Thompson Sampling for Stochastic Control: The Finite Parameter Case

Michael Jong Kim

Abstract—In this paper, we apply Thompson sampling to a class of average reward stochastic control problems with parameter uncertainty. Specifically, we study an average reward stochastic control problem over an infinite horizon in which both the reward and state transition distributions are parameterized by an unknown parameter taking values in a finite space. The main result of this paper is a proof showing that Thompson sampling achieves a worst-case average per period regret of $O(T^{-1})$, which is asymptotically optimal.

Index Terms—Bayesian learning, Thompson sampling, posterior convergence rate, average regret bounds.

I. INTRODUCTION

This paper is concerned with average reward stochastic control problems over an infinite horizon with parameter uncertainty. We adopt a Bayesian approach for parameter learning in which the decision maker begins with a prior distribution $\pi_0$ on the unknown parameter $\theta \in \mathcal{P}$. This paper is interested in adapting the well-known randomized policy known as Thompson sampling from the multi-armed bandit literature to stochastic control problems (see [1], [2], [3]). Specifically, the policy we study sequentially selects controls as follows: at each stage the decision maker samples a parameter value from the posterior and selects the control that maximizes the long-run expected average reward under the assumption that the sampled value is “correct”. Our goal is to establish theoretical performance guarantees for Thompson sampling as measured by (worst-case) average per period regret. Our main result is a proof showing that Thompson sampling achieves a worst-case average per period regret over $T$-periods of $O(T^{-1})$, which is asymptotically optimal.

We know of only a few recent papers that propose the use of Thompson sampling in a dynamic optimization environment. For example, the papers by Ortega and Braun [4], [5] propose using Thompson sampling for sequential decision problems and were able to establish asymptotic convergence to the optimal policy. Such a result can be viewed as a qualitative limit result, which ensures eventual convergence. In contrast, the regret bounds established in this paper can be viewed as a quantitative limit result, i.e., we characterize the “rate” at which this convergence is achieved. Establishing such a rate is important in real applications in order to quantify the possible loss when applying Thompson sampling over a planning horizon of interest. In this regard, the regret bound of $O(T^{-1})$ established in this paper provides a quantitative guarantee that Thompson sampling will quickly converge to the optimal policy which is a desirable property in practice. Osband et al. [6] consider Thompson sampling applied to a class of reinforcement learning problems over “repeated” finite horizons, and is perhaps the closest paper to our work in that the authors are able to establish a cumulative regret bound of $O(\sqrt{\tau T \log(SA\tau)})$, where $\tau$ is the finite time horizon and $S$ and $A$ are the cardinality of the state and action spaces. Our paper differs from [6] in a number of ways. First, our model formulation and setting are different. Specifically, Osband et al. [6] samples finite horizon policies (through Thompson sampling) at the beginning of each repeat and cannot change the selected policy until the beginning of the next finite horizon. Our paper on the other hand applies Thompson sampling at each stage for an infinite horizon objective and models uncertainty explicitly through a parameter set over which the posterior distributions are defined. Second, the method of proof in [6] is quite different that the method of proof taken in our paper – [6] uses dynamic programming analysis to achieve their regret bounds whereas we take a “sample path” approach to achieve our bound.

This paper complements two streams of literature on the topic of combined statistical learning and stochastic control. The first is the literature on stochastic adaptive control (see [7], [8], [9], [10]), i.e., problems in which the system under control depends on unknown parameters. This literature is also related to the class of problems known as partially observable Markov decision processes (POMDPs), in which (part of) the state process is also hidden from the decision maker and needs to be learned from data (see e.g. [11], [12]). When parameter learning is conducted in a Bayesian fashion, it is known that such problems can be converted into equivalent “fully observable” control problems where the system state is augmented by the posterior distribution for the unknown parameters. Consequently, the resulting state space is in general high-dimensional, which renders dynamic programming methodology intractable for even moderately sized problems (see [13]). While some work has been done on developing computational methods for approximating the optimal adaptive control policy (e.g. diffusion approximation [14], state-space discretization [8], approximate dynamic programming [15], value iteration [12]), very little work has been done on established quantitative performance bounds as measured by regret, which is the focus of this paper.

This paper is also related in spirit to the literature on optimal design of experiments (DOEs), specifically for selecting experimental inputs to learn about parameters in a control...
setting. Readers are referred to the recent contributions of Pronzato [16] and Spall [17] and the references therein, for a nice overview of this subject. From the perspective of optimal experimental design, Thompson sampling can be thought of as an input design that sequentially extracts data from the control problem so as to learn about the unknown parameter of interest and in turn, better control the system. In particular, Thompson sampling randomly selects controls according to the posterior probability they are optimal, and in response, data is generated in the form of state transitions and rewards, which are subsequently used to refine the estimate on the unknown parameter (i.e., update the posterior). If we regard the minimization of (i) posterior sampling error or (ii) average regret as a formal optimization criterion, the results of this paper show that as an input design, Thompson sampling is optimal with respect to both objectives (see Lemma 4 for (i), and Theorem 5 for (ii)).

II. Problem Setting

Consider the following average reward stochastic control problem over an infinite horizon. The system state and control applied at stage $t$ are denoted by $X_t \in \mathcal{X}$ and $U_t \in \mathcal{U}$, respectively, where the state space $\mathcal{X}$ and control space $\mathcal{U}$ are finite sets. We are interested in studying the problem in which both the reward and transition distributions are parameterized by an unknown (and fixed) parameter $\theta \in \mathcal{P}$, where the parameter space $\mathcal{P}$ is also finite.

In particular, at each stage $t$, after the state $X_t$ is observed and the decision maker applies control $U_t$, a reward $R_t \in \mathcal{R}$ is generated according to a positive continuous density

$$R_t \sim f_\theta(\cdot \mid X_t, U_t),$$

where $\mathcal{R}$ is a compact measurable subset of $\mathbb{R}$. We denote by

$$r_\theta(X_t, U_t) = \mathbb{E}_\theta[R_t \mid X_t, U_t] = \int r f_\theta(r \mid X_t, U_t) dr,$$

the expected reward at stage $t$ given $X_t$, $U_t$ and the unknown parameter is $\theta$. After reward $R_t$ is generated, the system state makes a transition from $X_t$ to $X_{t+1}$ according to distribution

$$\mathbb{P}_\theta(X_{t+1} = \cdot \mid X_t, U_t) = q_\theta(\cdot \mid X_t, U_t).$$

The objective is to construct policies that perform well regardless of the true underlying value of $\theta \in \mathcal{P}$. In this paper, we will study the admissible policy known as Thompson sampling, which will be shown to have good asymptotic performance as measured by worst-case average per period regret (see Theorem 5 for the main result of this paper).

Omniscient Decision Maker as a Benchmark

As a means of benchmarking the performance of a given admissible policy, we consider the omniscient decision maker who knows the true value of the unknown parameter $\theta \in \mathcal{P}$, and can make use of this value when constructing policies. The omniscient decision maker’s objective is to maximize over all admissible policies $\mu$, the long-run expected average reward per stage starting from a given initial state $X_0 = x_0$, given by

$$\lim_{T \to \infty} \frac{1}{T} \mathbb{E}_{\mu}^{\theta} \left[ \sum_{t=0}^{T-1} r_\theta(X_t, U_t) \right],$$

where $U_t$ is the control applied at stage $t$ under policy $\mu$. Here, we assume that the limit in (1) exists, and that an optimal stationary deterministic policy exists, which we denote by $\mu_\theta$. That is, the optimal policy $\mu_\theta$ is a mapping $\mu_\theta : \mathcal{X} \to \mathcal{U}$ that applies control $\mu_\theta(x)$ whenever $X_t = x$.

It follows that the state process $(X_t : t \geq 0)$ under the optimal policy $\mu_\theta$ constitutes a time-homogeneous Markov chain with transition probability matrix $Q_\theta$, where the $(x, y)$-entry of $Q_\theta$ is given by $Q_\theta(x,y) = q_\theta(y \mid x, \mu_\theta(x))$. We assume this Markov chain is ergodic in that there exists a $|\mathcal{X}| \times |\mathcal{X}|$ row-constant stochastic matrix $Q_\theta^\infty$ such that

$$\lim_{t \to \infty} \|Q_\theta^t - Q_\theta^\infty\| = 0,$$

where $\|A\|$ is the matrix norm defined by $\|A\| = \max_{x,y} \|a_{xy}\|$ for any real-valued matrix $A = (a_{xy})$.

We now state a key convergence rate result from the theory of ergodic time-homogeneous Markov chains, which will be useful in the proof of our main result.

**Proposition 1.** There exist constants $\alpha_0 > 0$ and $0 < \beta_0 < 1$ such that

$$\|Q_\theta^t - Q_\theta^\infty\| \leq \alpha_0 \beta_0^t.$$

A proof of Proposition 1 can be found in for example [18].

Average Regret and Asymptotic Efficiency

Since the decision maker does not in general know the true value of the parameter $\theta \in \mathcal{P}$, he can only select controls that are adapted to the history (data) available at each stage $t$ given by $H_t = X_0, U_0, \ldots, X_{t-1}, U_{t-1}, R_{t-1}, X_t$. (4)

The class of admissible policies $\mathcal{M}$ is the set of all sequences $\mu = (\mu_0, \mu_1, \ldots)$ such that $\mu_t \Rightarrow \mu_1(H_t)$ is a mapping from the history $H_t$ to a probability measure $\mu_1(H_t)$ over the control space $\mathcal{U}$, i.e., the control $U_t$ taken at stage $t$ is obtained by sampling from the probability measure $\mu_1(H_t)$.

We will measure the performance of a given admissible policy using the notion of average per period regret. Specifically, for any admissible policy $\mu \in \mathcal{M}$ and any value of the unknown parameter $\theta \in \mathcal{P}$, we define the average per period regret over $T$-periods as

$$AR(T, \mu, \theta) = \frac{1}{T} \mathbb{E}_\theta^{\mu_\theta} \left[ \sum_{t=0}^{T-1} r_\theta(Y_t, \mu_\theta(Y_t)) \right] - \mathbb{E}_\theta^{\mu_\theta} \left[ \sum_{t=0}^{T-1} r_\theta(X_t, U_t) \right],$$

where $Y_t$ represents the state realizations under the optimal control policy $\mu_\theta$ of the omniscient decision maker, whereas $X_t$ and $U_t$ represent the state and control realizations under the admissible policy $\mu \in \mathcal{M}$ that cannot make use of the true value of $\theta$.

The worst-case average per period regret is then defined as

$$AR(T, \mu) = \max_{\theta \in \mathcal{P}} AR(T, \mu, \theta),$$

and we want to find policies for which $AR(T, \mu)$ tends to zero as fast as possible (in $T$).
We say an admissible policy $\mu \in \mathcal{M}$ is asymptotically optimal if it achieves a worst-case average regret of $AR(T, \mu) = O(T^{-1})$. Our main result (Theorem 5) shows that Thompson sampling, which we denote by $\tau$, is asymptotically optimal.

III. THOMPSON SAMPLING AND MAIN RESULT

The decision maker accounts for uncertainty about $\theta \in \mathcal{P}$ by modeling it as a $\mathcal{P}$-valued random variable $\Theta$ with a prior distribution $\pi_0(\theta) = \mathbb{P}(\Theta = \theta | H_0)$, $\theta \in \mathcal{P}$.

Note that we make no assumptions on the structure of the prior distribution (e.g. no prior-likelihood conjugal relationship is assumed), nor do we impose any interpretation on the underlying parameter $\theta \in \mathcal{P}$ (e.g. we do not assume that $\theta$ represents an unknown mean/variation parameter). In this regard, our model is fairly general and applies to problems where the parameter space has a reasonable discrete representation (e.g. finite hypothesis testing, Bayesian mixture models). Alternatively, if one is interested in problems with a more structured parameter space (e.g. continuous or multi-dimensional spaces), the current model can still be applied in these situations if one views it as a suitable discretization of the problem of interest (see e.g. [8], p. 107). Nevertheless, establishing regret results for more general parameter spaces could be an interesting extension of this paper, which we discuss in Section VII.

At the beginning of stage $t$, given the history $H_t$ of past states, controls and rewards, the decision maker updates his belief about $\Theta$ by computing the posterior distribution $\pi_t(\theta) = \mathbb{P}(\Theta = \theta | H_t)$, $\theta \in \mathcal{P}$, given by Bayes’ Theorem:

$$\pi_t(\theta) = \frac{\mathcal{L}_t(H_t)\pi_0(\theta)}{\sum_{\gamma \in \mathcal{P}} \mathcal{L}_t(H_t)\pi_0(\gamma)}, \quad (5)$$

where

$$\mathcal{L}_t(H_t) = \prod_{s=1}^{t} f_\theta(R_{s-1} | X_{s-1}, U_{s-1}) q_\theta(X_s | X_{s-1}, U_{s-1}).$$

We note that from (5), the posterior distribution $\pi_t(\theta)$, $\theta \in \mathcal{P}$, is a random variable, as it is a function of the random history $H_t$ defined in (4).

Notice that for each $\theta \in \mathcal{P}$, $x \in \mathcal{X}$ and $u \in \mathcal{U}$, the joint density $f_\theta(\cdot | x, u)q_\theta(\cdot | x, u)$ specifies a joint probability measure on $\mathcal{R} \times \mathcal{X}$, i.e.,

$$v_\theta^{x,u}(A, B) := \int_A f_\theta(r | x, u)dr \sum_{y \in B} q_\theta(y | x, u),$$

for $A \subseteq \mathcal{R}$, $B \subseteq \mathcal{X}$. Then, for any $\gamma, \theta \in \mathcal{P}$, we can express the Radon-Nikodym derivative

$$\frac{dv_\gamma^{x,u}}{dv_\theta^{x,u}} = \frac{f_\theta(\cdot | x, u)q_\theta(\cdot | x, u)}{f_\gamma(\cdot | x, u)q_\gamma(\cdot | x, u)},$$

and define the relative entropy of $v_\theta^{x,u}$ with respect to $v_\gamma^{x,u}$ as

$$K(v_\theta^{x,u} | v_\gamma^{x,u}) = \mathbb{E}_\theta \left[ \log \left( \frac{dv_\theta^{x,u}}{dv_\gamma^{x,u}} \right) \right], \quad (6)$$

provided $v_\gamma^{x,u}$ is absolutely continuous with respect to $v_\theta^{x,u}$, and as $\infty$ otherwise (see e.g. [19], p. 26, for a general definition of relative entropy and its basic properties).

It can be shown that relative entropy is non-negative, i.e.,

$$K(v_\theta^{x,u} | v_\gamma^{x,u}) \geq 0,$$

and furthermore,

$$K(v_\theta^{x,u} | v_\gamma^{x,u}) = 0 \iff v_\theta^{x,u} = v_\gamma^{x,u},$$

so that relative entropy can be thought of as a “distance” from probability measure $v_\gamma^{x,u}$ to probability measure $v_\theta^{x,u}$.

Assumption 2: For any $x \in \mathcal{X}$, $u \in \mathcal{U}$, and any two distinct parameter values $\theta \neq \gamma \in \mathcal{P}$, there exists a positive constant $\epsilon(x, u, \theta, \gamma) > 0$ such that

$$K(v_\theta^{x,u} | v_\gamma^{x,u}) \geq \epsilon(x, u, \theta, \gamma). \quad (7)$$

In words, Assumption 2 states that provided $\theta \neq \gamma$, the two probability measures $v_\theta^{x,u}$ and $v_\gamma^{x,u}$ are distinguishable as measured by relative entropy.

Remark 3: We note that Assumption 2 is not particularly strong, as it is satisfied for many practical stochastic control problems of interest. To see this, let us characterize those problems for which Assumption 2 is not satisfied. In particular, if Assumption 2 fails to hold, then there exists an $x \in \mathcal{X}$, $u \in \mathcal{U}$ such that whenever $x_t = x$ and $u_t = u$, the posterior distribution $\pi_t$ will not update, regardless of the data observed at stage $t+1$, i.e., $\pi_{t+1} = \pi_t$ almost surely. In this regard, Assumption 2 excludes only those problems for which data is essentially useless for learning about the underlying parameter.

Examples of stochastic control problems that satisfy Assumption 2 include: Bayesian optimal stopping problems [20], sensor scheduling problems [21], inventory control problems with learning [22], and queuing control problems with arrival rate uncertainty [23] (a numerical example of this problem type will be given in Section VI).

We now describe the Thompson sampling policy $\tau$.

Thompson Sampling:

Step 1. Update the current posterior distribution $\pi_t$ according to (5).

Step 2. Sample $\theta_t \in \mathcal{P}$ from the current posterior $\pi_t$, i.e., choose $\theta_t = \theta$ with probability $\pi_t(\theta)$, $\theta \in \mathcal{P}$.

Step 3. Apply control $U_t = \mu_{\theta_t}(X_t)$.

Step 4. Observe new reward $R_t$ and state transition from $X_t$ to $X_{t+1}$, update history $H_{t+1}$, increment $t$ and go to Step 1.

Notice that Thompson sampling $\tau$ belongs to the class of admissible policies $\mathcal{M}$.

Lemma 4: Suppose Assumption 2 holds. Then, under Thompson sampling $\tau$, there exists constants $a_\theta, b_\theta > 0$ such that

$$\mathbb{E}_\theta \left[ 1 - \pi_t(\theta) \right] \leq a_\theta e^{-b_\theta t}. \quad (8)$$

A proof of Lemma 4 is given in Section IV. In Lemma 4 the operator $\mathbb{E}_\theta[\cdot]$ represents the conditional expectation given the true underlying parameter is $\theta \in \mathcal{P}$ and Thompson sampling $\tau$ is used to generate controls. Recall that at stage $t$, the parameter $\theta_t$ generated by Thompson sampling is sampled from the posterior distribution $\pi_t$, and hence the probability that the true parameter value $\theta$ is not selected at stage $t$ is
given by $1 - \pi_t(\theta)$. In this regard, the quantity $\mathbb{E}_\theta^\gamma [1 - \pi_t(\theta)]$ can be interpreted as the expected posterior belief error, and Lemma 4 states that under Thompson sampling, this error tends to zero exponentially fast. We now state the main result of this paper.

**Theorem 5 (Main Result):** Suppose Assumption 2 holds. Then, Thompson sampling $\tau$ is asymptotically optimal, i.e., it achieves a worst-case average per period regret of
\[
AR(T, \tau) \leq \max_{\theta \in \mathcal{P}} \left\{ \frac{|X| \alpha_\theta \sigma_\theta}{(1 - e^{-b_\theta})(1 - \beta_\theta)} + \frac{2\alpha_\theta}{1 - b_\theta} \right\} B T^{-1},
\]
where constants $\alpha_\theta, \beta_\theta$ are from Proposition 1 and constants $a_\theta, b_\theta$ are from Lemma 4.

**IV. PROOF OF LEMMA 4**

Fix an arbitrary $\theta \in \mathcal{P}$. We start by expressing the posterior probability $\pi_t(\theta)$ from Bayes’ Theorem 5 as
\[
\pi_t(\theta) = \frac{L_0(H_t) \pi_0(\theta)}{\sum_{\gamma \in \mathcal{P}} L_\gamma(H_t) \pi_0(\gamma)} = \frac{1}{1 + \sum_{\gamma \neq \theta} \frac{L_\gamma(H_t)}{L_\theta(H_t)}} = \frac{1}{1 + \sum c_\gamma \exp \left( - \sum_{s=0}^t \log \Lambda_s^\gamma \right)},
\]
where constant $c_\gamma = \pi_0(\gamma) \pi_0^{-1}(\theta)$ and
\[
\Lambda_s^\gamma = f_\theta(R_s-1 | X_{s-1}, U_{s-1}) q_\theta(X_s | X_{s-1}, U_{s-1}),
\]
for $0 < s \leq t$. For any $\gamma \neq \theta$, we define the following stochastic process
\[
Z_t^\gamma = \sum_{s=0}^t \log \Lambda_s^\gamma.
\]
It follows that if we define filtration $(\mathcal{H}_t : t \geq 0)$ by $\mathcal{H}_t = \sigma(H_t)$, where history $H_t$ is given in 4, the process $(Z_t^\gamma : t \geq 0)$ is a submartingale with respect to $\mathcal{H}_t$ under probability measure $\mathbb{P}_\theta$.

To see this, we write $Z_t^\gamma$ in terms of its Doob decomposition (see e.g. [23], p. 206):
\[
Z_t^\gamma = \sum_{s=0}^t \log \Lambda_s^\gamma = \sum_{s=0}^t (\log \Lambda_s^\gamma - \mathbb{E}_\theta^\gamma [\log \Lambda_s^\gamma | \mathcal{H}_{s-1}]) + \sum_{s=0}^t \mathbb{E}_\theta^\gamma [\Lambda_s^\gamma | \mathcal{H}_{s-1}].
\]

The first term on the right-hand side of (10)
\[
M_t^\gamma := \sum_{s=0}^t (\log \Lambda_s^\gamma - \mathbb{E}_\theta^\gamma [\log \Lambda_s^\gamma | \mathcal{H}_{s-1}]),
\]
is an $\mathcal{H}_t$-martingale under $\mathbb{P}_\theta$, by construction. Furthermore, its increments
\[
[\log \Lambda_s^\gamma - \mathbb{E}_\theta^\gamma [\log \Lambda_s^\gamma | \mathcal{H}_{s-1}]] \leq d,
\]
for some $d > 0$. To see this, recall that the conditional reward density $f_\theta(r | x, u)$ is positive and continuous with compact support. Therefore, $f_\theta(r | x, u)$ attains a maximum and minimum value that are both positive, which implies $[\log f_\theta(r | x, u)] < +\infty$, so that the above inequality holds almost surely.

The second term on the right-hand side of (10)
\[
A_t^\gamma := \sum_{s=0}^t \mathbb{E}_\theta^\gamma [\log \Lambda_s^\gamma | \mathcal{H}_{s-1}],
\]
is a predictable process (see e.g. [25], p. 191), i.e., $A_t^\gamma$ is $\mathcal{H}_{t-1}$-measurable, and satisfies $A_0^\gamma = 0$ since $A_0^\gamma = 1$. Furthermore, for any $s \leq t$, the summand
\[
\mathbb{E}_\theta^\gamma [\log \Lambda_s^\gamma | \mathcal{H}_{s-1}]
\]
is an increasing predictable process satisfying $A_s^\gamma = 0$ and
\[
A_t^\gamma = \sum_{s=0}^t \mathbb{E}_\theta^\gamma [\log \Lambda_s^\gamma | \mathcal{H}_{s-1}] \geq \epsilon t.
\]
Hence, $Z_t^\gamma = M_t^\gamma + A_t^\gamma$ is an $\mathcal{H}_t$-submartingale under $\mathbb{P}_\theta$.

Therefore, from (9) we have
\[
\mathbb{E}_\theta [\pi_t(\theta)] = \mathbb{E}_\theta \left[ \frac{1}{1 + \sum_{\gamma \neq \theta} c_\gamma \exp \left( - M_t^\gamma - A_t^\gamma \right)} \right] \geq \mathbb{E}_\theta \left[ \frac{1}{1 + \sum_{\gamma \neq \theta} c_\gamma \exp \left( - M_t^\gamma - \epsilon t \right)} \right].
\]
For any $\delta > 0$ and $\gamma \in \mathcal{P}$, if we define the event
\[
B_t^\gamma(\delta) = \{ |M_t^\gamma| \leq \delta t \},
\]
then, for any choice of $0 < \delta < \epsilon$,
\[
\mathbb{E}_\theta [\pi_t(\theta)] \geq \mathbb{E}_\theta \left[ \frac{1}{1 + \sum_{\gamma \neq \theta} c_\gamma \exp \left( - M_t^\gamma - \epsilon t \right)} \right] \geq \mathbb{E}_\theta \left[ \frac{1}{1 + \sum_{\gamma \neq \theta} c_\gamma \exp \left( - M_t^\gamma - \epsilon t \right)} \right] \geq \mathbb{P}_\theta \left( \cap_{\gamma \neq \theta} B_t^\gamma(\delta) \right) \geq 1 - \frac{\mathbb{P}_\theta \left( \cup_{\gamma \neq \theta} B_t^\gamma(\delta) \right)}{1 + \frac{1 - \mathbb{P}_\theta(\pi_t(\theta))}{\pi_t(\theta)} \exp \left( - (\epsilon - \delta)t \right)}.
\]
The second inequality follows since on the event $\cap_{\gamma \neq \theta} B_{t}^{\gamma}(\delta)$, we have $|M_{t}^{\gamma}| \leq \delta t$ which implies $\exp(-M_{t}^{\gamma} - ct) \leq \exp(\delta t - ct) = \exp(-c - \delta)t$. On the other hand, on the event $(\cap_{\gamma \neq \theta} B_{t}^{\gamma}(\delta))^{c}$, the term inside the expectation is always positive and therefore bounded below by zero.

The union bound gives

$$
E_{\theta}^{\gamma}[\pi_{t}(\theta)] \geq \frac{1 - \sum_{\gamma \neq \theta} \mathbb{P}_{\theta}^{\gamma}(B_{t}^{\gamma}(\delta))}{1 + \frac{1}{\pi_{0}(\theta)} \exp((-c - \delta)t)} = \frac{1 - \sum_{\gamma \neq \theta} \mathbb{P}_{\theta}^{\gamma}(|M_{t}^{\gamma}| \geq \delta t)}{1 + \frac{1}{\pi_{0}(\theta)} \exp((-c - \delta)t)}.
$$

Since $M_{t}^{\gamma}$ is an $\mathcal{H}_{t}$-martingale with bounded increments under $\mathbb{P}_{\theta}^{\gamma}$, by Azuma’s inequality (see e.g. [23], p. 198), we have

$$
\mathbb{P}_{\theta}^{\gamma}(|M_{t}^{\gamma}| \geq \delta t) \leq 2\exp\left(-\frac{\delta^{2}t}{2d^{2}}\right),
$$

which implies

$$
E_{\theta}^{\gamma}[\pi_{t}(\theta)] \geq \frac{1 - 2(|\mathcal{P}| - 1) \exp\left(-\frac{\delta^{2}t}{2d^{2}}\right)}{1 + \frac{1}{\pi_{0}(\theta)} \exp((-c - \delta)t)}.
$$

Therefore, if we choose $\delta = \frac{c}{2}$, and constants $a_{0} = 2 \max\left\{\frac{1 - \pi_{0}(\theta)}{\pi_{0}(\theta)}, 2(|\mathcal{P}| - 1)\right\}$ and $b_{0} = \min\left\{\frac{c}{2}, \frac{\delta^{2}}{8d^{2}}\right\}$, then

$$
E_{\theta}^{\gamma}[1 - \pi_{t}(\theta)] \leq 1 - \frac{1 - 2(|\mathcal{P}| - 1) \exp\left(-\frac{c^{2}t}{2d^{2}}\right)}{1 + \frac{1}{\pi_{0}(\theta)} \exp((-c - \delta)t)} = \frac{1 - \pi_{0}(\theta)}{\pi_{0}(\theta)} \exp((-c - \delta)t) + 2(|\mathcal{P}| - 1) \exp\left(-\frac{\delta^{2}t}{2d^{2}}\right)
\leq a_{0} \exp(-b_{0}t),
$$

which completes the proof.

**Remark 6:** We note that the choices of constants $a_{0}$ and $b_{0}$ given in the proof above do not depend on the current stage $t$. This is because parameters $\pi_{0}(\theta), |\mathcal{P}|, c$, and $d$ are all $t$-independent.

**Remark 7:** Note that Assumption 2 allowed us to bound the predictable process $A_{t}^{\gamma}$ defined in (11) by a deterministic linear function of $t$ in equation (12), and, by applying Azuma’s inequality, we achieved an estimate on the posterior sampling error $E_{\theta}^{\gamma}[1 - \pi_{t}(\theta)]$. In contrast, if one attempts to relax Assumption 2 this is no longer possible in general. Specifically, if we let $I_{t}$ represent the number of stages up to and including time $t$ at which $K(\gamma_{t}^{\gamma_{t-1}^{0}} | x_{t-1}^{\gamma_{t-1}^{0}}) > 0, \theta \neq \gamma_{t}$, then the predictable process $A_{t}^{\gamma}$ can only be bounded by a linear function of $I_{t}$, which is no longer deterministic. While, this will allow one to develop an estimate on the conditional posterior sampling error $E_{\theta}^{\gamma}[1 - \pi_{t}(\theta) | I_{t}]$ which is a function of random variable $I_{t}$, in order to estimate the unconditional posterior sampling error $E_{\theta}^{\gamma}[1 - \pi_{t}(\theta)]$, it is necessary to understand how $I_{t}$ increases over time. Clearly, characterizing the random behavior of $I_{t}$ under Thompson sampling cannot be done in general; it would highly depend on the particulars of a given model and would vary significantly from problem to problem. In this regard, while relaxing Assumption 2 may be an interesting extension to this paper, it would likely be at the cost of some loss of generality of the model.

**V. PROOF OF MAIN RESULT: THEOREM 5**

Fix an arbitrary parameter value $\theta \in \mathcal{P}$. Recall that $Y_{t}$ represents the state realizations under the optimal control policy $\mu_{0}$ of the omniscient decision maker, whereas $X_{t}$ and $U_{t}$ represent the state and control realizations under Thompson sampling $\tau \in \mathcal{M}$. To make notation more compact, when there is no confusion, for the remainder of the proof we will denote $r_{t}^{\mu_{0}}(x) = r_{t}(x, \mu_{0}(x)), x \in \mathcal{X}$.

The first difficulty one encounters when trying to estimate average regret is that the two state processes $X_{t}$ and $Y_{t}$ are not connected in any obvious way. To bridge the gap between these two processes, we decompose average per period regret as

$$
AR(T, \tau, \theta) = \frac{1}{T} \left| \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(Y_{t}) \right] - \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(X_{t}, U_{t}) \right] \right|
\leq \frac{1}{T} \left| \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(Y_{t}) \right] - \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(X_{t}) \right] \right|
+ \frac{1}{T} \left| \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(X_{t}) \right] - \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(X_{t}, U_{t}) \right] \right|,
$$

by applying the omniscient mapping $\mu_{0}$ to the state process $X_{t}$ generated by Thompson sampling.

We study the two terms on the right-hand side of (14) separately. We start with the first term given by

$$
\frac{1}{T} \left| \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(Y_{t}) \right] - \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(X_{t}) \right] \right|.
$$

Recall, the state process $Y_{t}$ under policy $\mu_{0}$ constitutes a time-homogeneous Markov chain with transition probability matrix $Q_{0}$. Then, by the Chapman-Kolmogorov equations (see e.g. [23], p. 9), we have

$$
\frac{1}{T} \left| \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(Y_{t}) \right] - \mathbb{E}_{\theta}^{\mu_{0}} \left[ \sum_{t=0}^{T-1} r_{t}^{\mu_{0}}(X_{t}) \right] \right|
= \frac{1}{T} \left| \sum_{t=0}^{T-1} \sum_{x_{t} \in \mathcal{X}} r_{t}^{\mu_{0}}(x_{t})(\mathbb{P}_{\theta}^{\mu_{0}}(Y_{t} = x_{t} | \theta_{t} = \theta_{t}) - \mathbb{P}_{\theta}^{\mu_{0}}(X_{t} = x_{t})) \right|
= \frac{1}{T} \left| \sum_{t=0}^{T-1} \sum_{x_{t} \in \mathcal{X}} r_{t}^{\mu_{0}}(x_{t})\bar{Q}_{0}(x_{t}, x_{t}) - \mathbb{P}_{\theta}^{\mu_{0}}(X_{t} = x_{t})) \right|.
$$

We now study the probability $\mathbb{P}_{\theta}^{\mu_{0}}(X_{t} = x_{t})$ appearing on the right-hand side of the equation above. Recall that $\theta_{t}$ is the randomly sampled parameter value generated by Thompson sampling at stage $t$. Then, by conditioning on the history $H_{t-1}$ at stage $t - 1$ we have

$$
\mathbb{P}_{\theta}^{\mu_{0}}(X_{t} = x_{t}) = \mathbb{E}_{\theta}^{\mu_{0}}[\mathbb{P}_{\theta}^{\mu_{0}}(X_{t} = x_{t} | H_{t-1})],
$$

where $\mathbb{P}_{\theta}^{\mu_{0}}(X_{t} = x_{t} | H_{t-1})$ is the posterior probability of $X_{t}$ given $H_{t-1}$.
where
\[ \mathbb{P}_\theta^T(X_t = x_t | H_{t-1}) \]
\[ = \sum_{\gamma \in \mathcal{P}} \pi_{t-1}(\gamma) \mathbb{P}_\theta^T(X_t = x_t | \theta_{t-1} = \gamma, H_{t-1}) \]
\[ = Q_\theta(X_{t-1}, x_t)
+ \sum_{\gamma \neq \theta} \pi_{t-1}(\gamma)(Q_\gamma(X_{t-1}, x_t) - Q_\theta(X_{t-1}, x_t)). \]

It follows that
\[ \mathbb{P}_\theta^\tau(X_t = x_t)
\]
\[ = \mathbb{E}_\theta[\mathbb{P}_\theta^T(X_t = x_t | H_{t-1})]
= \mathbb{E}_\theta[Q_\theta(X_{t-1}, x_t)]
+ \mathbb{E}_\theta\left[ \sum_{\gamma \neq \theta} \pi_{t-1}(\gamma)(Q_\gamma(X_{t-1}, x_t) - Q_\theta(X_{t-1}, x_t)) \right]
\]
\[ = \sum_{x_{t-1} \in \mathcal{X}} Q_\theta(x_{t-1}, x_t) \mathbb{P}_\theta^T(X_t = x_{t-1}) + \Delta_t(x_t), \tag{15}\]

where we have put
\[ \Delta_t(x_t) = \mathbb{E}_\theta^\tau\left[ \sum_{\gamma \neq \theta} \pi_{t-1}(\gamma)(Q_\gamma(X_{t-1}, x_t) - Q_\theta(X_{t-1}, x_t)) \right]. \tag{16}\]

Note that
\[ \sum_{x_{t} \in \mathcal{X}} \Delta_t(x_t) = 0. \tag{17}\]

We can then write equation (15) in matrix notation as follows
\[ P_t = P_{t-1}Q_\theta + \Delta_t, \tag{17}\]
where \(|\mathcal{X}|\)-dimensional row vectors \(P_t\) and \(\Delta_t\) have \(x\)-components equal to \(\mathbb{P}_\theta^T(X_t = x)\) and \(\Delta_t(x)\), respectively. Recursively substituting equation (17) into itself, we get
\[ P_t = P_0Q_\theta + \Delta_tQ_\theta^{t-1} + \cdots + \Delta_1Q_\theta + \Delta_t, \]
where \(P_0\) is the \(|\mathcal{X}|\)-dimensional row vector with 1 in component \(x_0\) and 0 elsewhere. We also denote by \(R_\theta^\mu\) the \(|\mathcal{X}|\)-dimensional column vector with \(x\)-component equal to \(r_\theta^\mu(x)\).

It follows that
\[ \frac{1}{T} \left| \mathbb{E}_\theta^\mu \left[ \sum_{t=0}^{T-1} r_\theta^\mu(Y_t) \right] - \mathbb{E}_\theta^\mu \left[ \sum_{t=0}^{T-1} r_\theta^\mu(X_t) \right] \right| \]
\[ = \frac{1}{T} \left| \sum_{t=0}^{T-1} \sum_{x_{t-1} \in \mathcal{X}} \Delta_t(x_t)(Q_\theta^\infty(x_{t-1}, x_t) - \mathbb{P}_\theta^T(X_t = x_t)) \right|
\]
\[ = \frac{1}{T} \left| \sum_{t=1}^{T-1} \left( \sum_{s=1}^{t} \Delta_sQ_\theta^{t-s} \right) R_\theta^\mu \right|
\]
\[ = \frac{1}{T} \left| \sum_{s=1}^{\infty} \left( \sum_{t=s}^{\infty} \Delta_t(Q_\theta^{t-s} - Q_\theta^\infty) \right) R_\theta^\mu \right|, \tag{18}\]

where \(Q_\theta^\infty\) is the row-right ergodic matrix defined in (21) and the last equality follows from property (16) and the fact that \(Q_\theta^\infty\) is row-right, i.e., every row is equal.

For any real-valued matrix \(A = (a_{ij})\), we define \([A] = (|a_{ij}|)\), where each component is replaced by its absolute value. Then, it is not hard to check that \([A + B] \leq [A] + [B]\) and \([AB] \leq [A][B]\), component-wise. Continuing with (18), it follows that
\[ \frac{1}{T} \left| \mathbb{E}_\theta^\mu \left[ \sum_{t=0}^{T-1} r_\theta^\mu(Y_t) \right] - \mathbb{E}_\theta^\mu \left[ \sum_{t=0}^{T-1} r_\theta^\mu(X_t) \right] \right| \]
\[ \leq \frac{1}{T} \left( \sum_{t=1}^{\infty} \sum_{s=1}^{\infty} [\Delta_t][Q_\theta^{t-s} - Q_\theta^\infty] \right) [R_\theta^\mu], \tag{19}\]

The following equality is easy to check.
\[ \sum_{t=1}^{\infty} \sum_{s=1}^{\infty} [\Delta_t][Q_\theta^{t-s} - Q_\theta^\infty] = \sum_{t=1}^{\infty} [\Delta_t] \sum_{s=0}^{\infty} [Q_\theta^s - Q_\theta^\infty] \]
\[ \leq \mathbb{E}_\theta \left[ 1 - \pi_{t-1}(\theta) \right], \]

for each \(x \in \mathcal{X}\).

We therefore need an estimate on the expected posterior sampling error under Thompson sampling. For this, we use the estimate on the expected posterior sampling error established in Lemma 4. In particular, from (8) of Lemma 4, it follows that for any \(x \in \mathcal{X}\),
\[ \sum_{t=1}^{\infty} [\Delta_t](x) \leq \sum_{t=1}^{\infty} \mathbb{E}_\theta \left[ 1 - \pi_{t-1}(\theta) \right] \leq \frac{\alpha \theta}{1 - e^{-\beta \theta}}. \tag{21}\]

We next apply the convergence rate result for ergodic time-homogeneous Markov chains given in Proposition 2 to show that the second infinite series on the right-hand side of (20) exists. In particular, the estimate given in (3) from Proposition 2 implies that for any \(x, y \in \mathcal{X}\),
\[ \sum_{s=0}^{\infty} [Q_\theta^s - Q_\theta^\infty](x, y) \leq \sum_{s=0}^{\infty} \|Q_\theta^s - Q_\theta^\infty\| \]
\[ \leq \sum_{s=0}^{\infty} \alpha \theta \beta \theta^t \]
\[ = \frac{\alpha \theta}{1 - \beta \theta}. \tag{22}\]

Therefore, equations (19), (20), (21) and (22) imply
\[ \frac{1}{T} \left| \mathbb{E}_\theta^\mu \left[ \sum_{t=0}^{T-1} r_\theta^\mu(Y_t) \right] - \mathbb{E}_\theta^\mu \left[ \sum_{t=0}^{T-1} r_\theta^\mu(X_t) \right] \right| \]
\[ \leq \frac{1}{T} \left( \sum_{t=1}^{\infty} \sum_{s=1}^{\infty} [\Delta_t][Q_\theta^{t-s} - Q_\theta^\infty] \right) [R_\theta^\mu] \]
\[ \leq \frac{1}{T} \sum_{t=1}^{\infty} [\Delta_t] \left( \sum_{s=0}^{\infty} [Q_\theta^s - Q_\theta^\infty] \right) [R_\theta^\mu] \]
\[ \leq \frac{B |\mathcal{X}| \alpha \theta \theta}{T (1 - e^{-\beta \theta})(1 - \beta \theta)}, \]

where \(B = \max_{x, u, \theta} |r_\theta(x, u)|\).
Thompson sampling achieves an optimal worst-case average regret of $O(T^{-1})$. The following experiment was performed. We ran the Thompson sampling algorithm on 5,000,000 simulated sample paths using model parameters: $N = 40$, $r = 10$, $c = 0.25$ and $\eta = 0.25$. In Figure 13 we report the worst-case average regret under Thompson sampling $AR(T, \tau)$ for $T = 1, \ldots, 500$, where the $T$th point on the graph is the arithmetic averages among all sample paths generated.

The total running time to produce this graph was 6746.11 seconds and was performed on a Dell Inspiron 7559, Intel core i7-6700 HQ processor with 4 cores with 8GB ram, and a base speed of 2.6 Ghz. It follows that implementing Thompson sampling in this example is extremely fast, as each sample path took on average 0.1499 milliseconds. In contrast, if one attempts to solve the problem using for example Bayesian dynamic programming the problem would take significantly longer to solve; perhaps even be intractable (see e.g [13]). We next illustrate the optimality of Thompson sampling in this example by validating our main result Theorem 5. In particular, in Figure 1b we graph the quantity $1/AR(T, \tau)$ and notice that this quantity grows linearly for $T$ large. This shows that indeed $AR(T, \tau) = O(T^{-1})$, for if average regret under Thompson sampling were suboptimal, the quantity $1/AR(T, \tau)$ would grow super-linearly and eventually depart from linear growth for $T$ large. Finally, we note that the quantities $AR(T, \tau)$ and $1/AR(T, \tau)$ in this example exhibit somewhat irregular behavior for $T$ (relatively) small, but eventually stabilize for large $T$, as the theory predicts.

VII. CONCLUDING REMARKS

We applied Thompson sampling to a class of average reward stochastic control problems with parameter uncertainty. Specifically, we studied an average reward stochastic control problem over an infinite horizon in which both the reward and state transition distributions are parameterized by an unknown fixed parameter taking values in a finite space. We proved that Thompson sampling achieves a worst-case average per period regret of $O(T^{-1})$, which is asymptotically optimal.

There are a number of open and challenging avenues for future research. In our problem, the state, control and parameter spaces were all assumed to be finite, so it was possible to obtain exponentially fast convergence of the posterior error (Lemma 4). When such spaces are general measurable spaces on the other hand, establishing exponential-rate convergence, and therefore optimal regret bounds, is no longer guaranteed. In fact, in the modern Bayesian statistics literature, to our knowledge, no readily available $L^1$ posterior convergence rate result is available over general measurable parameter spaces where data is not i.i.d. but comes from a stochastic control problem. Therefore, to obtain theoretical regret bounds along the lines of Theorem 5 would require developing new posterior convergence rate results and also a non-trivial generalization of the proof given in Section V.

ACKNOWLEDGMENTS

The author thanks the three anonymous referees, the associate editor, and the senior editor for their helpful comments and feedback. Special thanks goes to Andrew E.B. Lim.
for suggesting the idea of “dynamic Thompson sampling”,
Dragan Banjevic for carefully reading (multiple) early drafts
of the paper, and Michael Gimelfarb for helping with Section
VI. This research was partially supported by the NSERC
Discovery Grant RGPIN-2015-04019. The opinions, findings,
conclusions or recommendations expressed in this material are
those of the author and do not necessarily reflect the views of
the Natural Sciences and Engineering Research Council.

REFERENCES
multi-armed bandit problem,” Proceedings of the 25th Annual Confer-
sion sampling,” Journal of Machine Learning Research, forthcoming,
2014.
exceeds another in view of the evidence of two samples,” Biometrika,
learning and acting,” Journal of Artificial Intelligence Research, vol. 38,
sequential decision-making and causal inference,” Complex Adaptive
learning via posterior sampling,” Neural Information Processing Sys-
tems, 2013.
with unknown parameters,” IEEE Transactions on Automatic Control,
USA: Springer-Verlag, 1989.
[9] P.R. Kumar, “A survey of some results in stochastic adaptive control,”
partially observable Markov decision processes,” IEEE Transactions on
tion,” IEEE Transactions on Automatic Control, vol. 58, pp. 500–506,
2013.
processes,” Mathematics of Operations Research, vol. 12, pp. 441–450,
1987.
problems,” Journal of Economic Dynamics and Control, vol. 27, pp. 87–
[17] J.C. Spall, Introduction to Stochastic Search and Optimization: Estima-
tion, Simulation, and Control. Hoboken, USA: John Wiley & Sons,
2003.
[18] C.C. Huang, D. Isaacson, and B. Vinograde, “The rate of convergence of
certain nonhomogeneous Markov chains,” Zeitschrift Für Wahrschein-
mation and multiple means of learning,” Operations Research, vol. 63,
under censoring: stock-out timing is (almost) all you need.” Operations
with unknown delay cost characteristics,” Manufacturing & Service