote the noises in the first m observations of arm i. Since each arm is selected before its reward noise is observed, the sequence (ξ_r) is conditionally σ-sub-Gaussian, so Hoeffding's inequality gives

$$\Pr\left(\left|\sum_{r=1}^{m} \xi_r\right| > \sigma\sqrt{2m\log\frac{8K(m+1)^2}{\delta}}\right) \le \frac{\delta}{4K(m+1)^2}.$$

At any round up to n at which $T_i = m$, the simple reward model gives $|S_i/T_i - \mu_i| = \frac{1}{m}|\sum_{r=1}^{m} \xi_r|$. Therefore, a union bound over arms and observation times bounds the probability of any confidence failure through round n by

$$\sum_{i=1}^{K}\sum_{m=1}^{n} \frac{\delta}{4K(m+1)^2} \le \frac{\delta}{4}\sum_{m=1}^{\infty}\frac{1}{(m+1)^2} < \frac{\delta}{4}.$$

Letting $n \to \infty$ proves the simultaneous guarantee over all rounds. □

Lemma 8. *Let L_s be the number of exploitation rounds executed in phase s by horizon n, so $0 \le L_s \le H_s$. The number of uniform exploration rounds is at most $4\sqrt{Kn}$, and*

$$\sum_s \sqrt{L_s} \le 2\sqrt{n}. \tag{13}$$

Every phase with $L_s > 0$ satisfies $H_s \le 4n$. Moreover, $N_s = K(2^s - 1)$ satisfies $H_s \le N_s^2/K$, and for $u > 0$ and $q > 0$,

$$\sum_{s:N_s<u} N_s^q \le \frac{u^q}{1 - 2^{-q}}. \tag{14}$$

Proof. Let S be the last entered phase. If $S = 1$, the number of exploration rounds is at most $\min\{n, K\} \le \sqrt{Kn}$. If $S \ge 2$, phase $S-1$ is completed, so $n \ge \sum_{s=1}^{S-1}(E_s + H_s) \ge H_{S-1} = K4^{S-2}$. The number of exploration rounds is at most

$$\sum_{s=1}^{S} E_s = K(2^S - 1) < 4K2^{S-2} \le 4\sqrt{Kn}.$$

Let P be the last completed phase, or set $P = 0$ if no phase is completed. Each completed phase has $L_s = H_s = K4^{s-1}$. Therefore,

$$\sum_{s=1}^{P}\sqrt{H_s} = \sqrt{K}(2^P - 1), \qquad \sum_{s=1}^{P} H_s = \frac{K}{3}(4^P - 1).$$

Since $(2^P-1)^2 \le 4^P-1$,

$$\left(\sum_{s=1}^{P}\sqrt{H_s}\right)^2 \le 3\sum_{s=1}^{P} H_s.$$

There is at most one incomplete phase, whose exploitation length is L_{P+1}. Cauchy-Schwarz applied to the two remaining terms gives

$$\sum_s \sqrt{L_s} \le \sqrt{3\sum_{s=1}^{P} H_s} + \sqrt{L_{P+1}} \le 2\sqrt{\sum_{s=1}^{P} H_s + L_{P+1}} \le 2\sqrt{n}.$$

The last inequality uses that the total number of exploitation rounds is at most n. If phase one reaches exploitation, then $n \ge K = H_1$. Any later reached exploitation phase has its preceding block complete, so $n \ge H_{s-1} = H_s/4$. This proves the $H_s \le 4n$ bound.

Finally, $N_s \ge K2^{s-1}$ implies $H_s = K4^{s-1} \le N_s^2/K$, and $N_{s+1} = 2N_s + K > 2N_s$. If $u \le K = N_1$, no phase satisfies $N_s < u$, so the sum in Equation (14) is zero and the bound holds. Otherwise, let $p = \max\{s \ge 1 : N_s < u\}$. Repeatedly applying $N_{s+1} > 2N_s$ gives $N_s \le 2^{s-p}N_p$ for $s \le p$. Therefore,

$$\sum_{s:N_s<u} N_s^q \le N_p^q \sum_{j=0}^{p-1} 2^{-qj} = N_p^q \frac{1-2^{-qp}}{1-2^{-q}} \le \frac{u^q}{1-2^{-q}},$$

where the last inequality uses $N_p < u$ and $q > 0$. □

Theorem 3 (Regret guarantee of `SST-UCB`)**.** *Under Assumption 1, `SST-UCB` satisfies the following bounds for all time horizon $n \ge 1$ with probability at least $1-\delta$.*

If $\ell^* = 0$, then

$$R_n \le \left(10 + 4\sigma\sqrt{2\log\frac{8K(n+1)^2}{\delta}}\right)\sqrt{Kn} = \widetilde{\mathcal{O}}\Big(\sqrt{Kn}\Big). \tag{3}$$

If $\ell^* \ge 1$, then

$$R_n \le 16\sqrt{Kn} + C(1+\sigma)\Lambda_n^2\sqrt{(K+d_{\ell^*})n} + \frac{9A_n^2}{K\gamma_{\ell^*-1}^3} = \widetilde{\mathcal{O}}\left(\sqrt{(K+d_{\ell^*})n} + \frac{(K+d)^2}{K\widehat{\rho}^2\gamma_{\ell^*-1}^3}\right), \tag{4}$$

where C is a universal constant.

Proof. The conclusions of Lemmas 2 and 3 hold simultaneously with failure probability at most $\delta/2$, and Lemma 7 contributes at most $\delta/4$ in the simple model. Conditional on the history at the start of each exploitation block, the learner is fresh and its frozen filter is a fixed function of the current contexts, so Lemma 6 applies whenever the selected positive level is exact, with failure probability at most ε_s. Averaging over these histories and using $\sum_{s\ge1}\varepsilon_s = \delta/4$, a union bound gives probability at least $1-\delta$ that all applicable conclusions hold. We work throughout on this intersection and fix an arbitrary horizon n.

The simple model. If $\boldsymbol{\theta}^* = 0$, Lemma 4 gives $\widehat{\ell}_s = 0$ at the end of every phase. Every exploitation round therefore uses UCB. Lemma 5 ensures that optimal arms with mean $\mu^* = \max_i \mu_i$ are retained in $\mathcal{A}_{s,t}$. For a selected suboptimal arm i that has already been observed $m \ge 1$ times, optimism and Lemma 7 give

$$\mu^* \le U_i \le \mu_i + 2\sigma\sqrt{\frac{2\log(8K(m+1)^2/\delta)}{m}}.$$

Thus, if an observed arm i is selected, its UCB value satisfies $U_j \le U_i < +\infty$ for every $j \in \mathcal{A}_{s,t}$, including an optimal arm. Writing $\Delta_i = \mu^* - \mu_i$, we obtain

$$\Delta_i \le 2\sigma\sqrt{\frac{2\log(8K(n+1)^2/\delta)}{m}}.$$

If $n < K$, all executed rounds are uniform exploration and $R_n^s \le 2n \le 2\sqrt{Kn}$, so suppose that $n \ge K$. Selecting an arm i with $T_i = 0$ during exploitation incurs regret at most 2, since $\mu^* - \mu_i \le 2$. This can happen at most once for each arm, hence at most K times in total. For every subsequent exploitation selection of arm i, let $m \ge 1$ be its number of observations before that round. Observing its reward increases the counter from m to $m+1$, so each value of m occurs at most once among the exploitation selections of that arm. Let $T_i(n)$ be its total observation count by n. Since $x^{-1/2}$ is decreasing, we have

$$\sum_{m=1}^{T_i(n)-1} \frac{1}{\sqrt{m}} \le \int_0^{T_i(n)} \frac{\mathrm{d}x}{\sqrt{x}} = 2\sqrt{T_i(n)}.$$

Consequently, the exploitation regret is at most

$$\begin{aligned}
&2K + 2\sigma\sqrt{2\log(8K(n+1)^2/\delta)} \sum_{i=1}^{K} \sum_{m=1}^{T_i(n)-1} \frac{1}{\sqrt{m}} \\
&\le 2K + 4\sigma\sqrt{2\log(8K(n+1)^2/\delta)} \sum_{i=1}^{K} \sqrt{T_i(n)} \\
&\le 2\sqrt{Kn} + 4\sigma\sqrt{2\log(8K(n+1)^2/\delta)}\sqrt{Kn}.
\end{aligned}$$

The last line uses $K \le \sqrt{Kn}$, Cauchy-Schwarz, and $\sum_i T_i(n) = n$. Since uniform exploration costs at most 2 per round, so the exploration regret is at most $8\sqrt{Kn}$ due to Lemma 8. Adding them together proves Equation (3).

The contextual model. Denote $\gamma = \gamma_{\ell^*-1}$. Uniform exploration costs at most 4 per round, so it contributes at most $16\sqrt{Kn}$ regret due to Lemma 8. It remains to bound exploitation. By Lemma 4, every phase whose exploration endpoint is reached by n and satisfies $N_s \ge A_n/\gamma^2$ selects level ℓ^*. We first bound the exploitation regret of all other phases, without assuming that $\widehat{\ell}_s = \ell^*$.

For phases with $N_s < A_n$, we directly use the trivial instantaneous regret bound 4 and $L_s \le H_s \le N_s^2/K$. Equation (14) with $q = 2$ gives

$$\sum_{s:\, N_s < A_n} 4L_s \le \frac{4}{K} \sum_{s:\, N_s < A_n} N_s^2 \le \frac{16A_n^2}{3K}.$$

For the remaining phases with $A_n \le N_s < A_n/\gamma^2$, Lemmas 3 and 5 imply

$$R_s \le 4\sqrt{2}C_s L_s \le 2\sqrt{\frac{A_n}{N_s}\frac{N_s^2}{K}} = \frac{2\sqrt{A_n}}{K} N_s^{3/2},$$

where R_s denotes the regret of the executed exploitation rounds in phase s. Summing and applying Equation (14) with $q = 3/2$ yields

$$\sum_{s:\, A_n \le N_s < A_n/\gamma^2} R_s \le \frac{2\sqrt{A_n}}{K} \frac{(A_n/\gamma^2)^{3/2}}{1 - 2^{-3/2}} = \frac{2A_n^2}{K\gamma^3(1 - 2^{-3/2})}.$$

Let R_{pre} denote the total exploitation regret over phases with $N_s < A_n/\gamma^2$. Combining the preceding two bounds and using $\gamma \le 1$, we obtain

$$R_{\mathrm{pre}} \le \left(\frac{16}{3} + \frac{2}{1 - 2^{-3/2}}\right) \frac{A_n^2}{K\gamma^3} \le \frac{9A_n^2}{K\gamma^3}. \tag{15}$$

For every remaining phase with $N_s \ge A_n/\gamma^2$, the selected model is exact, i.e., $\widehat{\ell}_s = \ell^*$. Its mean reward can therefore be written as

$$\mu_i + \boldsymbol{\alpha}_{i,t}^\top \boldsymbol{\theta}^* = \mu_i + (\boldsymbol{U}_{\ell^*}^\top \boldsymbol{\alpha}_{i,t})^\top (\boldsymbol{U}_{\ell^*}^\top \boldsymbol{\theta}^*).$$

The algorithm uses a fresh SupLinUCB learner at level ℓ^*. By Lemma 5, the available arms include a globally optimal arm. Applying Lemma 6 with horizon H_s and confidence ε_s therefore gives

$$R_s \le 32(1+\sigma)\log\frac{32K(H_s+1)}{\varepsilon_s}\sqrt{(K + d_{\ell^*})L_s}.$$

For every phase with $L_s > 0$, we have $s \le n$ and $H_s \le 4n$ by Lemma 8. Using $\varepsilon_s = \delta/(4s(s+1))$,

$$\log \frac{32K(H_s+1)}{\varepsilon_s} = \log \frac{128Ks(s+1)(H_s+1)}{\delta} \le \log \frac{512K(n+1)^3}{\delta} \le 4\Lambda_n.$$

Let R_{post} denote the total exploitation regret over phases with $N_s \ge A_n/\gamma^2$. We have

$$\begin{aligned} R_{\mathrm{post}} = \sum_{s:\, N_s \ge A_n/\gamma^2} R_s &\le 128(1+\sigma)\Lambda_n \sqrt{K + d_{\ell^*}} \sum_{s:\, N_s \ge A_n/\gamma^2} \sqrt{L_s} \\ &\le 256(1+\sigma)\Lambda_n \sqrt{(K+d_{\ell^*})n}, \end{aligned}$$

where the last inequality follows from Lemma 8. Adding the uniform exploration regret $16\sqrt{Kn}$ and Equation (15), and using $\Lambda_n \ge 1$, proves Equation (4) with $C = 256$. □

D Proof of Theorem 4

In the following, let m_n be the largest number of candidates $|\mathcal{J}_s|$ among phases whose exploitation has begun by round n, with $m_n = 1$ if no exploitation has begun. Define

$$\Phi_n = (1+\sigma)\Lambda_n. \tag{16}$$

Lemma 9. *Let $\boldsymbol{V}_s$ denote the matrix $\boldsymbol{V}$ at the end of phase s's uniform exploration block in Algorithm 2 and $\widehat{\widetilde{\boldsymbol{\theta}}}_s := \boldsymbol{V}_s^{-1}\boldsymbol{b}$. With probability at least $1 - \delta/4$, the following statements hold simultaneously for every phase s.*

(1) The estimation error satisfies

$$\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\| \le r_s.$$

(2) The model test in Line 11 retains every candidate $j \in \mathcal{J}_s$ satisfying $\boldsymbol{\theta}^* \in V_j$. Equivalently,

$$\{j \in \mathcal{J}_s : \boldsymbol{\theta}^* \in V_j\} \subseteq \mathcal{I}.$$

(3) If $E_s \ge A_n$, then

$$r_s^2 < A_n/(4E_s).$$

(4) If $\ell^* \ge 1$ and $E_s \ge A_n/\gamma^2_{\ell^*-1}$, the model test retains only models containing the true parameter:

$$\boldsymbol{\theta}^* \in V_j \qquad \text{for every } j \in \mathcal{I}.$$

Proof. In the following, let $\mathcal{E}_s$ denote the uniform exploration rounds in phase s, and write $\boldsymbol{V}_s$ for the matrix $\boldsymbol{V}$ at the end of these rounds in Algorithm 2.

Similar to the proof of Lemma 2, we have

$$\begin{aligned} \widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^* &= \boldsymbol{V}_s^{-1} \sum_{t \in \mathcal{E}_s} \widetilde{\boldsymbol{x}}_{A_t,t} \left(\widetilde{\boldsymbol{x}}_{A_t,t}^\top \widetilde{\boldsymbol{\theta}}^* + \eta_{A_t,t} \right) - \widetilde{\boldsymbol{\theta}}^* \\ &= \boldsymbol{V}_s^{-1} \left[(\boldsymbol{V}_s - \lambda \boldsymbol{I}_{K+d}) \widetilde{\boldsymbol{\theta}}^* + \sum_{t \in \mathcal{E}_s} \widetilde{\boldsymbol{x}}_{A_t,t} \eta_{A_t,t} \right] - \widetilde{\boldsymbol{\theta}}^* \\ &= \boldsymbol{V}_s^{-1} \sum_{t \in \mathcal{E}_s} \widetilde{\boldsymbol{x}}_{A_t,t} \eta_{A_t,t} - \lambda \boldsymbol{V}_s^{-1} \widetilde{\boldsymbol{\theta}}^*. \end{aligned}$$

We apply the self-normalized bound (Theorem 1) of Abbasi-Yadkori et al. (2011) with failure probability $\delta/(8E_s)$ and the triangle inequality.

$$\begin{aligned} \|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\|_{\boldsymbol{V}_s} &\le \left\| \sum_{t \in \mathcal{E}_s} \widetilde{\boldsymbol{x}}_{A_t,t} \eta_{A_t,t} \right\|_{\boldsymbol{V}_s^{-1}} + \lambda \|\widetilde{\boldsymbol{\theta}}^*\|_{\boldsymbol{V}_s^{-1}} \\ &\le \sigma \sqrt{2\log(8E_s/\delta) + \log \frac{\det(\boldsymbol{V}_s)}{\lambda^{K+d}}} + \sqrt{\lambda} \|\widetilde{\boldsymbol{\theta}}^*\| \end{aligned}$$

$$\leq \sigma\sqrt{2\log(8E_s/\delta) + (K+d)\log\left(1+\frac{2E_s(K+1)}{K+d}\right)} + 1 = \beta_s.$$

The last line uses Lemma 11, $\|\widetilde{\boldsymbol{x}}_{A_t,t}\|^2 \leq 2$, and $\lambda\|\widetilde{\boldsymbol{\theta}}^*\|^2 \leq 1$. The Rayleigh quotient gives

$$\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\| \leq \frac{\|\widehat{\widetilde{\boldsymbol{\theta}}}_s - \widetilde{\boldsymbol{\theta}}^*\|_{\boldsymbol{V}_s}}{\sqrt{\lambda_{\min}(\boldsymbol{V}_s)}} \leq r_s.$$

In particular, the last d coordinates satisfy $\|\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^*\| \leq r_s$.

Independently of this noise argument, Lemma 1 gives, with failure probability at most $\delta/(8E_s)$,

$$\lambda_{\min}(\boldsymbol{V}_s) \geq \lambda + E_s\widetilde{\rho}/2 \quad \text{if} \quad E_s \geq \frac{4}{\widetilde{\rho}\log(e/2)}\log\frac{8(K+d)E_s}{\delta}. \tag{17}$$

Its conditions hold because the block calendar is deterministic, contexts are independent of the past, and the arm choices are independent uniform draws. For a phase not meeting the displayed sample-size condition, no Gram lower bound is needed.

We compare these phase quantities with Equation (9). If the endpoint has been reached by n, then $E_s \leq n$. Bernoulli's inequality with exponent $(K+d)/2 \geq 1$ gives

$$(K+d)\log\left(1+\frac{2n(K+1)}{K+d}\right) \geq 2\log(1+n(K+1)) \geq 2\log(2E_s).$$

Therefore,

$$\begin{aligned}
&2\log(8E_s/\delta) + (K+d)\log\left(1+\frac{2E_s(K+1)}{K+d}\right) \\
&\quad\leq 2\log(4/\delta) + 2(K+d)\log\left(1+\frac{2n(K+1)}{K+d}\right) \\
&\quad\leq 2\left[2\log(4/\delta) + (K+d)\log\left(1+\frac{2n(K+1)}{K+d}\right)\right],
\end{aligned}$$

so $\beta_s \leq \sqrt{2}\,\overline{\beta}_n$. Also, $2E_s \leq n(n+1)$ implies

$$\log\frac{8(K+d)E_s}{\delta} \leq \log\frac{4(K+d)n(n+1)}{\delta}.$$

If $E_s \geq A_n$, its definition implies the sample-size condition in Equation (17), since $16 > 4/\log(e/2)$. On the resulting Gram event,

$$r_s^2 \leq \frac{2\overline{\beta}_n^2}{\lambda + E_s\widetilde{\rho}/2} < \frac{4\overline{\beta}_n^2}{E_s\widetilde{\rho}} \leq \frac{A_n}{4E_s}. \tag{18}$$

If j is exact, projection contraction gives

$$\|(\boldsymbol{I} - \boldsymbol{\Pi}_j)\widehat{\boldsymbol{\theta}}_s\| = \|(\boldsymbol{I} - \boldsymbol{\Pi}_j)(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^*)\| \leq r_s.$$

Hence every such covered model passes. For an underfitted $j < \ell^*$, nestedness gives $\gamma_j \geq \gamma_{\ell^*-1}$. Indeed,

$$(\boldsymbol{I} - \boldsymbol{\Pi}_j)\boldsymbol{\theta}^* = (\boldsymbol{I} - \boldsymbol{\Pi}_{\ell^*-1})\boldsymbol{\theta}^* + (\boldsymbol{\Pi}_{\ell^*-1} - \boldsymbol{\Pi}_j)\boldsymbol{\theta}^*,$$

and the two terms are orthogonal, so their squared norms add. If that model passes, the triangle inequality consequently gives

$$\gamma_{\ell^*-1} \leq \gamma_j \leq \|(\boldsymbol{I} - \boldsymbol{\Pi}_j)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_s)\| + \|(\boldsymbol{I} - \boldsymbol{\Pi}_j)\widehat{\boldsymbol{\theta}}_s\| \leq 2r_s.$$

At $E_s \geq A_n/\gamma_{\ell^*-1}^2$, Equation (18) instead gives $2r_s < \gamma_{\ell^*-1}$. No underfitted candidate then passes. The default candidate in Line 11 of Algorithm 2 is also exact at this threshold. To see this, taking traces in Assumption 1 gives $d\rho_{\min} \leq 1$, whence

$$A_n \geq 16\max\{K, d\} \geq 8(K+d).$$

Since the largest cover threshold is E_s and $\gamma_{\ell^*-1} \le 1$, the full model is covered at the stated threshold and passes with zero residual.

Each phase used at most two events of failure $\delta/(8E_s)$. Their union has failure probability at most

$$2\sum_{s\ge 1}\frac{\delta}{8E_s} = \frac{\delta}{4K}\sum_{s\ge 1}2^{-(s-1)} = \frac{\delta}{2K} \le \frac{\delta}{4}.$$

The Gram events depend only on deterministic phase sample conditions, so their implications for every n require no additional union over horizons. □

Theorem 4 (SSC−UCB)**.** *Under Assumption 1, there is a universal constant C such that SSC−UCB satisfies the following bounds for all time horizon $n \ge 1$ with probability at least $1-\delta$.*

$$R_n \le C\Phi_n\left[\sqrt{m_n}(K+d_{\ell^*})\sqrt{n/K} + m_n^{3/2}\sqrt{(K+d_{\ell^*})n}\right] = \widetilde{\mathcal{O}}\left(\frac{K+d_{\ell^*}}{\sqrt{K}}\sqrt{n}\right). \tag{5}$$

If $\ell^* \ge 1$, it also satisfies

$$\begin{aligned} R_n &\le Cm_n\Phi_n\sqrt{(K+d_{\ell^*})n} + C\Phi_n\left[\frac{\sqrt{m_n}(K+d_{\ell^*})}{K} + m_n^{3/2}\sqrt{\frac{K+d_{\ell^*}}{K}}\right]\frac{A_n}{\gamma^2_{\ell^*-1}} \\ &= \widetilde{\mathcal{O}}\left(\sqrt{(K+d_{\ell^*})n} + \frac{(K+d_{\ell^*})(K+d)}{K\widetilde{\rho}\gamma^2_{\ell^*-1}}\right). \end{aligned} \tag{6}$$

If $\ell^* = 0$, it also satisfies

$$R_n \le Cm_n\Phi_n\sqrt{Kn} = \widetilde{\mathcal{O}}\left(\sqrt{Kn}\right) \tag{7}$$

Proof. Write $p_j = K + d_j$, $p_* = K + d_{\ell^*}$, and $J_s = |\mathcal{J}_s|$. Let $R_s(a)$ be the regret over the first a exploitation rounds of phase s.

Let q^* be the smallest nonnegative integer with $K2^{q^*} \ge p_*$, and define the reference model

$$\bar{j} = \max\{j : p_j \le K2^{q^*}\}.$$

Then

$$p_* \le p_{\bar{j}} < 2p_*. \tag{19}$$

The reference is exact and belongs to $\mathcal{J}_s$ precisely when $s \ge q^*+1$. For the simple model it is level zero. On the batch confidence event it passes whenever covered. Before it is covered, $E_s < p_*$ and $H_s = E_s^2/K$, so for an exploitation prefix $a \le H_s$,

$$R_s(a) \le 4a \le 4\sqrt{H_s a} \le 4p_*\sqrt{a/K}. \tag{20}$$

A positive candidate that first appears at threshold $K2^q$ has $q \ge 1$ and $p_j > K2^{q-1}$. Otherwise it would already be eligible at the preceding threshold. Distinct candidates have distinct first thresholds. Therefore every $\mathcal{I} \subseteq \mathcal{J}_s$ satisfies

$$\sum_{j\in\mathcal{I}}\frac{1}{K+p_j} \le \frac{1}{2K} + \frac{1}{2K} + \frac{1}{K}\sum_{q=2}^{\infty}2^{-(q-1)} = \frac{2}{K}. \tag{21}$$

The first two terms cover level zero and a candidate first appearing at $q = 1$.

For the reward comparisons, define $u_s = \log(1/\zeta_s)$. Fix a phase and abbreviate $H = H_s$, $\zeta = \zeta_s$, and $u = u_s$. An epoch starts with the initialization in Lines 12 to 13 or with the learner and statistic reset in Line 24. At an actual epoch start, condition on its preceding history. The candidate set $\mathcal{I}$, fresh learner states, and probabilities for selecting learners from Line 15,

$$q_j = \frac{1}{(K+p_j)Z}, \qquad Z = \sum_{i\in\mathcal{I}}\frac{1}{K+p_i}$$

are then fixed. For analysis, continue these learners for H rounds without deletions, independently selecting learner j with probability q_j at each round. The actual epoch is a prefix of this continuation.

Let $t \le H$ count calendar rounds in this continuation. Write $m_v^* = \max_i(\mu_i + \boldsymbol{\alpha}_{i,v}^\top \boldsymbol{\theta}^*)$, where the round index v is local to the continuation, and $\bar{m}_t = t^{-1}\sum_{v=1}^t m_v^*$. Let $R_j(t)$ be regret on candidate j's calls and $N_j(t)$ its call count. Let j_v denote the learner selected at round v. For positive counts, we will use

$$\left|\frac{S_j + R_j(t)}{N_j(t)} - \bar{m}_t\right| \le c_j := 4(1+\sigma)\sqrt{\frac{u}{N_j(t)}}. \tag{22}$$

Fix $t \le H$ and the context vectors $\boldsymbol{\alpha}_{i,v}$ for all $i \in [K]$ and $v \le t$, so that $m_1^*, \dots, m_t^* \in [-2, 2]$ and $\bar{m}_t$ are fixed. Conditional on learner j being selected $a > 0$ times, each subset of a rounds is equally likely because the learner is selected independently with the same probability at every round. Hoeffding's inequality for sampling without replacement (Hoeffding, 1963, Theorem 4) therefore gives, with conditional probability at least $1 - 2e^{-u}$,

$$\left|\frac{1}{a}\sum_{v \le t:\, j_v = j} m_v^* - \bar{m}_t\right| \le 2\sqrt{2u/a}.$$

Next, also fix which learner is selected at each round, so the a rounds assigned to learner j are fixed. The reward noises are independent of these context vectors and learner selections. Each arm is chosen before its reward noise is observed, so these a observed noises remain conditionally zero mean and σ-sub-Gaussian. The sub-Gaussian concentration inequality gives, with conditional probability at least $1 - 2e^{-u}$,

$$\left|\sum_{v \le t:\, j_v = j} \eta_{A_v, v}\right| \le \sigma\sqrt{2au}.$$

Averaging the two conditional failure bounds and taking a union bound shows that both inequalities hold whenever $N_j(t) > 0$, with failure probability at most $4e^{-u}$ for each fixed t. Using

$$S_j + R_j(t) = \sum_{v \le t:\, j_v = j} (m_v^* + \eta_{A_v, v}),$$

the two deviations sum to at most $(2+\sigma)\sqrt{2u/a} \le 4(1+\sigma)\sqrt{u/a}$. A union bound over $t = 1, \dots, H$ proves Equation (22) with failure probability at most $4H\zeta$ per candidate.

For $t \ge 1$, since $N_j(t) \sim \mathrm{Bin}(t, q_j)$, write $\mu = q_j t$. The bound $\log \mathbb{E}[e^{zN_j(t)}] \le \mu(e^z - 1)$ and exponential Markov's inequality give

$$\begin{aligned} \Pr(N_j(t) \ge a) &\le \exp(-a\log(a/\mu) + a - \mu), \qquad a \ge \mu, \\ \Pr(N_j(t) \le a) &\le \exp(-a\log(a/\mu) + a - \mu), \qquad 0 \le a \le \mu. \end{aligned}$$

Here $0 \log 0 = 0$, and the exponent follows by taking $z = \log(a/\mu)$, or its limit at $a = 0$. For $v = \sqrt{a/\mu}$,

$$a\log(a/\mu) - a + \mu - (\sqrt{a} - \sqrt{\mu})^2 = 2\mu v(v\log v - v + 1) \ge 0,$$

since $v \log v - v + 1$ has its minimum zero at $v = 1$. Applying these tails at $(\sqrt{\mu} + \sqrt{u})^2$ and, when $\mu > u$, at $(\sqrt{\mu} - \sqrt{u})^2$, gives failure at most $2e^{-u}$. Thus, with failure at most $2H\zeta$ per candidate, all calendar prefixes satisfy

$$\left|\sqrt{N_j(t)} - \sqrt{q_j t}\right| \le \sqrt{u}. \tag{23}$$

These inequalities include zero counts.

For an exact candidate, apply Lemma 6 conditionally on the epoch's starting history. The rounds assigned to this learner are independent of future contexts and noises, and all arms are offered, so the comparator is the global optimum. Since $H \ge K$ and $\zeta = \delta/(16H^2) < 1/(16H^2)$, Equation (12) gives, with conditional failure at most ζ,

$$R_j(t) \le \min\{4N_j(t), 22(1+\sigma)u\sqrt{p_j N_j(t)}\}, \qquad 0 \le t \le H.$$

Each deletion removes at least one candidate, and a sampled candidate with maximum empirical mean always remains. There are consequently at most J_s epochs in a phase on every trajectory.

Apply the preceding conditional bounds at each actual epoch start, with a vacuous success event if that start is not reached. Taking expectations and a union bound gives phase failure probability at most $(6H_s+1)J_s^2\zeta_s$. A union bound over these events in all phases gives failure probability at most

$$\sum_{s\geq 1}(6H_s+1)J_s^2\zeta_s \leq \frac{\delta}{2K}\sum_{s\geq 1}\frac{s^2}{4^{s-1}} = \frac{40\delta}{27K} \leq \frac{20\delta}{27} < \frac{3\delta}{4}.$$

Here we used $J_s \leq s$, $H_s = K4^{s-1} \geq 2$, $\zeta_s = \delta/(16H_s^2)$, and $\sum_{s\geq 1} s^2x^{s-1} = (1+x)/(1-x)^3$ at $x = 1/4$. Lemma 9 contributes at most $\delta/4$, so the total failure probability is less than δ. For the rest of the proof, work on the intersection of these events, which cover every phase and every prefix.

We next bound regret within an epoch. For positive counts write the comparison width and proof-only endpoints as

$$w_j = 24(1+\sigma)u\sqrt{p_j/N_j}, \qquad L_j = S_j/N_j - c_j, \qquad U_j = S_j/N_j + w_j,$$

and put $L_j = -\infty$, $U_j = +\infty$ when $N_j = 0$. The sampled-pair comparison in Lines 19 to 21 deletes j exactly when $U_j < \max_i L_i$. Since $p_j \geq 2$ and $u > 4$,

$$c_j \leq 2(1+\sigma)u\sqrt{p_j/N_j} \leq w_j.$$

Equation (22) implies $L_j \leq \bar{m}_t$ for every sampled candidate, since its regret is nonnegative. For a sampled exact candidate, Equation (12) gives

$$\frac{R_j(t)}{N_j} + c_j \leq 24(1+\sigma)u\sqrt{p_j/N_j} \leq w_j.$$

Consequently,

$$U_j \geq \bar{m}_t - \frac{R_j(t)}{N_j} - c_j + w_j \geq \bar{m}_t.$$

An unsampled candidate has an infinite upper endpoint. Thus no exact candidate is rejected.

Write $M = |\mathcal{I}|$ and $R_{\text{epoch}}(t) = \sum_j R_j(t)$. For nonnegative a, b, N, μ with $\sqrt{N} \leq \sqrt{\mu} + \sqrt{u}$,

$$\min\{bN, a\sqrt{N}\} = \sqrt{N}\min\{b\sqrt{N}, a\} \leq a\sqrt{\mu} + b\sqrt{uN}. \tag{24}$$

This follows by applying the bound on $\sqrt{N}$ to the first factor and bounding the second factor by a and $b\sqrt{N}$, respectively. Since $p_jq_j \leq 1/Z$ and $\sum_j \sqrt{N_j(t)} \leq \sqrt{Mt}$, Equations (23) and (24) give, when every candidate is exact,

$$\begin{aligned} R_{\text{epoch}}(t) &\leq 22(1+\sigma)uM\sqrt{t/Z} + 4\sqrt{uMt} \\ &\leq 24(1+\sigma)uM\sqrt{t/Z}. \end{aligned} \tag{25}$$

The last step uses $Z \leq M/(2K)$, $K \geq 2$, and $u > 4$.

Suppose the exact reference $\bar{j}$ is present. Consider any prefix with no deletion, excluding a deletion-triggering final observation. At such a prefix every active candidate satisfies $U_j \geq L_{\bar{j}}$. For $N_j, N_{\bar{j}} > 0$, Equation (22) and this comparison imply

$$\frac{R_j(t)}{N_j} \leq \bar{m}_t - \frac{S_j}{N_j} + c_j \leq \frac{R_{\bar{j}}(t)}{N_{\bar{j}}} + 2c_{\bar{j}} + w_j + c_j.$$

Set $B = \min\{4, R_{\bar{j}}(t)/N_{\bar{j}} + 2c_{\bar{j}}\}$ when $N_{\bar{j}} > 0$, and $B = 4$ otherwise. Combining the preceding inequality with $R_j(t) \leq 4N_j$ gives

$$R_{\text{epoch}}(t) \leq tB + \sum_{j:N_j>0}\min\{4N_j, (w_j+c_j)N_j\}. \tag{26}$$

Candidates with $N_j = 0$ contribute zero regret. For the reference term, Equation (23) and $B \leq 4$ give

$$B\sqrt{q_{\bar{j}}t} \leq B\sqrt{N_{\bar{j}}} + 4\sqrt{u}$$

$$\leq 22(1+\sigma)u\sqrt{p_{\bar{j}}} + (12+8\sigma)\sqrt{u}.$$

For $N_{\bar{j}} > 0$, the last step uses the exact-model regret bound and $2c_{\bar{j}}\sqrt{N_{\bar{j}}} = 8(1+\sigma)\sqrt{u}$. For $N_{\bar{j}} = 0$, it follows directly from $q_{\bar{j}}t \leq u$. For the sum in Equation (26), apply Equation (24) to $\min\{4N_j, w_jN_j\}$, then add c_jN_j. This yields

$$\sum_{j:N_j>0} \min\{4N_j, (w_j+c_j)N_j\} \leq 24(1+\sigma)uM\sqrt{t/Z} + 4(2+\sigma)\sqrt{uMt}.$$

Since $q_{\bar{j}}^{-1} = (K+p_{\bar{j}})Z$, combining these bounds gives

$$\begin{aligned} R_{\text{epoch}}(t) \leq (1+\sigma)u\left[\frac{24M}{\sqrt{Z}} + 22\sqrt{p_{\bar{j}}(K+p_{\bar{j}})Z}\right]\sqrt{t} \\ + 4(2+\sigma)\sqrt{uMt} + (12+8\sigma)\sqrt{u(K+p_{\bar{j}})Zt}. \end{aligned} \tag{27}$$

To sum over epochs, use $M \leq J_s$ and $1/(K+p_{\bar{j}}) \leq Z \leq 2/K$, which follow from reference survival and Equation (21). For a phase prefix of a exploitation rounds, remove the observation ending each completed epoch by a deletion and charge at most four regret for each removed observation. There are at most J_s remaining epoch prefixes. If their lengths are $t_1, \ldots, t_r$, Cauchy-Schwarz gives

$$\sum_{i=1}^{r}\sqrt{t_i} \leq \sqrt{r\sum_{i=1}^{r}t_i} \leq \sqrt{J_sa}.$$

The removed observations cost at most $4\min\{J_s, a\} \leq 4\sqrt{J_sa}$. Equation (27), including this cost, yields

$$R_s(a) \leq (1+\sigma)u_s\left[\begin{array}{l} 24J_s\sqrt{K+p_{\bar{j}}} + 22\sqrt{\dfrac{2p_{\bar{j}}(K+p_{\bar{j}})}{K}} \\ + 4\sqrt{J_s} + 6\sqrt{\dfrac{2(K+p_{\bar{j}})}{K}} + \dfrac{4}{(1+\sigma)u_s} \end{array}\right]\sqrt{J_sa}.$$

Here $4(2+\sigma)/[(1+\sigma)\sqrt{u_s}] \leq 4$ and $(12+8\sigma)/[(1+\sigma)\sqrt{u_s}] \leq 6$. If $\ell^* \geq 1$ and the reference is covered, then $J_s \geq 2$, since the cover contains both level zero and the positive-level reference. Using $p_{\bar{j}} < 2p_*$, $p_* \geq K \geq 2$, and $\sqrt{p_{\bar{j}}(K+p_{\bar{j}})} \leq p_{\bar{j}} + K/2$, we have

$$\begin{aligned} 24J_s\sqrt{K+p_{\bar{j}}} &\leq 24\sqrt{3}\,J_s\sqrt{p_*}, \\ 22\sqrt{\frac{2p_{\bar{j}}(K+p_{\bar{j}})}{K}} &\leq 44\sqrt{2}\,\frac{p_*}{\sqrt{K}} + 11\sqrt{2K} \leq 44\sqrt{2}\,\frac{p_*}{\sqrt{K}} + 8J_s\sqrt{p_*}, \\ 4\sqrt{J_s} + 6\sqrt{\frac{2(K+p_{\bar{j}})}{K}} + \frac{4}{(1+\sigma)u_s} &\leq (2+3\sqrt{3}+1)J_s\sqrt{p_*} < 9J_s\sqrt{p_*}. \end{aligned}$$

Since $24\sqrt{3}+17 < 64$ and $44\sqrt{2} < 64$,

$$R_s(a) \leq 64(1+\sigma)u_s\left[\frac{\sqrt{J_s}\,p_*}{\sqrt{K}} + J_s^{3/2}\sqrt{p_*}\right]\sqrt{a}. \tag{28}$$

If all candidates initially are exact, no deletion occurs. There is then a single epoch, so Equation (25) and $Z \geq 1/(K+p_{\bar{j}})$ give

$$R_s(a) \leq 24(1+\sigma)u_sJ_s\sqrt{(K+p_{\bar{j}})a}. \tag{29}$$

We now sum over phases for any horizon $n \geq 1$. Let $N_{\text{exp}}(n)$ be the number of uniform rounds through n, and sum only phases whose exploitation has begun. Let $a_s \leq H_s$ be the number of exploitation rounds executed in phase s through horizon n. Lemma 8 gives

$$N_{\text{exp}}(n) \leq 4\sqrt{Kn}.$$

We next bound the confidence factors. For a reached exploitation phase, $s \le 2+\log(n+1) \le \Lambda_n$ if $s \ge 2$, using $n \ge K4^{s-2}$. The case $s = 1$ follows from $\Lambda_n > 1$. This proves $m_n \le \min\{L+1, \Lambda_n\}$. The uniform block of phase s ends at the deterministic round

$$t_s = \sum_{r<s}(E_r + H_r) + E_s = 2E_s - K + \frac{H_s - K}{3} > \frac{H_s}{3}.$$

Hence $a_s > 0$ implies $H_s < 3n$, and

$$u_s - \Lambda_n = \log \frac{H_s^2}{4K(n+1)^2} < \log \frac{9}{4K} < 1. \tag{30}$$

If the exploitation block is completed, $H_s \le n$ instead gives $u_s \le \Lambda_n - \log(4K) \le \Lambda_n - 2$. Let S be the largest index with $a_S > 0$, and put $b = \sum_{s<S} H_s$. All preceding exploitation blocks are complete, and

$$\left(\sum_{s<S}\sqrt{H_s}\right)^2 = K(2^{S-1}-1)^2 \le K(4^{S-1}-1) = 3b.$$

Since $b + a_S \le n$, Cauchy-Schwarz gives

$$\begin{aligned}\sum_s u_s\sqrt{a_s} &\le (\Lambda_n - 2)\sqrt{3b} + (\Lambda_n + 1)\sqrt{a_S} \\ &\le \sqrt{3(\Lambda_n-2)^2 + (\Lambda_n+1)^2}\sqrt{n} \le (2\Lambda_n - 1)\sqrt{n}.\end{aligned} \tag{31}$$

The last inequality follows from $\Lambda_n > 2$ and $3(\Lambda_n - 2)^2 + (\Lambda_n + 1)^2 = (2\Lambda_n - 1)^2 - 6(\Lambda_n - 2)$.

For $v > 0$, let r be the largest reached exploitation phase with $E_r < v$. Then $\sqrt{a_s} \le E_s/\sqrt{K}$, $u_s \le u_r - \log 16$ for $s < r$, and $\sum_{s<r} E_s = E_r - K$. Consequently,

$$\begin{aligned}\sqrt{K}\sum_{s:E_s<v} u_s\sqrt{a_s} &\le \sum_{s=1}^{r} u_s E_s \\ &\le u_r E_r + (u_r - \log 16)(E_r - K) \\ &\le (2u_r - \log 16)E_r \le (2\Lambda_n + 2 - \log 16)E_r < 2\Lambda_n v.\end{aligned} \tag{32}$$

Here $u_r > \log 16$, Equation (30) applies because phase r has been reached, and $\log 16 > 2$. If the relevant set of phases is empty, its sum is zero and the corresponding bound holds directly.

The simple model. All candidates are exact and level zero passes every batch test. There is one epoch in each phase, and Equation (29) applies with $p_{\bar{j}} = K$. Uniform rounds cost at most two each. Consequently,

$$\begin{aligned}R_n^s &\le 2N_{\exp}(n) + 24\sqrt{2}(1+\sigma)m_n\sqrt{K}\sum_s u_s\sqrt{a_s} \\ &\le \left[8 + 24\sqrt{2}(1+\sigma)m_n(2\Lambda_n - 1)\right]\sqrt{Kn} \le 48\sqrt{2}\, m_n\Phi_n\sqrt{Kn}.\end{aligned}$$

The last step uses $24\sqrt{2}(1+\sigma)m_n \ge 8$.

The separation-free bound. Suppose $\ell^* \ge 1$. When the reference is available, Equation (28) and $J_s \le m_n$ imply

$$R_s(a) \le 64(1+\sigma)u_s\left(\frac{\sqrt{m_n}\,p_*}{\sqrt{K}} + m_n^{3/2}\sqrt{p_*}\right)\sqrt{a}.$$

Before the reference is available, Equation (20) implies the same bound. Summing over the actual prefixes using Equation (31) and charging uniform rounds at most four each gives

$$\begin{aligned}R_n &\le 16\sqrt{Kn} + 64(1+\sigma)(2\Lambda_n - 1)\left[\frac{\sqrt{m_n}\,p_*}{\sqrt{K}} + m_n^{3/2}\sqrt{p_*}\right]\sqrt{n} \\ &\le 128\Phi_n\left[\sqrt{m_n}\,p_*\sqrt{n/K} + m_n^{3/2}\sqrt{p_* n}\right].\end{aligned}$$

The bracket in the first line is at least $2\sqrt{K}$, so the term subtracted through $2\Lambda_n - 1$ absorbs $16\sqrt{Kn}$. For $\ell^* = 0$, the preceding simple-model bound also implies this inequality, since $\sqrt{m_n} + m_n^{3/2} \geq m_n$.

The separation-dependent bound. Assume $\ell^* \geq 1$, and put $\gamma = \gamma_{\ell^*-1}$. Apply the same phase bound to phases with $E_s < A_n/\gamma^2$. Equation (32) bounds their total exploitation regret by

$$128\Phi_n \left[\frac{\sqrt{m_n}\, p_*}{K} + m_n^{3/2} \sqrt{\frac{p_*}{K}} \right] \frac{A_n}{\gamma^2}.$$

In every remaining phase, Lemma 9 ensures that all candidates are exact and the reference is available. The sharper bound in Equation (29) gives, including all uniform exploration,

$$\begin{aligned} 4N_{\exp}(n) + \sum_{s:\, E_s \geq A_n/\gamma^2} R_s(a_s) &\leq 16\sqrt{Kn} + 24\sqrt{3}(1+\sigma) m_n (2\Lambda_n - 1)\sqrt{p_* n} \\ &\leq 48\sqrt{3}\, m_n \Phi_n \sqrt{p_* n}. \end{aligned}$$

Here $K + p_{\bar{j}} \leq 3p_*$, and $24\sqrt{3}(1+\sigma) m_n \sqrt{p_*} \geq 16\sqrt{K}$ absorbs uniform exploration. All three conclusions hold with the common explicit choice $C = 128$. □

E Proof of Theorem 5

Theorem 5 (Lower bounds). *There exists a universal constant $c_{\mathrm{lb}} > 0$ such that the following statements hold. If $K \geq 2$ and $n \geq K$, then for every policy π, there exists an instance satisfying Assumption 1 with $\boldsymbol{\theta}^* = 0$ such that*

$$\mathbb{E}[R_n] \geq c_{\mathrm{lb}} \sqrt{Kn}.$$

For any $\ell^* \geq 1$ with $d_{\ell^*} \geq 2$, $K \geq 2$, and $n \geq \max\{K, d_{\ell^*}^2\}$, for every policy π, there exists an instance satisfying Assumption 1 with $\boldsymbol{\theta}^* \in V_{\ell^*} \setminus V_{\ell^*-1}$, such that

$$\mathbb{E}[R_n] \geq c_{\mathrm{lb}} \sqrt{(K + d_{\ell^*})n}.$$

Proof. All lower-bound instances below use Gaussian noise with variance one. When $\boldsymbol{\theta}^* = 0$, the problem reduces to the standard K-armed stochastic bandits, and therefore the proof follows from the standard minimax lower bound argument (Lattimore & Szepesvári, 2020). Let $c_{\mathrm{MAB}} > 0$ be the universal constant in the standard minimax lower bound, we have for every policy there exists a simple-model instance with expected regret at least $c_{\mathrm{MAB}}\sqrt{Kn}$.

In the following, we focus on the case when $\ell^* \geq 1$. Fix an arbitrary policy π. All expectations below include the randomness of this fixed policy. We prove two separate lower bounds, one of order $\sqrt{Kn}$ and the other of order $\sqrt{d_{\ell^*} n}$.

The $\Omega(\sqrt{Kn})$ lower bound. Since the subspaces are strictly nested, there exists a unit vector $\boldsymbol{v} \in V_{\ell^*} \setminus V_{\ell^*-1}$. Set $\boldsymbol{\theta}^* = \varepsilon \boldsymbol{v}$ with $\varepsilon > 0$. Let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ be an orthonormal basis of $\mathbb{R}^d$. Draw each arm context $\boldsymbol{\alpha} \in \mathbb{R}^d$ uniformly from $\mathcal{S}_d = \{\pm \boldsymbol{e}_r : r \in [d]\}$. This gives

$$\mathbb{E}[\boldsymbol{\alpha}] = 0, \qquad \mathbb{E}[\boldsymbol{\alpha}\boldsymbol{\alpha}^\top] = \sum_{S=\pm 1} \sum_{r=1}^{d} \boldsymbol{e}_r \boldsymbol{e}_r^\top \cdot \frac{1}{2d} = \frac{\boldsymbol{I}_d}{d}, \qquad \|\boldsymbol{\alpha}\| = 1.$$

So the context satisfies Assumption 1. Let ν denote a problem instance with $\mu_i = 0$ for every arm $i \in [K]$. For each arm $j \in [K]$, let ν'_j denote an alternative problem instance with $\mu_j = \Delta$ and $\mu_i = 0$ for all $i \neq j$. The parameter $\Delta > 0$ is chosen below and we take $0 < \varepsilon \leq \Delta/8$. Because $\|\boldsymbol{\alpha}_{i,t}\| \leq 1$, the contextual contribution can change the gap between two arms by at most 2ε, so the arm with bias Δ remains uniformly optimal up to a constant fraction of the gap. More explicitly, in alternative instance ν'_j, for every arm $i \neq j$ and every context realization,

$$\begin{aligned} &\mu_j + \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{j,t} \rangle - \mu_i - \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t} \rangle \\ &= \Delta + \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{j,t} \rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t} \rangle \end{aligned}$$

$$\geq \Delta - |\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{j,t}\rangle| - |\langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle|$$
$$\geq \Delta - 2\varepsilon \geq 3\Delta/4.$$

We now show that no policy can reliably distinguish which arm has intercept Δ when Δ is of order $\sqrt{K/n}$. Similar to Section 4.6 of Lattimore & Szepesvári (2020), we construct the canonical bandit model of our setting as follows. Let $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'_j}$ denote the probability measures induced by the reference instance ν and the alternative instance ν'_j, respectively. Both measures are defined on the history $H_n = \big(\{\boldsymbol{\alpha}_{i,t}\}_{i=1}^K, A_t, g_t\big)_{t=1}^n$ generated by the same policy π. Since each arm context is drawn uniformly from $\mathcal{S}_d$, for every feasible context $\{\boldsymbol{\alpha}_{i,t}\}_{i=1}^K$, its probability mass is

$$\mathbb{P}(\boldsymbol{\alpha}_{1:K,t}) = \prod_{i=1}^K \frac{1}{|\mathcal{S}_d|} = (2d)^{-K}.$$

For a realized feasible history h_n, define

$$m_{i,t}^{\nu} = \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle, \qquad m_{i,t}^{\nu'_j} = \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle + \Delta \cdot \mathbb{1}\{i = j\}.$$

Let $\pi_t(i \mid h_{t-1}, \boldsymbol{\alpha}_{1:K,t})$ be the policy's conditional probability of selecting arm i after observing the current contexts. Let $\mathcal{N}(m, 1)$ denote the one-dimensional Gaussian distribution with mean m and variance 1, and write φ_m for its density. By the chain-rule factorization of the joint likelihood, the likelihood of h_n under $\mathbb{P}_\nu$ is

$$L_\nu(h_n) = \prod_{t=1}^n (2d)^{-K} \pi_t(A_t \mid h_{t-1}, \boldsymbol{\alpha}_{1:K,t}) \varphi_{m_{A_t,t}^{\nu}}(g_t).$$

Similarly, the likelihood of the same feasible history under $\mathbb{P}_{\nu'_j}$ is

$$L_{\nu'_j}(h_n) = \prod_{t=1}^n (2d)^{-K} \pi_t(A_t \mid h_{t-1}, \boldsymbol{\alpha}_{1:K,t}) \varphi_{m_{A_t,t}^{\nu'_j}}(g_t).$$

The context mass $(2d)^{-K}$ and the policy factor are identical under the two instances, so only the reward density can differ. Therefore, by the definition of KL divergence, we have

$$D_{\mathrm{KL}}\Big(\mathbb{P}_\nu \,\Big\|\, \mathbb{P}_{\nu'_j}\Big) = \int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'_j}}\right) \mathrm{d}\mathbb{P}_\nu = \mathbb{E}_\nu\left[\log \frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'_j}}\right].$$

By the chain rule for Radon-Nikodym derivatives, we have

$$\log \frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'_j}}(H_n) = \log \frac{L_\nu(H_n)}{L_{\nu'_j}(H_n)} = \log\left(\prod_{t=1}^n \frac{\varphi_{m_{A_t,t}^{\nu}}(g_t)}{\varphi_{m_{A_t,t}^{\nu'_j}}(g_t)}\right) = \sum_{t=1}^n \log \frac{\varphi_{m_{A_t,t}^{\nu}}(g_t)}{\varphi_{m_{A_t,t}^{\nu'_j}}(g_t)}.$$

Let $\mathcal{G}_t = (H_{t-1}, \boldsymbol{\alpha}_{1:K,t}, A_t)$. Taking expectation under $\mathbb{P}_\nu$ and conditioning on $\mathcal{G}_t$ yields

$$\begin{aligned}
D_{\mathrm{KL}}\Big(\mathbb{P}_\nu \,\Big\|\, \mathbb{P}_{\nu'_j}\Big) &= \sum_{t=1}^n \mathbb{E}_\nu\left[\log \frac{\varphi_{m_{A_t,t}^{\nu}}(g_t)}{\varphi_{m_{A_t,t}^{\nu'_j}}(g_t)}\right] = \sum_{t=1}^n \mathbb{E}_\nu\left[\mathbb{E}_\nu\left[\log \frac{\varphi_{m_{A_t,t}^{\nu}}(g_t)}{\varphi_{m_{A_t,t}^{\nu'_j}}(g_t)}\,\middle|\,\mathcal{G}_t\right]\right] \\
&= \sum_{t=1}^n \mathbb{E}_\nu\Big[D_{\mathrm{KL}}\Big(\mathcal{N}(m_{A_t,t}^{\nu}, 1) \,\Big\|\, \mathcal{N}(m_{A_t,t}^{\nu'_j}, 1)\Big)\Big] \qquad (33) \\
&= \sum_{i=1}^K \mathbb{E}_\nu\left[\sum_{t=1}^n \Big(\mathbb{1}\{A_t = i\} D_{\mathrm{KL}}\Big(\mathcal{N}(m_{A_t,t}^{\nu}, 1) \,\Big\|\, \mathcal{N}(m_{A_t,t}^{\nu'_j}, 1)\Big)\Big)\right] \\
&= \mathbb{E}_\nu\left[\sum_{t=1}^n \big(\mathbb{1}\{A_t = j\} D_{\mathrm{KL}}\big(\mathcal{N}(m_{j,t}^{\nu}, 1) \,\big\|\, \mathcal{N}(m_{j,t}^{\nu} + \Delta, 1)\big)\big)\right] \\
&= \mathbb{E}_\nu[N_j] \cdot \frac{\Delta^2}{2}, \qquad (34)
\end{aligned}$$

where Equation (33) uses the definition of KL divergence, Equation (34) uses the fact that the KL divergence is non-zero only when $A_t = j$ and the Gaussian KL formula (Lemma 12), and $N_j = \sum_{t=1}^{n} \mathbb{1}\{A_t = j\}$ is the total number of pulls of arm j.

For each $j \in [K]$, define $R^{\pi}_{\nu'_j}(n)$ as the expected cumulative regret of policy π under the alternative instance ν'_j, then

$$
\begin{aligned}
R^{\pi}_{\nu'_j}(n) &\coloneqq \mathbb{E}_{\nu'_j}\left[\sum_{t=1}^{n}\left(\max_{i\in[K]}\{\mu_i + \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle\} - \mu_{A_t} - \langle \boldsymbol{\theta}^*, \boldsymbol{\alpha}_{A_t,t}\rangle\right)\right] \\
&\geq \frac{3\Delta}{4}\,\mathbb{E}_{\nu'_j}\left[\sum_{t=1}^{n}\mathbb{1}\{A_t \neq j\}\right] \qquad (35) \\
&= \frac{3\Delta}{4}\,\mathbb{E}_{\nu'_j}[n - N_j] \\
&\geq \frac{3\Delta n}{8}\,\mathrm{Pr}_{\nu'_j}\left[n - N_j \geq \frac{n}{2}\right] \qquad (36) \\
&\geq \frac{3\Delta n}{8}\,\mathrm{Pr}_{\nu'_j}\left[N_j < \frac{n}{2}\right].
\end{aligned}
$$

Here Equation (35) uses the fact that, under ν'_j, pulling any arm other than j incurs regret at least $3\Delta/4$. Equation (36) uses $X \geq a\mathbb{1}\{X \geq a\}$ for $X = n - N_j \geq 0$ and $a = n/2$. The last inequality uses $\{N_j < n/2\} \subseteq \{n - N_j \geq n/2\}$.

Averaging the above display over $j \in [K]$, we obtain

$$
\frac{1}{K}\sum_{j=1}^{K} R^{\pi}_{\nu'_j}(n) \geq \frac{3\Delta n}{8K}\sum_{j=1}^{K}\mathrm{Pr}_{\nu'_j}\left[N_j < \frac{n}{2}\right]. \qquad (37)
$$

For each $j \in [K]$, applying the Bretagnolle-Huber inequality (Lemma 13) to the event $\{N_j \geq n/2\}$,

$$
\mathrm{Pr}_{\nu}\left[N_j \geq \frac{n}{2}\right] + \mathrm{Pr}_{\nu'_j}\left[N_j < \frac{n}{2}\right] \geq \frac{1}{2}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{P}_{\nu}\,\middle\|\,\mathbb{P}_{\nu'_j}\right)\right). \qquad (38)
$$

Therefore,

$$
\begin{aligned}
\frac{1}{K}\sum_{j=1}^{K}\mathrm{Pr}_{\nu'_j}\left[N_j < \frac{n}{2}\right] &\geq \frac{1}{2K}\sum_{j=1}^{K}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{P}_{\nu}\,\middle\|\,\mathbb{P}_{\nu'_j}\right)\right) - \frac{1}{K}\sum_{j=1}^{K}\mathrm{Pr}_{\nu}\left[N_j \geq \frac{n}{2}\right] \\
&\geq \frac{1}{2}\exp\left(-\frac{1}{K}\sum_{j=1}^{K}D_{\mathrm{KL}}\left(\mathbb{P}_{\nu}\,\middle\|\,\mathbb{P}_{\nu'_j}\right)\right) - \frac{2}{Kn}\,\mathbb{E}_{\nu}\left[\sum_{j=1}^{K}N_j\right] \qquad (39) \\
&= \frac{1}{2}\exp\left(-\frac{n\Delta^2}{2K}\right) - \frac{2}{K}. \qquad (40)
\end{aligned}
$$

Here Equation (39) uses Jensen's inequality for the convex function e^{-x}, and Markov's inequality. The last equality uses $\sum_{j=1}^{K} N_j = n$ and the KL identity (Equation (34)).

Now choose $\Delta = c_\Delta\sqrt{\frac{K}{n}}$, where $c_\Delta \in (0, 1/2]$ is a universal constant. Since $n \geq K$, this choice gives $\Delta \leq c_\Delta \leq 1/2$. Let $K_0 = 8$. For $K \geq K_0$, Equation (40) gives

$$
\begin{aligned}
\frac{1}{K}\sum_{j=1}^{K}\mathrm{Pr}_{\nu'_j}\left[N_j < \frac{n}{2}\right] &\geq \frac{1}{2}\exp\left(-\frac{n\Delta^2}{2K}\right) - \frac{2}{K} \\
&= \frac{1}{2}\exp\left(-\frac{c_\Delta^2}{2}\right) - \frac{2}{K} \\
&\geq \frac{1}{2}\exp\left(-\frac{1}{8}\right) - \frac{1}{4} \triangleq c_{\mathrm{fail}}.
\end{aligned}
$$

Combining this bound with Equation (37), we obtain

$$\frac{1}{K}\sum_{j=1}^{K} R^{\pi}_{\nu'_j}(n) \geq \frac{3\Delta n}{8}\cdot\frac{1}{K}\sum_{j=1}^{K}\Pr_{\nu'_j}\left[N_j < \frac{n}{2}\right] \geq \frac{3c_{\mathrm{fail}}\Delta n}{8} = \frac{3c_{\mathrm{fail}}c_\Delta}{8}\sqrt{Kn}.$$

Therefore, for some $j \in [K]$,

$$R^{\pi}_{\nu'_j}(n) \geq \frac{3c_{\mathrm{fail}}c_\Delta}{8}\sqrt{Kn}.$$

It remains to handle the finitely many cases $2 \leq K < K_0$. For these cases, we use a two-arm subfamily. Keep the same context distribution and the same nonzero contextual parameter $\boldsymbol{\theta}^* = \varepsilon v \in V_{\ell^*} \setminus V_{\ell^*-1}$. Define two instances $\bar{\nu}_1$ and $\bar{\nu}_2$ as follows. In $\bar{\nu}_1$,

$$\mu_1 = \Delta_0, \qquad \mu_2 = 0, \qquad \mu_a = -1 \quad \text{for } a \geq 3,$$

while in $\bar{\nu}_2$,

$$\mu_1 = 0, \qquad \mu_2 = \Delta_0, \qquad \mu_a = -1 \quad \text{for } a \geq 3.$$

Here

$$\Delta_0 = c'_0\frac{1}{\sqrt{n}},$$

where $c'_0 \in (0, 1]$ is a sufficiently small universal constant, and we choose $0 < \varepsilon \leq \Delta_0/8$. Then, under $\bar{\nu}_1$, arm 1 is uniformly better than every other arm by at least $3\Delta_0/4$. Similarly, under $\bar{\nu}_2$, arm 2 is uniformly better than every other arm by at least $3\Delta_0/4$. Let N_1 denote the number of pulls of arm 1 up to time n.

Starting from the regret definition under $\bar{\nu}_1$, we have

$$\begin{aligned} R^{\pi}_{\bar{\nu}_1}(n) &= \mathbb{E}_{\bar{\nu}_1}\left[\sum_{t=1}^{n}\left(\max_{i\in[K]}\{\mu_i + \langle\boldsymbol{\theta}^*, \boldsymbol{\alpha}_{i,t}\rangle\} - \mu_{A_t} - \langle\boldsymbol{\theta}^*, \boldsymbol{\alpha}_{A_t,t}\rangle\right)\right] \\ &\geq \frac{3\Delta_0}{4}\,\mathbb{E}_{\bar{\nu}_1}\left[\sum_{t=1}^{n}\mathbb{1}\{A_t \neq 1\}\right] \\ &= \frac{3\Delta_0}{4}\,\mathbb{E}_{\bar{\nu}_1}[n - N_1] \\ &\geq \frac{3\Delta_0 n}{8}\Pr_{\bar{\nu}_1}\left[N_1 \leq \frac{n}{2}\right]. \end{aligned}$$

Similarly, under $\bar{\nu}_2$, pulling arm 1 is suboptimal, and hence

$$R^{\pi}_{\bar{\nu}_2}(n) \geq \frac{3\Delta_0 n}{8}\Pr_{\bar{\nu}_2}\left[N_1 > \frac{n}{2}\right].$$

Adding the two inequalities gives

$$\begin{aligned} R^{\pi}_{\bar{\nu}_1}(n) + R^{\pi}_{\bar{\nu}_2}(n) &\geq \frac{3\Delta_0 n}{8}\left(\Pr_{\bar{\nu}_1}\left[N_1 \leq \frac{n}{2}\right] + \Pr_{\bar{\nu}_2}\left[N_1 > \frac{n}{2}\right]\right) \\ &\geq \frac{3\Delta_0 n}{16}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{\bar{\nu}_1} \,\|\, \mathbb{P}_{\bar{\nu}_2})), \end{aligned} \tag{41}$$

where the last step uses the Bretagnolle–Huber inequality with the event $\{N_1 \leq n/2\}$.

We now bound the KL divergence. The two instances $\bar{\nu}_1$ and $\bar{\nu}_2$ have the same context distribution, the same policy kernel, and the same contextual parameter $\boldsymbol{\theta}^*$. Their reward distributions differ only when the learner pulls arm 1 or arm 2. For either of these two arms, the Gaussian mean differs by Δ_0. Therefore, by the same KL decomposition as in Equation (34),

$$D_{\mathrm{KL}}(\mathbb{P}_{\bar{\nu}_1} \,\|\, \mathbb{P}_{\bar{\nu}_2}) = \frac{\Delta_0^2}{2}\,\mathbb{E}_{\bar{\nu}_1}[N_1 + N_2] \leq \frac{n\Delta_0^2}{2} = \frac{(c'_0)^2}{2}.$$

Plugging this into Equation (41) yields

$$R^{\pi}_{\bar{\nu}_1}(n) + R^{\pi}_{\bar{\nu}_2}(n) \geq \frac{3\Delta_0 n}{16} \exp\left(-\frac{(c_0')^2}{2}\right).$$

Hence

$$\max\{R^{\pi}_{\bar{\nu}_1}(n), R^{\pi}_{\bar{\nu}_2}(n)\} \geq \frac{3\Delta_0 n}{32} \exp\left(-\frac{(c_0')^2}{2}\right) = \frac{3c_0'}{32} \exp\left(-\frac{(c_0')^2}{2}\right)\sqrt{n}.$$

Since $K < K_0$, we have $\sqrt{n} \geq K_0^{-1/2}\sqrt{Kn}$. Therefore,

$$\max\{R^{\pi}_{\bar{\nu}_1}(n), R^{\pi}_{\bar{\nu}_2}(n)\} \geq \frac{3c_0'}{32\sqrt{K_0}} \exp\left(-\frac{(c_0')^2}{2}\right)\sqrt{Kn}.$$

Combining the cases $K \geq K_0$ and $2 \leq K < K_0$, and defining

$$c_1 = \min\left\{\frac{3c_{\text{fail}}c_\Delta}{8}, \frac{3c_0'}{32\sqrt{K_0}} \exp\left(-\frac{(c_0')^2}{2}\right)\right\},$$

we conclude that there exists an instance with expected regret at least $c_1\sqrt{Kn}$.

The $\Omega(\sqrt{d_{\ell^*} n})$ lower bound. For convenience, let $m = d_{\ell^*}$. Choose an orthonormal basis so that $V_{\ell^*} = \text{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$ and $V_{\ell^*-1} \subseteq \text{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m-1})$. Consider arm intercepts $\mu_1 = \mu_2 = 0$ and $\mu_3 = \mu_4 = \cdots = \mu_K = -1$.

For every arm and round, draw contexts independently as follows.

$$\boldsymbol{\alpha} = \begin{cases} 0, & \text{with probability } 1/2 \\ \text{draw uniformly from } \{\pm\boldsymbol{e}_r : r \in [m]\}, & \text{with probability } 3/8 \\ \text{draw uniformly from } \{\pm\boldsymbol{e}_r : r \in [d]\}, & \text{with probability } 1/8. \end{cases}$$

So we have

$$\mathbb{E}[\boldsymbol{\alpha}] = 0, \qquad \|\boldsymbol{\alpha}\| \leq 1, \qquad \mathbb{E}[\boldsymbol{\alpha}\boldsymbol{\alpha}^\top] = \frac{3}{8m}\sum_{r=1}^{m} \boldsymbol{e}_r\boldsymbol{e}_r^\top + \frac{1}{8d}\boldsymbol{I}_d.$$

Thus Assumption 1 holds with $\rho_{\min} = 1/(8d)$.

For a sign vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m) \in \{-1, 1\}^m$, let $\varepsilon = \frac{m}{2\sqrt{n}}$ and define parameter

$$\boldsymbol{\theta}_{\boldsymbol{\omega}} = \frac{\varepsilon}{\sqrt{m}} \sum_{j=1}^{m} \omega_j \boldsymbol{e}_j.$$

Since the theorem assumes $n \geq m^2$, we have $\varepsilon \leq 1/2$. So $\|\boldsymbol{\theta}_{\boldsymbol{\omega}}\| = \varepsilon \leq 1/2$ and $\boldsymbol{\theta}_{\boldsymbol{\omega}} \in V_{\ell^*} \setminus V_{\ell^*-1}$.

For $j \in [m]$, let $\boldsymbol{\omega}^{(j)}$ be obtained from $\boldsymbol{\omega}$ by flipping coordinate j. Then

$$\boldsymbol{\theta}_{\boldsymbol{\omega}} - \boldsymbol{\theta}_{\boldsymbol{\omega}^{(j)}} = \frac{2\varepsilon\omega_j}{\sqrt{m}} \boldsymbol{e}_j.$$

We first bound the KL divergence between neighboring hypotheses $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^{(j)}$. Let $\mathbb{P}_{\boldsymbol{\omega}}$ denote the probability law induced by the policy under parameter $\boldsymbol{\theta}_{\boldsymbol{\omega}}$. The context distribution and the policy kernel are the same under $\mathbb{P}_{\boldsymbol{\omega}}$ and $\mathbb{P}_{\boldsymbol{\omega}^{(j)}}$. Only the reward density can differ. Hence, similar to the derivation of Equation (33), we have

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}}) &= \sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{\omega}}[D_{\text{KL}}(\mathcal{N}\left(\mu_{A_t} + \langle \boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{A_t,t}\rangle, 1\right) \| \mathcal{N}\left(\mu_{A_t} + \langle \boldsymbol{\theta}_{\boldsymbol{\omega}^{(j)}}, \boldsymbol{\alpha}_{A_t,t}\rangle, 1\right))] \\ &= \frac{1}{2}\sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{\omega}}\Big[\left(\langle \boldsymbol{\theta}_{\boldsymbol{\omega}} - \boldsymbol{\theta}_{\boldsymbol{\omega}^{(j)}}, \boldsymbol{\alpha}_{A_t,t}\rangle\right)^2\Big] \\ &= \frac{1}{2}\sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{\omega}}\Big[\mathbb{1}\{(\boldsymbol{\alpha}_{A_t,t})_j \neq 0\}\left(\langle \boldsymbol{\theta}_{\boldsymbol{\omega}} - \boldsymbol{\theta}_{\boldsymbol{\omega}^{(j)}}, \boldsymbol{\alpha}_{A_t,t}\rangle\right)^2\Big] \end{aligned} \tag{42}$$

$$= \frac{1}{2}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{\boldsymbol{\omega}}\left[\mathbb{1}\{(\boldsymbol{\alpha}_{A_t,t})_j \neq 0\}\left(\frac{2\varepsilon\omega_j}{\sqrt{m}}(\boldsymbol{\alpha}_{A_t,t})_j\right)^2\right]$$
$$= \frac{2\varepsilon^2}{m}\mathop{\mathbb{E}}_{\boldsymbol{\omega}}[N_j]. \tag{43}$$

where Equation (42) uses the fact that if $(\boldsymbol{\alpha}_{A_t,t})_j = 0$, the inner product $\langle \boldsymbol{\theta}_{\boldsymbol{\omega}} - \boldsymbol{\theta}_{\boldsymbol{\omega}^{(j)}}, \boldsymbol{\alpha}_{A_t,t}\rangle$ is zero, and in Equation (43), we define $N_j = \sum_{t=1}^{n}\mathbb{1}\{(\boldsymbol{\alpha}_{A_t,t})_j \neq 0\}$ as the number of rounds on which the selected arm context has a nonzero j-th coordinate.

Averaging Equation (43) over a uniform $\boldsymbol{\omega} \in \{-1,1\}^m$ and a uniform coordinate $j \in [m]$, we get

$$\frac{1}{2^m m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{j=1}^{m} D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}}) \leq \frac{2\varepsilon^2}{m}\cdot\frac{1}{2^m m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{j=1}^{m}\mathop{\mathbb{E}}_{\boldsymbol{\omega}}[N_j]$$
$$\leq \frac{2\varepsilon^2}{m}\cdot\frac{n}{m} = \frac{2n\varepsilon^2}{m^2} = \frac{1}{2}. \tag{44}$$

where in Equation (44), we use the fact that for every time step t, the selected context vector $\boldsymbol{\alpha}_{A_t,t}$ has at most one nonzero coordinate among the first m coordinates, thus $\sum_{j=1}^{m} N_j = \sum_{t=1}^{n}\sum_{j=1}^{m}\mathbb{1}\{(\boldsymbol{\alpha}_{A_t,t})_j \neq 0\} \leq n$.

Next we define the diagnostic events that convert sign-estimation error into regret. For each round t and coordinate $j \in [m]$, let $D_{t,j}$ be the event that arm 1 has context $\boldsymbol{e}_j$, and arm 2 has context 0:

$$D_{t,j} = \{\boldsymbol{\alpha}_{1,t} = \boldsymbol{e}_j,\ \boldsymbol{\alpha}_{2,t} = 0\}.$$

By independence of the arm contexts, we have

$$\Pr(D_{t,j}) = \Pr(\boldsymbol{\alpha}_{1,t} = \boldsymbol{e}_j)\cdot\Pr(\boldsymbol{\alpha}_{2,t} = 0) = \left(\frac{3}{8}\cdot\frac{1}{2m} + \frac{1}{8}\cdot\frac{1}{2d}\right)\cdot\frac{1}{2} \geq \frac{3}{32m}. \tag{45}$$

For a fixed hypothesis $\boldsymbol{\omega}$, define the mistake event at round t, i.e., the event where the selected arm A_t is suboptimal:

$$M_t(\boldsymbol{\omega}) := \left\{A_t \notin \mathop{\arg\max}_{a\in[K]}\left(\mu_a + \langle\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{a,t}\rangle\right)\right\}.$$

We now show that, on $D_{t,j}$, choosing a suboptimal arm implies instantaneous regret at least $\varepsilon/\sqrt{m}$. On event $D_{t,j}$, since $\|\boldsymbol{\theta}_{\boldsymbol{\omega}}\| \leq \varepsilon$ and $\|\boldsymbol{\alpha}_{A_t,t}\| \leq 1$, the mean reward of each arm is

$$\begin{cases}\mu_1 + \langle\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{1,t}\rangle = 0 + \left\langle \frac{\varepsilon}{\sqrt{m}}\sum_{r=1}^{m}\omega_r\boldsymbol{e}_r, \boldsymbol{e}_j\right\rangle = \frac{\varepsilon\omega_j}{\sqrt{m}}, & \text{for arm 1,}\\ \mu_2 + \langle\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{2,t}\rangle = 0, & \text{for arm 2,}\\ \mu_a + \langle\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{a,t}\rangle \leq -1 + \|\boldsymbol{\theta}_{\boldsymbol{\omega}}\|\|\boldsymbol{\alpha}_{a,t}\| \leq -1 + \varepsilon < 0, & \text{for arm } a \geq 3.\end{cases}$$

For a fixed hypothesis $\boldsymbol{\omega}$, let the instantaneous regret at round t be

$$r_t(\boldsymbol{\omega}) = \max_{a\in[K]}\{\mu_a + \langle\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{a,t}\rangle\} - \left(\mu_{A_t} + \langle\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\alpha}_{A_t,t}\rangle\right).$$

If $\omega_j = 1$, then arm 1 is optimal and has mean $\varepsilon/\sqrt{m}$, while suboptimal arms have mean at most 0. Hence choosing suboptimal arms incurs instantaneous regret at least $\varepsilon/\sqrt{m}$.

Similarly, if $\omega_j = -1$, then arm 1 has mean $-\varepsilon/\sqrt{m}$, arm 2 has mean 0, and all other arms have mean less than $-1+\varepsilon \leq -1/2$. Hence arm 2 is optimal, choosing arm 1 incurs regret $\varepsilon/\sqrt{m}$, and choosing arms $a \geq 3$ incurs regret at least $1/2 \geq \varepsilon/\sqrt{m}$.

Therefore, on $D_{t,j}$,

$$r_t(\boldsymbol{\omega}) \geq \frac{\varepsilon}{\sqrt{m}}\mathbb{1}\{M_t(\boldsymbol{\omega})\}.$$

Finally, for a fixed round t, the events $D_{t,j}$, $j \in [m]$, are disjoint, because arm 1's context cannot be equal to $\boldsymbol{e}_j$ and $\boldsymbol{e}_{j'}$ for $j \neq j'$. Since $r_t(\boldsymbol{\omega}) \geq 0$, we have

$$r_t(\boldsymbol{\omega}) \geq r_t(\boldsymbol{\omega})\sum_{j=1}^{m}\mathbb{1}\{D_{t,j}\} \geq \sum_{j=1}^{m}\mathbb{1}\{D_{t,j}\}\frac{\varepsilon}{\sqrt{m}}\mathbb{1}\{M_t(\boldsymbol{\omega})\}. \tag{46}$$

We now lower bound the conditional mistake probability. Fix $t \in [n]$, $j \in [m]$, and two neighboring hypotheses $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^{(j)}$. Let H_{t-1} be the history before observing the contexts of round t, and let U_t denote the policy's internal randomization at round t, and let $W_t = (\{\boldsymbol{\alpha}_{a,t}\}_{a=1}^K, U_t, A_t)$. For a probability law $\mathbb{P}$ and random objects X, Y, we write $(\mathbb{P})_X$ for the marginal law of X under $\mathbb{P}$, and $(\mathbb{P})_{X|Y=y}$ for a regular conditional law of X given $Y = y$.

Given $H_{t-1} = h$, the current contexts and the policy randomization are generated by common distributions under all hypotheses. The arm is then selected by the same policy rule. Hence the conditional law of W_t given $H_{t-1} = h$ is the same under both $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^{(j)}$. Hence

$$(\mathbb{P}_{\boldsymbol{\omega}})_{W_t|H_{t-1}=h} = (\mathbb{P}_{\boldsymbol{\omega}^{(j)}})_{W_t|H_{t-1}=h}.$$

Since the conditional law of W_t given $H_{t-1} = h$ is identical under both hypotheses, and $D_{t,j}$ is an event determined by W_t with positive probability under this common law, conditioning on $D_{t,j}$ preserves the equality of the two conditional laws.

$$(\mathbb{P}_{\boldsymbol{\omega}})_{W_t|H_{t-1}=h,D_{t,j}} = (\mathbb{P}_{\boldsymbol{\omega}^{(j)}})_{W_t|H_{t-1}=h,D_{t,j}}.$$

Therefore, we have

$$\begin{aligned}
&D_{\mathrm{KL}}\Big((\mathbb{P}_{\boldsymbol{\omega}})_{H_{t-1},W_t|D_{t,j}} \,\Big\|\, (\mathbb{P}_{\boldsymbol{\omega}^{(j)}})_{H_{t-1},W_t|D_{t,j}}\Big) \\
&= D_{\mathrm{KL}}\Big((\mathbb{P}_{\boldsymbol{\omega}})_{H_{t-1}|D_{t,j}} \,\Big\|\, (\mathbb{P}_{\boldsymbol{\omega}^{(j)}})_{H_{t-1}|D_{t,j}}\Big) \\
&= D_{\mathrm{KL}}\Big((\mathbb{P}_{\boldsymbol{\omega}})_{H_{t-1}} \,\Big\|\, (\mathbb{P}_{\boldsymbol{\omega}^{(j)}})_{H_{t-1}}\Big) \\
&\le D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}}),
\end{aligned} \tag{47}$$

where in the first equality we apply Lemma 14 under the two conditional laws given $D_{t,j}$, with $Z = H_{t-1}$ and $W = W_t$. The second equality uses the independence of $D_{t,j}$ from H_{t-1} under each hypothesis. For the final inequality, we use the data-processing intuition that observing only the history H_{t-1} cannot make the two hypotheses easier to distinguish than observing the full trajectory.

Fix t, j and $\boldsymbol{\omega}$ with $\omega_j = 1$. On $D_{t,j}$, arm 1 is optimal, hence $M_t(\boldsymbol{\omega}) = \{A_t \neq 1\}$. Under the neighboring hypothesis $\boldsymbol{\omega}^{(j)}$, since the j-th sign is flipped, so arm 2 is optimal and $M_t(\boldsymbol{\omega}^{(j)}) = \{A_t \neq 2\}$, thus $\{A_t = 1\} \subseteq M_t(\boldsymbol{\omega}^{(j)})$. Therefore,

$$\begin{aligned}
&\Pr_{\boldsymbol{\omega}}\left(M_t(\boldsymbol{\omega}) \mid D_{t,j}\right) + \Pr_{\boldsymbol{\omega}^{(j)}}\left(M_t(\boldsymbol{\omega}^{(j)}) \mid D_{t,j}\right) \\
&\ge \Pr_{\boldsymbol{\omega}}\left(A_t \neq 1 \mid D_{t,j}\right) + \Pr_{\boldsymbol{\omega}^{(j)}}\left(A_t = 1 \mid D_{t,j}\right) \\
&\ge \frac{1}{2}\exp\left(-D_{\mathrm{KL}}\left((\mathbb{P}_{\boldsymbol{\omega}})_{H_{t-1},W_t|D_{t,j}} \,\|\, (\mathbb{P}_{\boldsymbol{\omega}^{(j)}})_{H_{t-1},W_t|D_{t,j}}\right)\right) \qquad (48) \\
&\ge \frac{1}{2}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}})), \qquad (49)
\end{aligned}$$

where Equation (48) applies the Bretagnolle-Huber inequality (Lemma 13) to the two conditional experiments with the event $\{A_t \neq 1\}$.

Averaging the above inequality over all $j \in [m]$ and all $\boldsymbol{\omega}$ with $\omega_j = 1$, we obtain

$$\begin{aligned}
&\frac{1}{2^m m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{j=1}^m \Pr_{\boldsymbol{\omega}}(M_t(\boldsymbol{\omega}) \mid D_{t,j}) \\
&= \frac{1}{2^m m}\sum_{j=1}^m \sum_{\boldsymbol{\omega}:\omega_j=1}\left[\Pr_{\boldsymbol{\omega}}\left(M_t(\boldsymbol{\omega}) \mid D_{t,j}\right) + \Pr_{\boldsymbol{\omega}^{(j)}}\left(M_t(\boldsymbol{\omega}^{(j)}) \mid D_{t,j}\right)\right] \\
&\ge \frac{1}{2^{m+1} m}\sum_{j=1}^m \sum_{\boldsymbol{\omega}:\omega_j=1}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}})) \qquad (50) \\
&= \frac{1}{4}\cdot\frac{1}{2^{m-1} m}\sum_{j=1}^m \sum_{\boldsymbol{\omega}:\omega_j=1}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}}))
\end{aligned}$$

$$\geq \frac{1}{4}\exp\left(-\frac{1}{2^{m-1}m}\sum_{j=1}^{m}\sum_{\boldsymbol{\omega}:\omega_j=1} D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}})\right) \tag{51}$$

$$\geq \frac{1}{4}\exp\left(-\frac{1}{2^{m-1}m}\sum_{j=1}^{m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m} D_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\omega}} \,\|\, \mathbb{P}_{\boldsymbol{\omega}^{(j)}})\right) \tag{52}$$

$$\geq \frac{1}{4e}. \tag{53}$$

where Equation (50) uses Equation (49), Equation (51) uses Jensen's inequality for the convex function e^{-x}, Equation (52) uses nonnegativity of KL divergence, and Equation (53) uses Equation (44).

Finally, we average the regret over the 2^m hypotheses. In the following, Equation (54) uses Equation (46). Equation (55) uses Equations (45) and (53).

$$\begin{aligned}
\frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\mathbb{E}_{\boldsymbol{\omega}}[R_n] &= \frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{t=1}^{n}\mathbb{E}_{\boldsymbol{\omega}}[r_t(\boldsymbol{\omega})] \\
&\geq \frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{t=1}^{n}\mathbb{E}_{\boldsymbol{\omega}}\left[\sum_{j=1}^{m}\mathbb{1}\{D_{t,j}\}\frac{\varepsilon}{\sqrt{m}}\mathbb{1}\{M_t(\boldsymbol{\omega})\}\right] \qquad (54)\\
&= \frac{\varepsilon}{\sqrt{m}}\frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{t=1}^{n}\sum_{j=1}^{m}\mathbb{E}_{\boldsymbol{\omega}}[\mathbb{1}\{D_{t,j}\}\mathbb{1}\{M_t(\boldsymbol{\omega})\}] \\
&= \frac{\varepsilon}{\sqrt{m}}\frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\sum_{t=1}^{n}\sum_{j=1}^{m}\Pr_{\boldsymbol{\omega}}[D_{t,j}, M_t(\boldsymbol{\omega})] \\
&= \frac{\varepsilon}{\sqrt{m}}\sum_{t=1}^{n}\sum_{j=1}^{m}\Pr(D_{t,j})\frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\Pr_{\boldsymbol{\omega}}(M_t(\boldsymbol{\omega}) \mid D_{t,j}) \\
&\geq \frac{\varepsilon}{\sqrt{m}}\sum_{t=1}^{n}\frac{3}{32m}\sum_{j=1}^{m}\frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\Pr_{\boldsymbol{\omega}}(M_t(\boldsymbol{\omega}) \mid D_{t,j}) \qquad (55)\\
&\geq \frac{\varepsilon}{\sqrt{m}}\sum_{t=1}^{n}\frac{3}{32}\frac{1}{4e} = \frac{3}{128e}\frac{n\varepsilon}{\sqrt{m}}.
\end{aligned}$$

Substituting $\varepsilon = m/(2\sqrt{n})$, we obtain

$$\frac{1}{2^m}\sum_{\boldsymbol{\omega}\in\{-1,1\}^m}\mathbb{E}_{\boldsymbol{\omega}}[R_n] \geq \frac{3}{256e}\sqrt{mn}.$$

Therefore, at least one hypothesis $\boldsymbol{\omega} \in \{-1,1\}^m$ satisfies

$$\mathbb{E}_{\boldsymbol{\omega}}[R_n] \geq c_2\sqrt{mn}, \qquad c_2 \coloneqq \frac{3}{256e}.$$

Putting two lower bounds together. For the fixed policy π, we have proven that there exists an instance with expected regret at least $c_1\sqrt{Kn}$, and there exists an instance with expected regret at least $c_2\sqrt{mn}$. Choosing the harder of these two instances gives

$$\begin{aligned}
\mathbb{E}[R_n] &\geq \max\{c_1\sqrt{Kn}, c_2\sqrt{mn}\} \\
&\geq \min\{c_1, c_2\}\max\{\sqrt{Kn}, \sqrt{mn}\} \\
&\geq \frac{\min\{c_1, c_2\}}{\sqrt{2}}\sqrt{(K+m)n} \\
&= \frac{\min\{c_1, c_2\}}{\sqrt{2}}\sqrt{(K+d_{\ell^*})n}.
\end{aligned}$$

Taking $c_{\mathrm{lb}} = \min\{c_{\mathrm{MAB}}, \min\{c_1, c_2\}/\sqrt{2}\}$ concludes the proof. □

F TECHNICAL INEQUALITIES

We present the technical inequalities used throughout the proofs. For inequalities from existing literature, we provide detailed references for readers' convenience.

Lemma 10 (Matrix Chernoff, Corollary 5.2 in Tropp (2011))**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies*

$$\boldsymbol{X}_k \succeq \mathbf{0} \quad \textit{and} \quad \lambda_{max}(\boldsymbol{X}_k) \leq R \quad \textit{almost surely.}$$

Define

$$\boldsymbol{Y} := \sum_k \boldsymbol{X}_k \quad \textit{and} \quad \mu_{min} := \lambda_{\min}(\mathbb{E}[\boldsymbol{Y}]) = \lambda_{\min}\left(\sum_k \mathbb{E}[\boldsymbol{X}_k]\right).$$

Then, for any $\varepsilon \in (0, 1)$,

$$\Pr\left[\lambda_{\min}\left(\sum_k \boldsymbol{X}_k\right) \leq (1-\varepsilon)\mu_{\min}\right] \leq d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/R}.$$

Lemma 11 (Determinant-trace inequality, Lemma 10 in Abbasi-Yadkori et al. (2011))**.** *Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|\boldsymbol{X}_s\|_2 \leq L$. Let $\overline{\boldsymbol{V}}_t = \lambda \boldsymbol{I} + \sum_{s=1}^t \boldsymbol{X}_s \boldsymbol{X}_s^\mathsf{T}$ for some $\lambda > 0$. Then,*

$$\det(\overline{\boldsymbol{V}}_t) \leq \left(\lambda + \frac{tL^2}{d}\right)^d.$$

Lemma 12 (KL divergence between Gaussian distributions)**.** *If $P \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Q \sim \mathcal{N}(\mu_2, \sigma^2)$, with $\sigma > 0$, then*

$$D_{\mathrm{KL}}(P \,\|\, Q) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

Lemma 13 (Bretagnolle & Huber (1978))**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D_{\mathrm{KL}}(P \,\|\, Q)),$$

where $D_{\mathrm{KL}}(P \,\|\, Q) = \int_\Omega \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]$ is the KL divergence between P and Q. $A^c = \Omega \setminus A$ is the complement of A.

Lemma 14. *Let P and Q be two probability laws on a common measurable space. For a random object R, write P_R and Q_R for the marginal laws of R under P and Q, respectively. Let Z and W be random objects with regular conditional laws. Assume that the conditional laws of W given Z under P and Q satisfy*

$$P_{W|Z=z} = Q_{W|Z=z} \qquad \textit{for } P_Z\textit{-almost every } z.$$

Then

$$D_{\mathrm{KL}}(P_{Z,W} \,\|\, Q_{Z,W}) = D_{\mathrm{KL}}(P_Z \,\|\, Q_Z).$$

Proof. By the chain rule for KL divergence,

$$D_{\mathrm{KL}}(P_{Z,W} \,\|\, Q_{Z,W}) = D_{\mathrm{KL}}(P_Z \,\|\, Q_Z) + \mathop{\mathbb{E}}_{Z \sim P_Z}\left[D_{\mathrm{KL}}\left(P_{W|Z=z} \,\middle\|\, Q_{W|Z=z}\right)\right].$$

The second term is zero because $P_{W|Z=z} = Q_{W|Z=z}$ for P_Z-almost every z. Hence

$$D_{\mathrm{KL}}(P_{Z,W} \,\|\, Q_{Z,W}) = D_{\mathrm{KL}}(P_Z \,\|\, Q_Z).$$

This proves the claim. □