Empirical Bound Information-Directed Sampling

Anonymous authors
Paper under double-blind review

Keywords: bandit algorithms, information-directed sampling, parameter bounds, heteroskedastic noise

Summary

Information-directed sampling (IDS) is a powerful framework for solving bandit problems which has shown strong results in both Bayesian and frequentist settings. However, frequentist IDS, like many other bandit algorithms, requires that one have prior knowledge of a (relatively) tight upper bound on the norm of the true parameter vector governing the reward model in order to achieve good performance. Unfortunately, this requirement is rarely satisfied in practice. As we demonstrate, using a poorly calibrated bound can lead to significant regret accumulation. To address this issue, we introduce a novel frequentist IDS algorithm that iteratively refines a high-probability upper bound on the true parameter norm using accumulating data. We focus on the linear bandit setting with heteroskedastic subgaussian noise. Our method leverages a mixture of relevant information gain criteria to balance exploration aimed at tightening the parameter norm bound and directly searching for the optimal action. We establish regret bounds for our algorithm that do not depend on an initially assumed parameter norm bound and demonstrate that our method outperforms state-of-the-art IDS and UCB algorithms.

Contribution(s)

1. This paper introduces a novel frequentist information-directed sampling (IDS) algorithm that does not require prior knowledge of a tight upper bound of the true parameter norm to achieve good performance. Our method uses accumulating data to generate a sequence of high-probability upper bounds on the parameter norm and accounts for potential heteroskedasticity of the rewards.
   Context: The performance of many frequentist bandit algorithms, including various IDS (Kirschner & Krause, 2018; Kirschner et al., 2021) and UCB methods (Auer, 2002; Abbasi-Yadkori et al., 2011), relies heavily on a (at least relatively) tight upper bound on the true parameter norm being available to the algorithm. This is almost never the case in practice which can lead to significant regret accumulation. Recently, some norm-agnostic bandit algorithms have been proposed to address this issue (Gales et al., 2022), however, they do not account for potential heteroskedasticity of the rewards.
2. We introduce a new composite information criterion that balances improving the requisite upper bound on the parameter norm and direct search for the optimal action.
   Context: To the best of our knowledge, no other IDS algorithm uses a mixture of information gain criteria to balance acquiring information about different aspects of the environment's dynamics. We are also not aware of any existing method that uses an information gain criterion aimed at improving the upper bound on the parameter norm.
3. We establish anytime sublinear regret bounds for our algorithm which eventually do not depend on the initially assumed parameter norm bound.
   Context: Previously proposed norm-agnostic bandits (Gales et al., 2022) rely on an initial burn-in during which regret accumulation is not controlled, e.g., it need not be sublinear.

Empirical Bound Information-Directed Sampling

Anonymous authors
Paper under double-blind review

Abstract

Information-directed sampling (IDS) is a powerful framework for solving bandit problems which has shown strong results in both Bayesian and frequentist settings. However, frequentist IDS, like many other bandit algorithms, requires that one have prior knowledge of a (relatively) tight upper bound on the norm of the true parameter vector governing the reward model in order to achieve good performance. Unfortunately, this requirement is rarely satisfied in practice. As we demonstrate, using a poorly calibrated bound can lead to significant regret accumulation. To address this issue, we introduce a novel frequentist IDS algorithm that iteratively refines a high-probability upper bound on the true parameter norm using accumulating data. We focus on the linear bandit setting with heteroskedastic subgaussian noise. Our method leverages a mixture of relevant information gain criteria to balance exploration aimed at tightening the estimated parameter norm bound and directly searching for the optimal action. We establish regret bounds for our algorithm that do not depend on an initially assumed parameter norm bound and demonstrate that our method outperforms state-of-the-art IDS and UCB algorithms.

1 Introduction

We consider linear stochastic bandits (Lattimore & Szepesvári, 2020) with heteroskedastic noise (see Weltz et al., 2023, for applications of such models in marketing and other areas). In this setting, information directed sampling (IDS) and upper confidence bound (UCB) algorithms have been shown to be extremely effective (Auer, 2002; Abbasi-Yadkori et al., 2011; Kirschner & Krause, 2018; Kirschner et al., 2021). However, many of these methods require strong prior information that can be used to inform a high-quality upper bound on the Euclidean norm of the parameter vector indexing the reward model. The choice of this bound is critical to algorithm performance. If the bound is too large, the algorithm risks incurring excess risk due to needless exploration, and if the bound is too small, the algorithm may fail to identify the optimal arm and thus suffer linear regret.

To reduce sensitivity on a user-specified bound, we propose a novel version of frequentist IDS that uses accumulating data to generate a sequence of high-probability upper bounds on the norm of the reward model parameters. A key component of our method is a new information gain criterion that balances improving the requisite upper bound and regret minimization. Because improving the bound is critical to avoid over-exploration in early rounds of the bandit process, we develop a two-phase procedure that uses our new information criterion in the first phase and then defaults to a more standard IDS information criterion in the second phase.

Unlike other bandit strategies, such as UCB (Auer, 2002; Garivier & Cappé, 2011; Cappé et al., 2013; Zhou et al., 2020) or Thompson sampling (TS) (Thompson, 1933; Agrawal & Goyal, 2013; Phan et al., 2019), which encourage exploration indirectly by leveraging uncertainty about the optimal arm, IDS explicitly balances exploration and exploitation. It selects actions that minimize estimated instantaneous regret while maximizing expected information gain about model parameters. As shown by Russo & Van Roy (2014) and Kirschner & Krause (2018), this approach allows IDS to avoid pitfalls inherent in UCB and TS-based algorithms, particularly in scenarios where certain suboptimal actions provide valuable information about the environment's dynamics. In such cases, UCB and

TS tend to overlook these actions, whereas IDS plays them early on, enabling faster learning of the optimal policy and ultimately achieving superior long-term performance. IDS was first introduced for Bayesian bandits by Russo & Van Roy (2014) and later adapted to the frequentist setting by Kirschner & Krause (2018). Beyond the standard bandit setting, IDS has been applied to problems such as linear partial monitoring (Kirschner et al., 2020) — a generalization of bandits where the observed signal on the environment model parameters is not necessarily the same as the reward to be optimized — as well as reinforcement learning (Nikolov et al., 2019; Lindner et al., 2021; Hao & Lattimore, 2022), where the actions taken by the agent influence the state of the environment and the reward dynamics.

To the best of our knowledge, no previous work has considered either the strategy of iteratively refining and utilizing a high-probability upper bound on the parameter norm in the heteroskedastic subgaussian linear bandit setting we work with here, or the use of the information gain criterion for tightening the bound on the parameter norm we introduce. We are also not aware of any work utilizing a mixture of information gain criteria to encourage simultaneously obtaining different types of information about the dynamics of the environment. We note that while we introduce this idea in the form of an IDS algorithm, the approach of iteratively refining and utilizing a high-probability upper bound of the true parameter norm can be regarded as a more general design principle beyond its IDS implementation in this setting.

The remainder of this manuscript is structured as follows. The next section provides a brief review of related work. Section 3 introduces the problem setup and notation used throughout the paper. In Section 4, we present the necessary background on IDS. Section 5 introduces the novel empirical bound information-directed sampling (EBIDS) algorithm, which removes the need for a tight parameter norm bound to be known *a priori*. Section 6 establishes regret bound guarantees for EBIDS, and finally, Section 7 evaluates its empirical performance against competitor algorithms in a simulation study.

2 Related works

The assumption that the norm of the parameter indexing the reward model is known or that one has a (relatively) tight upper bound on this quantity is abundant in the IDS and UCB literature (Auer, 2002; Abbasi-Yadkori et al., 2011; Kirschner & Krause, 2018; Hung et al., 2021); it has also been used in Thompson sampling (Xu et al., 2023). This assumption commonly arises through the use of self-normalized martingale bounds and related concentration results (Abbasi-Yadkori et al., 2011). Consequently, algorithms constructed through these concentration results require a user-specified upper bound on the norm or the true parameter vector. Critically, as noted previously, the performance of these algorithms can be highly sensitive to the choice of these bounds. Despite this, only a handful of papers have attempted to alleviate this sensitivity.

Gales et al. (2022) propose norm-agnostic linear bandits which construct a series of confidence ellipsoids for the true parameter vector along with a projection interval to construct a UCB-type algorithm. However, their algorithms rely on an initial burn-in during which regret accumulation is not controlled, e.g., it need not be sublinear. In our simulation experiments, we find that the impact of this initial exploration on accumulated regret is not negligible. Furthermore, as UCB algorithms, their methods do not explicitly make use of heteroskedasticity in the reward distributions across arms.

The algorithm proposed by Ghosh et al. (2021) shares some underlying ideas with our method in the sense that they use multi-phase exploration to iteratively update the bound on the unknown parameter norm. However, their algorithm is limited to the specialized setting of stochastic linear bandits introduced by Chatterji et al. (2020) with restrictive assumptions on the structure of the rewards which makes their methods generally not applicable to the settings we consider here. Similarly, Dani et al. (2008), Orabona & Cesa-Bianchi (2011), and Gentile & Orabona (2014) do not assume that one has a high-quality (i.e., relatively tight) bound on the norm of the parameter; however, they require bounded rewards for all arms. Other attempts to alleviate the assumption of known parameter norm bound have been made in spectral bandits (Kocák et al., 2020), and deep active learning (Wang et al., 2021). However, it is not clear how to port these methods to the setup we consider here.

3 Setup and notation

We denote the inner product of two vectors of the same dimension as $\langle \cdot, \cdot \rangle$ so that the squared Euclidean norm of vector $\boldsymbol{v}$ is $||\boldsymbol{v}||_2^2 = \langle \boldsymbol{v}, \boldsymbol{v} \rangle$. For a symmetric positive definite or semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, we denote the associated matrix norm (or semi-norm) of a vector $\boldsymbol{v} \in \mathbb{R}^d$ as $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 = \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{v} \rangle$. We let $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote the largest and the smallest eigenvalues of $\boldsymbol{A}$. Throughout, $\log(x)$ denotes the natural logarithm of $x \in \mathbb{R}_+$.

At each time step $t \in \{1, \dots, T\}$ the agent selects an action $A_t \in \mathcal{A}$ and observes the outcome $Y_t \in \mathbb{R}$ which is generated from the linear model

$$Y_t(A_t) = \langle \boldsymbol{\phi}(A_t), \boldsymbol{\theta}^* \rangle + \eta_t, \tag{1}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is a vector of unknown parameters and $\boldsymbol{\phi} : \mathcal{A} \to \mathbb{R}^d$ is a feature mapping, such that for any $a \in \mathcal{A}$ we have $\|\boldsymbol{\phi}(a)\|_2 \in [L, U]$ for some positive constants $L \leq U$. The noise term η_t is assumed to be subgaussian and conditionally mean zero, i.e., for every $c \in \mathbb{R}$ we assume that

$$\mathbb{E}\left\{\exp\left(c\eta_t\right) \mid A_t = a\right\} \leq \exp\left\{c^2 \rho\left(a\right)^2 / 2\right\}, \tag{2}$$

where $0 < \rho_{\min} \leq \rho(a) \leq \rho_{\max} < \infty$ for all $a \in \mathcal{A}$ and $\mathbb{E}\left(\eta_t \mid A_1, \dots, A_t, \eta_1, \dots, \eta_{t-1}\right) = 0$. Define $B^* := \|\boldsymbol{\theta}^*\|_2$, in some of our theoretical results we assume that one has a conservative upper bound B such that $B^* \leq B$ but that this bound may be quite conservative, i.e., it may be that $B^* \ll B$.

The available history to inform action selection at time t is $\boldsymbol{H}_t = \{(A_1, Y_1), \dots, (A_{t-1}, Y_{t-1})\}$ of past actions and rewards. A bandit algorithm is thus formalized as a map from histories to distributions over actions $\pi_t(a|\boldsymbol{h}_t) = \mathbb{P}(A_t = a | \boldsymbol{H}_t = \boldsymbol{h}_t)$. Let

$$\Delta(A_t) = \langle \boldsymbol{\phi}(a^*), \boldsymbol{\theta}^* \rangle - \langle \boldsymbol{\phi}(A_t), \boldsymbol{\theta}^* \rangle$$

be the gap between the action A_t and the optimal action $a^* = \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\phi}(a), \boldsymbol{\theta}^* \rangle$. Our goal is to design an algorithm $\pi_t(\cdot \mid \boldsymbol{h}_t)$ which maximizes the cumulative expected reward $\mathbb{E}\left\{\sum_{t=1}^T Y_t\right\}$, or equivalently, minimizes the regret, defined as $\mathcal{R}_T = \mathbb{E}\left\{\sum_{t=1}^T \Delta(A_t)\right\}$. While regret is a standard performance metric for bandit algorithms, it involves taking expectation over both the randomness in the policy and the noise in the rewards so it can be a poor indicator of the risk associated with the policy (Lattimore & Szepesvári, 2020). For this reason in this paper we also study the probabilistic bounds on the pseudo-regret defined as $\mathcal{PR}_T = \sum_{t=1}^T \Delta(A_t)$.

4 Review of information-directed sampling

Information-directed sampling (IDS Russo & Van Roy, 2014) is an algorithm design principle that balances minimizing the gap of an action with its potential for information gain. Let $\mathcal{P}(\mathcal{A})$ denote the space of distributions of $\mathcal{A}$. For any $\mu \in \mathcal{A}$ let $\widehat{\Delta}_t(\mu)$ be an estimator of the expected gap $\mathbb{E}_\mu \Delta := \mathbb{E}_{A \sim \mu} \Delta(A)$ constructed from the history $\boldsymbol{H}_t$, and, similarly, let $I_t(\mu)$ be a measure of information again, e.g., the reduction of entropy in the posterior or sampling distribution the parameter indexing the mean reward model (see below for additional details). For any function $f : \mathcal{P}(\mathcal{A}) \to \mathbb{R}$, if the argument is a point mass at a single action, e.g., where μ is the Dirac delta δ_a, we write $f(a)$ rather than $f(\delta_a)$. The IDS distribution is defined as

$$\mu_t^{\text{IDS}} = \underset{\mu \in \mathcal{P}(\mathcal{A})}{\arg\min} \frac{\left\{\widehat{\Delta}_t(\mu)\right\}^2}{I_t(\mu)}. \tag{3}$$

The quantity $\Psi_t(\mu) := \left\{\widehat{\Delta}_t(\mu)\right\}^2 / I_t(\mu)$ being minimized is known as the *information ratio*. An IDS algorithm samples the action $A_t \sim \mu_t^{\text{IDS}}$ at each time step t. Note that this results in a randomized algorithm, which, as shown by Russo & Van Roy (2014) and Kirschner & Krause (2018), always has at most two actions in its support. However, it is also possible to restrict the optimization in (3) to Dirac delta functions on the individual actions, thus obtaining what is often referred to as *deterministic IDS* (Kirschner & Krause, 2018)

$$\widehat{A}_t^{\text{DIDS}} = \arg\min_{a \in \mathcal{A}} \frac{\left\{\widehat{\Delta}_t(a)\right\}^2}{I_t(a)}. \tag{4}$$

Deterministic IDS is typically computationally cheaper, retains the same theoretical regret bounds as its randomized counterpart, and in simulation experiments was shown to be competitive with or superior to randomized IDS (Kirschner & Krause, 2018; Kirschner, 2021). Furthermore, deterministic IDS may be appealing in settings where randomized policies are unpalatable such as public health (Weltz et al., 2022) and site selection (Ahmadi-Javid et al., 2017).

The information ratio provides a natural way of bounding regret within a Bayesian setting (Russo & Van Roy, 2014). Notably, the information ratio can also be used to bound the regret under a frequentist paradigm (Kirschner & Krause, 2018) as illustrated by the following result based on the work of Kirschner (2021) which we prove in Section 10.1 of the Supplementary Materials.

Theorem 1 (Kirschner)**.** *For any T let G be a fixed subset of $\{1, \ldots, T\}$ and let $\{A_t\}_{t=1}^T$ be an $\boldsymbol{H}_t$-adapted sequence in $\mathcal{A}$. Then*

$$\mathbb{E}\left\{\sum_{t \in G} \widehat{\Delta}_t\left(A_t\right)\right\} \leq \sqrt{\mathbb{E}\left\{\sum_{t \in G} \Psi_t\left(A_t\right)\right\} \mathbb{E}\left\{\sum_{t \in G} I_t\left(A_t\right)\right\}},$$

and if $\widehat{\Delta}_t(A_t) \geq \Delta(A_t)$ for all $t \in G$ then with probability 1 we have

$$\sum_{t \in G} \Delta\left(A_t\right) \leq \sqrt{\left\{\sum_{t \in G} \Psi_t(A_t)\right\}\left\{\sum_{t \in G} I_t(A_t)\right\}}.$$

Kirschner & Krause (2018) used weighted ridge regression to estimate $\boldsymbol{\theta}^*$ at each time step t so that

$$\widehat{\boldsymbol{\theta}}_t^{\text{wls}} = \boldsymbol{W}_t^{-1} \sum_{s=1}^{t-1} \frac{1}{\rho(A_s)^2} \boldsymbol{\phi}(A_s) Y_s, \quad \text{where} \quad \boldsymbol{W}_t = \sum_{s=1}^{t-1} \frac{1}{\rho(A_s)^2} \boldsymbol{\phi}(A_s) \boldsymbol{\phi}(A_s)^\top + \gamma \boldsymbol{I}_d, \tag{5}$$

and $\gamma \geq 0$ is a constant chosen by the user. The following result, proposed by Abbasi-Yadkori et al. (2011) and extended by Kirschner & Krause (2018), provides a means to perform inference using this estimator.

Theorem 2. *Suppose that the generative model follows the linear bandit model $Y_t = \langle \boldsymbol{\phi}(A_t), \boldsymbol{\theta}^* \rangle + \eta_t$ given in (1), where the actions A_t are $\boldsymbol{H}_t$-adapted and the errors η_t have conditional mean of zero and satisfy the subgaussian condition in (2). Let $B \geq ||\boldsymbol{\theta}^*||_2$ be a (potentially conservative) bound on the norm of the parameters indexing the reward model and define*

$$\mathcal{E}_{t,\delta}^{\text{wls}} := \left\{\boldsymbol{\theta} \in \mathbb{R}^d : \left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_{\boldsymbol{W}_t}^2 \leq \beta_{t,\delta}(B)\right\},$$

where

$$\beta_{t,\delta}(B) = \left[\sqrt{2\log\frac{1}{\delta} + \log\left\{\frac{\det\left(\boldsymbol{W}_t\right)}{\det\left(\boldsymbol{W}_1\right)}\right\}} + \sqrt{\gamma} B\right]^2. \tag{6}$$

Then

$$\mathbb{P}\left(\bigcap_{t=1}^{\infty}\left\{\boldsymbol{\theta}^* \in \mathcal{E}_{t,\delta}^{\text{wls}}\right\}\right) \geq 1-\delta,$$

i.e., $\mathcal{E}_{t,\delta}^{\text{wls}}$ *is a* $(1-\delta)\times 100\%$ *confidence ellipsoid for* $\boldsymbol{\theta}^*$.

Kirschner & Krause (2018) use Theorem 2 to formulate a weighted UCB algorithm which at each time step t takes the action

$$A_t^{\text{UCB}(\delta_t)} = \underset{a\in\mathcal{A}}{\arg\max}\left\langle \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\delta_t}^{1/2}(B)\|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}, \tag{7}$$

maximizing the $(1-\delta_t)\times 100\%$ upper confidence bound on the expected reward based on the $\mathcal{E}_{t,\delta_t}^{\text{wls}}$ confidence set. Then they use

$$\check{\Delta}_{t,\delta_t}(a) = \left\langle \boldsymbol{\phi}\left(A_t^{\text{UCB}(\delta_t)}\right) - \phi(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\delta_t}^{1/2}(B)\left(\left\|\boldsymbol{\phi}\left(A_t^{\text{UCB}(\delta_t)}\right)\right\|_{\boldsymbol{W}_t^{-1}} + \|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}\right).$$

as the gap estimate. This ensures that $\Delta(a) \leq \check{\Delta}_{t,\delta_t}(a)$ for all $a\in\mathcal{A}$ whenever $\boldsymbol{\theta}^* \in \mathcal{E}_{t,\delta_t}^{\text{wls}}$ holds.

The choice of the information gain criterion is crucial when designing an IDS algorithm. Kirschner & Krause (2018) introduce the following criterion

$$I_t^{\text{UCB}(\delta_t)}(a) = \frac{1}{2}\log\left(\frac{\left\|\boldsymbol{\phi}\left(a_t^{\text{UCB}(\delta_t)}\right)\right\|_{\boldsymbol{W}_t^{-1}}^2}{\left\|\boldsymbol{\phi}\left(a_t^{\text{UCB}(\delta_t)}\right)\right\|_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)^{-1}}^2}\right),$$

for any $a\in\mathcal{A}$. We present the resulting procedure in Algorithm 1, which we hereafter refer to as IDS-UCB. It can be shown that if one chooses $\delta_t = 1/t^2$, the regret of IDS-UCB satisfies

Algorithm 1 IDS-UCB

Input: Action set $\mathcal{A}$, penalty parameter $\gamma > 0$, noise function $\rho : \mathcal{A}\to\mathbb{R}_+$, feature function $\boldsymbol{\phi} : \mathcal{A}\to\mathbb{R}$, sequence of confidence levels $\{\delta_t\}_{t\geq 1}\subset(0,1)$, assumed true parameter norm bound B.

For $t = 1, 2, \ldots, T$:

Compute $\boldsymbol{W}_t$ and $\widehat{\boldsymbol{\theta}}_t^{\text{wls}}$ using (5)

$A_t^{\text{UCB}(\delta_t)} \leftarrow \arg\max_{a\in A}\left\{\left\langle\boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\delta_t}^{1/2}(B)\|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}\right\}$

$I_t^{\text{UCB}(\delta_t)}(a) \leftarrow \frac{1}{2}\log\left(\left\|\boldsymbol{\phi}\left(A_t^{\text{UCB}(\delta_t)}\right)\right\|_{\boldsymbol{W}_t^{-1}}^2\right) - \frac{1}{2}\log\left(\left\|\boldsymbol{\phi}\left(A_t^{\text{UCB}(\delta_t)}\right)\right\|_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)^{-1}}^2\right)$

$\check{\Delta}_{t,\delta_t}(a) \leftarrow \left\langle \boldsymbol{\phi}\left(A_t^{\text{UCB}(\delta_t)}\right) - \phi(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\delta_t}^{1/2}(B)\left(\left\|\boldsymbol{\phi}\left(A_t^{\text{UCB}(\delta_t)}\right)\right\|_{\boldsymbol{W}_t^{-1}} + \|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}\right)$

$\mu_t \leftarrow \arg\min_{\mu\in\mathcal{P}(\mathcal{A})} \check{\Delta}_{t,\delta_t}^2(\mu_t)/I_t^{\text{UCB}(\delta_t)}(\mu_t)$

Sample $A_t \sim \mu_t$

Play A_t, observe $Y_t = \langle\boldsymbol{\phi}(A_t), \boldsymbol{\theta}^*\rangle + \eta_t$

$$\mathcal{R}_T \leq O\left(\max\{U/\sqrt{\gamma}, \rho_{\max}\}\sqrt{\gamma}dB\sqrt{T}\log T\right),$$

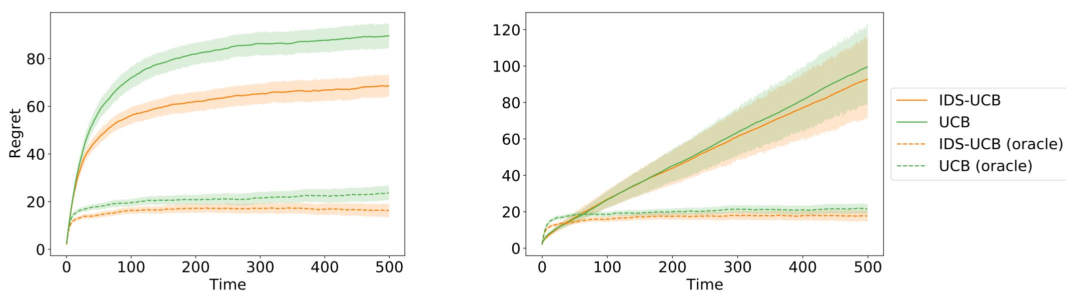
while the pseudo-regret $\mathcal{PR}_T$ of IDS-UCB with fixed $\delta_t = \delta$ satisfies with probability at least $1-\delta$

$$\mathcal{PR}_T \leq O(\max\{U/\sqrt{\gamma}, \rho_{\max}\}\sqrt{\gamma}dB\sqrt{T}\log(T/\delta));$$

critically, both regret bounds scale directly with the assumed bound B on the Euclidean norm of the true parameter (see Kirschner, 2021, for a formal statement of the preceding results and additional discussion).

We now demonstrate via a simple illustrative simulation experiment that the choice of B can have a significant impact on the finite time performance of IDS-UCB. Large values of B relative to B^* lead to excess exploration and large regret in early rounds of the algorithm, whereas small values of B can prevent the algorithm from identifying the optimal arm thus incurring linear regret. In this experiment we also include the weighted UCB policy given by (7). We evaluate versions of IDS-UCB and UCB that use a conservative value of $B > B^*$, and those which use an anti-conservative value $B < B^*$. The parameters indexing the generative model are $\boldsymbol{\theta}^* = [-5, 1, 1, 1.5, 2]^\top$ so that $B^* = \|\boldsymbol{\theta}^*\|_2 \approx 5.77$. We take $B = 100$ for the conservative bound and $B = 1$ for the anti-conservative bound. For reference, we also include *oracle* versions of UCB and IDS-UCB that have access to the true value of B^*. However, we emphasize that these procedures are not generally possible in practice.

We consider a setting with ten arms. Features for each arm are sampled from Unif$[-1/\sqrt{5}, 1/\sqrt{5}]$. The error distribution for the first five arms are standard normal and for the remaining five arms they are normal with mean zero and variance 0.2. Figure 1 shows the mean regret averaged over 200 repeated experiments with $T = 500$ steps along with 95% pointwise confidence bounds. As anticipated, using a conservative bound of $B = 100$ achieves sublinear regret but pays a strong initial cost due to excess exploration. Algorithms that used the anti-conservative bound of $B = 1$ fail to identify the optimal arm thus sustain linear regret.



(a) Conservative $B = 100$

(b) Anti-conservative $B = 1$

Figure 1: Regret incurred by IDS-UCB and UCB with: (a) conservative $B = 100$; (b) anti-conservative $B = 1$. In both plots we include the oracle versions of IDS-UCB and UCB using $B = B^*$ for reference. However, note that it is not feasible to implement them in most practical settings. The solid and dashes lines represent the regret averaged over 200 repeated experiments, while the shaded bounds are 95% pointwise confidence bands.

5 Empirical bound information-directed sampling

We propose the empirical bound information-directed sampling (EBIDS) algorithm, which, like existing IDS algorithms, relies on a conservative upper bound B, but, unlike existing algorithms, EBIDS refines this value with accruing data to obtain a tighter high-probability bound on B^*. Our algorithm proceeds in two phases. Throughout the first T_B steps, which we will refer to as the *bound exploration phase*, the goal is to gather initial information on the optimal action as well as to improve the bound on B^*. At each time step t in this first phase, we use

$$\widehat{B}_t = \min\left\{B, \|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\|_2 + \beta_{t,\zeta_t(\delta)}^{1/2}(B)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}\right\}. \tag{8}$$

as the upper bound on B^*. The term $\beta_{t,\zeta_t(\delta)}(B)$ is defined in (6) and $\zeta_t(\delta) = \min\{\delta, 1/t^2\}$, where $\delta > 0$ is a user-specified parameter that determines the confidence level for the upper bound on B^*.

The geometric motivation for this estimator stems from the fact that the confidence set $\mathcal{E}^{\text{wls}}_{t,\zeta_t(\delta)}$ is an ellipsoid centered at $\widehat{\boldsymbol{\theta}}^{\text{wls}}_t$ with the longest semi-axis of length $\beta^{1/2}_{t,\zeta_t(\delta)}(B)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}$, so by adding it to $\|\widehat{\boldsymbol{\theta}}^{\text{wls}}_t\|_2$, by the triangle inequality, we obtain a conservative upper bound on the distance between the origin and the point of $\mathcal{E}^{\text{wls}}_{t,\zeta_t(\delta)}$ furthest from it. We prove in the Supplementary Materials that

$$\mathbb{P}\left(\bigcap_{t=1}^{\infty}\left\{\widehat{B}_t \geq B^*\right\}\right) \geq 1-\delta.$$

Continuing our description of the bound exploration phase, for any $t \leq T_B$ we use $\widehat{B}_t$ to obtain a UCB algorithm, which we will refer to as empirical bound UCB (EB-UCB) via

$$A_t^{\text{EB-UCB}(\zeta_t(\delta))} = \arg\max_{a\in\mathcal{A}} \left\langle \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}^{\text{wls}}_t \right\rangle + \beta^{1/2}_{t,\zeta_t(\delta)}(\widehat{B}_t)\|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}. \tag{9}$$

Subsequently, we use

$$\begin{aligned}\widehat{\Delta}_{t,\zeta_t(\delta)}(a) = &\left\langle \boldsymbol{\phi}\left(A_t^{\text{EB-UCB}(\zeta_t(\delta))}\right) - \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}^{\text{wls}}_t \right\rangle \\ &+ \beta^{1/2}_{t,\zeta_t(\delta)}(\widehat{B}_t)\left(\left\|\boldsymbol{\phi}\left(A_t^{\text{EB-UCB}(\zeta_t(\delta))}\right)\right\|_{\boldsymbol{W}_t^{-1}} + \|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}\right)\end{aligned} \tag{10}$$

as the gap estimate for any $a \in \mathcal{A}$. We define a new information gain criterion that combines model improvement (classic information gain) with bound improvement. The first component of our new information gain criterion is given by

$$I_t^{\text{EB-UCB}(\zeta_t(\delta))}(a) = \frac{1}{2}\log\left(\frac{\left\|\boldsymbol{\phi}\left(A_t^{\text{EB-UCB}(\zeta_t(\delta))}\right)\right\|^2_{\boldsymbol{W}_t^{-1}}}{\left\|\boldsymbol{\phi}\left(A_t^{\text{EB-UCB}(\zeta_t(\delta))}\right)\right\|^2_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)^{-1}}}\right), \tag{11}$$

for any $a \in \mathcal{A}$. It can be seen that this is analogous to the IDS-UCB information gain criterion considered by Kirschner & Krause (2018). To ensure that improvement in the bound on B^*, we introduce the second component of our information gain criterion I_t^B which is given by

$$I_t^B(a) = \frac{1}{2}\log\left(\|\boldsymbol{v}_t^{\min}\|^2_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)}\right) - \frac{1}{2}\log\left\{\lambda_{\min}(\boldsymbol{W}_t)\right\},$$

where $\boldsymbol{v}_t^{\min}$ is the unit-length eigenvector of $\boldsymbol{W}_t$ associated with the smallest eigenvalue $\lambda_{\min}(\boldsymbol{W}_t)$. The maximizer of $I_t^B(a)$ corresponds to the feature vector $\phi(a)$ which generates the most (weighted) information in the direction of the minimum eigenvector of the current information matrix. This direction corresponds to the longest axis of the confidence ellipsoid defined by the inverse information and is closely related to E-optimal experimental designs (Dette & Studden, 1993).

In order to balance exploration aimed at reducing the uncertainty about B^* and directly searching for the optimal arm in the initial phase, we use a mixture of information gain criteria, which we refer to as the bound-action mixture (BAM) criterion:

$$I_t^{\text{BAM}(\zeta_t(\delta))}(a) = \alpha I_t^B(a) + (1-\alpha)I_t^{\text{EB-UCB}(\zeta_t(\delta))}(a),$$

where $\alpha \in (0,1)$ is a parameter chosen by the user. Note that while we use the $I_t^{\text{EB-UCB}(\zeta_t(\delta))}$ information gain criterion in this instance, we could use any information gain criterion of choice instead. For notational convenience we drop the $\zeta_t(\delta)$ term and write $I_t^{\text{EB-UCB}}$ for $I_t^{\text{EB-UCB}(\zeta_t(\delta))}$ and I_t^{BAM} for $I_t^{\text{BAM}(\zeta_t(\delta))}$ since we will use $\zeta_t(\delta) = \min\{\delta, 1/t^2\}$ in the remainder of this manuscript.

Given the advantages of deterministic IDS and its strong performance in various experimental settings, we focus on this variant of IDS. Hence, we always select the action which minimizes the information ratio on the set $\mathcal{A}$, as given in (4). So at each time step $t \in \{1, \dots, T_B\}$ of the bound exploration phase we choose the action

$$A_t^{\mathrm{BAM}} = \arg\min_{a \in \mathcal{A}} \left\{ \Psi_t^{\mathrm{BAM}}(a) := \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}(a)}{I_t^{\mathrm{BAM}}(a)} \right\}.$$

Throughout the second phase, which we refer to as the *bound exploitation phase*, for any $t \geq T_B + 1$ we use

$$\tilde{B}_t = \min\left\{ B, \min_{\tau \leq t} \left\{ \|\widehat{\boldsymbol{\theta}}_\tau^{\mathrm{wls}}\|_2 + \beta^{1/2}_{\tau,\zeta_\tau(\delta)}(\widehat{B}_\tau) \lambda_{\min}(\boldsymbol{W}_\tau)^{-1/2} \right\} \right\}$$

as the upper bound on B^*, with $\widehat{B}_t$ defined in (8). During this phase we drop the bound information gain criterion I_t^B from the mixture and use only the $I_t^{\text{EB-UCB}}$ criterion. The quantity $\tilde{B}_t$ is used as the upper bound for B^* for both the gap estimate $\widehat{\Delta}_{t,\zeta_t(\delta)}$ and $I_t^{\text{EB-UCB}}$, which are defined in the same way as in equations (9), (10), and (11) with $\tilde{B}_t$ in place of $\widehat{B}_t$. We summarize this method in Algorithm 2. Note that in the second phase we could use any algorithm which requires explicit use of an upper bound on B^* by taking $B = \tilde{B}_t$ as that upper bound. Furthermore, we formulate this procedure specifically in the context of IDS, however, the approach of estimating a high-probability upper bound on the true parameter norm and using it to guide decision making can be thought of as a more general technique, rather than something specific only to IDS.

6 Regret analysis of EBIDS algorithm

In this section we present the regret and pseudo-regret bounds for both phases of the EBIDS algorithm. We defer the proofs of these propositions and relevant lemmas to the Supplementary Materials. For any t and $\xi_t > 0$, let E_{t,ξ_t} be the event

$$E_{t,\xi_t} = \left\{ \left\| \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}} \right\|^2_{\boldsymbol{W}_t} \leq \beta_{t,\xi_t}(B^*) \right\}, \tag{12}$$

and define $E_\delta = \bigcap_{t=1}^{\infty} E_{t,\delta}$. Note that by Theorem 2 we have $\mathbb{P}(E_\delta) \geq 1 - \delta$. The following proposition summarizes the regret and pseudo-regret bounds for EBIDS during the bound exploration phase.

Proposition 1. *For any $2 \leq T \leq T_B$ the regret $\mathcal{R}_T$ of Algorithm 2 is bounded above by*

$$\mathcal{R}_T \leq O\left(\frac{d \max\{U/\sqrt{\gamma}, \rho_{\max}\}}{\sqrt{1-\alpha}} \sqrt{T} \log T \sqrt{\log(1/\delta) + \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) + \gamma B^2} \right)$$

and whenever event E_δ holds the pseudo-regret $\mathcal{PR}_T$ is bounded above by the same rate.

We also provide guarantees on the estimated upper bound on B^* after the bound exploration phase. This, in turn, will allow us to obtain an improved bound for the regret and pseudo-regret in the subsequent phase.

Proposition 2. *For any constant $g > 0$, with sufficiently large T_B and sufficiently large α, whenever event E_δ holds we have $B^* \leq \tilde{B}_t \leq (1+g)B^*$ for any $t \geq T_B + 1$.*

Please see Section 10.6 in the Supplementary Materials for the exact constants required as lower bounds for T_B and α depending on g. Finally, using the results of Proposition 2 we are able to establish a regret bound for the second phase of EBIDS which is independent of the original conservative bound B.

Proposition 3. *For any constant $g > 0$, with sufficiently large T_B and sufficiently large α, with probability at least $1 - \delta$ the regret and pseudo-regret of Algorithm 2 are both bounded above by $O\left(dU\rho_{\max}(1+g)B^*\sqrt{T}\log T\right)$, for any $T \geq T_B + 1$.*

Algorithm 2 EBIDS

Input: Action set $\mathcal{A}$, penalty parameter $\gamma > 0$, noise function $\rho : \mathcal{A} \to \mathbb{R}_+$, feature function $\boldsymbol{\phi} : \mathcal{A} \to \mathbb{R}$, conservative true parameter norm bound B, number of bound exploration steps T_B, information gain mixture parameter $\alpha \in (0, 1)$, error tolerance parameter $\delta \in (0, 1)$.

For $t = 1, 2, \ldots, T_B$:

Compute $\boldsymbol{W}_t$ and $\widehat{\boldsymbol{\theta}}_t^{\text{wls}}$ using (5)

$\widehat{B}_t \leftarrow \min \left\{ B, \|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\|_2 + \beta_{t,\zeta_t(\delta)}^{1/2}(B)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2} \right\}$

$A_t^{\text{EB-UCB}} \leftarrow \arg\max_{a\in\mathcal{A}} \left\langle \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}} \right\rangle + \beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}$

$I_t^{\text{EB-UCB}}(a) \leftarrow \frac{1}{2}\log\left(\left\|\boldsymbol{\phi}(A_t^{\text{EB-UCB}})\right\|_{\boldsymbol{W}_t^{-1}}^2\right) - \frac{1}{2}\log\left(\left\|\boldsymbol{\phi}(A_t^{\text{EB-UCB}})\right\|_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)^{-1}}^2\right)$

$I_t^B(a) \leftarrow \frac{1}{2}\log\left(\|\boldsymbol{v}_t^{\min}\|_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)}^2\right) - \frac{1}{2}\log\left\{\lambda_{\min}(\boldsymbol{W}_t)\right\}$

$I_t^{\text{BAM}}(a) \leftarrow \alpha I_t^B(a) + (1-\alpha)I_t^{\text{EB-UCB}}(a)$

$\widehat{\Delta}_{t,\zeta_t(\delta)}(a) \leftarrow \left\langle \boldsymbol{\phi}(A_t^{\text{EB-UCB}}) - \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}} \right\rangle + \beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\left(\|\boldsymbol{\phi}(A_t^{\text{EB-UCB}})\|_{\boldsymbol{W}_t^{-1}} + \|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}\right)$

$A_t \leftarrow \arg\min_{a\in\mathcal{A}} \widehat{\Delta}_{t,\zeta_t(\delta)}^2(a)/I_t^{\text{BAM}}(a)$

Play A_t, observe $Y_t = \langle \boldsymbol{\phi}(A_t), \boldsymbol{\theta}^* \rangle + \eta_t$

For $t = T_B + 1, T_B + 2, \ldots, T$:

Compute $\boldsymbol{W}_t$ and $\widehat{\boldsymbol{\theta}}_t^{\text{wls}}$ using (5)

$\widehat{B}_t \leftarrow \min \left\{ B, \|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\|_2 + \beta_{t,\zeta_t(\delta)}^{1/2}(B)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2} \right\}$

$\tilde{B}_t \leftarrow \min \left\{ B, \min_{\tau\le t}\left\{ \|\widehat{\boldsymbol{\theta}}_\tau^{\text{wls}}\|_2 + \beta_{\tau,\zeta_\tau(\delta)}^{1/2}(\widehat{B}_t)\lambda_{\min}(\boldsymbol{W}_\tau)^{-1/2} \right\}\right\}$

$A_t^{\text{EB-UCB}} \leftarrow \arg\max_{a\in\mathcal{A}} \left\langle \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}} \right\rangle + \beta_{t,\zeta_t(\delta)}(\tilde{B}_t)^{1/2}\|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}$

$I_t^{\text{EB-UCB}}(a) \leftarrow \frac{1}{2}\log\left(\left\|\boldsymbol{\phi}(A_t^{\text{EB-UCB}})\right\|_{\boldsymbol{W}_t^{-1}}^2\right) - \frac{1}{2}\log\left(\left\|\boldsymbol{\phi}(A_t^{\text{EB-UCB}})\right\|_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)^{-1}}^2\right)$

$\widehat{\Delta}_{t,\zeta_t(\delta)}(a) \leftarrow \left\langle \boldsymbol{\phi}(A_t^{\text{EB-UCB}}) - \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}} \right\rangle + \beta_{t,\zeta_t(\delta)}^{1/2}(\tilde{B}_t)\left(\|\boldsymbol{\phi}(A_t^{\text{EB-UCB}})\|_{\boldsymbol{W}_t^{-1}} + \|\boldsymbol{\phi}(a)\|_{\boldsymbol{W}_t^{-1}}\right)$

$A_t \leftarrow \arg\min_{a\in\mathcal{A}} \widehat{\Delta}_{t,\zeta_t(\delta)}^2(a)/I_t^{\text{EB-UCB}}(a)$

Play A_t, observe $A_t = \langle \boldsymbol{\phi}(A_t), \boldsymbol{\theta}^* \rangle + \eta_t$

Similarly, we give the exact constants required as lower bounds for T_B and α in Supplementary Materials, in Section 10.7. Thus, Propositions 1 and 3 together give us regret and pseudo-regret guarantees for both bound exploration phase and the subsequent bound exploitation phase of EBIDS. This is different from Gales et al. (2022) who do not control the regret in the initial stages of their norm-agnostic algorithms.

7 Simulation study

We evaluate the performance of EBIDS using simulation studies and compare it with the norm-agnostic competitor algorithms NAOFUL and OLSOFUL by Gales et al. (2022) which also aim at alleviating the dependence on access to a high-quality bound on the true parameter norm. We include the EB-UCB algorithm to demonstrate the advantage of using the IDS strategy in addition to utilizing the empirical norm bound. We run the comparison also against the oracle versions of EBIDS, IDS-UCB and UCB with access to the true value of B^*. We use the same setting as in the simulation illustration in Section 4 with $\boldsymbol{\theta}^* = [-5, 1, 1, 1.5, 2]^\top$ as the true parameter and ten arms with features sampled from Unif$[-1/\sqrt{5}, 1/\sqrt{5}]$. The error distribution for the first five arms are standard normal and for the remaining five arms they are normal with mean zero and variance 0.2. We take the conservative $B = 100$ as the assumed upper bound on B^*. Both the oracle and non-oracle versions of EBIDS use $\alpha = 0.5$, giving equal weight to both components of the BAM criterion, and run the bound exploration phase for $T_B = 50$ steps.

Figure 2 shows the mean regret averaged over 200 repeated experiments with $T = 500$ steps along with 95% pointwise confidence bounds. As we can see, EB-UCB is competitive with NAOFUL and OLSOFUL, while EBIDS performs best among all the algorithms which do not have access to the true parameter norm. It achieves significantly lower regret than IDS-UCB and UCB. Meanwhile, the performance of oracle EBIDS is better than that of oracle UCB and almost indistinguishable from the one achieved by oracle IDS-UCB.

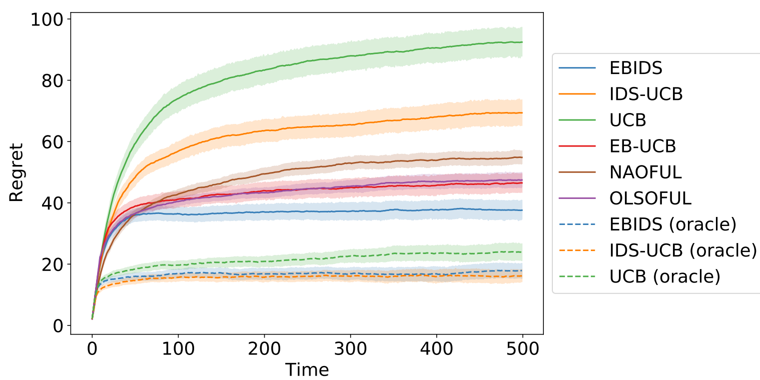


Figure 2: Regret incurred by EBIDS, EB-UCB, NAOFUL, OLSOFUL, IDS-UCB and UCB with conservative $B = 100$. We include the oracle versions of EBIDS, IDS-UCB and UCB using $B = B^*$ for reference. The solid and dashes lines represent the regret averaged over 200 repeated experiments, while the shaded bounds represent 95% pointwise confidence bounds.

We also perform an ablation study to determine the sensitivity of EBIDS to the tuning parameter α and the length T_B of the bound exploration phase. We consider all combinations of $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$ and $T_B \in \{50, 100\}$. We use the same setting as above and present the results for $T = 500$ steps averaged over 200 repeated experiments in Figure 3. Using $T_B = 50$ leads to somewhat better results than $T_B = 100$ and $\alpha = 0.3$ performs best for both values of T_B. However, the performance is similar for all considered combinations of the tuning parameters, especially compared to the differences in performance of the competitor algorithms. This shows that while EBIDS, like most other bandit algorithms, uses tuning parameters, its performance is not very sensitive to their choice, with several considered combinations of α and T_B achieving practically indistinguishable regret.

8 Discussion

Bandit algorithms often require access to a high-quality upper bound on the Euclidean norm of the true parameter vector in order to achieve good performance. In practice, such information is rarely available *a priori*, which can lead to significant regret accumulation. Despite its prevalence, this problem has received relatively little attention in the bandit literature. We introduced the empirical bound information-directed sampling (EBIDS) algorithm which addresses this challenge by iteratively refining a high-probability upper bound on the true parameter norm. We developed a novel information criterion that balances tightening the bound on the true parameter norm and explicitly searching for the optimal arm. In simulation experiments, EBIDS showed improved performance compared to the competing norm-agnostic algorithms. Furthermore, we proved regret bounds that

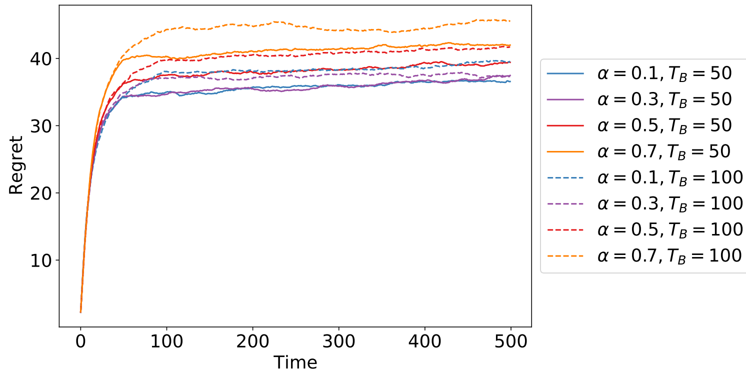


Figure 3: Average regret for EBIDS averaged over 200 repeated experiments with $T = 500$ steps under different values of the tuning parameter α and the length T_B of the bound exploration phase.

eventually do not depend on the initially assumed bound for the parameter norm, and unlike prior regret guarantees, our bounds are anytime in that they apply to all phases of the algorithm.

Broader Impact Statement

This paper introduces novel methodology for frequentist IDS that does not require strong prior information on the norm of the true parameter indexing the reward model. Our methodology, which involves a novel information criterion, can be viewed as a general approach to balancing bound improvement and regret minimization that is applicable in a wide range of UCB and IDS bandit algorithms.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Amir Ahmadi-Javid, Pardis Seyedi, and Siddhartha S Syam. A survey of healthcare facility location. *Computers & Operations Research*, 79:223–263, 2017.

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

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

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, pp. 1844–1854. PMLR, 2020.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

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

Holger Dette and William J Studden. Geometry of E-optimality. *The Annals of Statistics*, 21(1): 416–433, 1993.

Joel N. Franklin. *Matrix Theory*. Prentice-Hall, 1968.

Spencer B. Gales, Sunder Sethuraman, and Kwang-Sung Jun. Norm-agnostic linear bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 73–91. PMLR, 2022.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19, pp. 359–376. PMLR, 2011.

Claudio Gentile and Francesco Orabona. On multilabel classification and ranking with bandit feedback. *Journal of Machine Learning Research*, 15(70):2451–2487, 2014.

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, pp. 1396–1404. PMLR, 2021.

Botao Hao and Tor Lattimore. Regret bounds for information-directed reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yu-Heng Hung, Ping-Chun Hsieh, Xi Liu, and P. R. Kumar. Reward-biased maximum likelihood estimation for linear stochastic bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7874–7882, 2021.

Johannes Kirschner. *Information-Directed Sampling-Frequentist Analysis and Applications*. PhD thesis, ETH Zurich, 2021.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pp. 358–384. PMLR, 2018.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 2328–2369. PMLR, 2020.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvari. Asymptotically optimal information-directed sampling. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 2777–2821. PMLR, 2021.

Tomáš Kocák, Rémi Munos, Branislav Kveton, Shipra Agrawal, and Michal Valko. Spectral bandits. *Journal of Machine Learning Research*, 21(218):1–44, 2020.

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

David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause. Information directed reward learning for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *International Conference on Learning Representations*, 2019.

Francesco Orabona and Nicolo Cesa-Bianchi. Better algorithms for selective sampling. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 433–440, 2011.

My Phan, Yasin Abbasi Yadkori, and Justin Domke. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Zhilei Wang, Pranjal Awasthi, Christoph Dann, Ayush Sekhari, and Claudio Gentile. Neural active learning with performance guarantees. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Justin Weltz, Alex Volfovsky, and Eric B. Laber. Reinforcement learning methods in public health. *Clinical Therapeutics*, 44(1):139–154, 2022.

Justin Weltz, Tanner Fiez, Alexander Volfovsky, Eric Laber, Blake Mason, Houssam Nassif, and Lalit Jain. Experimental designs for heteroskedastic variance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ruitu Xu, Yifei Min, and Tianhao Wang. Noise-adaptive thompson sampling for linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 11492–11502. PMLR, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

In these supplementary materials we provide the proofs to the propositions we have stated in the paper.

9 Notation and lemmas

We begin by introducing some notation and basic facts. For any unit vector $\boldsymbol{v} \in \mathbb{R}^d$ and any $a \in \mathcal{A}$, let $\psi_{\boldsymbol{v}}(\boldsymbol{\phi}(a)), \psi_{\boldsymbol{v}}^{\perp}(\boldsymbol{\phi}(a)) \in \mathbb{R}$ denote the orthogonal decomposition of $\boldsymbol{\phi}(a)$, i.e.,

$$\boldsymbol{\phi}(a) = \psi_{\boldsymbol{v}}(\boldsymbol{\phi}(a))\boldsymbol{v} + \psi_{\boldsymbol{v}}^{\perp}(\boldsymbol{\phi}(a))\boldsymbol{v}^{\perp},$$

where $\|\boldsymbol{v}^{\perp}\|_2 = 1$ and $\boldsymbol{v}^{\perp} \perp \boldsymbol{v}$. Let

$$\kappa = \min_{\boldsymbol{v}\in\mathbb{R}^d \text{ s.t. } \|\boldsymbol{v}\|_2=1} \max_{a\in\mathcal{A}} \left\{\rho(a)^{-2}\psi_{\boldsymbol{v}}(\boldsymbol{\phi}(a))^2\right\}. \tag{13}$$

Note that $\kappa > 0$. Let

$$\omega_t(a) = \rho(a)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\phi(a))^2. \tag{14}$$

Also, note that for any $a \in \mathcal{A}$ we have

$$\|\boldsymbol{\phi}(a)\|^2_{\boldsymbol{W}_t^{-1}} = \sum_{i=1}^{d} \psi_{\boldsymbol{v}_i}(\boldsymbol{\phi}(a))^2\lambda_i^{-1}$$

where $\{(\lambda_i, \boldsymbol{v}_i)\}_{i=1}^d$ are the eigenvalue-eigenvector pairs of $\boldsymbol{W}_t$. Hence for every $t \geq 1$ and $a \in \mathcal{A}$ we have

$$\|\boldsymbol{\phi}(a)\|_2^2\lambda_{\max}(\boldsymbol{W}_t)^{-1} \leq \|\boldsymbol{\phi}(a)\|^2_{\boldsymbol{W}_t^{-1}} \leq \|\boldsymbol{\phi}(a)\|_2^2\lambda_{\min}(\boldsymbol{W}_t)^{-1},$$

so

$$L^2\lambda_{\max}(\boldsymbol{W}_t)^{-1} \leq \|\boldsymbol{\phi}(a)\|^2_{\boldsymbol{W}_t^{-1}} \leq U^2\lambda_{\min}(\boldsymbol{W}_t)^{-1}. \tag{15}$$

Also from Cauchy-Schwarz inequality

$$\left\langle \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}} \right\rangle^2 \leq \|\boldsymbol{\phi}(a)\|_2^2 \left\|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_2^2 \leq U^2 \left\|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_2^2 \tag{16}$$

From Weyl's inequality (Franklin, 1968), for any positive semi-definite matrices $\boldsymbol{A}, \boldsymbol{B}$ we have

$$\lambda_{\max}(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{A}) + \lambda_{\max}(\boldsymbol{B}).$$

Thus, for every $t \geq 1$ we have

$$\lambda_{\max}(\boldsymbol{W}_t) \leq \lambda_{\max}(\gamma\boldsymbol{I}_d) + \sum_{\tau=1}^{t-1}\lambda_{\max}(\rho(a_\tau)^{-2}\boldsymbol{\phi}(a_\tau)\boldsymbol{\phi}(a_\tau)^\top) \leq \gamma + (t-1)\rho_{\min}^{-2}U^2, \tag{17}$$

so from (15), for any $t \geq 1$ we have

$$\|\boldsymbol{\phi}(a)\|^2_{\boldsymbol{W}_t^{-1}} \geq \frac{L^2}{\gamma + (t-1)\rho_{\min}^{-2}U^2} \geq \frac{L^2}{t(\gamma + \rho_{\min}^{-2}U^2)}. \tag{18}$$

Also from (17) for $T \geq 2$ we have

$$\begin{aligned}
\log\left(\frac{\det(\boldsymbol{W}_T)}{\det(\boldsymbol{W}_1)}\right) =& \log(\det(\boldsymbol{W}_T)) - \log(\det(\gamma \boldsymbol{I}_d)) \leq d\log(\gamma + (T-1)\rho_{\min}^{-2}U^2) - d\log\gamma \\
=& d\log\left[1 + (T-1)\frac{\rho_{\min}^{-2}U^2}{\gamma}\right] \leq d\log\left[(T-1)\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right)\right] \\
=& d\log(T-1) + d\log\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right). \qquad (19)
\end{aligned}$$

Applying the data processing inequality (Cover & Thomas, 2012) in an analogous way as Kirschner & Krause (2018), we obtain

$$I_t^{\text{EB-UCB}}(a) \leq \frac{1}{2}\log\left(\frac{\det(\boldsymbol{W}_t + \rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)}{\det(\boldsymbol{W}_t)}\right) = \frac{1}{2}\log\left(1 + \rho(a)^{-2}\|\boldsymbol{\phi}(a)\|^2_{\boldsymbol{W}_t^{-1}}\right) \quad (20)$$

for any $a \in \mathcal{A}$. So from (19) we get

$$\sum_{t=1}^{T} I_t^{\text{EB-UCB}}(a_t) \leq \frac{1}{2}\log\left(\frac{\det(\boldsymbol{W}_{T+1})}{\det(\boldsymbol{W}_1)}\right) \leq \frac{1}{2}d\log T + \frac{1}{2}d\log\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right) = O(d\log T), \quad (21)$$

for any sequence $\{a_t\}_{t=1}^T \subset \mathcal{A}$.

We now state and prove some additional lemmas that will be useful throughout the proofs of Propositions 1 - 3.

Lemma 1. *Let $\widehat{\Delta}_t : \mathcal{A} \to \mathbb{R}_+$ be a gap estimate function and let $I_t^X, I_t^Y : \mathcal{A} \to \mathbb{R}_+$ be two information gain criteria. Let I_t^{XY} be the mixture information gain criterion given by*

$$I_t^{XY}(a) = \alpha I_t^X(a) + (1-\alpha)I_t^Y(a)$$

for some $\alpha \in (0,1)$. Consider now a deterministic IDS algorithm which at each time step t plays action a_t^{XY} given by

$$a_t^{XY} = \arg\min_{a\in\mathcal{A}} \frac{\widehat{\Delta}_t^2(a)}{I_t^{XY}(a)}$$

Then at each time step t the information gain on to the first criterion I_t^X is lower-bounded by

$$I_t^X\left(a_t^{XY}\right) \geq \frac{\widehat{\Delta}_t^2\left(a_t^{XY}\right)}{\widehat{\Delta}_t^2\left(a_t^{I,X}\right)} I_t^X\left(a_t^{I,X}\right) - \frac{1-\alpha}{\alpha} I_t^Y\left(a_t^{XY}\right),$$

where $a_t^{I,X} = \arg\max_{a\in\mathcal{A}} I_t^X(a)$.

Lemma 2. *Recall the definition $\omega_t(a) = \rho(a)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\boldsymbol{\phi}(a))^2$. For any $T \geq 1$ and any sequence of actions $\{a_t\}_{t=1}^T$ we have*

$$\lambda_{\min}(\boldsymbol{W}_{T+1}) \geq \gamma - \rho_{\min}^{-2}U^2 + \frac{1}{d}\sum_{t=1}^{T}\omega_t(a_t).$$

Lemma 3. *Let $\{x_t\}_{t=1}^{T+1} \subset [0,U]$ be a bounded sequence for some constant $U > 0$. Then for any constant $c > 0$ we have*

$$\sum_{t=1}^{T} \frac{x_{t+1}}{c + \sum_{\tau=1}^{t} x_\tau} \leq \log T + \frac{U}{c} + 1.$$

10 Proofs of theoretical results

In this section we provide the proofs of Theorem 1, Lemmas 1 - 3, and Propositions 1 - 3.

10.1 Proof of Theorem 1

Recall that by Cauchy-Schwarz inequality, for any random variables $\{X_t\}_{t\in G}, \{Y_t\}_{t\in G}$ with nonnegative support, with probability 1 we have

$$\sum_{t\in G} \sqrt{X_t Y_t} \leq \sqrt{\left(\sum_{t\in G} X_t\right)\left(\sum_{t\in G} Y_t\right)},$$

and for any random variables X, Y with nonnegative support we have

$$\mathbb{E}\left[\sqrt{XY}\right] \leq \sqrt{\mathbb{E}[X]\mathbb{E}[Y]}.$$

Hence if $\widehat{\Delta}(A_t) \geq \Delta(A_t)$, for all $t \in G$, then with probability 1 we have

$$\sum_{t\in G} \Delta\left(A_t\right) \leq \sum_{t\in G} \widehat{\Delta}_t\left(A_t\right) = \sum_{t\in G} \sqrt{\Psi_t(A_t) I_t(A_t)} \leq \sqrt{\left[\sum_{t\in G} \Psi_t(A_t)\right]\left[\sum_{t\in G} I_t(A_t)\right]}.$$

Also

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in G} \widehat{\Delta}_t\left(A_t\right)\right] =& \mathbb{E}\left[\sum_{t\in G} \sqrt{\Psi_t(A_t) I_t\left(A_t\right)}\right] \leq \mathbb{E}\left(\sqrt{\left[\sum_{t\in G} \Psi_t\left(A_t\right)\right]\left[\sum_{t\in G} I_t\left(A_t\right)\right]}\right) \\ \leq& \sqrt{\mathbb{E}\left[\sum_{t\in G} \Psi_t\left(A_t\right)\right]\mathbb{E}\left[\sum_{t\in G} I_t\left(A_t\right)\right]}. \quad \square\end{aligned}$$

10.2 Proof of Lemma 1

By the definition of a_t^{XY} we have

$$\frac{\widehat{\Delta}_t^2(a_t^{XY})}{\alpha I_t^X(a_t^{XY}) + (1-\alpha) I_t^Y(a_t^{XY})} \leq \frac{\widehat{\Delta}_t^2(a_t^{I,X})}{\alpha I_t^X(a_t^{I,X}) + (1-\alpha) I_t^Y(a_t^{I,X})},$$

hence

$$\alpha I_t^X(a_t^{XY}) + (1-\alpha) I_t^Y(a_t^{XY}) \geq \frac{\widehat{\Delta}_t^2(a_t^{XY})}{\widehat{\Delta}_t^2(a_t^{\text{I-X}})}\left[\alpha I_t^X(a_t^{I,X}) + (1-\alpha) I_t^Y(a_t^{I,X})\right],$$

and thus

$$\begin{aligned}I_t^X(a_t^{XY}) \geq& \frac{\widehat{\Delta}_t^2(a_t^{XY})}{\widehat{\Delta}_t^2(a_t^{I,X})} I_t^X(a_t^{I,X}) + \frac{(1-\alpha)}{\alpha} \cdot \frac{\widehat{\Delta}_t^2(a_t^{XY})}{\widehat{\Delta}_t^2(a_t^{I,X})} I_t^Y(a_t^{I,X}) - \frac{1-\alpha}{\alpha} I_t^Y(a_t^{XY}) \\ \geq& \frac{\widehat{\Delta}_t^2(a_t^{XY})}{\widehat{\Delta}_t^2(a_t^{I,X})} I_t^X(a_t^{I,X}) - \frac{1-\alpha}{\alpha} I_t^Y(a_t^{XY}). \quad \square\end{aligned}$$

10.3 Proof of Lemma 2

Recall that we define $\lambda_1^{(t)}, \ldots, \lambda_d^{(t)}$ as the (not necessarily ordered) eigenvalues of $\boldsymbol{W}_t$. Let

$$i^*(t) = \underset{1\leq i\leq d}{\arg\min}\, \lambda_i^{(t)}.$$

By Weyl's inequality (Franklin, 1968), for any symmetric positive semi-definite matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times m}$ we have

$$\lambda_{(i)}(\boldsymbol{A}+\boldsymbol{B}) \geq \lambda_{(i)}(\boldsymbol{A}) \tag{22}$$

where $\lambda_{(i)}(\boldsymbol{A})$ is the i-th largest eigenvalue of $\boldsymbol{A}$ for any $1 \leq i \leq m$. Let $\boldsymbol{v}_t^{\min}$ be the unit eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{W}_t$. Then for any $1 \leq i \leq d$ we have

$$\begin{aligned}\lambda_{(i)}(\boldsymbol{W}_{t+1}) =& \lambda_{(i)}\left(\boldsymbol{W}_t + \rho(a_t)^{-2}\boldsymbol{\phi}(a_t)\boldsymbol{\phi}(a_t)^\top\right)\\ =& \lambda_{(i)}\Big(\boldsymbol{W}_t + \rho(a_t)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\boldsymbol{\phi}(a))\boldsymbol{v}_t^{\min}(\boldsymbol{v}_t^{\min})^\top\\ &+ \rho(a_t)^{-2}\psi^{\perp}_{\boldsymbol{v}_t^{\min\perp}}(\boldsymbol{\phi}(a))\boldsymbol{v}_t^{\min\perp}(\boldsymbol{v}_t^{\min\perp})^\top\Big)\\ \geq& \lambda_{(i)}\left(\boldsymbol{W}_t + \rho(a_t)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\boldsymbol{\phi}(a))\boldsymbol{v}_t^{\min}(\boldsymbol{v}_t^{\min})^\top\right)\\ =& \lambda_{(i)}\left(\boldsymbol{W}_t + \omega_t(a_t)\boldsymbol{v}_t^{\min}(\boldsymbol{v}_t^{\min})^\top\right).\end{aligned}$$

Note that the matrix $\boldsymbol{W}_t + \omega_t(a_t)\boldsymbol{v}_t^{\min}(\boldsymbol{v}_t^{\min})^\top$ has the same eigenvectors as $\boldsymbol{W}_t$ and the smallest eigenvalue of $\boldsymbol{W}_t$, i.e., the one corresponding to $\boldsymbol{v}_t^{\min}$ is increased by $\omega_t(a_t)$. So for any t we can order the eigenvalues $\lambda_1^{(t+1)}, \ldots, \lambda_d^{(t+1)}$ in such way that $\lambda_i^{(t+1)} \geq \lambda_i^{(t)}$ and

$$\lambda_{i^*(t)}^{(t+1)} \geq \lambda_{i^*(t)}^{(t)} + \omega_t(a_t).$$

Since we have d eigenvalues and at each time step t we add at least $\omega_t(a_t)$ to the smallest eigenvalue at that time step without reducing the other ones we have

$$\lambda_{i^*(T)}^{(T)} - \lambda_{i^*(1)}^{(1)} + \omega_T(a_T) \geq \frac{1}{d}\sum_{t=1}^{T}\omega_t(a_t).$$

Note that $\lambda_1^{(1)} = \ldots = \lambda_d^{(1)} = \gamma$ and $\omega_T(a_T) \leq \rho_{\min}^{-2}U^2$, so

$$\lambda_{\min}(\boldsymbol{W}_{T+1}) = \lambda_{i^*(T+1)}^{(T+1)} \geq \lambda_{i^*(T)}^{(T)} \geq \gamma - \rho_{\min}^{-2}U^2 + \frac{1}{d}\sum_{t=1}^{T}\omega_t(a_t). \quad \square$$

10.4 Proof of Lemma 3

Let

$$f(x_1, \ldots, x_{T+1}) = \sum_{t=1}^{T}\frac{x_{t+1}}{c + \sum_{\tau=1}^{t}x_\tau}$$

We use induction to show that the maximum of f is achieved at $x_1 = 0$ and

$$x_2 = x_3 = \ldots = x_{T+1} = U.$$

Note that for any $\tilde{x}_1, \ldots, \tilde{x}_T \in [0, U]$ we have

$$\operatorname*{arg\,max}_{x_{T+1}\in[0,U]} f(\tilde{x}_1, \ldots, \tilde{x}_T, x_{T+1}) = U.$$

Suppose that for any $t \geq 2$ it holds that for any $t \leq k \leq T$ and any $\tilde{x}_1, \ldots, \tilde{x}_k \in [0, U]$ we have

$$(x_{k+1}^*, \ldots, x_{T+1}^*) := \operatorname*{arg\,max}_{x_{k+1},\ldots,x_{T+1}\in[0,U]} f(\tilde{x}_1, \ldots, \tilde{x}_k, x_{k+1}, \ldots, x_{T+1}) = (U, \ldots, U) \in \mathbb{R}^{T-k+1}. \tag{23}$$

Take any $\tilde{x}_1, \ldots, \tilde{x}_{t-1} \in [0, U]$. Then by taking $k = t+1$ the above statement gives us

$$\max_{x_t,x_{t+1},\ldots,x_{T+1}\in[0,U]} f(\tilde{x}_1, \ldots, \tilde{x}_{t-1}, x_t, x_{t+1}, \ldots, x_{T+1}) = \max_{x_t,x_{t+1}\in[0,U]} f(\tilde{x}_1, \ldots, \tilde{x}_{t-1}, x_t, x_{t+1}, U, \ldots, U).$$

Let

$$(\check{x}_t, \check{x}_{t+1}) = \operatorname*{arg\,max}_{x_t,x_{t+1}\in[0,U]} f(\tilde{x}_1, \ldots, \tilde{x}_{t-1}, x_t, x_{t+1}, \ldots, x_{T+1}).$$

Note that $\check{x}_{t+1} = x^*_{t+1} = U$ by taking the induction statement with $k = t$. For notational convenience let $b = c + \sum_{\tau=1}^{t-1} \tilde{x}_\tau$. Then

$$(\check{x}_t, \check{x}_{t+1}) = \underset{x_t, x_{t+1} \in [0,U]}{\arg\max} \left\{ \frac{x_t}{b} + \frac{x_{t+1}}{b + x_t} + \sum_{\tau=0}^{T-t-1} \frac{U}{b + x_t + x_{t+1} + \tau U} \right\}.$$

Let

$$g_t(x_t, x_{t+1}) = \frac{x_t}{b} + \frac{x_{t+1}}{b + x_t} + \sum_{\tau=0}^{T-t-1} \frac{U}{b + x_t + x_{t+1} + \tau U}.$$

Suppose that $\check{x}_t = x$ for some $0 \leq x < U$. Note that

$$g_t(U, x) - g_t(x, U) = \left(\frac{U}{b} + \frac{x}{b+U} \right) - \left(\frac{x}{b} + \frac{U}{b+x} \right) = \frac{Ux(U-x)}{b(b+U)(b+x)} > 0.$$

So $g_t(U, x) > g_t(x, U) = g_t(\check{x}_t, \check{x}_{t+1})$ which is a contradiction, since $(\check{x}_t, \check{x}_{t+1})$ is the maximizer of $g_t(x_t, x_{t+1})$. So $\check{x}_t = U$. Thus, we have shown that for any $\tilde{x}_1, \ldots, \tilde{x}_{t-1} \in [0, U]$ we have

$$(x^*_t, \ldots, x^*_{T+1}) = \underset{x_t, \ldots, x_{T+1} \in [0,U]}{\arg\max} f(\tilde{x}_1, \ldots, \tilde{x}_{t-1}, x_t, \ldots, x_{T+1}) = (U, \ldots, U) \in \mathbb{R}^{T-t+2}.$$

Hence by induction we get that for any $\tilde{x}_1 \in [0, U]$ we have

$$\underset{x_2, \ldots, x_{T+1} \in [0,U]}{\arg\max} f(\tilde{x}_1, x_2, \ldots, x_{T+1}) = (U, \ldots, U) \in \mathbb{R}^T.$$

Clearly

$$\underset{x_1 \in [0,U]}{\arg\max} f(x_1, U, \ldots, U) = 0,$$

so

$$\begin{aligned} \max_{x_1, \ldots, x_{T+1} \in [0,U]} f(x_1, \ldots, x_{T+1}) = & f(0, U, \ldots, U) = \sum_{t=1}^{T} \frac{U}{c + (t-1)U} \\ \leq & \frac{U}{c} + \sum_{t=2}^{T} \frac{1}{t-1} < \log T + \frac{U}{c} + 1. \quad \square \end{aligned}$$

10.5 Proof of Proposition 1

From Theorem 1, for any $T \leq T_B$ we have

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \widehat{\Delta}_{t,\zeta_t(\delta)}(A_t^{\text{BAM}}) \right] \leq & \sqrt{\left(\mathbb{E}\left[\sum_{t=1}^{T} \Psi_t^{\text{BAM}}\left(A_t^{\text{BAM}}\right) \right] \right) \left(\mathbb{E}\left[\sum_{t=1}^{T} I_t^{\text{BAM}}\left(A_t^{\text{BAM}}\right) \right] \right)} \\ \leq & \sqrt{\left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{BAM}}\right)}{I_t^{\text{BAM}}\left(A_t^{\text{BAM}}\right)} \right] \right) \left(\mathbb{E}\left[\sum_{t=1}^{T} I_t^{\text{BAM}}\left(A_t^{\text{BAM}}\right) \right] \right)} \\ = & \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{BAM}}\right)}{\alpha I_t^{B}\left(A_t^{\text{BAM}}\right) + (1-\alpha) I_t^{\text{EB-UCB}}\left(A_t^{\text{BAM}}\right)} \right]} \\ & \times \sqrt{\alpha \mathbb{E}\left[\sum_{t=1}^{T} I_t^{B}\left(A_t^{\text{BAM}}\right) \right] + (1-\alpha) \mathbb{E}\left[\sum_{t=1}^{T} I_t^{\text{EB-UCB}}\left(A_t^{\text{BAM}}\right) \right]}. \end{aligned}$$

By the definition of A_t^{BAM} we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{BAM}}\right)}{\alpha I_t^B\left(A_t^{\text{BAM}}\right)+(1-\alpha)I_t^{\text{EB-UCB}}\left(A_t^{\text{BAM}}\right)}\right] \leq& \mathbb{E}\left[\sum_{t=1}^{T}\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{EB-UCB}}\right)}{\alpha I_t^{\text{EB-UCB}}\left(A_t^{\text{EB-UCB}}\right)+(1-\alpha)I_t^B\left(A_t^{\text{EB-UCB}}\right)}\right] \\ \leq& \frac{1}{1-\alpha}\mathbb{E}\left[\sum_{t=1}^{T}\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{EB-UCB}}\right)}{I_t^{\text{EB-UCB}}\left(A_t^{\text{EB-UCB}}\right)}\right]. \end{aligned} \tag{24}$$

The next couple of steps are similar to the analysis by Kirschner (2021). Let $a_t^{\text{EB-UCB}}$ be the realization of $A_t^{\text{EB-UCB}}$. From the Sherman-Morrison formula, we obtain

$$\left(\boldsymbol{W}_t+\rho(a_t^{\text{EB-UCB}})^{-2}\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\boldsymbol{\phi}(a_t^{\text{EB-UCB}})^\top\right)^{-1}=\boldsymbol{W}_t^{-1}-\frac{\rho(a_t^{\text{EB-UCB}})^{-2}\boldsymbol{W}_t^{-1}\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\boldsymbol{\phi}(a_t^{\text{EB-UCB}})^\top\boldsymbol{W}_t^{-1}}{1+\rho(a_t^{\text{EB-UCB}})^{-2}\boldsymbol{\phi}(a_t^{\text{EB-UCB}})^\top\boldsymbol{W}_t^{-1}\boldsymbol{\phi}(a_t^{\text{EB-UCB}})}$$

so

$$\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\left(\boldsymbol{W}_t+\rho(a_t^{\text{EB-UCB}})^{-2}\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\boldsymbol{\phi}(a_t^{\text{EB-UCB}})^\top\right)^{-1}}=\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\boldsymbol{W}_t^{-1}}-\frac{\rho(a_t^{\text{EB-UCB}})^{-2}\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^4_{\boldsymbol{W}_t^{-1}}}{1+\rho(a_t^{\text{EB-UCB}})^{-2}\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\boldsymbol{W}_t^{-1}}}.$$

Thus

$$I_t^{\text{EB-UCB}}\left(a_t^{\text{EB-UCB}}\right)=\frac{1}{2}\log\left(1+\rho(a_t^{\text{EB-UCB}})^{-2}\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\boldsymbol{W}_t^{-1}}\right).$$

From (15), we have $\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|_{\boldsymbol{W}_t^{-1}}\leq U^2\gamma^{-1}$. Thus, using the fact that $\log(1+x)\geq\frac{x}{2q}$ for $q\geq 1$ and $x\in[0,q]$ we get

$$I_t^{\text{EB-UCB}}\left(a_t^{\text{EB-UCB}}\right)\geq\frac{1}{4}\min\{U^{-2}\gamma,\rho(a_t^{\text{EB-UCB}})^{-2}\}\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\boldsymbol{W}_t^{-1}}.$$

So

$$\begin{aligned}\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{\text{EB-UCB}}\right)}{I_t^{\text{EB-UCB}}\left(a_t^{\text{EB-UCB}}\right)}\leq&\frac{4\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\boldsymbol{W}_t^{-1}}}{\frac{1}{4}\min\{U^{-2}\gamma,\rho(a_t^{\text{EB-UCB}})^{-2}\}\left\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\right\|^2_{\boldsymbol{W}_t^{-1}}}\\ =&16\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\max\{U^2\gamma^{-1},\rho(a_t^{\text{EB-UCB}})^2\}\\ \leq&16\beta_{t,\zeta_t(\delta)}(B)\max\{U^2\gamma^{-1},\rho(a_t^{\text{EB-UCB}})^2\}\\ \leq&16\beta_{T,\zeta_T(\delta)}(B)\max\{U^2\gamma^{-1},\rho^2_{\max}\}.\end{aligned} \tag{25}$$

So

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{EB-UCB}}\right)}{I_t^{\text{EB-UCB}}\left(A_t^{\text{EB-UCB}}\right)}\right]\leq 16T\beta_{T,\zeta_T(\delta)}(B)\max\{U^2\gamma^{-1},\rho^2_{\max}\}. \tag{26}$$

Since $1/\zeta_T(\delta)=\max\{1/\delta,T^2\}$, from (19) we have

$$\begin{aligned}\beta_{T,\zeta_T(\delta)}(B)=&\left(\sqrt{2\log\left(1/\zeta_t(\delta)\right)+\log\left(\frac{\det\left(\boldsymbol{W}_t\right)}{\det\left(\boldsymbol{W}_1\right)}\right)}+\sqrt{\gamma}B\right)^2\\ \leq&2\max\{2\log T,\log(1/\delta)\}+2\log\left(\frac{\det\boldsymbol{W}_T}{\det\boldsymbol{W}_1}\right)+2\gamma B^2\\ \leq&2\max\{2\log T,\log(1/\delta)\}+2d\log(T-1)+2d\log\left(1+\frac{\rho^{-2}_{\min}U^2}{\gamma}\right)+2\gamma B^2\\ <&(2d+4)\log T+2\log(1/\delta)+2d\log\left(1+\frac{\rho^{-2}_{\min}U^2}{\gamma}\right)+2\gamma B^2.\end{aligned} \tag{27}$$

428 So from (24), (26), and (27) we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\mathrm{BAM}}\right)}{\alpha I_t^B\left(A_t^{\mathrm{BAM}}\right)+(1-\alpha)I_t^{\text{EB-UCB}}\left(A_t^{\mathrm{BAM}}\right)}\right]\leq\frac{16}{1-\alpha}\max\{U^2\gamma^{-1},\rho_{\max}^2\}T$$
$$\times\left[(2d+4)\log T+2\log(1/\delta)\ +\ 2d\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)+2\gamma B^2\right]. \tag{28}$$

429 Also, for any sequence $\{a_t\}_{t=1}^T\subset\mathcal{A}$ we have

$$\begin{aligned}I_t^B(a_t)=&\frac{1}{2}\log\left(\|\boldsymbol{v}_t^{\min}\|^2_{(\boldsymbol{W}_t+\rho(a)^{-2}\boldsymbol{\phi}(a)\boldsymbol{\phi}(a)^\top)}\right)-\frac{1}{2}\log\left(\lambda_{\min}(\boldsymbol{W}_t)\right)\\
=&\frac{1}{2}\log\left(\frac{(\boldsymbol{v}_t^{\min})^\top\left(\boldsymbol{W}_t+\rho(a_t)^{-2}\boldsymbol{\phi}(a_t)\boldsymbol{\phi}(a_t)^\top\right)\boldsymbol{v}_t^{\min}}{\lambda_{\min}(\boldsymbol{W}_t)}\right)\\
=&\frac{1}{2}\log\left(\frac{\lambda_{\min}(\boldsymbol{W}_t)+\rho(a_t)^{-2}\boldsymbol{v}_t^{\min}\boldsymbol{\phi}(a_t)\boldsymbol{\phi}(a_t)^\top\boldsymbol{v}_t^{\min}}{\lambda_{\min}(\boldsymbol{W}_t)}\right)\\
=&\frac{1}{2}\log\left(1+\frac{\rho(a_t)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\boldsymbol{\phi}(a_t))^2}{\lambda_{\min}(\boldsymbol{W}_t)}\right)=\frac{1}{2}\log\left(1+\frac{\omega_t(a_t)}{\lambda_{\min}(\boldsymbol{W}_t)}\right).\end{aligned} \tag{29}$$

Let

$$T_0=\max\left\{t:\sum_{\tau=1}^{t}\omega_\tau(a_\tau)\leq d(\rho_{\min}^{-2}U^2-\gamma)\right\}.$$

430 Without loss of generality, assume that $T_0\leq T$. Then using Lemma 2 we get

$$\begin{aligned}\sum_{t=1}^{T}I_t^B(a_t)=&\sum_{t=1}^{T}\log\left(1+\frac{\omega_t(a_t)}{\lambda_{\min}(\boldsymbol{W}_t)}\right)\\
=&\sum_{t=1}^{T_0}\log\left(1+\frac{\omega_t(a_t)}{\lambda_{\min}(\boldsymbol{W}_t)}\right)+\sum_{t=T_0+1}^{T}\log\left(1+\frac{\omega_t(a_t)}{\lambda_{\min}(\boldsymbol{W}_t)}\right)\\
\leq&\sum_{t=1}^{T_0}\frac{\omega_t(a_t)}{\lambda_{\min}(\boldsymbol{W}_t)}+\sum_{t=T_0+1}^{T}\log\left(1+\frac{\omega_t(a_t)}{\gamma-\rho_{\min}^{-2}U^2+\frac{1}{d}\sum_{\tau=1}^{t-1}\omega_\tau(a_\tau)}\right)\\
\leq&\frac{1}{\gamma}\sum_{t=1}^{T_0}\omega_t(a_t)+\sum_{t=T_0+1}^{T}\log\left(1+\frac{d\omega_t(a_t)}{d(\gamma-\rho_{\min}^{-2}U^2)+\sum_{\tau=1}^{t-1}\omega_\tau(a_\tau)}\right)\\
\leq&\frac{d(\rho_{\min}^{-2}U^2-\gamma)}{\gamma}+\sum_{t=T_0+1}^{T}\frac{d\omega_t(a_t)}{d(\gamma-\rho_{\min}^{-2}U^2)+\sum_{\tau=1}^{T_0}\omega_\tau(a_\tau)+\sum_{\tau=T_0+1}^{t-1}\omega_\tau(a_\tau)}.\end{aligned}$$

Let

$$c=d(\gamma-\rho_{\min}^{-2}U^2)+\sum_{\tau=1}^{T_0}\omega_\tau(a_\tau)$$

and

$$x_t=\omega_{T_0+t}(a_{T_0+t}).$$

431 Then from Lemma 3, since $c>0$ and $x_t\in[0,\rho_{\min}^{-2}U^2]$ for all t we have

$$\begin{aligned}\sum_{t=1}^{T}I_t^B(a_t)\leq&\frac{d(\rho_{\min}^{-2}U^2-\gamma)}{\gamma}+\sum_{t=T_0+1}^{T}\frac{d\omega_t(a_t)}{c+\sum_{\tau=T_0+1}^{t-1}\omega_\tau(a_\tau)}\\
=&\frac{d(\rho_{\min}^{-2}U^2-\gamma)}{\gamma}+d\sum_{t=1}^{T-T_0}\frac{x_t}{c+\sum_{\tau=1}^{t-1}x_\tau}\\
\leq&O(d\log(T-T_0))\leq O(d\log T).\end{aligned} \tag{30}$$

Thus, from (21) we have

$$\alpha\mathbb{E}\left[\sum_{t=1}^{T} I_t^B\left(A_t^{\mathrm{BAM}}\right)\right]+(1-\alpha)\mathbb{E}\left[\sum_{t=1}^{T} I_t^{\text{EB-UCB}}\left(A_t^{\mathrm{BAM}}\right)\right]\leq O(d\log T).$$

So from Theorem 1 and (28) we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\Delta}_{t,\zeta_t(\delta)}(A_t^{\mathrm{BAM}})\right]\leq & O\Bigg(\sqrt{d\frac{16}{1-\alpha}\max\{U^2\gamma^{-1},\rho_{\max}^2\}T\log T}\\ &\times\sqrt{(4+2d)\log T+2\log(1/\delta)+2d\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)+2\gamma B^2}\Bigg)\\ \leq & O\Bigg(\frac{d\max\{U/\sqrt{\gamma},\rho_{\max}\}}{\sqrt{1-\alpha}}\sqrt{T}\log T\\ &\times\sqrt{\log(1/\delta)+\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)+\gamma B^2}\Bigg). \end{aligned} \tag{31}$$

Take any $t\geq 1$ and suppose that the event $E_{t,\zeta_t(\delta)}$, as defined in (12), holds. Note that the set

$$\left\{\boldsymbol{\theta}\in\mathbb{R}:\left\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}\right\|_{\boldsymbol{W}_t}^2\leq\beta_{t,\zeta_t(\delta)}(B^*)\right\}$$

is an ellipsoid in $\mathbb{R}^d$ centered at $\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}$ with the longest semi-axis of length $\beta_{t,\zeta_t(\delta)}^{1/2}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}$, so

$$\left\|\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}-\boldsymbol{\theta}^*\right\|_2\leq\beta_{t,\zeta_t(\delta)}^{1/2}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}. \tag{32}$$

Since $B\geq B^*$ we have $\beta_{t,\zeta_t(\delta)}(B)\geq\beta_{t,\zeta_t(\delta)}(B^*)$, so by the triangle inequality we get

$$B^*=\|\boldsymbol{\theta}^*\|_2\leq\left\|\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}\right\|_2+\beta_{t,\zeta_t(\delta)}^{1/2}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}\leq\left\|\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}\right\|_2+\beta_{t,\zeta_t(\delta)}^{1/2}(B)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}=\widehat{B}_t. \tag{33}$$

So $B^*\leq\widehat{B}_t$ for all $t\geq 1$ and thus $\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\geq\beta_{t,\zeta_t(\delta)}(B^*)$, so

$$\boldsymbol{\theta}^*\in\left\{\boldsymbol{\theta}\in\mathbb{R}:\left\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}\right\|_{\boldsymbol{W}_t}^2\leq\beta_{t,\zeta_t(\delta)}(B^*)\right\}\subseteq\left\{\boldsymbol{\theta}\in\mathbb{R}:\left\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_t^{\mathrm{wls}}\right\|_{\boldsymbol{W}_t}^2\leq\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\right\}$$

Hence $\Delta(a)\leq\widehat{\Delta}_{t,\zeta_t(\delta)}(a)$ for all $a\in\mathcal{A}$. So for any $a\in\mathcal{A}$ we have

$$\mathbb{P}\left(\Delta(a)>\widehat{\Delta}_{t,\zeta_t(\delta)}(a)\right)\leq 1-\mathbb{P}(E_{t,\zeta_t(\delta)})\leq\zeta_t(\delta)\leq 1/t^2.$$

Thus, letting $\Delta_{\max}=\max_{a\in\mathcal{A}}\Delta(a)$, for any sequence $\{a_t\}_{t=1}^T\subset\mathcal{A}$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\Delta(a_t)-\widehat{\Delta}_{t,\zeta_t(\delta)}(a_t)\right]\leq\Delta_{\max}\sum_{t=1}^{T}\mathbb{P}\left(\Delta(a_t)>\widehat{\Delta}_{t,1/t^2}(a_t)\right)\leq\Delta_{\max}\sum_{t=1}^{T}\frac{1}{t^2}\leq O(\Delta_{\max}). \tag{34}$$

So from (31), for $T\leq T_B$ the regret of EBIDS is bounded above by

$$\mathcal{R}_T\leq O\left(\frac{d\max\{U/\sqrt{\gamma},\rho_{\max}\}}{\sqrt{1-\alpha}}\sqrt{T}\log T\sqrt{\log(1/\delta)+\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)+\gamma B^2}\right).$$

From (21) and (30) with probability 1 we have

$$\sum_{t=1}^{T} I_t^{\text{BAM}}(A_t^{\text{BAM}}) \leq O(d \log T). \tag{35}$$

Following the same steps as in (24), using (25) and (28) we have

$$\begin{aligned}
\sum_{t=1}^{T} \Psi_t^{\text{BAM}}\left(A_t^{\text{BAM}}\right) &= \sum_{t=1}^{T} \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{BAM}}\right)}{I_t^{\text{BAM}}\left(A_t^{\text{BAM}}\right)} \leq \sum_{t=1}^{T} \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{BAM}}\right)}{\alpha I_t^{B}\left(A_t^{\text{BAM}}\right) + (1-\alpha) I_t^{\text{EB-UCB}}\left(A_t^{\text{BAM}}\right)} \\
&\leq \sum_{t=1}^{T} \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{EB-UCB}}\right)}{\alpha I_t^{\text{EB-UCB}}\left(A_t^{\text{EB-UCB}}\right) + (1-\alpha) I_t^{B}\left(A_t^{\text{EB-UCB}}\right)} \\
&\leq \frac{1}{1-\alpha} \sum_{t=1}^{T} \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(A_t^{\text{EB-UCB}}\right)}{I_t^{\text{EB-UCB}}\left(A_t^{\text{EB-UCB}}\right)} \\
&\leq \frac{16}{1-\alpha} \sum_{t=1}^{T} \beta_{T,\zeta_T(\delta)}(B) \max\{U^2\gamma^{-1}, \rho_{\max}^2\} \\
&\leq \frac{16}{1-\alpha} \max\{U^2\gamma^{-1}, \rho_{\max}^2\} T \\
&\quad \times \left[(4+2d)\log T + 2\log(1/\delta) + 2d\log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) + 2\gamma B^2\right].
\end{aligned} \tag{36}$$

Following analogous steps as above, since $1/\zeta_t(\delta) \geq 1/\delta$ we have $\beta_{t,\zeta_t(\delta)}(B) \geq \beta_{t,\delta}(B) \geq \beta_{t,\delta}(B^*)$. So for any $t \geq 1$, whenever event $E_{t,\delta}$ holds, the inequality $B^* \leq \widehat{B}_t$ holds as well and thus $\Delta(a) \leq \widehat{\Delta}_{t,\zeta_t(\delta)}(a)$, for all $a \in \mathcal{A}$. So if $E_\delta = \bigcap_{t=1}^{\infty} E_{t,\delta}$ holds, then $\Delta(a) \leq \widehat{\Delta}_{t,\zeta_t(\delta)}(a)$, for all $a \in \mathcal{A}$ and for all $t \geq 1$. So from (36) and (35), by Theorem 1 we have

$$\mathcal{PR}_T \leq O\left(\frac{d \max\{U/\sqrt{\gamma}, \rho_{\max}\}}{\sqrt{1-\alpha}} \sqrt{T} \log T \sqrt{\log(1/\delta) + \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) + \gamma B^2}\right). \quad \square$$

10.6 Proof of Proposition 2

In order to precisely state the conditions on T_B and α, i.e., how large each of them needs to be for $B^* \leq \tilde{B}_t \leq (1+g)B^*$ to hold for all $t \geq T_B + 1$, we will first define several constants for notational convenience.

Let

$$c_0 = L^2 \left[U^2(\gamma + \rho_{\min}^{-2} U^2)\left(\frac{1}{\kappa} + \frac{1}{\gamma}\right)\right]^{-1} \tag{37}$$

and

$$h_0 = 8\log(5/4) + 4\log(1/\delta) + 2d\log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) + 2\gamma B^2. \tag{38}$$

Then let

$$u_0 = \frac{c_0}{6 + 16g^{-2}} \log 2 + \frac{1-\alpha}{\alpha} d \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) \tag{39}$$

$$u_1 = \frac{c_0}{12 + 32g^{-2}} - \frac{1-\alpha}{2\alpha} d \tag{40}$$

$$w_0 = \frac{c_0}{6 + 16g^{-2}} + \frac{1-\alpha}{\alpha} d \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) \tag{41}$$

$$w_1 = \frac{c_0}{12 + 32g^{-2}} - \frac{1-\alpha}{\alpha} d. \tag{42}$$

and finally let

$$b_0 = \frac{1}{d}\left[w_0\left(\frac{\gamma}{d}u_0 - \gamma + \rho_{\min}^{-2}U^2\right) - \gamma u_0\right] + \gamma - \rho_{\min}^{-2}U^2 \tag{43}$$

$$b_1 = \frac{1}{d}\left(\gamma u_1 - \frac{\gamma}{d}u_1 w_0 - \frac{\gamma}{d}u_0 w_1 + \gamma w_1 - \rho_{\min}^{-2}U^2 w_1\right) \tag{44}$$

$$b_2 = \frac{\gamma}{d^2}u_1 w_1 \tag{45}$$

We make the following assumptions.

Assumption 1. $B \geq B^*$.

Assumption 2.

$$T_B \geq \max\left\{4, \exp\left[\frac{h_0 + 2d + 8}{b_2}\left(4g^{-2}B^{*-2} + \frac{|b_1|}{2d+8} + \frac{|b_0|}{h_0 + 2d + 8}\right)\right]\right\}.$$

Assumption 3.

$$\alpha \geq \frac{d}{d + \frac{c_0}{12+32g^{-2}}}.$$

We will now show that if Assumptions 1-3 are satisfied and event E_δ holds then $B^* \leq \tilde{B}_t \leq (1+g)B^*$ for all $t \geq T_B + 1$.

Proof. Suppose that event E_δ holds. For any t let

$$s(t) = \arg\min_{\tau \leq t} \beta^{1/2}_{\tau,\zeta_\tau(\delta)}(\widehat{B}_\tau)\lambda_{\min}(\boldsymbol{W}_\tau)^{-1/2} \tag{46}$$

From (32) in the proof of Proposition 1, using the triangle inequality we get

$$\left\|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_2 \leq \|\boldsymbol{\theta}^*\|_2 + \beta^{1/2}_{t,\zeta_t(\delta)}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2} = B^* + \beta^{1/2}_{t,\zeta_t(\delta)}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}. \tag{47}$$

From (33) in the proof of Proposition 1, for any t we have $\widehat{B}_t \geq B^*$, so

$$\left\|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_2 \leq B^* + \beta^{1/2}_{t,\zeta_t(\delta)}(\widehat{B}_t)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}.$$

Hence

$$\begin{aligned}\tilde{B}_t &= \min_{\tau \leq t}\left\{\left\|\widehat{\boldsymbol{\theta}}_\tau^{\text{wls}}\right\|_2 + \beta^{1/2}_{\tau,\zeta_\tau(\delta)}(\widehat{B}_\tau)\lambda_{\min}(\boldsymbol{W}_\tau)^{-1/2}\right\} \\ &\leq \left\|\widehat{\boldsymbol{\theta}}_{s(t)}^{\text{wls}}\right\|_2 + \beta^{1/2}_{s(t),\zeta_{s(t)}(\delta)}(\widehat{B}_{s(t)})\lambda_{\min}(\boldsymbol{W}_{s(t)})^{-1/2} \\ &\leq B^* + 2\beta^{1/2}_{s(t),\zeta_{s(t)}(\delta)}(\widehat{B}_{s(t)})\lambda_{\min}(\boldsymbol{W}_{s(t)})^{-1/2}.\end{aligned} \tag{48}$$

Also, analogously as in (33), using (32) and the triangle inequality, for any $t \geq 1$ we have

$$B^* = \|\boldsymbol{\theta}^*\|_2 \leq \left\|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_2 + \beta^{1/2}_{t,\zeta_t(\delta)}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2} \leq \left\|\widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\|_2 + \beta^{1/2}_{t,\zeta_t(\delta)}(\widehat{B}_t)\lambda_{\min}(\boldsymbol{W}_t)^{-1/2}.$$

So

$$B^* \leq \tilde{B}_t \tag{49}$$

for any $t \geq 1$.

From Lemma 1, for any $t \leq T_B$ we have

$$I_t^B(a_t^{\text{BAM}}) \geq \frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}(a_t^{\text{BAM}})}{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{I,B}\right)} I_t^B\left(a_t^{I,B}\right) - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}\left(a_t^{\text{BAM}}\right) \tag{50}$$

where $a_t^{I,B} = \arg\max_{a\in\mathcal{A}} I_t^B(a)$. For any $t \le T_B$ we have

$$\begin{aligned}
\widehat{\Delta}_{t,\zeta_t(\delta)}\left(a_t^{\text{BAM}}\right) =& \max_{b\in\mathcal{A}}\left\{\left\langle \boldsymbol{\phi}(b), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\|\boldsymbol{\phi}(b)\|_{\boldsymbol{W}_t^{-1}}\right\} \\
&- \left(\left\langle \boldsymbol{\phi}\left(a_t^{\text{BAM}}\right), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle - \beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\left\|\boldsymbol{\phi}\left(a_t^{\text{BAM}}\right)\right\|_{\boldsymbol{W}_t^{-1}}\right) \\
=& \max_{b\in\mathcal{A}}\left\{\left\langle \boldsymbol{\phi}(b), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\|\boldsymbol{\phi}(b)\|_{\boldsymbol{W}_t^{-1}}\right\} \\
&- \left(\left\langle \boldsymbol{\phi}\left(a_t^{\text{BAM}}\right), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle + \beta_{t,\zeta_t(\delta)}(\widehat{B}_t)^{1/2}\left\|\boldsymbol{\phi}\left(a_t^{\text{BAM}}\right)\right\|_{\boldsymbol{W}_t^{-1}}\right) \\
&+ 2\beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\left\|\boldsymbol{\phi}\left(a_t^{\text{BAM}}\right)\right\|_{\boldsymbol{W}_t^{-1}} \\
\ge& 2\beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\left\|\boldsymbol{\phi}\left(a_t^{\text{BAM}}\right)\right\|_{\boldsymbol{W}_t^{-1}}.
\end{aligned}$$

So from (18)

$$\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{\text{BAM}}\right) \ge 4\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\left\|\boldsymbol{\phi}\left(a_t^{\text{BAM}}\right)\right\|^2_{\boldsymbol{W}_t^{-1}} \ge 4\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\frac{L^2}{t(\gamma+\rho_{\min}^{-2}U^2)} \tag{51}$$

Also

$$\begin{aligned}
\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{I,B}\right) =& \beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t)\left(\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\|_{\boldsymbol{W}_t^{-1}} + \left\|\boldsymbol{\phi}(a_t^{I,B})\right\|_{\boldsymbol{W}_t^{-1}}\right) \\
&+ \left\langle \boldsymbol{\phi}(a_t^{\text{EB-UCB}}), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle - \left\langle \boldsymbol{\phi}(a_t^{I,B}), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle,
\end{aligned}$$

so

$$\begin{aligned}
\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{I,B}\right) \le& 4\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)\left(\|\boldsymbol{\phi}(a_t^{\text{EB-UCB}})\|^2_{\boldsymbol{W}_t^{-1}} + \left\|\boldsymbol{\phi}(a_t^{I,B})\right\|^2_{\boldsymbol{W}_t^{-1}}\right) \\
&+ 4\left\langle \boldsymbol{\phi}(a_t^{\text{EB-UCB}}), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle^2 + 4\left\langle \boldsymbol{\phi}(a_t^{I,B}), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle^2.
\end{aligned}$$

Since E_δ holds, from (16) and (47) for any t and any $a \in \mathcal{A}$ we have

$$\left\langle \boldsymbol{\phi}(a), \widehat{\boldsymbol{\theta}}_t^{\text{wls}}\right\rangle^2 \le 2U^2(B^{*2} + \beta_{t,\zeta_t(\delta)}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1})$$

so from (15) we have

$$\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{I,B}\right) \le 8\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)U^2\lambda_{\min}(\boldsymbol{W}_t)^{-1} + 16U^2(B^{*2} + \beta_{t,\zeta_t(\delta)}(B^*)\lambda_{\min}(\boldsymbol{W}_t)^{-1}). \tag{52}$$

From (29) from the proof of Proposition 1, for any $a \in \mathcal{A}$ we have

$$I_t^B(a) = \frac{1}{2}\log\left(1 + \frac{\rho(a)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\boldsymbol{\phi}(a))^2}{\lambda_{\min}(\boldsymbol{W}_t)}\right), \tag{53}$$

so

$$\begin{aligned}
I_t^B\left(a_t^{I,B}\right) =& \max_{a\in\mathcal{A}} I_t^B(a) = \max_{a\in\mathcal{A}}\left\{\frac{1}{2}\log\left(1 + \frac{\rho(a)^{-2}\psi_{\boldsymbol{v}_t^{\min}}(\boldsymbol{\phi}(a))^2}{\lambda_{\min}(\boldsymbol{W}_t)}\right)\right\} \\
\ge& \frac{1}{2}\log\left(1 + \frac{\kappa}{\lambda_{\min}(\boldsymbol{W}_t)}\right).
\end{aligned}$$

Thus, since $\log x \ge 1 - \frac{1}{x}$ for all $x > 0$, we have

$$\begin{aligned}
I_t^B\left(a_t^{I,B}\right) \ge& \frac{\kappa}{2(\lambda_{\min}(\boldsymbol{W}_t) + \kappa)} = \left[2\lambda_{\min}(\boldsymbol{W}_t)\left(\frac{1}{\kappa} + \frac{1}{\lambda_{\min}(\boldsymbol{W}_t)}\right)\right]^{-1} \\
\ge& \left[2\lambda_{\min}(\boldsymbol{W}_t)\left(\frac{1}{\kappa} + \frac{1}{\gamma}\right)\right]^{-1}.
\end{aligned} \tag{54}$$

So combining (50), (51), (52), and (54), for any $t \leq T_B$ we have

$$
\begin{aligned}
I_t^B(a_t^{\text{BAM}}) \geq & \frac{L^2 \lambda_{\min}(\boldsymbol{W}_t)^{-1} \left[2t(\gamma + \rho_{\min}^{-2} U^2) \left(\frac{1}{\kappa} + \frac{1}{\gamma}\right)\right]^{-1}}{2U^2 \lambda_{\min}(\boldsymbol{W}_t)^{-1} + 4U^2 \beta_{t,\zeta_t(\delta)}(\widehat{B}_t)^{-1} \left(B^{*2} + \beta_{t,\delta}(B^*) \lambda_{\min}(\boldsymbol{W}_t)^{-1}\right)} \\
& - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}\left(a_t^{\text{BAM}}\right) = \\
= & \frac{1}{t} L^2 \left[U^2(\gamma + \rho_{\min}^{-2} U^2) \left(\frac{1}{\kappa} + \frac{1}{\gamma}\right) \left(4 + 8B^{*2} \frac{\lambda_{\min}(\boldsymbol{W}_t)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)} + 8 \frac{\beta_{t,\delta}(B^*)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)}\right)\right]^{-1} \\
& - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}) \geq \\
\geq & \frac{1}{t} L^2 \left[U^2(\gamma + \rho_{\min}^{-2} U^2) \left(\frac{1}{\kappa} + \frac{1}{\gamma}\right) \left(12 + 8B^{*2} \frac{\lambda_{\min}(\boldsymbol{W}_t)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)}\right)\right]^{-1} \\
& - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}),
\end{aligned}
$$

where the last inequality follows from the fact that $\widehat{B}_t \geq B^*$ and $1/\zeta_t(\delta) \geq 1/\delta$ which gives us

$$\frac{\beta_{t,\delta}(B^*)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)} \leq 1.$$

So from (37) we have

$$I_t^B(a_t^{\text{BAM}}) \geq \frac{1}{t} c_0 \left(12 + 8B^{*2} \frac{\lambda_{\min}(\boldsymbol{W}_t)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)}\right)^{-1} - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}). \tag{55}$$

From (53) we have

$$I_t^B(a_t^{\text{BAM}}) = \frac{1}{2} \log\left(1 + \frac{\omega_t(a_t^{\text{BAM}})}{\lambda_{\min}(\boldsymbol{W}_t)}\right) \leq \frac{\omega_t(a_t^{\text{BAM}})}{2\lambda_{\min}(\boldsymbol{W}_t)}.$$

So

$$\omega_t(a_t^{\text{BAM}}) \geq 2\lambda_{\min}(\boldsymbol{W}_t) I_t^B(a_t^{\text{BAM}}). \tag{56}$$

If

$$\beta_{t,\zeta_t(\delta)}^{1/2}(\widehat{B}_t) \lambda_{\min}(\boldsymbol{W}_t)^{-1/2} \leq \frac{1}{2} g B^* \tag{57}$$

for some $t \leq T_B + 1$ then

$$\beta_{s(t),\zeta_{s(t)}(\delta)}^{1/2}(\widehat{B}_{s(t)}) \lambda_{\min}(\boldsymbol{W}_{s(t)})^{-1/2} \leq \frac{1}{2} g B^*,$$

so from (48) and (49), since event E_δ holds, for any $t \geq T_B + 1$ we have

$$B^* \leq \tilde{B}_t \leq B^* + 2\beta_{s(t),\zeta_{s(t)}(\delta)}^{1/2}(\widehat{B}_t) \lambda_{\min}(\boldsymbol{W}_{s(t)})^{-1/2} = (1+g)B^* \tag{58}$$

which is what we want to show. We will prove by contradiction that since E_δ holds, (57) holds as well for some $t \leq T_B + 1$. Suppose that (57) does not hold. Then for all $t \leq T_B + 1$ we have

$$\frac{\lambda_{\min}(\boldsymbol{W}_t)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)} < 4g^{-2} B^{*-2}, \tag{59}$$

so from (55) we have

$$
\begin{aligned}
I_t^B(a_t^{\text{BAM}}) \geq & \frac{1}{t} c_0 \left(12 + 8B^{*2} \frac{\lambda_{\min}(\boldsymbol{W}_t)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)}\right)^{-1} - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}) \\
> & \frac{1}{t} \cdot \frac{c_0}{12 + 32g^{-2}} - \frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}).
\end{aligned}
$$

Hence, from (56) for any $t \leq T_B$ we have

$$
\omega_t(a_t^{\text{BAM}}) \geq \lambda_{\min}(\boldsymbol{W}_t) \left(\frac{1}{t} \cdot \frac{c_0}{6 + 16g^{-2}} - 2\frac{1-\alpha}{\alpha} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}})\right).
$$

Let $\lfloor x \rfloor$ denote the largest integer smaller than or equal to x for any $x \in \mathbb{R}$. From Weyl's inequality (Franklin, 1968)

$$
\lambda_{\min}(\boldsymbol{W}_{t+1}) \geq \lambda_{\min}(\boldsymbol{W}_t) \geq \gamma \tag{60}
$$

for any t. Also note that $\omega_t(a) \geq 0$ for any t and any $a \in \mathcal{A}$. So

$$
\sum_{t=1}^{\lfloor \sqrt{T_B} \rfloor} \omega_t(a_t^{\text{BAM}}) \geq \gamma \left(\frac{c_0}{6 + 16g^{-2}} \sum_{t=1}^{\lfloor \sqrt{T_B} \rfloor} \frac{1}{t} - 2\frac{1-\alpha}{\alpha} \sum_{t=1}^{\lfloor \sqrt{T_B} \rfloor} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}})\right).
$$

From (21) we have

$$
\begin{aligned}
\sum_{t=1}^{\lfloor \sqrt{T_B} \rfloor} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}) \leq & \frac{1}{2} d \log \lfloor \sqrt{T_B} \rfloor + \frac{1}{2} d \log \left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right) \\
\leq & \frac{1}{4} d \log T_B + \frac{1}{2} d \log \left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right).
\end{aligned}
$$

Also since $T_B \geq 4$ we have $\lfloor \sqrt{T_B} \rfloor \geq \sqrt{T_B} - 1 \geq \sqrt{T_B}/2$, so

$$
\sum_{t=1}^{\lfloor \sqrt{T_B} \rfloor} \frac{1}{t} > \log \lfloor \sqrt{T_B} \rfloor \geq \log\left(\frac{1}{2}\sqrt{T_B}\right) = \frac{1}{2} \log T_B - \log 2.
$$

So

$$
\begin{aligned}
\sum_{t=1}^{\lfloor \sqrt{T_B} \rfloor} \omega_t(a_t^{\text{BAM}}) \geq & \gamma \left(\frac{c_0}{6 + 16g^{-2}} \left[\frac{1}{2} \log T_B - \log 2\right] - \frac{1-\alpha}{\alpha} d \left[\frac{1}{2} \log T_B + \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right)\right]\right) \\
\geq & \gamma \left(\left[\frac{c_0}{12 + 32g^{-2}} - \frac{1-\alpha}{2\alpha} d\right] \log T_B - \left[\frac{c_0}{6 + 16g^{-2}} \log 2 + \frac{1-\alpha}{\alpha} d \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right)\right]\right) \\
= & \gamma (u_1 \log T_B - u_0),
\end{aligned} \tag{61}
$$

where the constants u_0 and u_1 were defined in (39) and (40), respectively. Similarly from (60) we have

$$
\sum_{t=\lfloor \sqrt{T_B} \rfloor + 1}^{T_B} \omega_t(a_t^{\text{BAM}}) \geq \lambda_{\min}(\boldsymbol{W}_{\lfloor \sqrt{T_B} \rfloor + 1}) \left(\frac{c_0}{6 + 16g^{-2}} \sum_{t=\lfloor \sqrt{T_B} \rfloor + 1}^{T_B} \frac{1}{t} - 2\frac{1-\alpha}{\alpha} \sum_{t=\lfloor \sqrt{T_B} \rfloor + 1}^{T_B} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}})\right)
$$

Note hat

$$
\sum_{t=\lfloor \sqrt{T_B} \rfloor + 1}^{T_B} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}) \leq \sum_{t=1}^{T_B} I_t^{\text{EB-UCB}}(a_t^{\text{BAM}}) \leq \frac{1}{2} d \log T_B + \frac{1}{2} d \log\left(1 + \frac{\rho_{\min}^{-2} U^2}{\gamma}\right)
$$

and

$$\sum_{t=\lfloor\sqrt{T_B}\rfloor+1}^{T_B}\frac{1}{t}=\sum_{t=1}^{T_B}\frac{1}{t}-\sum_{t=1}^{\lfloor\sqrt{T_B}\rfloor}\frac{1}{t}>\log T_B-(\log\sqrt{T_B}+1)=\frac{1}{2}\log T_B-1.$$

So

$$\begin{aligned}\sum_{t=\lfloor\sqrt{T_B}\rfloor+1}^{T_B}\omega_t(a_t^{\mathrm{BAM}})\geq&\lambda_{\min}(\boldsymbol{W}_{\lfloor\sqrt{T_B}\rfloor+1})\\&\times\left(\frac{c_0}{6+16g^{-2}}\left[\frac{1}{2}\log T_B-1\right]-\frac{1-\alpha}{\alpha}d\left[\log T_B+\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)\right]\right)\\=&\lambda_{\min}(\boldsymbol{W}_{\lfloor\sqrt{T_B}\rfloor+1})\\&\times\left(\left[\frac{c_0}{12+32g^{-2}}-\frac{1-\alpha}{\alpha}d\right]\log T_B-\left[\frac{c_0}{6+16g^{-2}}+\frac{1-\alpha}{\alpha}d\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)\right]\right)\\=&\lambda_{\min}(\boldsymbol{W}_{\lfloor\sqrt{T_B}\rfloor+1})(w_1\log T_B+w_0),\end{aligned}$$

where the constants w_0 and w_1 were defined in (41) and (42), respectively.

From Lemma 2 and (61) we have

$$\lambda_{\min}(\boldsymbol{W}_{\lfloor\sqrt{T_B}\rfloor+1})\geq\gamma-\rho_{\min}^{-2}U^2+\frac{1}{d}\sum_{t=1}^{\lfloor\sqrt{T_B}\rfloor}\omega_t(a_t^{\mathrm{BAM}})\geq\frac{\gamma}{d}(u_1\log T_B-u_0)+\gamma-\rho_{\min}^{-2}U^2.$$

So

$$\begin{aligned}\sum_{t=1}^{T_B}\omega_t(a_t^{\mathrm{BAM}})=&\sum_{t=1}^{\lfloor\sqrt{T_B}\rfloor}\omega_t(a_t^{\mathrm{BAM}})+\sum_{t=\lfloor\sqrt{T_B}\rfloor+1}^{T_B}\omega_t(a_t^{\mathrm{BAM}})\\\geq&\gamma(u_1\log T_B-u_0)+\left(\frac{\gamma}{d}(u_1\log T_B-u_0)+\gamma-\rho_{\min}^{-2}U^2\right)(w_1\log T_B-w_0)\\=&\frac{\gamma}{d}u_1w_1(\log T_B)^2+\left(\gamma u_1-\frac{\gamma}{d}u_1w_0-\frac{\gamma}{d}u_0w_1+\gamma w_1-\rho_{\min}^{-2}U^2w_1\right)\log T_B\\&+w_0\left(\frac{\gamma}{d}u_0-\gamma+\rho_{\min}^{-2}U^2\right)-\gamma u_0\\=&db_2(\log T_B)^2+db_1\log T_B+d(b_0-\gamma+\rho_{\min}^{-2}U^2),\end{aligned}$$

where the constants b_0, b_1 and b_2 were defined in (43), (44), and (45), respectively.

Then, applying Lemma 2 again we get

$$\lambda_{\min}(\boldsymbol{W}_{T_B+1})\geq\gamma-\rho_{\min}^{-2}U^2+\frac{1}{d}\sum_{t=1}^{T_B}\omega_t(a_t^{\mathrm{BAM}})\geq b_2(\log T_B)^2+b_1\log T_B+b_0. \tag{62}$$

Note that by Assumption 3, we have $u_1>0$ and $w_1>0$, so $b_2>0$.

From (19) we have

$$\begin{aligned}&\beta_{T_B+1,\zeta_{T_B+1}(\delta)}(\widehat{B}_t)=\left(\sqrt{2\log(1/\zeta_{T_B+1}(\delta))+\log\left(\frac{\det\boldsymbol{W}_{T_B+1}}{\det\boldsymbol{W}_1}\right)}+\sqrt{\gamma}\widehat{B}_{T_B+1}\right)^2\leq\\&\leq 4\log(1/\zeta_{T_B+1}(\delta))+2\log\left(\frac{\det\boldsymbol{W}_{T_B+1}}{\det\boldsymbol{W}_1}\right)+2\gamma\widehat{B}_{T_B+1}^2\leq\\&\leq 4\max\{\log(1/\delta),2\log(T_B+1)\}+2d\log T_B+2d\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right)+2\gamma B^2.\end{aligned}$$

Since $T_B \geq 4$ we have

$$\log(T_B+1) \leq \log\left(\frac{5}{4}T_B\right) = \log T_B + \log(5/4),$$

so

$$\begin{aligned}\beta_{T_B+1,\zeta_{T_B+1}(\delta)}(\widehat{B}_t) \leq &(2d+8)\log T_B + 8\log(5/4) + 4\log(1/\delta) + 2d\log\left(1+\frac{\rho_{\min}^{-2}U^2}{\gamma}\right) + 2\gamma B^2 \\ =&(2d+8)\log T_B + h_0, \end{aligned} \tag{63}$$

with h_0 defined in (38). Note that $h_0 > 0$. Also, since $T_B \geq 4$ we have $\log T_B > 1$ so from (62), (63) and the fact that $b_2 > 0$ we get

$$\begin{aligned}\frac{\lambda_{\min}(\boldsymbol{W}_{T_B+1})}{\beta_{T_B+1,\zeta_{T_B+1}(\delta)}(\widehat{B}_t)} \geq &\frac{b_2(\log T_B)^2 + b_1\log T_B + b_0}{(2d+8)\log T_B + h_0} \\ =&\frac{b_2}{2d+8+\frac{h_0}{\log T_B}}\log T_B + \frac{b_1}{(2d+8)+\frac{h_0}{\log T_B}} + \frac{b_0}{(2d+8)\log T_B + h_0} \\ \geq&\frac{b_2}{h_0+2d+8}\log T_B - \frac{|b_1|}{2d+8} - \frac{|b_0|}{h_0+2d+8}.\end{aligned}$$

Note that by Assumption 2 we have

$$T_B \geq \exp\left[\frac{h_0+2d+8}{b_2}\left(4g^{-2}B^{*-2} + \frac{|b_1|}{2d+8} + \frac{|b_0|}{h_0+2d+8}\right)\right]$$

so

$$\frac{\lambda_{\min}(\boldsymbol{W}_{T_B+1})}{\beta_{T_B+1,\zeta_{T_B+1}(\delta)}(\widehat{B}_t)} \geq 4g^{-2}B^{*-2}$$

which is the required contradiction to (59). So there exists $t \leq T_B + 1$ such that

$$\frac{\lambda_{\min}(\boldsymbol{W}_t)}{\beta_{t,\zeta_t(\delta)}(\widehat{B}_t)} \geq 4g^{-2}B^{*-2}$$

and thus, since E_δ holds, from (58) for any $t \geq T_B + 1$ we have

$$B^* \leq \tilde{B}_t \leq (1+g)B^*. \quad \square$$

10.7 Proof of Proposition 3

The exact assumptions made by Propositions 3 are as follows. We assume that T_B and α are sufficiently large so Assumptions 1 - 3 hold and $(T_B+1)^2 \geq 1/\delta$. We can now proceed to the proof.

Proof. Suppose that event E_δ holds.

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\Delta}_{t,\zeta_t(\delta)}(A_t^{\text{BEIDS}})\right] = \mathbb{E}\left[\sum_{t=1}^{T_B}\widehat{\Delta}_{t,\zeta_t(\delta)}(A_t^{\text{BAM}})\right] + \mathbb{E}\left[\sum_{t=T_B+1}^{T}\widehat{\Delta}_{t,\zeta_t(\delta)}(A_t^{\text{EB-UCB}})\right].$$

From (21) with probability 1 we have

$$\sum_{t=T_B+1}^{T} I_t^{\text{EB-UCB}}(A_t) \leq O(d\log T). \tag{64}$$

Let $a_t^{\text{EB-UCB}}$ be the realization of $A_t^{\text{EB-UCB}}$. Since event E_δ holds and Assumptions 1 - 3 hold, from Proposition 2 we have $B^* \leq \tilde{B}_t \leq (1+g)B^*$ for all $t \geq T_B + 1$. Also from the assumptions of this proposition, $2\log T \geq \log(1/\delta)$, so analogously as in (25) and (27), for any $t \in \{T_B+1, T_B+2, \ldots, T\}$ we have

$$
\begin{aligned}
\frac{\widehat{\Delta}^2_{t,\zeta_t(\delta)}\left(a_t^{\text{EB-UCB}}\right)}{I_t^{\text{EB-UCB}}\left(a_t^{\text{EB-UCB}}\right)} \leq & 16\beta_{T,\zeta_T(\delta)}(\tilde{B}_t)\max\{U^2\gamma^{-1}, \rho^2_{\max}\} \\
\leq & 16\max\{U^2\gamma^{-1}, \rho^2_{\max}\} \\
& \times \left[2\max\{2\log T, \log(1/\delta)\} + 2d\log(T-1) + 2d\log\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right) + 2\gamma\tilde{B}_t^2\right] \\
\leq & 16\max\{U^2\gamma^{-1}, \rho^2_{\max}\} \\
& \times \left[(2d+4)\log T + 2d\log\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right) + 2\gamma\tilde{B}_t^2\right] \\
\leq & 16\max\{U^2\gamma^{-1}, \rho^2_{\max}\} \\
& \times \left[(2d+4)\log T + 2d\log\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right) + 2\gamma\left((1+g)B^*\right)^2\right].
\end{aligned}
$$

Hence from Theorem 1 and (64) we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=T_B+1}^{T}\widehat{\Delta}_{t,\zeta_t(\delta)}(A_t^{\text{EB-UCB}})\right] \leq & O\Bigg(d\max\{U/\sqrt{\gamma}, \rho_{\max}\}\sqrt{T}\log T \\
& \times\sqrt{\log\left(1 + \frac{\rho_{\min}^{-2}U^2}{\gamma}\right) + \gamma\left((1+g)B^*\right)^2}\Bigg) \\
\leq & O\left(dU\rho_{\max}(1+g)B^*\sqrt{T}\log T\right),
\end{aligned}
$$

and thus from (34) we get that

$$
\mathbb{E}\left[\sum_{t=T_B+1}^{T}\Delta(A_t^{\text{EB-UCB}})\right] \leq O\left(dU\rho_{\max}(1+g)B^*\sqrt{T}\log T\right).
$$

and similarly with probability 1 we have

$$
\sum_{t=T_B+1}^{T}\Delta(A_t^{\text{EB-UCB}}) \leq O\left(dU\rho_{\max}(1+g)B^*\sqrt{T}\log T\right).
$$

Thus, since T_B is fixed with respect to T with probability at least $\mathbb{P}(E_\delta) \geq 1-\delta$ we have

$$
\mathcal{R}_T \leq O\left(dU\rho_{\max}(1+g)B^*\sqrt{T}\log T\right)
$$

and

$$
\mathcal{PR}_T \leq O\left(dU\rho_{\max}(1+g)B^*\sqrt{T}\log T\right). \quad \square
$$