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

Jackie Baek
Operations Research Center
Massachusetts Institute of Technology
baek@mit.edu

Vivek F. Farias
Sloan School of Management
Massachusetts Institute of Technology
vivekf@mit.edu

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. This proves particularly valuable in heuristics for deep contextual bandits: we show that TS-UCB achieves materially lower regret on all problem instances in 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 the benchmark set of deep bandit problems studied in Riquelme et al. (2018). Importantly, these improvements come with essentially zero additional computation. In contrast, an implementation of IDS (a state-of-the-art algorithm) (Russo and Van Roy 2018) did not provide consistent improvements over TS on this benchmark set, and required approximately three orders of magnitude more sampling (and thus compute) than either TS or TS-UCB.

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 more 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, viz. literature focused on the development and analysis of upper confidence bound algorithms, literature analyzing Thompson sampling (TS), and literature on 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. It is worth noting that neither the UCB or Bayes UCB algorithms are competitive on the benchmark set of problems in Riquelme et al. (2018).

As discussed, TS is a randomized Bayesian algorithm that chooses an action with the same probability that the action is optimal. Though it was initially proposed in Thompson (1933), TS 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.

On the deep learning front, one key idea that has been used to apply deep learning 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, if the posterior can be approximated in some way, then TS can be readily applied. 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, and we show that TS-UCB provides significant improvements over the use of TS on the deep bandit benchmark in Riquelme et al. (2018).

2. Model

An agent is given a compact set of actions $\mathcal{A}$ in which they must choose one to play at every time step $t \geq 1$. 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 $H_t = (A_1, R_1(A_1), \ldots, A_{t-1}, R_{t-1}(A_{t-1}))$ 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 $A^* : \Theta \to \mathcal{A}$ be a function satisfying $A^*(\theta) \in \operatorname{argmax}_{a\in\mathcal{A}} f_\theta(a)$, which represents the optimal action if θ was known. We use A^* to denote the random variable $A^*(\theta)$, where θ is the true parameter.

The T-period regret of policy π is defined as

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

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

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

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.

2.1. K-armed Bandit

In this setting, $|\mathcal{A}| = K$, and each of the entries of the unknown parameter $\theta \in \mathbb{R}^K$ correspond to the mean of each action. That is, for the i'th action, $f_\theta(i) = \theta_i$. 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.

2.2. Linear Bandit

In the linear bandit, there is a known vector $X(a) \in \mathbb{R}^d$ associated with each action, 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).

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 $\text{radius}_t(a) \triangleq U_t(a) - \hat{\mu}_t(a)$ 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 $\tilde{f}_i = f_{\tilde{\theta}_i}(A^*(\tilde{\theta}_i))$. $\tilde{f}_i$ is 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

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

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,

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

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

[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.

$m \to \infty$). 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.

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 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 evaluate the method described in the above example on a range of real-world datasets in Section 4.2.

3.1. Main Idea of Regret Analysis

We now give an outline of the regret analysis for TS-UCB, which provide intuition on both the form of the ratio (1) and the performance of the algorithm.

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

$$\bar{\Psi}_t(\nu) \triangleq \frac{\tilde{f}_t - \mathbb{E}_{A_t\sim\nu}[\hat{\mu}_t(A_t)]}{\mathbb{E}_{A_t\sim\nu}[\text{radius}_t(A_t)]}. \tag{3}$$

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

$$\Psi_t(A_t^{\text{TS-UCB}}) \leq \bar{\Psi}_t(\pi_t). \tag{4}$$

Now, assume the following two approximations hold at every time step:

(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

$$\begin{aligned}
\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[f_\theta(A^*) - f_\theta(A_t^{\text{TS-UCB}})|H_t]] \\
&\approx \sum_{t=1}^{T} \mathbb{E}[\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}})] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(A_t^{\text{TS-UCB}})\text{radius}_t(A_t^{\text{TS-UCB}})\right], \qquad (5)
\end{aligned}$$

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

(5) 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 light of (4), this implies $\Psi_t(A_t^{\text{TS-UCB}}) \lessapprox 1$. Plugging this back into (5) 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 (5)) 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.

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. The formal proofs of the theorems can be found in the supplementary materials.

3.2. K-armed Bandit

We assume $T \geq K$, and we slightly modify the algorithm to 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,

$$\hat{\mu}_t(a) \triangleq \sum_{s=1}^{t-1} \mathbb{1}(A_s = a) R_s(a) \qquad\qquad U_t(a) \triangleq \hat{\mu}_t(a) + \sqrt{\frac{3 \log T}{N_t(a)}}. \tag{6}$$

This implies $\text{radius}_t(a) = \sqrt{\frac{3 \log T}{N_t(a)}}$.

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

$$\sqrt{N_t(a)}(\tilde{f} - \hat{\mu}_t(a)). \tag{7}$$

This implies that TS-UCB does not have to know the time horizon T a priori.

Remark 1. *For UCB algorithms, it is well known that tuning the parameter $\alpha > 0$ in the radius $\alpha\sqrt{\frac{\log T}{N_t(a)}}$ can vastly change empirical performance Russo and Van Roy (2014). One benefit of TS compared to UCB is that it does not require any such tuning. We see from* (7) *that such tuning is also not needed for* TS-UCB*.*

We now state our main result for this setting.

Theorem 1. *For the K-armed bandit, using the UCBs as defined in* (6)*,*

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq 4\sqrt{3KT \log T} + T^{-2} + 3\sqrt{T} + K = O(\sqrt{KT \log T}). \tag{8}$$

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

It is worth noting that TS has been shown to match the lower bound exactly Bubeck and Liu (2013); we believe that the logarithmic gap is a shortcoming of our analysis.

3.3. 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

$$\hat{\mu}_t(a) \triangleq \langle X(a), \hat{\theta}_t \rangle \qquad\qquad U_t(a) \triangleq \max_{\rho \in C_t} \langle X(a), \rho \rangle. \tag{9}$$

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

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

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, there is no parameter tuning required and 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* (9)*, if* $||X(a)||_2 = 1$ *for all* $a \in \mathcal{A}$*,*

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq B + T^{-2} + 12\sqrt{2T} = O(d \log T \sqrt{T}). \tag{10}$$

where

$$B = 8\sqrt{Td\log(1+TL/d)}(S + r\sqrt{6\log(T) + d\log(1+T/d)}) = O(d\log T\sqrt{T}).$$

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 $|\mathcal{A}|$.

Proofs of both Theorem 1 and 2 can be found in Section 5.

4. Computational Results

We conduct two sets of experiments. The first set is entirely synthetic for an ensemble of linear 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 (a) natural problem features such as dimension and the the level of noise and (b) algorithmic parameters for TS-UCB such as the choice of UCB and the number of posterior samples. The second set of experiments then considers a deep bandit benchmark that consists of substantially more complex bandit problems. Here our goal is to show that TS-UCB provides both state of the art performance (by comparing it not just to TS but also IDS) while being computationally cheap.

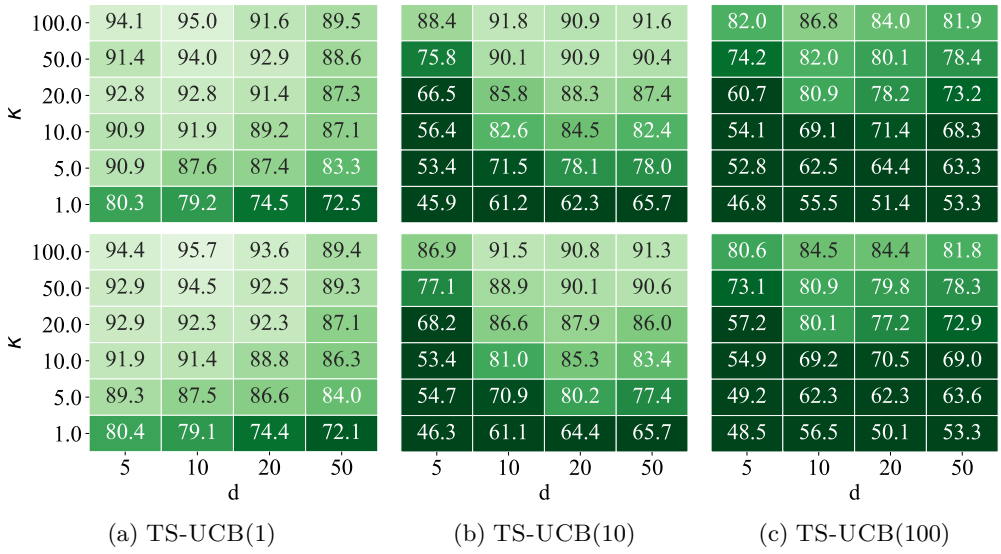
4.1. Synthetic Experiments

First, we simulate synthetic instances of the linear bandit with varying dimension and size of the prior covariance. Let d be the dimension. The number of actions is set to $2d$. For each action, we choose a vector $X(a) \in \mathbb{R}^d$ uniformly at random over the unit sphere. We set the prior for θ as $N(0, \kappa I_d)$ where I_d is the d-dimensional identity matrix, and $\kappa > 0$. The rewards are distributed as $\langle \theta, X(a)\rangle + \epsilon$, where $\epsilon \sim N(0,1)$. We vary the dimension $d \in \{5, 10, 20, 50\}$ and $\kappa \in \{1, 5, 10, 20, 50, 100\}$ to get a total of 24 instances. For each instance, we simulate 500 runs over

a time horizon of $T = 1000$. Note that instances become "easier" when κ increases. This is because when κ is bigger, the norm of θ is also bigger but since the variance of the noise ϵ stays constant at 1, the signal-to-noise ratio is higher in this case.

As for the parameters of the algorithm, we vary the number of samples $m \in \{1, 10, 100\}$, and we also vary how we define the UCBs. In particular, we use the UCBs as (9) and also use Bayes UCBs (Kaufmann et al. 2012a). For the Bayes UCBs, at every time step, we define $U_t(a)$ as the $1 - 1/t$'th percentile of the posterior of $f_\theta(a)$ for every action a.

For each algorithm and each problem instance, we report the median regret as a percentage of the regret from the TS policy. The results are shown in Figure 1.



(a) TS-UCB(1) (b) TS-UCB(10) (c) TS-UCB(100)

Figure 1: TS-UCB improves on TS across the board, particularly on harder instances (bottom). Grid reports median regret of each policy as a percentage of regret of Thompson Sampling over 500 runs. TS-UCB(m) refers to the algorithm using m samples. The top row uses the UCBs defined in (9), while the bottom row uses Bayes UCBs.

4.1.1. Synthetic Experiment Results

Performance Gain relative to TS: We see that TS-UCB outperforms TS across the board, in some cases halving regret. The general trend is that TS-UCB has a greater performance improvement over TS when κ is lower, which correspond to the "harder" instances.

Impact of m and UCB type: We see that performance improves as m increases; on average, the regret decreased by 10.8% from TS-UCB(1) to TS-UCB(10), and 8.8% from TS-UCB(10) to

TS-UCB(100). That said, there do exist problem settings for which a smaller m performs better; characterizing the dependence of regret on the number of samples is an interesting (but challenging) direction for future research. Lastly, performance was very similar across both UCB definitions, suggesting TS-UCB is robust to the specific UCB used.

4.2. Real-World Datasets

In challenging bandit models such as the deep contextual bandit discussed in Example 1, computing a posterior is challenging. 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 on the benchmark problems in Riquelme et al. (2018) and compare its performance to TS and IDS.

For a finite action set, Neural Linear maintains one neural network, $h(\cdot) : \mathbb{R}^{d'} \to \mathbb{R}^d$ and posterior distributions on $|\mathcal{A}|$ parameter vectors β_a. At time t, the posteriors on β_a are computed *ignoring the uncertainty* in the estimate of $h(\cdot)$[2] so that this computation is equivalent to bayesian linear regression. Denoting by $\boldsymbol{\beta}_a$ the random variable distributed according to the posterior on β_a at time t, the action picked by TS is described by the random variable $\operatorname{argmax} h(X_t)^\top \boldsymbol{\beta}_a$. For TS-UCB, we compute $U_t(a)$ as the $1 - 1/t$'th percentile of the random variable $h(X_t)^\top \boldsymbol{\beta}_a$ (note that this can be computed in closed form in the Gaussian-Gaussian model used in the neural-linear approach Riquelme et al. (2018)). Also, $\hat{\mu}_t(a) = h(X_t)^\top \mathbb{E}\boldsymbol{\beta}_a$. We then pick the action that minimizes $\Psi_t(a)$. Finally it is straightforward to implement the IDS algorithm (specifically, the variance-based approximation given by Algorithm 6 in Russo and Van Roy (2018)) given access to draws of $\boldsymbol{\beta}_a$.

We replicate the experiments from Riquelme et al. (2018) with the same real-world datasets, and evaluate the performance of TS, IDS and TS-UCB. These datasets vary widely in their properties; see Appendix A of Riquelme et al. (2018) for the details of each dataset. We use the same parameters and neural network structure as in their paper. While d' varies across experiments, the last layer of the neural network has dimension $d = 50$. For each dataset, one "run" is defined as 2000 data points randomly drawn from the entire dataset; that is, there are 2000 time steps, and each data point (or "context") arrives sequentially in a random order. Lastly, we also run the IDS policy, using

[2] $h(\cdot)$ can be updated at every time step or at scheduled intervals simply by fitting the network to observed rewards.

Table 1: Deep Bandit benchmark Riquelme et al. (2018) results for Neural-Linear and Linear posterior approximation methods. TS-UCB provides an improvement over TS across the board. For each posterior approximation approach, regret of TS-UCB is reported as a percentage of regret of Thompson Sampling (with 95% confidence intervals) for that approach. IDS(5000) requires five thousand samples from the posterior at each epoch; TS-UCB(1) and TS-UCB(10) require one and ten respectively.

Dataset	d'	K	TS-UCB(1)	TS-UCB(10)	IDS(5000)
Statlog	9	7	85.4 ± 0.7	79.0 ± 0.7	149.0 ± 10.7
Adult	14	86	98.0 ± 0.3	97.9 ± 0.3	94.8 ± 0.4
Financial	21	8	57.5 ± 0.7	53.5 ± 0.7	82.6 ± 5.7
Jester	32	8	98.5 ± 0.4	98.3 ± 0.5	92.7 ± 0.7
Covertype	54	7	93.5 ± 0.5	93.2 ± 0.5	79.3 ± 0.7
Mushroom	117	2	92.7 ± 3.0	86.0 ± 2.8	152.7 ± 8.7
Census	369	9	97.5 ± 0.5	97.5 ± 0.4	81.3 ± 0.5

the variance-based approximation given by Algorithm 6 in Russo and Van Roy (2018). Finally, we report the performance of TS-UCB for $m = 1$ and $m = 10$ posterior samples. To get meaningful performance, we require $m = 5000$ for IDS. We report the regret of each policy as a percentage of the regret of TS using the same method, shown in Table 1.

4.2.1. Deep Bandit Benchmark Results

Performance Relative to TS: 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. Moreover, TS-UCB(10) always outperforms TS-UCB(1). Finally, it is worth noting that TS-UCB requires essentially no additional computation over TS.

IDS: IDS is inconsistent across datasets; it performs well on some like Covertype and Census, but quite poorly on others like Statlog and Mushroom and thus does not provide a consistent improvement over TS. The algorithm also requires substantially more posterior samples. These results suggest that both TS and TS-UCB are perhaps more robust to posterior approximation than IDS. There has been some recent work (Phan et al. 2019) on analyzing TS with approximate inference; it is an interesting future direction to study the robustness of other arm selection rules (and regret) to posterior approximation.

In summary, these experiments suggest that TS-UCB consistently improves upon state-of-the-art performance on a challenging deep contextual bandit benchmark.

5. 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: $L_t(a) \triangleq \hat{\mu}_t(a) - (U_t(a) - \hat{\mu}_t(a))$.

5.1. Known Results

We first state three known results used in the analysis. The first result 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* (6) *in the K-armed setting and* (9) *in the linear bandit setting, for any 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 Section 5.3. The following corollary is immediate using the law of total expectation and the fact that $f_\theta(A) > -1$.

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

The next two results upper bound $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t)]$. In particular, for the K-armed setting, the proof of Proposition 2 of Russo and Van Roy (2014) implies the following result.

Theorem 3. *For the K-armed bandit, using the UCBs as defined in* (6)*,*

$$\sum_{t=K+1}^{T} \mathbb{E}[\text{radius}_t(A_t)] \leq 2\sqrt{3KT\log T},$$

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* (9)*,*

$$\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t)] \leq 4\sqrt{Td\log(1+TL/d)}(S + r\sqrt{6\log(T) + d\log(1+T/d)}) = O(d\log T\sqrt{T})$$

for any sequence of actions A_t.

5.2. Proof of Main Result

There are two main steps of the proof which are stated in the following two propositions. These results apply to both the K-armed and linear bandit settings.

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* (6) *for the* K*-armed bandit, and* (9) *for the linear bandit,*

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + \frac{r_{\max}}{r_{\min}}\left(1 + \frac{2T}{\sqrt{m}}\right) + T^{-2}. \tag{11}$$

Proposition 2. *Using the UCBs as defined in* (6) *for the* K*-armed bandit, and* (9) *for the linear bandit,*

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

The proof sketch from Section 3.1 refers to the proof of Proposition 1. 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 (11), which goes to zero as $m \to \infty$. Proposition 2 has the opposite relationship with respect to m, so it applies when m is small. The final step of showing Theorems 1 and 2 involves combining these two propositions to remove the dependence on m, showing $\frac{r_{\max}}{r_{\min}} = O(\sqrt{T})$, and plugging in the known bounds for $\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t)]$ from Theorems 3 and 4. Details of this final step are in Section 5.2.3.

5.2.1. Proof of Proposition 1.

We first show the result claimed in (4) whose proof is deferred to Section 5.3.

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

(12)
$$\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}}) \leq \text{radius}_t(A_t^{\text{TS-UCB}})(1 + \frac{1}{r_{\min}}(\Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])).$$

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 $\mathcal{A}$ 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

$$\bar{\Psi}_t(\pi_t^{\text{TS}}) = \frac{\tilde{f}_t - \mathbb{E}[\hat{\mu}_t(A_t)|H_t]}{\mathbb{E}[U_t(A_t) - \hat{\mu}_t(A_t)|H_t]} = \frac{\tilde{f}_t - \mathbb{E}[\hat{\mu}_t(A^*)|H_t]}{\mathbb{E}[U_t(A^*) - \hat{\mu}_t(A^*)|H_t]}. \tag{13}$$

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

$$\mathbb{E}[f_\theta(A^*)|H_t] \leq \mathbb{E}[U_t(A^*)|H_t] + \Pr(f_\theta(A^*) > U_t(A^*)|H_t). \tag{14}$$

Consider the numerator of (13). We add and subtract $\mathbb{E}[f_\theta(A^*)|H_t]$ and use (14):

$$\begin{aligned}\tilde{f}_t - \mathbb{E}[\hat{\mu}_t(A^*)|H_t] &= \mathbb{E}[f_\theta(A^*) - \hat{\mu}_t(A^*)|H_t] + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t] \\ &\leq \mathbb{E}[U_t(A^*) - \hat{\mu}_t(A^*)|H_t] + \Pr(f_\theta(A^*) > U_t(A^*)|H_t) + \tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t]. \end{aligned} \tag{15}$$

The first term of (15) is equal to the denominator of $\bar{\Psi}_t(\pi^{\text{TS}})$. Therefore,

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

■

The next lemma simplifies the expectation of (12) using Cauchy-Schwarz.

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

Proof. Taking the expectation of (12) gives us

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

We will now upper bound (17) and (18) 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 (17). Using Cauchy-Schwarz yields

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

where the second step uses the following.

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

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

Similarly, we apply Cauchy-Schwarz to (18).

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

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

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

The last inequality follows since $\tilde{f}_i \in [-1, 1]$. Combining this with (22), we get

$$\begin{aligned}\frac{1}{r_{\min}}\mathbb{E}[\mathrm{radius}_t(A_t^{\text{TS-UCB}})(\tilde{f}_t - \mathbb{E}[f_\theta(A^*)|H_t])] &\leq \frac{2}{r_{\min}\sqrt{m}}\sqrt{\mathbb{E}[\mathrm{radius}_t(A_t^{\text{TS-UCB}})^2]} \\ &\leq \frac{2}{\sqrt{m}} \cdot \frac{r_{\max}}{r_{\min}} \qquad (23)\end{aligned}$$

Substituting (20) and (23) into (18) 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

$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) = \sum_{t=1}^T \mathbb{E}[\tilde{f}_t - \hat{\mu}_t(A_t^{\text{TS-UCB}})] + \sum_{t=1}^T \mathbb{E}[\hat{\mu}_t(A_t^{\text{TS-UCB}}) - f_\theta(A_t^{\text{TS-UCB}})]. \qquad (24)$$

The first sum in (24) can be bounded by $\sum_{t=1}^T \mathbb{E}[\mathrm{radius}_t(A_t^{\text{TS-UCB}})] + \frac{r_{\max}}{r_{\min}}\left(1 + \frac{2T}{\sqrt{m}}\right)$ using Lemma 4. Using Corollary 1, the second sum in (24) can be bounded by $\sum_{t=1}^T(\mathbb{E}[\hat{\mu}_t(A_t^{\text{TS-UCB}}) - L_t(A_t^{\text{TS-UCB}})] + T^{-3}) \leq \sum_{t=1}^T \mathbb{E}[\mathrm{radius}_t(A_t^{\text{TS-UCB}})] + T^{-2}$. Substituting these two bounds results in

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

as desired. ■

5.2.2. 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 $a \in \mathcal{A}$, 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 $\tilde{f}_t \leq \max_{a\in\mathcal{A}} U_t(a)$, it will be that $U_t(A_t^{\text{TS-UCB}}) \geq \tilde{f}_t$. Since $\tilde{f}_t \leq 1$, we have

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

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

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

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

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

The last inequality uses Lemma 1. ∎

Proof of Proposition 2.

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

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

5.2.3. Final step of proof.

Proof of Theorem 1. The UCBs in (6) 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.[3] ■

Proof of Theorem 2. The following lemma, whose proof is deferred to Section 5.3, 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* (9)*, 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*.*

[3]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.

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

(25)
$$\text{BayesRegret}(T, \pi^{\text{TS-UCB}}) \leq 2\sum_{t=1}^{T} \mathbb{E}[\text{radius}_t(A_t^{\text{TS-UCB}})] + 4\sqrt{2T} + T^{-2} + \min\left\{8\sqrt{\frac{2T^3}{m}}, \frac{m}{T^2}\right\}.$$

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 (10), finishing the proof of Theorem 2. ■

5.3. Proofs of Lemmas

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 $\mathcal{A}$. Then,

(26)
$$\bar{\Psi}_t(\nu) = \frac{\mathbb{E}_{a\sim\nu}[\Delta_a]}{\mathbb{E}_{a\sim\nu}[\text{radius}_t(a)]}.$$

$\text{radius}_t(a) > 0$ 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 (26), 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 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 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

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

■

References

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

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

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

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

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

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

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

Calandra R, Peters J, Rasmussen CE, Deisenroth MP (2016) Manifold gaussian processes for regression. *2016 International Joint Conference on Neural Networks (IJCNN)*, 3338–3345 (IEEE).

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

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback .

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

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

Hinton GE, Salakhutdinov RR (2008) Using deep belief nets to learn covariance kernels for gaussian processes. *Advances in neural information processing systems*, 1249–1256.

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

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

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

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

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

Phan M, Yadkori YA, Domke J (2019) Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 8801–8811.

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

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

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

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

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

Scott SL (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, Adams R (2015) Scalable bayesian optimization using deep neural networks. *International conference on machine learning*, 2171–2180.

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