Old Dog Learns New Tricks: Randomized UCB for Bandit Problems

Sharan Vaswani
Mila, Université de Montréal

Abbas Mehrabian
McGill University

Audrey Durand
Mila, Université Laval

Branislav Kveton
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 $\widetilde{O}(\sqrt{KT})$ 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 $\widetilde{O}(d\sqrt{T})$ 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 ε-*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-

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

cally optimal in many 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 $\widetilde{O}(\sqrt{KT})$ 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 $\widetilde{O}(d\sqrt{T})$ 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 μ_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 $\widehat{\mu}_i(t)$, defined as the average of rewards received from arm i until round t. The algorithm also maintains a confidence interval of size $\mathcal{C}_i(t)$ around the estimated mean. The value of $\mathcal{C}_i(t)$ decreases as an arm is pulled more, and indicates how accurate $\widehat{\mu}_i(t)$ is at estimating μ_i. Although the exact values of $\widehat{\mu}_i(t)$ and $\mathcal{C}_i(t)$ 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

$$i_t = \arg\max_{i \in \mathcal{A}} \left\{ \widehat{\mu}_i(t) + \beta\, \mathcal{C}_i(t) \right\}. \tag{1}$$

The parameter β 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

$$i_t = \arg\max_{i \in \mathcal{A}} \left\{ \widehat{\mu}_i(t) + Z_t\, \mathcal{C}_i(t) \right\}, \tag{2}$$

where the deterministic quantity β 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 $p_m := \mathbf{P}(Z = \alpha_m)$. 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 α_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 $\mathbf{P}(Z \geq \beta) > 0$. 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 $\varepsilon, \sigma > 0$. The former is the constant probability to be put on the highest point: $\alpha_M = U$ with $p_M = \varepsilon$. For the remaining $M - 1$ points, we use a discretized Gaussian distribution; formally, for $1 \leq m \leq M-1$, let $\overline{p_m} := \exp(-\alpha_m^2/2\sigma^2)$ and let p_m denote the normalized probabilities, that is, $p_m := (1-\varepsilon)\,\overline{p_m}/(\sum_m \overline{p_m})$. 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.

sider *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 $i_t = \arg\max_i \{\widehat{\mu}_i(t) + Z_{i,t}\, \mathcal{C}_i(t)\}$. 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 ε and σ. 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, ε, and σ will be specified for each experiment.

4 Multi-Armed Bandit

In this section, we consider a stochastic multi-armed bandit (MAB) with $|\mathcal{A}| = K$ 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_t in round t, then it incurs an expected (cumulative) regret of $R(T) \coloneqq \sum_{t=1}^{T} \mathbf{E}[\mu_1 - \mu_{i_t}] = \sum_{t=1}^{T} \mathbf{E}[\Delta_{i_t}]$.

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 $\widehat{\mu}_i(t) = Y_i(t)/s_i(t)$ (we set $\widehat{\mu}_i(t) = 0$ if arm i has never been pulled). The confidence interval corresponds to the standard deviation in the estimation of μ_i and is given as $\mathcal{C}_i(t) = \sqrt{\frac{1}{s_i(t)}}$. To ensure that $s_i(t) > 0$, RandUCB begins by pulling each arm once and in each subsequent round $t > K$, selects

$$i_t = \arg\max_i \left\{ \widehat{\mu}_i(t) + Z_t \sqrt{\frac{1}{s_i(t)}} \right\}. \tag{3}$$

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 as TS (Appendix D.2). We show that RandUCB with $M \to \infty$ approaches OTS with a Gaussian prior and posterior. First, observe that uncoupled 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 ε 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 ε-greedy strategy. Recall that the classical ε-greedy (EG) strategy [Auer et al., 2002; Langford and Zhang, 2008] chooses a random action with probability ε and the greedy action with probability $1 - \varepsilon$. For a constant ε, EG might result in linear regret, whereas decreasing ε over time results in a sub-optimal $O(T^{2/3})$ regret [Auer et al., 2002]. An *adaptive ε-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 ε. 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 ε. Theorem 1 below implies that the regret of this modification of the ε-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* $c_1 \coloneqq 1 + \sqrt{\ln(KT^2)}$ *and* $c_3 \coloneqq 2K\ln\left(1 + \frac{T}{K}\right)$. *For any* $c_2 > c_1$, *the regret* $R(T)$ *of* RandUCB *for MAB is bounded by*

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

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)** $\mathbf{P}\left(Z > 1 + \sqrt{\ln(KT^2)}\right) > 0$ 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 ε at $U = 2\sqrt{\ln(T)}$ by design, it ensures that $\mathbf{P}\left(Z > c_1\right)$ 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* $O\left(\sum_{\Delta_i > 0} \Delta_i^{-1}\right) \times \left(\frac{M}{p_M} + Te^{-2\alpha_M^2} + \alpha_M^2\right)$.

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 $O\left(\ln T \times \left(\sum \Delta_i^{-1}\right)\right)$ (see, e.g., Lattimore and Szepesvári [2020, Theorem 16.4]). By a standard reduction, Theorem 2 implies that uncoupled `RandUCB` achieves the problem-independent $\widetilde{O}(\sqrt{KT})$ 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 $i \in \mathcal{A}$. 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 $\theta^\star$ with $\|\theta^\star\| \leq 1$, and that the rewards lie in $[0, 1]$.[5] We first consider the linear feature-reward mapping.

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_t is the reward obtained in round t, then $\mathbf{E}[Y_t | i_t = i] = \langle x_i, \theta^\star \rangle$. 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 $\Delta_i = \langle x_1 - x_i, \theta^\star \rangle$,

$$R(T) := \sum_{t=1}^{T} \mathbf{E}\left[\langle x_1 - x_{i_t}, \theta^\star \rangle\right] = \sum_{t=1}^{T} \mathbf{E}[\Delta_{i_t}]. \qquad (4)$$

Let us denote $X_t := x_{i_t}$ and define the Gram matrix $M_t := \lambda I_d + \sum_{\ell=1}^{t-1} X_\ell X_\ell^\mathsf{T}$. Here, $\lambda > 0$ is the ℓ_2 *regularization parameter*. We define the norm $\|x\|_M := \sqrt{x^\mathsf{T} M x}$ for any positive definite M.

5.1.1 Instantiating `RandUCB`

Given the observations $(X_\ell, Y_\ell)_{\ell=1}^{t-1}$ gathered until round t, the maximum likelihood estimator (MLE) for linear regression is $\widehat{\theta}_t := M_t^{-1} \sum_{\ell=1}^{t-1} Y_\ell X_\ell$. The estimated mean for the reward of arm i is $\widehat{\mu}_i(t) = \langle \widehat{\theta}_t, x_i \rangle$ and the corresponding confidence interval is $\mathcal{C}_i(t) = \|x_i\|_{M_t^{-1}}$. Thus, `RandUCB` chooses arm

$$i_t := \arg\max_{i \in \mathcal{A}} \left\{ \langle \widehat{\theta}_t, x_i \rangle + Z_t \|x_i\|_{M_t^{-1}} \right\}. \qquad (5)$$

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

5.1.2 Regret of `RandUCB` for linear bandits

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

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

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

$$E^{\mathrm{ls}} := \left\{ \forall i \in [K], \forall t \in [T]; \quad |\langle x_i, \widehat{\theta}_t - \theta^\star \rangle| \leq c_1 \|x_i\|_{M_t^{-1}} \right\},$$

$$E_t^{\mathrm{conc}} := \left\{ \forall i \in [K]; \quad |\widetilde{f}_t(x_i) - \langle x_i, \widehat{\theta}_t \rangle| \leq c_2 \|x_i\|_{M_t^{-1}} \right\},$$

$$E_t^{\mathrm{anti}} := \left\{ \widetilde{f}_t(x_1) - \langle x_1, \widehat{\theta}_t \rangle > c_1 \|x_1\|_{M_t^{-1}} \right\},$$

and assume for now that we have the following bounds for their probabilities: $\mathbf{P}\left(E^{\mathrm{ls}}\right) \geq 1 - p_1$, $\mathbf{P}(E_t^{\mathrm{conc}}) \geq 1 - p_2$, and $\mathbf{P}(E_t^{\mathrm{anti}}) \geq p_3$.

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 $\widetilde{f}_t(x)$, i.e., $i_t = \arg\max_i \widetilde{f}_t(x_i)$. Theorem 7 in Appendix B

[5] It is easy to generalize our results to reward distributions bounded in any known interval. Similarly, the analyses can be adapted to handle sub-gaussian distributions.

bounds the regret of such an algorithm by

$$(c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right)\sqrt{c_3 T} + T\,(p_1 + p_2). \qquad (6)$$

For `RandUCB`, we have $\widetilde{f}_t(x) = \langle \widehat{\theta}_t, x \rangle + Z_t \, \|x\|_{M_t^{-1}}$. Event E^{ls} 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 $\widetilde{f}_t$, $\mathbf{P}\left(E_t^{\text{anti}}\right) = \mathbf{P}\left(Z_t > c_1\right) =: p_3$ and $\mathbf{P}\left(\overline{E_t^{\text{conc}}}\right) = \mathbf{P}\left(|Z_t| > c_2\right) =: p_2$. These relations combined with the bound Eq. (6) concludes the proof. □

Similar to the MAB case, we choose $U = 3c_1$ for the default variant of `RandUCB`. This ensures $\mathbf{P}\left(Z > c_1\right)$ is a positive constant and $\mathbf{P}\left(|Z| > c_2\right) = 0$, resulting in the promised $\widetilde{O}(d\sqrt{T})$ regret bound. We reiterate that this bound does not have a dependence on K, and thus holds in the infinite-arms case.

5.1.3 Proof for regret bound for MAB

We now present the proof of Theorem 1 by using a reduction from linear bandits to multi-armed bandits. Consider a linear bandit (LB) with dimension $d = K$, where arm features correspond to standard basis vectors, i.e., x_i is a one-hot vector with the ith component set to 1, and the true parameter vector is $\theta^\star = (\mu(1), \ldots, \mu(K))$. Now consider using `RandUCB` with $\lambda = 1$ for this problem.

We claim that `RandUCB` for MAB (Eq. (3)) selects the same action on round $t > K$ as `RandUCB` for LB (Eq. (5)) on round $t - K$. In fact, for any $t > K$, let $s_i(t)$ denote the number of times arm i has been pulled during rounds $K+1, \ldots, t$, and so M_t^{-1} is a diagonal matrix with the ith diagonal entry $(s_i(t) + 1)^{-1}$. `RandUCB` for MAB in round t pulls $i_t = \arg\max_i \left\{\widehat{\mu}_t(i) + Z_t/\sqrt{s_i(t) + 1}\right\}$, while `RandUCB` for LB in round t pulls $i_t = \arg\max_i \{\left\langle x_i, \widehat{\theta}_t \right\rangle + Z_t \|x_i\|_{M_t^{-1}}\}$. Observe that these expressions are identical, hence `RandUCB` for MAB exactly corresponds to `RandUCB` for LB. So, Theorem 7 in Appendix B applies with $\widetilde{f}_t(x) := \left\langle x, \widehat{\theta}_t \right\rangle + Z_t \|x\|_{M_t^{-1}}$.

We next bound the probabilities p_1, p_2, and p_3 define in Theorem 7. By Hoeffding's inequality, for each arm i and round t:

$$\begin{aligned}
&\mathbf{P}\left(\left|\left\langle x_i, \widehat{\theta}_t - \theta^\star \right\rangle\right| > c_1 \|x_i\|_{M_t^{-1}}\right) \\
&= \mathbf{P}\left(|\widehat{\mu}_i(t) - \mu_i| > \frac{c_1}{\sqrt{1 + s_i(t)}}\right) < 1/KT^2.
\end{aligned}$$

Thus by a union bound over the arms and the rounds, we get $\mathbf{P}\left(\overline{E^{\text{ls}}}\right) \leq 1/T =: p_1$. We can bound p_2 and p_3 by definition of $\widetilde{f}_t$: we have $\mathbf{P}\left(E_t^{\text{anti}}\right) = \mathbf{P}\left(Z > c_1\right) =: p_3$ and $\mathbf{P}\left(\overline{E_t^{\text{conc}}}\right) = \mathbf{P}\left(|Z| > c_2\right) =: p_2$. Using these bounds together with the bound in 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 $\mathbf{E}[Y_t | i_t = i] = g(\langle x_i, \theta^\star \rangle) \in [0, 1]$, 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, whereas if $g(x) = 1/(1 + \exp(-x))$, we get logistic bandits. Assuming arm 1 is optimal, the regret is $R(T) := \sum_{t=1}^{T} \mathbf{E}\left[g(x_1, \theta^\star) - g(x_{i_t}, \theta^\star)\right]$ and the *effective* gap of arm i is $\Delta_i := g(x_1, \theta^\star) - g(x_i, \theta^\star)$.

5.2.1 Instantiating `RandUCB`

As before, we denote $X_t = x_{i_t}$. Given previous observations $(X_\ell, Y_\ell)_{\ell=1}^{t-1}$, the MLE in round t can be computed as [McCullagh, 1984] $\widehat{\theta}_t := \arg\min_\theta \sum_{\ell=1}^{t-1} \left[Y_\ell \langle X_\ell, \theta \rangle - b(\langle X_\ell, \theta \rangle)\right]$, where b is a strictly convex function such that its derivative is g. Let $H_t(\theta) := \sum_{\ell=1}^{t-1} g'(\langle X_\ell, \theta \rangle) X_\ell X_\ell^\mathsf{T}$ denote the Hessian at point θ on round t, and $H_t := H_t(\widehat{\theta}_t)$. We assume that g is $\mathcal{L}$-Lipschitz, i.e., $|g(x) - g(y)| \leq \mathcal{L}|x - y|$, implying $0 < g'(x) \leq \mathcal{L}$ 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 $\{v_j\}_{j=1}^{d} \subseteq \{x_i\}_{i \in \mathcal{A}}$ with $\sum_{j=1}^{d} v_j v_j^T \succeq \rho I$ 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

$$\mu := \inf\{g'(\langle x, \theta \rangle) : \|x\| \leq 1, \|\theta - \theta^\star\| \leq 1\} > 0.$$

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 $\|\widehat{\theta}_t - \theta^\star\| \leq 1$ 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, $M_t = \sum_{\ell=1}^{t-1} X_\ell X_\ell^\mathsf{T}$).

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 $\beta = \frac{1}{\mu}\sqrt{\frac{d}{2}\ln(1+2T/d)+\ln(T)}$. Let $c_1 := \sqrt{d\ln(T/d)+2\ln(T)}/2\mu$, and choose $U = 2\sqrt{\mathcal{L}}\, c_1$ for RandUCB; the following theorem, proved in Appendix C, gives the promised $\widetilde{O}(d\sqrt{T})$ regret bound by choosing $c_2 = 3\sqrt{\mathcal{L}}\, c_1$.

Theorem 4. *Let $c_1 = \sqrt{d\ln(T/d)+2\ln(T)}/2\mu$, $c_3 := 2d\ln\left(1+\frac{T}{d}\right)$. For any $c_2 > c_1$, the regret $R(T)$ of* RandUCB *for generalized linear bandits is bounded by*

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

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 $\sum_{t=1}^{T}\left[Y_t^\star - Y_t\right]$, where $Y_t^\star$ 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 $|\mathcal{A}| = K = 100$ 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 σ), and observe in Figure 2 that lower values of σ 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 β. 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).

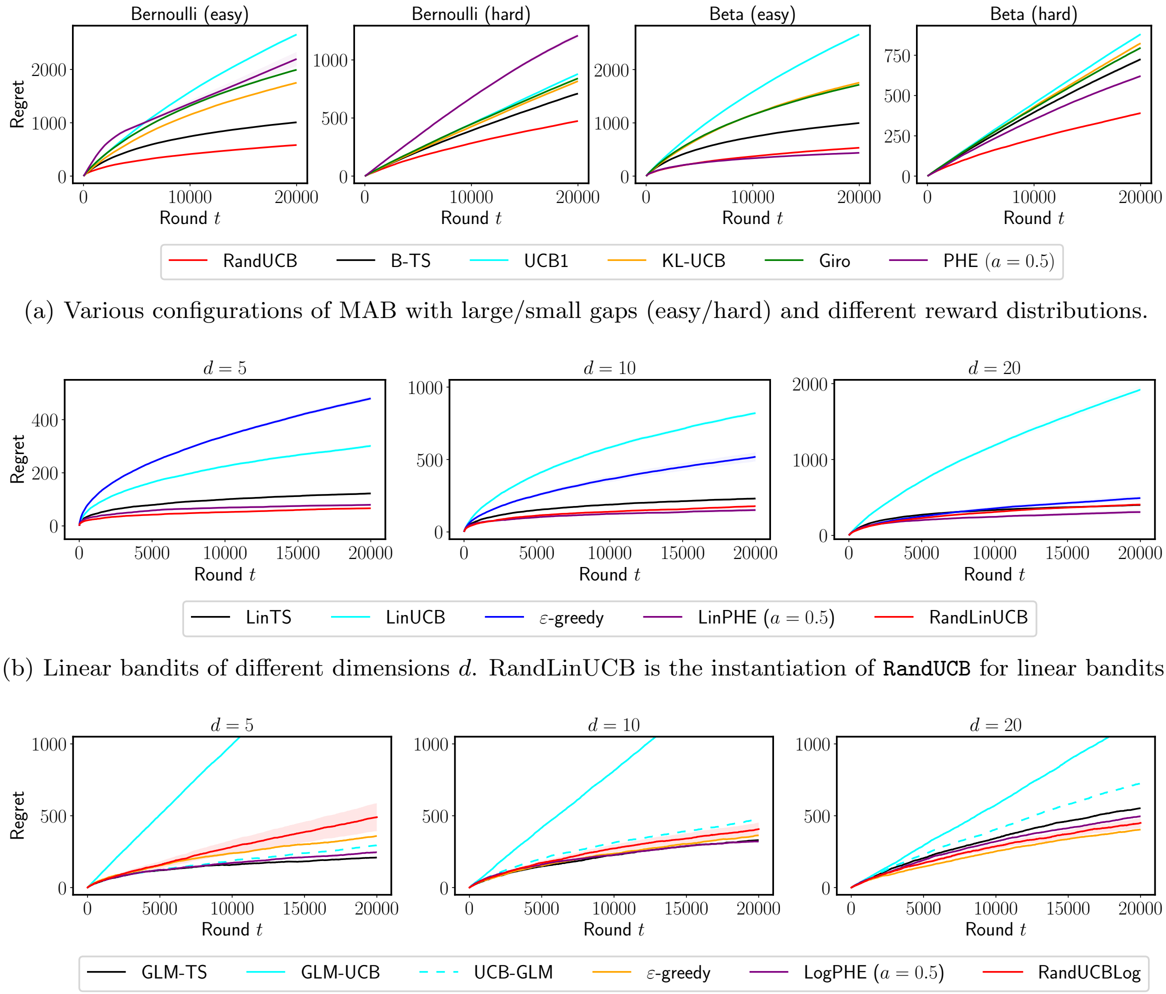
Following the above ablation study, in the following experiments we initiate RandUCB with a (discretized, optimistic) Gaussian sampling distribution and coupled arms with parameters $\varepsilon = 10^{-7}$, $\sigma = 1/8$, $L = 0$, $U = 2\sqrt{\ln(T)}$, and $M = 20$. Figure 1(a) compares RandUCB against 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 $\theta^\star$ and K arms. We consider Bernoulli $\{0, 1\}$ rewards.[7]

For RandUCB in the linear bandit setting, we use the same hyper-parameters as before, but set $U = \beta = \sqrt{\lambda} + \frac{1}{2}\sqrt{\ln(T^2\lambda^{-d}\det(M_t))}$, 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], ε-greedy [Langford and Zhang, 2008], and the best performing variant of the randomized strategy LinPHE [Kveton et al., 2019a]. For ε-greedy, we chose the best performing value of $\varepsilon = 0.05$ and anneal it as $\varepsilon_t = \frac{\varepsilon\sqrt{T}}{2\sqrt{t}}$.

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

[7] To make sure the expected rewards lie in $[0, 1]$, we choose each of $\theta^\star$ 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.



(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 logastic 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

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

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson sampling. *J. ACM*, 2017. Conference version in AISTATS 2013.

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

Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer, 2014.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in neural information processing systems*, pages 6284–6293, 2017.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st annual conference on learning theory (COLT)*, pages 355–366, 2008.

Dean Eckles and Maurits Kaptein. Thompson sampling with the online bootstrap. *arXiv:1410.4009*, 2014.

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

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

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*, pages 359–376, 2011.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of The 31st International Conference on Machine Learning*, pages 100–108, 2014.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

Chih-Wei Hsu, Branislav Kveton, Ofer Meshi, Martin Mladenov, and Csaba Szepesvári. Empirical Bayes regret minimization. *arXiv:1904.02664*, 2019.

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

Jaya Kawale, Hung Hai Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *Neural Information Processing Systems*, 2015.

Baekjin Kim and Ambuj Tewari. On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. In *Advances in Neural Information Processing Systems*, pages 2691–2700, 2019.

Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. *arXiv:1903.09132*, 2019a.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 2019b.

Branislav Kveton, Csaba Szepesvári, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610, 2019c.

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

Boutilier. Randomized exploration in generalized linear bandits. *arXiv:1906.08947*, 2019d.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. Book draft, available at https://tor-lattimore.com/downloads/book/book.pdf.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 19–36, 2012.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 2071–2080. JMLR.org, 2017. URL http://dl.acm.org/citation.cfm?id=3305890.3305895.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3260–3268, 2017.

Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

Ryan McNellis, Adam N. Elmachtoub, Sechan Oh, and Marek Petrik. A practical method for solving contextual bandit problems using decision trees. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017.

Ian Osband and Benjamin Van Roy. Bootstrapped Thompson sampling and deep exploration. *arXiv:1507.00300*, 2015.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. *arXiv:1802.09127*, 2018.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 323–332. ACM, 2015.

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.

Sharan Vaswani, Branislav Kveton, Zheng Wen, Anup Rao, Mark Schmidt, and Yasin Abbasi-Yadkori. New insights into bootstrapping for bandits. *arXiv:1805.09793*, 2018.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 392–401, 2016. URL http://jmlr.org/proceedings/papers/v48/zhangb16.html.