Thompson Sampling for Complex Online Problems

Aditya Gopalan ADITYA@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel

Shie Mannor SHIE@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel

Yishay Mansour MANSOUR@TAU.AC.IL
School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Abstract

We consider stochastic multi-armed bandit problems with complex actions over a set of basic arms, where the decision maker plays a complex action rather than a basic arm in each round. The reward of the complex action is some function of the basic arms' rewards, and the feedback observed may not necessarily be the reward per-arm. For instance, when the complex actions are subsets of the arms, we may only observe the maximum reward over the chosen subset. Thus, feedback across complex actions may be coupled due to the nature of the reward function. We prove a frequentist regret bound for Thompson sampling in a very general setting involving parameter, action and observation spaces and a likelihood function over them. The bound holds for discretely-supported priors over the parameter space without additional structural properties such as closed-form posteriors, conjugate prior structure or independence across arms. The regret bound scales logarithmically with time but, more importantly, with an improved constant that non-trivially captures the coupling across complex actions due to the structure of the rewards. As applications, we derive improved regret bounds for classes of complex bandit problems involving selecting subsets of arms, including the first nontrivial regret bounds for nonlinear MAX reward feedback from subsets. Using particle filters for computing posterior distributions which lack an explicit closed-form, we present numerical results for the performance of Thompson sampling for subset-selection and job scheduling problems.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

1. Introduction

The stochastic Multi-Armed Bandit (MAB) is a classical framework in machine learning and optimization. In the basic MAB setting, there is a finite set of actions, each of which has a reward derived from some stochastic process, and a learner selects actions to optimize long-term performance. The MAB model gives a crystallized abstraction of a fundamental decision problem – whether to explore or exploit in the face of uncertainty. Bandit problems have been extensively studied and several well-performing methods are known for optimizing the reward (Gittins et al., 2011; Auer et al., 2002; Audibert & Bubeck, 2009; Garivier & Cappé, 2011). However, the requirement that the actions' rewards be independent is often a severe limitation, as seen in these examples:

Web Advertising: Consider a website publisher selecting at each time a subset of ads to be displayed to the user. As the publisher is paid per click, it would like to maximize its revenue, but dependencies between different ads could mean that the problem does not "decompose nicely". For instance, showing two car ads might not significantly increase the click probability over a single car ad.

Job Scheduling: Assume we have a small number of resources or machines, and in each time step we receive a set of jobs (the "basic arms"), where the duration of each job follows some fixed but unknown distribution. The latency of a machine is the sum of the latencies of the jobs (basic arms) assigned to it, and the makespan of the system is the maximum latency across machines. Here, the decision maker's complex action is to partition the jobs (basic arms) between the machines, to achieve the least makespan on average.

Routing: Consider a multi-commodity flow problem, where for each source-destination pair, we need to select a route (a complex action). In this setting the capacities of the edges (the basic arms) are random variables, and the reward is the total flow in the system at each time. In this example, the rewards of different paths are inter-dependent, since the flow on one path depends on which other paths where selected.

These examples motivate settings where a model more complex than the simple MAB is required. Our high-level goal is to describe a methodology that can tackle bandit problems with complex action/reward structure, and also guarantee high performance. A crucial complication in the problems above is that it is unlikely that we will get to observe the reward of each basic action chosen. Rather, we can hope to receive only an aggregate reward for the complex action taken. Our approach to complex bandit problems stems from the idea that when faced with uncertainty, pretending to be Bayesian can be advantageous. A purely Bayesian view of the MAB assumes that the model parameters (i.e., the arms' distributions) are drawn from a prior distribution. We argue that even in a frequentist setup, in which the stochastic model is unknown but fixed, working with a fictitious prior over the model (i.e., being *pseudo-Bayesian*) helps solve very general bandit problems with complex actions and observations.

Our algorithmic prescription for complex bandits is Thompson sampling (Thompson, 1933; Scott, 2010; Agrawal & Goyal, 2012): Start with a fictitious prior distribution over the parameters of the basic arms of the model, whose posterior gets updated as actions are played. A parameter is randomly drawn according to the posterior and the (complex) action optimal for the parameter is played. The rationale behind this is twofold: (1) Updating the posterior adds useful information about the true unknown parameter. (2) Correlations among complex bandit actions (due to their dependence on the basic parameters) are implicitly captured by posterior updates on the space of basic parameters.

The main advantage of a pseudo-Bayesian approach like Thompson sampling, compared to other MAB methodologies such as UCB, is that it can handle a wide range of information models that go beyond observing the individual rewards alone. For example, suppose we observe only the final makespan in the multi-processor job scheduling problem above. In Thompson sampling, we merely need to compute a posterior given this observation and its likelihood. In contrast, it seems difficult to adapt an algorithm such as UCB for this case without a naive exponential dependence on the number of basic arms[1]. Besides, the deterministic approach of optimizing over regions of the parameter space that UCB-like algorithms follow (Dani et al., 2008; Abbasi-Yadkori et al., 2011) is arguably harder to apply in practice, as opposed to optimizing over the action space given a sampled parameter in Thompson sampling – often an efficient polynomial-time routine like a sort. The Bayesian view that motivates Thompson sampling also allows us to use efficient numerical algorithms such as particle filtering (Ristic et al., 2004; Doucet et al., 2001) to approximate complicated posterior distributions in practice.

Our main analytical result is a general regret bound for Thompson sampling in complex bandit settings. No specific structure is imposed on the initial (fictitious) prior, except that it be discretely supported and put nonzero mass on the true model. The bound for this general setting scales logarithmically with time[2], as is well-known. But more interestingly, the preconstant for this logarithmic scaling can be explicitly characterized in terms of the bandit's KL divergence geometry and represents the *information complexity* of the bandit problem. The standard MAB imposes no structure among the actions, thus its information complexity simply becomes a sum of terms, one for each separate action. However, in a complex bandit setting, rewards are often more informative about other parameters of the model, in which case the bound reflects the resulting coupling across complex actions.

Recent work has shown the regret-optimality of Thompson sampling for the basic MAB (Agrawal & Goyal, 2012; Kaufmann et al., 2012), and has even provided regret bounds for a specific complex bandit setting – the linear bandit case where the reward is a linear function of the actions (Agrawal & Goyal, 2011). However, the analysis of complex bandits in general poses challenges that cannot be overcome using the specialized techniques in these works. Indeed, these existing analyses rely crucially on the conjugacy of the prior and posterior distributions – either independent Beta or exponential family distributions for basic MAB or standard normal distributions for linear bandits. These methods break down when analyzing the

[1]The work of Dani et al. (Dani et al., 2008) first extended the UCB framework to the case of linear cost functions. However, for more complex, nonlinear rewards (e.g., multi-commodity flows or makespans), it is unclear how UCB-like algorithms can be applied other than to treat all complex actions independently.

[2]More precisely, we obtain a bound of the form $\mathsf{B} + \mathsf{C} \log T$, in which C is a non-trivial preconstant that captures precisely the structure of correlations among actions and thus is often better than the decoupled sum-of-inverse-KL-divergences bounds seen in literature (Lai & Robbins, 1985). The additive constant (wrt time) B, though potentially large and depending on the total number of complex actions, appears to be merely an artifact of our proof technique tailored towards extracting the time scaling C. This is borne out, for instance, from numerical experiments on complex bandit problems in Section 5. We remark that such additive constants, in fact, often appear in regret analyses of basic Thompson sampling (Kaufmann et al., 2012; Agrawal & Goyal, 2012).

evolution of complicated posterior distributions which often lack even a closed form expression.

In contrast to existing regret analyses, we develop a novel proof technique based on looking at the form of the Bayes posterior. This allows us to track the posterior distributions that result from general action and feedback sets, and to express the concentration of the posterior as a constrained optimization problem in path space. It is rather surprising that, with almost no specific structural assumptions on the prior, our technique yields a regret bound that reduces to Lai and Robbins' classic lower bound for standard MAB, and also gives non-trivial and improved regret scalings for complex bandits. In this vein, our results represent a generalization of existing performance results for Thompson sampling.

We complement our theoretical findings with numerical studies of Thompson sampling. The algorithm is implemented using a simple particle filter (Ristic et al., 2004) to maintain and sample from posterior distributions. We evaluate the performance of the algorithm on two complex bandit scenarios – subset selection from a bandit and job scheduling.

Related Work: Bayesian ideas for the multi-armed bandit date back nearly 80 years ago to the work of W. R. Thompson (Thompson, 1933), who introduced an elegant algorithm based on posterior sampling. However, there has been relatively meager work on using Thompson sampling in the control setup. A notable exception is (Ortega & Braun, 2010) that develops general Bayesian control rules and demonstrates them for classic bandits and Markov decision processes (i.e., reinforcement learning). On the empirical side, a few recent works have demonstrated the success of Thompson sampling (Scott, 2010; Chapelle & Li, 2011). Recent work has shown frequentist-style regret bounds for Thompson sampling in the standard bandit model (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Korda et al., 2013), and Bayes risk bounds in the purely Bayesian setting (Osband et al., 2013). Our work differs from this literature in that we go beyond simple, decoupled actions/observations – we focus on the performance of Thompson setting in a general action/feedback model, and show novel frequentist regret bounds that account for the structure of complex actions.

Regarding bandit problems with actions/rewards more complex than the basic MAB, a line of work that deserves particular mention is that of linear bandit optimization (Auer, 2003; Dani et al., 2008; Abbasi-Yadkori et al., 2011). In this setting, actions are identified with decision vectors in a Euclidean space, and the obtained rewards are random linear functions of actions, drawn from an unknown distribution. Here, we typically see regret bounds for generalizations of the UCB algorithm that show poly-logarithmic regret for this setting. However, the methods and bounds are highly tailored to the specific linear feedback structure and do not carry over to other kinds of feedback.

2. Setup and Notation

We consider a general stochastic model $X_1, X_2, ...$ of independent and identically distributed random variables living in a space $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}^{\mathsf{N}}$ if there is an underlying N-armed basic bandit – we will revisit this in detail in Section 4.1). The distribution of each X_t is parametrized by $\theta^* \in \Theta$, where Θ denotes the parameter space. At each time t, an action A_t is played from an action set $\mathcal{A}$, following which the decision maker obtains a stochastic observation $Y_t = f(X_t, A_t) \in \mathcal{Y}$, the observation space, and a scalar reward $g(f(X_t, A_t))$. Here, f and g are general fixed functions, and we will often denote $g \circ f$ by the function[3] h. We denote by $l(y; a, \theta)$ the likelihood of observing y upon playing action a when the distribution parameter is θ, i.e.,[4] $l(y; a, \theta) := \mathbb{P}_\theta[f(X_1, a) = y]$.

For $\theta \in \Theta$, let $a^*(\theta)$ be an action that yields the highest expected reward for a model with parameter θ, i.e., $a^*(\theta) := \arg\max_{a \in \mathcal{A}} \mathbb{E}_\theta[h(X_1, a)]$.[5]. We use $e^{(j)}$ to denote the j-th unit vector in finite-dimensional Euclidean space.

The goal is to play an action at each time t to minimize the (expected) *regret* over T rounds: $R_T := \sum_{t=1}^{T} h(X_t, a^*(\theta^*)) - h(X_t, A_t)$, or alternatively, the number of plays of suboptimal actions[6]: $\sum_{t=1}^{T} \mathbf{1}\{A_t \neq a^*\}$.

Remark: Our main result also holds in a more general stochastic bandit model $(\Theta, \mathcal{Y}, \mathcal{A}, l, \hat{h})$ without the need for the underlying "basic arms" $\{X_i\}_i$ and the basic ambient space $\mathcal{X}$. In this case we require $l(y; a, \theta) := \mathbb{P}_\theta[Y_1 = y | A_1 = a]$, $\hat{h} : \mathcal{Y} \to \mathbb{R}$ (the reward function), $a^*(\theta) := \arg\max_{a \in \mathcal{A}} \mathbb{E}_\theta[\hat{h}(Y_1) | A_1 = a]$, and the regret $R_T := T\hat{h}(Y_0) - \sum_{t=1}^{T} \hat{h}(Y_t)$ where $\mathbb{P}[Y_0 = \cdot] = l(\cdot; a^*(\theta^*), \theta^*)$.

For each action $a \in \mathcal{A}$, define $S_a := \{\theta \in \Theta : a^*(\theta) = a\}$ to be the *decision region* of a, i.e., the set of models in Θ whose optimal action is a. We use θ_a to denote the marginal probability distribution, under model θ, of the output upon

[3] e.g., when A_t is a subset of basic arms, $h(X_t, A_t)$ could denote the maximum reward from the subset of coordinates of X_t corresponding to A_t.

[4] Finiteness of $\mathcal{Y}$ is implicitly assumed for the sake of clarity. In general, when $\mathcal{Y}$ is a Borel subset of $\mathbb{R}^{\mathsf{N}}$, $l(\cdot; a, \theta)$ will be the corresponding N-dimensional density, etc.

[5] The absence of a subscript is to be understood as working with the parameter θ^*.

[6] We refer to the latter objective as regret since, under bounded rewards, both the objectives scale similarly with the problem size.

Algorithm 1 Thompson Sampling

Input: Parameter space Θ, action space $\mathcal{A}$, output space $\mathcal{Y}$, likelihood $l(y; a, \theta)$.
Parameter: Distribution π over Θ.
Initialization: Set $\pi_0 = \pi$.

for each $t = 1, 2, \ldots$

1. Draw $\theta_t \in \Theta$ according to the distribution π_{t-1}.
2. Play $A_t = a^*(\theta_t) := \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta_t}[h(X_1, a)]$.
3. Observe $Y_t = f(X_t, A_t)$.
4. (Posterior Update) Set the distribution π_t over Θ to

$$\forall S \subseteq \Theta : \quad \pi_t(S) = \frac{\int_S l(Y_t; A_t, \theta)\pi_{t-1}(d\theta)}{\int_\Theta l(Y_t; A_t, \theta)\pi_{t-1}(d\theta)}.$$

end for

playing action a, i.e., $\{l(y; a, \theta) : y \in \mathcal{Y}\}$. Moreover, set $D_\theta := (D(\theta_a^*||\theta_a))_{a \in \mathcal{A}}$. Within S_a, let S_a' be the models that *exactly match* θ^* in the sense of the marginal distribution of action a^*, i.e., $S_a' := \{\theta \in S_a : D(\theta_{a^*}^*||\theta_{a^*}) = 0\}$, where $D(\phi||\zeta)$ is the standard Kullback-Leibler divergence between probability distributions ϕ and ζ. Let $S_a'' := S_a \setminus S_a'$ be the remaining models in S_a.

3. Regret Performance: Overview

We propose using Thompson sampling (Algorithm 1) to play actions in the general bandit model. Before formally stating the regret bound, we present an intuitive explanation of how Thompson sampling learns to play good actions in a general setup where observations, parameters and actions are related via a general likelihood. To this end, let us assume that there are finitely many actions $\mathcal{A}$. Let us also index the actions in $\mathcal{A}$ as $\{1, 2, \ldots, |\mathcal{A}|\}$, with the index $|\mathcal{A}|$ denoting the optimal action a^* (we will require this indexing later when we associate each coordinate of $|\mathcal{A}|$-dimensional space with its respective action).

When action A_t is played at time t, the prior density gets updated to the posterior as $\pi_t(d\theta) \propto \exp\left(-\log \frac{l(Y_t; A_t, \theta^*)}{l(Y_t; A_t, \theta)}\right) \pi_{t-1}(d\theta)$. Observe that the conditional expectation of the "instantaneous" log-likelihood ratio $\log \frac{l(Y_t; A_t, \theta^*)}{l(Y_t; A_t, \theta)}$, is simply the appropriate marginal KL divergence, i.e., $\mathbb{E}\left[\log \frac{l(Y_t; A_t, \theta^*)}{l(Y_t; A_t, \theta)} \mid A_t\right] = \sum_{a \in \mathcal{A}} \mathbf{1}\{A_t = a\} D(\theta_a^*||\theta_a)$. Hence, up to a coarse approximation,

$$\log \frac{l(Y_t; A_t, \theta^*)}{l(Y_t; A_t, \theta)} \approx \sum_{a \in \mathcal{A}} \mathbf{1}\{A_t = a\} D(\theta_a^*||\theta_a),$$

with which we can write

$$\pi_t(d\theta) \overset{\propto}{\sim} \exp\left(-\sum_{a \in \mathcal{A}} N_t(a) D(\theta_a^*||\theta_a)\right) \pi_0(d\theta), \quad (1)$$

with $N_t(a) := \sum_{i=1}^t \mathbf{1}\{A_i = a\}$ denoting the play count of a. The quantity in the exponent can be interpreted as a "loss" suffered by the model θ up to time t, and each time an action a is played, θ incurs an additional loss of essentially the marginal KL divergence $D(\theta_a^*||\theta_a)$.

Upon closer inspection, the posterior approximation (1) yields detailed insights into the dynamics of posterior-based sampling. First, since $\exp\left(-\sum_{a \in \mathcal{A}} N_t(a) D(\theta_a^*||\theta_a)\right) \leq 1$, the true model θ^* always retains a significant share of posterior mass: $\pi_t(d\theta^*) \gtrsim \frac{\exp(0)\, \pi_0(d\theta^*)}{\int_\Theta 1\, \pi_0(d\theta)} = \pi_0(d\theta^*)$. This means that Thompson sampling samples θ^*, and hence plays a^*, with at least a constant probability each time, so that $N_t(a^*) = \Omega(t)$.

Suppose we can show that each model in any S_a'', $a \neq a^*$, is such that $D(\theta_{a^*}^*||\theta_{a^*})$ is bounded strictly away from 0 with a gap of $\xi > 0$. Then, our preceding calculation immediately tells us that any such model is sampled at time t with a probability exponentially decaying in t: $\pi_t(d\theta) \lesssim \frac{e^{-\xi\Omega(t)}\pi_0(d\theta)}{\pi_0(d\theta^*)}$; the regret from such S_a''-sampling is *negligible*. On the other hand, how much does the algorithm have to work to make models in S_a', $a \neq a^*$ suffer *large* ($\approx \log T$) losses and thus rid them of significant posterior probability?[7]

A model $\theta \in S_a'$ suffers loss whenever the algorithm plays an action a for which $D(\theta_a^*||\theta_a) > 0$. Hence, several actions can help in making a bad model (or set of models) suffer large enough loss. Imagine that we track the play count vector $N_t := (N_t(a))_{a \in \mathcal{A}}$ in the integer lattice from $t = 0$ through $t = T$, from its initial value $N_0 = (0, \ldots, 0)$. There comes a first time τ_1 when some action $a_1 \neq a^*$ is eliminated (i.e., when all its models' losses exceed $\log T$). The argument of the preceding paragraph indicates that the play count of a_1 will stay fixed at $N_{\tau_1}(a_1)$ for the remainder of the horizon up to T. Moving on, there arrives a time $\tau_2 \geq \tau_1$ when another action $a_2 \notin \{a*, a_1\}$ is eliminated, at which point its play count ceases to increase beyond $N_{\tau_2}(a_2)$, and so on.

To sum up: Continuing until all actions $a \neq a^*$ (i.e., the regions S_a') are eliminated, we have a path-based bound for the total number of times suboptimal actions can be played. If we let $z_k = N_{\tau_k}$, i.e., the play counts of all actions at time τ_k, then for all $i \geq k$ we must have the constraint $z_i(a_k) = z_k(a_k)$ as plays of a_k do not occur after time τ_k.

[7] Note: Plays of a^* do *not* help increase the losses of these models.

Moreover, $\min_{\theta \in S'_{a_k}} \langle z_k, D_\theta \rangle \approx \log T$: action a_k is eliminated precisely at time τ_k. A bound on the total number of bad plays thus becomes

$$\begin{aligned} \max \quad & ||z_k||_1 \\ \text{s.t.} \quad & \exists \text{ play count sequence } \{z_k\}, \\ & \exists \text{ suboptimal action sequence } \{a_k\}, \\ & z_i(a_k) = z_k(a_k), i \geq k, \\ & \min_{\theta \in S'_{a_k}} \quad \langle z_k, D_\theta \rangle \approx \log T, \quad \forall k. \end{aligned} \tag{2}$$

The final constraint above ensures that an action a_k is eliminated at time τ_k, and the penultimate constraint encodes the fact that the eliminated action a_k is not played after time τ_k. The bound not only depends on $\log T$ but also on the KL-divergence geometry of the bandit, i.e., the marginal divergences $D(\theta^*_a || \theta_a)$. Notice that no specific form for the prior or posterior was assumed to derive the bound, save the fact that $\pi_0(d\theta^*) \gtrsim 0$, i.e., that the prior puts "enough" mass on the truth.

In fact, all our approximate calculations leading up to the bound (2) hold rigorously – Theorem 1, to follow, states that under reasonable conditions on the prior, the number of suboptimal plays/regret scales as (2) with high probability. We will also see that the general bound (2) is non-trivial in that (a) for the standard multi-armed bandit, it gives essentially the optimum known regret scaling, and (b) for a family of complex bandit problems, it can be significantly less than the one obtained by treating all actions separately.

4. Regret Performance: Formal Results

Our main result is a high-probability large-horizon regret bound[8] for Thompson sampling. The bound holds under the following mild assumptions about the parameter space Θ, action space $|\mathcal{A}|$, observation space $|\mathcal{Y}|$, and the fictitious prior π.

Assumption 1 (Finitely many actions, observations)**.** $|\mathcal{A}|, |\mathcal{Y}| < \infty$.

Assumption 2 (Finitely supported, "Grain of truth" prior)**.** *(a) The prior distribution π is supported over a finite set: $|\Theta| < \infty$, (b) $\theta^* \in \Theta$ and $\pi(\theta^*) > 0$. Furthermore, (c) there exists $\Gamma \in (0, 1/2)$ such that $\Gamma \leq l(y; a, \theta) \leq 1 - \Gamma$ $\forall \theta \in \Theta, a \in \mathcal{A}, y \in \mathcal{Y}$.*

Remark: We emphasize that the finiteness assumption on the prior is made primarily for technical tractability, without compromising the key learning dynamics of Thompson sampling perform well. In a sense, a continuous prior can be approximated by a fine enough discrete prior without affecting the geometric structure of the parameter space. The core ideas driving our analysis explain why Thompson sampling provably performs well in very general action-observation settings, and, we believe, can be made general enough to handle even continuous priors/posteriors. However, the issues here are primarily measure-theoretic: much finer control will be required to bound and track posterior probabilities in the latter case, perhaps requiring the design of *adaptive* neighbourhoods of θ^* with sufficiently large posterior probability that depend on the evolving history of the algorithm. It is not clear to us how such regions may be constructed for obtaining regret guarantees in the case of continuous priors. We thus defer this highly nontrivial task to future work.

Assumption 3 (Unique best action)**.** *The optimal action in the sense of expected reward is unique*[9]*, i.e.,* $\mathbb{E}[h(X_1, a^*)] > \max_{a \in \mathcal{A}, a \neq a^*} \mathbb{E}[h(X_1, a)]$.

We now state the regret bound for Thompson sampling for general stochastic bandits. The bound is a rigorous version of the path-based bound presented earlier, in Section 3.

Theorem 1 (General Regret Bound for Thompson Sampling)**.** *Under Assumptions 1-3, the following holds for the Thompson Sampling algorithm. For $\delta, \epsilon \in (0, 1)$, there exists $T^\star \geq 0$ such that for all $T \geq T^\star$, with probability at least $1 - \delta$, $\sum_{a \neq a^*} N_T(a) \leq \mathsf{B} + \mathsf{C}(\log T)$, where $\mathsf{B} \equiv \mathsf{B}(\delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi)$ is a problem-dependent constant that does not depend on T, and* [10]*:*

$$\begin{aligned} & \mathsf{C}(\log T) := \\ \max \quad & \sum_{k=1}^{|\mathcal{A}|-1} z_k(a_k) \\ \textit{s.t.} \quad & z_k \in \mathbb{Z}_+^{|\mathcal{A}|-1} \times \{0\}, a_k \in \mathcal{A} \setminus \{a^*\}, k < |\mathcal{A}|, \\ & z_i \succeq z_k, z_i(a_k) = z_k(a_k), i \geq k, \\ & \forall 1 \leq j, k \leq |\mathcal{A}| - 1 : \\ & \quad \min_{\theta \in S'_{a_k}} \quad \langle z_k, D_\theta \rangle \geq \frac{1+\epsilon}{1-\epsilon} \log T, \\ & \quad \min_{\theta \in S'_{a_k}} \quad \langle z_k - e^{(j)}, D_\theta \rangle < \frac{1+\epsilon}{1-\epsilon} \log T. \end{aligned} \tag{3}$$

The proof is in the Appendix of the supplementary material, and uses a recently developed self-normalized concentration inequality (Abbasi-Yadkori et al., 2011) to help

[8] More precisely, we bound the number of plays of suboptimal actions. A bound on the standard regret can also be obtained easily from this, via a self-normalizing concentration inequality we use in this paper (see Appendices). However, we avoid stating this in the interest of minimizing clutter in the presentation, since there will be additional $O(\sqrt{\log T})$ terms in the bound on standard regret.

[9] This assumption is made only for the sake of notational ease, and does not affect the paper's results in any significant manner.

[10] $\mathsf{C}(\log T) \equiv \mathsf{C}(T, \delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi)$ in general, but we suppress the dependence on the problem parameters $\delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi$ as we are chiefly concerned with the time scaling.

track the sample path evolution of the posterior distribution in its general form. The power of Theorem 1 lies in the fact that it accounts for coupling of information across complex actions and give improved structural constants for the regret scaling than the standard decoupled case, as we show[11] in Corollaries 1 and 2. We also prove Proposition 2, which explicitly quantifies the improvement over the naive regret scaling for general complex bandit problems as a function of marginal KL-divergence separation in the parameter space Θ.

4.1. Playing Subsets of Bandit Arms and Observing "Full Information"

Let us take a standard N-armed Bernoulli bandit with arm parameters $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_N$. Suppose the (complex) actions are all size M subsets of the N arms. Following the choice of a subset, we get to observe the rewards of *all* M chosen arms (also known as the "semi-bandit" setting (Audibert et al., 2011)) and receive some bounded reward of the chosen arms (thus, $\mathcal{Y} = \{0,1\}^M$, $\mathcal{A} = \{S \subset [N] : |S| = M\}$, $f(\cdot, A)$ is simply the projection onto coordinates of $A \in \mathcal{A}$, and $g : \mathbb{R}^M \to [0,1]$, e.g., average or sum).

A natural finite prior for this problem can be obtained by discretizing each of the N basic dimensions and putting uniform mass over all points: $\Theta = \left\{\beta, 2\beta, \ldots \left(\lfloor \frac{1}{\beta} \rfloor - 1\right)\beta\right\}^N$, $\beta \in (0,1)$, and $\pi(\theta) = \frac{1}{|\Theta|}$ $\forall \theta \in \Theta$. We can then show, using Theorem 1, that

Corollary 1 (Regret for playing subsets of basic arms, Full feedback)**.** *Suppose $\mu \equiv (\mu_1, \mu_2, \ldots, \mu_N) \in \Theta$ and $\mu_{N-M} < \mu_{N-M+1}$. Then, the following holds for the Thompson sampling algorithm for $\mathcal{Y}$, $\mathcal{A}$, f, g, Θ and π as above. For $\delta, \epsilon \in (0,1)$, there exists $T^\star \geq 0$ such that for all $T \geq T^\star$, with probability at least $1-\delta$, $\sum_{a \neq a^*} N_T(a) \leq \mathsf{B}_2 + \left(\frac{1+\epsilon}{1-\epsilon}\right) \sum_{i=1}^{N-M} \frac{1}{D(\mu_i || \mu_{N-M+1})} \log T$, where $\mathsf{B}_2 \equiv \mathsf{B}_2(\delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi)$ is a problem-dependent constant that does not depend on T.*

This result, proved in the Appendix of the supplementary material, illustrates the power of additional information from observing several arms of a bandit at once. Even though the total number of actions $\binom{N}{M}$ is at worst exponential in M, the regret bound scales only as $O((N-M)\log T)$. Note also that for $M = 1$ (the standard MAB setting), the regret scaling is essentially $\sum_{i=1}^{N-M} \frac{1}{D(\mu_i || \mu_{N-M+1})} \log T$, which is interestingly the optimal regret scaling for standard Bernoulli bandits obtained by specialized algorithms for decoupled bandit arms such as KL-UCB (Garivier & Cappé, 2011) and, more recently, Thompson Sampling with the independent Beta prior (Kaufmann et al., 2012).

[11] We remark that though the non-scaling (with T) additive constant B might appear large, we believe it is an artifact of our proof technique tailored to extract the time scaling of the regret. Indeed, numerical results in Section 5 show practically no additive factor behaviour.

4.2. A General Regret Improvement Result & Application to MAX Subset Regret

Using the same setting and size-M subset actions as before but *not* being able to observe all the individual arms' rewards results in much more challenging bandit settings. Here, we consider the case where we get to observe as the reward *only* the maximum value of M chosen arms of a standard N-armed Bernoulli bandit (i.e., $f(x, A) := \max_{i \in A} x_i$ and $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x$). The feedback is still aggregated across basic arms, but at the same time very different from the full information case, e.g., observing a reward of 0 is very uninformative whereas a value of 1 is highly informative about the constituent arms.

We can again apply the general machinery provided by Theorem 1 to obtain a non-trivial regret bound for observing the highly nonlinear MAX reward. Along the way, we derive the following consequence of Theorem 1, useful in its own right, that explicitly guarantees an improvement in regret directly based on the Kullback-Leibler resolvability of parameters in the parameter space – a measure of coupling across complex actions.

Proposition 2 (Explicit Regret Improvement Based on Marginal KL-divergences)**.** *Let T be large enough such that $\max_{\theta \in \Theta, a \in \mathcal{A}} D(\theta_a^* || \theta_a) \leq \frac{1+\epsilon}{1-\epsilon} \log T$. Suppose $\Delta \leq \min_{a \neq a^*, \theta \in S_a'} D(\theta_a^* || \theta_a)$, and that the integer L is such that for every $a \neq a^*$ and $\theta \in S_a'$, $|\{\hat{a} \in \mathcal{A} : \hat{a} \neq a^*, D(\theta_{\hat{a}}^* || \theta_{\hat{a}}) \geq \Delta\}| \geq L$, i.e., at least L coordinates of D_θ (excluding the $|\mathcal{A}|$-th coordinate a^*) are at least Δ. Then, $\mathsf{C}(\log T) \leq \left(\frac{|\mathcal{A}|-L}{\Delta}\right) \frac{2(1+\epsilon)}{1-\epsilon} \log T$.*

Note that the result assures a non-trivial additive reduction of $\Omega\left(\frac{L}{\Delta} \log T\right)$ from the naive decoupled regret, whenever any suboptimal model in Θ can be resolved apart from θ^* by at least L actions in the sense of marginal KL-divergences of their observations. Its proof is contained in the Appendix in the supplementary material.

Turning to the MAX reward bandit, let $\beta \in (0,1)$, and suppose that $\Theta = \{1-\beta^R, 1-\beta^{R-1}, \ldots, 1-\beta^2, 1-\beta\}^N$, for positive integers R and N. As before, let $\mu \in \Theta$ denote the basic arms' parameters, and let $\mu_{min} := \min_{a \in \mathcal{A}} \prod_{i \in a}(1-\mu_i)$, and $\pi(\theta) = \frac{1}{|\Theta|}$ $\forall \theta \in \Theta$. The action and observation spaces $\mathcal{A}$ and $\mathcal{Y}$ are the same as those in Section 4.1, but the feedback function here is $f(x, a) := \max_{i \in a} x_i$, and g is the identity on $\mathbb{R}$. An application of our general regret improvement result (Proposition 2) now gives, for the highly nonlinear MAX reward

function,

Corollary 2 (Regret for playing subsets of basic arms, MAX feedback)**.** *The following holds for the Thompson sampling algorithm for $\mathcal{Y}$, $\mathcal{A}$, f, g, Θ and π as above. For $0 \leq M \leq N$, $M \neq \frac{N}{2}$, $\delta, \epsilon \in (0,1)$, there exists $T^\star \geq 0$ such that for all $T \geq T^\star$, with probability at least $1-\delta$, $\sum_{a \neq a^*} N_T(a) \leq \mathsf{B}_3 + (\log 2)\left(\frac{1+\epsilon}{1-\epsilon}\right)\left[1+\binom{N-1}{M}\right]\frac{\log T}{\mu^2_{\min}(1-\beta)}$.*

Observe that this regret bound is of the order of $\binom{N-1}{M}\frac{\log T}{\mu^2_{\min}}$, which is significantly less than the standard decoupled bound of $|\mathcal{A}|\frac{\log T}{\mu^2_{\min}} = \binom{N}{M}\frac{\log T}{\mu^2_{\min}}$ by a multiplicative factor of $\frac{\binom{N-1}{M}}{\binom{N}{M}} = \frac{N-M}{N}$, or by an additive factor of $\binom{N-1}{M-1}\frac{\log T}{\mu^2_{\min}}$. In fact, though this is a provable reduction in the regret scaling, the actual reduction is likely to be much better in practice – the experimental results in Section 5 attest to this. The proof of this result uses sharp combinatorial estimates relating to vertices on the N-dimensional hypercube (Ahlswede et al., 2003), and can be found in the Appendix in the supplementary material.

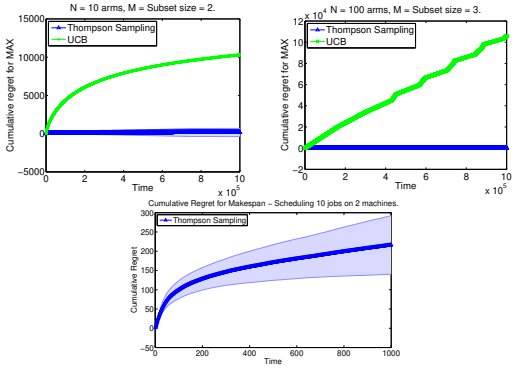


Figure 1. **Top Left and Top Right:** Cumulative regret with observing the maximum of a pair out of 10 arms (left), and that of a triple out of 100 arms (center), for (a) Thompson sampling using a particle filter, and (b) UCB treating each subset as a separate actions. The arm means are chosen to be equally spaced in $[0,1]$. The regret is averaged across 150 runs, and the confidence intervals shown are ± 1 standard deviation. **Bottom:** Cumulative regret with respect to the best makespan with particle-filter-based Thompson sampling, for scheduling 10 jobs on 2 machines. The job means are chosen to be equally spaced in $[0, 10]$. The best job assignment gives an expected makespan of 31. The regret is averaged across 150 runs, and the confidence intervals shown are ± 1 standard deviation.

5. Numerical Experiments

We evaluate the performance of Thompson sampling (Algorithm 1) on two complex bandit settings – (a) Playing subsets of arms with the MAX reward function, and (b) Job scheduling over machines to minimize makespan. Where the posterior distribution is not closed-form, we approximate it using a particle filter (Ristic et al., 2004; Doucet et al., 2001), allowing efficient updates after each play.

1. Subset Plays, MAX Reward: We assume the setup of Section 4.2 where one plays a size-M subset in each round and observes the maximum value. The individual arms' reward parameters are taken to be equi-spaced in $(0,1)$. It is observed that Thompson sampling outperforms standard "decoupled" UCB by a wide margin in the cases we consider (Figure 1, left and center). The differences are especially pronounced for the larger problem size $N = 1000, M = 3$, where UCB, that sees $\binom{N}{M}$ separate actions, appears be in the exploratory phase throughout.

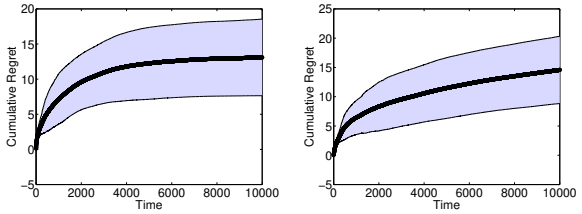
Figure 2 affords a closer look at the regret for the above problem, and presents the results of using a flat prior over a uniformly discretized grid of models in $[0,1]^{10}$ – the setting of Theorem 1.

2. Subset Plays, Average Reward: We apply Thompson sampling again to the problem of choosing the best M out of N basic arms of a Bernoulli bandit, but this time receiving a reward that is the *average value* of the chosen subset. This specific form of the feedback makes it possible to use a *continuous, Gaussian prior* density over the space of basic parameters that is updated to a Gaussian posterior assuming a fictitious Gaussian likelihood model (Agrawal & Goyal, 2011). This is a fast, practical alternative to UCB-style deterministic methods (Dani et al., 2008; Abbasi-Yadkori et al., 2011) which require performing a convex optimization every instant. Figure 3 shows the regret of Thompson sampling with a Gaussian prior/posterior for choosing various size M subsets $(5, 10, 20, 50)$ out of $N = 100$ arms. It is practically impossible to naively apply a decoupled bandit algorithm over such a problem due to the very large number of complex actions (e.g., there are $\approx 10^{13}$ actions even for $M = 10$)[12]. However, Thompson sampling merely samples from a $N = 100$ dimensional Gaussian and picks the best M coordinates of the sample, which yields a dramatic reduction in running time. The constant factors in the regret curves are seen to be modest when compared to the total number of complex actions.

3. Job Scheduling: We consider a stochastic job-scheduling problem in order to illustrate the versatility of Thompson sampling for bandit settings more complicated than subset actions. There are $N = 10$ types of jobs and 2

[12]Both the ConfidenceBall algorithm of Dani et al. (Dani et al., 2008) and the OFUL algorithm (Abbasi-Yadkori et al., 2011) are designed for linear feedback from coupled actions via the use of tight confidence sets. However, as stated, they require searching over the space of all actions/subsets. Thus, we remain unclear about how one might efficiently apply them here.

machines. Every job type has a different, unknown mean duration, with the job means taken to be equally spaced in $[0, N]$, i.e., $\frac{iN}{N+1}$, $i = 1, \ldots, N$. At each round, one job of each type arrives to the scheduler, with a random duration that follows the exponential distribution with the corresponding mean. All jobs must be scheduled on one of two possible machines. The loss suffered upon scheduling is the *makespan*, i.e., the maximum of the two job durations on the machines. Once the jobs in a round are assigned to the machines, only the *total* durations on the machines machines can be observed, instead of the individual job durations. Figure 1 (right) shows the results of applying Thompson sampling with an exponential prior for the jobs' means along with a particle filter.



(a) $N = 10, M = 3$ (b) $N = 10, M = 4$

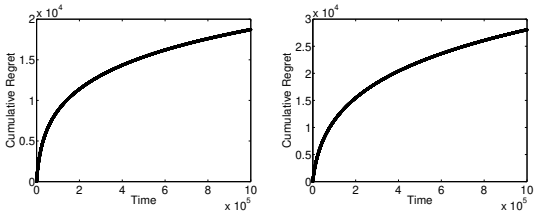
(c) $N = 10, M = 5$ (d) $N = 10, M = 6$

Figure 2. Cumulative regret with observing the maximum value of M out of $N = 10$ arms for Thompson sampling. The prior is uniform over the discrete domain $\{0.1, 0.3, 0.5, 0.7, 0.9\}^N$, with the arms' means lying in the same domain (setting of Theorem 1). The regret is averaged across 10 runs, and the confidence intervals shown are ± 1 standard deviation.

6. Discussion & Future Work

We applied Thompson sampling to balance exploration and exploitation in bandit problems where the action/observation space is complex. Using a novel technique of viewing posterior evolution as a path-based optimization problem, we developed a generic regret bound for Thompson sampling with improved constants that capture the structure of the problem. In practice, the algorithm is easy to implement using sequential Monte-Carlo methods such as particle filters.

Moving forward, the technique of converting posterior concentration to an optimization involving exponentiated KL divergences could be useful in showing *adversarial* regret bounds for Bayesian-inspired algorithms. It is reasonable



(a) $(100, 5)$ (b) $(100, 10)$

(c) $(100, 20)$ (d) $(100, 50)$

Figure 3. Cumulative regret for (N, M): Observing the average value of M out of $N = 10$ arms for Thompson sampling. The prior is a standard normal independent density over N dimensions, and the posterior is also normal under a Gaussian likelihood model. The regret is averaged across 10 runs. Confidence intervals are ± 1 standard deviation.

to posit that Thompson sampling would work well in a range of complex learning settings where a suitable point estimate is available. As an example, optimal bidding for online repeated auctions depending on continuous bid reward functions can be potentially learnt by constructing an estimate of the bid curve.

Another unexplored direction is handling large scale reinforcement learning problems with complex, state-dependent Markovian dynamics. It would be promising if computationally demanding large-state space MDPs could be solved using a form of Thompson sampling by policy iteration after sampling from a parameterized set of MDPs; this has previously been shown to work well in practice (Poupart, 2010; Ortega & Braun, 2010). We can also attempt to develop a theoretical understanding of pseudo-Bayesian learning for complex spaces like the X-armed bandit problem (Srinivas et al., 2010; Bubeck et al., 2011) with a continuous state space. At a fundamental level, this could result in a rigorous characterization of Thompson sampling/pseudo-Bayesian procedures in terms of the value of information per learning step.

Acknowledgements: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC Grant Agreement No 306638. It has also been supported in part by The Israeli Centers of Research Excellence (I-CORE) program (Center No. 4/11), by a grant from the Israel Science Foundation, and by a grant from the United States-Israel Binational Science Foundation (BSF).

References

Abbasi-Yadkori, Yasin, Pal, David, and Szepesvari, Csaba. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Agrawal, Shipra and Goyal, Navin. Thompson sampling for contextual bandits with linear payoffs. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Agrawal, Shipra and Goyal, Navin. Analysis of Thompson sampling for the multi-armed bandit problem. *Journal of Machine Learning Research - Proceedings Track*, 23: 39.1–39.26, 2012.

Ahlswede, R., Aydinian, H., and Khachatrian, L. Maximum number of constant weight vertices of the unit n-cube contained in a k-dimensional subspace. *Combinatorica*, 23(1):5–22, 2003. ISSN 0209-9683.

Audibert, Jean-Yves and Bubeck, Sébastien. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory (COLT)*, pp. 773–818, 2009.

Audibert, Jean-Yves, Bubeck, Sébastien, and Lugosi, Gábor. Minimax policies for combinatorial prediction games. In *Conference on Learning Theory (COLT)*, pp. 107–132, 2011.

Auer, Peter. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2003.

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

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695, 2011.

Chapelle, Olivier and Li, Lihong. An empirical evaluation of Thompson sampling. In *NIPS-11*, 2011.

Dani, Varsha, Hayes, Thomas P., and Kakade, Sham M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Doucet, A., Freitas, N. De, and Gordon, N. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

Garivier, Aurélien and Cappé, Olivier. The KL-UCB algorithm for bounded stochastic bandits and beyond. *Journal of Machine Learning Research - Proceedings Track*, 19:359–376, 2011.

Gittins, J. C., Glazebrook, K. D., and Weber, R. R. *Multi-Armed Bandit Allocation Indices*. Wiley, 2011.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson sampling: An asymptotically optimal finite-time analysis. In *Conference on Algorithmic Learning Theory (ALT)*, 2012.

Korda, Nathaniel, Kaufmann, Emilie, and Munos, Remi. Thompson sampling for 1-dimensional exponential family bandits. In *NIPS*, 2013.

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

Ortega, P. A. and Braun, D. A. A minimum relative entropy principle for learning and acting. *JAIR*, 38:475–511, 2010.

Osband, I., Russo, D., and Roy, B. Van. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, 2013.

Poupart, Pascal. *Encyclopedia of Machine Learning*. Springer, 2010. ISBN 978-0-387-30768-8.

Ristic, B., Arulampalam, S., and Gordon, N. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004.

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

Srinivas, Niranjan, Krause, Andreas, Kakade, Sham, and Seeger, Matthias. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, pp. 1015–1022, 2010.

Thompson, William R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 24(3–4):285–294, 1933.