Old Dog Learns New Tricks: Randomized UCB for Bandit Problems

Sharan Vaswani  Abbas Mehrabian  Audrey Durand  Branislav Kveton
Mila, Université de Montréal  McGill University  Mila, Université Laval  Google Research

Abstract

We propose RandUCB, a bandit strategy that builds on theoretically derived confidence intervals similar to upper confidence bound (UCB) algorithms, but akin to Thompson sampling (TS), it uses randomization to trade off exploration and exploitation. In the K-armed bandit setting, we show that there are infinitely many variants of RandUCB, all of which achieve the minimax-optimal $\tilde{O}(\sqrt{KT})$ regret after $T$ rounds. Moreover, for a specific multi-armed bandit setting, we show that both UCB and TS can be recovered as special cases of RandUCB. For structured bandits, where each arm is associated with a $d$-dimensional feature vector and rewards are distributed according to a linear or generalized linear model, we prove that RandUCB achieves the minimax-optimal $O(d\sqrt{T})$ regret even in the case of infinitely many arms. Through experiments in both the multi-armed and structured bandit settings, we demonstrate that RandUCB matches or outperforms TS and other randomized exploration strategies. Our theoretical and empirical results together imply that RandUCB achieves the best of both worlds.

1 Introduction

The multi-armed bandit (MAB) [Woodroofe, 1979; Lai and Robbins, 1985; Auer et al., 2002] is a sequential decision-making problem with arms corresponding to actions available to a learning agent to choose from. For example, the arms may correspond to potential treatments in a clinical trial or ads available for display on a website. When an arm is chosen (pulled), the agent receives a reward from the environment. In the stochastic MAB, which is our focus, this reward is sampled from an underlying distribution associated with that particular arm. The agent’s goal is to maximize its expected reward accumulated across interactions with the environment (rounds). As the agent does not know the arms’ reward distributions, she faces an exploration-exploitation dilemma: explore and learn more about the arms, or exploit and choose the arm with the highest estimated mean thus far.

Structured bandits [Li et al., 2010; Filippi et al., 2010; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013; Li et al., 2017] are generalizations of the MAB problem in which each arm is associated with a known feature vector. These features encode properties of the arms; for example, they may represent the properties of a drug being tested in a clinical trial, or the meta-data of an advertisement on a website. In structured bandits, the expected reward of an arm is an unknown function of its feature vector. This function is often assumed to be parametric; an important special case is the linear bandit [Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011], where the function is linear and the expected reward is the dot product of the feature vector and an unknown parameter vector. Similarly, in the generalized linear bandit [Filippi et al., 2010; Li et al., 2017; Kveton et al., 2019d], the expected reward follows a generalized linear model [McCullagh, 1984].

1.1 Classic exploration strategies

In both the multi-armed and structured bandit settings, classic strategies to trade off exploration and exploitation include $\varepsilon$-greedy (EG) [Sutton and Barto, 1998; Auer et al., 2002], optimism in the face of uncertainty (OFU) [Auer et al., 2002; Abbasi-Yadkori et al., 2011], and Thompson sampling (TS) [Thompson, 1933; Agrawal and Goyal, 2017]. The EG policy is simple, can be applied to any MAB or structured bandit setting, and is thus widely used in practice. However, it is statistically sub-optimal, does not explore in a problem dependent manner, and its practical performance is sensitive to hyper-parameter tuning. On the other hand, deterministic strategies based on OFU, such as the celebrated UCB1 algorithm [Auer et al., 2002], construct closed-form high-probability confidence sets. OFU-based algorithms are theoreti-
Randomized UCB for Bandit Problems

Randomized strategies are computationally expensive even in the standard MAB or structured bandit settings, including MAB and linear bandits. However, since these confidence sets are constructed to obtain good worst-case performance, they often have poor empirical performance on typical problem instances. Moreover, for structured bandits, when the feature-reward mapping is non-linear (e.g., generalized linear models), we can only construct coarse confidence sets [Filippi et al., 2010; Zhang et al., 2016; Jun et al., 2017; Li et al., 2017], which are often too conservative in practice.

In contrast, TS is a randomized strategy that maintains a posterior distribution over the unknown parameters, and samples from it in order to choose actions. When the posterior has a closed form, as in the Bernoulli or Gaussian MAB or linear bandits, it is possible to sample exactly from it. In these cases, TS is computationally efficient and have good empirical performance [Chapelle and Li, 2011]. However, when there is no closed form posterior, one has to resort to approximate sampling techniques, which are typically expensive [Gopalan et al., 2014; Kawale et al., 2015; Riquelme et al., 2018] and limit the practical applicability of TS. From a theoretical point of view, TS results in near-optimal regret bounds for the MAB problem [Agrawal and Goyal, 2017], but current analyses result in a sub-optimal dependence on the feature dimension for structured bandits [Abeille and Lazaric, 2017; Agrawal and Goyal, 2013].

1.2 Randomized exploration strategies

There has been substantial recent research on using bootstrapping [Baransi et al., 2014; Eckles and Kaptein, 2014; Osband and Van Roy, 2015; Tang et al., 2015; McNellis et al., 2017; Vaswani et al., 2018] or designing general randomized exploration schemes [Kveton et al., 2019a,b,c,d; Kim and Tewari, 2019]. These data-driven strategies do not rely on problem-specific confidence sets, neither do they require a posterior distribution. Moreover, they are applicable even when the feature to reward mappings is a complex one (e.g., a neural networks) [Osband and Van Roy, 2015; Vaswani et al., 2018; Kveton et al., 2019c].

However, these strategies suffer from theoretical and practical drawbacks. In particular, for typical bootstrapping strategies, theoretical guarantees have been derived only for linear bandits and MAB with Gaussian or Bernoulli rewards [Lu and Roy, 2017; Osband and Van Roy, 2015; Vaswani et al., 2018]. General randomized strategies [Kveton et al., 2019b,c; Kim and Tewari, 2019] achieve near-optimal regret bounds in the general MAB setting; however, the degree of exploration is difficult to control, complicating their proofs. For structured bandits, randomized strategies have been proposed in the linear [Kveton et al., 2019a] and generalized linear [Kveton et al., 2019d] settings.

However, their analysis for linear bandits closely follows that of TS and inherits its sub-optimality in the feature dimension [Kveton et al., 2019a], and proving regret bounds for the generalized linear case [Kveton et al., 2019d] requires additional assumptions.

From a practical perspective, the advantage of these randomized strategies is that they do not rely on closed form posterior distributions like TS, but they “sample” from an implicit distribution. This distribution could be induced via bootstrapping [Osband and Van Roy, 2015; Lu and Roy, 2017; Vaswani et al., 2018], adding pseudo-observations [Kveton et al., 2019c], or randomizing the observed data [Kveton et al., 2019b,a,d; Kim and Tewari, 2019]. These choices complicate the resulting algorithm. Moreover, in order to generate a “sample”, these strategies require solving a maximum likelihood estimation problem in each round. Unlike computing an upper confidence bound (as in OFU) or sampling from the posterior (as for TS), this estimation problem cannot be solved in an efficient, online manner while preserving regret guarantees [Jun et al., 2017]. For computational efficiency, these strategies resort to heuristics for approximating the maximum likelihood estimator (MLE) [Vaswani et al., 2018; Kveton et al., 2019c; Osband and Van Roy, 2015; Lu and Roy, 2017]. Unfortunately, these approximations do not have rigorous theoretical guarantees and add another layer of complexity to the algorithm design.

2 Our Contribution

As general randomized strategies are complicated and computationally expensive even in the standard MAB or structured bandit settings, we consider randomizing simple OFU-based algorithms. To this end, we propose the RandUCB (meta-)algorithm, which relies on existing theoretically derived confidence sets, but similar to TS, it uses randomization to trade off exploration and exploitation. In Section 3, we describe the general framework of the RandUCB meta-algorithm.

In Section 4, we instantiate RandUCB in the MAB setting. We show that TS can be viewed as a special case of RandUCB in a specific MAB setting (Section 4.2). Furthermore, by reasoning about the algorithmic choices in RandUCB, we derive variants of classic exploration strategies. For example, we formulate optimistic Thompson sampling, a variant of TS which only generates posterior samples greater than the mean, and show that it results in comparable theoretical and empirical performance as TS (Appendix D.2). More generally, we show that there are infinitely many variants of RandUCB, all of which achieve the minimax-optimal $O(\sqrt{KT})$ regret for an MAB with $K$ arms over $T$ rounds (Section 4.3).

For structured bandits, we present an instantiation of the RandUCB meta-algorithm when the rewards follow a
linear (Section 5.1) or a generalized linear model (Section 5.2). We show that RandUCB achieves the optimal \( \tilde{O}(d\sqrt{T}) \) regret for \( d \)-dimensional feature vectors, even with infinitely many arms. In both these settings, RandUCB matches the theoretical regret bounds of the corresponding OFU-based algorithms [Abbasi-Yadkori et al., 2011; Li et al., 2017] up to constant factors. To the best of our knowledge, RandUCB is the first randomized algorithm that results in the near-optimal dependence on the dimension in the infinite-armed case. For all the above settings, the algorithm design of RandUCB enables simple proofs that extend naturally from the existing TS and OFU analyses.

Finally, we conduct experiments in the MAB and structured bandit settings\(^1\), investigating the impact of algorithmic design choices through an ablation study (Appendix D.1), and demonstrating the practical effectiveness and efficiency of RandUCB (Section 6).

In all settings, the performance of RandUCB is either comparable to or better than that of TS and the more complex, computationally expensive generalized randomized strategies.

3 The RandUCB Meta-Algorithm

In this section, we describe the general form of RandUCB and detail the design decisions. Consider a bandit setting with action set \( \mathcal{A} \). When arm \( i \in \mathcal{A} \) is pulled, a reward is drawn from its underlying distribution, with mean \( \mu_i \) and support \([0, 1]\), and is presented to the learner. The learner’s objective is to maximize its expected cumulative reward across \( T \) rounds.

An OFU-based bandit algorithm keeps track of the estimated mean \( \hat{\mu}_i(t) \), defined as the average of rewards received from arm \( i \) until round \( t \). The algorithm also maintains a confidence interval of size \( C_i(t) \) around the estimated mean. The value of \( C_i(t) \) decreases as an arm is pulled more, and indicates how accurate \( \hat{\mu}_i(t) \) is at estimating \( \mu_i \). Although the exact values of \( \hat{\mu}_i(t) \) and \( C_i(t) \) depend on the bandit setting under consideration, OFU-based strategies [Auer et al., 2002; Abbasi-Yadkori et al., 2011] have the same general form: in round \( t \), they choose the arm

\[
i_t = \arg \max_{i \in \mathcal{A}} \{ \hat{\mu}_i(t) + \beta \cdot C_i(t) \}. \tag{1}
\]

The parameter \( \beta \) is carefully chosen to trade off exploration and exploitation optimally. We will instantiate this algorithm for the multi-armed (Section 4), linear (Section 5.1), and generalized linear (Section 5.2) bandit settings. As a simple modification, RandUCB randomizes the confidence intervals and chooses the arm

\[
i_t = \arg \max_{i \in \mathcal{A}} \{ \hat{\mu}_i(t) + Z_t \cdot C_i(t) \}, \tag{2}
\]

where the deterministic quantity \( \beta \) is replaced by a random variable \( Z_t \). Here, \( Z_1, \ldots, Z_T \) are i.i.d. samples from the sampling distribution that we describe next.

3.1 The sampling distribution

The random variables \( Z_1, \ldots, Z_T \) are i.i.d. and have the same distribution as a template random variable \( Z \), explained below. We consider a discrete distribution for \( Z \) on the interval \([L, U]\), supported on \( M \) points. Let \( \alpha_1 = L, \ldots, \alpha_M = U \) denote \( M \) equally spaced points in \([L, U]\), and define \( p_m := P(Z = \alpha_m) \). If \( M = 1 \) and \( L = U = \beta \), then we recover the OFU-based algorithm, Eq. (1). If \( L = 0 \) and \( U = \beta \), then RandUCB chooses between values in the \([0, \beta]\) range; in this case, the \( \alpha_m \) can be viewed as nested confidence intervals. We choose a constant value for \( M \) throughout this paper, but note that letting \( M \to \infty \) can simulate a fine discretization of an underlying continuous distribution supported on \([L, U]\). To obtain optimal theoretical guarantees, the probabilities \( p_1, \ldots, p_M \) in RandUCB must be chosen in a way that ensures \( P(Z \geq \beta) > 0 \). This guarantees that the algorithm has enough optimism and we will later prove that this constraint ensures that RandUCB attains optimal regret for all the bandit settings we consider.

Our choice of the sampling distribution (the \( p_m \) values) is inspired from a Gaussian distribution truncated in the \([0, U]\) interval and has tunable hyper-parameters \( \epsilon, \sigma > 0 \). The former is the constant probability to be put on the highest point: \( \alpha_M = U \) with \( p_M = \epsilon \).

For the remaining \( M - 1 \) points, we use a discretized Gaussian distribution; formally, for \( 1 \leq m \leq M - 1 \), let \( p_m := \exp(-\alpha_m^2/2\sigma^2) \) and let \( p_m \) denote the normalized probabilities, that is, \( p_m := (1 - \epsilon) p_m / \sum_{m=1}^{M} p_m \).

The above choice can be viewed as a truncated (between 0 and \( U \)) and discretized (into \( M \) points) Gaussian distribution. As we explain in Section 4.2, choosing this distribution resembles Gaussian TS.\(^2\)

3.2 Algorithmic decisions

Optimism By only considering positive values for \( Z \) (by setting \( L = 0 \)), we maintain the OFU principle [Auer et al., 2002; Abbasi-Yadkori et al., 2011] of the corresponding OFU-based algorithm. Although our theoretical results allows \( Z \) to take negative values, we experimentally observe that this does not significantly improve the empirical performance of RandUCB (see Figure 3 in Appendix D.1).

Coupling the arms By default, in each round \( t \), RandUCB samples a single value of \( Z_t \) that is shared between all the arms (see Eq. (2)) thus “coupling” the arms. Alternatively, we could con-

\(^1\)See code: https://github.com/vaswanis/randucb.

\(^2\)One might also consider a discretized uniform distribution on \([0, U]\), but our experiments in Appendix D show that this choice performs poorly in practice.
Consider uncoupled RandUCB where in each round $t$, each arm $i$ generates its own independent copy of $Z$, say $Z_{i,t}$, and the algorithm selects the arm $i_t = \arg \max_i \left\{ \hat{\mu}_i(t) + Z_{i,t} C_i(t) \right\}$. This is similar to the Boltzmann exploration algorithm in Cesa-Bianchi et al. [2017]. However, our experiments show that the uncoupled variant does not perform better than the default, coupled version (see Figure 3 in Appendix D.1).

In the next sections, we revisit these decisions, instantiate RandUCB, and analyze its performance in specific bandit settings. The subsequent theoretical results hold for $L = 0$, any positive integer $M$, and any positive constants $\varepsilon$ and $\sigma$. The value of $U$ depends on the specific bandit setting. For the empirical evaluation (Section 6 and Appendix D), the specific values of $L$, $U$, $M$, $\varepsilon$, and $\sigma$ will be specified for each experiment.

## 4 Multi-Armed Bandit

In this section, we consider a stochastic multi-armed bandit (MAB) with $|A| = K$ arms. Without loss of generality, we may assume that arm 1 is optimal, namely $\mu_1 = \max_i \mu_i$, and refer to $\Delta_i = \mu_1 - \mu_i$ as the gap of arm $i$. Maximizing the expected reward is equivalent to minimizing the expected regret across $T$ rounds. If a bandit algorithm pulls arm $i$ in round $t$, then it incurs an expected (cumulative) regret of

$$R(T) := \sum_{t=1}^{T} E[\mu_1 - \mu_{i_t}] = \sum_{i=1}^{T} E[\Delta_i].$$

### 4.1 Instantiating RandUCB

Let $s_i(t)$ denote the number of pulls and $Y_i(t)$ denote the total reward received from arm $i$ by round $t$. Then the estimated mean is simply $\hat{\mu}_i(t) = Y_i(t)/s_i(t)$ (we set $\hat{\mu}_i(t) = 0$ if arm $i$ has never been pulled). The confidence interval corresponds to the standard deviation in the estimation of $\mu_i$ and is given as $C_i(t) = \sqrt{\frac{1}{s_i(t)}}$.

To ensure that $s_i(t) > 0$, RandUCB begins by pulling each arm once and in each subsequent round $t > K$, selects

$$i_t = \arg \max_i \left\{ \hat{\mu}_i(t) + Z_{i,t} \frac{1}{\sqrt{s_i(t)}} \right\}. \quad (3)$$

The corresponding OFU-based algorithm [Auer et al., 2002, Figure 1] sets the constant $\beta = \sqrt{2 \ln(T)}$. For RandUCB, we choose $L = 0$ and $U = 2\sqrt{\ln(T)}$, that is, we inflate\(^3\) the confidence interval by $\sqrt{2}$.

### 4.2 Connections to TS and EG

We now describe how RandUCB relates to existing algorithms. Recall that TS [Agrawal and Goyal, 2017] may draw samples below the empirical mean for each arm, whereas RandUCB with $L \geq 0$ samples from a one-sided distribution above the mean. In order to make the connection from RandUCB to TS, we consider a variant of TS which only samples values above the mean for each arm,\(^4\) referred to as optimistic Thompson sampling (OTS). Our experiments show that OTS has similar empirical performance to TS (Appendix D.2). We show that RandUCB with $M \to \infty$ approaches OTS with a Gaussian prior and posterior.

First, observe that the regret of RandUCB with $Z \sim \mathcal{N}(0,1)$ without truncation or discretization exactly corresponds to TS. Now consider optimistic TS and further truncate the tail of the Gaussian posterior at $2\sqrt{\ln(T)}$. By putting a constant probability mass of $\varepsilon$ at $2\sqrt{\ln(T)}$ (the upper bound of the distribution) and discretizing the resulting distribution at $M - 1$ equally-spaced points, we obtain the uncoupled variant of RandUCB.

The flexibility of RandUCB also allows us to consider a variant that resembles an adaptive $\varepsilon$-greedy strategy. Recall that the classical $\varepsilon$-greedy (EG) strategy [Auer et al., 2002; Langford and Zhang, 2008] chooses a random action with probability $\varepsilon$ and the greedy action with probability $1 - \varepsilon$. For a constant $\varepsilon$, EG might result in linear regret, whereas decreasing $\varepsilon$ over time results in a sub-optimal $O(T^{2/3})$ regret [Auer et al., 2002]. An adaptive $\varepsilon$-greedy can be instantiated from RandUCB as follows: let $Z$ be a random variable that takes value 0 with probability $1 - \varepsilon$ and $2\sqrt{\ln(T)}$ with probability $\varepsilon$. This results in choosing the greedy action with probability $1 - \varepsilon$ and choosing the action that maximizes the data-dependent upper-confidence-bound with probability $\varepsilon$. Theorem 1 below implies that the regret of this modification of the $\varepsilon$-greedy algorithm is bounded by $O(\sqrt{KT \ln(KT)})$.

### 4.3 Regret of RandUCB for MAB

In this section, we first bound the regret of the default optimistic, coupled variant of RandUCB with a general distribution for $Z$ and then obtain a bound for the uncoupled variant.

**Theorem 1** (Minimax regret of RandUCB with coupled arms for MAB). Let $c_1 := 1 + \sqrt{\ln(KT^2)}$ and $c_3 := 2K \ln(1 + \frac{T}{K})$. For any $c_2 > c_1$, the regret $R(T)$ of RandUCB for MAB is bounded by

$$\left( c_1 + c_2 \right) \left( 1 + \frac{2}{P(Z > c_1) - P(|Z| > c_2)} \right) \times \sqrt{c_3 T} + T \left( P(|Z| > c_2) + K + 1 \right).$$

The proof for the above theorem uses a reduction from linear bandits; we defer it to Section 5.1.3. The above result implies that the regret of RandUCB can be bounded by $O(\sqrt{KT \ln(KT)})$ so long as (i) $P \left( Z > 1 + \sqrt{\ln(KT^2)} \right) > 0$ and (ii) $|Z| \leq c_2$ deter-

\(^3\)This inflation is a technicality needed for our analysis.

\(^4\)It samples from a conditional posterior distribution, conditioned on the sample being larger than the mean.
ministically. By choosing $U = 2\sqrt{\ln T}$, our sampling distribution would lie in $[0, 2\sqrt{\ln T}]$, so condition (ii) holds by setting $c_2 = 2\sqrt{\ln T}$ in Theorem 1. Since the considered sampling distribution in Section 3.1 has a constant probability mass of $\varepsilon$ at $U = 2\sqrt{\ln(T)}$ by design, it ensures that $P(Z > c_1)$ is a positive constant. Since any consistent algorithm for MAB has regret at least $\Omega(\sqrt{KT})$ (see, e.g., Lattimore and Szepesvári [2020, Theorem 15.2]), RandUCB is minimax-optimal up to logarithmic factors.

The next result states that uncoupled RandUCB achieves problem-dependent logarithmic regret, therefore also being nearly-optimal.

**Theorem 2** (Instance-dependent regret of uncoupled RandUCB for MAB). If $Z$ takes $M$ different values $0 \leq \alpha_1 \leq \cdots \leq \alpha_M$ with probabilities $p_1, p_2, \ldots, p_M$, the regret $R(T)$ of uncoupled RandUCB can be bounded as $O\left(\sum_{i=1}^{\beta} \varepsilon_i\right) \times \left(\frac{M}{p_M} + T e^{-2\alpha_M^2} + \alpha_M^2\right)$.

Since the sampling distribution of RandUCB satisfies $U = \alpha_M = 2\sqrt{\ln T}$ and $M$ and $p_M$ are constant, uncoupled RandUCB attains the optimal instance-dependent regret $O\left(\ln T \times \left(\sum_{i=1}^{\beta} \varepsilon_i\right)\right)$ (see, e.g., Lattimore and Szepesvári [2020, Theorem 16.4]). By a standard reduction, Theorem 2 implies that uncoupled RandUCB achieves the problem-independent $O(\sqrt{KT})$ regret. Please refer to Appendix A for the proof and the statement of Theorem 6 for a tighter regret bound.

5 Structured Bandits

In this section, we consider the structured bandit setting where each arm is associated with a $d$-dimensional feature vector and there exists an underlying parametric function that maps these features to rewards. Let $x_i \in \mathbb{R}^d$ denote the corresponding feature vector for arm $i \in A$. We assume that $d > 1$ and $\|x_i\| \leq 1$ for every arm $i$. We also assume that the function mapping a feature vector to the expected reward is parameterized by an unknown parameter vector $\theta^*$ with $\|\theta^*\| \leq 1$, and that the rewards lie in $[0, 1]$.

5.1 Linear bandits

In linear bandits, the expected reward of an arm is the dot product of its corresponding feature vector and the unknown parameter. Formally, if $Y_i$ is the reward obtained in round $t$, then $E[Y_i|t] = \langle x_i, \theta^* \rangle$. If $i_t$ is the arm pulled in round $t$ and arm $1$ is the optimal arm, then the regret can be defined similarly as in the MAB case, but with an “effective” gap $\Delta_i = \langle x_1 - x_i, \theta^* \rangle$.

$$R(T) := \sum_{t=1}^{T} E[(x_1 - x_{i_t}, \theta^*)] = \sum_{t=1}^{T} E[\Delta_{i_t}]. \quad (4)$$

Let us denote $X_t := x_i$, and define the Gram matrix $M_t := \lambda I_d + \sum_{t=1}^{T-1} X_t X_t^T$. Here, $\lambda > 0$ is the $\ell_2$ regularization parameter. We define the norm $\|x\|_M := \sqrt{x^T M x}$ for any positive definite $M$.

5.1.1 Instantiating RandUCB

Given the observations $(X_t, Y_t)_{t=1}^T$ gathered until round $t$, the maximum likelihood estimator (MLE) for linear regression is $\hat{\theta}_t = \text{arg max}_i \{\hat{\theta}_t, x_i\}$ and the corresponding confidence interval is $C_t = \|x_i\|_{M^{-1}_t}$. Thus, RandUCB chooses arm

$$i_t := \text{arg max}_{i \in A} \{\langle \hat{\theta}_t, x_i \rangle + Z_t \|x_i\|_{M^{-1}_t}\}. \quad (5)$$

Note that the corresponding OFU-based algorithm [Abbasi-Yadkori et al., 2011, Theorem 2] sets $\beta = \sqrt{\lambda + \frac{1}{2} \sqrt{\log(T^2\lambda^{-d}\det(M_t))}}$. We prove the following theorem for RandUCB.

5.1.2 Regret of RandUCB for linear bandits

**Theorem 3.** Let $c_1 = \sqrt{\lambda} + \frac{1}{2} \sqrt{d \ln (T + T^2/d\lambda)}$ and $c_3 := 2\ln (1 + \frac{1}{\lambda})$. For any $c_2 > c_1$, the regret of RandUCB for linear bandits is bounded by

$$(c_1 + c_2) \left(1 + \frac{2}{\log c_1 - \log c_2}\right) \times \sqrt{c_3 T} + T \mathbb{P}(\|Z\| > c_2).$$

**Proof.** Let $\tilde{f}_t(x) := \langle \hat{\theta}_t, x \rangle + Z_t \|x\|_{M^{-1}_t}$, and define the events

$$E^{\text{F}} := \{\forall i \in [K], \forall t \in [T]; \ |\langle x_i, \hat{\theta}_t - \theta^* \rangle| \leq c_1 \|x_i\|_{M^{-1}_t}\};$$

$$E^{\text{conc}}_t := \{\forall i \in [K]; \ |\tilde{f}_t(x_i) - \langle x_i, \hat{\theta}_t \rangle| \leq c_2 \|x_i\|_{M^{-1}_t}\};$$

$$E^{\text{anti}}_t := \{\tilde{f}_t(x_1) - \langle x_1, \hat{\theta}_t \rangle > c_1 \|x_1\|_{M^{-1}_t}\};$$

and assume for now that we have the following bounds for their probabilities: $\mathbb{P}(E^{\text{F}}) \geq 1 - p_1$, $\mathbb{P}(E^{\text{conc}}_t) \geq 1 - p_2$, and $\mathbb{P}(E^{\text{anti}}_t) \geq p_3$.

In Appendix B, we prove an upper bound for the regret of any index-based algorithm in terms of $p_1, p_2$, and $p_3$. An index-based algorithm is one that in each round $t$ chooses the arm $i_t$ that maximizes some function $\tilde{f}_t(x)$, i.e., $i_t = \text{arg max}_i \tilde{f}_t(x_i)$. Theorem 7 in Appendix B
bounds the regret of such an algorithm by
\[
(c_1 + c_2) \left(1 + \frac{2}{p_3 - p_2}\right) \sqrt{c_3 T} + T \left(p_1 + p_2\right).
\]

For RandUCB, we have \(\bar{f}_t(x) = \langle \hat{\theta}_t, x \rangle + Z_t \|x\|_{M_t^{-1}}\). Event \(E^a\) concerns the concentration of the MLE and does not depend on the algorithm. By Abbasi-Yadkori et al. [2011, Theorem 2], we have \(p_1 \leq 1/T\). By definition of \(\bar{f}_t\), \(P\left(E^\text{anti} \cap \{Z_t > c_1\}\right) =: p_3\) and \(P\left(E^\text{con}\right) = P\left(|Z_t| > c_2\right) =: p_2\). These relations combined with the bound Eq. (6) completes the proof of Theorem 1.

5.2 Generalized linear bandits

We next consider structured bandits where the feature to reward mapping is a generalized linear model [McCullagh, 1984], meaning that the expected reward in round \(t\) satisfies \(E[Y_t^{i_t}] = g(\langle x_t, \theta^* \rangle) \in [0, 1]\), where \(g\) is a known, strictly increasing, differentiable function, called the link function or the mean function. If \(g(x) = x\), we recover linear bandits. Assuming arm 1 is optimal, the regret is \(R(T) = \sum_{t=1}^T E[g(x_1, \theta^*) - g(x_t, \theta^*)]\) and the effective gap of arm \(i\) is \(\Delta_i := g(x_1, \theta^*) - g(x_i, \theta^*)\).

5.2.1 Instantiating RandUCB

As before, we denote \(X_t = x_{i_t}\). Given previous observations \((X_t, Y_t)_{t=1}^{\ell_t}\), the MSE in round \(t\) can be computed as [McCullagh, 1984] \(\hat{\theta}_t \triangleq \arg\min_{\theta} \sum_{\ell=1}^{t-1} [Y_\ell \langle X_\ell, \theta \rangle - b(\langle X_\ell, \theta \rangle)]\), where \(b\) is a strictly convex function such that its derivative is \(g\). Let \(H_t(\theta) := \sum_{\ell=1}^{t-1} g'(\langle X_\ell, \theta \rangle)X_\ell X_\ell^T \) denote the Hessian at point \(\theta\) on round \(t\), and \(H_t := H_t(\hat{\theta}_t)\). We assume that \(g\) is \(L\)-Lipschitz, i.e., \(|g(x) - g(y)| \leq L|x - y|\), implying \(0 \leq g'(x) \leq L\) for all \(x\).

Note that in general, matrix \(H_t\) is not guaranteed to be positive definite. To guarantee the positive definiteness of \(H_t\), we make the following assumptions.\(^6\)

(i) Feature vectors span the \(d\)-dimensional space. In particular, we assume that there exist basis vectors \(v_j^d_{i=1} \leq \{x_i\}_{i=1}^d\) with \(\sum_{j=1}^d v_j v_j^T \geq \rho I\) for some \(\rho > 0\). This assumption is natural as it would not hold only when the actual dimensionality of the problem is smaller than \(d\).

(ii) We assume
\[
\mu := \inf \{g'(x, \theta) : \|x\| \leq 1, \|\theta - \theta^*\| \leq 1\} > 0.
\]

This assumption holds for all interesting link functions, such as in linear and logistic regression.

RandUCB for GLB starts by pulling each of the \(v_i\) for \(O(d \ln(T)/\mu^2 \rho)\) many times. We shall show that after this initialization, with probability at least \(1 - 1/T\) we have that \(\|\hat{\theta}_t - \theta^*\| \leq 1\) and further \(H_t\) is positive-definite for all subsequent rounds. After this initialization, RandUCB follows the same algorithm as for linear bandits (Eq. (5)), except that there is no regularization in this case (so, \(M_t = \sum_{\ell=1}^{t-1} X_\ell X_\ell^T\)).

The corresponding OFU-based algorithm [Li et al.,

---

\(^6\)These assumptions are standard in the analysis of generalized linear bandits [Li et al., 2017; Kveton et al., 2019d].
2017] has \( \beta = \frac{1}{2} \sqrt{\frac{d}{T}} \ln(1 + 2T/d) + \ln(T) \). Let \( c_1 := \sqrt{d \ln(T/d)} + 2 \ln(T)/2 \mu \), and choose \( U = 2\sqrt{c_1} \) for RandUCB: the following theorem, proved in Appendix C, gives the promised \( O(d\sqrt{T}) \) regret bound by choosing \( c_2 = 3\sqrt{c_1} \).

**Theorem 4.** Let \( c_1 = \sqrt{d \ln(T/d)} + 2 \ln(T)/2 \mu \), \( c_2 := 2d \ln \left( 1 + \frac{T}{d} \right) \). For any \( c_2 > c_1 \), the regret \( R(T) \) of RandUCB for generalized linear bandits is bounded by

\[
\left( c_1 + \frac{c_2}{\sqrt{U}} \right) \left( 1 + \frac{2}{P \left( Z > c_1 \sqrt{U} \right) - P \left( |Z| > c_2 \right)} \right) \\
\times L \sqrt{c_3 T} + T P \left( |Z| > c_2 \right) + O(d^2 \ln(T)/\mu^2 \rho).
\]

6 Experiments

Finally, we empirically evaluate the performance of RandUCB on the bandit settings studied in this paper. We compare various algorithms based on their cumulative empirical regret \( \sum_{t=1}^T [Y_t^* - Y_t] \), where \( Y_t^* \) denotes the reward received by the optimal arm and \( Y_t \) is the reward received by the algorithm in round \( t \). For all the experiments, we consider \( |A| = K = 100 \) arms and set \( T = 20,000 \) rounds. We average our results over 50 randomly generated bandit instances.

**Multi-armed Bandits:** We first consider the MAB setting and investigate the impact of the gap sizes and the reward distribution. We consider an easy class and a hard class of problem instances: in the former, arm means are sampled uniformly in \([0.25, 0.75] \), while in the latter, they are sampled in \([0.45, 0.55] \). We consider both discrete, binary rewards sampled from Bernoulli distributions, as well as continuous rewards sampled from beta distributions. We present results for a Gaussian reward distribution in Appendix D.3. In Appendix D.1, we investigate the impact of the design choices and parameters of RandUCB in the MAB setting. Recall that RandUCB is characterized by the choice of sampling distribution (Section 3.1). We compare the performance of the uniform and Gaussian distributions (with different standard deviations \( \sigma \)), and observe in Figure 2 that lower values of \( \sigma \) result in better performance in all our experiments. We also observe in Figure 4 that RandUCB is robust to the value of \( M \), the extent of discretization. Note that previous work has also observed that the empirical performance of UCB1 can be improved by using smaller confidence intervals than suggested by theory [Hsu et al., 2019; Li et al., 2012], e.g., by tuning \( \sqrt{\beta} \). In contrast to our work, these heuristics do not have theoretical guarantees.

We then estimate the impact of optimism and coupling of the arms on the empirical performance of RandUCB. In Figure 3, we observe that coupling the arms is more determinant in improving the performance of RandUCB compared with optimism, which has only a minor effect. We notice that this phenomenon is also observed for TS: the optimistic variant of TS (OTS) has similar performance to TS (Figure 5 in Appendix D.2).

Following the above ablation study, in the following experiments we initiate RandUCB with a (discretized, optimistic) Gaussian sampling distribution and coupled arms with parameters \( \varepsilon = 10^{-7}, \sigma = 1/8, L = 0, U = 2\sqrt{\ln(T)} \), and \( M = 20 \). Figure 1(a) compares RandUCB against classical and state-of-the-art baselines. In particular, we compare against TS with Bernoulli-Beta conjugate priors (B-TS) [Agrawal and Goyal, 2017] and UCB1 [Auer et al., 2002]. We also consider the much tighter KL-UCB version [Garivier and Cappé, 2011], in addition to the recent GiRo [Kveton et al., 2019c] and PHE [Kveton et al., 2019b] algorithms and observe that RandUCB performs consistently well, clearly outperforming all baselines in three settings, while matching the performance of PHE in the remaining setting. Most importantly, it outperforms TS in all settings.

**Structured Bandits:** For structured bandits, we use the same setting of RandUCB described above but with the confidence intervals given by the specific bandit problem. We consider linear bandits as well as logistic regression for the generalized linear case.

For each of these problems, we vary the dimension \( d \in \{5, 10, 20\} \). Each problem is characterized by an (unknown) parameter \( \theta^* \) and \( K \) arms. We consider Bernoulli \( \{0, 1\} \) rewards.

For RandUCB in the linear bandit setting, we use the same hyper-parameters as before, but set \( U = \beta = \sqrt{\lambda} + \frac{1}{\sqrt{2}} \sqrt{\ln(T^2 \lambda^{-d} \det(M_\lambda))} \), which is the value from the corresponding OFU-based algorithm [Abbasi-Yadkori et al., 2011, Theorem 2], and \( \lambda = 10^{-4} \). For comparison, we consider two variants of LinTS [Abeille and Lazaric, 2017; Agrawal and Goyal, 2013]: a theoretically optimal variant with the covariance matrix “inflated” by a dimension-dependent quantity and the more commonly used variant without this additional inflation [Chapelle and Li, 2011]. We also consider LinUCB [Abbasi-Yadkori et al., 2011], \( \varepsilon \)-greedy [Langford and Zhang, 2008], and the best performing variant of the randomized strategy LinPHE [Kveton et al., 2019a]. For \( \varepsilon \)-greedy, we chose the best performing value of \( \varepsilon = 0.05 \) and anneal it as \( \varepsilon_t = \frac{\varepsilon}{\sqrt{T}} \).

For RandUCB in the GLB setting, we use the same hyper-parameters as before, but now set \( U = \beta = \frac{1}{\sqrt{2}} \sqrt{\ln(1 + 2T/d) + \ln(T)} \), which is the constant

\[ \frac{1}{\sqrt{2}} \sqrt{\ln(1 + 2T/d) + \ln(T)} \]

To make sure the expected rewards lie in \([0, 1]\), we choose each of \( \theta^* \) and the feature vectors by sampling a uniformly random \((d - 1)\)-dimensional vector of norm \( 1/\sqrt{2} \) and concatenate it with a \( 1/\sqrt{2} \) component.
Randomized UCB for Bandit Problems

Figure 1: Cumulative empirical regrets of RandUCB versus competitors on various bandit settings.

(a) Various configurations of MAB with large/small gaps (easy/hard) and different reward distributions.

(b) Linear bandits of different dimensions $d$. RandLinUCB is the instantiation of RandUCB for linear bandits.

(c) Logistic bandits of different dimensions $d$. RandUCBLog is the instantiation of RandUCB for logistic bandits.

Figure 1: Cumulative empirical regrets of RandUCB versus competitors on various bandit settings.

from the corresponding OFU-based algorithm [Li et al., 2017]. We compare against GLM-TS [Abeille and Lazaric, 2017; Kveton et al., 2019d], which samples from a Laplace approximation of the posterior distribution. We consider two OFU-based algorithms: GLM-UCB [Filippi et al., 2010] and UCB-GLM [Li et al., 2017]. For RandUCB, we chose to randomize the tighter confidence intervals in UCB-GLM by the same scheme in Eq. (5). We further compare against ε-greedy [Langford and Zhang, 2008] and the best performing variant of LogPHE [Kveton et al., 2019d].

Figure 1(b) shows that RandUCB matches the performance of the best strategies in linear bandits. Figure 1(c) shows that RandUCB is competitive against other state-of-the-art strategies in logistic bandits. These results confirm that RandUCB is robust to the problem configuration and is an effective randomized alternative to complicated strategies.

7 Conclusion

We introduced the RandUCB meta-algorithm as a generic strategy for randomizing OFU-based algorithms. Our results across bandit settings illustrate that RandUCB matches the empirical performance of TS (and often outperforms it) and yet attains the theoretically optimal regret bounds of OFU-based algorithms, thus achieving the best of both worlds. An additional advantage of RandUCB is its broad applicability: the same mechanism of randomizing upper confidence bounds can be potentially used to improve the performance of other OFU-based algorithms. This could be useful in domains such as Monte-Carlo tree search [Kocsis and Szepesvári, 2006] and risk-aware bandits [Galichet et al., 2013], where designing randomized exploration strategies is not straightforward, as well as for practical scenarios such as delayed rewards [Chapelle and Li, 2011], where randomization is crucial for robustness.
8 Acknowledgements

We would like to thank Aditya Ramdas for pointing out a mistake in an earlier version of the paper. This research was partially supported by an IVADO postdoctoral scholarship. Abbas Mehrabian was supported by an IVADO-Apogée-CFREF postdoctoral fellowship.

References


Branislav Kveton, Manzil Zaheer, Csaba Szepesvári, Lihong Li, Mohammad Ghavamzadeh, and Craig


A Regret Bound for Multi-Armed Bandits with Uncoupled Arms

**Proposition 1** (Hoeffding’s inequality [Hoeffding, 1963, Theorem 2]). Let $X_1, \ldots, X_n$ be independent random variables taking values in $[0, 1]$. Then, for any $t \geq 0$, 
\[
\Pr \left[ \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) > tn \right] < \exp(-2nt^2).
\]

Recall that RandUCB first pulls each arm once and then in round $t > K$, chooses an arm $i$ maximizing $\hat{\mu}_i(t) + \frac{Z_{i,t}}{\sqrt{s_i(t)}}$, where the $Z_{i,t}$ are i.i.d. and distributed like a given random variable $Z$. We may assume, without loss of generality, that arm $1$ is the unique optimal arm. For a random variable $X$, the notation $P_X$ means taking the probability with respect to the randomness in $X$. We will also use the shorthand $\hat{\mu}_{i,s} := \hat{\mu}_i(s)$.

We will follow the usual result, which follows from Theorem 1 from Kveton et al. [2019c].

**Theorem 5.** Let $\tau_1, \ldots, \tau_K$ be arbitrary but deterministic. The regret of RandUCB after $T$ rounds can be upper bounded by $K + \sum_{i=2}^{K} \Delta_i (a_i + b_i)$, where 
\[
a_i := \sum_{s=1}^{T-1} \mathbb{E}_{\hat{\mu}_i,s} \left[ \min \left\{ \frac{1}{P_{Z_{1,s}} \left( \hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i \right)} - 1, T \right\} \right],
\]
\[
b_i := 1 + \sum_{s=1}^{T-1} P_{\hat{\mu}_i,s} \left\{ P_{Z_{1,s}} \left( \hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i \right) > 1/T \right\}.
\]

In the rest of this section, we explain how the following result follows from the above theorem. Assume that $Z$ has a discrete distribution and takes value $\alpha_i$ with probability $p_i$.

**Theorem 6.** Assume that $0 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_M$, and $p_i > 0$, and suppose $p_M > 1/T$. Then, the regret of RandUCB after $T$ rounds is bounded by 
\[
K + \left( \sum_{n=1}^{M-1} \frac{p_1 + \cdots + p_n}{p_{n+1} + \cdots + p_M} e^{-2\alpha_n^2} + T e^{-2\alpha_M^2} + 4 + 3\alpha_M^2 \right) \cdot \left( \sum_{i=2}^{K} \left( \frac{6}{\Delta_i} \right) \right).
\]

Theorem 2 follows from Theorem 6 by crudely bounding $\frac{p_1 + \cdots + p_n}{p_{n+1} + \cdots + p_M} \leq 1/p_M$ and $e^{-2\alpha_n^2} \leq 1$.

Theorem 6 follows from Theorem 5 by setting $\tau_i = \mu_i + \Delta_i/2 = \mu_1 - \Delta_i/2$. We bound the $a_i$ and the $b_i$ in the following two sections.

**A.1 Bounding the $a_i$**

Let us define 
\[
a_{i,s} := \mathbb{E}_{\hat{\mu}_{i,s}} \left[ \min \left\{ \frac{1}{P_{Z_{1,s}} \left( \hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i \right)} - 1, T \right\} \right].
\]

Fix $s$ and for each $1 \leq j \leq M$, define the event $E_j$ as 
\[
E_j := \left\{ \hat{\mu}_{1,s} + \alpha_j/\sqrt{s} \geq \tau_i \right\},
\]

which is deterministic given the history. Note that $E_M \supset E_{M-1} \supset \cdots \supset E_1$. Then, define 
\[
N := \min\{j : E_j \text{ holds}\},
\]
and \( N = M + 1 \) if none of the \( E_j \) hold. Since the events \( N = 1, N = 2, \ldots, N = M, N = M + 1 \) partition the space, we can bound \( a_{i,s} \) as

\[
\begin{aligned}
a_{i,s} &= \sum_{n=1}^{M+1} P_{\hat{\mu}_{1,s}} \{ N = n \} E_{\hat{\mu}_{1,s}} \left[ \min \left\{ \frac{1}{P_{Z_{i,s}}(\hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i)} - 1, T \right\} : N = n \} \\
&\leq \sum_{n=1}^{M} P_{\hat{\mu}_{1,s}} \{ N = n \} E_{\hat{\mu}_{1,s}} \left[ \frac{1}{P_{Z_{i,s}}(\hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i)} - 1 \right| N = n \} + TP_{\hat{\mu}_{1,s}} \{ N = M + 1 \}.
\end{aligned}
\]

Next, observe that by definition of \( N \), under the event \( N = n \) with \( 2 \leq n \leq M \), we have

\[
\frac{1}{P_{Z_{i,s}}(\hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i)} - 1 = \frac{1}{p_1 + p_2 + \cdots + p_M} - 1 = \frac{p_1 + \cdots + p_{n-1}}{p_n + \cdots + p_M}.
\]

Moreover, if \( N = 1 \) then \( P_{Z_{i,s}}(\hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i) = 1 \) and thus \( \frac{1}{P_{Z_{i,s}}(\hat{\mu}_1(s) + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i)} - 1 = 0 \). Therefore, we have

\[
\begin{aligned}
a_{i,s} &\leq \sum_{n=2}^{M} P_{\hat{\mu}_{1,s}} \{ N = n \} \cdot \left( \frac{p_1 + \cdots + p_{n-1}}{p_n + \cdots + p_M} \right) + TP_{\hat{\mu}_{1,s}} \{ N = M + 1 \}.
\end{aligned}
\]

We next bound the probability \( P_{\hat{\mu}_{1,s}} \{ N = n \} \), for any \( 2 \leq n \leq M + 1 \). Note that \( N = n \) implies \( E_{n-1} \) did not happen. That is, if \( N = n \), then

\[
\hat{\mu}_{1,s} + \frac{\alpha_{n-1}}{\sqrt{s}} < \tau_1 = \mu_1 - \Delta_i/2,
\]

which is equivalent to

\[
\mu_1 - \hat{\mu}_{1,s} > \frac{\alpha_{n-1}}{\sqrt{s}} + \Delta_i/2.
\]

Since \( \alpha_{n-1} \geq 0 \), we apply Hoeffding’s inequality to conclude

\[
P_{\hat{\mu}_{1,s}} \{ N = n \} \leq P_{\hat{\mu}_{1,s}} \{ \mu_1 - \hat{\mu}_{1,s} > \frac{\alpha_{n-1}}{\sqrt{s}} + \frac{\Delta_i}{2} \} \leq \exp \left( -2s \left( \frac{\alpha_{n-1}}{\sqrt{s}} + \frac{\Delta_i}{2} \right)^2 \right) \leq \exp \left( -2 \alpha_{n-1}^2 - s \Delta_i^2/2 \right).
\]

Therefore, we have

\[
a_{i,s} \leq \sum_{n=2}^{M} \exp(-2\alpha_{n-1}^2 - s\Delta_i^2/2) \cdot \left( \frac{p_1 + \cdots + p_{n-1}}{p_n + \cdots + p_M} \right) + T \exp(-2\alpha_M^2 - s\Delta_i^2/2),
\]

and so

\[
a_i \leq \sum_{s=0}^{T-1} \left( \sum_{n=1}^{M-1} \exp(-2\alpha_n^2 - s\Delta_i^2/2) \cdot \left( \frac{p_1 + \cdots + p_n}{p_{n+1} + \cdots + p_M} \right) + T \exp(-2\alpha_M^2 - s\Delta_i^2/2) \right) \\
\leq \sum_{n=1}^{M-1} \left( \frac{p_1 + \cdots + p_n}{p_{n+1} + \cdots + p_M} \right) e^{-2\alpha_n^2} \sum_{s=0}^{\infty} e^{-s\Delta_i^2/2} \right) + T e^{-2\alpha_M^2} \sum_{s=0}^{\infty} e^{-s\Delta_i^2/2}.
\]

We now bound

\[
\sum_{s=0}^{\infty} e^{-s\Delta_i^2/2} = \frac{1}{1 - e^{-\Delta_i^2/2}} \leq \frac{1}{\Delta_i^2/6},
\]

which gives

\[
a_i \leq \frac{6}{\Delta_i^2} \cdot \left( \sum_{n=1}^{M-1} \frac{p_1 + \cdots + p_n}{p_{n+1} + \cdots + p_M} e^{-2\alpha_n^2} + Te^{-2\alpha_M^2} \right).
\]
A.2 Bounding the $b_i$

Recall that

$$ b_i - 1 = \sum_{s=0}^{T-1} P_{\tilde{\mu}_{i,s}} \left\{ P_{Z_{i,s}} \left( \tilde{\mu}_{i,s} + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i \right) > 1/T \right\}. $$

Let

$$ b_{i,s} := P_{\tilde{\mu}_{i,s}} \left\{ P_{Z_{i,s}} \left( \tilde{\mu}_{i,s} + \frac{Z_{i,s}}{\sqrt{s}} \geq \tau_i \right) > 1/T \right\}. $$

By the definition of $Z_{i,s}$, we have

$$ b_{i,s} = P_{\tilde{\mu}_{i,s}} \left( \sum_{n=1}^{M} p_n I \left\{ \tilde{\mu}_{i,s} + \frac{\alpha_n}{\sqrt{s}} \geq \tau_i \right\} > 1/T \right). $$

Since $p_M > 1/T$, we have

$$ b_{i,s} = P_{\tilde{\mu}_{i,s}} \left( \tilde{\mu}_{i,s} + \frac{\alpha M}{\sqrt{s}} \geq \tau_i \right) = P_{\tilde{\mu}_{i,s}} \left( \tilde{\mu}_{i,s} + \frac{\alpha M}{\sqrt{s}} \geq \mu_i + \Delta_i/2 \right). $$

For $s \leq [16\alpha^2_M / \Delta_i^2]$, we simply upper bound $b_{i,s} \leq 1$, while for $s \geq [16\alpha^2_M / \Delta_i^2]$, we have $\Delta_i/2 - \alpha M/\sqrt{s} \geq \Delta_i/4 > 0$, so by Hoeffding’s inequality,

$$ b_{i,s} \leq P_{\tilde{\mu}_{i,s}} \left\{ \tilde{\mu}_{i,s} - \mu_i \geq \Delta_i/4 \right\} \leq \exp(-s\Delta_i^2/8). $$

Therefore,

$$ b_i - 1 = \sum_{s=0}^{T-1} b_{i,s} \leq [16\alpha^2_M / \Delta_i^2] + \sum_{s=[16\alpha^2_M / \Delta_i^2]}^{T-1} b_{i,s} \leq 16\alpha^2_M / \Delta_i^2 + \sum_{s=0}^{\infty} e^{-s\Delta_i^2/8} \leq \frac{24}{\Delta_i^2}, $$

where we have used $\sum_{s=0}^{\infty} e^{-s\Delta_i^2/8} \leq \frac{24}{\Delta_i^2}$, proved as in (7).

B A regret bound for a general index-based algorithm for linear bandits

In this appendix, we obtain an analogous result as Theorem 1 from Kveton et al. [2019a] for any index-based algorithm for linear bandits which in each round $t$ selects the arm $i_t := \arg \max \tilde{f}_i(x_t)$, where $\tilde{f}_i$ is a stochastic function depending on the history up to round $t$, and possibly additional independent randomness. In RandUCB, we will set $f_t(x) := \langle \theta_t, x \rangle + Z_t \|x\|_{M_{t-1}}^{-1}$, where $Z_1, \ldots, Z_T$ are i.i.d. random variables. The proof is quite similar to that of Theorem 1 from Kveton et al. [2019a], where this result was proved for the special case when $\tilde{f}_i$ is linear. To extend the proof to a general $\tilde{f}_i$, the first step in the proof of Lemma 1 has been changed. We include the complete proof for completeness.

Theorem 7 (Generic regret bound for index-based algorithms for linear bandits). Let $c_3 := 2d \ln (1 + \frac{1}{\Delta_i})$. Suppose $c_1 < c_2$ are real numbers satisfying $c_1 + c_2 \geq 1$ and define the events

$$ E_i^{ls} := \left\{ \forall i \in A, \forall d + 1 \leq t \leq T; \ |\langle x_i, \hat{\theta}_t - \theta^* \rangle| \leq c_1 \|x_i\|_{M_{t-1}} \right\}, $$

$$ E_i^{conc} := \left\{ \forall i \in A; \ |\tilde{f}_t(x_i) - \langle x_i, \hat{\theta}_t \rangle| \leq c_2 \|x_i\|_{M_{t-1}} \right\}, $$

and

$$ E_i^{anti} := \left\{ \langle \tilde{f}_t(x_i), \langle x_i, \hat{\theta}_t \rangle > c_1 \|x_i\|_{M_{t-1}} \right\}. $$

Suppose $p_1, p_2, p_3 \in [0, 1]$ are such that $P(E_i^{ls}) \geq 1 - p_1$ and that for each given $t \leq T$ and for any possible history $H_{t-1}$ until the end of round $t - 1$, we have

$$ P(E_i^{anti}|H_{t-1}) \geq p_3, \text{ and } P(E_i^{conc}|H_{t-1}) \geq 1 - p_2. $$

Then, the regret after $T$ rounds is bounded by $(c_1 + c_2) \left( 1 + \frac{2}{p_3 - p_2} \right) \sqrt{c_3 T} + T (p_1 + p_2)$. 

In the rest of this section, we prove Theorem 7. Let us denote by \( E_t, P_t \) the randomness injected by the algorithm in round \( t \) (i.e., the randomness of \( Z_t \)), and denote by \( E_{H_{t-1}}, P_{H_{t-1}} \) the randomness in the history. The following lemma is an analogue of Lemma 2 from Kveton et al. [2019a]. Let \( \Delta_i := (x_i - x_i, \theta^*) \) denote the gap of arm \( i \).

**Lemma 1.** For any round \( t \) and any history \( H_{t-1} \), we have

\[
E_t[\Delta_i, I \{E^{ls}\} | H_{t-1}] \leq p_2 + (c_1 + c_2) \left( 1 + \frac{2}{p_3 - p_2} \right) E_t[\min\{1, \|X_t\|_{M_t^{-1}}\} | H_{t-1}].
\]

Before proving this lemma we show how it implies Theorem 7. We will be using Lemma 11 from Abbasi-Yadkori et al. [2011], which states that, deterministically,

\[
\sum_{t=1}^{T} \min\{1, \|X_t\|_{M_t^{-1}}\}^2 \leq c_3.
\]  

(8)

**Proof of Theorem 7.** Recall that \( X_t = x_i \) and \( Y_t \) denotes the reward received in round \( t \). By Eq. (8) and the Cauchy-Schwarz inequality, we have, deterministically,

\[
\sum_{t=1}^{T} \min\{1, \|X_t\|_{M_t^{-1}}\} \leq \sqrt{c_3 T}.
\]

Thus, since \( \Delta_i \leq 1 \) for all \( i \), and using Lemma 1 we have

\[
\text{Regret} = \sum_{i=1}^{T} E \Delta_i \\
\leq TP(\overline{E^{ls}}) + \sum_{t=1}^{T} E[\Delta_i, I \{E^{ls}\} | H_{t-1}]
\]

\[
\leq Tp_1 + \sum_{t=1}^{T} E_{H_{t-1}}[E_t[\Delta_i, I \{E^{ls}\} | H_{t-1}]]
\]

\[
\leq Tp_1 + \sum_{t=1}^{T} E_{H_{t-1}} \left[ p_2 + (c_1 + c_2) \left( 1 + \frac{2}{p_3 - p_2} \right) E_t[\min\{1, \|X_t\|_{M_t^{-1}}\} | H_{t-1}] \right]
\]

\[
= Tp_1 + \sum_{t=1}^{T} E \left[ p_2 + (c_1 + c_2) \left( 1 + \frac{2}{p_3 - p_2} \right) \min\{1, \|X_t\|_{M_t^{-1}}\} \right]
\]

\[
= Tp_1 + Tp_2 + (c_1 + c_2) \left( 1 + \frac{2}{p_3 - p_2} \right) E \sum_{t=1}^{T} \min\{1, \|X_t\|_{M_t^{-1}}\}
\]

\[
\leq T(p_1 + p_2) + (c_1 + c_2) \left( 1 + \frac{2}{p_3 - p_2} \right) \sqrt{c_3 T},
\]

completing the proof of the theorem.

We next prove Lemma 1.

**Proof of Lemma 1.** We fix a round \( t \) and an arbitrary history \( H_{t-1} \) satisfying \( E^{ls} \), and omit the conditioning on \( H_{t-1} \) henceforth. Let us denote \( N := M_t^{-1} \). Let \( c := c_1 + c_2 \) and define

\[
S_t := \{ i \in A : c \|x_i\|_N < \Delta_i \} \text{ and } \overline{S}_t = A \setminus S_t.
\]

Observe that \( S_t \) is deterministic (since we have fixed the history) and that \( 1 \notin S_t \). The arms in \( S_t \) are sufficiently
sampled, and the rest of the arms are undersampled. Also, let
\[ j_t := \arg\min_{i \notin S_t} \|x_i\|_N \]
be the least uncertain undersampled arm, which is deterministic given the history. We may write
\[
E[\Delta_{i_t}] = E[\Delta_{i_t} I\{E_t^{\text{conc}}\}] + E[\Delta_{i_t} I\{E_t^{\text{anti}}\}]
\leq E[\Delta_{i_t} I\{E_t^{\text{conc}}\}] + P(E_t^{\text{conc}})
\leq E[\Delta_{i_t} I\{E_t^{\text{conc}}\}] + p_2. \tag{9}
\]
Recall that \(E_t^{\text{conc}}\) is the event that for all arms \(i\), \(|f_t(x_i) - \langle x_i, \hat{\theta}_t \rangle| \leq c_2\|x_i\|_N\), and \(E_t^{\text{ls}}\) is the event that for all arms \(i\), \(|\langle x_i, \hat{\theta}_t - \theta^* \rangle| \leq c_1\|x_i\|_N\). Hence, on event \(E_t^{\text{conc}} \cap E_t^{\text{ls}}\) we have
\[
\tilde{f}_t(x_{i_t}) \leq c_2\|x_{i_t}\|_N + \langle x_{i_t}, \hat{\theta}_t \rangle \leq c_2\|x_{i_t}\|_N + c_1\|x_{i_t}\|_N + \langle x_{i_t}, \theta^* \rangle,
\]
and similarly,
\[
\tilde{f}_t(x_j) \geq \langle x_j, \theta^* \rangle - c\|x_j\|_N,
\]
which, since \(\tilde{f}_t(x_{i_t}) \geq \tilde{f}_t(x_{j_t})\), gives
\[
\Delta_{i_t} = \Delta_{j_t} + \langle x_{j_t} - x_{i_t}, \theta^* \rangle \leq c\|x_{j_t}\|_N + (c\|x_{i_t}\|_N + c\|x_{j_t}\|_N) = c\|x_{i_t}\|_N + 2\|x_{j_t}\|_N
\]
deterministically. On the other hand, since \(\Delta_{i_t} \leq 1 \leq c\), we also have
\[
\Delta_{i_t} \leq c((1 \wedge \|x_{i_t}\|_N) + 2(1 \wedge \|x_{j_t}\|_N)),
\]
where \(a \wedge b := \min\{a, b\}\). Plugging into Eq. (9), we obtain
\[
E[\Delta_{i_t}] \leq c(E(1 \wedge \|x_{i_t}\|_N) + 2(1 \wedge \|x_{j_t}\|_N)) + p_2. \tag{10}
\]
The next step is to bound \(1 \wedge \|x_{j_t}\|_N\) from above. Observe that,
\[
E(1 \wedge \|x_{i_t}\|_N) \geq E[(1 \wedge \|x_{i_t}\|_N) \mid i_t \in \overline{S}_t] P(i_t \in \overline{S}_t) \geq (1 \wedge \|x_{j_t}\|_N)P(i_t \in \overline{S}_t),
\]
where the last inequality is by the definition of \(j_t\). Rearranging gives
\[
1 \wedge \|x_{j_t}\|_N \leq E(1 \wedge \|x_{i_t}\|_N)/P(i_t \in \overline{S}_t).
\]
Next, we bound \(P(i_t \in \overline{S}_t)\) from below. By definition of \(i_t\) and since \(1 \in \overline{S}_t\), we have
\[
P(i_t \in \overline{S}_t) \geq P\left(\tilde{f}_t(x_1) > \max_{j \in S_t} \tilde{f}_t(x_j)\right) \geq P\left(\tilde{f}_t(x_1) > \max_{j \in S_t} \tilde{f}_t(x_j) \text{ and } E^{\text{conc}} \text{ holds}\right).
\]
If \(E^{\text{conc}}\) holds, then for any \(j \in S_t\) we have
\[
\tilde{f}_t(x_j) \leq \langle x_j, \theta^* \rangle + c\|x_j\|_N < \langle x_j, \theta^* \rangle + \Delta_{j_t} = \langle x_{j_t}, \theta^* \rangle,
\]
whence,
\[
P\left(\tilde{f}_t(x_1) > \max_{j \in S_t} \tilde{f}_t(x_j) \text{ and } E^{\text{conc}} \text{ holds}\right) \geq P\left(\tilde{f}_t(x_1) > \langle x_{j_t}, \theta^* \rangle \text{ and } E^{\text{conc}} \text{ holds}\right)
\geq P\left(\tilde{f}_t(x_1) > \langle x_{j_t}, \theta^* \rangle\right) - P\left(E^{\text{conc}}\right).
\]
Finally, note that if \(E^{\text{anti}} \cap E^{\text{ls}}\) holds, then
\[
\tilde{f}_t(x_1) > \langle x_{j_t}, \hat{\theta}_t \rangle + c_1\|x_1\|_N \geq \langle x_1, \theta^* \rangle,
\]
and thus
\[ P\left( f_t(x_1) > \langle x_1, \theta^* \rangle \right) - P\left( E_{\text{conc}} \right) \geq p_3 - p_2. \]

Hence, we find \( P\left( i_t \in \overline{S}_t \right) \geq p_3 - p_2 \), and plugging this back into Eq. (10) gives
\[ E[\Delta_{i_t}] \leq c(E(1 \wedge \|x_i\|_N) + 2(1 \wedge \|x_{ji}\|_N)) + p_2 \leq c(E(1 \wedge \|x_i\|_N) + \frac{2E(1 \wedge \|x_i\|_N)}{p_3 - p_2}) + p_2, \]
completing the proof of the lemma.

C Regret bounds for general index-based algorithms for generalized linear bandits

In this section, we prove Theorem 4. Recall that we denote by \( i_t \) the arm pulled in round \( t \). Let \( X_t := x_{i_t} \) and let \( Y_t \) denote the reward received in round \( t \). We define \( \hat{\theta}_t \) as the MLE for the generalized linear model (GLM) at round \( t \). We denote the Hessian matrix by \( H_t := \sum_{t=1}^{t-1} g'((X_t, \hat{\theta}_t))X_tX_t^T \). We also define the Gram matrix \( M_t := \sum_{t=1}^{T-1} X_tX_t^T \). For matrices \( A \) and \( B \), by \( A \succeq B \) we mean that \( A - B \) is positive semidefinite, and by \( A \succ B \) we mean that \( A - B \) is positive definite. Recall that \( \{v_j\}_{j=1}^d = \{x_i\}_{i \in A} \) forms a basis such that \( \sum_{j=1}^d v_jv_j^T \geq \rho I \).

Similar to the linear bandit case, we first obtain an analogous result as Theorem 1 from Kveton et al. [2019d] for any index-based algorithm that does some initialization, then in each subsequent round \( t \), selects the arm \( i_t = \arg \max_a f_t(x_i) \), where \( f_t \) is a stochastic function depending on the history up to round \( t \), and possibly additional independent randomness. For RandUCB, the initialization is pulling the \( v_j \) vectors sufficiently many times and we will set \( f_t(x) := \langle \hat{\theta}_t, x \rangle + Z_t \|x\|_{M_t^{-1}} \), where \( Z_1, \ldots, Z_T \) are i.i.d. random variables. Recall that \( L \) is the Lipschitz constant of \( g \), and that
\[ 0 < \mu = \inf \{ g'((x, \theta)): \|x\| \leq 1, \|\theta - \theta^*\| \leq 1 \}. \]

Theorem 8 (Generic regret bound for index-based algorithms for generalized linear bandits). Let \( c_3 := 2d\ln \left( 1 + \frac{T}{\tau} \right) \). Suppose \( c_1 < c_2 \) and \( \tau \) are real numbers and define the events
\[ E_{\text{mle}} := \left\{ \forall i \in A, \forall \tau < t: |\langle x_i, \hat{\theta}_t - \theta^* \rangle| \leq c_1 \|x_i\|_{M_t^{-1}} \right\}, \]
\[ E_{\text{bound}} := \left\{ \forall \tau < t: \|\hat{\theta}_t - \theta^*\| \leq 1 \text{ and } M_t \succeq I \text{ and } H_t \succ 0 \right\}, \]
\[ E_{\text{conc}} := \left\{ \forall i \in A: |\tilde{f}_t(x_i) - \langle x_i, \tilde{\theta}_t \rangle| \leq c_2 \|x_i\|_{H_t^{-1}} \right\}, \]
\[ E_{\text{anti}} := \left\{ \tilde{f}_t(x_1) - \langle x_1, \tilde{\theta}_t \rangle > \sqrt{\tau} c_1 \|x_1\|_{H_t^{-1}} \right\}. \]

Suppose \( p_1, p_2, p_3, p_4 \in [0, 1] \) are such that
\[ P(E_{\text{mle}}) \geq 1 - p_1, P(E_{\text{bound}}) \geq 1 - p_4, \text{ and that for any given } t > \tau \]
and for any possible history \( \mathcal{H}_{t-1} \) before the start of round \( t \), we have
\[ P(E_{\text{anti}} | \mathcal{H}_{t-1}) \geq p_3, \text{ and } P(E_{\text{conc}} | \mathcal{H}_{t-1}) \geq 1 - p_2. \]

Then, the regret after \( T \) rounds is bounded by
\[ R(T) \leq L \cdot \left( c_1 + \frac{c_2}{\sqrt{\rho}} \right) \left( 1 + \frac{2}{p_3 - p_2} \right) \sqrt{c_3T} + (p_1 + p_2 + p_4) T + \tau. \]

Before proving this theorem, we show how it implies Theorem 4.

Proof of Theorem 4. We first describe the initialization.

Define \( \tau_0 := \max \left\{ \frac{d \log(T/d) + 2 \log(T)}{\rho^2 \rho}, 1/\rho \right\} \). First, we pull each \( v_j \) for \( \tau_0 \) many times. By then, the smallest eigenvalue of the Gram matrix \( M_t \) becomes at least \( \rho \tau_0 = \max \left\{ \frac{d \log(T/d) + 2 \log(T)}{\rho^2 \rho}, 1 \right\} \), so by the arguments in Lemma 8 from Kveton et al. [2019d] and Theorem 1 from Li et al. [2017], \( \|\hat{\theta}_t - \theta^*\| \leq 1 \) in each subsequent round,
with probability at least $1 - 1/T$. In particular, by definition of $\mu$, this implies that in each subsequent round $t$ we have $g'\left(\left(X_t, \hat{\theta}_t\right)\right) \geq \mu$. We pull each $v_i$ one more time. After these pulls, we have $H_t \geq \mu \sum v_i v_i^T \geq \mu f > 0$. Therefore, with $\tau = d + \max \left\{ \frac{d^2 \log(T/d)}{\mu^2} + 2d \log(T), d/\rho \right\}$ initial rounds, the event $E_{\text{bound}}$ holds with probability at least $1 - 1/T = 1 - p_4$.

Note that for RandUCB, we have $f_t(x) = \langle \hat{\theta}_t, x \rangle + Z_t \|x\|_{M_{t-1}}$, so by definition, $P\left(E_{t}^{\text{anti}}\right) = P\left(Z_t > \sqrt{C_1}\right) =: p_3$ and $P\left(E_{t}^{\text{conc}}\right) = P\left(\|Z_t\| > c_2\right) =: p_2$. Moreover, by Lemma 5 from Kveton et al. [2019d] we have $P\left(E_{\text{mle}}\right) \geq 1 - 1/T = 1 - p_4$. These bounds combined with Theorem 8 give Theorem 4.

In the rest of this section we prove Theorem 8. Let use denote by $E_t, P_t$ the randomness injected by the algorithm in round $t$ (i.e., the randomness of $Z_t$), and denote by $E_{\mathcal{H}_{t-1}}, P_{\mathcal{H}_{t-1}}$ the randomness in the history (up to round $t - 1$). Let $\Delta_t = g(x_t, \theta^*) - g(x_t, \hat{\theta}_t)$ denote the gap of the arm $i$ under the generalized linear model. The following lemma is an analogue of Lemma 2 from Kveton et al. [2019d].

**Lemma 2.** For any round $t > \tau$ and any history $\mathcal{H}_{t-1}$, we have

$$E_t \left[ \Delta_t, I \left\{ E_{\text{mle}}, E_{\text{bound}} \right\} | \mathcal{H}_{t-1} \right] \leq p_2 + \mathcal{L} \left( c_1 + \frac{c_2}{\sqrt{\mu}} \right) \left( 1 + \frac{2}{p_3 - p_2} \right) E_t \left[ \min \{ 1, \|X_t\|_{M_{t-1}} \} | \mathcal{H}_{t-1} \right].$$

Before proving this lemma we show that it implies Theorem 8.

**Proof of Theorem 8.** On the event $E_{\text{bound}}$, for $t > \tau$ we have $M_t \geq I$, hence by (8) and the Cauchy-Schwarz inequality, we have, deterministically,

$$\sum_{t = \tau + 1}^{T} \min \{ 1, \|X_t\|_{M_{t-1}} \} \leq \sqrt{c_3 T}.$$

Thus, since $\Delta_i \leq 1$ for all $i$, and using Lemma 2 we have,

$$R(T) = \sum_{t=1}^{T} E_t [\Delta_i]$$

$$\leq \tau + \sum_{t = \tau + 1}^{T} E_t [\Delta_i]$$

$$\leq \tau + T \mathcal{P}(E_{\text{mle}}) + T \mathcal{P}(E_{\text{bound}}) + \sum_{t = \tau + 1}^{T} E_t [\Delta_i, I \left\{ E_{\text{mle}}, E_{\text{bound}} \right\} | \mathcal{H}_{t-1} |]$$

$$= \tau + T (p_1 + p_4) + \sum_{t = \tau + 1}^{T} E_t [\Delta_i, I \left\{ E_{\text{mle}}, E_{\text{bound}} \right\} | \mathcal{H}_{t-1} |]$$

$$\leq T (p_1 + p_4) + \tau + T \sum_{t = \tau + 1}^{T} E_t [\Delta_i, I \left\{ E_{\text{mle}}, E_{\text{bound}} \right\} | \mathcal{H}_{t-1} |]$$

$$\leq T (p_1 + p_4) + T \sum_{t = \tau + 1}^{T} \left( p_2 + \mathcal{L} \left( c_1 + \frac{c_2}{\sqrt{\mu}} \right) \left( 1 + \frac{2}{p_3 - p_2} \right) E_t \left[ \min \{ 1, \|X_t\|_{M_{t-1}} \} | \mathcal{H}_{t-1} \right] \right)$$

$$\leq T (p_1 + p_4) + T \sum_{t = \tau + 1}^{T} \left( c_1 + \frac{c_2}{\sqrt{\mu}} \right) \left( 1 + \frac{2}{p_3 - p_2} \right) \sqrt{c_3 T}$$

completing the proof of the theorem.
We finally prove Lemma 2.

**Proof of Lemma 2.** Fix a round $t$ and an arbitrary history $\mathcal{H}_{t-1}$ satisfying $E^\text{mle}$ and $E^\text{bound}$, and omit the conditioning on $\mathcal{H}_{t-1}$ henceforth. On the event $E^\text{bound}$ we have $M_t \geq I$, and since $\|X_t\| \leq 1$ we have $\|X_t\|_{M^{-1}_t} \leq 1$ deterministically, so we need only show that

$$E |\Delta_i, I \{E^\text{mle}, E^\text{bound}\}| \leq p_2 + \mathcal{L} \left( c_1 + \frac{c_2}{\sqrt{p}} \right) \left( 1 + \frac{2}{p_3 - p_2} \right) E \|X_t\|_{M^{-1}_t}.$$  

Let us define

$$h(x) := c_1 \|x\|_{M^{-1}_t} + c_2 \|x\|_{H^{-1}_t},$$

$$\bar{\Delta}_i := \langle x_i - x_i, \theta^* \rangle,$$

$$S_t := \{ i \in A : h(x_i) < \bar{\Delta}_i \}$$

and since $1 \notin S_t$. The arms in $S_t$ are sufficiently sampled, and the rest of the arms are undersampled. Also, let

$$j_t := \arg \min_{i \notin S_t} h(x_i)$$

be the least uncertain undersampled arm, which is deterministic since we have fixed the history. We may then write

$$E[\Delta_i] = E[\Delta_i, I \{E^\text{conc}_i\}] + E[\Delta_i, I \{E^\text{conc}_i\}]$$

$$\leq E[\Delta_i, I \{E^\text{conc}_i\}] + \mathbb{P} \{ E^\text{conc}_i \}$$

$$\leq E[\Delta_i, I \{E^\text{conc}_i\}] + p_2$$

$$= E[\|g((x_1, \theta^*)) - g((x_i, \theta^*))\| \{E^\text{conc}_i\}] + p_2$$

$$\leq \mathcal{L} E[\|\tilde{\Delta}_i - (x_i, \theta^*)\| \{E^\text{conc}_i\}] + p_2$$

$$\implies E[\Delta_i] \leq \mathcal{L} E[\bar{\Delta}_i, I \{E^\text{conc}_i\}] + p_2. \quad (11)$$

Recall that $E^\text{conc}_i$ is the event that for all arms $i$, $\|\tilde{f}_i(x_i) - \langle x_i, \tilde{\theta}_t \rangle\| \leq c_1 \|x_i\|_{H^{-1}_t}$, and $E^\text{mle}$ is the event that for all arms $i$, $\|\tilde{f}_i(x_i) - \langle x_i, \tilde{\theta}_t \rangle\| \leq c_1 \|x_i\|_{M^{-1}_t}$. Hence, on event $E^\text{conc}_i \cap E^\text{mle}$ we have

$$\tilde{f}_i(x_i) \leq c_2 \|x_i\|_{H^{-1}_t} + \langle x_i, \tilde{\theta}_t \rangle \leq c_2 \|x_i\|_{H^{-1}_t} + c_1 \|x_i\|_{M^{-1}_t} + (x_i, \theta^*) \implies \tilde{f}_i(x_i) \leq h(x_i) + (x_i, \theta^*),$$

and, similarly,

$$\tilde{f}_i(x_j) \geq (x_j, \theta^*) - h(x_j),$$

which, since $\tilde{f}_i(x_i) \geq \tilde{f}_i(x_j)$, gives

$$\bar{\Delta}_i = \bar{\Delta}_j + \langle x_i - x_i, \theta^* \rangle \leq h(x_j) + h(x_i) + h(x_j) = h(x_i) + 2 h(x_j)$$

deterministically. Plugging into (11), we obtain

$$E[\Delta_i] \leq \mathcal{L} \left( E h(x_i) + 2h(x_j) \right) + p_2. \quad (12)$$

The next step is to bound $h(x_j)$ from above. Observe that

$$E h(x_i) \geq E \left[ h(x_i) | i_t \in \bar{S}_t \right] \mathbb{P} \left( i_t \in \bar{S}_t \right) \geq h(x_j) \mathbb{P} \left( i_t \in \bar{S}_t \right),$$

where the last inequality is by the definition of $j_t$. Rearranging gives

$$h(x_j) \leq E h(x_i) / \mathbb{P} \left( i_t \in \bar{S}_t \right).$$
Next, we bound \( P \{ i_t \in S_t \} \) from below. By definition of \( i_t \) and since \( 1 \in S_t \), we have

\[
P \{ i_t \in S_t \} \geq P \left( \tilde{f}_t(x_t) > \max_{j \in S_t} \tilde{f}_t(x_j) \right) \geq P \left( \tilde{f}_t(x_t) > \max_{j \in S_t} \tilde{f}_t(x_j) \text{ and } E^\text{conc} \text{ holds} \right).
\]

If \( E^\text{conc} \text{ holds} \), then for any \( j \in S_t \) we have

\[
\tilde{f}_t(x_j) \leq \langle x_j, \theta^* \rangle + h(x_j) < \langle x_j, \theta^* \rangle + \Delta_j = \langle x_t, \theta^* \rangle,
\]

whence,

\[
P \left( \tilde{f}_t(x_t) > \max_{j \in S_t} \tilde{f}_t(x_j) \text{ and } E^\text{conc} \text{ holds} \right) \geq P \left( \tilde{f}_t(x_t) > \langle x_t, \theta^* \rangle \text{ and } E^\text{conc} \text{ holds} \right)
\]

\[
\geq P \left( \tilde{f}_t(x_t) > \langle x_t, \theta^* \rangle \right) - P \left( E^\text{conc} \right).
\]

Since \( E^\text{mle} \) holds,

\[
P \left( \tilde{f}_t(x_t) > \langle x_t, \theta^* \rangle \right) \geq P \left( \tilde{f}_t(x_t) - \langle x_t, \hat{\theta}_t \rangle > c_1 \| x_1 \|_{M_t^{-1}} \right).
\]

By Lemma 3 below we have \( \| x_1 \|_{M_t^{-1}} \leq \sqrt{\mathcal{L}} \| x_1 \|_{H_t^{-1}} \), implying

\[
P \left( \tilde{f}_t(x_t) - \langle x_t, \hat{\theta}_t \rangle > c_1 \| x_1 \|_{M_t^{-1}} \right) \geq P \left( \tilde{f}_t(x_t) - \langle x_t, \hat{\theta}_t \rangle > c_1 \sqrt{\mathcal{L}} \| x_1 \|_{H_t^{-1}} \right) = P \left( E^\text{anti} \right) \geq p_3,
\]

whence,

\[
P \left( \tilde{f}_t(x_t) > \langle x_t, \theta^* \rangle \right) - P \left( E^\text{conc} \right) \geq p_3 - p_2.
\]

Hence, we find \( P \{ i_t \in S_t \} \geq p_3 - p_2 \), and plugging this back into (12) gives

\[
\mathbb{E}[\Delta_{i_t}] \leq \mathcal{L} \left[ \mathbb{E} h(x_{i_t}) + \frac{2 \mathbb{E} h(x_{i_t})}{p_3 - p_2} \right] + p_2 = \mathcal{L} \left[ 1 + \frac{2}{p_3 - p_2} \right] \mathbb{E} h(x_{i_t}) + p_2.
\]

We now bound the quantity \( h(x_{i_t}) \). By Lemma 3 below we have

\[
\| X_t \|_{H_t^{-1}} \leq \frac{\| X_t \|_{M_t^{-1}}}{\sqrt{\mu}},
\]

thus,

\[
h(X_t) = c_1 \| X_t \|_{M_t^{-1}} + c_2 \| X_t \|_{H_t^{-1}} \leq \left( c_1 + \frac{c_2}{\sqrt{\mu}} \right) \| X_t \|_{M_t^{-1}}.
\]

Putting everything together, we find

\[
\mathbb{E}[\Delta_{i_t}] \leq \mathcal{L} \left[ 1 + \frac{2}{p_3 - p_2} \right] \left( c_1 + \frac{c_2}{\sqrt{\mu}} \right) \mathbb{E} \| X_t \|_{M_t^{-1}} + p_2,
\]

completing the proof of Lemma 2.

\[ \square \]

**Lemma 3.** Let \( V_1, \ldots, V_t \) be positive semi-definite matrices, and let \( \mu \leq b_1, \ldots, b_t \leq \mathcal{L} \) be real numbers such that both \( A := \sum_{s=1}^t V_s \) and \( B := \sum_{s=1}^t b_s V_s \) are positive definite. Then, for any vector \( x \) we have

\[
\sqrt{\mu} \| x \|_{B^{-1}} \leq \| x \|_{A^{-1}} \leq \sqrt{\mathcal{L}} \| x \|_{B^{-1}}.
\]

**Proof.** Note that \( \mathcal{L} A = \sum_{s=1}^t \mathcal{L} V_s \geq \sum_{s=1}^t b_s V_s = B \). Invertibility of the PSD ordering implies \( A^{-1}/\mathcal{L} \preceq B^{-1} \), whence for any vector \( x \) we have \( x^T A^{-1} x \leq \mathcal{L} x^T B^{-1} x \), which gives the inequality on the right. The left inequality is proven symmetrically, by choosing \( V'_t := b_t V_t \) and \( b'_t := 1/b_t \). \[ \square \]
D Additional experiments

All algorithms are compared based on their cumulative empirical regret, defined as

$$R(T) := \sum_{t=1}^{T} (Y_t^* - Y_t),$$

where $Y_t^*$ denotes the reward of the optimal arm in round $t$.

In all experiments, algorithms are run over $T = 20,000$ rounds on 50 randomly generated instances (the generated instances are the same for all algorithms).

D.1 Ablation study

We first investigate the impact of the RandUCB design choices and parameters on the MAB settings with, unless specified, $L = 0$, $U = 2\sqrt{\ln(T)}$, and $M = 20$.

Recall that RandUCB is characterized by the choice of sampling distribution (Section 3.1). We compare the performance of RandUCB using uniform and Gaussian sampling ($\varepsilon = 10^{-7}$, $\sigma \in \{1/16, 1/8, 1\}$) distributions. Figure 2 shows that increasing $\sigma$ brings us closer to the uniform distribution.

We then compare the default (optimistic, with coupled arms) RandUCB with non-optimistic and uncoupled variants. All use Gaussian sampling ($\varepsilon = 10^{-7}$, $\sigma = 1/8$). Non-optimistic considers $L = -2\sqrt{\ln(T)}$ and $M = 40$. Figure 3 shows that coupling the arms is more determinant in the performance of RandUCB compared with optimism. This is not surprising as the same happens for TS (Appendix D.2).

We also evaluate the impact of the support size of the sampling distribution $M$. We compare RandUCB using Gaussian sampling distribution ($\varepsilon = 10^{-7}$, $\sigma = 1/8$), $M = 20$, optimistic, and coupled arms, against alternatives with $M = 5$ and $M = 100$. Figure 4 shows that RandUCB is robust to the discretization induced by the support size.

D.2 Optimistic Thompson Sampling

In the last experiment, we empirically show that Optimistic TS is almost equivalent to TS. To this end, we compare both variants on the MAB setting. Figure 5 confirms that their performance is similar.
D.3 Gaussian MAB

We use the same number of arms $K = 100$. The rewards are generated from a $N(\mu, 0.1)$ distribution. For generating the mean rewards, we consider both the easy and hard settings as before. Both Gaussian TS and

UCB1 are theoretically optimal in this setting. From Fig. 6, we observe that similar to the other settings, the performance of RandUCB (with the same hyperparameter settings) is better than both TS and UCB1.