Learning to Optimize Via Information Directed Sampling

Daniel Russo and Benjamin Van Roy

March 25, 2014

Abstract

This paper proposes information directed sampling—a new algorithm for balancing between exploration and exploitation in online optimization problems in which a decision-maker must learn from partial feedback. The algorithm quantifies the amount learned by selecting an action through an information theoretic measure: the mutual information between the true optimal action and the algorithm’s next observation. Actions are then selected by optimizing a myopic objective that balances earning high immediate reward and acquiring information. We show this algorithm is provably efficient and is empirically efficient in simulation trials. We provide novel and general regret bounds that scale with the entropy of the optimal action distribution. Furthermore, as we highlight through several examples, information directed sampling sometimes dramatically outperforms popular approaches like UCB algorithms and Thompson sampling which don’t quantify the information provided by different actions.

1 Introduction

In this technical note, we propose information-directed sampling, a new approach to making sequential decisions in the presence of model uncertainty. This approach is motivated by our recent information-theoretic analysis of Thompson sampling [17], which led to a regret bound that scales with the entropy of the optimal action distribution. A distinctive feature of this bound is its dependence on soft knowledge, which conveys which models are more or less likely to match reality. Prior analyses focussed on the dependence of regret on hard knowledge, which identifies a family of possible models.

Soft knowledge is typically represented in terms of a probability distribution or confidence set and evolves as an agent learns from decision outcomes. A regret bound that depends on soft knowledge captures how new observations reduce subsequent regret and, thus, yields insight into how an agent should make decisions that effectively balance between exploration and exploitation. It is this kind of insight from our prior work [17] that led us to the design of the algorithms and analyses presented in this technical note.

We are currently extending the theory contained herein and carrying out computational studies to assess the efficacy of information-directed sampling and the benefits it offers over prior art. We plan to report further findings as this research progresses.

2 Problem Formulation

The decision-maker sequentially chooses actions \((A_t)_{t \in \mathbb{N}}\) from a finite action set \(A\) and observes the corresponding outcomes \((Y_t(A_t))_{t \in \mathbb{N}}\). To avoid measure-theoretic subtleties, we assume the set \(\mathcal{Y}\) of possible outcomes is a subset of a finite dimensional Euclidean space. Conditional on
the true outcome distribution $p^*$, the random variables $\{Y_t(a)\}_{t \in \mathbb{N}}$ are drawn i.i.d according to $p^*_a$ for each action $a \in \mathcal{A}$. In particular, we assume that for any $\tilde{\mathcal{Y}}_1, ..., \tilde{\mathcal{Y}}_t \subset \mathcal{Y}$ and $a_1, ..., a_t \in \mathcal{A}$, $\mathbb{P}(Y_t(a_1) \in \tilde{\mathcal{Y}}_1, ..., Y_t(a_t) \in \tilde{\mathcal{Y}}_t | p^*) = \prod_{k=1}^t p^*_a(\tilde{\mathcal{Y}}_k)$. The true outcome distribution $p^*$ is itself randomly drawn from the family $\mathcal{P}$ of distributions.

We define all random variables with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Actions are chosen based on the history of past observations and possibly some external source of randomness. To represent this external source of randomness more formally, we introduce an i.i.d. sequence of random variables $(U_t)_{t \in \mathbb{N}}$ that are jointly independent of the outcomes $\{Y_t(a)\}_{t \in \mathbb{N}, a \in \mathcal{A}}$ and $p^*$. We fix the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$ where $\mathcal{F}_{t-1} \subset \mathcal{F}$ is the sigma–algebra generated by $(U_1, A_1, Y_1(A_1), ..., U_{t-1}, A_{t-1}, Y_{t-1}(A_{t-1}))$. The action $A_t$ is measurable with respect to the sigma–algebra generated by $(\mathcal{F}_{t-1}, U_t)$. That is, given the history of past observations, $A_t$ is random only through its dependence on $U_t$.

The agent associates a reward $R(y)$ with each outcome $y \in \mathcal{Y}$, where the reward function $R : \mathcal{Y} \to \mathbb{R}$ is fixed and known. Uncertainty about $p^*$ induces uncertainty about the true optimal action, which we denote by $A^* \in \arg \max_{a \in \mathcal{A}} \mathbb{E}_{y \sim p^*_a} [R(y)]$. The $T$ period regret of the sequence of actions $A_1, ..., A_T$ is the random variable,

$$\text{Regret}(T) := \sum_{t=1}^T [R(Y_t(A^*)) - R(Y_t(A_t))],$$

which measures the cumulative difference between the reward earned by an algorithm that always chooses the optimal action, and actual accumulated reward up to time $T$. In this paper we study expected regret

$$\mathbb{E} [\text{Regret}(T)] = \mathbb{E} \left[ \sum_{t=1}^T [R(Y_t(A^*)) - R(Y_t(A_t))] \right],$$

where the expectation is taken over the randomness in the actions $A_t$ and the outcomes $Y_t$, and over the prior distribution over $p^*$. This measure of performance is commonly called Bayesian regret or Bayes risk.

The objective is to choose actions in a manner that minimizes expected regret. For this purpose, it’s useful to think of the actions as being chosen by a randomized policy $\pi$, which is an $\mathcal{F}_t$–predictable sequence $(\pi_t)_{t \in \mathbb{N}}$. An action is chosen at time $t$ by randomizing according to $\pi_t(\cdot) = \mathbb{P}(A_t \in \cdot | \mathcal{F}_{t-1})$, which specifies a probability distribution over $\mathcal{A}$. We denote the set of probability distributions over $\mathcal{A}$ by $\mathcal{D}(\mathcal{A})$. We explicitly display the dependence of regret on the policy $\pi$, letting $\mathbb{E} [\text{Regret}(T, \pi)]$ denote the expected value given by $[2]$ when the actions $(A_1, ..., A_T)$ are chosen according to $\pi$.

### 2.1 Further Notation

We set $\alpha_t(a) = \mathbb{P}(A^* = a | \mathcal{F}_{t-1})$ to be the posterior distribution of $A^*$. For two probability measures $P$ and $Q$ over a common measurable space, if $P$ is absolutely continuous with respect to $Q$, the Kullback-Leibler divergence between $P$ and $Q$ is

$$D_{\text{KL}}(P||Q) = \int_{\mathcal{Y}} \log \left( \frac{dP}{dQ} \right) dP$$

(3)

where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of $P$ with respect to $Q$. For a probability distribution $P$ over a finite set $\mathcal{X}$, the Shannon entropy of $P$ is defined as

$$H(P) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{1}{P(x)} \right).$$

(4)
The *mutual information* under the posterior distribution between two random variables $X_1 : \Omega \rightarrow X_1$, and $X_2 : \Omega \rightarrow X_2$

$$I_t(X_1; X_2) := D_{KL}(\mathbb{P}((X_1, X_2) \in \cdot | F_{t-1}) \Vert \mathbb{P}(X_1 \in \cdot | F_{t-1})\mathbb{P}(X_2 \in \cdot | F_{t-1}))$$

(5)

is the Kullback-Leibler divergence between the joint posterior distribution of $X_1$ and $X_2$ and the product of the marginal distributions. Note that $I_t(X_1; X_2)$ is a random variable because of its dependence on the conditional probability measure $\mathbb{P}(\cdot | F_{t-1})$.

To reduce notation, we define the *information gain* from an action $a$ to be $g_t(a) := I_t(A^*; Y_t(a))$. This is equal to the expected reduction in entropy of the posterior distribution of $A^*$ due to observing $Y_t(a)$:

$$g_t(a) = \mathbb{E}[H(\alpha_t) - H(\alpha_{t+1})| F_{t-1}, A_t = a],$$

(6)

which plays a crucial role in our results.

Let $g_t(\pi_t) = \sum_{a \in A} \pi_t(a) g_t(a)$ denote the expected information gain when actions are selected according to $\pi_t$. Let $\Delta_t(a) := \mathbb{E}[R_t(Y_t(A^*)) - R(Y_t(a))| F_{t-1}]$ denote the expected instantaneous regret of action $a$ at time $t$ and $\Delta_t(\pi_t) = \sum_{a \in A} \pi_t(a) \Delta_t(a)$.

## 3 Information Directed Sampling

Information directed sampling explicitly balances between attaining low expected immediate regret and acquiring new information about which action is optimal. It does this by maximizing over all action sampling distributions $\pi \in \mathcal{D}(A)$ a single-period objective that encourages low regret $\Delta_t(\pi)$ and high information gain $g_t(\pi)$ about the optimal action $A^*$. In particular, the policy $\pi_{IDS}(\lambda) = (\pi_{1IDS}(\lambda), \pi_{2IDS}(\lambda), \ldots)$ is chosen so that

$$\pi_{tIDS}(\lambda) \in \arg \min_{\pi \in \mathcal{D}(A)} \{ \Delta^2_t(\pi) - \lambda g_t(\pi) \}$$

(7)

where $\lambda \in \mathbb{R}$ is a tuning parameter.

Adjusting the the parameter $\lambda$ gives users a simple way to tune how aggressively the algorithm explores, but performance depends strongly on the choice of $\lambda$, which can be undesirable. In some settings, we prove regret bounds for information directed sampling when applied with particular values for $\lambda$. Our theoretical guarantees also apply to a policy that avoids the need for tuning. The policy $\pi_{IDS(R)} = (\pi_{1IDS(R)}, \pi_{2IDS(R)}, \ldots)$ is chosen to minimize the ratio between the square of expected regret and expected information gain:

$$\pi_{tIDS(R)} \in \arg \min_{\pi \in \mathcal{D}(A)} \frac{\Delta^2_t(\pi)}{g_t(\pi)}.$$

(8)

The solution to (7) or (8) may be a randomized policy, but the following proposition shows that the randomization will be over at most two actions.

**Proposition 1.** For every $t \in \mathbb{N}$ and $\lambda \geq 0$ there exists a distribution $\pi^* \in \arg \min_{\pi \in \mathcal{D}(A)} \{ \Delta_t(\pi)^2 - \lambda g_t(\pi) \}$ with support over at most 2 actions. Similarly, there exists a distribution $\pi^* \in \arg \min_{\pi \in \mathcal{D}(A)} \frac{\Delta_t(\pi)^2}{g_t(\pi)}$ with support over at most 2 actions.

By this result, if $\Delta_t(a)$ and $g_t(a)$ have been computed for each action $a \in A$, we can compute $\pi_{tIDS(\lambda)}$ in a straightforward manner. A distribution with support at most 2 is characterized by a
pair of actions \((a_1, a_2) \in \mathcal{A} \times \mathcal{A}\) and a probability \(p \in [0, 1]\) of selecting \(a_1\) instead of \(a_2\). We can compute the distribution that maximizes (7) effectively by looping over all pairs of actions and finding the optimal probability.

4 Regret bounds

The following proposition provides general regret bounds for information directed sampling that scale with the entropy of the optimal action distribution. They depend on what we call the information ratio, which is the following ratio between regret and information gain:

\[
\Psi_t(\pi) := \frac{\Delta_t(\pi)^2}{g_t(\pi)} \quad \Psi^*_t := \min_{\pi \in \mathcal{D}(\mathcal{A})} \frac{\Delta_t(\pi)^2}{g_t(\pi)}.
\]

We will call \(\Psi_t(\pi)\) the information ratio of a sampling distribution \(\pi \in \mathcal{D}(\mathcal{A})\) and \(\Psi^*_t\) the minimal information ratio.

**Proposition 2.** Fix a deterministic \(\lambda \in \mathbb{R}\) such that \(\Psi^*_t \leq \lambda\) almost surely for each \(t \in \{1, \ldots, T\}\). Then,

\[
\mathbb{E} \left[ \text{Regret} \left( \pi^{\text{IDS}(\lambda)}, T \right) \right] \leq \sqrt{\lambda H(\alpha_1)T}
\]

and

\[
\mathbb{E} \left[ \text{Regret} \left( \pi^{\text{IDS}(R)}, T \right) \right] \leq \sqrt{\lambda H(\alpha_1)T}.
\]

This proposition relies on a worst case bound on the minimal information ratio of the form \(\Psi^*_t \leq \lambda\), and we will provide several bounds of that form in the next section. For example, for “linear bandit” problems, we show \(\Psi^*_t \leq d/2\), where \(d\) is the dimension of the linear model.

The next proposition shows that one can also state regret bounds that depend on the average value of \(\Psi^*_t\) instead of a worst case bound.

**Proposition 3.**

\[
\mathbb{E} \left[ \text{Regret} \left( \pi^{\text{IDS}(R)}, T \right) \right] \leq \sqrt{\left( \frac{1}{T^2} \sum_{t=1}^{T} \Psi_t^* \right) H(\alpha_1)T}.
\]

4.1 Bounds on the minimal information ratio

This section establishes upper bounds on the minimal information ratio \(\Psi^*_t\) in several important settings. This yields explicit regret bounds when combined with Proposition 2. These worst–case bounds follow from our recent analysis of Thompson sampling, and the implied regret bounds are the same as those established for Thompson sampling. However, information directed sampling outperforms Thompson sampling in simulation, and, as we will highlight in the next section, is sometimes provably much more informationally efficient.

The bounds provided here also help to clarify the role the minimal information ratio \(\Psi^*_t\) plays in our results: it roughly captures the extent to which sampling some actions allows the decision maker to make inferences about other actions. In the worst case, the ratio depends on the number of actions, reflecting the fact that actions could provide no information about others. For problems with full information, the minimal information ratio is bounded by a numerical constant, reflecting that sampling one action perfectly reveals the rewards that would have been earned by selecting any other action. The problems of online linear optimization under “bandit feedback” and under “semi–bandit feedback” lie between these two extremes, and the ratio provides a natural measure
of each problem’s information structure. In each case, our bounds reflect that information directed
sampling is able to automatically exploit this structure.

The proofs of these bounds are omitted from this short draft, but each one follows from our
recent work on Thompson sampling [17]. In particular, since \(\Psi_t^* \leq \Psi_t(\pi_{TS})\) where \(\pi_{TS}\) is the
Thompson sampling policy, it is enough to bound \(\Psi_t(\pi_{TS})\). Several such bounds were provided in
Russo and Van Roy [17].

For each problem setting, we will compare our upper bounds on expected regret with known
lower bounds. Some of these lower bounds were developed and stated in an adversarial framework,
but were proved using the probabilistic method; authors fixed a family of distributions \(\mathcal{P}\) and an
initial distribution over \(p^*\) and lower bounded the expected regret under this environment of any
algorithm. This provides lower bounds on \(\inf_{\pi} \mathbb{E}[\text{Regret}(T, \pi)]\) in our framework.

To simplify the exposition, our results are stated under the assumption that rewards are uni-
formly bounded. This effectively controls the worst-case variance of the reward distribution, and
as shown in the appendix of Russo and Van Roy [17], our results can be extended to the case where
reward distributions are sub-Gaussian.

**Assumption 1.** \(\sup_{y \in \mathcal{Y}} R(y) - \inf_{y \in \mathcal{Y}} R(y) \leq 1\).

### 4.1.1 Worst Case Bound

The next proposition shows that \(\Psi^*_t\) is never larger than \(|A|/2\). That is, there is always an action
sampling distribution \(\pi\) such that \(\Delta_t(\pi)^2 \leq (|A|/2)g_t(\pi)\). In the next section, we will will show
that under different information structures the ratio between regret and information gain can be
much smaller, which leads to stronger theoretical guarantees.

**Proposition 4.** For any \(t \in \mathbb{N}\), \(\Psi^*_t \leq |A|/2\) almost surely.

Combining Proposition 4 with Proposition 2 shows that \(\mathbb{E}\left[\text{Regret}\left(\pi_{IDS(R)}, T\right)\right] \leq \sqrt{\frac{1}{2}|A|H(\alpha_1)T}\). Further,
a worst-case bound on the entropy of \(\alpha_1\) shows that \(\mathbb{E}\left[\text{Regret}(T, \pi_{IDS(R)})\right] \leq \sqrt{\frac{1}{2}\log(|A|)T}\). Dani
et al. [7] show this bound is order optimal, in the sense that for any time horizon \(T\) and number of ac-
tions \(|A|\) there exists a prior distribution over \(p^*\) under which \(\inf_{\pi} \mathbb{E}[\text{Regret}(T, \pi)] \geq c_0 \sqrt{\log(|A|)T}\)
where \(c_0\) is a numerical constant that does not depend on \(|A|\) or \(T\). The bound here improves upon

\[1\Psi_t(\pi_{TS})\) is exactly equal to the term \(\Gamma^2_t\), that is bounded in Russo and Van Roy [17].
this worst case bound since $H(\alpha_1)$ can be much smaller than $\log(|\mathcal{A}|)$ when the prior distribution in informative.

### 4.1.3 Linear Optimization Under Bandit Feedback

The stochastic linear bandit problem has been widely studied (e.g. [1, 8, 15]) and is one of the most important examples of a multi-armed bandit problem with “correlated arms.” In this setting, each action is associated with a finite dimensional feature vector, and the mean reward generated by an action is the inner product between its known feature vector and some unknown parameter vector. Because of this structure, observations from taking one action allow the decision-maker to make inferences about other actions. The next proposition bounds the minimal information ratio for such problems.

**Proposition 6.** If $\mathcal{A} \subset \mathbb{R}^d$ and for each $p \in \mathcal{P}$ there exists $\theta_p \in \mathbb{R}^d$ such that for all $a \in \mathcal{A}$

$$
\mathbb{E}_{y \sim p_a} [R(y)] = a^T \theta_p,
$$

then for all $t \in \mathbb{N}$, $\Psi_t^* \leq d/2$ almost surely.

This result shows that $\mathbb{E} \left[ \text{Regret}(T, \pi^{IDS}(R)) \right] \leq \sqrt{\frac{1}{2} H(\alpha_1) d T} \leq \sqrt{\frac{1}{2} \log(|\mathcal{A}|) d T}$ for linear bandit problems. Again, Dani et al. [7] show this bound is order optimal, in the sense that for any time horizon $T$ and dimension $d$ if the actions set is $\mathcal{A} = \{0, 1\}^d$, there exists a prior distribution over $p^*$ such that $\inf_{\pi} \mathbb{E} \left[ \text{Regret}(T, \pi) \right] \geq c_0 \sqrt{\log(|\mathcal{A}|) d T}$ where $c_0$ is a constant that is independent of $d$ and $T$. The bound here improves upon this worst case bound since $H(\alpha_1)$ can be much smaller than $\log(|\mathcal{A}|)$ when the prior distribution is informative.

### 4.1.4 Combinatorial Action Sets and “Semi–Bandit” Feedback

To motivate the information structure studied here, consider a simple resource allocation problem. There are $d$ possible projects, but the decision-maker can allocate resources to at most $m \leq d$ of them at a time. At time $t$, project $i \in \{1, ..., d\}$ yields a random reward $\theta_{t,i}$, and the reward from selecting a subset of projects $a \in \mathcal{A} \subset \{a' \in \{0, 1, ..., d\} : |a'| \leq m\}$ is $m^{-1} \sum_{i \in a} \theta_{t,i}$. In the linear bandit formulation of this problem, upon choosing a subset of projects $a$ the agent would only observe the overall reward $m^{-1} \sum_{i \in a} \theta_{t,i}$. It may be natural instead to assume that the outcome of each selected project $(\theta_{t,i} : i \in a)$ is observed. This type of observation structure is sometimes called “semi–bandit” feedback.

A naive application of Proposition 6 to address this problem would show $\Psi_t^* \leq d/2$. The next proposition shows that since the entire parameter vector $(\theta_{t,i} : i \in a)$ is observed upon selecting action $a$, we can provide an improved bound on the information ratio. The proof of the proposition is provided in the appendix.

**Proposition 7.** Suppose $\mathcal{A} \subset \{a \in \{0, 1, ..., d\} : |a| \leq m\}$, and that there are random variables $(\theta_{t,i} : t \in \mathbb{N}, i \in \{1, ..., d\})$ such that

$$
Y_t(a) = (\theta_{t,i} : i \in a) \quad \text{and} \quad R(Y_t(a)) = \frac{1}{m} \sum_{i \in a} \theta_{t,i}.
$$

Assume that the random variables $(\theta_{t,i} : i \in \{1, ..., d\})$ are independent conditioned on $\mathcal{F}_{t-1}$ and $\theta_{t,i} \in \left[ -\frac{1}{2}, \frac{1}{2} \right]$ almost surely for each $(t, i)$. Then for all $t \in \mathbb{N}$, $\Psi_t \leq \frac{d}{2m^2}$ almost surely.
In this problem, there are as many as $d^m$ actions, but because information directed sampling exploits the structure relating actions to one another, its regret is only polynomial in $m$ and $d$. In particular, combining Proposition 7 with Proposition 2 shows that:

$$E[\text{Regret}(T, \pi_{IDS(R)})] \leq \frac{1}{m}\sqrt{dH(\alpha_1)T}.$$  

Since $H(\alpha_1) \leq \log |A| = O(m \log(\frac{d}{m}))$ this also yields a bound of order $\sqrt{\frac{d}{m} \log \left(\frac{d}{m}\right)} T$. As shown by Audibert et al. [3], the lower bound for this problem is of order $\sqrt{\frac{d}{m}T}$, so our bound is order optimal up to a $\sqrt{\log(\frac{d}{m})}$ factor.

5 Beyond UCB and Thompson sampling

Upper confidence bound algorithms (UCB) and Thompson sampling are two of the most popular approaches to balancing between exploration and exploitation. In some cases, UCB algorithms and Thompson sampling are empirically effective, and have strong theoretical guarantees. Specific UCB algorithms and Thompson sampling are known to be asymptotically efficient for multi–armed bandit problems with independent arms [2, 6, 11, 13, 14] and satisfy strong regret bounds for some problems with dependent arms [5, 8, 9, 10, 15, 16, 18].

Each of these algorithms balances between exploration and exploitation by selecting actions that are likely to yield high reward, but encouraging the selection of actions with uncertain mean reward. However, in general, the degree of the algorithm’s uncertainty about an action’s mean reward is not an effective measure of the information it provides about the optimum. For this reason, as we will show in this section, such algorithms can be grossly suboptimal. We demonstrate this through several examples - each of which is designed to be simple and transparent. To set the stage for our discussion, we introduce UCB algorithms and Thompson sampling in the next subsection.

5.1 UCB Algorithms and Thompson sampling

Bayesian UCB algorithms

Upper confidence bound algorithms provide a simple method for balancing between exploration and exploitation. Actions are selected through two steps. First, for each action $a \in A$ and upper confidence bound $U_t(a)$ is constructed. Then the algorithm selects an action $A_t \in \arg\max_{a \in A} U_t(a)$ with maximal upper confidence bound. The upper confidence bound $U_t(a)$ represents the greatest mean reward value that is statistically plausible. In particular, $U_t(a)$ is typically constructed so that:

$$E_{y \sim p^*_a}[R(y)] \leq U_t(a)$$

with high probability, but that $U_t(a) \to E_{y \sim p^*_a}[R(y)]$ as data about action $a$ accumulates. One effective method involves choosing $U_t(a)$ to be a particular quantile of the posterior distribution of the mean reward [12].

---

**In their formulation, the reward from selecting action $a$ is $\sum_{i \in a} \theta_{t, i}$, which is $m$ times larger than in our formulation. The lower bound stated in their paper is therefore of order $\sqrt{mdT}$. They don’t provide a complete proof of their result, but note that it follows from standard lower bounds in the bandit literature. In the proof of Theorem 5 in that paper, they construct an example in which the decision maker plays $m$ bandit games in parallel, each with $d/m$ actions. Using that example, and the standard bandit lower bound (see Theorem 3.5 of Bubeck and Cesa-Bianchi [4]), the agent’s regret from each component must be at least $\sqrt{\frac{d}{m}T}$, and hence her overall expected regret is lower bounded by a term of order $m\sqrt{\frac{d}{m}T} = \sqrt{mdT}$.**

7
Thompson sampling

The Thompson sampling algorithm simply samples actions according to the posterior probability they are optimal. In particular, actions are chosen randomly at time $t$ according to the sampling distribution $\pi_t^{TS} = \alpha_t$. By definition, this means that for each $a \in \mathcal{A}$, $P(A_t = a|\mathcal{F}_{t-1}) = P(A^* = a|\mathcal{F}_{t-1}) = \alpha_t(a)$. This algorithm is sometimes called probability matching because the action selection distribution is matched to the posterior distribution of the optimal action.

5.2 Example 1: A revealing action

Let $\mathcal{A} = \{a_0, a_1, ..., a_K\}$ and suppose that $p^*$ is drawn uniformly at random from a finite set $\mathcal{P} = \{p^1, ..., p^K\}$ of alternatives. Consider a problem with bandit-feedback $Y_t(a) = R(Y_t(a))$. Under $p^i$, the reward of action $a_j$ is

$$R(Y_t(a)) = \begin{cases} 1 & j = i \\ 0 & j \neq i, j \neq 0 \\ \frac{1}{K} & j = 0 \end{cases}$$

Action $a_0$ is known to never yield the maximal reward, and is therefore never selected by Thompson sampling or any sensible UCB algorithm. Instead, these algorithms will select among $\{a_1, ..., a_n\}$ ruling out only a single action at a time until a reward 1 is earned and the optimal action is identified. Their expected regret therefore grows linearly in $n$.

Information directed sampling, on the other hand, is able to recognize that much more is learned by drawing action $a_0$ than by selecting one of the other arms. In fact, selecting action $a_0$ uniquely identifies the optimal action. Information directed sampling will select this action, learn which action is optimal, and select that action in all future periods. Its regret is therefore independent of $n$.

5.3 Example 2: Sparse linear bandits

Consider a linear bandit problem where $\mathcal{A} \subset \mathbb{R}^d$ and the reward from an action $a \in \mathcal{A}$ is $a^T\theta^*$ where $\theta^*$ is an unknown parameter. $\theta^*$ is drawn uniformly at random from the set of 1–sparse vectors $\Theta = \{\theta \in \{0,1\}^d : \|\theta\|_0 = 1\}$. For simplicity, assume $d = 2^m$ for some $m \in \mathbb{N}$. The action set is taken to be the set of vectors in $\{0,1\}^d$ normalized to be a unit vector in the $L^1$ norm: $\mathcal{A} = \{\frac{x}{\|x\|_1} : x \in \{0,1\}^d\}$. We will show that the expected number of time steps for Thompson sampling (or a UCB algorithm) to identify the optimal action grows linearly with $d$, whereas information directed sampling requires only $\log_2(d)$ time steps.

When an action $a$ is selected and $y = a^T\theta^* \in \{0,1/\|a\|_0\}$ is observed, each $\theta \in \Theta$ with $a^T\theta \neq y$ is ruled out. Let $\Theta_t$ denote the set of all parameters in $\Theta$ that are consistent the rewards observed up to time $t$ and let $I_t = \{i \in \{1, ..., d\} : \theta_i = 1, \theta \in \Theta_t\}$ denote the corresponding set of possible positive components.

For this problem, $A^* = \theta^*$. That is, if $\theta^*$ were known, the optimal action would be to choose the action $\theta^*$. Thompson sampling and UCB algorithms only choose actions from the support of $A^*$ and therefore will only sample actions $a \in \mathcal{A}$ that have only a single positive component. Unless that is also the positive component of $\theta^*$, the algorithm will observe a reward zero and rule out only one possible value for $\theta^*$. In the worst case, the algorithm requires $d$ samples to identify the optimal action.

Now, consider an application of information directed sampling to this problem. The algorithm essentially performs binary search: it selects $a \in \mathcal{A}$ with $a_i > 0$ for half of the components $i \in I_t$
and \( a_i = 0 \) for the other half as well as for any \( i \notin \mathcal{I}_t \). After just \( \log_2(d) \) time steps the true support of the parameter vector \( \theta^* \) is identified.

To see why this is the case, first note that all parameters in \( \Theta_t \) are equally likely and hence the expected reward of an action \( a \) is \( \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} a_i \). Since \( a_i \geq 0 \) and \( \sum_i a_i = 1 \) by construction, every action whose positive components are in \( \mathcal{I}_t \) yields the highest possible expected reward of \( 1/|\mathcal{I}_t| \). Therefore, binary search minimizes expected regret in period \( t \) for this problem. At the same time, binary search is assured to rule out half of the parameter vectors in \( \Theta_t \) at each time \( t \). This is the largest possible expected reduction, and also leads to the largest possible information gain about \( A^* \). Since binary search both minimizes expected regret in period \( t \) and uniquely maximizes expected information gain in period \( t \), it is the sampling strategy followed by \( \pi^{\text{IDS}(\lambda)} \) for any \( \lambda > 0 \) as well as by \( \pi^{\text{IDS}(R)} \).

In this setting we can explicitly calculate the information ratio of each policy, and the difference between them highlights the advantages of information directed sampling. The information ratio of Thompson sampling is

\[
\Psi_1(\pi^{\text{TS}}) = \frac{(d-1)^2/d^2}{\log(d)/d + \frac{d-1}{d} \log\left(\frac{d}{d-1}\right)} \sim \frac{d}{\log(d)}
\]

where \( h(d) \sim f(d) \) if \( h(d)/f(d) \to 1 \) as \( d \to \infty \). The minimal information ratio is

\[
\Psi_1^* = \Psi_1(\pi^{\text{IDS}(R)}) = \frac{(d-1)^2/d^2}{\log(d/2)} \sim \frac{1}{\log(d/2)}.
\]

Especially for large \( d \), \( \Psi_1(\pi^{\text{IDS}(R)}) \) is much smaller than \( \Psi_1(\pi^{\text{TS}}) \).

5.4 Example 3: Recommending products to a customer of unknown type

Consider the problem of repeatedly recommending an assortment of products to a customer. The customer has unknown type \( c^* \in C \) where \( |C| = n \). Each product is geared toward customers of a particular type, and the assortment \( a \in A = C^m \) of \( m \) products offered is characterized by the vector of product types \( a = (c_1, \ldots, c_m) \). We model customer responses through a random utility model in which customers are apriori more likely to derive high value from a product geared toward their type. When offered an assortment of products \( a \), the customer associates with the \( i \)th product utility \( U_{ci}^{(t)}(a) = \beta 1_{\{a_i = c\}} + \xi_{ci}^{(t)} \), where \( \xi_{ci} \) follows an extreme-value distribution and \( \beta \in \mathbb{R} \) is a known constant. This is a standard multinomial logit discrete choice model. The probability a customer of type \( c \) chooses product \( i \) is given by

\[
\frac{\exp\{\beta 1_{\{a_i = c\}}\}}{\sum_{j=1}^m \exp\{\beta 1_{\{a_j = c\}}\}}.
\]

When an assortment \( a \) is offered at time \( t \), the customer makes a choice \( I_t = \text{arg max}_i U_{ci}^{(t)}(a) \) and leaves a review \( U_{ci}^{(t)}(a) \) indicating the utility derived from the product, both of which are observed by the recommendation system. The reward to the recommendation system is the normalized utility of the customer \( \frac{1}{\beta} U_{ci}^{(t)}(a) \).

If the type \( c^* \) of the customer were known, then the optimal recommendation would be \( A^* = (c^*, c^*, \ldots, c^*) \), consisting only of products targeted at the customers type. Therefore, both Thompson sampling and UCB algorithms would only offer assortments consisting of a single type of product. Because of this, each type of algorithm requires order \( n \) samples to learn the customers.
true type. Information directed sampling will instead offer a diverse assortment of products to the customer, allowing it to learn much more quickly.

To make the presentation more transparent, suppose that \( c^* \) is drawn uniformly at random from \( C \) and consider the behavior of each type of algorithm in the limiting case where \( \beta \to \infty \). In this regime, the probability a customer chooses a product of type \( c^* \) if it available tends to 1, and the review \( U_{c^*}^{(t)}(a) \) tends to \( 1_{\{a_t = c^*\}} \), an indicator for whether the chosen product had type \( c^* \). The initial assortment offered by information directed sampling will consist of \( m \) different and previously untested product types. Such an assortment maximizes both the algorithm’s expected reward in the next period and the algorithm’s information gain, since it has the highest probability of containing a product of type \( c^* \). The customer’s response almost perfectly indicates whether one of those items was of type \( c^* \). The algorithm continues offering assortments containing \( m \) unique, untested, product types until a review near \( U_{c^*}^{(t)}(a) \approx 1 \) is received. With extremely high probability, this takes at most \( n/m \) time periods. By diversifying \( m \) products in the assortment, the algorithm learns a factor of \( m \) times faster.

As in the previous example, we can explicitly calculate the information ratio of each policy, and the difference between them highlights the advantages of information directed sampling. The information ratio of Thompson sampling is

\[
\Psi_1(\pi_{TS}) = \frac{(1 - \frac{1}{n})^2}{\log(n) + \frac{n-1}{n} \log \left( \frac{n}{n-m} \right)} \sim \frac{n}{\log(n)}.
\]

The information ratio of information directed sampling is

\[
\Psi_1(\pi_{IDS(R)}) = \frac{(1 - \frac{1}{m})^2}{\frac{m}{n} \log(n) + \frac{n-m}{n} \log \left( \frac{n}{n-m} \right)} \leq \frac{1}{\frac{m}{n} \log(n)} = \frac{n}{m \log(n)},
\]

which is more than \( m \) times smaller.

6 Computation with other measures of divergence

Information directed sampling requires calculating the the mutual information \( g_t(a) = I_t(A^*, Y_t(a)) \) between \( A^* \) and an observation \( Y_t(a) \). For guiding effective computation of \( g_t(a) \) it is useful to investigate its structure. Let \( p_{t,a} = \mathbb{P}(Y_t(a) \in \cdot | F_{t-1}) \) denote the posterior predictive distribution at an action \( a \), and let \( p_{t,a}(\cdot | a^*) = \mathbb{P}(Y_t(a) \in \cdot | F_{t-1}, A^* = a^*) \) denote the posterior predictive distribution conditional on the event that \( a^* \) is the optimal action.

It can be shown that

\[
g_t(a) = \mathbb{E}_{a^* \sim \alpha_t} [D_{KL}(p_{t,a}(\cdot \mid a^*) \mid \mid p_{t,a})].
\]

That is, the mutual information between \( A^* \) and \( Y_t(a) \) is the expected Kullback-Leibler divergence between the posterior predictive distribution \( p_{t,a} \) and the predictive distribution conditioned on the identity of the optimal action \( p_{t,a}(\cdot | a^*) \). Our analysis provides theoretical guarantees for an algorithm that uses a simpler measure of divergence: the squared divergence “in mean” between \( p_{t,a}(\cdot | a^*) \) and \( p_{t,a} \):

\[
(\mathbb{E}[R(Y_t(a)) | F_{t-1}, A^* = a^*] - \mathbb{E}[R(Y_t(a)) | F_{t-1}])^2.
\]

To clarify the exposition, consider the problem of linear optimization under bandit feedback. The action set \( A \) is a finite subset of \( \mathbb{R}^d \). Whenever an action \( a \) is sampled, only the resulting reward
$Y_t(a) = R(Y_t(a))$ is observed. There is an unknown parameter $\theta^* \in \mathbb{R}^d$ such that for each $a \in \mathcal{A}$ the expected reward of $a$ is $a^T \theta^*$. Write $\mu_t = \mathbb{E} [\theta^* | \mathcal{F}_{t-1}]$ and write $\mu_t (\pi^*) = \mathbb{E} [\theta^* | \mathcal{F}_{t-1}, A^* = a^*]$. With this notation, (9) can be rewritten as

$$\left( a^T [\mu_t (\pi^*) - \mu_t] \right)^2 = a^T \left( [\mu_t (\pi^*) - \mu_t][\mu_t (\pi^*) - \mu_t]^T \right) a. $$

Now, define

$$g_t^{ME}(a) = \mathbb{E}_{a^* \sim \alpha_t} \left[ a^T \left( [\mu_t (\pi^*) - \mu_t][\mu_t (\pi^*) - \mu_t]^T \right) a \right],$$

$$L_t = \mathbb{E}_{a^* \sim \alpha_t} [\mu_t (\pi^*) - \mu_t][\mu_t (\pi^*) - \mu_t]^T. $$

One can show that a policy $\pi^* = (\pi_1^*, \pi_2^*, ...)$ chosen so that

$$\pi_t^* \in \arg \min_{\pi \in D(A)} g_t^{ME}(\pi) ,$$

satisfies the regret bound

$$\mathbb{E} \left[ \text{Regret}(T, \pi^*) \right] \leq \sqrt{\frac{dH(\alpha_1)T}{2}}.$$ 

Algorithm [1] presents a simulation based procedure that computes $g_t^{ME}(a)$ and $\Delta_i(a)$ and selects an action according to the distribution $\pi_t^*$. This algorithm requires the ability to generate a large number of samples, denoted by $M \in \mathbb{N}$ in the algorithm, from the posterior distribution of $\theta^*$, which is denoted by $P(\cdot)$ in the algorithm. The actions set $\mathcal{A} = \{ a_1, ..., a_K \}$ is represented by a matrix $A \in \mathbb{R}^{K \times d}$ where the $i$th row of $A$ is the action feature vector $a_i \in \mathbb{R}^d$. The algorithm directly approximates the matrix $L_t$ that appears in equation (11). It does this by sampling parameters from the posterior distribution of $\theta^*$, and, for each action $a$, tracking the number of times $a$ was optimal and the sample average of parameters under which $a$ was optimal. From these samples, it can also compute an estimated vector $R \in \mathbb{R}^K$ of the mean reward from each action and an estimate $p^* \in \mathbb{R}$ of the expected reward from the optimal action $A^*$. Then, as in Proposition [1], the algorithm can compute the optimal distribution by searching over distributions with support at most 2.

### A Proof of Proposition [1]

**Proof.** To reduce notation, we will drop the subscript $t$ throughout this proof. Let $\mathcal{A} = \{ a_1, ..., a_K \}$ and use $\pi_i$ to denote $\pi(a_i)$, $\Delta_i$ to denote $\Delta(a_i)$ and $g_i$ to denote $g(a_i)$. Define

$$\rho(\pi) = \Delta(\pi)^2 - \lambda g(\pi) = \left( \sum_{i=1}^{K} \pi_i \Delta_i \right)^2 - \lambda \sum_{i=1}^{K} \pi_i g_i$$

and consider solving for

$$\pi^* \in \arg \min_{\pi \in D(A)} \rho(\pi).$$

Evaluating the partial derivative of $\rho(\pi)$ with respect to $\pi_i$ at $\pi = \pi^*$ we have

$$\frac{\partial}{\partial \pi_i} \rho(\pi^*) = 2\Delta_i \left( \sum_{j=1}^{K} \pi_j^* \Delta_j \right) - \lambda g_i$$

$$= 2(L^*)_i \Delta_i - \lambda g_i$$

| 11 |
Algorithm 1 Linear Information Directed Sampling

1: **Initialize**: Input $A \in \mathbb{R}^{K \times d}$, $M \in \mathbb{N}$ and posterior distribution $P(\theta)$.
2: for $i \in \{1,..,K\}$ do
3: \hspace{1em} $s^{(i)} \leftarrow 0 \in \mathbb{R}^d$
4: \hspace{1em} $n_i \leftarrow 0 \in \mathbb{R}$
5: end for
6: 
7: **Perform Monte Carlo**: 
8: for $m \in \{1,..,M\}$ do
9: \hspace{1em} Sample $\theta \sim P(\cdot)$
10: \hspace{1em} $I \leftarrow \text{arg max}_i \{(A\theta)_i\}$
11: \hspace{1em} $n_I += 1$
12: \hspace{1em} $s^{(I)} += \theta$
13: end for
14: 
15: **Calculate Problem Data From Monte Carlo Totals**
16: \hspace{1em} $\mu \leftarrow \frac{1}{M} \sum_{i=1}^{K} s^{(i)}$
17: \hspace{1em} $\mu^T \in \mathbb{R}^{d}$
18: \hspace{1em} $\mu^{(i)} \leftarrow s^{(i)}/n_i$
19: \hspace{1em} $\alpha_i \leftarrow n_i/M$
20: end for
21: $L \leftarrow \sum_{i=1}^{K} \alpha_i \left[ \mu^{(i)} - \mu \right] \left[ \mu^{(i)} - \mu \right]^T \in \mathbb{R}^{d\times d}$
22: $\rho^* \leftarrow \sum_{i=1}^{K} \alpha_i \left[ a_i^T \mu^{(i)} \right] \in \mathbb{R}$
23: for $i \in \{1,..,K\}$ do
24: \hspace{1em} $g_i \leftarrow a_i^T La_i \in \mathbb{R}$
25: end for
26: 
27: **Select Action**: 
28: \hspace{1em} $(i^*, j^*, q^*) \leftarrow \text{arg min}_{i,j,q} \left\{ \left[ \rho^* - q R_i - (1 - q) R_j \right]^2 / \left[ q g_i + (1 - q) g_j \right] \right\}$
29: Sample $U \sim \text{Uniform}[0, 1]$
30: if $U < q^*$ then
31: \hspace{2em} Play $i$*
32: else
33: \hspace{2em} Play $j$*
34: end if
35: end if
where $L^* = \sum_j \pi_j^* \Delta_j$ is the expected instantaneous regret of the policy $\pi^*$. Let $d^* = \min_i \frac{\partial}{\partial \pi_i} \rho(\pi^*)$. It must be the case that any $i$ with $\pi_i > 0$ satisfies $d^* = \frac{\partial}{\partial \pi_i} \rho(\pi^*)$, as otherwise transferring probability from action $a_i$ could lead to strictly lower cost. This shows that for each $i$ with $\pi_i^* > 0$, 

$$g_i = \frac{-d^*}{\lambda} + \frac{2L^*}{\lambda} \Delta_i.$$  

(12)

Let $i_1, \ldots, i_m$ be the indices such that $\pi_{i_k}^* > 0$ and 

$$g_{i_1} \geq g_{i_2} \geq \ldots \geq g_{i_m} \quad \Delta_{i_1} \geq \Delta_{i_2} \geq \ldots \geq \Delta_{i_m}.$$ 

Then we can choose a $\beta \in [0, 1]$ so that 

$$\sum_{k=1}^m g_{i_k} = \beta \pi_{i_1}^* + (1 - \beta) \pi_{i_m}^*.$$ 

By equation (12), this implies as well that 

$$\sum_{k=1}^m \Delta_{i_k} = \beta \Delta_{i_1} + (1 - \beta) \Delta_{i_m},$$ 

and hence that the sampling distribution that plays $a_{i_1}$ with probability $\beta$ and $a_{i_m}$ otherwise has the same instantaneous expected regret and the same expected information gain as $\pi^*$. That is, starting with a general sampling distribution $\pi^*$ that maximizes $\rho(\pi)$, we showed there is a sampling distribution with support over at most two actions attains the same objective value and hence that also maximizes $\rho(\pi)$.

Now consider solving for the policy 

$$D_t^{IDS(R)} \in \arg \min_{\pi \in D(A)} \frac{\Delta_t(\pi)^2}{g_t(\pi)}.$$

This is equivalent to solving for 

$$\pi^* \in \arg \min_{\pi \in D(A)} \left\{ \Delta(\pi)^2 - \lambda^*_t g(\pi) \right\}$$

where 

$$\lambda^*_t = \min_{\pi \in D(A)} \frac{\Delta_t(\pi)^2}{g_t(\pi)}$$

is minimal information ratio. Therefore, by the previous part of the proof, there must be a distribution $\pi^*$ with support over at most two actions that minimizes $\Delta_t(\pi)^2/g_t(\pi)$.

\[ \square \]

## B Proof of Proposition 2 and Proposition 3

**Fact 1.** *(Entropy reduction form of mutual information)*

$$I_t(A^*; Y_t(a)) = \mathbb{E} [H(\alpha_t) - H(\alpha_{t+1}) | A_t = a, F_{t-1}]$$

**Lemma 1.** If actions are selected according to a policy $\pi = (\pi_1, \pi_2, \ldots)$,

$$\mathbb{E} \sum_{t=1}^T g_t(\pi_t) \leq H(\alpha_1)$$

**Proof.**

$$\mathbb{E} \sum_{t=1}^T g_t(\pi_t) = \mathbb{E} \sum_{t=1}^T [H(\alpha_t) - H(\alpha_{t+1}) | F_{t-1}] = \mathbb{E} \sum_{t=1}^T (H(\alpha_t) - H(\alpha_{t+1})) = H(\alpha_1) - H(\alpha_{T+1}) \leq H(\alpha_1).$$

\[ \square \]
B.1 Proof of Proposition 2

Lemma 2. If policy $\pi = (\pi_1, \pi_2, ...)$ satisfies

$$\Delta_t(\pi_t) \leq \sqrt{\lambda g_t(\pi_t)} \quad (13)$$

almost surely for each $t \leq T$, then $\mathbb{E}[\text{Regret}(T, \pi)] \leq \sqrt{\lambda H(\alpha_1)T}$.

Proof.

$$\mathbb{E}[\text{Regret}(T, \pi)] = \mathbb{E} \sum_{t=1}^{T} \Delta_t(\pi_t) \leq \sqrt{\lambda} \sum_{t=1}^{T} g_t(\pi_t) \leq \sqrt{\lambda T} \sqrt{\mathbb{E} \sum_{t=1}^{T} g_t(\pi_t)} \leq \sqrt{\lambda H(\alpha_1)T}.$$  \hfill \Box

To conclude the proof of Proposition 2, we show that under the conditions of the proposition both $\pi^{IDS(R)}_t$ and $\pi^{IDS(\lambda)}_t$ satisfy (13).

Proof. Suppose $\Psi^*_t := \min_{\pi \in \mathcal{D}(A)} \frac{\Delta_t(\pi)^2}{g_t(\pi)} \leq \lambda$. Then $\pi^{IDS(R)}_t = \arg \min_{\pi \in \mathcal{D}(A)} \Delta_t^2(g_t(\pi))$ must satisfy

$$\frac{\Delta_t(\pi^{IDS(R)}_t)^2}{g_t(\pi^{IDS(R)}_t)} \leq \lambda \implies \Delta_t(\pi^{IDS(R)}_t) \leq \sqrt{\lambda \sqrt{g_t(\pi^{IDS(R)}_t)}}.$$

Similarly,

$$\Delta_t(\pi^{IDS(\lambda)}_t)^2 - \lambda g_t(\pi^{IDS(\lambda)}_t) \leq \Delta_t(\pi^{IDS(R)}_t)^2 - \lambda g_t(\pi^{IDS(R)}_t) \leq 0,$$

which implies

$$\Delta_t(\pi^{IDS(\lambda)}_t) \leq \sqrt{\lambda \sqrt{g_t(\pi^{IDS(\lambda)}_t)}}.$$

\hfill \Box

B.2 Proof of Proposition 3

Proof. By the definition of $\pi^{IDS(R)}_t$ given in equation (8) and the definition of $\Psi^*_t$ given in equation (9), $\Delta_t(\pi^{IDS(R)}_t)^2 = \lambda^*_t g_t(\pi^{IDS(R)}_t)$. This implies

$$\mathbb{E}[\text{Regret}(T, \pi^{IDS(R)}_t)] = \mathbb{E} \sum_{t=1}^{T} \Delta_t(\pi^{IDS(R)}_t) = \mathbb{E} \sum_{t=1}^{T} \sqrt{\lambda^*_t} \sqrt{g_t(\pi^{IDS(R)}_t)} = \sqrt{\mathbb{E} \sum_{t=1}^{T} \lambda^*_t} \sqrt{\mathbb{E} \sum_{t=1}^{T} g_t(\pi^{IDS(R)}_t)} \leq \sqrt{H(\alpha_1)} \sqrt{\mathbb{E} \sum_{t=1}^{T} \lambda^*_t} \leq \sqrt{\frac{H(\alpha_1)}{T} \mathbb{E} \sum_{t=1}^{T} \lambda^*_t} \leq \sqrt{\lambda H(\alpha_1)T},$$

where (a) follows from Holder’s inequality, and (b) follows from Lemma 1.  \hfill \Box
References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

[2] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013.

[3] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 2013.

[4] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[5] S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, June 2011. Submitted on 21/1/2010.

[6] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

[7] V. Dani, S.M. Kakade, and T.P. Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2007.

[8] V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 355–366, 2008.

[9] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23:1–9, 2010.

[10] A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex bandit problems. *arXiv preprint arXiv:1311.0466*, 2013.

[11] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite time analysis. In *International Conference on Algorithmic Learning Theory*, 2012.

[12] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

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

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

[15] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[16] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *CoRR*, abs/1301.2609, 2013.

[17] D. Russo and B. Van Roy. An information theoretic analysis of thompson sampling. 2014.

[18] N. Srinivas, A. Krause, S.M. Kakade, and M. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *Information Theory, IEEE Transactions on*, 58(5):3250–3265, may 2012. ISSN 0018-9448. doi: 10.1109/TIT.2011.2182033.