TS-UCB: Improving on Thompson Sampling With Little to No Additional Computation

Jackie Baek
NYU Stern School of Business

Vivek F. Farias
MIT Sloan School of Management

Abstract

Thompson sampling has become a ubiquitous approach to online decision problems with bandit feedback. The key algorithmic task for Thompson sampling is drawing a sample from the posterior of the optimal action. We propose an alternative arm selection rule we dub TS-UCB, that requires negligible additional computational effort but provides significant performance improvements relative to Thompson sampling. At each step, TS-UCB computes a score for each arm using two ingredients: posterior sample(s) and upper confidence bounds. TS-UCB can be used in any setting where these two quantities are available, and it is flexible in the number of posterior samples it takes as input. TS-UCB achieves materially lower regret on a comprehensive suite of synthetic and real-world datasets, including a personalized article recommendation dataset from Yahoo! and a suite of benchmark datasets from a deep bandit suite proposed in Riquelme et al. (2018). Finally, from a theoretical perspective, we establish optimal regret guarantees for TS-UCB for both the K-armed and linear bandit models.

1 INTRODUCTION

This paper studies the stochastic multi-armed bandit problem, a classical problem modeling sequential decision-making under uncertainty. This problem captures the inherent tradeoff between exploration and exploitation. We study the Bayesian setting, in which we are endowed with an initial prior on the mean reward for each arm.

Thompson sampling (TS) (Thompson, 1933), has in recent years come to be a solution of choice for the multi-armed bandit problem. This popularity stems from the fact that the algorithm performs well empirically (Scott, 2010; Chapelle and Li, 2011) and also admits near-optimal theoretical performance guarantees (Agrawal and Goyal, 2012, 2013b; Kaufmann et al., 2012b; Bubeck and Liu, 2013; Russo and Van Roy, 2014, 2016). Perhaps one of the most attractive features of Thompson sampling though, is the simplicity of the algorithm itself: the key algorithmic task of TS is to sample once from the posterior on arm means, a task that is arguably the simplest thing one can hope to do in a Bayesian formulation of the multi-armed bandit problem.

This Paper: Against the backdrop of Thompson sampling, we propose TS-UCB. Given one or more samples from the posterior on arm means, TS-UCB simply provides a distinct approach to scoring the possible arms. The only additional ingredient this scoring rule relies on is the availability of so-called upper confidence bounds (UCBs) on these arm means.

Now both sampling from a posterior, as well as computing a UCB can be a potentially hard task, especially in the context of bandit models where the payoff from an arm is a complex function of unknown parameters. A canonical example of such a hard problem variant is the contextual bandit problem wherein mean arm reward is given by a complicated function (say, a deep neural network) of the context. Riquelme et al. (2018) provide a recent benchmark comparison of ten different approaches to sampling from an approximate posterior on unknown arm parameters. They show that an approach that chooses to model the uncertainty in *only the last layer* of the neural network defining the mean reward from pulling a given arm at a given context is an effective and robust approach to posterior approximation. In such an approach, not only is (approximate) posterior sampling possible, but UCBs have a closed-form expression and can be easily computed, making possible the use of TS-UCB.

Our Contributions: We show that TS-UCB provides material improvements over Thompson sampling *across the board* on comprehensive sets of synthetic and real-world datasets. The real-world datasets include personalizing news article recommendations for the front page of Yahoo!, and a benchmark set of deep bandit problems studied

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

in Riquelme et al. (2018). Importantly, the performance of TS-UCB either matched or improved upon the state-of-the-art algorithm Information-Directed Sampling (IDS) (Russo and Van Roy, 2018), which requires approximately three orders of magnitude more sampling (and thus compute) than either TS or TS-UCB.

TS-UCB's arm scoring rule can be seen as a modification of the one used in IDS. In particular, TS-UCB essentially replaces the role of the "information gain" term used in IDS to a quantity that is much easier to compute: the radius of the confidence interval. This modification makes TS-UCB orders of magnitude cheaper in terms of computation than IDS, and the experimental results show that this does not come at the cost of any degradation in performance.

Theoretically, we analyze TS-UCB in two specific bandit settings: the K-armed bandit and the linear bandit. In the first setting, there are K independent arms. In the linear bandit, each arm is a vector in $\mathbb{R}^d$, and the rewards are linear in the chosen arm. In both settings, TS-UCB is agnostic to the time horizon. We prove the following Bayes regret bounds for TS-UCB:

For the K-armed bandit, the Bayes regret of TS-UCB *is at most* $O(\sqrt{KT\log T})$.

For the linear bandit of dimension d, the Bayes regret of TS-UCB *is at most* $O(d\log T\sqrt{T})$.

Both of these results match the lower bounds up to log factors. The results are stated formally in Theorems 1 and 2.

1.1 Related Literature

Given the vast literature on bandit algorithms, we restrict our review to literature heavily related to our work, namely, literature on UCB, TS, and methods of applying deep learning models to bandit problems.

The UCB algorithm (Auer et al., 2002) computes an upper confidence bound for every action, and plays the action whose UCB is the highest. In the Bayesian setting, 'Bayes UCB' is defined as the α'th percentile of this distribution, and Kaufmann et al. (2012a) show that using $\alpha = 1 - \frac{1}{t\log^c t}$ achieves the lower bound of Lai and Robbins (1985) for K-armed bandits. For linear bandits, Dani et al. (2008) prove a lower bound of $\Omega(d\sqrt{T})$ for infinite action sets, and the UCB algorithms from Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010); Abbasi-Yadkori et al. (2011) match this up to log factors.

TS, though it was initially proposed in Thompson (1933), has only recently gained a surge of interest, largely influenced by the strong empirical performance of TS demonstrated in Chapelle and Li (2011) and Scott (2010). Since then, many theoretical results on regret bounds for TS have been established (Agrawal and Goyal, 2012, 2013a,b, 2017; Kaufmann et al., 2012b). In the Bayesian setting, Russo and Van Roy (2014) prove a regret bound of $O(\sqrt{KT\log T})$ and $O(d\log T\sqrt{T})$ for TS in the K-armed and linear bandit setting respectively. Bubeck and Liu (2013) improve the regret in the Bayesian K-armed setting to $O(\sqrt{KT})$, and they show this is order-optimal.

The ideas in this paper were heavily influenced by our reading of Russo and Van Roy (2014, 2018). In the former paper, the authors use UCB algorithms as an *analytical* tool to analyze TS. This begs the natural question of whether an appropriate decomposition of regret can provide insight on *algorithmic* modifications that might improve upon TS. Russo and Van Roy (2018) provide such a decomposition and proposes Information Directed Sampling (IDS). IDS has been shown to provide significant performance improvement over TS in some cases, but has heavy sampling (and thus, computational) requirements. The present paper presents yet another decomposition, providing an arm selection rule that does not require additional sampling (i.e. a single sample from the posterior continues to suffice), but nonetheless provides significant improvements over TS while being competitive with IDS.

Zhang (2022) also proposes a modification of TS to handle special information structures by altering the sampling distribution to sample more from parameters that yield a large reward. On the other hand, our work does not change the posterior distribution, and it does not have the same motivation of being able to deal with special information structures.

On the deep learning front, one idea that has been used to sequential decision making problems is to use TS (Riquelme et al., 2018; Lu and Van Roy, 2017; Dwaracherla et al., 2020), since TS requires just a single sample from the posterior. Riquelme et al. (2018) use this idea and evaluates TS on ten different posterior approximation methods for neural networks, ranging from variational methods (Graves, 2011), MCMC methods (Neal, 2012), among others. The authors find that the approach of modeling uncertainty on just the last layer of the neural network (the 'Neural-Linear' approach) (Snoek et al., 2015; Hinton and Salakhutdinov, 2008; Calandra et al., 2016) was overall one of the most effective approaches. This neural linear approach provides not just a tractable approach to approximate posterior sampling, but further provides a tractable UCB for the problem as well. As such, the neural linear approach facilitates the use of the TS-UCB arm selection rule.

2 MODEL

Let $\mathcal{A}$ be a compact set of all possible actions. At time t, an agent is presented with a possibly random subset $\mathcal{A}_t \subseteq \mathcal{A}$ in which they choose an action to play from. If action a is chosen at time t, the agent immediately observes a random reward $R_t(a) \in \mathbb{R}$. For each action

a, the sequence $(R_t(a))_{t\geq 1}$ is i.i.d. and independent of plays of other actions. The mean reward of each action a is $f_\theta(a)$, where $\theta \in \Theta$ is an unknown parameter, and $\{f_\theta : \mathcal{A} \to \mathbb{R} | \theta \in \Theta\}$ is a known set of deterministic functions. That is, $\mathbb{E}[R_t(a)|\theta] = f_\theta(a)$ for all $a \in \mathcal{A}$ and $t \geq 1$.

Let $H_t = (\mathcal{A}_1, A_1, R_1(A_1), \ldots, \mathcal{A}_{t-1}, A_{t-1}, R_{t-1}(A_{t-1}), \mathcal{A}_t)$ denote the history of observations available when the agent is choosing the action for time t, and let $\mathcal{H}$ denote the set of all possible histories. We often refer to H_t as the "state" at time t. A policy $(\pi_t)_{t\geq 1}$ is a deterministic sequence of functions mapping the history to a distribution over actions. An agent employing the policy plays the random action A_t distributed according to $\pi_t(H_t)$, where H_t is the current history. We will often write $\pi_t(a)$ instead of $\pi_t(H_t)(a)$, where $\pi_t(a) = \Pr(A_t = a|H_t)$. Let $A_t^* : \Theta \to \mathcal{A}_t$ be a function satisfying $A_t^*(\theta) \in \operatorname{argmax}_{a\in\mathcal{A}_t} f_\theta(a)$, which represents the optimal action at time t if θ were known. We use A_t^* to denote the random variable $A_t^*(\theta)$, where θ is the true parameter. The T-period regret of policy π is defined as

$$\mathrm{Regret}(T, \pi, \theta) = \sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^*) - f_\theta(A_t)|\theta].$$

We study the Bayesian setting, in which we are endowed with a known prior q on the parameter θ. We take an expectation over this prior to define the T-period Bayes regret

$$\mathrm{BayesRegret}(T, \pi) = \sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^*) - f_\theta(A_t)].$$

We assume that the agent can perform a Bayesian update to their prior at each step after the reward is observed. Let $q(H_t)$ denote to the posterior distribution of θ given the history H_t. In our work, we assume that the agent is able to *sample* from the distribution $q(H_t)$ for any state H_t.

We end this section by describing two concrete bandit models that are the focus of our regret analysis.

K-armed Bandit: In this setting, $\mathcal{A}_t = \mathcal{A} = [K]$ for all t, and each of the entries of the unknown parameter $\theta \in \mathbb{R}^K$ correspond to the mean of each action. That is, $f_\theta(i) = \theta_i$ for every $i \in [K]$. We assume that $\theta_a \in [0, 1]$ for all a, and the rewards $R_t(a)$ are also bounded in $[0, 1]$ for all a and t. The prior distribution q on θ, supported on $[0, 1]^K$, can otherwise be arbitrary.

Linear Bandit: In the linear bandit setting, there is a known vector $X(a) \in \mathbb{R}^d$ associated with each $a \in \mathcal{A}$, and the mean reward takes on the form $f_\theta(a) = \langle \theta, X(a) \rangle$, for $\theta \in \Theta \subseteq \mathbb{R}^d$. We assume that $||\theta||_2 \leq S \leq \sqrt{d}$, $||X(a)|| \leq L$, and $f_\theta(a) \in [-1, 1]$ for all $a \in \mathcal{A}$. Lastly, we assume that $R_t(a) - f_\theta(a)$ is r-sub-Gaussian for every t and a for some $r \geq 1$. All of these assumptions are standard and are the same as in Abbasi-Yadkori et al. (2011).

A special case of linear bandits is contextual linear bandits in which there are K arms and there is an unknown parameter $\beta_k \in \mathbb{R}^d$ for each arm $k \in [K]$. A random context $X_t \in \mathbb{R}^d$ is observed at the start of each time step, and the mean reward for arm k at time t is $\langle \beta_k, X_t \rangle$. This is equivalent to a linear bandit problem of dimension dk where the action set (transposed) at time t is $\{(X_t^\top, 0_d, \ldots, 0_d), (0_d, X_t^\top, \ldots, 0_d), \ldots, (0_d, \ldots, X_t^\top)\}$, where 0_d is the transposed 0-vector of dimension d, and the unknown parameter is $\theta^\top = (\beta_1^\top, \ldots, \beta_K^\top)$.

3 ALGORITHM

TS-UCB requires a set of functions $U, \hat{\mu} : \mathcal{H} \times \mathcal{A} \to \mathbb{R}$ to first be specified, where $U(h, a)$ represents the upper confidence bound of action a at history h, and $\hat{\mu}(h, a)$ represents an estimate of $f_\theta(a)$ at history h. We require that $U(h, a) - \hat{\mu}(h, a) > 0$ on every input. We write $U_t(a) = U(H_t, a)$ and $\hat{\mu}_t(a) = \hat{\mu}(H_t, a)$, and we refer to the quantity $\mathrm{radius}_t(a) \triangleq U_t(a) - \hat{\mu}_t(a)$ as the *radius* of the confidence interval.

TS-UCB proceeds as follows. At state H_t, draw m independent samples from the posterior distribution $q(H_t)$, for some integer parameter $m \geq 1$. Denote these samples by $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$, and let $\tilde{f}_i = f_{\tilde{\theta}_i}(A_t^*(\tilde{\theta}_i))$ be the mean reward of the best arm when the true parameter is $\tilde{\theta}_i$. (Conditioned on H_t, the distribution of $\tilde{f}_i$ is the same as the distribution of $f_\theta(A^*)$.) Let $\tilde{f}_t = \frac{1}{m}\sum_{i=1}^m \tilde{f}_i$. For every action a, define the ratio $\Psi_t(a)$ as

$$\Psi_t(a) \triangleq \frac{\tilde{f}_t - \hat{\mu}_t(a)}{U_t(a) - \hat{\mu}_t(a)} = \frac{\tilde{f}_t - \hat{\mu}_t(a)}{\mathrm{radius}_t(a)}. \tag{1}$$

TS-UCB chooses an action that minimizes this ratio, which we assume exists.[1] That is, if $A_t^{\text{TS-UCB}}$ is the random variable for the action chosen by TS-UCB at time t, then,

$$A_t^{\text{TS-UCB}} \in \operatorname*{argmin}_{a\in\mathcal{A}_t} \Psi_t(a). \tag{2}$$

We parse the ratio $\Psi_t(a)$: $\hat{\mu}_t(a)$ is an estimate of the expected reward $\mathbb{E}[f_\theta(a)|H_t]$ from playing action a, and $\tilde{f}_t$ is an estimate of the optimal reward $\mathbb{E}[f_\theta(A^*)|H_t]$ (indeed, $\tilde{f}_t \to \mathbb{E}[f_\theta(A^*)|H_t]$ as $m \to \infty$). Then, the numerator of the ratio estimates the expected instantaneous regret from playing action a. We clearly want this to be small, but minimizing only the numerator would result in the greedy policy. The denominator enforces exploration by favoring actions with larger confidence intervals, corresponding to actions in which not much information is known about. This ratio is similar to the information ratio that is minimized in the IDS algorithm — the main difference is that the denominator in IDS is information gain.

[1] Clearly it exists if $\mathcal{A}$ is finite. Otherwise, since $\mathcal{A}$ is assumed to be compact, it exists if $\hat{\mu}_t$ and U_t are continuous functions.

Comparison to TS, IDS, and UCB. TS-UCB can be derived from optimizing a Bayes regret analysis of TS. In the regret analysis in Section 5, a crucial step is to upper bound the ratio $\Psi_t(a)$. We show an upper bound for $\Psi_t(a)$ under *TS*, while TS-UCB *optimizes* this quantity by definition. Second, TS-UCB can be thought of as an approximation to IDS, where the usual denominator of information gain is replaced with a confidence radius. In many settings, the information gain and confidence radius both represent uncertainty, where the latter is much simpler to compute. This is not the case in all settings; we note that TS-UCB does not have the same purpose of IDS of being able to exploit complex information structures. Lastly, TS-UCB can also be thought of as a UCB policy whose tuning parameter is *automatically* and *dynamically* adjusted — see Appendix A for a detailed description of this interpretation.

TS-UCB can be applied whenever the quantities $\tilde{f}_t = \frac{1}{m}\sum_{i=1}^{m} \tilde{f}_i$ and $\{U_t(a), \hat{\mu}_t(a)\}_{a\in\mathcal{A}}$ can be computed, which are exactly the quantities needed for TS ($m = 1$) and UCB respectively. The following example shows that TS-UCB can be applied in a general setting where the relationship between actions and rewards is modeled using a deep neural network.

Example 1 (Neural Linear (Riquelme et al., 2018))**.** *Consider a contextual bandit problem where a context $X_t \in \mathbb{R}^{d'}$ arrives at each time step, and the expected reward of taking action $a \in \mathcal{A}$ is $g(X_t, a)$, for an unknown function g. The 'Neural Linear' method models uncertainty in only the last layer of the network by considering a specific class of functions g. Specifically, consider that g allows the decomposition $g(X_t, a) = h(X_t)^\top \beta_a$ where $h(X_t) \in \mathbb{R}^d$ represent the outputs from the last layer of some neural network and $\beta_a \in \mathbb{R}^d$ is some parameter vector. If the function $h(\cdot)$ were known, then the resulting problem is a linear bandit problem for which both sampling from the posterior on β_a for all $a \in \mathcal{A}$ as well as computing a (closed form) UCB on β_a are easy. In reality $h(\cdot)$ is unknown but the Neural Linear method approximates this quantity from past observations and ignores uncertainty in the estimate. As such, it is clear that TS-UCB can be used as an alternative to TS in the Neural Linear approach.*

We now apply TS-UCB for the K-armed bandit and linear bandit using the standard definitions of upper confidence bounds found in the literature, and we formally state the main theorems.

3.1 K-armed Bandit

We assume $T \geq K$, and we pull every arm once in the first K time steps. Let $N_t(a) = \sum_{s=1}^{t-1} \mathbb{1}(A_s = a)$ be the number of times that action a was played up to but not including time t. We define the upper confidence bounds in a similar way to Auer et al. (2002); namely,

$$\hat{\mu}_t(a) \triangleq \frac{1}{N_t(a)} \sum_{s=1}^{t-1} \mathbb{1}(A_s = a) R_s(a) \tag{3}$$

$$U_t(a) \triangleq \hat{\mu}_t(a) + \sqrt{3\log T / N_t(a)}. \tag{4}$$

This implies $\text{radius}_t(a) = \sqrt{3\log T / N_t(a)}$. Because the term $\sqrt{3\log T}$ appears as a multiplicative factor in the radius and the same term is used for all actions and time steps, the algorithm is agnostic to this value. That is, TS-UCB reduces to picking the action which minimizes

$$\sqrt{N_t(a)}(\tilde{f}_t - \hat{\mu}_t(a)). \tag{5}$$

This implies that TS-UCB does not have to know the time horizon T a priori. We now state our main result for this setting.

Theorem 1. *For the K-armed bandit, using the UCBs as defined in* (4)*,*

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) = O(\sqrt{KT\log T}). \tag{6}$$

This result matches the $\Omega(\sqrt{KT})$ lower bound Bubeck and Liu (2013) up to a logarithmic factor.

3.2 Linear Bandit.

For the linear bandit, to define the functions $\hat{\mu}_t$ and U_t, we first need to define a confidence set $C_t \subseteq \Theta$, which contains θ with high probability. We use the confidence sets developed in Abbasi-Yadkori et al. (2011). Let $X_t = X(A_t)$ be the vector associated with the action played at time t. Let $\mathbf{X}_t$ be the $t \times d$ matrix whose s'th row is $X_s^\top$. Let $\mathbf{Y}_t \in \mathbb{R}^t$ be the vector of rewards seen up to and including time t. At time t, define the positive semi-definite matrix $V_t = I + \sum_{s=1}^{t} X_s X_s^\top = I + \mathbf{X}_t^\top \mathbf{X}_t$, and construct the estimate $\hat{\theta}_t = V_t^{-1}\mathbf{X}_t^\top \mathbf{Y}_t$. Using the notation $||x||_A = \sqrt{x^\top A x}$, let $C_t = \{\rho : ||\rho - \hat{\theta}_t||_{V_t} \leq \sqrt{\beta_t}\}$, where $\sqrt{\beta_t} = r\sqrt{d\log(T^2(1+tL))} + S$.

Using this confidence set, the functions needed for TS-UCB are defined as

$$\hat{\mu}_t(a) \triangleq \langle X(a), \hat{\theta}_t \rangle \qquad U_t(a) \triangleq \max_{\rho \in C_t} \langle X(a), \rho \rangle. \tag{7}$$

Since $U_t(a)$ is the solution to maximizing a linear function subject to an ellipsoidal constraint, it has a closed form solution: $U_t(a) = \langle X(a), \hat{\theta}_t \rangle + \sqrt{\beta_t}||X(a)||_{V_t^{-1}}$, which implies $\text{radius}_t(a) = \sqrt{\beta_t}||X(a)||_{V_t^{-1}}$. Then, TS-UCB reduces to picking the action which minimizes

$$\frac{\tilde{f}_t - \langle X(a), \hat{\theta}_t \rangle}{||X(a)||_{V_t^{-1}}}.$$

Note that the $\sqrt{\beta_t}$ term disappears, implying TS-UCB does not depend on the exact expression of this term. Like

the K-armed bandit, the algorithm does not have to know the time horizon T a priori.

We state our main result for this setting.

Theorem 2. *For the linear bandit, using the UCBs as defined in* (7)*, if* $||X(a)||_2 = 1$ *for all* $a \in \mathcal{A}$*,*

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) = O(d \log T \sqrt{T}). \quad (8)$$

This result matches the $\Omega(d\sqrt{T})$ lower bound (Dani et al., 2008) up to a logarithmic factor. We believe the additional assumption that $||X(a)||_2 = 1$ is an artifact our proof, which we believe can be likely removed with a more refined analysis. We note that TS and IDS has been shown to achieve a regret of $O(\sqrt{dT \log(|\mathcal{A}|)})$ (Russo and Van Roy, 2016, 2018), which is dependent on the total number of actions $|\mathcal{A}|$.

We give an outline of the proofs of Theorem 1 and 2 in Section 5, and we provide the full proof in the Appendix.

4 COMPUTATIONAL RESULTS

We conduct three sets of experiments for the contextual bandit problem. The first set is entirely synthetic for an ensemble of linear contextual bandit problems where exact posterior samples (and a regret analysis) are available for all methods considered. Our objective here is to understand the level of improvement TS-UCB can provide over TS and how the level of this improvement depends on natural problem features such as the number of arms and the level of noise. The next two experiments are on real-world datasets, where the exact Bayesian structure is not available. The second set of experiments studies personalizing news article recommendations on the front page of the Yahoo! website, and the last set of experiments considers a deep bandit benchmark on seven different real-world datasets. Our goal is to show that TS-UCB provides state of the art performance while being computationally cheap and robust to prior misspecification. In all experiments, we compare TS-UCB to TS, UCB, Greedy, and IDS.

4.1 Synthetic Experiments

First, we simulate synthetic instances of the linear contextual bandit with varying number of actions and size of the prior covariance. Let d be the dimension and K be the number of actions. For each action $k \in [K]$, we independently sample $\beta_k \sim N(0, I_d)$, where I_d is the d-dimensional identity matrix. At each time t, a context X_t is drawn i.i.d. from $N(0, \frac{1}{d} I_d)$. The reward for arm k at time t is $\langle \beta_k, X_t \rangle + \epsilon_t$, where ϵ_t is drawn i.i.d. from $N(0, \sigma^2)$. We set $d = 10$ and vary the number of actions as $K \in \{3, 5, 10, 20, 60\}$. We also vary the magnitude of the noise as $\sigma \in \{0.05, 0.1, 0.5, 1, 2\}$, which results in a total of 25 instances.

We run the following algorithms:

- **TS**: We run Thompson Sampling as our baseline algorithm. All results are stated relative to the performance of TS.
- **TS-UCB**: We run our algorithm as defined in Section 3.2 with $m = 1$ and $m = 100$, denoted as TS-UCB(1) and TS-UCB(100) respectively.
- **Greedy**: We pull the arm that maximizes $\langle \hat{\beta}_k, X_t \rangle$, where $\hat{\beta}_k$ is the posterior mean of β_k.
- **UCB**: We pull the arm with the highest $U_t(a)$ as defined in Section 3.2 (this is the OFUL algorithm of Abbasi-Yadkori et al. (2011)).
- **IDS**: We run the sample variance-based IDS (Algorithm 6 from Russo and Van Roy (2018)) with $m = 1000$ samples.

All algorithms are given knowledge of the prior for the parameters and the variance of the noise, σ^2, so that correct posteriors can be computed if the algorithm requires it.

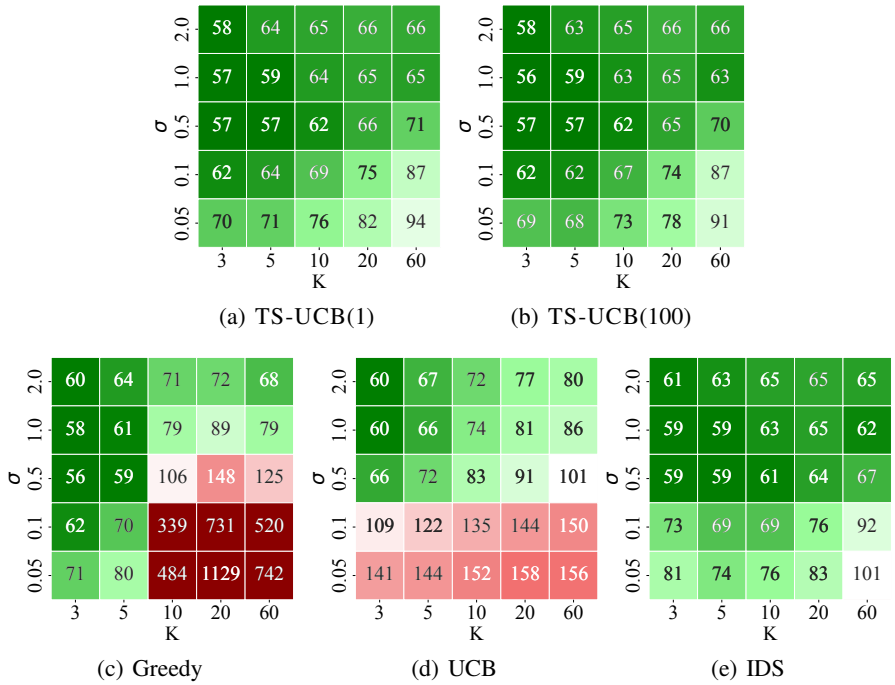
For each algorithm and problem instance, we simulate 200 runs over a time horizon of $T = 10,000$. We report the average regret as a percentage of the regret from the TS policy (that is, we estimate $100 \cdot \mathbb{E}\left[\frac{\text{Regret(ALG)}}{\text{Regret(TS)}}\right]$). The results are shown in Figure 1.

Results. We see that both TS-UCB(1) and TS-UCB(100) outperforms TS across the board, almost halving regret in many instances. The general trend is that TS-UCB has a greater performance improvement over TS when σ is higher, which correspond to the "harder" instances. Over the 25 instances, the regret from TS-UCB(1) and TS-UCB(100) was 67.9% and 66.6% of the regret of TS respectively.

We see that TS-UCB(100) performs better than TS-UCB(1) overall. However, this improvement is small relative to the performance gain over TS; most of the benefit of TS-UCB is captured by using just a single sample.

Both greedy and UCB have inconsistent performances relative to TS. The greedy algorithm outperforms TS and performs similarly to TS-UCB when the number of actions is small. This is consistent with the result of Bastani et al. (2020), which prove greedy is rate optimal when $K = 2$ and the contexts are diverse. However, when K is large, context diversity becomes insufficient to guarantee enough exploration for every arm, resulting in poor performance. For UCB, we see that performance is poor when the noise is small; this is likely due to UCB being too conservative.

IDS performs well across the board compared to TS. Its performance is similar to TS-UCB but slightly worse on



(a) TS-UCB(1) (b) TS-UCB(100)

(c) Greedy (d) UCB (e) IDS

Figure 1: TS-UCB improves on TS across the board. Grid reports mean regret of each policy as a percentage of regret of Thompson Sampling over 200 runs. A number smaller than 100 means the regret of that policy is smaller than TS; otherwise the regret is larger than TS. TS-UCB(m) refers to the algorithm using m samples.

average — across the 25 instances, regret for IDS was 5.6% higher than TS-UCB(100) and 3.8% higher than TS-UCB(1). IDS was expected perform well, as it is considered the state-of-the-art algorithm. However, we see that a much simpler algorithm, TS-UCB, performs as well, and often better, than IDS.

4.2 Personalized News Article Recommendation

We test the same set of bandit algorithms on a real-world dataset for personalizing news article recommendations for users that land on the front page of the Yahoo! website. In this setting, when a user goes on the website, the website must choose one article to recommend out of a pool of available articles at that time, in which the user may click on the article to read the full story. The pool of available articles changes throughout the day. The goal is to recommend articles that maximize the click-through rate.

We use a dataset that was generated by an experiment done by Yahoo! from 10 days in October 2011, made available through the Yahoo Webscope Program[2]. In the experiment, when a user appeared, the article that was recommended was chosen uniformly at random out of all available at the time, and whether the user clicked on the article was logged. Each of these users is associated with a context vector of dimension $d = 136$ that corresponds to user covariates such as gender and age. Refer to the Appendix for further details on the simulation setup. We report the regret of each policy as a percentage of the regret of TS, shown in Table 1.

Results. TS-UCB(100) performed the best overall, having the lowest average regret in 7 out of 10 days. We see that TS-UCB(1), TS-UCB(100) and IDS significantly outperform TS in all instances — reducing regret by more than 10% on average. The relative performance of these three algorithms are comparable, where TS-UCB(100) slightly outperforms IDS, and IDS slightly outperforms TS-UCB(1). Greedy performs similarly to TS, while UCB is clearly outperformed by TS. Overall, we see a similar pattern in performance as compared to the synthetic experiments; TS-UCB clearly outperforms TS, and moreover, often outperforms IDS while being much cheaper than IDS computationally.

4.3 Deep Bandit Benchmark

In challenging bandit models such as the deep contextual bandit discussed in Example 1, computing exact posteriors are difficult. Riquelme et al. (2018) evaluate a large number of posterior approximation methods on a variety of real-world datasets for such a contextual bandit problem. Their results suggest that performing posterior sampling using the "Neural Linear" method, described in Example 1, is an effective and robust approach. We evaluate TS-UCB

[2] `https://webscope.sandbox.yahoo.com/`

Table 1: Yahoo! article recommendation simulation results from 10 days in October 2011. TS-UCB provides an improvement over TS across the board. For each policy, we report the regret as a percentage of regret of Thompson Sampling (with 95% confidence intervals) for that approach. For each day, the policy with the lowest average regret is bolded. IDS requires one thousand samples from the posterior at each time step; TS-UCB(1) and TS-UCB(100) requires one and one hundred samples respectively.

Day	TS-UCB(1)	TS-UCB(100)	IDS	Greedy	UCB
1	91.0 ± 1.5	$\mathbf{89.1 \pm 1.6}$	90.0 ± 1.5	106.1 ± 6.2	106.2 ± 1.1
2	86.3 ± 1.3	$\mathbf{82.2 \pm 1.9}$	84.8 ± 2.2	100.6 ± 9.8	121.9 ± 1.7
3	85.8 ± 1.9	$\mathbf{84.1 \pm 1.4}$	84.8 ± 1.7	122.3 ± 7.0	124.9 ± 1.6
4	92.5 ± 1.7	91.6 ± 2.0	$\mathbf{90.9 \pm 1.6}$	107.0 ± 6.2	123.8 ± 2.1
5	91.1 ± 1.8	$\mathbf{89.8 \pm 1.7}$	90.9 ± 1.7	100.7 ± 3.2	110.7 ± 1.4
6	85.1 ± 1.4	83.7 ± 0.7	$\mathbf{83.2 \pm 1.2}$	105.6 ± 4.4	102.2 ± 1.1
7	96.2 ± 1.5	96.3 ± 1.9	94.0 ± 2.2	$\mathbf{88.5 \pm 7.3}$	121.8 ± 1.2
8	90.7 ± 2.4	$\mathbf{89.5 \pm 2.1}$	90.0 ± 2.4	106.9 ± 4.3	119.8 ± 2.0
9	92.3 ± 1.7	$\mathbf{88.8 \pm 2.4}$	90.4 ± 2.0	92.4 ± 7.7	116.4 ± 2.1
10	88.1 ± 1.9	$\mathbf{86.4 \pm 3.0}$	86.7 ± 1.3	93.1 ± 5.9	122.8 ± 1.6
Overall	89.9 ± 0.7	$\mathbf{88.2 \pm 0.8}$	88.6 ± 0.7	102.3 ± 2.4	117.1 ± 1.2

on the benchmark problems in Riquelme et al. (2018) and compare its performance to TS, IDS, greedy, and UCB.

For a finite action set of size K, Neural Linear maintains one neural network, $h_t : \mathbb{R}^{d'} \to \mathbb{R}^d$, as well as posterior distributions on K parameter vectors $\beta_a \in \mathbb{R}^d$ and K scalar parameters $\sigma_a^2 \in \mathbb{R}$. At time t, the posteriors on β_a and σ_a^2 are computed *ignoring the uncertainty* in the estimate of $h_t(\cdot)$ so that this computation is equivalent to bayesian linear regression. Specifically, we assume a linear contextual bandit model on the context *representation* $h_t(X)$.

While the original dimension, d', varies across datasets, the last layer of the neural network has same dimension $d = 50$ for every dataset. We use a neural network with two fully connected layers of size 50 for $h_t(\cdot)$. The network is updated every 50 time steps, in which the network minimizes mean squared error for the observed rewards using the RMSProp optimizer (Hinton et al., 2012).

We replicate the experiments from Riquelme et al. (2018) with the same real-world datasets. These datasets vary widely in their properties; see Appendix A of Riquelme et al. (2018) for the details of each dataset. We simulate 200 runs for each dataset and algorithm. For each dataset, one "run" is defined as 10,000 data points randomly drawn from the entire dataset; that is, there are 10,000 time steps[3], and each data point (or "context") arrives sequentially in a random order. We report the regret of each policy as a percentage of the regret of TS, shown in Table 2.

Results. Riquelme et al. (2018) establish TS along with the neural linear approach to posterior sampling as a benchmark algorithm for deep contextual bandits. We see here that TS-UCB improves upon TS on every dataset except possibly mushroom, and it offers significant improvements in datasets financial and statlog. Similarly to the synthetic experiments, TS-UCB(100) always outperforms TS-UCB(1).

The performance of the other algorithms relative to both TS and TS-UCB is also similar to the results of the synthetic experiments. IDS usually outperforms TS, and has a similar performance to TS-UCB but slightly worse in some cases. On average, the regret for IDS was 4.1% higher than TS-UCB(100) and 0.02% higher than TS-UCB(1). Greedy has a very inconsistent performance across datasets. It outperforms all other algorithms in three datasets (census, covertype, jester), suggesting that no exploration is needed in these cases. However, its poor performance in mushroom and statlog suggest that exploration is indeed necessary in several real-world settings. UCB is consistently outperformed by both TS and TS-UCB.

In summary, both the synthetic and real-world experiments suggest the same conclusion: TS-UCB outperforms TS across a comprehensive suite of experiments with essentially no additional computation. Moreover, TS-UCB consistently matches or improves upon the state-of-the-art algorithm IDS, while being a much simpler algorithm than IDS both computationally and conceptually.

5 OUTLINE OF REGRET ANALYSIS

In this section, we give an outline of the proofs of Theorem 1 and 2. The full proofs can be found in the Appendix.

We first state two known results results on upper bounding $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t)]$ for the two bandit settings that follow from standard UCB analyses. For the K-armed setting, the proof of Proposition 2 of Russo and Van Roy (2014) implies the following result.

[3]The financial dataset did not have 10,000 data points, so we used $3,000$ data points for this dataset only.

Table 2: Deep Bandit benchmark Riquelme et al. (2018) results for the Neural Linear posterior approximation method. For each posterior approximation approach, the regret is reported as a percentage of regret of Thompson Sampling (with 95% confidence intervals) for that approach. For each dataset, the policy with the lowest average regret is bolded.

Dataset	d'	K	TS-UCB(1)	TS-UCB(100)	IDS	Greedy	UCB
adult	14	86	98.8 ± 0.2	$\mathbf{98.6 \pm 0.2}$	$\mathbf{98.6 \pm 0.2}$	103.4 ± 0.7	101.0 ± 0.2
census	369	9	99.4 ± 0.5	99.2 ± 0.5	99.2 ± 0.5	$\mathbf{92.2 \pm 0.5}$	105.2 ± 0.5
covertype	54	7	98.7 ± 0.6	98.6 ± 0.6	98.4 ± 0.6	$\mathbf{91.5 \pm 0.6}$	110.3 ± 0.5
financial	21	8	60.0 ± 0.9	$\mathbf{54.7 \pm 0.7}$	56.3 ± 0.8	101.6 ± 9.8	160.7 ± 2.3
jester	32	8	99.4 ± 0.3	99.2 ± 0.3	99.6 ± 0.3	$\mathbf{96.6 \pm 0.4}$	113.1 ± 0.8
mushroom	117	2	108.0 ± 11.1	$\mathbf{98.6 \pm 5.1}$	118.6 ± 15.7	189.5 ± 32.4	918.3 ± 53.0
statlog	9	7	89.7 ± 0.6	74.9 ± 0.6	$\mathbf{73.9 \pm 0.6}$	317.5 ± 27.8	322.9 ± 2.0

Theorem 3. *For the K-armed bandit, using the UCBs as defined in* (4)*, $\sum_{t=K+1}^{T} \mathbb{E}[\text{radius}_t(A_t)] \leq 2\sqrt{3KT \log T}$, for any sequence of actions A_t.*

Similarly, in the linear bandit setting, the proof of Theorem 3 of Abbasi-Yadkori et al. (2011) (using the parameters $\delta = T^{-3}, \lambda = 1$) implies the following result.

Theorem 4. *For the linear bandit, using the UCBs as defined in* (7)*, for any sequence of actions A_t, $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t)] = O(d \log T \sqrt{T})$.*

Next, it is useful to extend the definition of Ψ_t to randomized actions. If ν is a probability distribution over $\mathcal{A}$, define

$$\bar{\Psi}_t(\nu) \triangleq \frac{\tilde{f}_t - \mathbb{E}_{A_t \sim \nu}[\hat{\mu}_t(A_t)]}{\mathbb{E}_{A_t \sim \nu}[\text{radius}_t(A_t)]}. \tag{9}$$

Using this definition, we show (Lemma 2 in the Appendix) that for any policy $(\pi_t)_{t \geq 1}$, surely,

$$\Psi_t(A_t^{\text{TS-UCB}}) \leq \bar{\Psi}_t(\pi_t). \tag{10}$$

Now, suppose the following two approximations hold at every time step (a rigorous version of these approximations are used in the proof in the Appendix):

(i) $\tilde{f}_t$ approximates the expected optimal reward: $\tilde{f}_t \approx \mathbb{E}[f_\theta(A^*)|H_t]$.

(ii) $\hat{\mu}_t(a)$ approximates the expected reward of action a: $\hat{\mu}_t(a) \approx \mathbb{E}[f_\theta(a)|H_t]$.

The Bayes regret for TS-UCB can be decomposed as

$$\begin{aligned}
&\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \\
&= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[f_\theta(A_t^*) - f_\theta(A_t^{\text{TS-UCB}})|H_t]] \\
&\approx \sum_{t=1}^{T} \mathbb{E}[\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}})] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(A_t^{\text{TS-UCB}}) \text{radius}_t(A_t^{\text{TS-UCB}})\right], \qquad (11)
\end{aligned}$$

where the second step uses (i)-(ii), and the third step uses the definition (1).

(11) decomposes the regret into the product of two terms: the ratio $\Psi_t(A_t^{\text{TS-UCB}})$ and the radius of the action taken. For the second piece, standard analyses for the UCB algorithm found in the literature bound regret by bounding the sum $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t)]$ for any sequence of actions A_t. Therefore, if $\Psi_t(A_t^{\text{TS-UCB}})$ can be upper bounded by a constant, the regret bounds found for UCB can be directly applied.

We show $\bar{\Psi}_t(\pi_t^{\text{TS}}) \lessapprox 1$, where TS is the Thompson Sampling policy (this is stated formally and shown in Lemma 3 in the Appendix). In light of (10), this implies $\Psi_t(A_t^{\text{TS-UCB}}) \lessapprox 1$. Plugging this back into (11) gives us $\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \lessapprox \sum_{t=1}^{T} \mathbb{E}\left[\text{radius}_t(A_t^{\text{TS-UCB}})\right]$, which lets us apply UCB regret bounds from the literature and finishes the proof.

This method of decomposing the regret into the product of two terms (as in (11)) and minimizing one of them was used in Russo and Van Roy (2018) for the IDS policy. The optimization problem in IDS is difficult, as the term that is minimized involves evaluating the information gain, requiring computing integrals over high-dimensional spaces. The optimization problem for TS-UCB is almost trivial, but it trades off on the ability to incorporate complicated information structures as IDS can.

The above proof outline can be used to prove the following proposition.

Proposition 1. *Suppose $\text{radius}_t(a) \in [r_{\min}, r_{\max}]$ for all $a \in \mathcal{A}$ and $t \geq 1$. Using the UCBs as defined in* (4) *for the K-armed bandit, and* (7) *for the linear bandit,*

$$\begin{aligned}
\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) &\leq 2 \sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] \\
&+ \frac{r_{\max}}{r_{\min}} \left(1 + \frac{2T}{\sqrt{m}}\right) + T^{-2}. \qquad (12)
\end{aligned}$$

The approximation $\tilde{f}_t \approx \mathbb{E}[f_\theta(A^*)|H_t]$ used in the proof

sketch only holds when m is large; the fact that this doesn't hold contributes to the $\frac{1}{\sqrt{m}}$ term in (12), which goes to zero as $m \to \infty$. To cover the case when m is small, we also show the following proposition, which has the opposite relationship with respect to m.

Proposition 2. *Using the UCBs as defined in* (4) *for the* K*-armed bandit, and* (7) *for the linear bandit,*

$$\begin{aligned}&\mathrm{BayesRegret}(T, \pi^{\text{TS-UCB}}) \\ &\leq 2\sum_{t=1}^{T} \mathbb{E}[\mathrm{radius}_t(A_t^{\text{TS-UCB}})] + (m+1)T^{-2}.\end{aligned}$$

The main results follow from combining the above two propositions with the known bounds of Theorems 3 and 4. The formal proofs of the theorems can be found in the Appendix.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions. Both authors were partially supported by NSF Grant CMMI 1727239.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1.

Agrawal, S. and Goyal, N. (2013a). Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107.

Agrawal, S. and Goyal, N. (2013b). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24.

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

Bastani, H., Bayati, M., and Khosravi, K. (2020). Mostly exploration-free algorithms for contextual bandits. *Management Science*.

Bubeck, S. and Liu, C.-Y. (2013). Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pages 638–646.

Calandra, R., Peters, J., Rasmussen, C. E., and Deisenroth, M. P. (2016). Manifold gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345. IEEE.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Dwaracherla, V., Lu, X., Ibrahimi, M., Osband, I., Wen, Z., and Van Roy, B. (2020). Hypermodels for exploration. In *International Conference on Learning Representations*.

Graves, A. (2011). Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356.

Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8).

Hinton, G. E. and Salakhutdinov, R. R. (2008). Using deep belief nets to learn covariance kernels for gaussian processes. In *Advances in neural information processing systems*, pages 1249–1256.

Kaufmann, E., Cappé, O., and Garivier, A. (2012a). On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600.

Kaufmann, E., Korda, N., and Munos, R. (2012b). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer.

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

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

Lu, X. and Van Roy, B. (2017). Ensemble sampling. In *Advances in neural information processing systems*, pages 3258–3266.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Riquelme, C., Tucker, G., and Snoek, J. (2018). Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*.

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

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

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Russo, D. and Van Roy, B. (2018). Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180.

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

Yahoo! (2020). Yahoo! webscope program. `https://webscope.sandbox.yahoo.com/`. Accessed: 2020-11-28.

Zhang, T. (2022). Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857.

A INTERPRETATION OF TS-UCB AS A DYNAMIC UCB ALGORITHM

Recall that a UCB algorithm chooses the arm with the highest upper confidence bound. Often, the radius of the confidence interval takes the form $\alpha \cdot \text{radius}_t(a)$, where α is a scalar parameter. For example, the UCB1 algorithm of Auer et al. (2002) plays the arm that maximizes $\hat{\mu}_t(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$; here we can think of $\alpha = \sqrt{2}$. It is well known that tuning this parameter can vastly improve empirical performance (Russo and Van Roy, 2014).

TS-UCB can be interpreted as a UCB algorithm whose α parameter is dynamically tuned. TS-UCB plays the arm with the highest $\hat{\mu}_t(a) + \alpha_t \cdot \text{radius}_t(a)$, where

$$\alpha_t = \min\{\alpha : \max_{a \in \mathcal{A}} \{\hat{\mu}_t(a) + \alpha \cdot \text{radius}_t(a)\} \geq \tilde{f}_t\}. \tag{13}$$

That is, after sampling $\tilde{f}_t$, α_t is the smallest α such that there exists an arm whose UCB, $\hat{\mu}_t(a) + \alpha \cdot \text{radius}_t(a)$, is at least as large as $\tilde{f}_t$. This ends up being equivalent to setting $\alpha_t = \Psi_t(A_t^{\text{TS-UCB}})$. Indeed, for any $a \in \mathcal{A}_t$, since $\Psi_t(A_t^{\text{TS-UCB}}) \leq \Psi_t(a)$ by definition of the algorithm, we have

$$\hat{\mu}_t(a) + \Psi_t(A_t^{\text{TS-UCB}}) \cdot \text{radius}_t(a) \leq \hat{\mu}_t(a) + \Psi_t(a) \cdot \text{radius}_t(a) = \tilde{f}_t,$$

where the inequality is an equality if and only if $a \in \text{argmin}_{a \in \mathcal{A}} \Psi_t(a)$. In other words, for the action that TS-UCB chooses, its (dynamically tuned) UCB is exactly $\tilde{f}_t$; for other actions, their UCB is smaller. In this sense, TS-UCB is a method of *automatically* (since the parameter α_t is adjusted using posterior samples) and *dynamically* (since α_t changes with t) tuning the UCB algorithm.

B REGRET ANALYSIS

For our analysis, we introduce lower confidence bounds $(L_t)_{t \geq 1}$, which we define in a symmetric way to upper confidence bounds: $L_t(a) \triangleq \hat{\mu}_t(a) - (U_t(a) - \hat{\mu}_t(a))$.

We first state a lemma that says that the confidence bounds are valid with high probability

Lemma 1. *Using the functions $\{\hat{\mu}_t\}_{t \geq 1}$, $\{U_t\}_{t \geq 1}$ as defined in* (4) *in the K-armed setting and* (7) *in the linear bandit setting, for any $t \leq T$, $\Pr(f_\theta(A) < U_t(A)) \leq T^{-3}$, where A is any deterministic or random action. The analogous bounds hold for lower confidence bounds, i.e. $\Pr(f_\theta(A) > L_t(A)) \leq T^{-3}$.*

For completeness, the proof of Lemma 1 can be found in B.4. The following corollary is immediate using the law of total expectation and the fact that $f_\theta(A) \geq -1$.

Corollary 1. *For any $t \leq T$, $\mathbb{E}[-f_\theta(A)] \leq \mathbb{E}[-L_t(A)] + T^{-3}$, where A is any deterministic or random action.*

The next two subsections prove Proposition 1 and 2 respectively. The final step of the proof combines these propositions with the known bounds for $\sum_{t=1}^T \mathbb{E}[\text{radius}_t(A_t)]$ from Theorems 3 and 4, and can be found in Appendix B.3.

B.1 Proof of Proposition 1.

We first state the result claimed in (10) whose proof is deferred to Appendix B.4.

Lemma 2. *For any distribution τ over $\mathcal{A}_t$, $\Psi_t(A_t^{\text{TS-UCB}}) \leq \bar{\Psi}_t(\tau)$ almost surely.*

Next, we upper bound the ratio $\Psi_t(A_t^{\text{TS-UCB}})$ by analyzing the Thompson Sampling policy.

Lemma 3. $\Psi_t(A_t^{\text{TS-UCB}}) \leq 1 + \frac{1}{r_{\min}}(\Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])$ *almost surely. Equivalently, using* (1)*,*

$$\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}}) \leq \text{radius}_t(A_t^{\text{TS-UCB}})(1 + \frac{1}{r_{\min}}(\Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])). \tag{14}$$

Proof. Let π^{TS} be the Thompson sampling policy. We show the inequality for $\bar{\Psi}_t(\pi_t^{\text{TS}})$ instead, and then use $\Psi_t(A_t^{\text{TS-UCB}}) \leq \bar{\Psi}_t(\pi_t^{\text{TS}})$ from Lemma 2 to get the desired result.

By definition of TS, $\pi_t^{\text{TS}} = \pi_t^{\text{TS}}(H_t)$ is the distribution over $\mathcal{A}_t$ corresponding to the posterior distribution of A^* conditioned on H_t. Then, if A_t is the action chosen by TS at time t, we have $\mathbb{E}[U_t(A_t)|H_t] = \mathbb{E}[U_t(A^*)|H_t]$ and $\mathbb{E}[\hat{\mu}_t(A_t)|H_t] = \mathbb{E}[\hat{\mu}_t(A^*)|H_t]$. Using this, we can write $\bar{\Psi}_t(\pi_t^{\text{TS}})$ as

$$\bar{\Psi}_t(\pi_t^{\text{TS}}) = \frac{\tilde{f}_t - \mathbb{E}[\hat{\mu}_t(A_t)|H_t]}{\mathbb{E}[U_t(A_t) - \hat{\mu}_t(A_t)|H_t]} = \frac{\tilde{f}_t - \mathbb{E}[\hat{\mu}_t(A^*)|H_t]}{\mathbb{E}[U_t(A^*) - \hat{\mu}_t(A^*)|H_t]}. \tag{15}$$

By conditioning on the event $\{f_\theta(A^*) \le U_t(A^*)\}$, the following inequality follows from the fact that $f_\theta(A^*) \le 1$.

$$\mathbb{E}[f_\theta(A^*)|H_t] \le \mathbb{E}[U_t(A^*)|H_t] + \Pr(f_\theta(A^*) > U_t(A^*)|H_t). \tag{16}$$

Consider the numerator of (15). We add and subtract $\mathbb{E}[f_\theta(A^*)|H_t]$ and use (16):

$$\begin{aligned}\tilde{f}_t - \mathbb{E}[\hat{\mu}_t(A^*)|H_t] &= \mathbb{E}[f_\theta(A^*) - \hat{\mu}_t(A^*)|H_t] + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t] \\ &\le \mathbb{E}[U_t(A^*) - \hat{\mu}_t(A^*)|H_t] + \Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t]. \end{aligned} \tag{17}$$

The first term of (17) is equal to the denominator of $\bar{\Psi}_t(\pi^{\text{TS}})$. Therefore,

$$\begin{aligned}\bar{\Psi}_t(\pi^{\text{TS}}) &\le 1 + \frac{\Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t]}{\mathbb{E}[U_t(A^*) - \hat{\mu}_t(A^*)|H_t]} \\ &\le 1 + \frac{1}{r_{\min}}(\Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t]).\end{aligned}$$

□

The next lemma simplifies the expectation of (14) using Cauchy-Schwarz.

Lemma 4. *For any t, $\mathbb{E}[\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}})] \le \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \frac{r_{\max}}{r_{\min}}\left(\frac{1}{T} + \frac{2}{\sqrt{m}}\right)$.*

Proof. Taking the expectation of (14) gives us

$$\begin{aligned}&\mathbb{E}[\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}})] \\ \le&\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})(1 + \frac{1}{r_{\min}}(\Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t]))] \\ =&\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] \\ &+ \frac{1}{r_{\min}}\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})\Pr(f_\theta(A^*) > U_t(A^*)|H_t)] &(18)\\ &+ \frac{1}{r_{\min}}\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])]. &(19)\end{aligned}$$

We will now upper bound (18) and (19) with $\frac{r_{\max}}{r_{\min}} \cdot \frac{1}{T}$ and $\frac{r_{\max}}{r_{\min}} \cdot \frac{2}{\sqrt{m}}$ respectively, in which case the result will follow. First, consider (18). Using Cauchy-Schwarz yields

$$\begin{aligned}&\frac{1}{r_{\min}}\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})\Pr(f_\theta(A^*) > U_t(A^*)|H_t)] \\ \le&\frac{1}{r_{\min}}\sqrt{\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})^2]\mathbb{E}[\Pr(f_\theta(A^*) > U_t(A^*)|H_t)^2]} \\ \le&\frac{1}{r_{\min}T}\sqrt{\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})^2]} \\ \le&\frac{1}{T} \cdot \frac{r_{\max}}{r_{\min}},\end{aligned} \tag{20}$$

where the second step uses the following.

$$\begin{aligned}\mathbb{E}[\Pr(f_\theta(A^*) > U_t(A^*)|H_t)^2] &= \mathbb{E}[\mathbb{E}[\mathbb{1}(f_\theta(A^*) > U_t(A^*))|H_t]^2] \\ &\le \mathbb{E}[\mathbb{E}[\mathbb{1}(f_\theta(A^*) > U_t(A^*))^2|H_t]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}(f_\theta(A^*) > U_t(A^*))|H_t]] \\ &\le \Pr(f_\theta(A^*) > U_t(A^*)) \\ &\le \frac{1}{T^2},\end{aligned}$$

where the first inequality uses Jensen's inequality, and the last inequality uses Lemma 1.

Similarly, we apply Cauchy-Schwarz to (19).

$$\frac{1}{r_{\min}}\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])] \leq \frac{1}{r_{\min}}\sqrt{\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})^2]\mathbb{E}[(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])^2]}. \tag{21}$$

Recall that $\tilde{f}_t = \frac{1}{m}\sum_{i=1}^m \tilde{f}_i$, and $\tilde{f}_i$ has the same distribution as $f_\theta(A^*)$ conditioned on H_t. Therefore, $\mathbb{E}[\tilde{f}_t|H_t] = \mathbb{E}[f_\theta(A^*)|H_t]$. Then, we have

$$\begin{aligned}
\mathbb{E}[(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])^2] &= \mathbb{E}[\mathbb{E}[(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])^2|H_t]] \\
&= \mathbb{E}[\text{Var}(\tilde{f}_t|H_t)] \\
&= \mathbb{E}[\frac{1}{m}\text{Var}(\tilde{f}_i|H_t)] \\
&\leq \frac{4}{m}.
\end{aligned}$$

The last inequality follows since $\tilde{f}_i \in [-1, 1]$. Combining this with (21), we get

$$\begin{aligned}
\frac{1}{r_{\min}}\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])] &\leq \frac{2}{r_{\min}\sqrt{m}}\sqrt{\mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})^2]} \\
&\leq \frac{2}{\sqrt{m}} \cdot \frac{r_{\max}}{r_{\min}}
\end{aligned} \tag{22}$$

Substituting (20) and (22) into (19) yields the desired result. □

Proof of Proposition 1. Conditioned on H_t, the expectation of $f_\theta(A^*)$ and $\tilde{f}_t$ is the same, implying $\mathbb{E}[f_\theta(A^*)] = \mathbb{E}[\tilde{f}_t]$ for any t. Therefore, the Bayes regret can be written as $\sum_{t=1}^T \mathbb{E}[\tilde{f}_t - f_\theta(A_t^{\text{TS-UCB}})]$. By adding and subtract $\hat{\mu}_t(A_t^{\text{TS-UCB}})$, we derive

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) = \sum_{t=1}^T \mathbb{E}[\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}})] + \sum_{t=1}^T \mathbb{E}[\hat{\mu}_t(A_t^{\text{TS-UCB}}) - f_\theta(A_t^{\text{TS-UCB}})]. \tag{23}$$

The first sum in (23) can be bounded by $\sum_{t=1}^T \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \frac{r_{\max}}{r_{\min}}\left(1 + \frac{2T}{\sqrt{m}}\right)$ using Lemma 4. Using Corollary 1, the second sum in (23) can be bounded by $\sum_{t=1}^T (\mathbb{E}[\hat{\mu}_t(A_t^{\text{TS-UCB}}) - L_t(A_t^{\text{TS-UCB}})] + T^{-3}) \leq \sum_{t=1}^T \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + T^{-2}$. Substituting these two bounds results in

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq 2\sum_{t=1}^T \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \frac{r_{\max}}{r_{\min}}\left(1 + \frac{2T}{\sqrt{m}}\right) + T^{-2}$$

as desired. □

B.2 Proof of Proposition 2.

The main idea of this proof is captured in the following lemma, which says that we can essentially replace the term $\mathbb{E}[f_\theta(A^*)]$ with $\mathbb{E}[U_t(A_t^{\text{TS-UCB}})]$.

Lemma 5. *For every t,* $\mathbb{E}[f_\theta(A^*)] \leq \mathbb{E}[U_t(A_t^{\text{TS-UCB}})] + mT^{-3}$.

Proof. Fix t, H_t, and $\tilde{f}_t$. For an action $a \in \mathcal{A}_t$, if $U_t(a) \geq \tilde{f}_t$, then $\Psi_t(a) \leq 1$ since the denominator of the ratio is always positive. Otherwise, if $U_t(a) < \tilde{f}_t$, then $\Psi_t(a) > 1$. This implies that an action whose UCB is higher than $\tilde{f}_t$ will always be chosen over an action whose UCB is smaller than $\tilde{f}_t$. Therefore, in the case that $\tilde{f}_t \leq \max_{a\in\mathcal{A}_t} U_t(a)$, it will be that $U_t(A_t^{\text{TS-UCB}}) \geq \tilde{f}_t$. Since $\tilde{f}_t \leq 1$, we have

$$\begin{aligned}
\mathbb{E}[\tilde{f}_t|H_t] &\leq U_t(A_t^{\text{TS-UCB}})\Pr(\tilde{f}_t \leq \max_{a\in\mathcal{A}_t} U_t(a)|H_t) + \Pr(\tilde{f}_t > \max_{a\in\mathcal{A}_t} U_t(a)|H_t) \\
&\leq U_t(A_t^{\text{TS-UCB}}) + \Pr(\tilde{f}_t > \max_{a\in\mathcal{A}_t} U_t(a)|H_t).
\end{aligned}$$

Since $\tilde{f}_t = \frac{1}{m}\sum_{i=1}^{m} \tilde{f}_i$, if $\tilde{f}_t$ is larger than $\max_{a\in\mathcal{A}_t} U_t(a)$, it must be that at least one of the elements $\tilde{f}_i$ is larger than $\max_{a\in\mathcal{A}_t} U_t(a)$. Then, the union bound gives us $\Pr(\tilde{f}_t > \max_{a\in\mathcal{A}_t} U_t(a)|H_t) \leq \sum_{i=1}^{m} \Pr(\tilde{f}_i > \max_{a\in\mathcal{A}_t} U_t(a)|H_t)$. By definition of $\tilde{f}_i$, the distribution of $\tilde{f}_i$ and $f_\theta(A_t^*)$ are the same conditioned on H_t.Therefore,

$$\mathbb{E}[\tilde{f}_t|H_t] \leq U_t(A_t^{\text{TS-UCB}}) + m\Pr(f_\theta(A_t^*) > \max_{a\in\mathcal{A}_t} U_t(a)|H_t).$$

Using the fact that $\mathbb{E}[\tilde{f}_t|H_t] = \mathbb{E}[f_\theta(A_t^*)|H_t]$ and taking expectations on both sides, we have

$$\begin{aligned}
\mathbb{E}[f_\theta(A_t^*)] &\leq \mathbb{E}[U_t(A_t^{\text{TS-UCB}})] + m\Pr(f_\theta(A_t^*) > \max_{a\in\mathcal{A}_t} U_t(a)) \\
&\leq \mathbb{E}[U_t(A_t^{\text{TS-UCB}})] + m\Pr(f_\theta(A_t^*) > U_t(A_t^*)) \\
&\leq \mathbb{E}[U_t(A_t^{\text{TS-UCB}})] + mT^{-3}.
\end{aligned}$$

The last inequality uses Lemma 1. □

Proof of Proposition 2.

$$\begin{aligned}
\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) &= \sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^*) - f_\theta(A_t^{\text{TS-UCB}})] \\
&\leq \sum_{t=1}^{T} (\mathbb{E}[U_t(A_t^{\text{TS-UCB}}) - f_\theta(A_t^{\text{TS-UCB}})] + mT^{-3}) \\
&\leq \sum_{t=1}^{T} (\mathbb{E}[U_t(A_t^{\text{TS-UCB}}) - L_t(A_t^{\text{TS-UCB}})] + T^{-3}) + mT^{-2} \\
&= 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + (m+1)T^{-2},
\end{aligned}$$

where the first inequality uses Lemma 5 and the second inequality uses Corollary 1. □

B.3 Final step of proof.

Proof of Theorem 1. The UCBs in (4) imply that $\text{radius}_t(a) \in [\sqrt{\frac{3\log T}{T}}, \sqrt{3\log T}]$ for all a and t, therefore $\frac{r_{\max}}{r_{\min}} \leq \sqrt{T}$. Then, Propositions 1 and 2 result in the following two inequalities respectively:

$$\begin{aligned}
\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) &\leq 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \sqrt{T} + 2\sqrt{\frac{T^3}{m}} + T^{-2}, \\
\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) &\leq 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \frac{m}{T^2} + T^{-2}.
\end{aligned}$$

Combining these two bounds results in

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \sqrt{T} + T^{-2} + \min\left\{2\sqrt{\frac{T^3}{m}}, \frac{m}{T^2}\right\}.$$

For any value of $m > 0$, $\min\left\{2\sqrt{\frac{T^3}{m}}, \frac{m}{T^2}\right\} \leq 2\sqrt{T}$. Plugging in the known bound for $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})]$ from Theorem 3 finishes the proof of Theorem 1.[4] □

[4]The statement of Theorem 1 has an additional $+K$ term since the first K time steps are used to pull each arm once, which we did not include in the proof to simplify exposition.

Proof of Theorem 2. The following lemma, whose proof is deferred to Section B.4, allows us to bound $\frac{r_{\max}}{r_{\min}}$ by $4\sqrt{2T}$.

Lemma 6. *For the linear bandit, using the UCBs as defined in* (7)*, if* $||X(a)||_2 = 1$ *for every* a*, then* $\text{radius}_t(a) \in [r\sqrt{\frac{d\log T}{T}}, 4r\sqrt{2d\log T}]$ *for every* t *and* a*.*

Then, using the same steps from the proof of Theorem 1, we derive

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + 4\sqrt{2T} + T^{-2} + \min\left\{8\sqrt{\frac{2T^3}{m}}, \frac{m}{T^2}\right\}. \tag{24}$$

For any m, $\min\left\{8\sqrt{\frac{2T^3}{m}}, \frac{m}{T^2}\right\} \leq 8\sqrt{2T}$. Plugging in the known bound for $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})]$ from Theorem 4 gives us (8), finishing the proof of Theorem 2. □

B.4 Deferred Proofs of Lemmas

Proof of Lemma 1. In the linear bandit, this lemma follows directly from Theorem 2 of Abbasi-Yadkori et al. (2011) (using the parameters $\delta = T^{-3}, \lambda = 1$). In the K-armed setting, if $\hat{\mu}(n, a)$ is the empirical mean of the first n plays of action a, Hoeffding's inequality implies $\Pr(f_\theta(a) - \hat{\mu}(n, a) \geq \sqrt{\frac{3\log T}{n}}) \leq T^{-6}$ for any n. Then, since the number of plays of a particular action is no larger than T, we have

$$\Pr(f_\theta(a) - \hat{\mu}_t(a) \geq \sqrt{\frac{3\log T}{N_t(a)}}) \leq \Pr(\cup_{n=1}^{T}\{f_\theta(a) - \hat{\mu}(n, a) \geq \sqrt{\frac{3\log T}{n}}\}) \leq T^{-5}.$$

Since $|\mathcal{A}| = K \leq T$ and $A^*, A_t \in \mathcal{A}$, the result follows after taking another union bound over actions (which proves a stronger bound of T^{-4}). □

Proof of Lemma 2. Fix H_t and $\tilde{f}_t$. For every action a, let $\Delta_a = \tilde{f}_t - \hat{\mu}_t(a)$, and hence $\Psi_t(a) = \frac{\Delta_a}{\text{radius}_t(a)}$. Let ν be a distribution over $\mathcal{A}_t$. Then,

$$\bar{\Psi}_t(\nu) = \frac{\mathbb{E}_{a\sim\nu}[\Delta_a]}{\mathbb{E}_{a\sim\nu}[\text{radius}_t(a)]}. \tag{25}$$

$\text{radius}_t(a) > 0$ for all a, but Δ_a can be negative. We claim that the above ratio is minimized when τ puts all of its mass on one action — in particular, the action $a^* \in \text{argmin}_a \frac{\Delta_a}{\text{radius}_t(a)}$.

For $a \neq a^*$, let $c_a = \frac{\text{radius}_t(a)}{\text{radius}_t(a^*)} > 0$. Then, since $\Psi_t(a) \geq \Psi_t(a^*)$, we can write $\Delta_a = c_a\Delta_{a^*} + \delta_a$ for $\delta_a \geq 0$ for all a. Let $p_{a^*} = \Pr(a = a^*)$. Let $E = \{a \neq a^*\}$ Substituting into (25), we get

$$\begin{aligned}
\bar{\Psi}_t(\nu) &= \frac{\mathbb{E}[c_a\Delta_{a^*} + \delta_a]}{\mathbb{E}[c_a\text{radius}_t(a^*)]} \\
&= \frac{p_{a^*}\Delta_{a^*} + \mathbb{E}[c_a\Delta_{a^*} + \delta_a|E]\Pr(E)}{p_{a^*}\text{radius}_t(a^*) + \mathbb{E}[c_a\text{radius}_t(a^*)|E]\Pr(E)} \\
&= \frac{\Delta_{a^*}(p_{a^*} + \mathbb{E}[c_a|E]\Pr(E)) + \mathbb{E}[\delta_a|E]\Pr(E)}{\text{radius}_t(a^*)(p_{a^*} + \mathbb{E}[c_a|E]\Pr(E))} \\
&= \frac{\Delta_{a^*}}{\text{radius}_t(a^*)} + \frac{\mathbb{E}[\delta_a|E]\Pr(E)}{\text{radius}_t(a^*)(p_{a^*} + \mathbb{E}[c_a|E]\Pr(E))} \\
&\geq \frac{\Delta_{a^*}}{\text{radius}_t(a^*)} \\
&= \Psi_t(a^*)
\end{aligned}$$

□

Proof of Lemma 6. We have

$$\text{radius}_t(a) = \sqrt{\beta_t}||X(a)||_{V_t^{-1}} = \sqrt{\beta_t}||V_t^{-1/2}X(a)||_2.$$

Then, since $||X(a)||_2 = 1$ for all a,

$$\sqrt{\beta_t}\sigma_{\min}(V_t^{-1/2}) \leq \text{radius}_t(a) \leq \sqrt{\beta_t}\sigma_{\max}(V_t^{-1/2}).$$

First, we lower bound $\sigma_{\min}(V_t^{-1/2})$. To do this, we can instead upper bound $||V_t||_2$, since $\sigma_{\min}(V_t^{-1/2}) = \sqrt{\sigma_{\min}(V_t^{-1})} = \frac{1}{\sqrt{\sigma_{\max}(V_t)}} = \frac{1}{\sqrt{||V_t||_2}}$. The triangle inequality gives $||V_t||_2 \leq ||I||_2 + \sum_{s=1}^t ||X_s X_s^\top||_2$. Since $X_s X_s^\top$ is a rank-1 matrix, the only non-zero eigenvalue is $||X_s||_2^2 = 1$ with eigenvector X_s, since $(X_s X_s^\top)X_s = X_s(X_s^\top X_s)$. Therefore, $||V_t||_2 \leq ||I||_2 + \sum_{s=1}^t ||X_s||_2^2 \leq 1 + T$, which implies $\sigma_{\min}(V_t^{-1/2}) \geq \frac{1}{\sqrt{T+1}} \geq \frac{1}{\sqrt{2T}}$. Recall $\sqrt{\beta_t} = r\sqrt{d\log(T^2(1+t))} + S \geq r\sqrt{d\log T}$, implying $\text{radius}_t(a) \geq r\sqrt{\frac{d\log T}{2T}}$.

Next, we upper bound $\sigma_{\max}(V_t^{-1/2}) = \frac{1}{\sqrt{\sigma_{\min}(V_t)}}$ by lower bounding $\sigma_{\min}(V_t)$. $\sigma_{\min}(V_t) \geq \sigma_{\min}(I) = 1$. Therefore, $\sigma_{\max}(V_t^{-1/2}) \leq 1$. We can upper bound $\sqrt{\beta_t}$ by $r\sqrt{d\log(T^4)} + S \leq 2r\sqrt{4d\log(T)}$, since we assumed $r \geq 1$ and $S \leq \sqrt{d}$. Therefore, we have

$$\text{radius}_t(a) \leq \sqrt{\beta_t} \leq 4r\sqrt{d\log(T)}.$$

□

C SIMULATION SETUP FOR PERSONALIZED NEWS ARTICLE RECOMMENDATION

Each sample in the dataset corresponds to the user context, the set of articles available, the article recommended, and whether the user clicked on the article. There are 1.3 - 2.2 million samples for each of the 10 days. A very similar dataset was used in Chapelle and Li (2011), which was one of the first papers to display superior empirical performance of Thompson Sampling.

Given this dataset, we use the following method to evaluate a bandit policy. For each article, we first learn a mapping from user features to their click probabilities using a logistic regression model on the entire dataset. We only considered articles that had more than 5000 samples so that we could learn an accurate mapping. We then use this logistic regression model to compute $\hat{p}_{ua}$, an estimate of the probability that a user u will click an article a. We then simulate a bandit policy, where the reward observed from the chosen arm is a Bernoulli random variable with parameter $\hat{p}_{ua}$[5]. Then, the total regret is computed as $\sum_{t=1}^T (\max_{a\in\mathcal{A}_t} \hat{p}_{ua} - \hat{p}_{uA_t})$. We consider each of the 10 days as separate bandit problem instances, and we also randomly sample 2% of the dataset so that we have around 25,000-45,000 samples for each problem instance (each "run" used a new random sample). Furthermore, at each time step, out of the articles that were marked to be available in the dataset at that time (this pool contained 20-40 articles), we chose only 10 articles to be available to the bandit algorithm, chosen as the first 10 alphabetically[6].

We model this setting as a linear contextual bandit and use the Bayesian structure of Riquelme et al. (2018). Each article corresponds to an arm, and each arm is associated with unknown parameters $\beta_a \in \mathbb{R}^d$ and $\sigma_a^2 \in \mathbb{R}$. The reward for arm a corresponding to context X is modeled as $Y = \beta_a^\top X + \epsilon$ where $\epsilon \sim N(0, \sigma_a^2)$. We model the joint distribution of the parameters β_a and σ_a^2, where we assume they are distributed according to a Gaussian and an Inverse Gamma distribution respectively. At time t, suppose there have been t_a pulls of arm a, corresponding to the contexts $\boldsymbol{X_t} \in \mathbb{R}^{t_a \times d}$ and rewards $\boldsymbol{Y}_t \in \mathbb{R}^{t_a}$.

Then, the posterior distributions are $\sigma_a^2 \sim \text{IG}(a_t, b_t)$, and $\beta_a|\sigma_a^2 \sim N(\mu_t, \sigma_a^2\Sigma_t)$, where

$$\Sigma_t = (\boldsymbol{X}_t^\top \boldsymbol{X}_t + \Lambda_0)^{-1} \qquad \mu_t = \Sigma_t(\Lambda_0\mu_0 + \boldsymbol{X}_t^\top \boldsymbol{Y}_t), \tag{26}$$

$$a_t = a_0 + t_a/2 \qquad b_t = b_0 + \frac{1}{2}(\boldsymbol{Y}_t^\top \boldsymbol{Y}_t + \mu_0\Sigma_0\mu_0 - \mu_t^\top \Sigma_t^{-1}\mu_t). \tag{27}$$

[5]The motivation for this bandit evaluation method was solely to speed up computation for the IDS algorithm. We initially tried the offline evaluation of Li et al. (2011), which does not require learning a separate regression model; however, running this method once for one day's worth of data using the IDS policy took over 3 days. Using the logistic regression model reduced the simulation time by more than 20×.

[6]Reducing the number of arms was done to also to speed up the computation for IDS, as the runtime of IDS is quadratic in the number of arms.

We initialize the prior parameters to be $a_0 = b_0 = 6$, $\mu_0 = 0_d$, and $\Lambda_0 = 4I_d$. At each time step, for each arm, we first sample $\tilde{\sigma}^2$ from its posterior, and then we use the conditional posterior for β_a, $N(\mu_t, \tilde{\sigma}^2\Sigma_t)$, to run the following linear contextual bandit algorithms: TS, TS-UCB(1), TS-UCB(100), IDS, Greedy, and UCB. For each of the 10 days, we simulated 20 runs for each policy.