Random Effect Bandits

Rong Zhu
Institute of Science and Technology for Brain-Inspired Intelligence
Fudan University

Branislav Kveton
Amazon*

Abstract

This paper studies regret minimization in a multi-armed bandit. It is well known that side information, such as the prior distribution of arm means in Thompson sampling, can improve the statistical efficiency of the bandit algorithm. While the prior is a blessing when correctly specified, it is a curse when misspecified. To address this issue, we introduce the assumption of a random-effect model to bandits. In this model, the mean arm rewards are drawn independently from an unknown distribution, which we estimate. We derive a random-effect estimator of the arm means, analyze its uncertainty, and design a UCB algorithm `ReUCB` that uses it. We analyze `ReUCB` and derive an upper bound on its n-round Bayes regret, which improves upon not using the random-effect structure. Our experiments show that `ReUCB` can outperform Thompson sampling, without knowing the prior distribution of arm means.

1 INTRODUCTION

We study stochastic multi-armed bandits (Lai and Robbins, 1985; Auer et al., 2002; Lattimore and Szepesvari, 2019), where the learning agent sequentially takes actions in order to maximize its cumulative reward. As the agent learns through experience, it faces a trade-off between exploration and exploitation: exploiting actions that maximize immediate rewards, as estimated by its current model; or improving its future rewards by exploring and learning a better model. Side information, such as the *prior distribution* of arm means in *Thompson sampling (TS)* (Thompson, 1933; Chapelle and Li, 2011; Agrawal and Goyal, 2012, 2013; Russo and Van Roy, 2014; Abeille and Lazaric, 2017), can improve the statistical efficiency of the bandit algorithm and make it more practical.

While the prior is a blessing when correctly specified, a misspecified prior is a curse. Take online advertising as an example. It is well known that click probabilities of ads are low. Therefore, when estimating the click probability of a cold-start ad, it is important to model this structure. One approach would be Bayesian modeling, where the prior distribution is beta with a low mean. The shortcoming of this approach is that the prior needs to be specified, and is potentially misspecified. Therefore, design of bandit algorithms that depend less on exact priors is an important direction.

To address this issue, we study *random-effect models* (Henderson, 1975; Robinson, 1991) in the bandit setting, and refer to the setting as a *random-effect bandit*. Random-effect models were developed in statistics and econometrics (Diggle et al., 2013; Wooldridge, 2001), and are frequentist counterparts of hierarchical Bayesian models (Carlin and Louis, 2000). In our model, the *arm means are sampled i.i.d. from a fixed unknown distribution*. The estimator of arm means is a weighted sum of two terms. The first term is the average of observed rewards of the arm. The second term estimates the common mean from all observations. The weights are chosen adaptively based on data, and balance the common mean estimate with that of the specific arm. Due to this structure, the resulting estimator of arm means is more statistically efficient than in the classical setting.

Our proposed bandit algorithm uses *upper confidence bounds (UCBs)*, which is a popular approach to exploration with guarantees (Lai and Robbins, 1985; Auer et al., 2002; Audibert et al., 2009; Garivier and Cappe, 2011). In round $t \in [n]$, it pulls the arm with the highest UCB, observes its reward, and then updates its estimated arm means and their high-probability confidence intervals. The main difference from the classical algorithms is that all estimates are based on the random-effect model. Our method is essentially a

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

*The work started while being at Google Research.

random-effect UCB1 (Auer et al., 2002), and thus we call it ReUCB.

Since our arm means are stochastic, ReUCB is related to both TS and Bayes-UCB (Kaufmann et al., 2012), which rely on posterior distributions. TS is popular in practice, but the assumption of knowing the prior exactly is rarely satisfied. In ReUCB, we do not require that the prior is *fully specified*, and thus we relax this assumption.

We make the following contributions. First, we introduce the assumption of random-effect models to multi-armed bandits, and properly formulate the corresponding bandit problem. Second, we propose a UCB-like algorithm for this problem, which we call ReUCB. ReUCB estimates arm means using the *best linear unbiased predictor (BLUP)* (Henderson, 1975; Robinson, 1991), a method of estimating random effects without assumptions on distributions. The BLUP estimates leverage the structure of our problem and yield tighter confidence intervals than those of UCB1. Third, we analyze ReUCB and derive an upper bound on its n-round Bayes regret (Russo and Van Roy, 2014) that reflects the structure of our problem. The main challenge in our regret analysis is the underspecified prior. Specifically, ReUCB estimates the distribution of arm means from all observations and then uses it to estimate the mean of each arm. As a result, the estimated arm means are correlated, unlike in a typical multi-armed bandit. Finally, we evaluate ReUCB empirically on a range of problems, such as Gaussian and Bernoulli bandits, and a movie recommendation problem. We observe that ReUCB outperforms or is comparable to TS while using less prior knowledge.

2 RANDOM-EFFECT BANDITS

We study a stochastic K-armed bandit (Lai and Robbins, 1985; Auer et al., 2002; Lattimore and Szepesvari, 2019) where the number of arms can be large but finite. Because the mean rewards of some arms may not be reliably estimated due to many arms, it is challenging to explore all suboptimal arms efficiently. To overcome this challenge, we introduce a novel modeling assumption to multi-armed bandits.

We assume that the *mean reward* of arm $k \in [K]$ follows a *random-effect model*

$$\mu_k = \mu_0 + \delta_k \,, \tag{1}$$

where μ_0 is a common mean, $\delta_k \sim P^{(\mu)}(0, \sigma_0^2)$ is a random offset from that mean, and $P^{(\mu)}(0, \sigma_0^2)$ is a distribution with zero mean and variance σ_0^2. Thus μ_k is a random variable with mean μ_0 and variance σ_0^2. With a lower variance, the differences among the arms are smaller. We improve over traditional bandit designs (Auer et al., 2002) by using the stochasticity of μ_k. Unlike in Thompson sampling (Thompson, 1933; Chapelle and Li, 2011; Russo and Van Roy, 2014) or Bayes-UCB (Kaufmann et al., 2012), we do not assume that the prior of arm means is *conjugate* or *fully specified*. We only require that $P^{(\mu)}(0, \sigma_0^2)$ has a finite second-order moment.

The reward of arm k after the j-th pull is denoted by $r_{k,j}$ and we assume that it is generated i.i.d. as

$$r_{k,j} \sim P^{(r)}(\mu_k, \sigma^2) \,, \tag{2}$$

where $P^{(r)}(\mu_k, \sigma^2)$ is a distribution with mean μ_k and variance σ^2. Similarly to $P^{(\mu)}(0, \sigma_0^2)$ in (1), we only require that its second-order moment is finite.

Our bandit has K arms and a horizon of n rounds. Before the first round, the mean reward of each arm is generated according to (1). In round $t \in [n]$, the agent pulls arm $I_t \in [K]$ and observes its stochastic reward, drawn according to (2). For any arm k and round t, we denote by $n_{k,t}$ the number of pulls of arm k up to round t, and by $r_{k,1}, \dots, r_{k,n_{k,t}}$ the sequence of associated rewards. We call this problem a *random-effect bandit*.

3 MODEL ESTIMATION

This section describes our estimators of arms. In Section 3.1, we estimate μ_k under the assumption that μ_0 is known. In Section 3.2, we provide an estimator for μ_k when μ_0 is unknown. Additionally, we show how to estimate the variance parameters σ_0^2 and σ^2 in Appendix D. Because this section is devoted to estimating means and their variances at a fixed round t, we drop subindexing by t to reduce clutter.

3.1 Estimating μ_k When μ_0 Is Known

We estimate μ_k using the *best linear unbiased prediction (BLUP)*, which is a common method for estimating random effects (Henderson, 1975; Robinson, 1991). The BLUP estimator of μ_k minimizes the mean squared error among the class of linear unbiased estimators that do not depend on the distribution of model error.

We call the sample mean of arm k its *direct estimator*, and define it as $\bar{r}_k = n_k^{-1} \sum_{j=1}^{n_k} r_{k,j}$. From (2), we get that $\text{Var}(\bar{r}_k) = n_k^{-1}\sigma^2$. We improve upon this estimator with a class of linear unbiased estimators of form

$$\breve{\mu}_k := \mu_0 + a(\bar{r}_k - \mu_0) \,,$$

where $a \in \mathbb{R}$ is a to-be-optimized coefficient. Since μ_k is random rather than fixed, BLUP minimizes the mean squared error of $\breve{\mu}_k$ with respect to μ_k, which is $\min_a \mathbb{E}\left[(\breve{\mu}_k - \mu_k)^2\right]$. Note that

$$\mathbb{E}\left[(\breve{\mu}_k - \mu_k)^2\right] = \mathbb{E}\left[[a(\bar{r}_k - \mu_k) + (1-a)(\mu_0 - \mu_k)]^2\right] = a^2 n_k^{-1}\sigma^2 + (1-a)^2\sigma_0^2\,, \quad (3)$$

where the last equality is from (1) and (2), and that the reward noise is independent of δ_k. When (3) is minimized with respect to a, the optimal value of a is

$$w_k = \sigma_0^2/(\sigma_0^2 + n_k^{-1}\sigma^2) = 1/(1 + n_k^{-1}\sigma^2/\sigma_0^2)\,. \quad (4)$$

Thus, if μ_0, σ_0^2, and σ^2 were known; and we plugged our derived w_k into the definition of $\breve{\mu}_k$, we would get the following BLUP estimator of μ_k

$$\tilde{\mu}_k = \mu_0 + w_k(\bar{r}_k - \mu_0) = (1 - w_k)\mu_0 + w_k\bar{r}_k\,. \quad (5)$$

From (4), we have that $\sigma^{-2}\sigma_0^2 \leq w_k < 1$ for $n_k \geq 1$, and that $w_k \to 1$ as n_k increases. We also have

$$w_k n_k^{-1}\sigma^2 = (1 - w_k)\sigma_\mu^2\,. \quad (6)$$

These properties are important in our analysis.

The estimator $\tilde{\mu}_k$ in (5) is biased. The degree of this bias depends on both n_k and σ^2/σ_0^2 in (4). If the arm has not been pulled enough, w_k is low and $\tilde{\mu}_k$ is biased towards μ_0. So we are not as aggressive in exploring as if $w_k = 1$. As the arm is pulled more, $w_k \to 1$ and the bias reduces to zero. When σ_0^2 decreases, the gaps among the arms decrease, and the effect of μ_0 increases. Similarly, as σ^2 increases, the uncertainty in the direct estimator $\bar{r}_k$ increases, and so does the effect of μ_0.

Now we set $a = w_k$ in (3) and get

$$\mathbb{E}\left[(\tilde{\mu}_k - \mu_k)^2\right] = w_k^2 n_k^{-1}\sigma^2 + (1 - w_k)^2\sigma_0^2 = w_k n_k^{-1}\sigma^2 =: \tilde{\tau}_k^2\,, \quad (7)$$

where the last step is from (6). As $\mathrm{Var}(\bar{r}_k) = n_k^{-1}\sigma^2$, (7) shows that $\tilde{\mu}_k$ is a better estimator of μ_k than $\bar{r}_k$, since $w_k < 1$.

3.2 Estimating μ_k When μ_0 Is Unknown

When σ_0^2 and σ^2 are known, the mean of arm means μ_0 can be estimated by the generalized least squares estimator (Rao, 2001). That estimator is

$$\bar{r}_0 = \left[\sum_{k=1}^K (1 - w_k)n_k\right]^{-1} \sum_{k=1}^K (1 - w_k)n_k\bar{r}_k \quad (8)$$

and we derive it in Appendix A. The estimator is more statistically efficient than the ordinary least squares because it weights the mean estimates of individual arms by their heteroscedasticity. Since $(1 - w_k)n_k = \sigma^2/(\sigma_0^2 + n_k^{-1}\sigma^2) \to \sigma_0^{-2}\sigma^2$ as $n_k \to \infty$, we get $\bar{r}_0 - K^{-1}\sum_{k=1}^K \mu_k \to 0$ as $n_k \to \infty$ for all k. This means that $\bar{r}_0$ is a consistent estimator of μ_0.

Algorithm 1 `ReUCB` for random-effect bandits.

1: **for** $t = 1, \ldots, n$ **do**
2: **for** $k = 1, \ldots, K$ **do**
3: $U_{k,t} \leftarrow \hat{\mu}_{k,t} + c_{k,t}$
4: **end for**
5: **if** $t \leq K$ **then** $I_t \leftarrow t$
6: **else** $I_t \leftarrow \arg\max_{k \in [K]} U_{k,t}$
7: Pull arm I_t and observe its reward $r_{I_t, n_{I_t,t}+1}$
8: Update all statistics
9: **end for**

Now we plug the estimator $\bar{r}_0$ of μ_0 into (5) and get a *synthetic estimator* of μ_k,

$$\hat{\mu}_k = (1 - w_k)\bar{r}_0 + w_k\bar{r}_k\,. \quad (9)$$

The key point underlying the synthetic estimator is the weight w_k, which automatically balances variation among the arms and the uncertainty of $\bar{r}_k$. The variance of $\hat{\mu}_k$ is

$$\mathbb{E}\left[(\hat{\mu}_k - \mu_k)^2\right] = w_k n_k^{-1}\sigma^2 + \frac{(1 - w_k)^2}{\sum_{k=1}^K n_k(1 - w_k)}\sigma^2 =: \tau_k^2\,. \quad (10)$$

The derivation of (10) is in Appendix B. The classical estimator of arm means in multi-armed bandits can be compared to that in random-effect bandits as follows.

Proposition 1. *For any arm $k \in [K]$, and any $\sigma^2 > 0$ and $n_k \geq 1$, we have $\tau_k^2 < \sigma^2/n_k$.*

The proof is in Appendix C. Proposition 1 shows that τ_k^2 is always lower than σ^2/n_k when $\sigma^2 > 0$, where the latter is the variance estimate in the classical bandit setting. In the worst case, for $\sigma^2 = 0$, we get $\tau_k^2 = \sigma^2/n_k$, implying that the variance of $\hat{\mu}_k$ equals to that of $\bar{r}_k$. Thus, by using the synthetic estimator $\hat{\mu}_k$, we can be less optimistic than `UCB1`.

4 ALGORITHM

We propose a UCB algorithm for random-effect bandits. The key idea in UCB algorithms (Auer et al., 2002; Audibert et al., 2009) is to pull the arm with the highest sum of its mean reward estimate and a weighted standard deviation of that estimate. In the setting of Section 3.2, the estimated mean reward of arm k is $\hat{\mu}_k$ in (9) and its variance is τ_k^2 in (10). Due to space constraints, we do not present the algorithm for the setting in Section 3.1. In this case, $\hat{\mu}_k$ would be replaced by $\tilde{\mu}_k$ and τ_k^2 would be replaced by $\tilde{\tau}_k^2$.

Our algorithm is presented in Algorithm 1 and we call it ReUCB, which stands for *random-effect UCB*. We subindex all statistics in Section 3 with an additional t, to make clear that we refer to round t. As an example, $\hat{\mu}_{k,t}$ and $\tau_{k,t}^2$ are the respective values of (9) and (10) at the beginning of round t. ReUCB works as follows. It is initialized by pulling each arm once. The *upper confidence bound (UCB)* of arm k in round t is

$$U_{k,t} = \hat{\mu}_{k,t} + c_{k,t}\,,$$

where $c_{k,t} = \sqrt{a\tau_{k,t}^2 \log t}$ is its uncertainty bonus and $a > 0$ is a tunable parameter. In Section 5, we prove regret bounds for $a \geq 1$. In round t, ReUCB pulls the arm with the highest UCB $I_t = \arg\max_{k \in [K]} U_{k,t}$. To break ties, any fixed rule can be used.

4.1 Related Algorithm Designs

ReUCB extends UCB1 to a better BLUP estimator. For $w_{k,t} = 1$ and $a = 1$, ReUCB has a similar UCB to UCB1, $U_{k,t} = \bar{r}_{k,t} + \sqrt{n_{k,t}^{-1}\sigma^2 \log t}$. We call this algorithm $\texttt{ReUCB}^\infty$ and evaluate it empirically in Figure 5 in Appendix I. Our results show that $\texttt{ReUCB}^\infty$ is comparable to TS, but worse than ReUCB. This shows the benefit of our model. Specifically, the estimate of μ_k in ReUCB borrows information from other arms. This increases its statistical efficiency (Proposition 1), since the confidence interval of μ_k in ReUCB can be narrower than in the classical setting. Note that ReUCB with $a = 1$ reduces to UCB1 only if all weights $w_{k,t}$ are one. This could happen only if all arms were pulled infinitely often. So ReUCB with $a = 1$ does not behave like UCB1.

Due to assuming random arm means, ReUCB is related to both TS (Thompson, 1933; Chapelle and Li, 2011; Russo and Van Roy, 2014) and Bayes-UCB (Kaufmann et al., 2012). Both Bayes-UCB and TS maintain posterior distributions. The computation of the posteriors requires that the mean of the prior μ_0 is known. ReUCB employs an alternative random-effect estimator that does not need it.

Li et al. (2011) proposed a hybrid model, where some coefficients are shared by all arms. However, this model is still traditional in the sense that the coefficients that are not shared are estimated separately in each arm. Gupta et al. (2021) recently proposed correlated multi-armed bandits, where the learning agent knows an upper bound on the mean reward of each arm given the mean reward of any other single arm. Such side information could be derived in our setting. However, it is also clearly not as powerful as using the observations of all arms jointly, as in (9) and (10).

5 REGRET ANALYSIS

We derive an upper bound on the n-round regret of ReUCB. In our setting, μ_k are random variables. Under the assumption that $r_{k,j} \sim \mathcal{N}(\mu_k, \sigma^2)$ and $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$, which is used in one of our analyses, $\hat{\mu}_{k,t}$ is the *maximum a posteriori (MAP)* estimate of μ_k given history, meaning that $\hat{\mu}_{k,t}$ can be viewed as a Bayesian estimator. Because of that, we adopt the Bayes regret (Russo and Van Roy, 2014) to analyze ReUCB. The main novelty in our analysis is addressing the unknown mean of the prior.

Let $H_t = (I_\ell, r_{I_\ell, n_{I_\ell,\ell}+1})_{\ell=1}^{t-1}$ be the *history* at the beginning of round t and I_t be the pulled arm in round t. The regret is the difference between the rewards we would have obtained by pulling the optimal arm $I_* = \arg\max_{i \in [K]} \mu_i$ and the rewards that we did obtain in n rounds. Our goal is to bound the Bayes regret $R_n = \mathbb{E}\left[\sum_{t=1}^n \mu_{I_*} - \mu_{I_t}\right]$, where the expectation is over stochastic rewards and random $\mu_1, \ldots, \mu_K$. Our main result is stated below.

Theorem 2. *Consider a K-armed Gaussian bandit with rewards $r_{k,j} \sim \mathcal{N}(\mu_k, \sigma^2)$ and $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Let* ReUCB *use σ_0^2 and σ^2. Then (1) for any $a \geq 1$, the n-round Bayes regret of* ReUCB *is*

$$R_n \leq 2\sqrt{\frac{a\log(1+\sigma^{-2}\sigma_0^2 n)}{\log(1+\sigma^{-2}\sigma_0^2)}\left(1+\frac{\sigma^2}{K\sigma_0^2}\right)\sigma_0^2 K n \log n} + \frac{K\sigma_0^2+\sigma^2}{\sigma_0^2}\sqrt{\frac{8n\sigma_0^2\sigma^2}{\pi(\sigma_0^2+\sigma^2)}}.$$

(2) for any $a \geq 2$, the n-round Bayes regret of ReUCB *is obtained by replacing the last term above with $(1 + \log n)(K + \sigma^2\sigma_0^{-2})\sqrt{2\sigma_0^2\sigma^2/(\pi(\sigma^2+\sigma_0^2))}$.*

5.1 Discussion

Up to logarithmic factors, Theorem 2 shows that the n-round Bayes regret of ReUCB is $O(K\sqrt{n})$ for $a \in [1, 2]$ and $O(\sqrt{Kn})$ for $a \geq 2$. So the regret is sublinear in n for any $a \geq 1$. Since both bounds increase in a, we suggest using $a = 1$, which performs extremely well in practice. Also note that the mean reward estimate in (9) is a weighted sum of the estimate of μ_0 (Term 1) and the per-arm reward mean (Term 2). The variance of the former is linear in σ_0^2, which gives rise to the linear dependence on σ_0 in Theorem 2. Note that this dependence is standard in Bayes regret analyses (Lu and Van Roy, 2019; Basu et al., 2021), and it is due to using similar techniques in our proofs.

A Bayes regret lower bound exists for a K-armed bandit (Lai, 1987). However, it has not been generalized to structured problems yet, including in seminal works on Bayes regret minimization (Russo and Van Roy, 2014). Similarly, we also do not provide a matching lower bound in this work. Instead, we argue that our regret bound reflects the structure of our problem by comparing it to agents that use more information or less structure.

Theorem 2 is proved under the assumption that ReUCB estimates μ_0. Now consider a variant of ReUCB where μ_0 is known. This agent with more information can be analyzed similarly to ReUCB. In this analysis, $\hat{\mu}_{k,t}$ and $\tau_{k,t}^2$ would be replaced by $\tilde{\mu}_{k,t}$ in (5) and $\tilde{\tau}_{k,t}^2$ in (7), respectively. The resulting regret bound would be the same as in Theorem 2, except for the extra factor of $1 + \sigma^2/(K\sigma_0^2)$. Therefore, this factor can be viewed as the price for learning μ_0. As it is $O(1 + 1/K)$, its impact on the Bayes regret of ReUCB is small when K is large.

Now suppose that $\mu_0 \sim \mathcal{N}(0, \sigma_q^2)$. However, the structure that μ_0 is the same for all arms is not modeled. This problem is equivalent to a Bayesian bandit with a per-arm prior $\mathcal{N}(0, \sigma_q^2 + \sigma_0^2)$ and ReUCB with known $\mu_0 = 0$ can solve it. When analyzed, the leading term in Theorem 2 would be

$$2\sqrt{\frac{a\log(1 + \sigma^{-2}(\sigma_q^2 + \sigma_0^2)n)}{\log(1 + \sigma^{-2}(\sigma_q^2 + \sigma_0^2))}(\sigma_q^2 + \sigma_0^2)Kn\log n}\,.$$

Thus, up to logarithmic factors, our regret bound is lower whenever $(1 + \sigma^2/(K\sigma_0^2))\sigma_0^2 \leq \sigma_q^2 + \sigma_0^2$, and it is beneficial to learn the common μ_0 in this case. For any $\sigma_q > 0$, this is guaranteed as K increases.

Theorem 2 can be extended in several ways. First, we generalize the model in (2) to arm-dependent reward noise. Specifically, the reward of arm k after the j-th pull is drawn i.i.d. as $r_{k,j} \sim \mathcal{N}(\mu_k, \sigma_k^2)$, where the variance σ_k^2 may depend on k. In Appendix E, we show that the Bayes regret bound in Theorem 2 still holds for $\sigma^2 = \max_{k\in[K]} \sigma_k^2$. Second, the Gaussian assumption in Theorem 2 is replaced with bounded sub-Gaussianity in Appendix G.

Finally, we would like to point out the limitations of our results. First, our proofs rely on well-behaved posterior distributions, either Gaussian or bounded sub-Gaussian. This is due to limitations of existing Bayes regret analyses, which use it to bound tail events conditioned on history (Russo and Van Roy, 2014). We observe that it is not needed for good practical performance and believe that better analyses will be possible in the future. Second, our proofs are under the assumption that σ_0^2 and σ^2 are known. This is akin to existing Bayes regret analyses. We experiment with estimating σ_0^2 and σ^2 in Section 6.

Now we are ready to prove Theorem 2.

5.2 Proof of Theorem 2

Let the confidence interval of arm k in round t be

$$c_{k,t} = \sqrt{a\tau_{k,t}^2 \log t} = \sqrt{2\tau_{k,t}^2 \log(1/\delta_t)}\,, \tag{11}$$

where $\delta_t = t^{-a/2}$. Define the events that all confidence intervals in round t hold as

$$E_{R;t} = \{\forall k \in [K] : \hat{\mu}_{k,t} - \mu_k \leq c_{k,t}\}\,,$$
$$E_{L;t} = \{\forall k \in [K] : \mu_k - \hat{\mu}_{k,t} \leq c_{k,t}\}\,.$$

Fix round t. The regret in round t is decomposed as

$$\begin{aligned}\mathbb{E}[\mu_{I_*} - \mu_{I_t}] &= \mathbb{E}[\mathbb{E}[\mu_{I_*} - \mu_{I_t} \mid H_t]] \\ &= \mathbb{E}[\mathbb{E}[\mu_{I_*} - \hat{\mu}_{I_*,t} - c_{I_*,t} \mid H_t]] \\ &\quad + \mathbb{E}[\mathbb{E}[\hat{\mu}_{I_*,t} + c_{I_*,t} - \mu_{I_t} \mid H_t]] \\ &\leq \mathbb{E}[\mathbb{E}[\mu_{I_*} - \hat{\mu}_{I_*,t} - c_{I_*,t} \mid H_t]] \\ &\quad + \mathbb{E}[\mathbb{E}[\hat{\mu}_{I_t,t} + c_{I_t,t} - \mu_{I_t} \mid H_t]]\,. \end{aligned} \tag{12}$$

The first equality is by the tower rule. The second is from the fact $\hat{\mu}_{k,t}$ and $c_{k,t}$ are deterministic given H_t. The inequality is from the fact that I_t maximizes $\hat{\mu}_{k,t} + c_{k,t}$ over $k \in [K]$ given H_t. For each term in (12), we get

$$\begin{aligned}&\mathbb{E}[\mu_{I_*} - \hat{\mu}_{I_*,t} - c_{I_*,t} \mid H_t] \\ &\quad \leq \mathbb{E}\left[(\mu_{I_*} - \hat{\mu}_{I_*,t})\mathbb{1}\{\bar{E}_{L;t}\} \,\middle|\, H_t\right], \\ &\mathbb{E}[\hat{\mu}_{I_t,t} + c_{I_t,t} - \mu_{I_t} \mid H_t] \\ &\quad \leq 2\mathbb{E}[c_{I_t,t} \mid H_t] + \mathbb{E}\left[(\hat{\mu}_{I_t,t} - \mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right],\end{aligned}$$

where the inequalities are from the fact that $\mu_{I_*} - \hat{\mu}_{I_*,t} \leq c_{I_*,t}$ on $E_{L;t}$, and that $\hat{\mu}_{I_t,t} - \mu_{I_t} \leq c_{I_t,t}$ on $E_{R;t}$. By chaining all inequalities, the regret is bounded as

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^n (\mu_{I_*} - \mu_{I_t})\right] \leq 2\mathbb{E}\left[\sum_{t=1}^n c_{I_t,t}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[(\hat{\mu}_{I_t,t} - \mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right]\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[(\mu_{I_*} - \hat{\mu}_{I_*,t})\mathbb{1}\{\bar{E}_{L;t}\} \,\middle|\, H_t\right]\right]. \end{aligned} \tag{13}$$

We start with the first term in (13). This term depends on $\tau_{k,t}^2$, which depends on the pulls of all arms, as defined in (10). Therefore, it is challenging to analyze. To do that, we use Lemma 3 of Appendix F, which shows that

$$\tau_{k,t}^2 \leq \beta\tilde{\tau}_{k,t}^2 = \beta\frac{1}{\sigma_0^{-2} + \sigma^{-2}n_{k,t}}$$

for $\beta = 1 + \sigma^2/(K\sigma_0^2)$. This means that we can bound $\tau_{k,t}^2$ by only considering arm k. Then

$$\begin{aligned}\sum_{t=1}^n c_{I_t,t} &\leq \sum_{t=1}^n \sqrt{a\tau_{I_t,t}^2 \log n} \\ &\leq \sqrt{a\beta n\log n}\sqrt{\sum_{t=1}^n \frac{1}{\sigma_0^{-2} + \sigma^{-2}n_{I_t,t}}}\,,\end{aligned}$$

where the first inequality is from the definition of $c_{k,t}$ and $\log t \leq \log n$, and we used the Cauchy-Schwarz inequality in the second one.

Note that $x/\log(1+x) \leq m/\log(1+m)$ for $x \in [0, m]$, because $x = \log(1 + x)$ at $x = 0$ and x grows faster than $\log(1 + x)$ on $[0, m]$. Now we apply this bound for $x = \sigma^{-2}/(\sigma_0^{-2} + \sigma^{-2} n_{I_t,t})$ and $m = \sigma^{-2}\sigma_0^2$, and get

$$\frac{1}{\sigma_0^{-2} + \sigma^{-2} n_{I_t,t}} \leq \gamma \log\left(1 + \frac{\sigma^{-2}}{\sigma_0^{-2} + \sigma^{-2} n_{I_t,t}}\right) = \gamma \log \frac{\sigma_0^{-2} + \sigma^{-2}(n_{I_t,t} + 1)}{\sigma_0^{-2} + \sigma^{-2} n_{I_t,t}},$$

where $\gamma = \sigma_0^2/\log(1 + \sigma^{-2}\sigma_0^2)$. The above leads to telescoping and

$$\sum_{t=1}^{n} \frac{1}{\sigma_0^{-2} + \sigma^{-2} n_{I_t,t}} \leq \gamma K \left[\log(\sigma_0^{-2} + \sigma^{-2} n) - \log(\sigma_0^{-2})\right] = \gamma K \log(1 + \sigma^{-2}\sigma_0^2 n),$$

where we used that any arm is pulled at most n times. Now we put everything together and get

$$\sum_{t=1}^{n} c_{I_t,t} \leq \sqrt{\frac{a \log(1 + \sigma_0^2 n)}{\log(1 + \sigma^{-2}\sigma_0^2)} \beta \sigma^{-2}\sigma_0^2 K n \log n}\,. \quad (14)$$

The next step is the second term in (13). To bound it, we show in Lemma 4 of Appendix F that $\mu_k | H_t \sim \mathcal{N}(\hat{\mu}_{k,t}, \tau_{k,t}^2)$, under the assumption of $r_{k,j} \sim \mathcal{N}(\mu_k, \sigma^2)$ and $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$. By using this property,

$$\mathbb{E}\left[(\hat{\mu}_{I_t,t} - \mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right] \leq \sum_{k=1}^{K} \frac{1}{\sqrt{2\pi\tau_{k,t}^2}} \int_{x \geq c_{k,t}} x \exp\left(-\frac{x^2}{2\tau_{k,t}^2}\right) dx \leq \frac{\delta_t}{\sqrt{2\pi}} \sum_{k=1}^{K} \tau_{k,t}\,,$$

where the last inequality is from (11). It follows that for $a \geq 1$,

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}\left[(\hat{\mu}_{I_t,t} - \mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right]\right] \leq \frac{1}{\sqrt{2\pi}} \sum_{t=1}^{n} t^{-1/2} \beta K \sqrt{\frac{\sigma^2}{1 + \sigma^2\sigma_0^{-2}}} \leq \frac{1}{\sqrt{\pi}} \beta K \sqrt{\frac{2n\sigma^2}{1 + \sigma^2\sigma_0^{-2}}}\,, \quad (15)$$

where the first inequality follows from $\delta_t \leq t^{-1/2}$ for $a \geq 1$ (Lemma 3 of Appendix F), and the last one is from $\sum_{t=1}^{n} t^{-1/2} \leq 2\sqrt{n}$. Similarly, when $a \geq 2$,

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}\left[(\hat{\mu}_{I_t,t} - \mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right]\right] \leq \frac{\beta K (1 + \log n)}{\sqrt{2\pi}} \sqrt{\frac{\sigma^2}{1 + \sigma^2\sigma_0^{-2}}}\,, \quad (16)$$

where the inequality is from $\delta_t \leq t^{-1}$ for $a \geq 2$ and $\sum_{t=1}^{n} t^{-1} \leq 1 + \log n$.

At last, we study the third term in (13). The result in (15) or (16) holds for the term. Therefore, by combing (13), (14), (15), and (16), the theorem is proved.

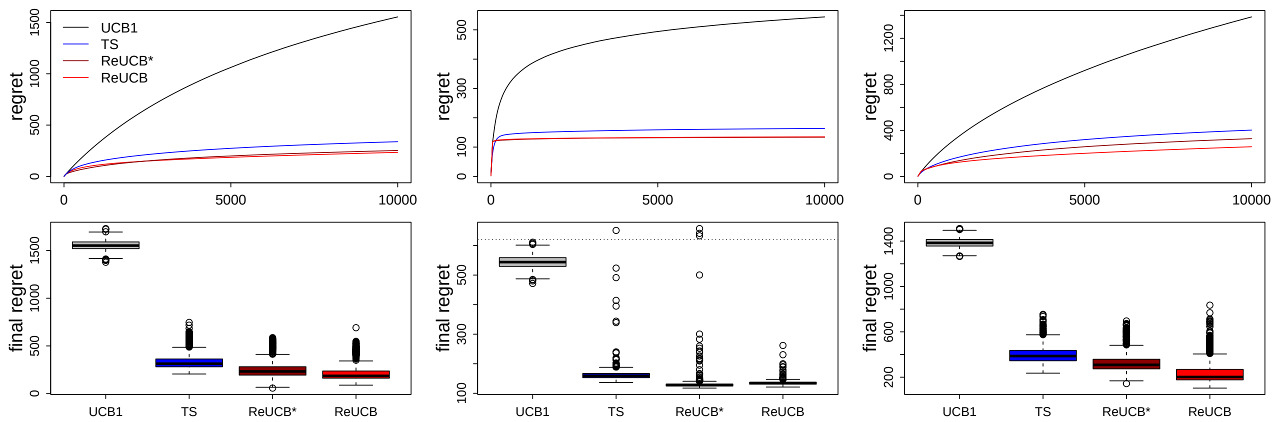
6 SYNTHETIC EXPERIMENTS

We study two bandit settings: Gaussian (Section 6.1) and Bernoulli (Section 6.2). Moreover, in Section 6.3, we study misspecified priors. `ReUCB` is compared to `UCB1` (Auer et al., 2002) and `TS` (Thompson, 1933). `TS` is chosen because it uses the same structure as `ReUCB`, that arm means are random. However, it needs more knowledge, the prior distribution of μ_0. Since `ReUCB` is a UCB algorithm, it is natural to compare it to other UCB algorithms. We focus on `UCB1` due to its simplicity and popularity, but also compare to `Bayes-UCB` (Kaufmann et al., 2012) and `KL-UCB` (Garivier and Cappe, 2011) in Figure 5 of Appendix I. Both `Bayes-UCB` and `KL-UCB` improve over `UCB1`, but are not better than `TS`. This is consistent with other reported results in the literature (Kveton et al., 2019). There are many other potential baselines, such as Giro (Kveton et al., 2018) and PHE (Kveton et al., 2019). Our `ReUCB` is fundamentally different from these methods, since our arm means are random. Also, when compared to these methods, `TS` is typically a strong baseline (Kveton et al., 2019). Therefore, to make our empirical studies clean and focused, we compare to `UCB1` and `TS`.

We evaluate two variants of `ReUCB`: (1) `ReUCB`*, where μ_0 is estimated, and σ_0^2 and σ^2 are known; and (2) `ReUCB`, where all of μ_0, σ_0^2, and σ^2 are estimated. The variance estimators $\hat{\sigma}_{0,t}^2$ and $\hat{\sigma}_t^2$ are provided in (22) and (23), respectively, of Appendix D. Unless specified, the default priors in Gaussian and Bernoulli `TS` are $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathrm{Beta}(1, 1)$, respectively. The upper confidence bound in `UCB1` is $\bar{r}_{k,t} + \sqrt{8 n_{k,t}^{-1} \sigma^2 \log t}$. This is a generalization of the original algorithm to σ^2-sub-Gaussian rewards. In the original algorithm, $\sigma^2 = 1/4$. In Gaussian bandits, we set σ to Gaussian noise. In Bernoulli bandits, this reduces to the `UCB1` index since $\sigma^2 = 1/4$. All simulations are averaged over 1000 independent runs.

6.1 Gaussian Bandits

Our first experiment is on K-armed Gaussian bandits. The reward distribution of arm k is $\mathcal{N}(\mu_k, \sigma^2)$ where $\sigma = 0.5$. We generate μ_k in two ways. First, μ_k are drawn independently from Gaussian prior $\mathcal{N}(\mu_0, \sigma_0^2)$, where we study two settings of (μ_0, σ_0^2): $(1, 0.04)$ (low coefficient of variation 0.2) and $(1, 1)$ (high coefficient



(a) $\mu_k \sim \mathcal{N}(1, 0.04)$

(b) $\mu_k \sim \mathcal{N}(1, 1)$

(c) $\mu_k \sim \mathcal{U}[1, 2]$

Figure 1: K-armed Gaussian bandits with μ_k from $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{U}[1, 2]$. Upper row: Regret as a function of round n. Lower row: Distribution of the regret at the final round.

of variation 1). Second, μ_k are drawn from uniform distribution $\mathcal{U}[1, 2]$. The number of arms is $K = 50$. The horizon is $n = 10^4$ rounds.

Figures 1(a) and 1(b) report results for Gaussian priors, while Figure 1(c) shows results for the uniform prior. We observe that `TS` works well and outperforms `UCB1`. `ReUCB` has a much lower regret than both `UCB1` and `TS`. Besides good average performance, the distribution of the regret in the final round (lower row in Figure 1) shows good stability. The good performance of `ReUCB` in Figure 1(c) indicates that `ReUCB` works for various priors. `ReUCB` performs well empirically because our high-probability confidence intervals are narrower than in the classical setting (Proposition 1). It outperforms `TS` with more information in Gaussian bandits because its confidence interval widths $\tau_{k,t}$ are narrower than the posterior widths of `TS`.

Now we compare the regret of `ReUCB`* and `ReUCB` in Figure 1. Clearly the estimation of σ^2 and σ_0^2 does not have a major impact on the regret of `ReUCB`. In fact, the regret slightly decreases. We believe that this is due to the additional randomness in our method-of-moments estimators of σ^2 and σ_0^2. These results suggest that one limitation of our analysis, that σ^2 and σ_0^2 are known, is not a limitation in practice.

Finally, we report the run times of all algorithms. All experiments are conducted in R, on a PC with 3GHz Intel i7 CPU, 8GB RAM, and OS X operating system. In Figure 1(a), a single run of `ReUCB`, `UCB1`, and `TS` takes on average 1.27s, 1.16s, and 3.19s, respectively. So `ReUCB` is slightly slower than `UCB1` but much faster than `TS`, which is slower due to posterior sampling.

6.2 Bernoulli Bandits

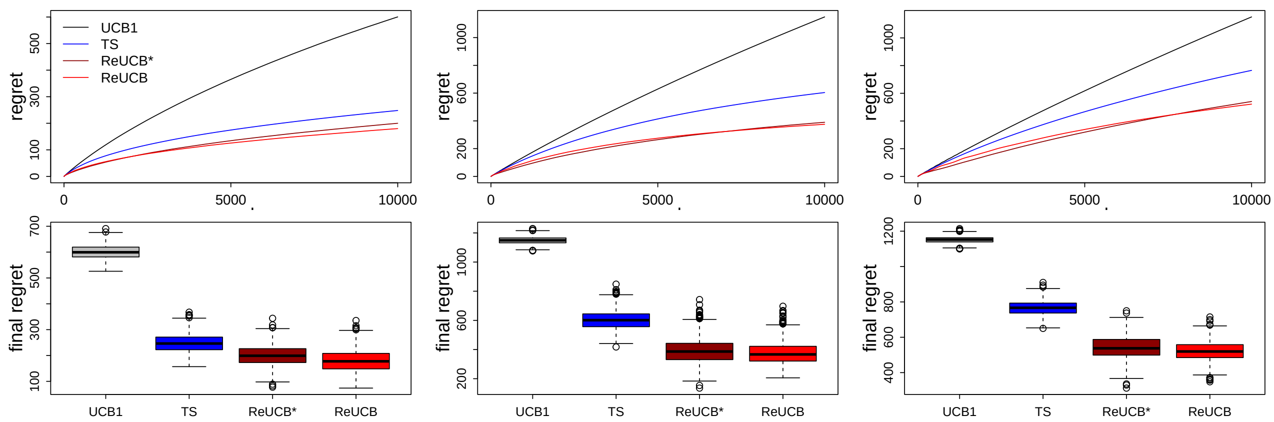
The second experiment is conducted on K-armed Bernoulli bandits, where the reward distribution of arm k is $\mathrm{Bern}(\mu_k)$. The arm means μ_k are drawn i.i.d. from uniform distribution $\mathcal{U}[0.2, 0.5]$. We experiment with three settings for the number of arms $K \in \{20, 50, 100\}$, to show that `ReUCB` performs well across all of them. The horizon is $n = 10^4$ rounds.

As in Figure 1, we observe in Figure 2 that `ReUCB` has a much lower regret than `UCB1` and `TS`, and performs similarly to `ReUCB`*. To implement `ReUCB`*, we set the maximum variance to $\sigma^2 = 1/4$, as suggested in Appendix G. Different from Figure 1, Figure 2 shows the regret for various K. As the number of arms K increases, the gap between our approaches and the baselines increases.

6.3 Model Misspecification

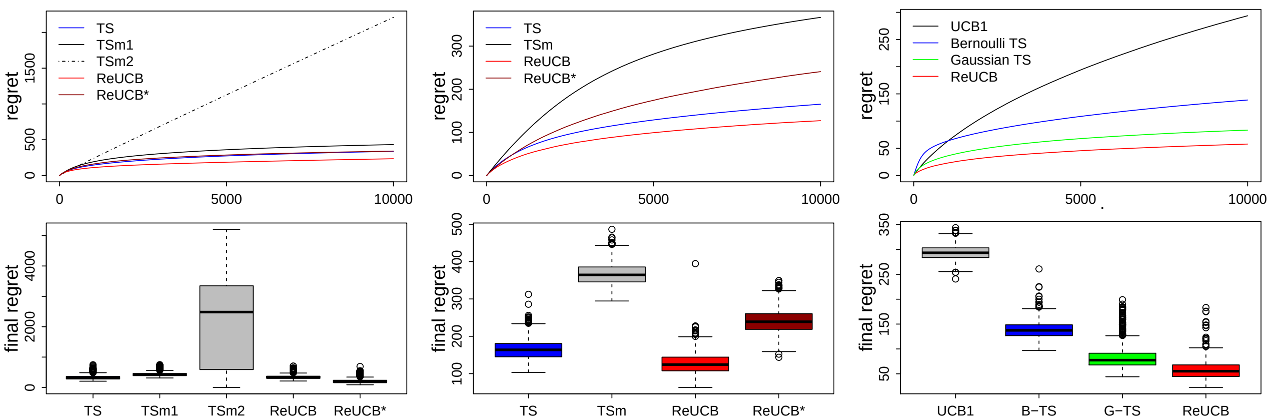
Now we study what happens when `TS` and `ReUCB`* are applied to misspecified models. Note that `ReUCB` is also misspecified in Section 6.2, where the reward noise in Bernoulli bandits depends on the mean of the arm, meaning that it is not identical across the arms.

In the first experiment, we have a 50-armed Gaussian bandit with $\mu_k \sim \mathcal{N}(1, 0.04)$, as in Figure 1(a). We implement two variants of Thompson sampling with misspecified priors: `TSm1` with prior $\mathcal{N}(1, 1)$ (misspecified σ_0^2) and `TSm2` with prior $\mathcal{N}(0, 0.04)$ (misspecified μ_0). In `ReUCB`*, $\sigma_0^2 = 1$ and thus is also misspecified. Our results are reported in Figure 3(a), where `TSm2` fails and has linear regret. The reason is that the misspecified prior has low variance and is downwards biased. Therefore, any initially pulled suboptimal arm is likely



(a) $K = 20$ (b) $K = 50$ (c) $K = 100$

Figure 2: K-armed Bernoulli bandits with $\mu_k \sim \mathcal{U}[0.2, 0.5]$. Upper row: Regret as a function of round n. Lower row: Distribution of the regret at the final round.



(a) Gaussian $\mu_k \sim \mathcal{N}(1, 0.04)$. (b) Bernoulli $\mu_k \sim \text{Beta}(1, 9)$. (c) Truncated Gaussian

Figure 3: Model misspecification experiments. Upper row: Regret as a function of round n. Lower row: Regret at the final round.

to be pulled again. `TSm1` also performs much worse than `TS` with the correct prior. In contrast, `ReUCB` estimates the unknown mean μ_0 and outperforms `TS` that knows μ_0. Even `ReUCB`* with misspecified σ_0^2 is comparable to `TS` with the correct prior.

In the second experiment, we have a Bernoulli bandit with $K = 20$ arms and $\mu_k \sim \text{Beta}(1, 9)$. We study two variants of Thompson sampling: `TS` with a correct prior and `TSm` with misspecified prior $\text{Beta}(9, 1)$. In `ReUCB`*, we set $\sigma_0^2 = 0.00818$ to match the variance of $\text{Beta}(9, 1)$ and $\sigma^2 = 0.25$ because this is the maximum reward variance. This makes `ReUCB`* misspecified. Our results are reported in Figure 3(b) and are similar to Figure 3(a). We observe that `TSm` performs worse than `ReUCB`*, and that `TS` has much higher regret than `ReUCB`.

In the last experiment, we study reward-model misspecification. We have a 20-armed Gaussian bandit where the rewards are truncated to $[0, 1]$ as $r_{k,j} = \min\{\max\{r'_{k,j}, 0\}, 1\}$ for $r'_{k,j} \sim \mathcal{N}(\mu_k, 0.04)$. The mean arm rewards are generated as $\mu_k \sim \mathcal{N}(0.3, 0.01)$. We implement Bernoulli `TS` with prior $\text{Beta}(6, 14)$ to match the moments of $\mathcal{N}(0.3, 0.01)$ and Gaussian `TS` with the correct prior $\mathcal{N}(0.3, 0.01)$. Our results are reported in Figure 3(c). `ReUCB` performs robustly and outperforms Thompson sampling.

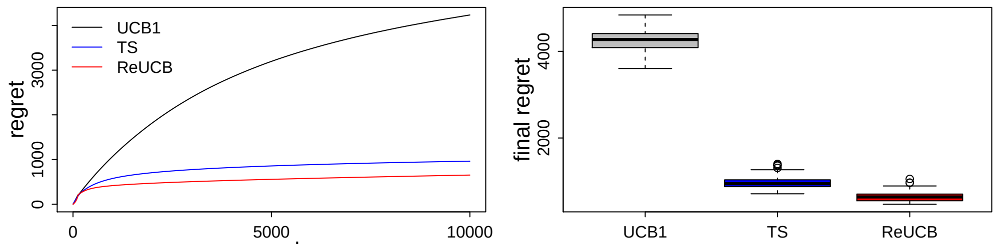


Figure 4: MovieLens experiment. Left: Regret as a function of round n. Right: Distribution of the regret at the final round.

7 EXPERIMENTS ON REAL DATA

In the last experiment, we evaluate ReUCB on a recommendation problem. The goal is to identify the movie that has the highest expected rating. We experiment with the MovieLens dataset (Lam and Herlocker, 2016), where we used a subset of $K = 128$ user groups and $L = 128$ movies randomly chosen from the full dataset, as described in Katariya et al. (2017). For each user group and movie, we average the ratings of all users in the group that rated the movie, and obtain the expected rating matrix $\mathbf{M}$ of rank 5, which is learned by a low-rank approximation on the underlying rating matrix of the user groups and movies. See details of the pre-processing in Katariya et al. (2017).

Our results are averaged over 200 runs. In each run, user j is chosen uniformly at random from $[128]$ and it represents a bandit instance in that run. The goal is to learn the most rewarding movie for user j. We treat this problem as a random-effect bandit with $K = 128$ arms, one per movie, where the mean reward of movie k by user j is $M_{j,k}$. The rewards are generated from $\mathcal{N}(M_{j,k}, 0.796^2)$, where the variance 0.796^2 is estimated from data.

Our approach is compared to UCB1 and TS. We implement Gaussian TS with a prior $\mathcal{N}(\mu_0, \sigma_0^2)$ that is estimated from the empirical mean rewards of all 128 arms. That is, for each user j, μ_0 and σ_0^2 are the empirical mean and variance of $M_{j,1}, \dots, M_{j,128}$. We implement UCB1 by taking the upper confidence bound with $\sigma = 0.796$. Our results are reported in Figure 4. We observe that ReUCB has a much lower regret than both UCB1 and TS. This indicates that ReUCB can learn the biases of different bandit instances, which represent individual users.

8 CONCLUSIONS

We propose a random-effect bandit, a novel setting where the arm means are sampled i.i.d. from an unknown distribution. Using this model, we obtain an improved estimator of arm means and design an efficient UCB-like algorithm ReUCB. ReUCB is prior-free and we show empirically that it can outperform Thompson sampling. We analyze ReUCB and prove a Bayes regret bound on its n-round regret, which improves over not using the random-effect structure.

Our initial results with random-effect models are encouraging. One limitation of our current approach is that ReUCB is not contextual. In the future work, we plan to propose random-effect contextual bandits and provide an algorithm for them. Another limitation is that our regret analysis is under the assumption that ReUCB knows σ_0^2 and σ^2. While this seems limiting, it is a weaker assumption than knowing the common mean μ_0, which would be a standard assumption in the analysis of TS and Bayes-UCB.

Acknowledgements

This work was partially supported by Science and Technology Innovation 2030 - Brain Science and Brain-Inspired Intelligence Project (No. 2021ZD0200204), by National Natural Science Foundation of China (No.11871459), and by Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01). The authors would like to thank anonymous reviewers for their helpful comments.

References

M. Abeille and A. Lazaric. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11:5165–5197, 2017.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 1–39, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013.

J.-Y. Audibert, R. Munosc, and C. Szepesvari. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

S. Basu, B. Kveton, M. Zaheer, and C. Szepesvari. No regrets for learning the prior in bandits. In *Advances in neural information processing systems*, 2021.

B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, second edition edition, 2000.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, second edition edition, 2013.

A. Garivier and O. Cappe. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceeding of the 24th Annual Conference on Learning Theory*, pages 359–376, 2011.

S. Gupta, S. Chaudhari, G. Joshi, and O. Yağan. Multi-armed bandits with correlated arms. *IEEE Transactions on Information Theory*, 67(10):6711–6732, 2021.

C. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31: 423–447, 1975.

R. Kachar and D. Harville. Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association*, 79:853–862, 1984.

S. Katariya, B. Kveton, C. Szepesvari, C. Vernade, and W. Z. Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

E. Kaufmann, O. Cappa, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, 2012.

B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, M. Ghavamzadeh, and T. Lattimore. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proceedings of the 36rd International Conference on Machine Learning*, 2018.

B. Kveton, C. Szepesvari, M. Ghavamzadeh, and C. Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

T. L. Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091 – 1114, 1987.

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

S. Lam and J. Herlocker. Movielens dataset, 2016.

T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

L. Li, W. Chu, J. Langford, , and X. Wang. Unbiased offline evaluation of contextual-banditbased news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

X. Lu and B. Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

C. Rao. *Linear Statistical Inference and its Applications*. Wiley, second edition edition, 2001.

G. Robinson. That blup is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6:15–51, 1991.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, first edition edition, 2001.

A Derivation of (8)

Let $\mathbf{r}_k = (r_{k,1}, r_{k,2}, \dots, r_{k,n_k})^\top$ be a column vector of rewards obtained by pulling arm k. From modeling assumptions (1) and (2), we get

$$\mathbf{r}_k = \mu_k \mathbf{1}_k + \mathbf{e}_k \,,$$

where $\mathbf{e}_k = (e_{k,1}, e_{k,2}, \dots, e_{k,n_k})^\top$ and $e_{k,j} \sim P_r(0, \sigma^2)$. The covariance matrix for vector $\mathbf{r}_k$ is $\mathbf{V}_k = \sigma_0^2 \mathbf{1}_k \mathbf{1}_k^\top + \sigma^2 \mathbf{I}_k$, where $\mathbf{1}_k$ is an all-ones vector of length n_k and $\mathbf{I}_k$ is a $n_k \times n_k$ identity matrix. The generalized least squares estimator of μ_0 minimizes the following loss

$$L(\mu_0) = \sum_{k=1}^{K} (\mathbf{r}_k - \mu_0 \mathbf{1}_k)^\top \mathbf{V}_k^{-1} (\mathbf{r}_k - \mu_0 \mathbf{1}_k)$$

with respect to μ_0. Using the Sherman-Morrison formula,

$$\begin{aligned}\mathbf{V}_k^{-1} &= \sigma^{-2}\mathbf{I}_k - \sigma^{-4}(\sigma_0^{-2} + n_k \sigma^{-2})^{-1} \mathbf{1}_k \mathbf{1}_k^\top = \sigma^{-2}\mathbf{I}_k - \sigma^{-2}\sigma_0^2(\sigma^2 + n_k \sigma_0^2)^{-1} \mathbf{1}_k \mathbf{1}_k^\top \\ &= \sigma^{-2}\mathbf{I}_k - n_k^{-1}\sigma^{-2} w_k \mathbf{1}_k \mathbf{1}_k^\top \,.\end{aligned}$$

Inserting the above formula into $L(\mu_0)$ yields

$$L(\mu_0) = \sigma^{-2} \sum_{k=1}^{K} \left[\|\mathbf{r}_k - \mu_0 \mathbf{1}_k\|^2 - n_k w_k (\bar{r}_k - \mu_0)^2 \right] \,.$$

The first-order derivative of $L(\mu_0)$ with respect to μ_0 is

$$\frac{\partial L(\mu_0)}{\partial \mu_0} = 2\sigma^{-2} \sum_{k=1}^{K} (1 - w_k) n_k (\bar{r}_k - \mu_0) \,.$$

Thus we get that μ_0 is estimated by

$$\bar{r}_0 = \frac{\sum_{k=1}^{K} (1 - w_k) n_k \bar{r}_k}{\sum_{k=1}^{K} (1 - w_k) n_k} \,. \tag{17}$$

B Derivation of (10)

Now we derive the variance of $\hat{\mu}_k$. We have

$$\mathrm{E}[(\hat{\mu}_k - \mu_k)^2] = \mathrm{E}[(\tilde{\mu}_k - \mu_k + \hat{\mu}_k - \tilde{\mu}_k)^2] = \mathrm{E}[(\tilde{\mu}_k - \mu_k)^2] + \mathrm{E}[(\hat{\mu}_k - \tilde{\mu}_k)^2] \,, \tag{18}$$

where, we recall, $\tilde{\mu}_k$ in (5) and $\hat{\mu}_k$ in (9) differ only in that μ_0 is estimated by $\bar{r}_0$, and the first step follows from $\mathrm{E}[(\tilde{\mu}_k - \mu_k)(\hat{\mu}_k - \tilde{\mu}_k)] = 0$ (Kachar and Harville, 1984). Now we derive the two terms of the right-hand of (18). Note that the first term is shown in (7). For the other term,

$$\mathrm{E}[(\hat{\mu}_k - \tilde{\mu}_k)^2] = (1 - w_k)^2 \mathrm{E}[(\bar{r}_0 - \mu_0)^2] \,. \tag{19}$$

From $\bar{r}_0$ in (8),

$$\begin{aligned}\mathrm{Var}(\bar{r}_0) &= \left[\sum_{k=1}^{K} (1 - w_k) n_k \right]^{-2} \sum_{k=1}^{K} (1 - w_k)^2 n_k^2 \mathrm{Var}(\bar{r}_k) \\ &= \left[\sum_{k=1}^{K} (1 - w_k) n_k \right]^{-2} \sum_{k=1}^{K} (1 - w_k)^2 n_k^2 (\sigma_0^2 + \sigma^2 / n_k) \\ &= \sigma^2 \left[\sum_{k=1}^{K} (1 - w_k) n_k \right]^{-1} \,,\end{aligned}$$

where the last step is from $\sigma^2 + n_k\sigma_0^2 = (1-w_k)^{-1}\sigma^2$. Inserting the result above into (19), we have

$$\mathrm{E}[(\hat{\mu}_k - \tilde{\mu}_k)^2] = \frac{(1-w_k)^2\sigma^2}{\sum_{k=1}^{K} n_k(1-w_k)}. \tag{20}$$

Therefore, inserting (7) & (20) into (18),

$$\mathrm{E}[(\hat{\mu}_k - \mu_k)^2] = w_k n_k^{-1}\sigma^2 + \frac{(1-w_k)^2\sigma^2}{\sum_{k=1}^{K} n_k(1-w_k)} =: \tau_k^2, \tag{21}$$

where the reason that we use the squares notation in τ_k^2 is because τ_k^2 is a mean squared error not smaller than 0.

C Proofs of Proposition 1

Note that

$$\sigma^2/n_k - \tau_k^2 = (1-w_k)n_k^{-1}\sigma^2 - \frac{(1-w_k)^2\sigma^2}{\sum_{i=1}^{K} n_i(1-w_i)} = (1-w_k)n_k^{-1}\sigma^2\left[1 - \frac{n_k(1-w_k)}{\sum_{i=1}^{K} n_i(1-w_i)}\right].$$

Obviously, $n_k(1-w_k) < \sum_{i=1}^{K} n_i(1-w_i)$ as long as $n_i \geq 1$ for all $i \in [K]$. Thus, when $\sigma^2 > 0$, we get $\tau_k^2 < \sigma^2/n_k$.

D Estimation of σ_0^2 and σ^2

Our BLUP estimators depend on σ_0^2 and σ^2, which may be unknown. We can estimate these quantities, and replace σ_0^2 and σ^2 in w_k with these estimates. Various methods for obtaining consistent estimators of $\hat{\sigma}_0^2$ and $\hat{\sigma}^2$ are available, including the method of moments, maximum likelihood, and restricted maximum likelihood. See Robinson (1991) for details.

We use the method of moments, which does not rely on the assumption of distributions. Unbiased quadratic estimates of σ^2 and σ_0^2 are given by

$$\hat{\sigma}^2 = \left[\sum_{k=1}^{K}(n_k - 1)\right]^{-1} \sum_{k=1}^{K}\sum_{j=1}^{n_k}(r_{k,j} - \bar{r}_k)^2 \tag{22}$$

and

$$\hat{\sigma}_0^2 = n_*^{-1}\sum_{k=1}^{K} n_k u_k^2\,, \tag{23}$$

where $n_* = \sum_{k=1}^{K} n_k - \left(\sum_{k=1}^{K} n_k\right)^{-1} \sum_{k=1}^{K} n_k^2$ and $u_k = \bar{r}_k - \left(\sum_{k=1}^{K} n_k\right)^{-1} \sum_{k=1}^{K}\sum_{j=1}^{n_k} r_{k,j}\,.$

E Varying Reward Noise

We can generalize the standard random effect model in (2) by eliminating the assumption of identical observation noise across all arms. Instead, we allow the noise vary across arms. Specifically, the reward of arm k after the j-th pull is assumed to be generated i.i.d. as

$$r_{k,j} \sim \mathcal{N}(\mu_k, \sigma_k^2)\,,$$

where, compared to model (2), variance σ_k^2 is allowed to depend on k. Accordingly, we have the estimate $\hat{\mu}_k^h = (1-w_k^h)\bar{r}_0 + w_k^h \bar{r}_k$ and $\hat{\mu}_k^h - \mu_k \mid H_t \sim \mathcal{N}(0, \tau_{h;k,t}^2)$, where the superscript "h" means the heteroscedasticity of reward noise among arms,

$$w_k^h = \sigma_0^2/(\sigma_0^2 + \sigma_k^2/n_k) \text{ and } \tau_{h;k,t}^2 = w_k^h n_k^{-1}\sigma_k^2 + (1-w_k^h)^2\sigma_k^2 / \sum_{k=1}^K n_k(1-w_k^h).$$

We consider an upper bound of these σ_k^2, e.g., $\max_k \sigma_k^2$, denoted by σ^2. Notice

$$\tau_{h;k,t}^2 = n_k^{-1}\sigma_k^2\sigma_0^2/(\sigma_0^2 + n_k^{-1}\sigma_k^2) + (1-w_k^h)^2\sigma_0^2/(\sum_{k=1}^K w_k^h).$$

Because $n_k^{-1}\sigma_k^2\sigma_0^2/(\sigma_0^2 + n_k^{-1}\sigma_k^2)$ is increasing of σ_k^2 and w_k^h is decreasing of σ_k^2, we have that

$$\tau_{h;k,t}^2 \le \tau_{k,t}^2.$$

Therefore, we uniformly use the upper variance σ^2 across arms replacing of σ_k^2. By this way, the Bayes regret bound in Theorem 2 still holds.

F Lemmas

Lemma 3. *We have that*

$$\tau_k^2 \le \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2 + \sigma^2}(1 + K^{-1}\sigma^2\sigma_0^{-2}).$$

Proof. Now we provide an upper bound on τ_k^2. By using $n_k(1-w_k)\sigma_0^2 = w_k\sigma^2$, we have

$$\begin{aligned}
\tau_k^2 &= w_k n_k^{-1}\sigma^2 + \frac{(1-w_k)^2\sigma^2}{\sum_{k=1}^K n_k(1-w_k)} \\
&= w_k n_k^{-1}\sigma^2 + \frac{(1-w_k)^2\sigma_0^2}{\sum_{k=1}^K w_k} \\
&\le w_k n_k^{-1}\sigma^2 + K^{-1}(1-w_k)^2(\sigma_0^2+\sigma^2) \\
&= w_k n_k^{-1}\sigma^2 + K^{-1}n_k^{-1}(1-w_k)w_k\sigma^2\sigma_0^{-2}(\sigma_0^2+\sigma^2) \\
&\le w_k n_k^{-1}\sigma^2(1 + K^{-1}\sigma^2\sigma_0^{-2}) \\
&= \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2+\sigma^2}(1 + K^{-1}\sigma^2\sigma_0^{-2}),
\end{aligned}$$

where the first inequality is from $w_k \ge \sigma_0^2/(\sigma_0^2+\sigma^2)$ and the last inequality is from $1 - w_k \le \sigma^2/(\sigma_0^2+\sigma^2)$ due to $n_k \ge 1$ for all $k \in [K]$. □

Lemma 4. *Let $r_{k,j} \sim \mathcal{N}(\mu_k, \sigma^2)$ and $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Assuming that σ^2 and σ_0^2 are known, and μ_0 is an improper flat prior, i.e., $p(\mu_0) \propto 1$, we have that $\mu_k \mid H_t \sim \mathcal{N}(\hat{\mu}_{k,t}, \tau_{k,t}^2)$.*

Proof. Recall the following well-known identity. Let $Y|X = x \sim \mathcal{N}(ax+b, \sigma^2)$ and $X \sim \mathcal{N}(\mu, \sigma_x^2)$. Then

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma_x^2 + \sigma^2).$$

Obviously, if we set X to $\mu_0 \mid H_t$ and Y to $\mu_k \mid H_t$, we can apply this result to obtain the distribution of $\mu_k \mid H_t$ from the distributions of $\mu_k \mid \mu_0, H_t$ and $\mu_0 \mid H_t$. The distribution of $\mu_k \mid \mu_0, H_t$ is studied in Section 3.1. Under the assumptions of Lemma 4, $\mu_k \mid \mu_0, H_t$ is a Gaussian with mean in (5) and variance in (7).

Now we derive the distribution of $\mu_0 \mid H_t$. Note that $p(\mu_0) \propto 1$ is the extreme case of $\mu_0 \sim \mathcal{N}(0, \lambda)$ as $\lambda \to \infty$. Assuming $\mu_0 \sim \mathcal{N}(0, \lambda)$, $\bar{r}_k$ can be considered to be generated from the following Bayesian model:

$$\bar{r}_k \mid \mu_0 \sim \mathcal{N}(\mu_0, \sigma_0^2 + \sigma^2/n_k), \quad \mu_0 \sim \mathcal{N}(0, \lambda).$$

Thus, the distribution of $\mu_0 \mid H_t$ is easily obtained as

$$\mu_0 \mid H_t \sim \mathcal{N}(\bar{r}_0, \sigma_0^2),$$

where

$$\bar{r}_0 = \sigma_0^2[\sigma^{-2}\sum_{k=1}^{K}(1-w_k)n_k\bar{r}_k] \quad \text{and} \quad \sigma_0^2 = [\lambda^{-1} + \sigma^{-2}\sum_{k=1}^{K}(1-w_k)n_k]^{-1}.$$

Taking $\lambda \to \infty$, we have that

$$\mu_0 \mid H_t \sim \mathcal{N}(\bar{r}_0, \sigma^2[\sum_{k=1}^{K}(1-w_k)n_k]^{-1}).$$

Using the above results, we obtain the distribution of $\mu_k \mid H_t$ from the distributions of $\mu_k \mid \mu_0, H_t$ and $\mu_0 \mid H_t$. That distribution is a Gaussian with mean in (9) and variance in (10). This completes the proof. ☐

G Extension to sub-Gaussian

Our analysis can be extended to bounded sub-Gaussian random variables. Without loss of generality, we consider the support of $[0,1]$ below.

Theorem 5. *Consider* ReUCB *in a K-armed bandit with sub-Gaussian rewards $r_{k,j}-\mu_k \sim subG(\sigma^2)$ and $\mu_k-\mu_0 \sim subG(\sigma_0^2)$ with support in $[0,1]$. Let σ_0^2 and σ^2 be known and used by* ReUCB. *Define $m = [1+K^{-1}\sigma_0^2/(\sigma_0^2+\sigma^2)]^{-1}[1+K^{-1/2}\sigma/\sigma_0]^2$. Then (1) for any $a \geq m$, the n-round Bayes regret of* ReUCB *is*

$$R_n \leq 2\left(1+\frac{\sigma^2}{K\sigma_0^2}\right)\sqrt{\frac{a\sigma_0^2\log(1+\sigma^{-2}\sigma_0^2 n)}{\log(1+\sigma^{-2}\sigma_0^2)}Kn\log n} + \frac{K\sigma_0^2+\sigma^2}{\sigma_0^2}\sqrt{\frac{8n\sigma_0^2\sigma^2}{\pi(\sigma_0^2+\sigma^2)}}\,.$$

(2) for any $a \geq 2m$, the n-round Bayes regret of ReUCB *is obtained by replacing the last term above with $2K(1+\log n)$.*

Proof. Under the sub-Gaussian assumptions that $\mu_k - \mu_0 \sim \text{subG}(\sigma_0^2)$ and $\bar{r}_{k,t} - \mu_k \sim \text{subG}(\sigma^2/n_k)$, Lemma 6 shows that $\mu_k - \hat{\mu}_{k,t} \sim \text{subG}(\tau_{k,t}^{*2})$ for $k \in [K]$, where

$$\tau_{k,t}^{*2} = \sigma^2\left[\sqrt{\frac{w_k}{n_k}} + \sqrt{\frac{(1-w_k)^2}{\sum_{k=1}^{K}(1-w_k)n_k}}\right]^2. \tag{24}$$

Lemma 7 of the Appendix shows that

$$\tau_{k,t}^{*2} \leq \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2+\sigma^2}\left(1+\sqrt{K^{-1}\sigma^2\sigma_0^{-2}}\right)^2. \tag{25}$$

Note that $c_{k,t} = \sqrt{2\tau_{k,t}^2\log(1/\delta_t)}$, $E_{R;t}$, and $E_{L;t}$. We have that

$$\mathbb{E}\left[(\hat{\mu}_{I_t,t}-\mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right] \leq \mathbb{E}\left[\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right], \tag{26}$$

where the inequality is from $(\hat{\mu}_{I_t,t}-\mu_{I_t}) \in [0,1]$ on $\bar{E}_{R;t}$ due to the support $[0,1]$. It follows that when $a \geq m$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}\left[(\hat{\mu}_{I_t,t}-\mu_{I_t})\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right]\right] &\leq \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\{\bar{E}_{R;t}\} \,\middle|\, H_t\right]\right] = \sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\{\bar{E}_{R;t}\}\right] \\ &\leq K\sum_{t=1}^{n}\delta_t^{1/m} \leq K\sum_{t=1}^{n}t^{-1/2} \leq 2K\sqrt{n},\end{aligned}$$

where the second inequality is from $\tau_{k,t}^2/\tau_{k,t}^{*2} \geq 1/m$ shown in Lemma 9, and the third inequality is from the $a \geq m$. Similarly we have $\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}\left[(\mu_{I_*}-\hat{\mu}_{I_*,t})\mathbb{1}\{\bar{E}_{L;t}\} \,\middle|\, H_t\right]\right] \leq 2K\sqrt{n}$.

Similarly to (14), we have that

$$\mathbb{E}\left[\sum_{t=1}^{n}\sqrt{2\tau_{I_t,t}^2\log(1/\delta_t)}\right] \le \left(1+\frac{\sigma^2}{K\sigma_0^2}\right)\sqrt{\frac{a\sigma_0^2\log(1+\sigma^{-2}\sigma_0^2 n)}{\log(1+\sigma^{-2}\sigma_0^2)}Kn\log n}.$$

Therefore, when $a \ge m$, the regret is bounded as

$$\mathbb{E}\left[R_n\right] \le 2\left(1+\frac{\sigma^2}{K\sigma_0^2}\right)\sqrt{\frac{a\sigma_0^2\log(1+\sigma^{-2}\sigma_0^2 n)}{\log(1+\sigma^{-2}\sigma_0^2)}Kn\log n} + 4K\sqrt{n}.$$

Similarly, when $a \ge 2m$, the regret is bounded as

$$\mathbb{E}\left[R_n\right] \le 2\left(1+\frac{\sigma^2}{K\sigma_0^2}\right)\sqrt{\frac{a\sigma_0^2\log(1+\sigma^{-2}\sigma_0^2 n)}{\log(1+\sigma^{-2}\sigma_0^2)}Kn\log n} + 4K(1+\log n).$$

□

When comparing Theorems 2 and 5, the regret is of the same order. Since the assumption of $r_{k,j}-\mu_k \sim \mathrm{subG}(\sigma^2)$ allows for modeling arm-dependent reward noise, such as $\sigma^2 = 1/4$ in Bernoulli bandits, Theorem 5 holds for Bernoulli bandits. In Section 6, we experiment with Bernoulli bandits. Unlike Theorem 2, Theorem 5 requires that $a \ge m$ or $a \ge 2m$. We note that m in Theorem 5 is typically small, and approaches 1 as K and σ_0^2/σ^2 increase.

Lemma 6. *We have that for $k \in [K]$*

$$\mu_k - \hat{\mu}_{k,t} \sim subG(\tau_{k,t}^{*2}),$$

where

$$\tau_{k,t}^{*2} = \sigma^2\left[\sqrt{\frac{w_k}{n_k}} + \sqrt{\frac{(1-w_k)^2}{\sum_{k=1}^{K}(1-w_k)n_k}}\right]^2.$$

Notice

$$\hat{\mu}_{k,t} - \mu_k = \tilde{\mu}_{k,t} - \mu_k + \hat{\mu}_{k,t} - \tilde{\mu}_{k,t},$$

where $\tilde{\mu}_{k,t} - \mu_k = w_k(\bar{r}_{k,t} - \mu_k) + (1-w_k)(\mu_k - \mu_0)$ and $\hat{\mu}_{k,t} - \tilde{\mu}_{k,t} = (1-w_k)(\bar{r}_{0,t} - \mu_0)$. From the properties of sub-Gaussian (Fact 2) and the independence between $\mu_0 - \mu_k$ and $\bar{r}_{k,t} - \mu_k$, we have

$$\tilde{\mu}_{k,t} - \mu_k \sim \mathrm{subG}\left(w_k\sigma^2/n_k\right);$$
$$\bar{r}_{0,t} - \mu_0 \sim \mathrm{subG}\left(\sigma^2[\sum_{k=1}^{K}(1-w_k)n_k]^{-1}\right).$$

Thus, the properties of sub-Gaussian (Fact 2) tell us that

$$\tau_{k,t}^{*2} = \sigma^2\left[\sqrt{\frac{w_k}{n_k}} + \frac{1-w_k}{\sqrt{\sum_{k=1}^{K}(1-w_k)n_k}}\right]^2,$$

i.e., $\hat{\mu}_{k,t} - \mu_k$ is sub-Gaussian with variance proxy $\tau_{k,t}^{*2}$.

Lemma 7.

$$\tau_{k,t}^{*2} \le \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2+\sigma^2}\left(1+\sqrt{K^{-1}\sigma^2\sigma_0^{-2}}\right)^2.$$

Proof. Now we provide an upper bound on τ_k^{*2}. By making using of $n_k(1-w_k)\sigma_0^2 = w_k\sigma^2$, we have that

$$\begin{aligned}
\tau_{k,t}^{*2} &= \sigma^2\left[\sqrt{w_k n_k^{-1}} + \sqrt{\frac{(1-w_k)^2}{\sum_{k=1}^K n_k(1-w_k)}}\right]^2 \\
&= \left[\sqrt{\sigma^2 w_k n_k^{-1}} + \sqrt{\frac{(1-w_k)^2\sigma_0^2}{\sum_{k=1}^K w_k}}\right]^2 \\
&\le \left[\sqrt{\sigma^2 w_k n_k^{-1}} + \sqrt{K^{-1}(1-w_k)^2(\sigma_0^2+\sigma^2)}\right]^2 \\
&= \left[\sqrt{\sigma^2 w_k n_k^{-1}} + \sqrt{K^{-1}n_k^{-1}(1-w_k)w_k\sigma^2\sigma_0^{-2}(\sigma_0^2+\sigma^2)}\right]^2 \\
&\le \left[\sqrt{\sigma^2 w_k n_k^{-1}}(1+\sqrt{K^{-1}\sigma^2\sigma_0^{-2}})\right]^2 \\
&= \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2+\sigma^2}\left(1+\sqrt{K^{-1}\sigma^2\sigma_0^{-2}}\right)^2,
\end{aligned}$$

where the first inequality is from $w_k \ge \sigma_0^2/(\sigma_0^2+\sigma^2)$, and the last inequality is from $1-w_k \le \sigma^2/(\sigma_0^2+\sigma^2)$. □

Lemma 8.

$$\tau_{k,t}^2 \ge \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2+\sigma^2}(1+K^{-1}\sigma_0^2/(\sigma_0^2+\sigma^2)).$$

Proof. Now we provide a lower bound on τ_k^2. Similarly, we have that

$$\begin{aligned}
\tau_{k,t}^2 &= w_k n_k^{-1}\sigma^2 + \frac{(1-w_k)^2\sigma_0^2}{\sum_{k=1}^K w_k} \\
&\ge w_k n_k^{-1}\sigma^2 + K^{-1}(1-w_k)^2\sigma_0^2 \\
&= w_k n_k^{-1}\sigma^2 + K^{-1}n_k^{-1}(1-w_k)w_k\sigma^2 \\
&\ge w_k n_k^{-1}\sigma^2(1+K^{-1}\sigma_0^2/(\sigma_0^2+\sigma^2)) \\
&= \frac{\sigma_0^2\sigma^2}{n_k\sigma_0^2+\sigma^2}(1+K^{-1}\sigma_0^2/(\sigma_0^2+\sigma^2)),
\end{aligned}$$

where the first inequality is from $w_k \le 1$, and the last inequality is from $1-w_k \ge \sigma_0^2/(\sigma_0^2+\sigma^2)$. □

Lemma 9.

$$\frac{\tau_{k,t}^2}{\tau_k^{*2}} \ge \frac{1+\sigma_0^2/(K(\sigma_0^2+\sigma^2))}{(1+K^{-1/2}\sigma/\sigma_0)^2}.$$

Proof. From Lemmas 8 & 7, we use the upper bound of $\tau_{k,t}^{*2}$ and the lower bound of $\tau_{k,t}^2$. Then the result is proved. □

H Some Facts

Let X and Y be sub-Gaussian with variance proxies σ^2 and τ^2, respectively. Then (1) aX is sub-Gaussian with variance proxy $a^2\sigma^2$; (2) $X+Y$ is sub-Gaussian with variance proxy $(\sigma+\tau)^2$; and (3) if X and Y are independent, $X+Y$ is sub-Gaussian with variance proxy $\sigma^2+\tau^2$.

I Additional Experiments

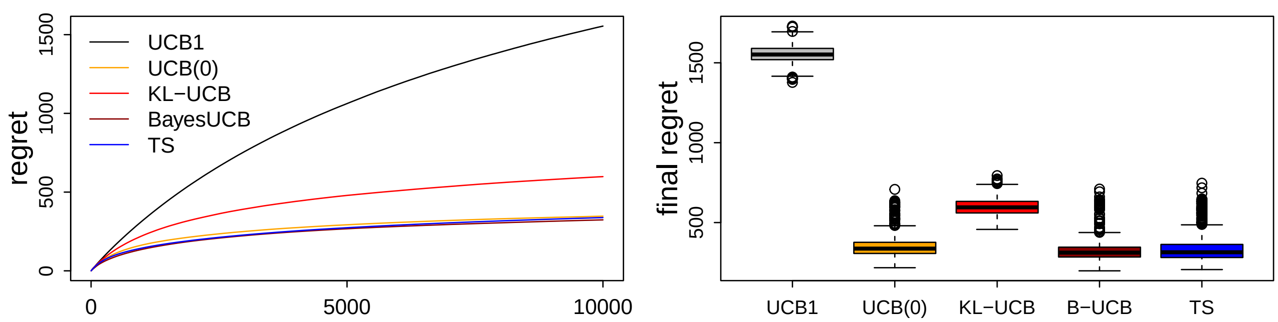


Figure 5: Performance of UCB algorithms on the 50-armed Gaussian bandit with $\mu_k \sim \mathcal{N}(1, 0.04)$, as in Figure 1(a). Left column: Regret performance as a function of round n. Right column: Distribution of the regret at the final round. "UCB(0)" denotes the extreme case of `ReUCB`, i.e., `ReUCB`$^\infty$ by taking $w_k = 1$ that behaves as `UCB1` with $a = 1$. "B-UCB" in the right figure denotes BayesUCB. The results are summarized over 1000 runs.