Variational Thompson sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Thompson sampling (TS) is an efficient online learning algorithm that balances the exploration-exploitation tradeoff in sequential decision problems. However, it requires the ability to sample from the exact posterior at every iteration, which may be challenging. In this work we develop a variational approximation to TS that maintains the good performance properties while making the inference procedure tractable. At every iteration we solve a convex optimization problem and use the solution as though it was the true posterior probability of each action being optimal. The resulting procedure does not require the posterior to have any special properties, such as log-concavity or smoothness. The optimization view of the problem has several useful properties, including the ability to tune the exploration-exploitation tradeoff, add regularization, incorporate constraints, and extend to games. Moreover, since the variational representation of TS is *differentiable*, it permits the use of TS in a differentiable computation framework, such as a layer in a neural network.

1 Introduction and related work

In this manuscript we consider online learning, in particular the multi-armed stochastic bandit problem [17] and its extension to bilinear saddle-point problems. Thompson sampling (TS) is one of the most famous Bayesian algorithms to tackle this problem [29, 28]. At each iteration TS samples an action according to the posterior probability that it is the optimal action. However, it never computes these probabilities, instead TS samples from the posterior of each arm and then acts greedily with respect to that sample, which is an *implicit* sample from the posterior probability of optimality. TS enjoys strong theoretical guarantees [1, 28] as well as excellent practical performance [6]. Moreover it has an elegant information-theoretic interpretation of an agent that is making a tradeoff between information accumulation and regret [27]. That being said it does have several weaknesses. Firstly, TS requires samples from the true posterior, which may be intractable for some problems. It is known that TS with even small error in its posterior samples can suffer linear regret [25]. In many cases of interest the posterior distribution is highly intractable to sample from and running MCMC chains, one for each arm, can be wasteful and slow and still not yield precise enough samples for good performance, though recent work has attempted to improve this situation [20]. Moreover, TS is unable to take into account other preferences, such as constraints or the presence of other agents. In particular, it was recently shown that TS can suffer linear regret in two-player zero sum games [22] and we shall demonstrate the same phenomenon here empirically for constrained bandits. Finally, when performing TS the user never has direct access to the posterior probability of optimality, only samples from that distribution (though the distribution can be estimated by taking many samples). In some practical applications having access to the policy is desirable, for example to ensure safety constraints or allocate budgets.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In this paper we present *variational* Thompson sampling (VTS). At every step the VTS algorithm solves a convex optimization problem over the simplex. The solution to this problem is an approximation to the TS policy and satisfies a regret bound in the sub-Gaussian reward noise case. The problem is always convex, no matter what form the posteriors take. The only requirement is the ability to compute or estimate the cumulant generating function of the posteriors, which in many cases is far easier than sampling. VTS has similar performance to TS in the standard stochastic multi-armed bandit setting but can dramatically outperform TS in other settings. In particular, the variational formulation allows us to incorporate constraints and opponents in games.

1.1 Contributions

1. We derive a set of 'stochastically optimistic' Bayesian policies, which includes TS, that satisfy a particular optimism inequality and present a new, simple proof of a $\tilde{O}(\sqrt{AT})$ stochastic multi-armed bandit regret-bound for a problem with A arms after T timesteps in the sub-Gaussian noise case that covers any policy in the optimistic set ($\tilde{O}$ ignores logarithmic factors).
2. We derive *variational* TS, which is a policy in the stochastically optimistic set referenced above that can be computed by solving a convex optimization problem.
3. We show that, unlike TS, the variational formulation easily handles more complicated problems, such as bilinear saddle-point problems, which include as special cases zero-sum two player games and constrained bandits.

1.2 Preliminaries

A stochastic multi-armed bandit problem is a sequential learning problem in which an agent interacts with an environment in order to maximize its total cumulative reward. Initially the agent does not know how the rewards are distributed and must learn about them from experience. In each round $t \in \mathbb{N}$ the agent selects an action $a_t \in \{1, \ldots, A\}$ and receives an observation $o_t \in \mathcal{O}$ and a reward $r_t = \mu_{i_t} + \eta_{a_t}$ where $\eta_{a_t} \in \mathbb{R}$ is zero-mean noise. In the bandit feedback case the observation is simply the reward, *i.e.*, $o_t = r_t$, but in other cases it may be different (we shall present an example where the reward is not observed). We define $\mathcal{F}_t = (a_1, o_1, \ldots, a_{t-1}, o_{t-1})$ to be the sequence of actions and observations available to the agent prior to round t, and as shorthand we shall use the notation $\mathbb{E}^t(\cdot) = \mathbb{E}(\,\cdot \mid \mathcal{F}_t)$.

An algorithm alg is a measurable mapping that takes the entire history at time t and produces a probability distribution, the *policy*, from which the agent samples its action. In order to assess the quality of an algorithm alg we consider the *regret*, or shortfall in cumulative rewards relative to the optimal value,

$$\mathrm{Regret}(\mu, \mathrm{alg}, T) = \mathbb{E}_{\eta,\mathrm{alg}}\left[\sum_{t=1}^{T} \max_i \mu_i - r_t\right], \tag{1}$$

where $\mathbb{E}_{\eta,\mathrm{alg}}$ denotes the expectation with respect to the noise process η and any randomness in the algorithm. This quantity (1) depends on the unknown expected reward $\mu \in \mathbb{R}^A$, which is fixed at the start of play and kept the same throughout. To assess the quality of learning algorithms designed to work across some *family* of problems $\mu \in \mathcal{M}$ we define

$$\mathrm{BayesRegret}(\phi, \mathrm{alg}, T) = \mathbb{E}_{\mu\sim\phi}\mathrm{Regret}(\mu, \mathrm{alg}, T) \tag{2}$$

where ϕ is a prior probability measure over $\mu \in \mathcal{M}$ that assigns relative importance to each problem instance which we assume is known to the algorithm.

2 Thompson sampling and a variational approximation

The algorithm that minimizes the Bayesian regret is called the *Bayesian optimal* policy, but it is believed to be intractable to compute in most cases [9]. With that in mind we would like an algorithm that is computationally tractable and still guarantees a good bound on the Bayesian regret. TS is a Bayesian algorithm that at each time step samples possible values of each arm from their respective posteriors and acts greedily with respect to those samples; see Algorithm 1. This simple strategy provides a principled approach to the exploration-exploitation tradeoff in the stochastic multi-armed

bandit problem. In this section we derive a set of 'stochastically optimistic' policies, which includes the TS policy, and we shall show that any algorithm that produces policies in this set satisfies a guaranteed Bayesian regret bound. For brevity we defer all proofs to the appendices.

2.1 Bounding the conditional expectation

The Bayesian regret (2) depends on the expectation of the maximum over a collection of random variables. Here we shall derive an upper bound on this quantity, starting with a generic upper bound on the conditional expectation. If $X : \Omega \to \mathbb{R}^n$ is a random variable with finite expectation then we denote by $\Psi_X : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ the cumulant generating function of $X - \mathbb{E}X$, *i.e.*,

$$\Psi_\mu(\beta) = \log \mathbb{E} \exp(\beta^\top (X - \mathbb{E}X)).$$

This function is always convex, and we shall assume throughout this manuscript that it is also closed and proper for any random variables we encounter. With this definition we can present a bound on the conditional expectation.

Lemma 1. *Let $\mu : \Omega \to \mathbb{R}$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and $A \in \mathcal{F}$ be an event with $\mathbb{P}(A) > 0$. Then, for any $\tau \geq 0$*

$$\mathbb{E}[\mu|A] \leq \mathbb{E}\mu + \tau \Psi_\mu(1/\tau) - \tau \log \mathbb{P}(A). \tag{3}$$

The right hand side of (3) involves the perspective of Ψ_μ and so is convex in τ [4] and we can minimize over $\tau \geq 0$ to find the tightest bound, which we do next. Recall that the *convex conjugate* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is denoted $f^* : \mathbb{R}^n \to \mathbb{R}$ and is given by $f^*(y) = \sup_x \left(x^\top y - f(x)\right)$ [4].

Theorem 1. *Let $\mu : \Omega \to \mathbb{R}$ be a random variable such that the interior of the domain of Ψ_μ is non-empty, then under the same assumptions as Lemma 1 we have* [1]

$$\mathbb{E}[\mu|A] \leq \mathbb{E}\mu + (\Psi_\mu^*)^{-1}(-\log \mathbb{P}(A)).$$

This inequality has clear connections to Kullback's inequality [7] and the Donsker-Varadhan variational representation of KL-divergence [10, Thm. 3.2]. The function Ψ^* is referred to as the *Cramér function* or the *rate function* in large deviations theory [31]. The inverse of the rate function comes up in several contexts, such as queuing theory and calculating optimal insurance premiums [19]. Cramér's theorem tells us that the probability of a large deviation of the empirical average of a collection of IID random variables decays exponentially with more data, and that the decay constant is the rate function. It is no surprise then that the regret bound we derive will depend on the inverse of the rate function, *i.e.*, when the tails decay faster we incur less regret.

2.2 The stochastically optimistic set

Using Theorem 1 we next derive the maximal inequality we shall use to bound the regret.

Lemma 2. *Let $\mu_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, A$, be a collection of random variables, let $i^\star = \operatorname{argmax}_i \mu_i$ (ties broken arbitrarily), and $p_i^\star = \mathbb{P}(i^\star = i)$ which we assume is positive and denote by $\Psi_i \coloneqq \Psi_{\mu_i}$, then*

$$\mathbb{E} \max_i \mu_i \leq \sum_{i=1}^{A} p_i^\star \left(\mathbb{E}\mu_i + (\Psi_i^*)^{-1}(-\log p_i^\star)\right).$$

Note that Lemma 2 makes no assumption of independence between arms, the bound holds even in the case of dependency. In the context of online learning $p_i^\star = \mathbb{P}(i^\star = i)$ is exactly the TS policy when μ_i have the laws of the posterior. For ease of notation we introduce function $\mathcal{G}_\mu : \Delta_A \to \mathbb{R}$ for a collection of random variables $\mu_1, \ldots, \mu_A$, given by

$$\mathcal{G}_\mu(\pi) \coloneqq \sum_{i=1}^{A} \pi_i \left(\mathbb{E}\mu_i + (\Psi_i^*)^{-1}(-\log \pi_i)\right), \tag{4}$$

where Δ_A denotes the probability simplex of dimension $A - 1$. Note that $\mathcal{G}_\mu$ is concave since Ψ_i is convex [4]. With this notation we can write $\mathbb{E} \max_i \mu_i \leq \mathcal{G}_\mu(p^\star)$. This brings us to the 'stochastically optimistic' set of probability distributions, which corresponds to the set of policies that satisfy that same inequality.

[1] If $\Psi_\mu = 0$, then we take $(\Psi_\mu^*)^{-1} = 0$.

Definition 1. *Let μ_i, $i = 1, \dots, A$ be a collection of random variables, we denote by $\mathcal{P}_\mu$ the set*

$$\mathcal{P}_\mu = \{\pi \in \Delta_A \mid \mathbb{E} \max_i \mu_i \leq \mathcal{G}_\mu(\pi)\}.$$

Immediately we have two facts about the set $\mathcal{P}_\mu$. First, it is non-empty since $p^\star \in \mathcal{P}_\mu$, and second, it is a convex set since $\mathcal{G}_\mu$ is concave.

2.3 Variational Thompson sampling

The TS policy is one element of the stochastically optimistic set, but it may not be the only element. Since the TS policy is in $\mathcal{P}_\mu$, it implies that we can maximize the bound and also obtain a policy in $\mathcal{P}_\mu$, since

$$\mathbb{E} \max_i \mu_i \leq \mathcal{G}_\mu(p^\star) \leq \max_{\pi \in \Delta_A} \mathcal{G}_\mu(\pi).$$

We refer to the policy that maximizes $\mathcal{G}_\mu$ as the *variational* Thompson sampling policy (VTS) and we will use $\pi^\star$ to represent it hereafter. It is guaranteed to exist since Δ_A is compact and $\mathcal{G}_\mu$ is continuous. The problem of maximizing $\mathcal{G}_\mu$ is a concave maximization problem, and is thus computationally tractable so long as each Ψ_i is readily accessible. This holds *no matter what form the posterior distribution takes*. This means we don't rely on any properties like log-concavity or smoothness of the distribution for tractability of the optimization problem that yields the VTS policy. For the case where each $\mu_i \sim \mathcal{N}(0, \sigma^2)$ independently we recover the familiar bound of $\mathbb{E} \max_i \mu_i \leq \mathcal{G}_\mu(p^\star) = \max_{\pi \in \Delta_A} \mathcal{G}_\mu(\pi) = \sigma\sqrt{2 \log A}$. The variational form extends this bound to the case of differently distributed random variables.

Algorithm 1 TS for bandits

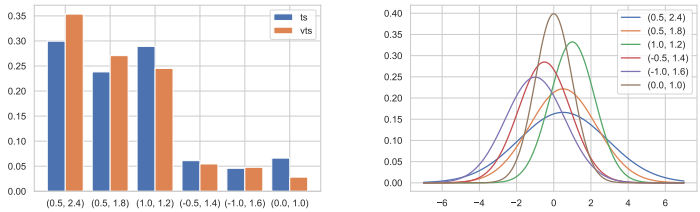
for round $t = 1, 2, \dots, T$ **do**
  sample $\hat{\mu} \sim \phi(\cdot \mid \mathcal{F}_t)$
  choose $a_t \in \operatorname{argmax}_i \hat{\mu}_i$
end for

Algorithm 2 VTS for bandits

for round $t = 1, 2, \dots, T$ **do**
  compute $\pi^t = \operatorname{argmax}_{\pi \in \Delta_A} \mathcal{G}_\mu^t(\pi)$
  sample $a_t \sim \pi^t$
end for

We use the term *variational* to make it clear that this algorithm produces a policy by solving a convex optimization problem. We call it variational *Thompson sampling* because it was motivated via TS and in practice for multi-armed bandits it performs similarly. The name is misleading in two respects. Firstly, variational Bayesian methods typically attempt to maximize the so-called evidence lower bound (ELBO) and will recover the true posterior distribution with a flexible enough parameterization - this is not true in our case. Secondly, it is easy to create examples where the VTS policy is quite different to vanilla TS. In fact the second half of this manuscript is focused on cases where VTS outperforms vanilla TS, sometimes provably so.

In Figure 1 we show an example comparing the true probability of optimality and the variational probabilities, on a collection of Gaussian random variables. We compare the estimated true probability of optimality (by drawing 10^7 samples) to the probabilities as determined by maximizing $\mathcal{G}_\mu$. As can be seen the probability distributions are similar, though not exactly the same. In this case $\mathcal{G}_\mu(p^\star) = 1.75$ and $\mathcal{G}_\mu(\pi^\star) = 1.78$, so $p^\star$ is close to maximizing $\mathcal{G}_\mu$.



(a) Distributions $p^\star$ and $\pi^\star$.

(b) Distributions.

Figure 1: Comparing the true and variational approximation to posterior probability of optimality. The labels correspond to the parameters (μ, σ) for each distribution.

2.4 Regret analysis

Now we shall present regret bounds for algorithms that produce policies in the stochastically optimistic set. These bounds will cover both TS and variational TS. In the following we denote by $\mathcal{P}_\mu^t$ the stochastically optimistic set for random variables $\mu_1, \ldots, \mu_A$ conditioned on $\mathcal{F}_t$.

Lemma 3. *Let* alg *produce any sequence of policies* π^t*,* $t = 1, \ldots, T$*, that satisfy* $\pi^t \in \mathcal{P}_\mu^t$*, then*

$$\text{BayesRegret}(\phi, \text{alg}, T) \leq \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{A} \pi_i^t (\Psi_i^{t*})^{-1} (-\log \pi_i^t).$$

This bound is generic in that it holds for *any* set of posteriors and *any* algorithm producing policies in $\mathcal{P}_\mu^t$. The regret depends on the inverse rate functions of the posteriors, no matter how complicated those posteriors may be, and faster concentrating posteriors, *i.e.*, larger rate functions, will have lower regret. Loosely speaking, as more data accumulates the posteriors concentrate towards the true value (under benign conditions [30, Ch. 10.2]), implying $(\Psi_i^{t*})^{-1} \to 0$. By carefully examining the rate of concentration of the posteriors we can derive concrete regret bounds in specific cases, such as the standard sub-Gaussian reward noise case, which we do next.

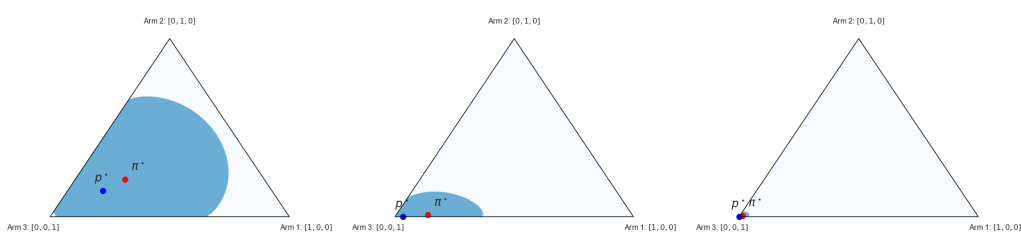
Definition 2. *A random variable* $X : \Omega \to \mathbb{R}$ *is* σ*-sub-Gaussian if* $\Psi_X(\beta) \leq (1/2)\sigma^2\beta^2$.

Now we are ready to prove a concrete regret bound.

Theorem 2. *Let* alg *produce any sequence of policies* π^t*,* $t = 1, \ldots, T$*, that satisfy* $\pi^t \in \mathcal{P}_\mu^t$ *and assume that both the prior and reward noise are* 1*-sub-Gaussian for each arm, then*

$$\text{BayesRegret}(\phi, \text{alg}, T) \leq \sqrt{2AT \log A(1 + \log T)} = \tilde{O}(\sqrt{AT}).$$

The assumption that the prior is 1-sub-Gaussian is not really necessary and can be removed if, for example, we assume the algorithm pulls each arm once at the start of the procedure. Theorem 2 bounds the regret of *both* TS and VTS, as well as any other algorithm producing stochastically optimistic policies. The point of this theorem is not to provide the tightest regret bound but to provide some insight into the question of how close to the true TS policy another policy has to be in order to achieve similar (in particular sub-linear) regret. We have provided one answer - any policy in the optimistic set $\mathcal{P}_\mu^t$ for all t, is close enough. However, this set contracts as more data accumulates (as the posteriors concentrate) so it becomes harder to find good approximating policies. That being said, the VTS policy is always sufficiently close because it is always in the stochastically optimistic set. In Figure 2 we show the progression of a single realization of $(p^\star)^t$, $(\pi^\star)^t$ and the set $\mathcal{P}_\mu^t$ over time for a three-armed Gaussian bandit with standard normal prior and where the policy being followed is TS. Initially $\mathcal{P}_\mu$ is quite wide, but gradually as more data accrues it concentrates toward the optimal arm vertex of the simplex.



(a) $t = 1$ (b) $t = 32$ (c) $t = 1024$

Figure 2: Three-armed bandit: shaded area is $\mathcal{P}_\mu^t$, $p^\star$ is TS policy, and $\pi^\star$ is VTS policy.

3 Connections and interpretations of VTS

Exploration-exploitation tradeoff. Examining the problem that is solved to determine the VTS policy we can break it into two components

$$\mathcal{G}_\mu(\pi) = \underbrace{\sum_{i=1}^{A} \pi_i \mathbb{E}^t \mu_i}_{\text{exploit}} + \underbrace{\sum_{i=1}^{A} \pi_i (\Psi_i^{t*})^{-1}(-\log \pi_i)}_{\text{explore}},$$

the first term of which includes the expected reward (under the posterior) from pulling each arm and is the agent's current understanding of the world, and the second term consists only of the uncertainty in each arm. The policy that maximizes this function is balancing this exploration and exploitation tradeoff, with a tradeoff parameter of one. This is the analog of the ELBO used in variational inference, where the variational distribution is chosen to maximize the sum of two terms, one involving the observed data and the other involving the prior uncertainty, again with a tradeoff parameter of one [33].

Upper confidence bounds. Here we show a connection between (variational) Thompson sampling and upper confidence bound (UCB) methods [16, 2]. The Chernoff inequality tells us that the tail probability of $\mu : \Omega \to \mathbb{R}$ can be bounded as $\mathbb{P}(\mu - \mathbb{E}\mu > \epsilon) \leq \exp(-\Psi_\mu^*(\epsilon))$, for any $\epsilon \geq 0$. In UCB style methods we build upper confidence intervals, which yield high-probability bounds that the random variable lies in that interval [16, 2, 14]. We want the bounds to contain the true value of μ with probability at least $1 - \delta$, so we solve for ϵ yielding $\epsilon = (\Psi_\mu^*)^{-1}(-\log \delta)$ and therefore we have $\mathbb{E}\mu + (\Psi_\mu^*)^{-1}(-\log \delta) \geq \mu$ with high probability. Standard UCB approaches to the stochastic multi-armed bandit typically choose the largest upper confidence bound (the 'optimism in the face of uncertainty' approach), that is, at time t they select action $a_t \in \operatorname{argmax}_{i \in [A]}(\mathbb{E}^t \mu_i + (\Psi_i^{t*})^{-1}(-\log \delta_t))$, for some schedule δ_t. A lot of work goes into tuning δ_t for various applications. By contrast, VTS samples from policy $a_t \sim \operatorname{argmax}_{\pi \in \Delta_A} \sum_{i=1}^{A} \pi_i \left(\mathbb{E}^t \mu_i + (\Psi_i^{t*})^{-1}(-\log \pi_i)\right)$. It is clear that VTS is building upper confidence bounds on the values of each arm that depend on the value of the policy, since if we denote $\pi^\star$ the maximizer to the above problem, then $\mathbb{P}_t(\mathbb{E}^t \mu_i + (\Psi_i^{t*})^{-1}(-\log \pi_i^\star) < \mu_i) \leq \pi_i^\star$. This procedure gives higher 'bonuses' to arms that end up with lower probability of being pulled, and lower bonuses to arms with high probability of being pulled, which is similar to what is done in more sophisticated UCB algorithms such as KL-UCB+ [15, 8, 12]. This is desirable because if the mean of an arm is low then there is a chance this is simply because we have under-estimated it and we should pay more attention to that arm. In this light VTS can be interpreted as a stochastic Bayesian UCB approach [14]. One advantage is that we do not have to tune a schedule δ_t, like TS our algorithm is *stationary*, *i.e.*, it does not depend on how many time-steps have elapsed other than implicitly through the posteriors.

K-learning. K-learning (knowledge learning) is an algorithm originally developed for efficient reinforcement learning in MDPs where it enjoys a similar regret bound to TS [21, 24]. For bandits, the agent policy is given by

$$\tau^\star = \operatorname*{argmin}_{\tau \geq 0} \left(\tau \log \sum_{i=1}^{A} \exp(\mathbb{E}\mu_i/\tau + \Psi_i(1/\tau)) \right), \qquad \pi_i^\star \propto \exp(\mathbb{E}\mu_i/\tau^\star + \Psi_i(1/\tau^\star)). \tag{5}$$

It can be shown that this is equivalent to VTS when the 'temperature' parameters τ_i that we optimize over are constrained to be equal for each arm. Although K-learning and VTS have essentially the same regret bound, this can make a large difference in practice. This connection also suggests that recent reinforcement learning algorithms that rely on the log-sum-exp operator, the 'soft' max family [11, 18, 23, 26], are also in some sense approximating the TS policy, although as made clear in [24] these approaches do not correctly model the uncertainty and so will suffer linear regret in general.

4 Bilinear saddle-point problems

In this section we extend the above analysis to constrained bilinear saddle-point problems, which include as special cases zero-sum two player games and constrained multi-armed bandits, as well as

vanilla stochastic multi-armed bandits. Equipped with the tools from the previous sections we shall formulate a convex optimization problem that when solved yields the VTS policy, and prove that in two special cases this strategy yields a sub-linear regret bound. It is known, and we shall demonstrate empirically, that TS (Algorithm 3) can fail, *i.e.*, suffer linear regret, in this more general regime [22, §3.2]. Consequently, this is the first setting where the variational approach is substantially more powerful than TS.

A finite dimensional, constrained, bilinear saddle-point problem is defined by a matrix $R \in \mathbb{R}^{m\times n}$ and two non-empty, compact, convex sets $\Pi \subseteq \mathbb{R}^n$, $\Lambda \subseteq \mathbb{R}^m$ which are the feasible sets for the maximizer and minimizer respectively. Here we shall assume that the agent is the maximizer, *i.e.*, that $\Pi = \Delta_A$, and that the agent is uncertain about the entries of R and must learn about them from experience. The saddle-point problem has optimal value given by

$$V_R^\star := \min_{\lambda\in\Lambda} \max_{\pi\in\Delta_n} \lambda^\top R\pi = \max_{\pi\in\Delta_n} \min_{\lambda\in\Lambda} \lambda^\top R\pi \tag{6}$$

where we can swap the order of the min-max due to the minimax theorem [32]. Unlike the stochastic multi-armed bandit case the maximizing policy in the bilinear saddle-point is typically stochastic. We shall consider the case where at each time period the agent selects a column of R and receives reward $r_t \in \mathbb{R}$ and observation $o_t \in \mathcal{O}$ where the exact reward, observation, and opponent space Λ will depend on the problem setup. For instance, the vanilla stochastic multi-armed bandit is a special case of this problem where Λ is a singleton and the agent receives bandit feedback associated the the action it selected. We define the Bayes regret to be

$$\mathrm{BayesRegret}(\phi, \mathrm{alg}, T) = \mathbb{E}\sum_{t=1}^{T}(V_R^\star - r_t).$$

The key insight is that we can bound $\mathbb{E}V_R^\star$ using the techniques we have previously developed.

Lemma 4. *Let $R_j : \Omega \to \mathbb{R}^m$ denote the random variable corresponding to the jth row of R, denote by $\Psi_j^t : \mathbb{R}^m \to \mathbb{R}$ the cumulant generating function of $R_j - \mathbb{E}^t R_j$, then*

$$\mathbb{E}^t \min_{\lambda\in\Lambda} \max_{\pi\in\Delta_n} \lambda^\top R\pi \le \min_{\lambda\in\Lambda,\tau\ge 0} \max_{\pi\in\Delta_n} \sum_{j=1}^{n} \pi_j(\lambda^\top \mathbb{E}^t R_j + \tau_j \Psi_j^t(\lambda/\tau_j) - \tau_j \log \pi_j). \tag{7}$$

The right hand side of (7) is a saddle-point problem which is jointly convex-concave in variables $(\lambda, \tau) \in \Lambda \times \mathbb{R}_+^n, \pi \in \Delta_n$. Maximizing over π yields dual convex optimization problem:

$$\begin{array}{ll} \text{minimize} & V + \sum_{i=1}^n \tau_j \exp(s_j/\tau_j) \\ \text{subject to} & s_j \ge \lambda^\top \mathbb{E}^t R_j + \tau_i \Psi_j^t(\lambda/\tau_j) - V - \tau_j, \quad j = 1, \dots, n \\ & \tau \ge 0, \lambda \in \Lambda, \end{array} \tag{8}$$

over variables $\tau \in \mathbb{R}^n, \lambda \in \mathbb{R}^m, s \in \mathbb{R}^n, V \in \mathbb{R}$. The VTS distribution can be recovered using $\pi_i^\star = \exp(s_i^\star/\tau_i^\star)$, this is presented as Algorithm 4.

Algorithm 3 TS for saddle-points

for round $t = 1, 2, \dots, T$ **do**
  sample $\hat{R} \sim \phi(\cdot \mid \mathcal{F}_t)$
  choose $a_t \sim \pi = \operatorname*{argmax}_{\pi\in\Delta_A} \min_{\lambda\in\Lambda} \lambda^\top \hat{R}\pi$
end for

Algorithm 4 VTS for saddle-points

for round $t = 1, 2, \dots, T$ **do**
  solve Equation (8) for $s^\star$, $\tau^\star$
  choose $a_t \sim \pi^t = \exp(s^\star/\tau^\star)$
end for

4.1 Zero-sum two-player games

Consider the problem of an agent playing a two-player zero-sum matrix game against an opponent. In this setting the 'row player' selects a row index $i \in \{1, \dots, m\}$ and simultaneously the 'column player' selects a column index $j \in \{1, \dots, n\}$, *i.e.*, in this case $\Lambda = \Delta_m$. These choices are revealed to the players and the row player makes a payment of $r_t = R_{ij} + \eta_t$ to the column player, where η_t is a zero-mean noise term, and the observation the agents receive is the reward. The optimal strategies

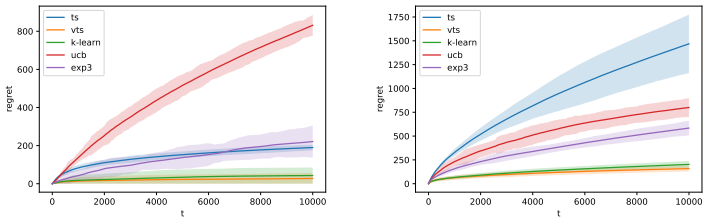
$(\pi^\star, \lambda^\star)$ are a Nash equilibrium for the game [32, 5]. We shall assume that the agent we control is the column player, and that we have no control over the actions of the opponent. It was recently shown that both K-learning and UCB variants enjoy sub-linear regret bounds in this context [22], which stand in contrast to TS which can suffer *linear* regret. Here we prove a regret bound for VTS under a sub-Gaussian assumption.

Theorem 3. *Assuming that the prior over each entry of R and the reward noise are 1-sub-Gaussian, then the VTS Algorithm 4 has regret bound*

$$\text{BayesRegret}(\phi, \text{VTS}, T) \leq \sqrt{2nmT \log n(1 + \log T)} = \tilde{O}(\sqrt{nmT}).$$

Unlike the stochastic multi-armed bandit case, the above regret bound does not imply a regret bound for TS. This is due to the fact that in some time-steps TS is open to exploitation by the opponent and, unlike the vanilla bandit, the opponent can exploit TS without revealing information about R, so the exploitation can continue forever. This argument is formalized via a counter-example in [22].

In Figure 3 we show how five agents perform on a 10×10 game in self-play and against a best-response opponent. The entries of R were sampled from prior $\mathcal{N}(0, 1)$ and the noise term η_t at each time-period was also sampled from $\mathcal{N}(0, 1)$. In self-play the opponent has the same knowledge and is applying the same algorithm to generate the policy at each time-step. The best-response opponent knows the matrix M exactly as well as the policy that the agent produces and picks the action that minimizes the reward to the agent. The matrix is the same in both cases. We compare the K-learning and UCB algorithms as described in [22] as well as EXP3 [3], a generic adversarial bandit algorithm, VTS Algorithm 4, and TS Algorithm 3 over 8 seeds. As can be seen VTS is the best performing algorithm in both cases, followed closely by K-learning. Although performing moderately well in self-play, it is no surprise that TS does very poorly against the best-response opponent since it is susceptible to exploitation.



(a) Self-play. (b) Against best-response opponent.

Figure 3: Regret on 10×10 random game.

4.2 Constrained bandits

In constrained bandits we want to find the policy that maximizes expected reward subject to a set of constraints. The constraints specify that the expected value of other reward functions associated with the same arms, which are also initially unknown, are larger than zero. Formally, we have the usual unknown reward function $\mu_0 \in \mathbb{R}^d$ as well as k other unknown reward functions $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$. When there is zero uncertainty we want to solve the following primal-dual linear-programming pair

$$\begin{array}{ll} \text{maximize} & \mu_0^\top \pi \\ \text{subject to} & \mu_i^\top \pi \geq 0, \quad i = 1, \ldots, k \\ & \pi \in \Delta_d \end{array} \qquad \begin{array}{ll} \text{minimize} & V \\ \text{subject to} & \sum_{i=0}^k \mu_i \lambda_i \leq V\mathbf{1} \\ & \lambda \geq 0, \lambda_0 = 1, \end{array}$$

over primal variable $\pi \in \Delta_A$ and dual variables $V \in \mathbb{R}$ and $\lambda \in \mathbb{R}^{k+1}$, and where $\mathbf{1}$ denotes the vector of all ones. For brevity, let $m = k + 1$ and $R \in \mathbb{R}^{m \times n}$ be the matrix with the reward vectors $\mu_0, \ldots, \mu_m$ stacked row-wise, then we can write the Lagrangian for the above primal-dual pair as Equation (6), *i.e.*, this problem is a constrained bilinear saddle-point problem.

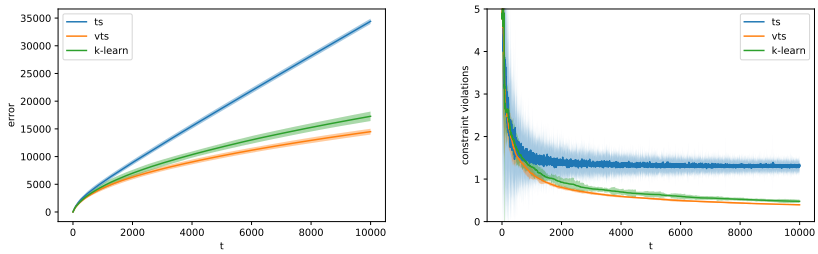
At each time-period the agent selects column i according to its policy after which it observes vector $o_t = (\mu_{0i} + \eta_{0i}^t, \mu_{1i} + \eta_{1i}^t, \ldots, \mu_{mi} + \eta_{mi}^t)$ where η_{ij}^t is zero-mean noise for entry i, j. In this case we must be careful to ensure that the Bayes regret is well-defined. If the support of the prior

includes infeasible problems, which have $V_R^\star = -\infty$, then the expectation of the regret under that prior is unbounded below. However, if we make the assumption that $\|\lambda^\star\|_2 \leq C$ for some $C > 0$ almost surely, which implies that the problem is feasible almost surely, then the Bayes regret is well-defined and we can derive a regret bound that depends on C. With this in mind we define the set $\Lambda = \{\lambda \in \mathbb{R}^k \mid \lambda_0 = 1, \lambda \geq 0, \|\lambda\|_2 \leq C\}$ for some $C > 0$. We define the reward at each time-period to be $r_t = \min_{\lambda \in \Lambda} \lambda R \pi_t$, where π_t is the policy of the agent at time t. We choose this reward for two reasons, firstly the regret with this reward is always non-negative, and secondly if the regret is zero then the corresponding π_t is optimal for the linear program since it is a maximizer of the saddle-point problem [4]. Unlike the previous examples the agent never observes this reward since it depends on exact knowledge of R to compute, the reward is only used to measure regret. Nevertheless, VTS satisfies a Bayes regret upper bound.

Theorem 4. *Assuming that the prior over each entry of R and the observation noise terms are 1-sub-Gaussian and that $\|\lambda^\star\|_2 \leq C$ almost surely, then the VTS Algorithm 4 has regret bound*

$$\text{BayesRegret}(\phi, \text{VTS}, T) \leq 2C(\sqrt{m/2} + \sqrt{2\log n})(1 + \log T)\sqrt{nT} = \tilde{O}(\sqrt{mnT}).$$

Next we present a numerical experiment with $n = 50$ variables and $m = 40$ constraints. The entries of R were sampled from prior $\mathcal{N}(-0.15, 1)$, and we repeatedly sampled from this distribution until a feasible problem was found. We chose a slightly negative bias in the prior for R in order to make the problem of finding a feasible policy more challenging. All noise was sampled from $\mathcal{N}(0, 1)$. For this experiment we set $C = 10$; the true optimal dual variable satisfied $\|\lambda^\star\| \approx 7$. In Figure 4 we compare the performance of TS Algorithm 3, VTS Algorithm 4, and K-learning [21] over 8 seeds. We also plot how much the policy produced by each algorithm violated the constraints at each time-period, as measured by ℓ_∞ norm. As can be seen both VTS and K-learning enjoy sub-linear regret and have their constraint violation converging to zero, with VTS slightly outperforming K-learning on both metrics. TS on the other hand appears to suffer from linear regret, and the constraint violation decreases up to a point after which it saturates at a non-zero value. It is clear that TS is unable to adequately trade off exploration and exploitation in this problem.



(a) Regret. (b) Constraint violation.

Figure 4: Performance on a random 50×40 constrained bandit problem.

5 Conclusion

In this manuscript we introduced *variational* Thompson sampling, which is a policy that approximates the Thompson sampling policy, enjoys low regret in certain cases, and is computed by solving an optimization problem. The optimization problem is always convex, no matter how complicated the posteriors are. We then extended the analysis of the algorithm to the bilinear saddle-point problem, a more general problem that includes the vanilla stochastic multi-armed bandit as a special case, but also includes zero-sum two-player games and constrained bandits and where Thompson sampling often fails.

We conclude with some discussion about future directions for this work. First, we note that TS applied to Markov decision processes is also amenable to the variational approximation with similar regret guarantees, though the proofs are more delicate. Second, note that VTS permits the use of parameterized policies, where we can find the optimal policy in the family by solving the VTS optimization problem, and since the VTS objective is differentiable this opens the door to exotic architectures, such as deep neural networks. These suggest the possibility of applying VTS in deep reinforcement learning to tackle the challenging exploration-exploitation tradeoff in those domains.

References

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

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

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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

[7] A. Fuchs and G. Letta. L'inégalité de kullback. application à la théorie de l'estimation. *Séminaire de probabilités de Strasbourg*, 4:108–131, 1970.

[8] A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[9] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

[10] R. M. Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[11] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017.

[12] J. Honda. A note on kl-ucb+ policy for the stochastic bandit. *arXiv preprint arXiv:1903.07839*, 2019.

[13] B. Jorgensen. *The theory of dispersion models*. CRC Press, 1997.

[14] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, pages 592–600, 2012.

[15] T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

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

[17] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[18] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

[19] J. T. Lewis and R. Russell. An introduction to large deviations for teletraffic engineers, 1997.

[20] E. Mazumdar, A. Pacchiano, Y.-a. Ma, P. L. Bartlett, and M. I. Jordan. On thompson sampling with langevin algorithms. *arXiv preprint arXiv:2002.10002*, 2020.

[21] B. O'Donoghue. Variational Bayesian reinforcement learning with regret bounds. *arXiv preprint arXiv:1807.09647*, 2018.

[22] B. O'Donoghue, T. Lattimore, and I. Osband. Stochastic matrix games with bandit feedback. *arXiv preprint arXiv:2006.05145*, 2020.

[23] B. O'Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih. Combining policy gradient and Q-learning. In *International Conference on Learning Representations (ICLR)*, 2017.

[24] B. O'Donoghue, I. Osband, and C. Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations (ICLR)*, 2020.

[25] M. Phan, Y. A. Yadkori, and J. Domke. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems*, pages 8801–8811, 2019.

[26] P. H. Richemond and B. Maginnis. A short variational proof of equivalence between policy gradients and soft q learning. *arXiv preprint arXiv:1712.08650*, 2017.

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

[28] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

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

[30] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[31] S. S. Varadhan. *Large deviations and applications*. SIAM, 1984.

[32] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

[33] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a primarily theoretical paper with limited societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We included a description of the data generation process for the simulations we ran.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]