Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Learning to Optimize via Posterior Sampling

Daniel Russo, Benjamin Van Roy

To cite this article:
Daniel Russo, Benjamin Van Roy (2014) Learning to Optimize via Posterior Sampling. Mathematics of Operations Research 39(4):1221-1243. https://doi.org/10.1287/moor.2014.0650

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 39, No. 4, November 2014, pp. 1221–1243
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2014.0650

Learning to Optimize via Posterior Sampling

Daniel Russo
Department of Management Science and Engineering, Stanford University, Stanford, California 94305,
djrusso@stanford.edu

Benjamin Van Roy
Departments of Management Science and Engineering and Electrical Engineering, Stanford University,
Stanford, California 94305, bvr@stanford.edu

This paper considers the use of a simple posterior sampling algorithm to balance between exploration and exploitation when learning to optimize actions such as in multiarmed bandit problems. The algorithm, also known as *Thompson Sampling* and as *probability matching*, offers significant advantages over the popular upper confidence bound (UCB) approach, and can be applied to problems with finite or infinite action spaces and complicated relationships among action rewards. We make two theoretical contributions. The first establishes a connection between posterior sampling and UCB algorithms. This result lets us convert regret bounds developed for UCB algorithms into Bayesian regret bounds for posterior sampling. Our second theoretical contribution is a Bayesian regret bound for posterior sampling that applies broadly and can be specialized to many model classes. This bound depends on a new notion we refer to as the *eluder dimension*, which measures the degree of dependence among action rewards. Compared to UCB algorithm Bayesian regret bounds for specific model classes, our general bound matches the best available for linear models and is stronger than the best available for generalized linear models. Further, our analysis provides insight into performance advantages of posterior sampling, which are highlighted through simulation results that demonstrate performance surpassing recently proposed UCB algorithms.

Keywords: online optimization; multiarmed bandits; Thompson sampling
MSC2000 subject classification: Primary: 93E35; secondary: 62L05
OR/MS subject classification: Primary: decision analysis: sequential
History: Received February 26, 2013; revised November 21, 2013. Published online in *Articles in Advance* April 23, 2014.

1. Introduction. We consider an optimization problem faced by an agent who is uncertain about how his actions influence performance. The agent selects actions sequentially, and upon each action observes a reward. A *reward function* governs the mean reward of each action. The agent represents his initial beliefs through a prior distribution over reward functions. As rewards are observed, the agent learns about the reward function, and this allows him to improve his behavior. Good performance requires adaptively sampling actions in a way that strikes an effective balance between exploring poorly understood actions and exploiting previously acquired knowledge to attain high rewards. In this paper, we study a simple algorithm for selecting actions and provide finite-time performance guarantees that apply across a broad class of models.

The problem we study has attracted a great deal of recent interest and is often referred to as the multiarmed bandit (MAB) problem with dependent arms. We refer to the problem as one of *learning to optimize* to emphasize its divergence from the classical MAB literature. In the typical MAB framework, there are a finite number of actions that are modeled independently; sampling one action provides no information about the rewards that can be gained through selecting other actions. In contrast, we allow for infinite action spaces and for general forms of model uncertainty, captured by a prior distribution over a set of possible reward functions. Recent papers have addressed this problem in cases where the relationship among action rewards takes a known parametric form. For example, Abbasi-Yadkori et al. [2], Dani et al. [16], and Rusmevichientong and Tsitsiklis [31] study the case where actions are described by a finite number of features and the reward function is linear in these features. Other authors have studied cases where the reward function is Lipschitz continuous (Bubeck et al. [13], Kleinberg et al. [23], Valko et al. [36]), sampled from a Gaussian process (Srinivas et al. [35]), or takes the form of a generalized (Filippi et al. [18]) or sparse (Abbasi-Yadkori et al. [3]) linear model.

Each paper cited above studies an upper confidence bound (UCB) algorithm. Such an algorithm forms an optimistic estimate of the mean reward value for each action, taking it to be the highest statistically plausible value. It then selects an action that maximizes among these optimistic estimates. Optimism encourages selection of poorly understood actions, which leads to informative observations. As data accumulates, optimistic estimates are adapted, and this process of exploration and learning converges toward optimal behavior.

We study an alternative algorithm that we refer to as *posterior sampling*. It is also known as *Thompson sampling* and as *probability matching*. The algorithm randomly selects an action according to the probability it is optimal. Although posterior sampling was first proposed almost 80 years ago, it has until recently received little attention in the literature on MABs. While its asymptotic convergence has been established in some generality (May et al. [30]),

not much else is known about its theoretical properties in the case of dependent arms, or even in the case of independent arms with general prior distributions. Our work provides some of the first theoretical guarantees.

Our interest in posterior sampling is motivated by several potential advantages over UCB algorithms, which we highlight in §4.3. While particular UCB algorithms can be extremely effective, performance and computational tractability depends critically on the confidence sets used by the algorithm. For any given model, there is a great deal of design flexibility in choosing the structure of these sets. Because posterior sampling avoids the need for confidence bounds, its use greatly simplifies the design process and admits practical implementations in cases where UCB algorithms are computationally onerous. In addition, we show through simulations that posterior sampling outperforms various UCB algorithms that have been proposed in the literature.

In this paper, we make two theoretical contributions. The first establishes a connection between posterior sampling and UCB algorithms. In particular, we show that while the regret of a UCB algorithm can be bounded in terms of the confidence bounds used by the algorithm, the Bayesian regret of posterior sampling can be bounded in an analogous way by *any* sequence of confidence bounds. In this sense, posterior sampling preserves many of the appealing theoretical properties of UCB algorithms without requiring explicit, designed optimism. We show that, due to this connection, existing analysis available for specific UCB algorithms immediately translates to Bayesian regret bounds for posterior sampling.

Our second theoretical contribution is a Bayesian regret bound for posterior sampling that applies broadly and can be specialized to many specific model classes. Our bound depends on a new notion of dimension that measures the degree of dependence among actions. We compare our notion of dimension to the Vapnik-Chervonenkis (VC) dimension and explain why that and other measures of dimension used in the supervised learning literature do not suffice when it comes to analyzing posterior sampling.

The remainder of this paper is organized as follows. The next section discusses related literature. Section 3 then provides a formal problem statement. We describe UCB and posterior sampling algorithms in §4. We then establish in §5 a connection between them, which we apply in §6 to convert existing bounds for UCB algorithms to bounds for posterior sampling. Section 7 develops a new notion of dimension and presents Bayesian regret bounds that depend on it. Section 8 presents simulation results. A closing section makes concluding remarks.

2. Related literature.

One distinction of results presented in this paper is that they center around Bayesian regret as a measure of performance. In the next subsection, we discuss this choice and how it relates to performance measures used in other work. Following that, we review prior results and their relation to results of this paper.

2.1. Measures of performance.

Several recent papers have established theoretical results on posterior sampling. One difference between this work and ours is that we focus on a different measure of performance. These papers all study the algorithm's regret, which measures its cumulative loss relative to an algorithm that always selects the optimal action, for some fixed reward function. To derive these bounds, each paper fixes an uninformative prior distribution with a convenient analytic structure, and studies posterior sampling assuming this particular prior is used. With one exception (Agrawal and Goyal [6]), the focus is on the classical MAB problem, where sampling one action provides no information about others.

Posterior sampling can be applied to a much broader class of problems, and one of its greatest strengths is its ability to incorporate prior knowledge in a flexible and coherent way. We therefore aim to develop results that accommodate the use of a wide range of models. Accordingly, most of our results allow for an *arbitrary* prior distribution over a particular class of mean reward functions. To derive meaningful results at this level of generality, we study the algorithm's expected regret, where the expectation is taken with respect to the prior distribution over reward functions. This quantity is sometimes called the algorithm's *Bayesian regret*. We find this to be a practically relevant measure of performance and find this choice allows for more elegant analysis. Further, as we discuss in §3, the Bayesian regret bounds we provide in some cases immediately yield regret bounds.

In addition, studying Bayesian regret reveals deep connections between posterior sampling and the principle of *optimism in the face of uncertainty*, which we feel provides new conceptual insight into the algorithm's performance. Optimism in the face of uncertainty is a general principle and is not inherently tied to any measure of performance. Indeed, algorithms based on this principle have been shown to be asymptotically efficient in terms of both regret (Lai and Robbins [26]) and Bayesian regret (Lai [25]), to satisfy order optimal minimax regret bounds (Audibert and Bubeck [8]), to satisfy order optimal bounds on regret and Bayesian regret when the reward function is linear (Rusmevichientong and Tsitsiklis [31]), and to satisfy strong bounds when the reward function is sampled from a Gaussian process prior (Srinivas et al. [35]). We take a very general view of optimistic algorithms, allowing upper confidence bounds to be constructed in an essentially arbitrary way based on the algorithm's observations and possibly the prior distribution over reward functions.

2.2. Related results. Though it was first proposed in 1933, posterior sampling has until recently received relatively little attention. Interest in the algorithm grew after empirical studies (Chapelle and Li [15], Scott [34]) demonstrated performance exceeding state-of-the-art methods. An asymptotic convergence result was established by May et al. [30], but finite time guarantees remain limited. The development of further performance bounds was raised as an open problem at the 2012 Conference on Learning Theory (Li and Chapelle [29]).

Three recent papers (Agrawal and Goyal [4, 5]; Kauffmann et al. [22]) provide regret bounds for posterior sampling when applied to MAB problems with finitely many independent actions and rewards that follow Bernoulli processes. These results demonstrate that posterior sampling is asymptotically optimal for the class of problems considered. A key feature of the bounds is their dependence on the difference between the optimal and second-best mean reward values. Such bounds tend not to be meaningful when the number of actions is large or infinite unless they can be converted to bounds that are independent of this gap, which is sometimes the case.

In this paper, we establish distribution independent bounds. When the action space $\mathcal{A}$ is finite, we establish a finite-time Bayesian regret bound of order $\sqrt{|\mathcal{A}|T\log T}$. This matches what is implied by the analysis of Agrawal and Goyal [5]. However, our bound does not require actions are modeled independently, and our approach also leads to meaningful bounds for problems with large or infinite action sets.

Only one other paper has studied posterior sampling in a context involving dependent actions (Agrawal and Goyal [6]). That paper considers a contextual bandit model with arms whose mean reward values are given by a d-dimensional linear model. The cumulative T-period regret is shown to be of order $(d^2/\epsilon)\sqrt{T^{1+\epsilon}}\ln(Td)\ln(1/\delta)$ with probability at least $1-\delta$. Here $\epsilon \in (0,1)$ is a parameter used by the algorithm to control how quickly the posterior distribution concentrates. The Bayesian regret bounds we will establish are stronger than those implied by the results of Agrawal and Goyal [6]. In particular, we provide a Bayesian regret bound of order $d\sqrt{T}\ln T$ that holds for any compact set of actions. This is order optimal up to a factor of $\ln T$ (Rusmevichientong and Tsitsiklis [31]).

We are also the first to establish finite-time performance bounds for several other problem classes. One applies to linear models when the vector of coefficients is likely to be sparse; this bound is stronger than the aforementioned one that applies to linear models in the absence of sparsity assumptions. We establish the first bounds for posterior sampling when applied to generalized linear models and to problems with a general Gaussian prior. Finally, we establish bounds that apply very broadly and depend on a new notion of dimension.

Unlike most of the relevant literature, we study MAB problems in a general framework, allowing for complicated relationships between the rewards generated by different actions. The closest related work is that of Amin et al. [7], who consider the problem of learning the optimum of a function that lies in a known, but otherwise arbitrary set of functions. They provide bounds based on a new notion of dimension, but unfortunately, this notion does not provide a bound for posterior sampling. Other work provides general bounds for contextual bandit problems where the context space is allowed to be infinite, but the action space is small (see, e.g., Beygelzimer et al. [10]). Our model captures contextual bandits as a special case, but we emphasize problem instances with large or infinite action sets, and where the goal is to learn without sampling every possible action.

A focus of our paper is the connection between posterior sampling and UCB approaches. We discuss UCB algorithms in some detail in §4. UCB algorithms have been the primary approach considered in the segment of the stochastic MAB literature that treats models with dependent arms. Other approaches are the knowledge gradient algorithm (Ryzhov et al. [32]), forced exploration schemes for linear bandits (Abbasi-Yadkori et al. [1], Rusmevichientong and Tsitsiklis [31], Deshpande and Montanari [17]), and exponential weighting schemes (Beygelzimer et al. [10]).

There is an immense and rapidly growing literature on bandits with independent arms and on adversarial bandits. Theoretical work on stochastic bandits with independent arms often focuses on UCB algorithms (Auer et al. [9], Lai and Robbins [26]), or on the Gittin's index approach (Gittins and Jones [19]). Bubeck and Cesa-Bianchi [11] provide a review of work on UCB algorithms and on adversarial bandits. Gittins et al. [20] cover work on Gittin's indices and related extensions.

Since an initial version of this paper was made publicly available, the literature on the analysis of posterior sampling has rapidly grown. Korda et al. [24] extend their earlier work (Kauffmann et al. [22]) to the case where reward distributions lie in the one-dimensional exponential family. Bubeck and Liu [12] combine the regret decomposition we derive in §5 with the confidence bound analysis of Audibert and Bubeck [8] to tighten the bound provided in §6.1, and also consider a problem setting where the regret of posterior sampling is bounded uniformly over time. Li [28] explores a connection between posterior sampling and exponential weighting schemes, and Gopalan et al. [21] study the asymptotic growth rate of regret in problems with dependent arms.

3. Problem formulation. We consider a model involving a set of actions $\mathcal{A}$ and a set of real-valued functions $\mathcal{F} = \{f_\rho : \mathcal{A} \mapsto \mathbb{R} \mid \rho \in \Theta\}$, indexed by a parameter that takes values from an index set Θ. We will define random variables with respect to a probability space $(\Omega, \mathbb{F}, \mathbb{P})$. A random variable θ indexes the true reward function f_θ. At each time t, the agent is presented with a possibly random subset $\mathcal{A}_t \subseteq \mathcal{A}$ and selects an action $A_t \in \mathcal{A}_t$, after which she observes a reward R_t.

We denote by H_t the history $(\mathcal{A}_1, A_1, R_1, \ldots, \mathcal{A}_{t-1}, A_{t-1}, R_{t-1}, \mathcal{A}_t)$ of observations available to the agent when choosing an action A_t. The agent employs a policy $\pi = \{\pi_t \mid t \in \mathbb{N}\}$, which is a deterministic sequence of functions, each mapping the history H_t to a probability distribution over actions $\mathcal{A}$. For each realization of H_t, $\pi_t(H_t)$ is a distribution over $\mathcal{A}$ with support $\mathcal{A}_t$, though with some abuse of notation, we will often write this distribution as π_t. The action A_t is selected by sampling from the distribution π_t, so that $\mathbb{P}(A_t \in \cdot \mid \pi_t) = \mathbb{P}(A_t \in \cdot \mid H_t) = \pi_t(\cdot)$. We assume that $\mathbb{E}[R_t \mid H_t, \theta, A_t] = f_\theta(A_t)$. In other words, the realized reward is the mean reward value corrupted by zero mean noise. We will also assume that for each $f \in \mathcal{F}$ and $t \in \mathbb{N}$, $\arg\max_{a \in \mathcal{A}_t} f(a)$ is nonempty with probability one, though algorithms and results can be generalized to handle cases where this assumption does not hold.

The T-period regret of a policy π is the random variable defined by

$$\text{Regret}(T, \pi, \theta) = \sum_{t=1}^{T} \mathbb{E}\left[\max_{a \in \mathcal{A}_t} f_\theta(a) - f_\theta(A_t) \,\middle|\, \theta\right].$$

The T-period Bayesian regret is defined by $\mathbb{E}[\text{Regret}(T, \pi, \theta)]$, where the expectation is taken with respect to the prior distribution over θ. Hence

$$\text{BayesRegret}(T, \pi) = \sum_{t=1}^{T} \mathbb{E}\left[\max_{a \in \mathcal{A}_t} f_\theta(a) - f_\theta(A_t)\right].$$

This quantity is also called Bayes risk, or simply expected regret.

Remark 1. Measurability assumptions are required for the above expectations to be well defined. To avoid technicalities that do not present fundamental obstacles in the contexts we consider, we will not explicitly address measurability issues in this paper and instead simply assume that functions under consideration satisfy conditions that ensure relevant expectations are well defined.

Remark 2. All equalities between random variables in this paper hold almost surely with respect to the underlying probability space.

3.1. On regret and Bayesian regret. To interpret results about the regret and Bayesian regret of various algorithms and to appreciate their practical implications, it is useful to take note of several properties of and relationships between these performance measures. For starters, asymptotic bounds on Bayesian regret are essentially asymptotic bounds on regret. In particular, if $\text{BayesRegret}(T, \pi) = O(g(T))$ for some nonnegative function g, then an application of Markov's inequality shows $\text{Regret}(T, \pi, \theta) = O_P(g(T))$. Here O_P indicates that $\text{Regret}(T, \pi, \theta)/g(T)$ is stochastically bounded under the prior distribution. In other words, for all $\epsilon > 0$, there exists $M > 0$ such that

$$\mathbb{P}\left(\frac{\text{Regret}(T, \pi, \theta)}{g(T)} \geq M\right) \leq \epsilon \quad \forall\, T \in \mathbb{N}.$$

This observation can be further extended to establish a sense in which Bayesian regret is robust to prior mis-specification. In particular, if the agent's prior over θ is μ but for convenience he selects actions as though his prior were an alternative $\tilde{\mu}$, the resulting Bayesian regret satisfies

$$\mathbb{E}_{\theta_0 \sim \mu}[\text{Regret}(T, \pi, \theta_0)] \leq \left\| \frac{d\mu}{d\tilde{\mu}} \right\|_{\tilde{\mu}, \infty} \mathbb{E}_{\theta_0 \sim \tilde{\mu}}[\text{Regret}(T, \pi, \theta_0)],$$

where $d\mu/d\tilde{\mu}$ is the Radon-Nikodym derivative[1] of μ with respect to $\tilde{\mu}$ and $\|\cdot\|_{\tilde{\mu}, \infty}$ is the essential supremum magnitude with respect to $\tilde{\mu}$. Note that the final term on the right-hand side is the Bayesian regret for a problem with prior $\tilde{\mu}$ without mis-specification.

It is also worth noting that an algorithm's Bayesian regret can only differ significantly from its worst-case regret if regret varies significantly depending on the realization of θ. This provides one method of converting

[1] Note that the Radon-Nikodym derivative is only well defined when μ is absolutely continuous with respect to $\tilde{\mu}$.

Bayesian regret bounds to regret bounds. For example, consider the linear model $f_\theta(a) = \theta^T \phi(a)$, where $\Theta = \{\rho \in \mathbb{R}^d : \|\rho\|_2 = S\}$ is the boundary of a hypersphere in $\mathbb{R}^d$. Let $\mathcal{A}_t = \mathcal{A}$ for each t and let the set of feature vectors be $\{\phi(a) \mid a \in \mathcal{A}\} = \{u \in \mathbb{R}^d \mid \|u\|_2 \le 1\}$. Consider a problem instance where θ is uniformly distributed over Θ, and the noise terms $R_t - f_\theta(A_t)$ are independent of θ. By symmetry, the regret of most reasonable algorithms for this problem should be the same for all realizations of θ, and indeed this is the case for posterior sampling. Therefore, in this setting, Bayesian regret is equal to worst-case regret. This view also suggests that to attain strong minimax regret bounds, one should not choose a uniform prior as in Agrawal and Goyal [5], but should instead place more prior weight on the worst-possible realizations of θ (see the discussion of "least favorable" prior distributions in Lehmann and Casella [27]).

3.2. On changing action sets. Our stochastic model of action sets $\mathcal{A}_t$ is distinct relative to most of the MAB literature, which assumes that $\mathcal{A}_t = \mathcal{A}$. This construct allows our formulation to address a variety of practical issues that are usually viewed as beyond the scope of standard MAB formulations. Let us provide three examples.

Example 1 (Contextual Models). The contextual MAB model is a special case of the formulation presented above. In such a model, an exogenous Markov process X_t taking values in a set $\mathcal{X}$ influences rewards. In particular, the expected reward at time t is given by $f_\theta(a, X_t)$. However, this is mathematically equivalent to a problem with stochastic time-varying decision sets $\mathcal{A}_t$. In particular, one can define the set of actions to be the set of state-action pairs $\mathcal{A} := \{(x, a) : x \in \mathcal{X},\ a \in \mathcal{A}(x)\}$, and the set of available actions to be $\mathcal{A}_t = \{(X_t, a) : a \in \mathcal{A}(X_t)\}$.

Example 2 (Cautious Actions). In some applications, one may want to explore without risking terrible performance. This can be accomplished by restricting the set $\mathcal{A}_t$ to conservative actions. Then, the instantaneous regret in our framework is the gap between the reward from the chosen action and the reward from the best conservative action. In many settings, the Bayesian regret bounds we will establish for posterior sampling imply that the algorithm either attains near-optimal performance or converges to a point where any better decision is unacceptably risky.

A number of formulations of this flavor are amenable to efficient implementations of posterior sampling. For example, consider a problem where $\mathcal{A}$ is a polytope or ellipsoid in $\mathbb{R}^d$ and $f_\theta(a) = \langle a, \theta \rangle$. Suppose θ has a Gaussian prior and that reward noise is Gaussian. Then, the posterior distribution of θ is Gaussian. Consider an ellipsoidal confidence set $\mathcal{U}_t = \{u \mid \|u - \mu_t\|_{\Sigma_t} \le \beta\}$, for some scalar constant $\beta > 0$, where μ_t and Σ_t are the mean and covariance matrix of θ, conditioned on H_t. One can attain good worst-case performance with high probability by solving the robust optimization problem $V_{\text{robust}} \equiv \max_{a \in \mathcal{A}} \min_{u \in \mathcal{U}_t} \langle a, u \rangle$, which is a tractable linear saddle point problem. Letting our cautious set be given by

$$\mathcal{A}_t = \left\{ a \in \mathcal{A} \;\middle|\; \min_{u \in \mathcal{U}_t} \langle a, u \rangle \ge V_{\text{robust}} - \alpha \right\}$$

for some scalar constant $\alpha > 0$, we can then select an optimal cautious action given θ by solving $\max_{a \in \mathcal{A}_t} \langle a, \theta \rangle$, which is equivalent to

$$\begin{aligned} \text{maximize} \quad & \langle a, \theta \rangle \\ \text{subject to} \quad & a \in \mathcal{A} \\ & \|a\|_{\Sigma_t^{-1}} \le \frac{1}{\beta}(\langle a, \mu_t \rangle - V_{\text{robust}} + \alpha). \end{aligned}$$

This problem is computationally tractable, which accommodates efficient implementation of posterior sampling.

Example 3 (Adaptive Adversaries). Consider a model in which rewards are influenced by the choices of an adaptive adversary. At each time period, the adversary selects an action A_t^- from some set $\mathcal{A}^-$ based on past observations. The agent observes this action, responds with an action A_t^+ selected from a set $\mathcal{A}^+$, and receives a reward that depends on the pair of actions (A_t^+, A_t^-). This fits our framework if the action A_t is taken to be the pair (A_t^+, A_t^-), and the set of actions available to the agent is $\mathcal{A}_t = \{(a, A_t^-) \mid a \in \mathcal{A}^+\}$.

4. Algorithms. We will establish finite-time performance bounds for posterior sampling by leveraging prior results pertaining to UCB algorithms and a connection we will develop between the two classes of algorithms. To set the stage for our analysis, we discuss the algorithms in this section.

4.1. UCB algorithms. UCB algorithms have received a great deal of attention in the MAB literature. Such an algorithm makes use of a sequence of UCBs $U = \{U_t \mid t \in \mathbb{N}\}$, each of which is a function that takes the history H_t as its argument. For each realization of H_t, $U_t(H_t)$ is a function mapping $\mathcal{A}$ to $\mathbb{R}$. With some abuse of notation, we will often write this function as U_t and its value at $a \in \mathcal{A}$ as $U_t(a)$. The UCB $U_t(a)$ represents the greatest value of $f_\theta(a)$ that is statistically plausible given H_t. A UCB algorithm selects an action $\bar{A}_t \in \arg\max_{a \in \mathcal{A}_t} U_t(a)$ that maximizes the UCB. We will assume that the argmax operation breaks ties among optima in a deterministic way. As such, each action is determined by the history H_t, and for the policy $\pi = \{\pi_t \mid t \in \mathbb{N}\}$ followed by a UCB algorithm, each action distribution π_t concentrates all probability on a single action.

As a concrete example, consider Algorithm 1 proposed by Auer et al. [9] to address MAB problems with a finite number of independent actions. For such problems, $\mathcal{A}_t = \mathcal{A}$, θ is a vector with one independent component per action, and the reward function is given by $f_\theta(a) = \theta_a$. The algorithm begins by selecting each action once. Then, for each subsequent time $t > |\mathcal{A}|$, the algorithm generates point estimates of action rewards, defines UCBs based on them, and selects actions accordingly. For each action a, the point estimate $\hat{\theta}_t(a)$ is taken to be the average reward obtained from samples of action a taken prior to time t. The UCB is produced by adding an "uncertainty bonus" $\beta\sqrt{\log t / N_t(a)}$ to the point estimate, where $N_t(a)$ is the number of times action a was selected prior to time t and β is an algorithm parameter generally selected based on reward variances. This uncertainty bonus leads to an optimistic assessment of expected reward when there is uncertainty, and it is this optimism that encourages exploration that reduces uncertainty. As $N_t(a)$ increases, uncertainty about action a diminishes and so does the uncertainty bonus. The $\log t$ term ensures that the agent does not permanently rule out any action, which is important as there is always some chance of obtaining an overly pessimistic estimate by observing an unlikely sequence of rewards.

Algorithm 1 (Independent UCB)

1. **Initialize**: Select each action once
2. **Update Statistics**: For each $a \in \mathcal{A}$,
   $\hat{\theta}_t(a) \leftarrow$ sample average of observed rewards
   $N_t(a) \leftarrow$ number of times a sampled so far
3. **Select Action**:
   $$\bar{A}_t \in \arg\max_{a \in \mathcal{A}} \left\{ \hat{\theta}_t(a) + \beta \sqrt{\frac{\log t}{N_t(a)}} \right\}$$
4. **Increment** t **and Go to Step 2**.

Our second example treats a linear bandit problem. Here, we assume θ is drawn from a normal distribution $N(\mu_0, \Sigma_0)$ but without assuming that the covariance matrix is diagonal. We consider a linear reward function $f_\theta(a) = \langle \phi(a), \theta \rangle$ and assume the reward noise $R_t - f_\theta(A_t)$ is normally distributed and independent from (H_t, A_t, θ). One can show that, conditioned on the history H_t, θ remains normally distributed. Algorithm 2 presents a UCB algorithm for this problem. The expectations can be computed efficiently via Kalman filtering. The algorithm employs UCB $\langle \phi(a), \mu_t \rangle + \beta \log(t) \|\phi(a)\|_{\Sigma_t}$. The term $\|\phi(a)\|_{\Sigma_t}$ captures the posterior variance of θ in the direction $\phi(a)$, and, as with the case of independent arms, causes the uncertainty bonus $\beta \log(t) \|\phi(a)\|_{\Sigma_t}$ to diminish as the number of observations increases.

Algorithm 2 (Linear-Gaussian UCB)

1. **Update Statistics**:
   $\mu_t \leftarrow \mathbb{E}[\theta \mid H_t]$
   $\Sigma_t \leftarrow \mathbb{E}[(\theta - \mu_t)(\theta - \mu_t)^\top \mid H_t]$
2. **Select Action**:
   $\bar{A}_t \in \arg\max_{a \in \mathcal{A}} \{ \langle \phi(a), \mu_t \rangle + \beta \log(t) \|\phi(a)\|_{\Sigma_t} \}$
3. **Increment** t **and Go to Step 1**.

4.2. Posterior sampling. The posterior sampling algorithm simply samples each action according to the probability it is optimal. In particular, the algorithm applies action sampling distributions $\pi_t = \mathbb{P}(A_t^* \in \cdot \mid H_t)$, where A_t^* is a random variable that satisfies $A_t^* \in \arg\max_{a \in \mathcal{A}_t} f_\theta(a)$. Practical implementations typically operate by, at each time t, sampling an index $\hat{\theta}_t \in \Theta$ from the distribution $\mathbb{P}(\theta \in \cdot \mid H_t)$ and then generating an action $A_t \in \arg\max_{a \in \mathcal{A}_t} f_{\hat{\theta}_t}(a)$. To illustrate, let us provide concrete examples that address problems analogous to Algorithms 1 and 2.

Our first example involves a model with independent arms. In particular, suppose θ is drawn from a normal distribution $N(\mu_0, \Sigma_0)$ with a diagonal covariance matrix Σ_0, the reward function is given by $f_\theta(a) = \theta_a$, and

the reward noise $R_t - f_\theta(A_t)$ is normally distributed and independent from (H_t, A_t, θ). It then follows that, conditioned on the history H_t, θ remains normally distributed with independent components. Algorithm 3 presents an implementation of posterior sampling for this problem. The expectations are easy to compute and can also be computed recursively.

Algorithm 3 (Independent posterior sampling)

1. **Sample Model**:
   $\hat{\theta}_t \sim N(\mu_{t-1}, \Sigma_{t-1})$
2. **Select Action**:
   $A_t \in \arg\max_{a \in \mathcal{A}} \hat{\theta}_t(a)$
3. **Update Statistics**: For each a,
   $\mu_{ta} \leftarrow \mathbb{E}[\theta_a \mid H_t]$
   $\Sigma_{taa} \leftarrow \mathbb{E}[(\theta_a - \mu_{ta})^2 \mid H_t]$
4. **Increment t and Go to Step 1**.

Our second example treats a linear bandit problem. Algorithm 4 presents a posterior sampling analog to Algorithm 2. As before, we assume θ is drawn from a normal distribution $N(\mu_0, \Sigma_0)$. We consider a linear reward function $f_\theta(a) = \langle \phi(a), \theta \rangle$ and assume the reward noise $R_t - f_\theta(A_t)$ is normally distributed and independent from (H_t, A_t, θ).

Algorithm 4 (Linear posterior sampling)

1. **Sample Model**:
   $\hat{\theta}_t \sim N(\mu_{t-1}, \Sigma_{t-1})$
2. **Select Action**:
   $A_t \in \arg\max_{a \in \mathcal{A}} \langle \phi(a), \hat{\theta}_t \rangle$
3. **Update Statistics**:
   $\mu_t \leftarrow \mathbb{E}[\theta \mid H_t]$
   $\Sigma_t \leftarrow \mathbb{E}[(\theta - \mu_t)(\theta - \mu_t)^\top \mid H_t]$
4. **Increment t and Go to Step 1**.

4.3. Potential advantages of posterior sampling. Well-designed optimistic algorithms can be extremely effective. When simple and efficient UCB algorithms are available, they may be preferable to posterior sampling. Our work is primarily motivated by the challenges in designing optimistic algorithms to address very complicated problems, and the important advantages posterior sampling sometimes offers in such cases.

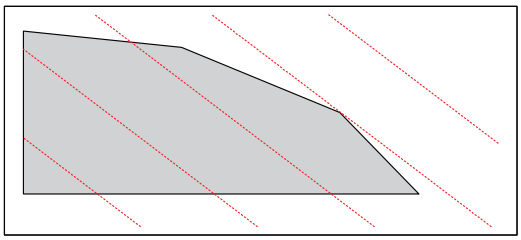
The performance of a UCB algorithm depends critically on the choice of UCBs $(U_t\colon t \in \mathbb{N})$. These functions should be chosen so that $U_t(A^*) \geq f_\theta(A^*)$ with high probability. However, unless the posterior distribution of $f_\theta(a)$ can be expressed in closed form, computing high quantiles of this distribution can require extensive Monte-Carlo simulation for each possible action. In addition, since A^* depends on θ, it isn't even clear that $U_t(a)$ should be set to a particular quantile of the posterior distribution of $f_\theta(a)$. Strong frequentist UCBs were recently developed (Cappé et al. [14]) for problems with independent arms, but it is often unclear how to generate tight confidence sets for more complicated models. In fact, even in the case of a linear model, the design of confidence sets has required sophisticated tools from the study of multivariate self-normalized martingale processes (Abbasi-Yadkori et al. [2]). We believe posterior sampling provides a powerful tool for practitioners, as a posterior sampling approach can be designed for complicated models without sophisticated statistical analysis. Further, posterior sampling does not require computing posterior distributions but only sampling from posterior distributions. As such, Markov Chain Monte- Carlo methods can often be used to efficiently generate samples even when the posterior distribution is complex.

Posterior sampling can also offer critical computational advantages over UCB algorithms when the action space is intractably large. Consider the problem of learning to solve a linear program, where the action set $\mathcal{A}$ is a polytope encoded in terms of linear inequalities, and $f_\theta(a) := \langle \phi(a), \theta \rangle$ is a linear function. Algorithm 2 becomes impractical, because as observed by Dani et al. [16], the action selection step entails solving a problem equivalent to linearly constrained negative definite quadratic optimization, which is NP-hard (Sahni [33]).[2] By contrast, the action selection step of Algorithm 4 only requires solving a linear program. The figure below displays the level sets of the linear objective $\langle \phi(a), \hat{\theta} \rangle$ and of the UCBs used by Algorithm 2. While posterior sampling preserves

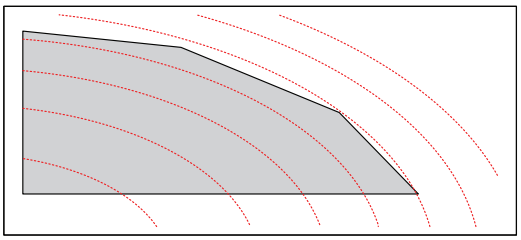
[2] Dani et al. [16] studies a slightly different UCB algorithm, but the optimization step shares the same structure.

the linear structure of the functions $f_\theta(a)$, it is challenging to maximize the UCBs of Algorithm 2, which are strictly convex.

(a) *Tractable* problem of maximizing a linear function over a polytope



(b) *Intractable* problem of maximizing an UCB over a polytope



It is worth mentioning that because posterior sampling requires specifying a fully probabilistic model of the underlying system, it may not be ideally suited for every practical setting. In particular, when dealing with some complex classes of functions, specifying an appropriate prior can be a challenge, while there may be alternative algorithms that address the problem in an elegant and practically effective way.

5. Confidence bounds and regret decompositions. Unlike UCB algorithms, posterior sampling does not make use of UCBs to encourage exploration and instead relies on randomization. As such, the two classes of algorithm seem very different. However, we will establish in this section a connection that will enable us in §6 to derive performance bounds for posterior sampling from those that apply to UCB algorithms. Since UCB algorithms have received much more attention, this leads to a number of new results about posterior sampling. Further, the relationship yields insight into the performance advantages of posterior sampling.

5.1. UCB regret decomposition. Consider a UCB algorithm with a UCB sequence $U = \{U_t \mid t \in \mathbb{N}\}$. Recall that $\bar{A}_t \in \arg\max_{a \in \mathcal{A}_t} U_t(a)$ and $A_t^* \in \arg\max_{a \in \mathcal{A}_t} f_\theta(a)$. We have the following simple regret decomposition:

$$\begin{aligned} f_\theta(A_t^*) - f_\theta(\bar{A}_t) &= f_\theta(A_t^*) - U_t(\bar{A}_t) + U_t(\bar{A}_t) - f_\theta(\bar{A}_t) \\ &\le [f_\theta(A_t^*) - U_t(A_t^*)] + [U_t(\bar{A}_t) - f_\theta(\bar{A}_t)]. \end{aligned} \tag{1}$$

The inequality follows from the fact that $\bar{A}_t$ is chosen to maximize U_t. If the UCB is an upper bound with high probability, as one would expect from a UCB algorithm, then the first term is negative with high probability. The second term, $U_t(\bar{A}_t) - f_\theta(\bar{A}_t)$, penalizes for the width of the confidence interval. As actions are sampled, U_t should diminish and converge on f_θ. As such, both terms of the decomposition should eventually vanish. An important feature of this decomposition is that, so long as the first term is negative, it bounds regret in terms of uncertainty about *the current action* $\bar{A}_t$.

Taking the expectation of (1) establishes that the T-period Bayesian regret of a UCB algorithm satisfies

$$\mathrm{BayesRegret}(T, \pi^U) \le \mathbb{E}\sum_{t=1}^{T}[U_t(\bar{A}_t) - f_\theta(\bar{A}_t)] + \mathbb{E}\sum_{t=1}^{T}[f_\theta(A_t^*) - U_t(A_t^*)], \tag{2}$$

where π^U is the policy derived from U.

5.2. Posterior sampling regret decomposition. As established by the following proposition, the Bayesian regret of posterior sampling decomposes in a way analogous to what we have shown for UCB algorithms. Recall that, with some abuse of notation, for a UCB sequence $\{U_t \mid t \in \mathbb{N}\}$, we denote by $U_t(a)$ the random variable $U_t(H_t)(a)$. The following proposition allows U_t to be an arbitrary real-valued function of H_t and $a \in \mathcal{A}$. Let π^{PS} denote the policy followed by posterior sampling.

Proposition 1. *For any UCB sequence* $\{U_t \mid t \in \mathbb{N}\}$,

$$\mathrm{BayesRegret}(T, \pi^{PS}) = \mathbb{E}\sum_{t=1}^{T}[U_t(A_t) - f_\theta(A_t)] + \mathbb{E}\sum_{t=1}^{T}[f_\theta(A_t^*) - U_t(A_t^*)] \tag{3}$$

for all $T \in \mathbb{N}$.

PROOF. Note that, conditioned on H_t, the optimal action A_t^* and the action A_t selected by posterior sampling are identically distributed, and U_t is deterministic. Hence $\mathbb{E}[U_t(A_t^*) \mid H_t] = \mathbb{E}[U_t(A_t) \mid H_t]$. Therefore

$$\begin{aligned}\mathbb{E}[f_\theta(A_t^*) - f_\theta(A_t)] &= \mathbb{E}[\mathbb{E}[f_\theta(A_t^*) - f_\theta(A_t) \mid H_t]] \\ &= \mathbb{E}[\mathbb{E}[U_t(A_t) - U_t(A_t^*) + f_\theta(A_t^*) - f_\theta(A_t) \mid H_t]] \\ &= \mathbb{E}[\mathbb{E}[U_t(A_t) - f_\theta(A_t) \mid H_t] + \mathbb{E}[f_\theta(A_t^*) - U_t(A_t^*) \mid H_t]] \\ &= \mathbb{E}[U_t(A_t) - f_\theta(A_t)] + \mathbb{E}[f_\theta(A_t^*) - U_t(A_t^*)].\end{aligned}$$

Summing over t gives the result. □

To compare (2) and (3), consider the case where f_θ takes values in $[0, C]$. Then

$$\text{BayesRegret}(T, \pi^U) \le \mathbb{E}\sum_{t=1}^{T}[U_t(\bar{A}_t) - f_\theta(\bar{A}_t)] + C\sum_{t=1}^{T}\mathbb{P}(f_\theta(A_t^*) > U_t(A_t^*))$$

and

$$\text{BayesRegret}(T, \pi^{PS}) \le \mathbb{E}\sum_{t=1}^{T}[U_t(A_t) - f_\theta(A_t)] + C\sum_{t=1}^{T}\mathbb{P}(f_\theta(A_t^*) > U_t(A_t^*)).$$

An important difference to take note of is that the Bayesian regret bound of π^U depends on the specific UCB sequence U used by the UCB algorithm in question, whereas the bound of π^{PS} applies simultaneously for *all* UCB sequences. This suggests that, while the Bayesian regret of a UCB algorithm depends critically on the specific choice of confidence sets, posterior sampling depends on the best-possible choice of confidence sets. This is a crucial advantage when there are complicated dependencies among actions, as designing and computing with appropriate confidence sets presents significant challenges. This difficulty is likely the main reason that posterior sampling significantly outperforms recently proposed UCB algorithms in the simulations presented in §8.

We have shown how UCBs characterize Bayesian regret bounds for posterior sampling. We will leverage this concept in the next two sections. Let us emphasize though, that while our *analysis* of posterior sampling will make use of UCBs, the actual *performance* of posterior sampling does not depend on UCBs used in the analysis.

6. From UCB to posterior sampling regret bounds.

In this section, we present Bayesian regret bounds for posterior sampling that can be derived by combining our regret decomposition (3) with results from prior work on UCB regret bounds. Each UCB regret bound was established through a common procedure, which entailed specifying lower and UCBs $L_t\colon \mathcal{A} \mapsto \mathbb{R}$ and $U_t\colon \mathcal{A} \mapsto \mathbb{R}$ so that $L_t(a) \le f_\theta(a) \le U_t(a)$ with high probability for each t and a, and then providing an expression that dominates the sum $\sum_1^T (U_t - L_t)(a_t)$ for all sequences of actions $a_1, \ldots, a_T$. As we will show, each such analysis together with our regret decomposition (3) leads to a Bayesian regret bound for posterior sampling.

6.1. Finitely many actions.

We consider in this section a problem with $|\mathcal{A}| < \infty$ actions and rewards satisfying $R_t \in [0, 1]$ for all t almost surely. We note, however, that the results we discuss can be extended to cases where R_t is not bounded but where instead its distribution is "light tailed." It is also worth noting that we make no further assumptions on the class of reward functions $\mathcal{F}$ or on the prior distribution over θ.

In this setting, Algorithm 1, which was proposed by Auer et al. [9], is known to satisfy a problem independent regret bound of order $\sqrt{|\mathcal{A}|T\log T}$. Under an additional assumption that action rewards are independent and take values in $\{0, 1\}$, an order $\sqrt{|\mathcal{A}|T\log T}$ regret bound for posterior sampling is also available (Agrawal and Goyal [5]).

Here, we provide a Bayesian regret bound that is also of order $\sqrt{|\mathcal{A}|T\log T}$ but does not require that action rewards are independent or binary. Our analysis, like that of Auer et al. [9], makes use of confidence sets that are Cartesian products of action-specific confidence intervals. The regret decomposition (3) lets us use such confidence sets to produce bounds for posterior sampling even when the algorithm itself may exploit dependencies among actions.

PROPOSITION 2. *If* $|\mathcal{A}| = K < \infty$ *and* $R_t \in [0, 1]$, *then for any* $T \in \mathbb{N}$

$$\text{BayesRegret}(T, \pi^{\text{PS}}) \le 2\min\{K, T\} + 4\sqrt{KT(2 + 6\log(T))}. \tag{4}$$

PROOF. Let $N_t(a) = \sum_{l=1}^{t} \mathbf{1}(A_t = a)$ denote the number of times a is sampled over the first t periods, and $\hat{\mu}_t(a) = N_t(a)^{-1} \sum_{l=1}^{t} \mathbf{1}(A_t = a) R_t$ denote the empirical average reward from these samples. Define upper and lower confidence bounds as follows:

$$U_t(a) = \min\left\{\hat{\mu}_{t-1}(a) + \sqrt{\frac{2+6\log(T)}{N_{t-1}(a)}}, 1\right\} \qquad L_t(a) = \max\left\{\hat{\mu}_{t-1}(a) - \sqrt{\frac{2+6\log(T)}{N_{t-1}(a)}}, 0\right\}. \tag{5}$$

The next lemma, which is a consequence of analysis in Abbasi-Yadkori et al. [2], shows these hold with high probability. Were actions sampled in an iid fashion, this lemma would follow immediately from the Hoeffding inequality. For more details, see Appendix A.

LEMMA 1. *If $U_t(a)$ and $L_t(a)$ are defined as in (5), then $\mathbb{P}(\bigcup_{t=1}^{T}\{f_\theta(a) \notin [L_t(a), U_t(a)]\}) \le 1/T$.*

First, consider the case where $T \le K$. Since $f_\theta(a) \in [0,1]$, $\text{BayesRegret}(T, \pi^{PS}) \le T = \min\{K, T\}$. Now, assume $T > K$. Then

$$\begin{aligned}\text{BayesRegret}(T, \pi^{\text{PS}}) &\le \mathbb{E}\left[\sum_{t=1}^{T}(U_t - L_t)(A_t)\right] + T\,\mathbb{P}\left(\bigcup_{a\in\mathcal{A}}\bigcup_{t=1}^{T}\{f_\theta(a) \notin [L_t(a), U_t(a)]\}\right)\\ &\le \mathbb{E}\left[\sum_{t=1}^{T}(U_t - L_t)(A_t)\right] + K.\end{aligned}$$

We now turn to bounding $\sum_{t=1}^{T}(U_t - L_t)(A_t)$. Let $\mathcal{T}_a = \{t \le T\colon A_t = a\}$ denote the periods in which a is selected. Then $\sum_{t=1}^{T}(U_t - L_t)(A_t) = \sum_{a\in\mathcal{A}}\sum_{t\in\mathcal{T}_a}(U_t - L_t)(a)$. We show

$$\sum_{t\in\mathcal{T}_a}(U_t - L_t)(a) \le 1 + 2\sqrt{2+6\log(T)}\sum_{t\in\mathcal{T}_a}(1+N_{t-1}(a))^{-1/2} = 1 + 2\sqrt{2+6\log(T)}\sum_{j=0}^{N_T(a)-1}(j+1)^{-1/2}$$

and

$$\sum_{j=0}^{N_T(a)-1}(j+1)^{-1/2} \le \int_{x=0}^{N_T(a)} x^{-1/2}\,dx = 2\sqrt{N_T(a)}.$$

Summing over actions and applying the Cauchy-Schwartz inequality yields,

$$\begin{aligned}\text{BayesRegret}(T, \pi^{PS}) \le 2K + 4\sqrt{2+6\log(T)}\sum_{a\in\mathcal{A}}\sqrt{N_T(a)} &\overset{\text{(a)}}{\le} 2K + 4\sqrt{(2+6\log(T))K\sum_{a}N_T(a)}\\ &= 2K + 4\sqrt{KT(2+6\log(T))}\\ &\overset{\text{(b)}}{=} 2\min\{K,T\} + 4\sqrt{KT(2+6\log(T))},\end{aligned}$$

where (a) follows from the Cauchy-Shwartz inequality and (b) follows from the assumption that $T > K$. □

6.2. Linear and generalized linear models. We now consider function classes that represent linear and generalized linear models. The bound of Proposition 2 applies so long as the number of actions is finite, but we will establish alternative bounds that depend on the dimension of the function class rather than the number of actions. Such bounds accommodate problems with infinite action sets and can be much stronger than the bound of Proposition 2 if there are many actions.

The Bayesian regret bounds we provide in this section derive from regret bounds of the UCB literature. In §7, we will establish a Bayesian regret bound that is as strong for the case of linear models and stronger for the case of generalized linear models. Since the results of §7 to a large extent supersede those we present here, we aim to be brief and avoid formal proofs in this section's discussion of the bounds and how they follow from results in the literature.

6.2.1. Linear models. In the "linear bandit" problem studied by Abbasi-Yadkori et al. [2, 3], Dani et al. [16], and Rusmevichientong and Tsitsiklis [31], reward functions are parameterized by a vector $\theta \in \Theta \subset \mathbb{R}^d$, and there is a known feature mapping $\phi\colon \mathcal{A} \mapsto \mathbb{R}^d$ such that $f_\theta(a) = \langle \phi(a), \theta\rangle$. The following proposition establishes Bayesian regret bounds for such problems. The proposition uses the term σ-sub-Gaussian to describe any random variable X that satisfies $\mathbb{E}\exp(\lambda X) \le \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$.

PROPOSITION 3. *Fix positive constants σ, c_1, and c_2. If $\Theta \subset \mathbb{R}^d$, $f_\theta(a) = \langle \phi(a), \theta \rangle$ for some ϕ: $\mathcal{A} \mapsto \mathbb{R}$, $\sup_{\rho \in \Theta} \|\rho\|_2 \le c_1$, and $\sup_{a \in \mathcal{A}} \|\phi(a)\|_2 \le c_2$, and for each t, $R_t - f_\theta(A_t)$ conditioned on (H_t, A_t, θ) is σ-sub-Gaussian, then*

$$\text{BayesRegret}(T, \pi^{PS}) = O(d \log T \sqrt{T})$$

and

$$\text{BayesRegret}(T, \pi^{PS}) = \tilde{O}(\mathbb{E}\sqrt{\|\theta\|_0 dT}).$$

The second bound essentially replaces the dependence on the dimension d with one on $\mathbb{E}\sqrt{\|\theta\|_0 d}$. The "zero norm" $\|\theta\|_0$ is the number of nonzero components, which can be much smaller than d when the reward function admits a sparse representation. Note that $\tilde{O}$ ignores logarithmic factors. Both bounds follow from our regret decomposition (3) together with the analysis of Abbasi-Yadkori et al. [2] in the case of the first bound, and the analysis of Abbasi-Yadkori et al. [3] in the case of the second bound. We now provide a brief sketch of how these bounds can be derived.

If f_θ takes values in $[-C, C]$, then (3) implies

$$\text{BayesRegret}(T, \pi^{PS}) \le \mathbb{E}\sum_{t=1}^{T}[U_t(A_t) - L_t(A_t)] + 2C\sum_{t=1}^{T}[\mathbb{P}(f_\theta(A_t^*) > U_t(A_t^*)) + \mathbb{P}(f_\theta(A_t) < L_t(A_t))]. \quad (6)$$

The analyses of Abbasi-Yadkori et al. [2] and [3] follow two steps that can be used to bound the right-hand side of this equation. In the first step, an ellipsoidal confidence set $\Theta_t := \{\rho \in \mathbb{R}^d : \|\rho - \hat{\theta}_t\|_{V_t} \le \sqrt{\beta_t}\}$ is constructed, where for some $\lambda \in \mathbb{R}$, $V_t := \sum_{k=1}^{t} \phi(A_t)\phi(A_t)^T + \lambda I$ captures the amount of exploration carried out in each direction up to time t. The upper and lower bounds induced by the ellipsoid are $U_t(a) := \min\{C, \max_{\rho \in \Theta_t}(\rho^T \phi(a))\}$ and $L_t(a) := \max\{-C, \min_{\rho \in \Theta_t}(\rho^T \phi(a))\}$. If the sequence of confidence parameters $\beta_1, \ldots, \beta_T$ is selected so that $\mathbb{P}(\theta \notin \Theta_t \mid H_t) \le 1/T$, then the second term of the regret decomposition is less than $4C$. For these confidence sets, the second step establishes a bound on $\sum_1^T (U_t - L_t)(a_t)$ that holds for any sequence of actions. The analyses presented on pages 7–8 of Dani et al. [16] and pages 14–15 of Abbasi-Yadkori et al. [2] each implies such a bound of order $\sqrt{d \max_{t \le T} \beta_t T \log(T/\lambda)}$. Plugging in closed-form expressions for β_t provided in these papers leads to the bounds of Proposition 3.

6.2.2. Generalized linear models. In a generalized linear model, the reward function takes the form $f_\theta(a) := g(\langle \phi(a), \theta \rangle)$, where the inverse link function g is strictly increasing and continuously differentiable. The analysis of Filippi et al. [18] can be applied to establish a Bayesian regret bound for posterior sampling, but with one caveat. The algorithm considered in Filippi et al. [18] begins by selecting a sequence of actions $a_1, \ldots, a_d$ with linearly independent feature vectors $\phi(a_1), \ldots, \phi(a_d)$. Until now, we haven't even assumed such actions exist or that they are guaranteed to be feasible over the first d time periods. After this period of designed exploration, the algorithm selects at each time an action that maximizes a UCB. What we will establish using the results from Filippi et al. [18] is a bound on a similarly modified version of posterior sampling, in which the first d actions taken are $a_1, \ldots, a_d$, while subsequent actions are selected by posterior sampling. Note that the posterior distribution employed at time $d+1$ is conditioned on observations made over the first d time periods. We denote this modified posterior sampling algorithm by $\pi^{IPS}_{a_1, \ldots, a_d}$. It is worth mentioning here, that in §7, we present a result with a stronger bound that applies to the standard version of posterior sampling, which does not include a designed exploration period.

PROPOSITION 4. *Fix positive constants c_1, c_2, C, and λ. If $\Theta \subset \mathbb{R}^d$, $f_\theta(a) = g(\langle \phi(a), \theta \rangle)$ for some strictly increasing continuously differentiable function g: $\mathbb{R} \mapsto \mathbb{R}$, $\sup_{\rho \in \Theta} \|\rho\|_2 \le c_1$, $\sup_{a \in \mathcal{A}} \|\phi(a)\|_2 \le c_2$, $\mathcal{A}_t = \mathcal{A}$ for all t, $\sum_{i=1}^{d} \phi(a_i)\phi(a_i)^T \succeq \lambda I$ for some $a_1, \ldots, a_d \in \mathcal{A}$, and $R_t \in [0, C]$ for all t, then*

$$\text{BayesRegret}(T, \pi^{IPS}_{a_1, \ldots, a_d}) = O(rd \log^{3/2} T \sqrt{T}),$$

where $r = \sup_{\rho, a} g'(\langle \phi(a), \rho \rangle) / \inf_{\rho, a} g'(\langle \phi(a), \rho \rangle)$.

Like the analyses of Abbasi-Yadkori et al. [2] and [3], which apply to linear models, the analysis of Filippi et al. [18] follows two steps that together bound both terms of our regret decomposition (6). First, an ellipsoidal confidence set Θ_t is constructed, centered around a quasi-maximum likelihood estimator. This confidence set is designed to contain θ with high probability. Given confidence bounds $U_t(a) := \min\{C, \max_{\rho \in \Theta_t} g(\langle \phi(a), \rho \rangle)\}$ and $L_t(a) := \max\{0, \min_{\rho \in \Theta_t} g(\langle \phi(a), \rho \rangle)\}$, a worst-case bound on $\sum_1^T (U_t - L_t)(a_t)$ is established. The bound is similar to those established for the linear case, but there is an added dependence on the slope of g.

6.3. Gaussian processes. In this section, we consider the case where the reward function f_θ is sampled from a Gaussian process. That is, the stochastic process $(f_\theta(a)\colon a \in \mathcal{A})$ is such that for any $a_1, \ldots, a_k \in \mathcal{A}$, the collection $f_\theta(a_1), \ldots, f_\theta(a_k)$ follows a multivariate Gaussain distribution. Srinivas et al. [35] study a UCB algorithm designed for such problems and provide general regret bounds. Again, through the regret decomposition (3), their analysis provides a Bayesian regret bound for posterior sampling.

For simplicity, we focus our discussion on the case where $\mathcal{A}$ is finite, so that $(f_\theta(a)\colon a \in \mathcal{A})$ follows a multivariate Gaussian distribution. As shown by Srinivas et al. [35], the results extend to infinite action sets through a discretization argument as long as certain smoothness conditions are satisfied.

When confidence bounds hold, a UCB algorithm incurs high regret from sampling an action only when the confidence bound at that action is loose. In that case, one would expect the algorithm to learn a lot about f_θ based on the observed reward. This suggests the algorithm's cumulative regret may be bounded in an appropriate sense by the total amount it is expected to learn. Leveraging the structure of the Gaussian distribution, Srinivas et al. [35] formalize this idea. They bound the regret of their UCB algorithm in terms of the maximum amount that *any* algorithm could learn about f_θ. They use an information-theoretic measure of learning: the information gain. This is defined to be the difference between the entropy of the prior distribution of $(f_\theta(a)\colon a \in \mathcal{A})$ and the entropy of the posterior. The maximum possible information gain is denoted γ_T, where the maximum is taken over all sequences $a_1, \ldots, a_T$.[3] Their analysis also supports the following result on posterior sampling.

Proposition 5. *If $\mathcal{A}$ is finite, $(f_\theta(a)\colon a \in \mathcal{A})$ follows a multivariate Gaussian distribution with marginal variances bounded by 1, $R_t - f_\theta(A_t)$ is independent of (H_t, θ, A_t), and $\{R_t - f_\theta(A_t) \mid t \in \mathbb{N}\}$ is an iid sequence of zero mean Gaussian random variables with variance σ^2, then*

$$\mathrm{BayesRegret}(T, \pi^{PS}) \le 1 + 2\sqrt{T\gamma_T \ln(1+\sigma^{-2})^{-1} \ln\left(\frac{(T^2+1)|\mathcal{A}|}{\sqrt{2\pi}}\right)}$$

for all $T \in \mathbb{N}$.

Srinivas et al. [35] also provide bounds on γ_T for kernels commonly used in Gaussian process regression, including the linear kernel, radial basis kernel, and Matérn kernel. Combined with the above proposition, this yields explicit Bayesian regret bounds in these cases.

We will briefly comment on their analysis and how it provides a bound for posterior sampling. First, note that the posterior distribution is Gaussian, which suggests a UCB of the form $U_t(a) := \mu_{t-1}(a) + \sqrt{\beta_t}\sigma_{t-1}(a)$, where $\mu_{t-1}(a)$ is the posterior mean, $\sigma_{t-1}(a)$ is the posterior standard deviation of $f_\theta(a)$, and β_t is a confidence parameter. We can provide a Bayesian regret bound by bounding both terms of (3). The next lemma, which follows bounds the second term. Some new analysis is required since the Gaussian distribution is unbounded, and we study Bayesian regret, whereas Srinivas et al. [35] bound regret with high probability under the prior.

Lemma 2. *If $U_t(a) := \mu_{t-1}(a) + \sqrt{\beta_t}\sigma_{t-1}(a)$ and $\beta_t := 2\ln(((t^2+1)|\mathcal{A}|)/\sqrt{2\pi})$, then for all $T \in \mathbb{N}$ $\mathbb{E}\sum_{t=1}^{T}[f_\theta(A_t^*) - U_t(A_t^*)] \le 1$.*

Proof. First, if $X \sim N(\mu, \sigma^2)$, then if $\mu \le 0$, $\mathbb{E}[X\mathbf{1}\{X > 0\}] = \int_0^\infty (x/(\sigma\sqrt{2\pi}))\exp\{(-(x-\mu)^2)/(2\sigma^2)\}\,dx = (\sigma/\sqrt{2\pi})\exp\{-\mu^2/(2\sigma^2)\}$.

Then since the posterior distribution of $f_\theta(a) - U_t(a)$ is normal with mean $-\sqrt{\beta_t}\sigma_{t-1}(a)$ and variance $\sigma_{t-1}^2(a)$,

$$\mathbb{E}[\mathbf{1}\{f_\theta(a) - U_t(a) \ge 0\}[f_\theta(a) - U_t(a)] \mid H_t] = \frac{\sigma_{t-1}(a)}{\sqrt{2\pi}}\exp\left\{\frac{-\beta_t}{2}\right\} = \frac{\sigma_{t-1}(a)}{(t^2+1)|\mathcal{A}|} \le \frac{1}{(t^2+1)|\mathcal{A}|}. \tag{7}$$

The final inequality above follows from the assumption that $\sigma_0(a) \le 1$. The claim follows from (7) since

$$\mathbb{E}\sum_{t=1}^{T}[f_\theta(A_t^*) - U_t(A_t^*)] \le \sum_{t=1}^{\infty}\sum_{a\in\mathcal{A}}\mathbb{E}[\mathbf{1}\{f_\theta(a) - U_t(a) \ge 0\}[f_\theta(a) - U_t(a)]] \le \sum_{t=1}^{\infty}\frac{1}{(t^2+1)} \le 1. \quad \square$$

Now, consider the first term of (3), which is

$$\mathbb{E}\sum_{t=1}^{T}(U_t - f_\theta)(A_t) = \mathbb{E}\sum_{t=1}^{T}(U_t - \mu_{t-1})(A_t) = \mathbb{E}\sum_{t=1}^{T}\sqrt{\beta_t}\sigma_{t-1}(A_t) \le \mathbb{E}\sqrt{T\max_{t\le T}\beta_t}\sqrt{\sum_{t=1}^{T}\sigma_{t-1}^2(A_t)}.$$

[3] An important property of the Gaussian distribution is that the information gain does not depend on the observed rewards. This is because the posterior covariance of a multivariate Gaussian is a deterministic function of the points that were sampled. For this reason, this maximum is well defined.

Here, the second equality follows by the tower property of conditional expectation, and the final step follows from the Cauchy-Schwartz inequality. Therefore, to establish a Bayesian regret bound, it is sufficient to provide a bound on the sum of posterior variances $\sum_{t=1}^{T} \sigma_{t-1}^2(a_t)$ that holds for any $a_1, \ldots, a_T$. Under the assumption that $\sigma_0(a) \le 1$, the proof of Lemma 5.4 of Srinivas et al. [35] shows that $\sigma_{t-1}^2(a_t) \le \alpha^{-1} \log(1 + \sigma^{-2}\sigma_{t-1}^2(a_t))$, where $\alpha = 1 + \sigma^{-2}$. At the same time, Lemma 5.3 of Srinivas et al. [35] shows the information gain from selecting $a_1, \ldots, a_T$ is equal to $\frac{1}{2}\sum_{t=1}^{T} \log(1 + \sigma^{-2}\sigma_{t-1}^2(a_t))$. This shows that for any actions $a_1, \ldots, a_T$, the sum of posterior variances $\sum_{t=1}^{T} \sigma_{t-1}^2(a_t)$ can be bounded in terms of the information gain from selecting $a_1, \ldots, a_T$. Therefore $\sum_{t=1}^{T} \sigma_{t-1}^2(A_t)$ can be bounded in terms of the *largest possible* information gain γ_T.

7. Bounds for general function classes. The previous section treated models in which the relationship among action rewards takes a simple and tractable form. Indeed, nearly all of the MAB literature focuses on such problems. Posterior sampling can be applied to a much broader class of models. As such, more general results that hold beyond restrictive cases are of particular interest. In this section, we provide a Bayesian regret bound that applies when the reward function lies in a known, but otherwise arbitrary class of uniformly bounded real-valued functions $\mathcal{F}$. Our analysis of this abstract framework yields a more general result that applies beyond the scope of specific problems that have been studied in the literature, and also identifies factors that unify more specialized prior results. Further, our more general result when specialized to linear models recovers the strongest known Bayesian regret bound and in the case of generalized linear models, yields a bound stronger than that established in prior literature.

If $\mathcal{F}$ is not appropriately restricted, it is impossible to guarantee any reasonably attractive level of Bayesian regret. For example, in a case where $\mathcal{A} = [0, 1]$, $f_\theta(a) = \mathbf{1}(\theta = a)$, $\mathcal{F} = \{f_\theta \mid \theta \in [0, 1]\}$, and θ is uniformly distributed over $[0, 1]$, it is easy to see that the Bayesian regret of *any* algorithm over T periods is T, which is no different from the worst level of performance an agent can experience.

This example highlights the fact that Bayesian regret bounds must depend on the function class $\mathcal{F}$. The bound we develop in this section depends on $\mathcal{F}$ through two measures of complexity. The first is the Kolmogorov dimension, which measures the growth rate of the covering numbers of $\mathcal{F}$ and is closely related to measures of complexity that are common in the supervised learning literature. It roughly captures the sensitivity of $\mathcal{F}$ to statistical overfitting. The second measure is a new notion we introduce, which we refer to as the eluder dimension. This captures how effectively the value of unobserved actions can be inferred from observed samples. We highlight in §7.3 why notions of dimension common to the supervised learning literature are insufficient for our purposes.

Though the results of this section are very general, they do not apply to the entire range of problems represented by the formulation we introduced in §3. In particular, throughout the scope of this section, we fix constants $C > 0$ and $\sigma > 0$ and impose two simplifying assumptions. The first concerns boundedness of reward functions.

ASSUMPTION 1. *For all $f \in \mathcal{F}$ and $a \in \mathcal{A}$, $f(a) \in [0, C]$.*

Our second assumption ensures that observation noise is light tailed. Recall that we say a random variable x is σ-sub-Gaussian if $\mathbb{E}[\exp(\lambda x)] \le \exp(\lambda^2\sigma^2/2)$ almost surely for all λ.

ASSUMPTION 2. *For all $t \in \mathbb{N}$, $R_t - f_\theta(A_t)$ conditioned on (H_t, θ, A_t) is σ-sub-Gaussian.*

It is worth noting that the Bayesian regret bounds we provide are distribution independent, in the sense that we show $\text{BayesRegret}(T, \pi^{PS})$ is bounded by an expression that does not depend on $\mathbb{P}(\theta \in \cdot)$.

Our analysis in some ways parallels those found in the literature on UCB algorithms. In the next section, we provide a method for constructing a set $\mathcal{F}_t \subset \mathcal{F}$ of functions that are statistically plausible at time t. Let $w_{\mathcal{F}}(a) := \sup_{\bar{f} \in \mathcal{F}} \bar{f}(a) - \inf_{\underline{f} \in \mathcal{F}} \underline{f}(a)$ denote the width of $\mathcal{F}$ at a. Based on these confidence sets, and using the regret decomposition (3), one can bound Bayesian regret in terms of $\sum_1^T w_{\mathcal{F}_t}(A_t)$. In §7.2, we establish a bound on this sum in terms of the Kolmogorov and eluder dimensions of $\mathcal{F}$.

7.1. Confidence bounds. The construction of tight confidence sets for specific classes of functions presents technical challenges. Even for the relatively simple case of linear bandit problems, significant analysis is required. It is therefore perhaps surprising that, as we show in this section, one can construct strong confidence sets for an arbitrary class of functions without much additional sophistication. While the focus of our work is on providing a Bayesian regret bound for posterior sampling, the techniques we introduce for constructing confidence sets may find broader use.

The confidence sets constructed here are centered around least squares estimates $\hat{f}_t^{LS} \in \arg\min_{f \in \mathcal{F}} L_{2,t}(f)$, where $L_{2,t}(f) = \sum_1^{t-1} (f(A_t) - R_t)^2$ is the cumulative squared prediction error.[4] The sets take the form $\mathcal{F}_t :=$

[4] The results can be extended to the case where the infimum of $L_{2,t}(f)$ is unattainable by selecting a function with squared prediction error sufficiently close to the infimum.

$\{f \in \mathcal{F}\colon \|f - \hat{f}_t^{LS}\|_{2,E_t} \le \sqrt{\beta_t}\}$ where β_t is an appropriately chosen confidence parameter, and the empirical 2-norm $\|\cdot\|_{2,E_t}$ is defined by $\|g\|_{2,E_t}^2 = \sum_1^{t-1} g^2(A_k)$. Hence $\|f - f_\theta\|_{2,E_t}^2$ measures the cumulative discrepancy between the previous predictions of f and f_θ.

The following lemma is the key to constructing strong confidence sets $(\mathcal{F}_t\colon t \in \mathbb{N})$. For an arbitrary function f, it bounds the squared error of f from below in terms of the empirical loss of the true function f_θ and the aggregate empirical discrepancy $\|f - f_\theta\|_{2,E_t}^2$ between f and f_θ. It establishes that for any function f, with high probability, the random process $(L_{2,t}(f)\colon t \in \mathbb{N})$ never falls below the process $(L_{2,t}(f_\theta) + \frac{1}{2}\|f - f_\theta\|_{2,E_t}^2\colon t \in \mathbb{N})$ by more than a fixed constant. A proof of the lemma is provided in the appendix.

Lemma 3. *For any $\delta > 0$ and $f\colon \mathcal{A} \mapsto \mathbb{R}$, with probability at least $1 - \delta$,*

$$L_{2,t}(f) \ge L_{2,t}(f_\theta) + \tfrac{1}{2}\|f - f_\theta\|_{2,E_t}^2 - 4\sigma^2 \log(1/\delta)$$

simultaneously for all $t \in \mathbb{N}$.

By Lemma 3, with high probability, f can enjoy lower squared error than f_θ only if its empirical deviation $\|f - f_\theta\|_{2,E_t}^2$ from f_θ is less than $8\sigma^2 \log(1/\delta)$. Through a union bound, this property holds uniformly for all functions in a finite subset of $\mathcal{F}$. Using this fact and a discretization argument, together with the observation that $L_{2,t}(\hat{f}_t^{LS}) \le L_{2,t}(f_\theta)$, we can establish the following result, which is proved in the appendix. Let $N(\mathcal{F}, \alpha, \|\cdot\|_\infty)$ denote the α-covering number of $\mathcal{F}$ in the sup-norm $\|\cdot\|_\infty$, and let

$$\beta_t^*(\mathcal{F}, \delta, \alpha) := 8\sigma^2 \log(N(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha t(8C + \sqrt{8\sigma^2 \ln(4t^2/\delta)}). \tag{8}$$

Proposition 6. *For all $\delta > 0$ and $\alpha > 0$, if*

$$\mathcal{F}_t = \{f \in \mathcal{F}\colon \|f - \hat{f}_t^{LS}\|_{2,E_t} \le \sqrt{\beta_t^*(\mathcal{F}, \delta, \alpha)}\}$$

for all $t \in \mathbb{N}$, then

$$\mathbb{P}\left(f_\theta \in \bigcap_{t=1}^{\infty} \mathcal{F}_t\right) \ge 1 - 2\delta.$$

While the expression (8) defining the confidence parameter is complicated, it can be bounded by simple expressions in important cases. We provide three examples.

Example 4 (Finite Function Classes). When $\mathcal{F}$ is finite, $\beta_t^*(\mathcal{F}, \delta, 0) = 8\sigma^2 \log(|\mathcal{F}|/\delta)$.

Example 5 (Linear Models). Consider the case of a d-dimensional linear model $f_\rho(a) := \langle \phi(a), \rho \rangle$. Fix $\gamma = \sup_{a \in \mathcal{A}} \|\phi(a)\|$ and $s = \sup_{\rho \in \Theta} \|\rho\|$. Hence, for all $\rho_1, \rho_2 \in \mathcal{F}$, we have $\|f_{\rho_1} - f_{\rho_2}\|_\infty \le \gamma \|\rho_1 - \rho_2\|$. An α-covering of $\mathcal{F}$ can therefore be attained through an (α/γ)-covering of $\Theta \subset \mathbb{R}^d$. Such a covering requires $O((1/\alpha)^d)$ elements, and it follows that, $\log N(\mathcal{F}, \alpha, \|\cdot\|_\infty) = O(d \log(1/\alpha))$. If α is chosen to be $1/t^2$, the second term in (8) tends to zero, and therefore, $\beta_t^*(\mathcal{F}, \delta, 1/t^2) = O(d \log(t/\delta))$.

Example 6 (Generalized Linear Models). Consider the case of a d-dimensional generalized linear model $f_\theta(a) := g(\langle \phi(a), \theta \rangle)$, where g is an increasing Lipschitz continuous function. Fix g, $\gamma = \sup_{a \in \mathcal{A}} \|\phi(a)\|$ and $s = \sup_{\rho \in \Theta} \|\rho\|$. Then, the previous argument shows $\log N(\mathcal{F}, \alpha, \|\cdot\|_\infty) = O(d \log(1/\alpha))$. Again, choosing $\alpha = 1/t^2$ yields a confidence parameter $\beta_t^*(\mathcal{F}, \delta, 1/t^2) = O(d \log(t/\delta))$.

The confidence parameter $\beta_t^*(\mathcal{F}, 1/t^2, 1/t^2)$ is closely related to the following concept.

Definition 1. The *Kolmogorov dimension* of a function class $\mathcal{F}$ is given by

$$\dim_K(\mathcal{F}) = \limsup_{\alpha \downarrow 0} \frac{\log N(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\log(1/\alpha)}.$$

In particular, we have the following result.

Proposition 7. *For any fixed class of functions $\mathcal{F}$,*

$$\beta_t^*(\mathcal{F}, 1/t^2, 1/t^2) = 16(1 + o(1) + \dim_K(\mathcal{F})) \log t.$$

Proof. By definition,

$$\beta_t^*(\mathcal{F}, 1/t^2, 1/t^2) = 8\sigma^2 \left[\frac{\log(N(\mathcal{F}, 1/t^2, \|\cdot\|_\infty))}{\log(t^2)} + 1\right] \log(t^2) + 2\frac{t}{t^2}(8C + \sqrt{8\sigma^2 \ln(4t^2\delta)}).$$

The result follows from the fact that $\limsup_{t \to \infty} \log(N(\mathcal{F}, 1/t^2, \|\cdot\|_\infty))/\log(t^2) = \dim_K(\mathcal{F})$. □

7.2. Bayesian regret bounds. In this section, we introduce a new notion of complexity—the *eluder dimension*—and then use it to develop a Bayesian regret bound. First, we note that, using the regret decomposition (3) and the confidence sets $(\mathcal{F}_t\colon t\in\mathbb{N})$ constructed in the previous section, we can bound the Bayesian regret of posterior sampling in terms confidence interval widths $w_{\mathcal{F}}(a) := \sup_{f\in\mathcal{F}} f(a) - \inf_{f\in\mathcal{F}} f(a)$. In particular, the following lemma follows from our regret decomposition (3).

Lemma 4. *For all $T\in\mathbb{N}$, if $\inf_{\rho\in\mathcal{F}_\tau} f_\rho(a) \le f_\theta(a) \le \sup_{\rho\in\mathcal{F}_\tau} f_\rho(a)$ for all $\tau\in\mathbb{N}$ and $a\in\mathcal{A}$ with probability at least $1-1/T$, then*

$$\mathrm{BayesRegret}(T,\pi^{PS}) \le C + \mathbb{E}\sum_{t=1}^{T} w_{\mathcal{F}_t}(A_t).$$

We can use the confidence sets constructed in the previous section to guarantee that the conditions of this lemma hold. In particular, choosing $\delta = 1/2T$ in (8) guarantees that $f_\theta \in \bigcap_{t=1}^{\infty}\mathcal{F}_t$ with probability at least $1-1/T$.

Our remaining task is to provide a worst-case bound on the sum $\sum_1^T w_{\mathcal{F}_t}(A_t)$. First, consider the case of a linearly parameterized model where $f_\rho(a) := \langle\phi(a),\rho\rangle$ for each $\rho\in\Theta\subset\mathbb{R}^d$. Then, it can be shown that our confidence set takes the form $\mathcal{F}_t := \{f_\rho\colon \rho\in\Theta_t\}$, where $\Theta_t\subset\mathbb{R}^d$ is an ellipsoid. When an action A_t is sampled, the ellipsoid shrinks in the direction $\phi(A_t)$. Here, the explicit geometric structure of the confidence set implies that the width $w_{\mathcal{F}_t}$ shrinks not only at A_t but also at any other action whose feature vector is not orthogonal to $\phi(A_t)$. Some linear algebra leads to a worst-case bound on $\sum_1^T w_{\mathcal{F}_t}(A_t)$. For a general class of functions, the situation is much subtler, and we need to measure the way in which the width at each action can be reduced by sampling other actions. To do this, we introduce the following notion of dependence.

Definition 2. An action $a\in\mathcal{A}$ is *ϵ-dependent* on actions $\{a_1,\dots,a_n\}\subseteq\mathcal{A}$ with respect to $\mathcal{F}$ if any pair of functions $f,\tilde f\in\mathcal{F}$, satisfying $\sqrt{\sum_{i=1}^n (f(a_i)-\tilde f(a_i))^2}\le\epsilon$ also satisfies $f(a)-\tilde f(a)\le\epsilon$. Further, a is ϵ-independent of $\{a_1,\dots,a_n\}$ with respect to $\mathcal{F}$ if a is not ϵ-dependent on $\{a_1,\dots,a_n\}$.

Intuitively, an action a is independent of $\{a_1,\dots,a_n\}$ if two functions that make similar predictions at $\{a_1,\dots,a_n\}$ can nevertheless differ significantly in their predictions at a. The above definition measures the "similarity" of predictions at ϵ-scale, and measures whether two functions make similar predictions at $\{a_1,\dots,a_n\}$ based on the cumulative discrepancy $\sqrt{\sum_{i=1}^n (f(a_i)-\tilde f(a_i))^2}$. This measure of dependence suggests using the following notion of dimension. In this definition, we imagine that the sequence of elements in $\mathcal{A}$ is chosen by an *eluder* that hopes to show the agent poorly understood actions for as long as possible.

Definition 3. The *ϵ-eluder dimension* $\dim_E(\mathcal{F},\epsilon)$ is the length d of the longest sequence of elements in $\mathcal{A}$ such that, for some $\epsilon'\ge\epsilon$, every element is ϵ'-independent of its predecessors.

Recall that a vector space has dimension d if and only if d is the length of the longest sequence of elements such that each element is linearly independent or equivalently, 0-independent of its predecessors. Definition 3 replaces the requirement of linear independence with ϵ-independence. This extension is advantageous as it captures both nonlinear dependence and approximate dependence. The following result uses our new notion of dimension to bound the number of times the width of the confidence interval for a selected action A_t can exceed a threshold.

Proposition 8. *If $(\beta_t\ge 0 \mid t\in\mathbb{N})$ is a nondecreasing sequence and $\mathcal{F}_t := \{f\in\mathcal{F}\colon \|f-\hat f_t^{LS}\|_{2,E_t}\le\sqrt{\beta_t}\}$, then*

$$\sum_{t=1}^{T}\mathbf{1}(w_{\mathcal{F}_t}(A_t)>\epsilon) \le \left(\frac{4\beta_T}{\epsilon^2}+1\right)\dim_E(\mathcal{F},\epsilon)$$

for all $T\in\mathbb{N}$ and $\epsilon>0$.

Proof. We begin by showing that if $w_t(A_t)>\epsilon$, then A_t is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(A_1,\dots,A_{t-1})$ for $T>t$. To see this, note that if $w_{\mathcal{F}_t}(A_t)>\epsilon$, there are $\bar f,\underline f\in\mathcal{F}_t$ such that $\bar f(A_t)-\underline f(A_t)>\epsilon$. By definition, since $\bar f(A_t)-\underline f(A_t)>\epsilon$ if A_t is ϵ-dependent on a subsequence $(A_{i_1},\dots,A_{i_k})$ of $(A_1,\dots,A_{t-1})$, then $\sum_{j=1}^k(\bar f(A_{i_j})-\underline f(A_{i_j}))^2>\epsilon^2$. It follows that, if A_t is ϵ-dependent on K disjoint subsequences of $(A_1,\dots,A_{t-1})$, then $\|\bar f-\underline f\|_{2,E_t}^2>K\epsilon^2$. By the triangle inequality, we have

$$\|\bar f-\underline f\|_{2,E_t}\le\|\bar f-\hat f_t^{LS}\|_{2,E_t}+\|\underline f-\hat f_t^{LS}\|_{2,E_t}\le 2\sqrt{\beta_t}\le 2\sqrt{\beta_T}$$

and it follows that $K<4\beta_T/\epsilon^2$.

Next, we show that, in any action sequence $(a_1,\dots,a_\tau)$, there is some element a_j that is ϵ-dependent on at least $\tau/d-1$ disjoint subsequences of $(a_1,\dots,a_{j-1})$, where $d:=\dim_E(\mathcal{F},\epsilon)$. To show this, for an integer K

satisfying $Kd+1 \le \tau \le Kd+d$, we will construct K disjoint subsequences $B_1, \ldots, B_K$. First let $B_i = (a_i)$ for $i = 1, \ldots, K$. If a_{K+1} is ϵ-dependent on each subsequence $B_1, \ldots, B_K$, our claim is established. Otherwise, select a subsequence B_i such that a_{K+1} is ϵ-independent and append a_{K+1} to B_i. Repeat this process for elements with indices $j > K+1$ until a_j is ϵ-dependent on each subsequence or $j = \tau$. In the latter scenario $\sum |B_i| \ge Kd$, and since each element of a subsequence B_i is ϵ-independent of its predecessors, $|B_i| = d$. In this case, a_τ must be ϵ-dependent on each subsequence.

Now, consider taking $(a_1, \ldots, a_\tau)$ to be the subsequence $(A_{t_1}, \ldots, A_{t_\tau})$ of $(A_1, \ldots, A_T)$ consisting of elements A_t for which $w_{\mathcal{F}_t}(A_t) > \epsilon$. As we have established, each A_{t_j} is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(A_1, \ldots, A_{t_j-1})$. It follows that each a_j is ϵ-dependent on fewer than $4\beta_T/\epsilon^2$ disjoint subsequences of $(a_1, \ldots, a_{j-1})$. Combining this with the fact that we have established that there is some a_j that is ϵ-dependent on at least $\tau/d - 1$ disjoint subsequences of $(a_1, \ldots, a_{j-1})$, we have $\tau/d - 1 \le 4\beta_T/\epsilon^2$. It follows that $\tau \le (4\beta_T/\epsilon^2 + 1)d$, which is our desired result. □

Using Proposition 8, one can bound the sum $\sum_{t=1}^T w_{\mathcal{F}_t}(A_t)$, as established by the following lemma.

Lemma 5. *If $(\beta_t \ge 0 \mid t \in \mathbb{N})$ is a nondecreasing sequence and $\mathcal{F}_t := \{f \in \mathcal{F}\colon \|f - \hat{f}_t^{LS}\|_{2,E_t} \le \sqrt{\beta_t}\}$, then*

$$\sum_{t=1}^T w_{\mathcal{F}_t}(A_t) \le 1 + \dim_E(\mathcal{F}, T^{-1})C + 4\sqrt{\dim_E(\mathcal{F}, T^{-1})\beta_T T}$$

for all $T \in \mathbb{N}$.

Proof. To reduce notation, write $d = \dim_E(\mathcal{F}, T^{-1})$ and $w_t = w_t(A_t)$. Reorder the sequence $(w_1, \ldots, w_T) \to (w_{i_1}, \ldots, w_{i_T})$, where $w_{i_1} \ge w_{i_2} \ge \cdots \ge w_{i_T}$. We have

$$\sum_{t=1}^T w_{\mathcal{F}_t}(A_t) = \sum_{t=1}^T w_{i_t} = \sum_{t=1}^T w_{i_t}\mathbf{1}\{w_{i_t} \le T^{-1}\} + \sum_{t=1}^T w_{i_t}\mathbf{1}\{w_{i_t} > T^{-1}\} \le 1 + \sum_{t=1}^T w_{i_t}\mathbf{1}\{w_{i_t} \ge T^{-1}\}.$$

We know $w_{i_t} \le C$. In addition, $w_{i_t} > \epsilon \iff \sum_{k=1}^T \mathbf{1}(w_{\mathcal{F}_k}(A_k) > \epsilon) \ge t$. By Proposition 8, this can only occur if $t < ((4\beta_T)/\epsilon^2 + 1)\dim_E(\mathcal{F}, \epsilon)$. For $\epsilon \ge T^{-1}$, $\dim_E(\mathcal{F}, \epsilon) \le \dim_E(\mathcal{F}, T^{-1}) = d$, since $\dim_E(\mathcal{F}, \epsilon')$ is nonincreasing in ϵ'. Therefore, when $w_{i_t} > \epsilon \ge T^{-1}$, $t < ((4\beta_T)/\epsilon^2 + 1)d$, which implies $\epsilon < \sqrt{(4\beta_T d)/(t-d)}$. This shows that if $w_{i_t} > T^{-1}$, then $w_{i_t} \le \min\{C, \sqrt{(4\beta_T d)/(t-d)}\}$. Therefore

$$\sum_{t=1}^T w_{i_t}\mathbf{1}\{w_{i_t} > T^{-1}\} \le dC + \sum_{t=d+1}^T \sqrt{\frac{4d\beta_T}{t-d}} \le dC + 2\sqrt{d\beta_T}\int_{t=0}^T \frac{1}{\sqrt{t}}\,dt = dC + 4\sqrt{d\beta_T T}. \quad \square$$

Our next result, which follows from Lemma 4, Lemma 5, and Proposition 6, establishes a Bayesian regret bound.

Proposition 9. *For all $T \in \mathbb{N}$, $\alpha > 0$ and $\delta \le 1/2T$,*

$$\text{BayesRegret}(T, \pi^{PS}) \le 1 + [\dim_E(\mathcal{F}, T^{-1}) + 1]C + 4\sqrt{\dim_E(\mathcal{F}, T^{-1})\beta_T^*(\mathcal{F}, \alpha, \delta)T}.$$

Using bounds on β_t^* provided in the previous section together with Proposition 9 yields Bayesian regret bounds that depend on $\mathcal{F}$ only through the eluder dimension and either the cardinality or Kolmogorov dimension. The following proposition provides such bounds.

Proposition 10. *For any fixed class of functions $\mathcal{F}$,*

$$\text{BayesRegret}(T, \pi^{PS}) \le 1 + [\dim_E(\mathcal{F}, T^{-1}) + 1]C + 16\sigma\sqrt{\dim_E(\mathcal{F}, T^{-1})(1 + o(1) + \dim_K(\mathcal{F}))\log(T)T}$$

for all $T \in \mathbb{N}$. Further, if $\mathcal{F}$ is finite, then

$$\text{BayesRegret}(T, \pi^{PS}) \le 1 + [\dim_E(\mathcal{F}, T^{-1}) + 1]C + 8\sigma\sqrt{2\dim_E(\mathcal{F}, T^{-1})\log(2|\mathcal{F}|T)T}$$

for all $T \in \mathbb{N}$.

The next two examples show how the first Bayesian regret bound of Proposition 10 specializes to d-dimensional linear and generalized linear models. For each of these examples, a bound on $\dim_E(\mathcal{F}, \epsilon)$ is provided in the appendix.

Example 7 (Linear Models). Consider the case of a d-dimensional linear model $f_\rho(a) := \langle \phi(a), \rho \rangle$. Fix $\gamma = \sup_{a \in \mathcal{A}} \|\phi(a)\|$ and $s = \sup_{\rho \in \Theta} \|\rho\|$. Then, $\dim_E(\mathcal{F}, \epsilon) = O(d \log(1/\epsilon))$ and $\dim_K(\mathcal{F}) = O(d)$. Proposition 10 therefore yields an $O(d\sqrt{T}\log(T))$ Bayesian regret bound. This is tight to within a factor of $\log T$ (Rusmevichientong and Tsitsiklis [31]), and matches the best available bound for a linear UCB algorithm (Abbasi-Yadkori et al. [2]).

Example 8 (Generalized Linear Models). Consider the case of a d-dimensional generalized linear model $f_\theta(a) := g(\langle \phi(a), \theta \rangle)$, where g is an increasing Lipschitz continuous function. Fix $\gamma = \sup_{a \in \mathcal{A}} \|\phi(a)\|$ and $s = \sup_{\rho \in \Theta} \|\rho\|$. Then, $\dim_K(\mathcal{F}) = O(d)$ and $\dim_E(\mathcal{F}, \epsilon) = O(r^2 d \log(r\epsilon))$, where $r = \sup_{\tilde{\theta}, a} g'(\langle \phi(a), \tilde{\theta} \rangle) / \inf_{\tilde{\theta}, a} g'(\langle \phi(a), \tilde{\theta} \rangle)$ bounds the ratio between the maximal and minimal slope of g. Proposition 10 yields an $O(rd\sqrt{T}\log(rT))$ Bayesian regret bound. We know of no other guarantee for posterior sampling when applied to generalized linear models. In fact, to our knowledge, this bound is a slight improvement over the strongest Bayesian regret bound available for any algorithm in this setting. The regret bound of Filippi et al. [18] translates to an $O(rd\sqrt{T}\log^{3/2}(T))$ Bayesian regret bound.

One advantage of studying posterior sampling in a general framework is that it allows bounds to be obtained for specific classes of models by specializing more general results. This advantage is highlighted by the ease of developing a performance guarantee for generalized linear models. The problem is reduced to one of bounding the eluder dimension, and such a bound follows almost immediately from the analysis of linear models. In prior literature, extending results from linear to generalized linear models required significant technical developments, as presented in Filippi et al. [18].

7.3. Relation to the VC dimension. To close our section on general bounds, we discuss important differences between our new notion of eluder dimension and complexity measures used in the analysis of supervised learning problems. We begin with an example that illustrates how a class of functions that is learnable in constant time in a supervised learning context may require an arbitrarily long duration when learning to optimize.

Example 9. Consider a finite class of binary-valued functions $\mathcal{F} = \{f_\rho : \mathcal{A} \mapsto \{0, 1\} \mid \rho \in \{1, \ldots, n\}\}$ over a finite action set $\mathcal{A} = \{1, \ldots, n\}$. Let $f_\rho(a) = \mathbf{1}(\rho = a)$, so that each function is an indicator for an action. To keep things simple, assume that $R_t = f_\theta(A_t)$, so that there is no noise. If θ is uniformly distributed over $\{1, \ldots, n\}$, it is easy to see that the Bayesian regret of posterior sampling grows linearly with n. For large n, until θ is discovered, each sampled action is unlikely to reveal much about θ and learning therefore takes very long.

Consider the closely related supervised learning problem in which at each time step, an action $\tilde{A}_t$ is sampled uniformly from $\mathcal{A}$ and the mean reward value $f_\theta(\tilde{A}_t)$ is observed. For large n, the time it takes to effectively learn to predict $f_\theta(\tilde{A}_t)$, given $\tilde{A}_t$ does not depend on t. In particular, prediction error converges to $1/n$ in constant time. Note that predicting 0 at every time already achieves this low level of error.

In the preceding example, the ϵ-eluder dimension is n for $\epsilon \in (0, 1)$. On the other hand, the VC dimension, which characterizes the sample complexity of supervised learning, is 1. To highlight conceptual differences between the eluder dimension and the VC dimension, we will now define VC dimension in a way analogous to how we defined eluder dimension. We begin with a notion of independence.

Definition 4. An action a is *VC-independent* of $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ if for any $f, \tilde{f} \in \mathcal{F}$, there exists some $\bar{f} \in \mathcal{F}$, which agrees with f on a and with $\tilde{f}$ on $\tilde{\mathcal{A}}$; that is, $\bar{f}(a) = f(a)$ and $\bar{f}(\tilde{a}) = \tilde{f}(\tilde{a})$ for all $\tilde{a} \in \tilde{\mathcal{A}}$. Otherwise, a is *VC-dependent* on $\tilde{\mathcal{A}}$.

By this definition, an action a is said to be VC-dependent on $\tilde{\mathcal{A}}$ if knowing the values $f \in \mathcal{F}$ takes on $\tilde{\mathcal{A}}$ *could restrict* the set of possible values at a. This notion of independence is intimately related to the VC dimension of a class of functions. In fact, it can be used to define VC dimension.

Definition 5. The VC dimension of a class of binary-valued functions with domain $\mathcal{A}$ is the largest cardinality of a set $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ such that every $a \in \tilde{\mathcal{A}}$ is VC-independent of $\tilde{\mathcal{A}} \backslash \{a\}$.

In the above example, any two actions are VC dependent because knowing the label of one action could completely determine the value of the other action. However, this only happens if the sampled action has label 1. If it has label 0, one cannot infer anything about the value of the other action. Instead of capturing the fact that one *could* gain useful information about the reward function through exploration, we need a stronger requirement that guarantees one *will* gain useful information through exploration. Such a requirement is captured by the following concept.

DEFINITION 6. An action a is *strongly dependent* on a set of actions $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ if any two functions $f, \tilde{f} \in \mathcal{F}$ that agree on $\tilde{\mathcal{A}}$ agree on a; that is, the set $\{f(a) : f(\tilde{a}) = \tilde{f}(\tilde{a}) \ \forall \tilde{a} \in \tilde{\mathcal{A}}\}$ is a singleton. An action a is *weakly independent* of $\tilde{\mathcal{A}}$ if it is not strongly dependent on $\tilde{\mathcal{A}}$.

According to this definition, a is strongly dependent on $\tilde{\mathcal{A}}$ if knowing the values of f on $\tilde{\mathcal{A}}$ completely determines the value of f on a. While the above definition is conceptually useful, for our purposes, it is important to capture approximate dependence between actions. Our definition of eluder dimension achieves this goal by focusing on the possible difference $f(a) - \tilde{f}(a)$ between two functions that approximately agree on $\tilde{\mathcal{A}}$.

8. Simulation results. In this section, we compare the performance in simulation of posterior sampling to that of UCB algorithms that have been proposed in the recent literature. Our results demonstrate that posterior sampling significantly outperforms these algorithms. Moreover, we identify a clear cause for the large discrepancy: confidence sets proposed in the literature are too loose to attain good performance.

We consider the linear model $f_\theta(a) = \langle \phi(a), \theta \rangle$, where $\theta \in \mathbb{R}^{10}$ follows a multivariate Gaussian distribution with mean vector $\mu = 0$ and covariance matrix $\Sigma = 10I$. The noise terms $\epsilon_t := R_t - f_\theta(A_t)$ follow a standard Gaussian distribution. There are 100 actions with feature vector components drawn uniformly at random from $[-1/\sqrt{10}, 1/\sqrt{10}]$, and $\mathcal{A}_t = \mathcal{A}$ for each t. Figure 1 shows the portion $\langle \phi(A_t^*), \theta \rangle - \langle \phi(A_t), \theta \rangle$ of regret attributable to each time period t in the first 1,000 time periods. The results are averaged across 5,000 trials.

Several UCB algoriths are suitable for such problems, including those of Abbasi-Yadkori et al. [2], Rusmevichientong and Tsitsiklis [31], and Srinivas et al. [35]. While the confidence bound of Rusmevichientong and Tsitsiklis [31] is stronger than that of Dani et al. [16], it is still too loose and the resulting linear UCB algorithm hardly improves its performance over the 1,000-period time horizon. We display the results only of the more competitive UCB algorithms. The line labeled "linear UCB" displays the results of the algorithm proposed in Abbasi-Yadkori et al. [2], which incurred average regret of 339.7. The algorithm of Srinivas et al. [35] is labeled "Gaussian UCB," and incurred average regret 198.7. Posterior sampling, on the other hand, incurred average regret of only 97.5.

Each of these UCB algorithms uses a confidence bound that was derived through stochastic analysis. The Gaussian linear model has a clear structure, however, which suggests UCBs should take the form $U_t(a) = \mu_{t-1}(a) + \sqrt{\beta}\sigma_{t-1}(a)$, where $\mu_{t-1}(a)$ and $\sigma_{t-1}(a)$ are the posterior mean and standard deviation at a. The final algorithm we consider ignores theoretical considerations, and tunes the parameter β to minimize the average regret over the first 1,000 periods. The average regret of the algorithm was only 68.9, a dramatic improvement over (Abbasi-Yadkori et al. [2], Srinivas et al. [35]), and even outperforming posterior sampling. On the plot shown below, these results are labeled "Gaussian UCB—Tuned Heuristic." Note such tuning requires the time horizon to be fixed and known.

In this setting, the problem of choosing UCBs reduces to choosing a single confidence parameter β. For more complicated problems, however, significant analysis may be required to choose a structural form for confidence sets. The results in this section suggest that it can be quite challenging to use such analysis to derive confidence bounds

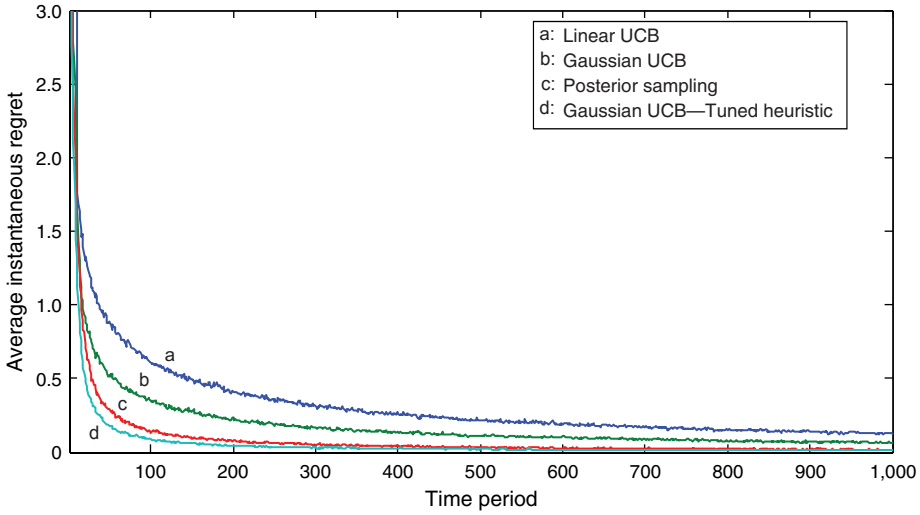


FIGURE 1. Portion of regret attributable to each time period.

that lead to strong empirical performance. In particular, this is challenging even for linear models. For example, the paper (Abbasi-Yadkori et al. [2]) uses sophisticated tools from the study of multivariate self-normalized martingales to derive a confidence bound that is stronger than those of Dani et al. [16] or Rusmevichientong and Tsitsiklis [31], but their algorithm still incurs about three and a half times the regret of posterior sampling. This highlights a crucial advantage of posterior sampling that we have emphasized throughout this paper; it effectively separates confidence bound *analysis* from algorithm *design*.

Finally, it should be noted that the algorithms of Abbasi-Yadkori et al. [2], and Srinivas et al. [35] have free parameters that must be chosen by the user. We have attempted to set these values in a way that minimizes average regret over the 1,000-period time horizon. Both algorithms construct confidence bounds that hold with a prespecified probability $1-\delta \in [0,1]$. Higher levels of δ lead to lower UCBs, which we find improves performance. We set $\delta = 1$ to minimize the average regret of the algorithms. The algorithm of Abbasi-Yadkori et al. [2] requires two other choices. We used a line search to set the algorithm's regularization parameter to the level $\lambda = 0.025$, which minimizes cumulative regret. The algorithm of Abbasi-Yadkori et al. [2] also requires a uniform upper bound on $\|\theta\|$, but the Gaussian distribution is unbounded. We avoid this issue by providing the actual realized value $\|\theta\|$ as an input to algorithm.

9. Conclusion. This paper has considered the use of a simple posterior sampling algorithm for learning to optimize actions when the decision maker is uncertain about how his actions influence performance. We believe that, particularly for difficult problem instances, this algorithm offers significant potential advantages because of its design simplicity and computational tractability. Despite its great potential, not much is known about posterior sampling when there are dependencies between actions. Our work has taken a significant step toward remedying this gap. We showed that the Bayesian regret of posterior sampling can be decomposed in terms of confidence sets, which allowed us to establish a number of new results on posterior sampling by leveraging prior work on UCB algorithms. We then used this regret decomposition to analyze posterior sampling in a very general framework, and developed Bayesian regret bounds that depend on a new notion of dimension.

In constructing these bounds, we have identified two factors that control the hardness of a particular MAB. First, an agent's ability to quickly attain near-optimal performance depends on the extent to which the reward value at one action can be inferred by sampling other actions. However, to select an action, the agent must make inferences about many possible actions, and an error in its evaluation of any one could result in large regret. Our second measure of complexity controls for the difficulty of maintaining appropriate confidence sets simultaneously at every action. While our bounds are nearly tight in some cases, further analysis is likely to yield stronger results in other cases. We hope, however, that our work provides a conceptual foundation for the study of such problems, and inspires further investigation.

Acknowledgments. Daniel Russo is supported by a Burt and Deedee McMurty Stanford Graduate Fellowship. This work was supported, in part, by the National Science Foundation Award [CMMI-0968707]

.

Appendix A. Details regarding Lemma 1. Lemma 1 follows as a special case of Theorem 1 of Abbasi-Yadkori et al. [2], which is much more general. Note that because reward noise $R_t - f_\theta(A_t)$ is bounded in $[-1,1]$, it is 1-sub-Gaussian. Equation (12) in Abbasi-Yadkori et al. [2] gives a specialization of Theorem 1 to the problem we consider. It states that, for any $\delta > 0$ with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{t} \mathbf{1}_{\{A_k=a\}}(R_k - f_\theta(a))\right| \le \sqrt{(1+N_t(a))\left(1+2\log\left(\frac{(1+N_t(a))^{1/2}}{\delta}\right)\right)} \quad \forall t \in \mathbb{N}.$$

We choose $\delta = 1/T$ and use that for $t \le T$, $N_t(a) \le T-1$ to show that with probability at least $1/T$,

$$\left|\sum_{k=1}^{t} \mathbf{1}_{\{A_k=a\}}(R_k - f_\theta(a))\right| \le \sqrt{(1+N_t(a))(1+3\log(T))} \quad \forall t \in \{1,\dots,T\}.$$

Since $1+N_t(a) \le 2N_t(a)$ whenever a has been played at least once, with probability at least $1/T$,

$$\left|\sum_{k=1}^{t} \mathbf{1}_{\{A_k=a\}}(R_k - f_\theta(a))\right| \le \sqrt{(N_t(a))(2+6\log(T))} \quad \forall t \in \{1,\dots,T\}.$$

Appendix B. Proof of confidence bound.

B.1. Preliminaries: Martingale exponential inequalities. Consider random variables $(Z_n \mid n \in \mathbb{N})$ adapted to the filtration $(\mathcal{H}_n\colon n = 0, 1, \dots)$. Assume $\mathbb{E}[\exp\{\lambda Z_i\}]$ is finite for all λ. Define the conditional mean $\mu_i = \mathbb{E}[Z_i \mid \mathcal{H}_{i-1}]$. We define the conditional cumulant generating function of the centered random variable $[Z_i - \mu_i]$ by $\psi_i(\lambda) = \log \mathbb{E}[\exp(\lambda[Z_i - \mu_i]) \mid \mathcal{H}_{i-1}]$. Let

$$M_n(\lambda) = \exp\left\{\sum_{i=1}^{n} \lambda[Z_i - \mu_i] - \psi_i(\lambda)\right\}.$$

Lemma 6. *$(M_n(\lambda) \mid n \in \mathbb{N})$ is a martingale, and $\mathbb{E}\, M_n(\lambda) = 1$.*

Proof. By definition,

$$\mathbb{E}[M_1(\lambda) \mid \mathcal{H}_0] = \mathbb{E}[\exp\{\lambda[Z_1 - \mu_1] - \psi_1(\lambda)\} \mid \mathcal{H}_0] = \mathbb{E}[\exp\{\lambda[Z_1 - \mu_1]\} \mid \mathcal{H}_0]/\exp\{\psi_1(\lambda)\} = 1.$$

Then, for any $n \geq 2$,

$$\begin{aligned}
\mathbb{E}[M_n(\lambda) \mid \mathcal{H}_{n-1}] &= \mathbb{E}\left[\exp\left\{\sum_{i=1}^{n-1} \lambda[Z_i - \mu_i] - \psi_i(\lambda)\right\} \exp\{\lambda[Z_n - \mu_n] - \psi_n(\lambda)\} \,\middle|\, \mathcal{H}_{n-1}\right] \\
&= \exp\left\{\sum_{i=1}^{n-1} \lambda[Z_i - \mu_i] - \psi_i(\lambda)\right\} \mathbb{E}[\exp\{\lambda[Z_n - \mu_n] - \psi_n(\lambda)\} \mid \mathcal{H}_{n-1}] \\
&= \exp\left\{\sum_{i=1}^{n-1} \lambda[Z_i - \mu_i] - \psi_i(\lambda)\right\} = M_{n-1}(\lambda). \quad \square
\end{aligned}$$

Lemma 7. *For all $x \geq 0$ and $\lambda \geq 0$, $\mathbb{P}(\sum_1^n \lambda Z_i \leq x + \sum_1^n [\lambda \mu_i + \psi_i(\lambda)]\ \forall n \in \mathbb{N}) \geq 1 - e^{-x}$.*

Proof. For any λ, $M_n(\lambda)$ is a martingale with $\mathbb{E}\, M_n(\lambda) = 1$. Therefore, for any stopping time τ, $\mathbb{E}\, M_{\tau \wedge n}(\lambda) = 1$. For arbitrary $x \geq 0$, define $\tau_x = \inf\{n \geq 0 \mid M_n(\lambda) \geq x\}$ and note that τ_x is a stopping time corresponding to the first time M_n crosses the boundary at x. Then $\mathbb{E}\, M_{\tau_x \wedge n}(\lambda) = 1$ and by Markov's inequality,

$$x\, \mathbb{P}(M_{\tau_x \wedge n}(\lambda) \geq x) \leq \mathbb{E}\, M_{\tau_x \wedge n}(\lambda) = 1.$$

We note that the event $\{M_{\tau_x \wedge n}(\lambda) \geq x\} = \bigcup_{k=1}^{n} \{M_k(\lambda) \geq x\}$. So we have shown that for all $x \geq 0$ and $n \geq 1$,

$$\mathbb{P}\left(\bigcup_{k=1}^{n} \{M_k(\lambda) \geq x\}\right) \leq \frac{1}{x}.$$

Taking the limit as $n \to \infty$, and applying the monotone convergence theorem shows $\mathbb{P}(\bigcup_{k=1}^{\infty} \{M_k(\lambda) \geq x\}) \leq 1/x$, or $\mathbb{P}(\bigcup_{k=1}^{\infty} \{M_k(\lambda) \geq e^x\}) \leq e^{-x}$. This then shows, using the definition of $M_k(\lambda)$, that

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \left\{\sum_{i=1}^{n} \lambda[Z_i - \mu_i] - \psi_i(\lambda) \geq x\right\}\right) \leq e^{-x}. \quad \square$$

B.2. Proof of Lemma 3.

Lemma 3. *$a = a$. For any $\delta > 0$ and $f\colon \mathcal{A} \mapsto \mathbb{R}$, with probability at least $1 - \delta$,*

$$L_{2,t}(f) \geq L_{2,t}(f_\theta) + \frac{1}{2}\|f - f_\theta\|_{2,E_t}^2 - 4\sigma^2 \log(1/\delta)$$

simultaneously for all $t \in \mathbb{N}$.

We will transform our problem to apply the general exponential martingale result shown above. We set $\mathcal{H}_{t-1}$ to be the σ-algebra generated by (H_t, A_t, θ). By previous assumptions, $\epsilon_t := R_t - f_\theta(A_t)$ satisfies $\mathbb{E}[\epsilon_t \mid \mathcal{H}_{t-1}] = 0$, and $\mathbb{E}[\exp\{\lambda \epsilon_t\} \mid \mathcal{H}_{t-1}] \leq \exp\{(\lambda^2 \sigma^2)/2\}$ a.s. for all λ. Define $Z_t = (f_\theta(A_t) - R_t)^2 - (f(A_t) - R_t)^2$.

Proof. By definition, $\sum_1^T Z_t = L_{2,T+1}(f_\theta) - L_{2,T+1}(f)$. Some calculation shows that $Z_t = -(f(A_t) - f_\theta(A_t))^2 + 2(f(A_t) - f_\theta(A_t))\epsilon_t$. Therefore the conditional mean and conditional cumulant generating function satisfy,

$$\begin{aligned}
\mu_t &= \mathbb{E}[Z_t \mid \mathcal{H}_{t-1}] = -(f(A_t) - f_\theta(A_t))^2 \\
\psi_t(\lambda) &= \log \mathbb{E}[\exp(\lambda[Z_t - \mu_t]) \mid \mathcal{H}_{t-1}] \\
&= \log \mathbb{E}[\exp(2\lambda(f(A_t) - f_\theta(A_t))\epsilon_t) \mid \mathcal{H}_{t-1}] \leq \frac{(2\lambda[f(A_t) - f_\theta(A_t)])^2 \sigma^2}{2}.
\end{aligned}$$

Applying Lemma 7 shows that, for all $x \geq 0$, $\lambda \geq 0$

$$\mathbb{P}\left(\sum_{k=1}^{t} \lambda Z_k \leq x - \lambda \sum_{k=1}^{t} (f(A_k) - f_\theta(A_k))^2 + \frac{\lambda^2}{2}(2f(A_k) - 2f_\theta(A_k))^2\sigma^2 \ \forall t \in \mathbb{N}\right) \geq 1 - e^{-x}.$$

Or rearranging terms

$$\mathbb{P}\left(\sum_{k=1}^{t} Z_k \leq \frac{x}{\lambda} + \sum_{k=1}^{t} (f(A_k) - f_\theta(A_k))^2(2\lambda\sigma^2 - 1) \ \forall t \in \mathbb{N}\right) \geq 1 - e^{-x}.$$

Choosing $\lambda = 1/(4\sigma^2)$, $x = \log(1/\delta)$, and using the definition of $\sum_1^t Z_k$ implies

$$\mathbb{P}\big(L_{2,t}(f) \geq L_{2,t}(f_\theta) + \tfrac{1}{2}\|f - f_\theta\|_{2,E_t}^2 - 4\sigma^2 \log(1/\delta) \ \forall t \in \mathbb{N}\big) \geq 1 - \delta. \quad \square$$

B.3. Least squares bound—Proof of Proposition 6.

Proposition 6. *For all $\delta > 0$ and $\alpha > 0$, if $\mathcal{F}_t = \{f \in \mathcal{F}\colon \|f - \hat{f}_t^{LS}\|_{2,E_t} \leq \sqrt{\beta_t^*(\mathcal{F}, \delta, \alpha)}\}$ for all $t \in \mathbb{N}$, then*

$$\mathbb{P}\left(f_\theta \in \bigcap_{t=1}^{\infty} \mathcal{F}_t\right) \geq 1 - 2\delta.$$

Proof. Let $\mathcal{F}^\alpha \subset \mathcal{F}$ be an α-cover of $\mathcal{F}$ in the sup norm in the sense that, for any $f \in \mathcal{F}$, there is an $f^\alpha \in \mathcal{F}^\alpha$ such that $\|f^\alpha - f\|_\infty \leq \epsilon$. By a union bound, with probability at least $1 - \delta$,

$$L_{2,t}(f^\alpha) - L_{2,t}(f_\theta) \geq \tfrac{1}{2}\|f^\alpha - f_\theta\|_{2,E_t} - 4\sigma^2 \log(|\mathcal{F}^\alpha|/\delta) \quad \forall t \in \mathbb{N}, \quad f \in \mathcal{F}^\alpha.$$

Therefore, with probability at least $1 - \delta$ for all $t \in \mathbb{N}$ and $f \in \mathcal{F}$,

$$L_{2,t}(f) - L_{2,t}(f_\theta) \geq \tfrac{1}{2}\|f - f_\theta\|_{2,E_t}^2 - 4\sigma^2 \log(|\mathcal{F}^\alpha|/\delta) + \underbrace{\min_{f^\alpha \in \mathcal{F}^\alpha} \left\{\tfrac{1}{2}\|f^\alpha - f_\theta\|_{2,E_t}^2 - \tfrac{1}{2}\|f - f_\theta\|_{2,E_t}^2 + L_{2,t}(f) - L_{2,t}(f^\alpha)\right\}}_{\text{Discretization error}}.$$

Lemma 8, which we establish in the next section, asserts that with probability at least $1 - \delta$, the discretization error is bounded for all t by $\alpha\eta_t$, where $\eta_t := t[8C + \sqrt{8\sigma^2 \ln(4t^2/\delta)}]$. Since the least squares estimate $\hat{f}_t^{LS}$ has lower squared error than f_θ by definition, we find with probability at least $1 - 2\delta$

$$\tfrac{1}{2}\|\hat{f}_t^{LS} - f_\theta\|_{2,E_t}^2 \leq 4\sigma^2 \log(|\mathcal{F}^\alpha|/\delta) + \alpha\eta_t.$$

Taking the infimum over the size of α covers implies

$$\|\hat{f}_t^{LS} - f_\theta\|_{2,E_t} \leq \sqrt{8\sigma^2 \log(N(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha\eta_t} \stackrel{\text{def}}{=} \sqrt{\beta_t^*(\mathcal{F}, \delta, \alpha)}. \quad \square$$

B.4. Discretization error.

Lemma 8. *If f^α satisfies $\|f - f^\alpha\|_\infty \leq \alpha$, then with probability at least $1 - \delta$,*

$$\left|\tfrac{1}{2}\|f^\alpha - f_\theta\|_{2,E_t}^2 - \tfrac{1}{2}\|f - f_\theta\|_{2,E_t}^2 + L_{2,t}(f) - L_{2,t}(f^\alpha)\right| \leq \alpha t[8C + \sqrt{8\sigma^2 \ln(4t^2/\delta)}] \quad \forall t \in \mathbb{N}. \tag{B1}$$

Proof. Since any two functions in $f, f^\alpha \in \mathcal{F}$ satisfy $\|f - f^\alpha\|_\infty \leq C$, it is enough to consider $\alpha \leq C$. We find

$$|(f^\alpha)^2(a) - (f)^2(a)| \leq \max_{-\alpha \leq y \leq \alpha} |(f(a) + y)^2 - f(a)^2| = 2f(a)\alpha + \alpha^2 \leq 2C\alpha + \alpha^2,$$

which implies

$$|(f^\alpha(a) - f_\theta(a))^2 - (f(a) - f_\theta(a))^2| = |[(f^\alpha)(a)^2 - f(a)^2] + 2f_\theta(a)(f(a) - f^\alpha(a))| \leq 4C\alpha + \alpha^2$$
$$|(R_t - f(a))^2 - (R_t - f^\alpha(a))^2| = |2R_t(f^\alpha(a) - f(a)) + f(a)^2 - f^\alpha(a)^2| \leq 2\alpha|R_t| + 2C\alpha + \alpha^2.$$

Summing over t, we find that the left-hand side of (B1) is bounded by

$$\sum_{k=1}^{t-1}\left(\frac{1}{2}[4C\alpha + \alpha^2] + [2\alpha|R_k| + 2C\alpha + \alpha^2]\right) \leq \alpha \sum_{k=1}^{t-1}(6C + 2|R_k|).$$

Because ϵ_k is sub-Gaussian, $\mathbb{P}(|\epsilon_k| > \sqrt{2\sigma^2 \ln(2/\delta)}) \leq \delta$. By a union bound,

$$\mathbb{P}(\exists k \text{ s.t. } |\epsilon_k| > \sqrt{2\sigma^2 \ln(4t^2/\delta)}) \leq \frac{\delta}{2}\sum_1^\infty \frac{1}{k^2} \leq \delta.$$

Since $|R_k| \leq C + |\epsilon_k|$, this shows that with probability at least $1 - \delta$ the discretization error is bounded for all t by $\alpha\eta_t$, where $\eta_t := t[8C + 2\sqrt{2\sigma^2 \ln(4t^2/\delta)}]$. $\square$

Appendix C. Bounds on eluder dimension for common function classes. Definition 3, which defines the eluder dimension of a class of functions, can be equivalently written as follows. The ϵ-eluder dimension of a class of functions $\mathcal{F}$ is the length of the longest sequence $a_1, \ldots, a_\tau$ such that for some $\epsilon' \geq \epsilon$,

$$w_k := \sup\left\{(f_{\rho_1} - f_{\rho_1})(a_k): \sqrt{\sum_{i=1}^{k-1}(f_{\rho_1} - f_{\rho_2})^2(a_i)} \leq \epsilon' \ \rho_1, \rho_2 \in \Theta\right\} > \epsilon' \tag{C1}$$

for each $k \leq \tau$.

C.1. Finite action spaces. Any action is ϵ'-dependent on itself since

$$\sup\{(f_{\rho_1} - f_{\rho_1})(a): \sqrt{(f_{\rho_1} - f_{\rho_2})^2(a)} \leq \epsilon' \ \rho_1, \rho_2 \in \Theta\} \leq \epsilon'.$$

Therefore, for all $\epsilon > 0$, the ϵ-eluder dimension of $\mathcal{A}$ is bounded by $|\mathcal{A}|$.

C.2. Linear case.

Proposition 11. *Suppose $\Theta \subset \mathbb{R}^d$ and $f_\theta(a) = \theta^T \phi(a)$. Assume there exist constants γ and S such that for all $a \in \mathcal{A}$ and $\rho \in \Theta$, $\|\rho\|_2 \leq S$, and $\|\phi(a)\|_2 \leq \gamma$. Then $\dim_E(\mathcal{F}, \epsilon) \leq 3d(e/(e-1))\ln\{3 + 3((2S)/\epsilon)^2\} + 1$.*

To simplify the notation, define w_k as in (C1), $\phi_k = \phi(a_k)$, $\rho = \rho_1 - \rho_2$, and $\Phi_k = \sum_{i=1}^{k-1} \phi_i \phi_i^T$. In this case, $\sum_{i=1}^{k-1}(f_{\rho_1} - f_{\rho_2})^2(a_i) = \rho^T \Phi_k \rho$, and by the triangle inequality $\|\rho\|_2 \leq 2S$. The proof follows by bounding the number of times $w_k > \epsilon'$ can occur.

Step 1. If $w_k \geq \epsilon'$, then $\phi_k^T V_k^{-1} \phi_k \geq \frac{1}{2}$ where $V_k := \Phi_k + \lambda I$ and $\lambda = (\epsilon'/(2S))^2$.

Proof. We find $w_k \leq \max\{\rho^T \phi_k\colon \rho^T \Phi_k \rho \leq (\epsilon')^2, \rho^T I \rho \leq (2S)^2\} \leq \max\{\rho^T \phi_k\colon \rho^T V_k \rho_k \leq 2(\epsilon')^2\} = \sqrt{2(\epsilon')^2}\|\phi_k\|_{V_k^{-1}}$. The second inequality follows because any ρ that is feasible for the first maximization problem must satisfy $\rho^T V_k \rho \leq (\epsilon')^2 + \lambda(2S)^2 = 2(\epsilon')^2$. By this result, $w_k \geq \epsilon'$ implies $\|\phi_k\|^2_{V_k^{-1}} \geq 1/2$. □

Step 2. If $w_i \geq \epsilon'$ for each $i < k$, then $\det V_k \geq \lambda^d(\frac{3}{2})^{k-1}$ and $\det V_k \leq ((\gamma^2(k-1))/d + \lambda)^d$.

Proof. Since $V_k = V_{k-1} + \phi_k \phi_k^T$, using the matrix determinant lemma,

$$\det V_k = \det V_{k-1}(1 + \phi_t^T V_k^{-1} \phi_t) \geq \det V_{k-1}\left(\frac{3}{2}\right) \geq \cdots \geq \det[\lambda I]\left(\frac{3}{2}\right)^{k-1} = \lambda^d\left(\frac{3}{2}\right)^{k-1}.$$

Recall that $\det V_k$ is the product of the eigenvalues of V_k, whereas $\mathrm{trace}[V_k]$ is the sum. As noted in Dani et al. [16], $\det V_k$ is maximized when all eigenvalues are equal. This implies $\det V_k \leq ((\mathrm{trace}[V_k])/d)^d \leq ((\gamma^2(t-1))/d + \lambda)^d$. □

Step 3. Complete proof.

Proof. Manipulating the result of Step 2 shows k must satisfy the inequality: $(\frac{3}{2})^{(k-1)/d} \leq \alpha_0[(k-1)/d] + 1$, where $\alpha_0 = \gamma^2/\lambda = (2S\gamma/\epsilon')^2$. Let $B(x, \alpha) = \max\{B\colon (1+x)^B \leq \alpha B + 1\}$. The number of times $w_k > \epsilon'$ can occur is bounded by $dB(1/2, \alpha_0) + 1$.

We now derive an explicit bound on $B(x, \alpha)$ for any $x \leq 1$. Note that any $B \geq 1$ must satisfy the inequality $\ln\{1+x\}B \leq \ln\{1+\alpha\} + \ln B$. Since $\ln\{1+x\} \geq x/(1+x)$, using the transformation of variables $y = B[x/(1+x)]$ gives

$$y \leq \ln\{1+\alpha\} + \ln\frac{1+x}{x} + \ln y \leq \ln\{1+\alpha\} + \ln\frac{1+x}{x} + \frac{y}{e} \quad \Longrightarrow \quad y \leq \frac{e}{e-1}\left(\ln\{1+\alpha\} + \ln\frac{1+x}{x}\right).$$

This implies $B(x, \alpha) \leq ((1+x)/x)(e/(e-1))(\ln\{1+\alpha\} + \ln((1+x)/x))$. The claim follows by plugging in $\alpha = \alpha_0$ and $x = 1/2$. □

C.3. Generalized linear models.

Proposition 12. *Suppose $\Theta \subset \mathbb{R}^d$ and $f_\theta(a) = g(\theta^T \phi(a))$ where $g(\cdot)$ is a differentiable and strictly increasing function. Assume that there exist constants $\underline{h}$, $\bar{h}$, γ, and S such that for all $a \in \mathcal{A}$ and $\rho \in \Theta$, $0 < \underline{h} \leq g'(\rho^T \phi(a)) \leq \bar{h}$, $\|\rho\|_2 \leq S$, and $\|\phi(a)\|_2 \leq \gamma$. Then $\dim_E(\mathcal{F}, \epsilon) \leq 3dr^2(e/(e-1))\ln\{3r^2 + 3r^2((2S\bar{h})/\epsilon)^2\} + 1$.*

The proof follows three steps that closely mirror those used to prove Proposition 11.

Step 1. If $w_k \geq \epsilon'$, then $\phi_k^T V_k^{-1} \phi_k \geq 1/(2r^2)$ where $V_k := \Phi_k + \lambda I$ and $\lambda = (\epsilon'/(2S\underline{h}))^2$.

Proof. By definition $w_k \leq \max\{g(\rho^T \phi_k)\colon \sum_{i=1}^{k-1} g(\rho^T \phi(a_i))^2 \leq (\epsilon')^2, \rho^T I \rho \leq (2S)^2\}$. By the uniform bound on $g'(\cdot)$ this is less than $\max\{\bar{h}\rho^T \phi_k\colon \underline{h}^2 \rho^T \Phi_k \rho \leq (\epsilon')^2, \rho^T I \rho \leq (2S)^2\} \leq \max\{\bar{h}\rho^T \phi_k\colon \underline{h}^2 \rho^T V_k \rho \leq 2(\epsilon')^2\} = \sqrt{2(\epsilon')^2/r^2}\|\phi_k\|_{V_k^{-1}}$. □

Step 2. If $w_i \geq \epsilon'$ for each $i < k$, then $\det V_k \geq \lambda^d(\frac{3}{2})^{k-1}$ and $\det V_k \leq ((\gamma^2(k-1))/d + \lambda)^d$.

Step 3. Complete proof.

Proof. The above inequalities imply k must satisfy $(1 + 1/(2r^2))^{(k-1)/d} \leq \alpha_0[(k-1)/d]$, where $\alpha_0 = \gamma^2/\lambda$. Therefore, as in the linear case, the number of times $w_k > \epsilon'$ can occur is bounded by $dB(1/(2r^2), \alpha_0) + 1$. Plugging these constants into the earlier bound $B(x, \alpha) \leq ((1+x)/x)(e/(e-1))(\ln\{1+\alpha\} + \ln((1+x)/x))$ and using $1 + x \leq 3/2$, yields the result. □

References

[1] Abbasi-Yadkori Y, Antos A, Szepesvári C (2009) Forced-exploration based algorithms for playing in stochastic linear bandits. *COLT Workshop: On-line Learn. Limited Feedback*.
[2] Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira FCN, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* (*NIPS*), (Currant Associates, Red Hook, NY), 2312–2320.
[3] Abbasi-Yadkori Y, Pal D, Szepesvári C (2012) Online-to-confidence-set conversions and application to sparse stochastic bandits. Lawrence ND, Girolami MA, eds. *Proc. 15th Internat. Conf. Artificial Intelligence Statist. (AISTATS)*, JMLR Workshop and Conference Proceedings, Vol. 22, 1–9.
[4] Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. Mannor S, Srebro N, Williamson RC, eds. *Proc. 25th Ann. Conf. Learn. Theory*, JMLR Workshop and Conference Proceedings, Vol. 23, 39.1–39.26
[5] Agrawal S, Goyal N (2012) Further optimal regret bounds for Thompson sampling. arXiv preprint arXiv:1209.3353.
[6] Agrawal S, Goyal N (2012) Thompson sampling for contextual bandits with linear payoffs. arXiv preprint arXiv:1209.3352.
[7] Amin K, Kearns M, Syed U (2011) Bandits, query learning, and the haystack dimension. Kakade SM, von Luxburg U, eds. *Proc. 24th Annual Conf. Learn. Theory (COLT)*, JMLR Workshop and Conference Proceedings, Vol. 19, 87–106.
[8] Audibert J-Y, Bubeck S (2009) Minimax policies for adversarial and stochastic bandits. *Proc. 22th Annual Conf. Learn. Theory (COLT)*, (Omnipress, Madison, WI), 773–818.
[9] Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2):235–256.
[10] Beygelzimer A, Langford J, Li L, Reyzin L, Schapire RE (2011) Contextual bandit algorithms with supervised learning guarantees. *Proc. 14th Internat. Conf. Artificial Intelligence Statist. (AISTATS)*, JMLR Workshop and Conference Proceedings, Vol. 15, 19–26.
[11] Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721.
[12] Bubeck S, Liu C-Y (2013) Prior-free and prior-dependent regret bounds for Thompson sampling. *Adv. Neural Inform. Processing Systems* (*NIPS*), 638–646.
[13] Bubeck S, Munos R, Stoltz G, Szepesvári C (2011) *X*-armed bandits. *J. Machine Learn. Res.* 12:1655–1695.
[14] Cappé O, Garivier A, Maillard O-A, Munos R, Stoltz G (2013) Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* 41(3):1516–1541.
[15] Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereria F, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* (*NIPS*) (Currant Associates, Red Hook, NY), 2249–2257.
[16] Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. *Proc. 21st Annual Conf. Learn. Theory (COLT)* (Omnipress, Madison, WI), 355–366.
[17] Deshpande Y, Montanari A (2012) Linear bandits in high dimension and recommendation systems. *50th Annual Allerton Conf. Communication, Control, Comput.* (IEEE, Piscataway, NJ), 1750–1754.
[18] Filippi S, Cappé O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. Lafferty J, Williams C, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Adv. Neural Inform. Processing Systems* (*NIPS*), (Currant Associates, Red Hook, NY), 586–594.
[19] Gittins JC, Jones DM (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* 66(3):561–565.
[20] Gittins J, Glazebrook K, Weber R (2011) *Multi-Armed Bandit Allocation Indices* (John Wiley & Sons, Chichester, UK).
[21] Gopalan A, Mannor S, Mansour Y (2013) Thompson sampling for complex bandit problems. arXiv preprint arXiv:1311.0466.
[22] Kauffmann E, Korda N, Munos R (2012) Thompson sampling: An asymptotically optimal finite time analysis. Bshouty NH, Stoltz G, Vayatis N, Zeugmann T, eds., *Algorithmic Learn. Theory*, Lecture Notes in Computer Science, Vol. 7568 (Springer-Verlag, Berlin, Heidelberg), 586–594.
[23] Kleinberg R, Slivkins A, Upfal E (2008) Multi-armed bandits in metric spaces. *Proc. 40th ACM Sympos. Theory Comput.* (ACM, New York), 681–690.
[24] Korda N, Kaufmann E, Munos R (2013) Thompson sampling for one-dimensional exponential family bandits. Burges CJC, Bottou L, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems*, 1448–1456.
[25] Lai TL (1987) Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* 15(3):1091–1114.
[26] Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.
[27] Lehmann EL, Casella G (1998) *Theory of Point Estimation*, 2nd ed. (Springer-Verlag, New York).
[28] Li L (2013) Generalized Thompson sampling for contextual bandits. arXiv preprint arXiv:1310.7163.
[29] Li L, Chapelle O (2012) Open problem: Regret bounds for Thompson sampling. Mannor S, Srebro B, Williamson RC, eds. *Proc. 25th Annual Conf. Learn. Theory (COLT)*, JMLR Workshop and Conference Proceedings, Vol. 23, 43.1–43.3.
[30] May BC, Korda N, Lee A, Leslie DS (2012) Optimistic Bayesian sampling in contextual-bandit problems. *J. Machine Learn. Res.* 13(1):2069–2106.
[31] Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Math. Oper. Res.* 35(2):395–411.
[32] Ryzhov IO, Powell WB, Frazier PI (2012) The knowledge gradient algorithm for a general class of online learning problems. *Oper. Res.* 60(1):180–195.
[33] Sahni A (1974) Computationally related problems. *SIAM J. Comput.* 3(4):262–279.
[34] Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Appl. Stochastic Models Bus. Indust.* 26(6):639–658.
[35] Srinivas N, Krause A, Kakade SM, Seeger M (2012) Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *Inform. Theory, IEEE Trans.* 58(5):3250–3265.
[36] Valko M, Carpentier A, Munos R (2013) Stochastic simultaneous optimistic optimization. Dasgupta S, Mcallester D, eds. *Proc. 30th Internat. Conf. Machine Learn. (ICML-13)*, JMLR Workshop and Conference Proceedings, Vol. 28, 19–27.