Learning to Optimize Via Information-Directed Sampling

Daniel Russo and Benjamin Van Roy

July 22, 2014

Abstract

We propose *information-directed sampling* – a new algorithm for online optimization problems in which a decision-maker must balance between exploration and exploitation while learning from partial feedback. Each action is sampled in a manner that minimizes the ratio between squared expected single-period regret and a measure of information gain: the mutual information between the optimal action and the next observation.

We establish an expected regret bound for information-directed sampling that applies across a very general class of models and scales with the entropy of the optimal action distribution. For the widely studied Bernoulli, Gaussian, and linear bandit problems, we demonstrate simulation performance surpassing popular approaches, including upper confidence bound algorithms, Thompson sampling, and the knowledge gradient algorithm. Further, we present simple analytic examples illustrating that, due to the way it measures information gain, information-directed sampling can dramatically outperform upper confidence bound algorithms and Thompson sampling.

1 Introduction

In the classical multi-armed bandit problem, a decision-maker repeatedly chooses from among a finite set of actions. Each action generates a random reward drawn independently from a probability distribution associated with the action. The decision-maker is uncertain about these reward distributions, but learns about them as rewards are observed. Strong performance requires striking a balance between *exploring* poorly understood actions and *exploiting* previously acquired knowledge to attain high rewards. Because selecting one action generates no information pertinent to other actions, effective algorithms must sample every action many times.

There has been significant interest in addressing problems with more complex *information structures*, in which sampling one action can inform the decision-maker's assessment of other actions. Effective algorithms must take advantage of the information structure to learn more efficiently. Recent work has extended popular algorithms for the classical multi-armed bandit problem, such as *upper confidence bound* (UCB) algorithms and *Thompson sampling*, to address such contexts.

In some cases, such as classical and linear bandit problems, strong performance guarantees have been established for UCB algorithms and Thompson sampling. However, as we will demonstrate through simple analytic examples, these algorithms can perform very poorly when faced with more complex information structures. The shortcoming lies in the fact that these algorithms do not adequately assess the information gain from selecting an action.

In this paper, we propose a new algorithm – *information-directed sampling* (IDS) – that preserves numerous guarantees of Thompson sampling for problems with simple information structures while offering strong performance in the face of more complex problems that daunt alternatives like Thompson sampling or UCB algorithms. IDS quantifies the amount learned by selecting an action

through an information theoretic measure: the mutual information between the true optimal action and the next observation. Each action is sampled in a manner that minimizes the ratio between squared expected single-period regret and this measure of information gain.

As we will show through simple analytic examples, the way in which IDS assesses information gain allows it to dramatically outperform UCB algorithms and Thompson sampling. Further, by leveraging the tools of our recent information theoretic analysis of Thompson sampling [47], we establish an expected regret bound for IDS that applies across a very general class of models and scales with the entropy of the optimal action distribution. We also specialize this bound to several classes of online optimization problems, including problems with full feedback, linear optimization problems with bandit feedback, and combinatorial problems with semi-bandit feedback, in each case establishing that bounds are order optimal up to a poly-logarithmic factor.

We benchmark the performance of IDS through simulations of the widely studied Bernoulli, Gaussian, and linear bandit problems, for which UCB algorithms and Thompson sampling are known to be very effective. We find that even in these settings, IDS outperforms UCB algorithms, Thompson sampling, and the knowledge gradient algorithm. This is particularly surprising for Bernoulli bandit problems, where Thompson sampling and UCB algorithms are known to be asymptotically optimal in the sense proposed by Lai and Robbins [39].

IDS is a stationary randomized policy. *Randomized* because the ratio between expected single-period regret and our measure of information gain can be smaller for a randomized action than for any deterministic action. *Stationary* because action probabilities are selected based only on the current posterior distribution; this is as opposed to UCB algorithms, for example, which selects actions in a manner that depends on the current time period. It is natural to wonder whether randomization plays a fundamental role. Our approach to bounding expected regret can be applied to other policies and scales with a policy-dependent statistic we call the *information ratio*. We establish that randomization is essential to our results because the information ratio can become arbitrarily large for any stationary deterministic policy.

IDS solves a single-period optimization problem as a proxy to an intractable multi-period problem. Solution of this single-period problem can itself be computationally demanding, especially in cases where the number of actions is enormous or mutual information is difficult to evaluate. To carry out computational experiments, we develop numerical methods for particular classes of online optimization problems. We also propose mean-based IDS – an approximate form of IDS that is suitable for some problems with bandit feedback, satisfies our regret bounds for such problems, employs an alternative information measure, and can sometimes facilitate design of more efficient numerical methods. More broadly, we hope that our development and analysis of IDS facilitate the future design of efficient algorithms that capture its benefits.

It is worth noting that the problem formulation we work with, which is presented in Section 3, is very general, encompassing not only problems with bandit feedback, but also a broad array of information structures for which observations can offer information about rewards of arbitrary subsets of actions or factors that influence these rewards. Because IDS and our analysis accommodate this level of generality, they can be specialized to problems that in the past have been studied individually, such as those involving pricing and assortment optimization (see, e.g., [13, 45, 49]), though in each case, developing an efficient version of IDS may require innovation.

2 Literature review

UCB algorithms are the primary approach considered in the segment of the stochastic multi-armed bandit literature that treats problems with dependent arms. UCB algorithms have been applied

to problems where the mapping from action to expected reward is a linear [1, 20, 44], generalized linear [21], or sparse linear [2] model; is sampled from a Gaussian process [51] or has small norm in a reproducing kernel Hilbert space [51, 53]; or is a smooth (e.g. Lipschitz continuous) model [15, 35, 52]. Recently, an algorithm known as Thompson sampling has received a great deal of interest. Agrawal and Goyal [6] provided the first analysis for linear contextual bandit problems. Russo and Van Roy [46] consider a more general class of models, and show that standard analysis of upper confidence bound algorithms leads to bounds on the expected regret of Thompson sampling. Very recent work of Gopalan et al. [27] provides asymptotic frequentist bounds on the growth rate of regret for problems with dependent arms. Both UCB algorithms and Thompson sampling have been applied to other types of problems, like reinforcement learning [31, 42] and Monte Carlo tree search [10, 36]. We will describe both UCB algorithms and Thompson sampling in more detail in Section 8.

In one of the first papers on multi-armed bandit problems with dependent arms, Agrawal et al. [3] consider a general model in which the reward distribution associated with each action depends on a common unknown parameter. When the parameter space is finite, they provide a lower bound on the asymptotic growth rate of the regret of any admisible policy as the time horizon tends to infinity and show that this bound is attainable. These results were later extended by Agrawal et al. [4] and Graves and Lai [28] to apply to the adaptive control of Markov chains and to problems with infinite parameter spaces. These papers provide results of fundemental importance, but seem to have been overlooked by much of the recent literature.

Two other papers [30, 54] have used the mutual information between the optimal action and the next observation to guide action selection. Each focuses on optimization of expensive-to-evaluate black-box functions. Here, *black–box* indicates the absence of strong structural assumptions such as convexity and that the algorithm only has access to function evaluations, while *expensive-to-evaluate* indicates that the cost of evaluation warrants investing considerable effort to determine where to evaluate. These papers focus on settings with low-dimensional continuous action spaces, and with a Gaussian process prior over the objective function, reflecting the belief that "smoother" objective functions are more plausible than others. This approach is often called "Bayesian optimization" in the machine learning community [12]. Both Villemonteix et al. [54] and Hennig and Schuler [30] propose selecting each sample to maximize the mutual information between the next observation and the optimal solution.

Several features distinguish our work from that of Villemonteix et al. [54] and Hennig and Schuler [30]. First, these papers focus on pure exploration problems: the objective is simply to learn about the optimal solution – not to attain high cumulative reward. Second, and more importantly, they focus only on problems with Gaussian process priors and continuous action spaces. For such problems, simpler approaches like UCB algorithms [51], probability of improvement [37], and expected improvement [40] are already extremely effective. As noted by Brochu et al. [12], each of these algorithms simply chooses points with "*potentially* high values of the objective function: whether because the prediction is high, the uncertainty is great, or both." By contrast, a major motivation of our work is that a richer information measure is needed to address problems with more complicated information structures. Finally, we provide a variety of general theoretical guarantees for information-directed sampling, whereas Villemonteix et al. [54] and Hennig and Schuler [30] propose their algorithms as heuristics without guarantees. Appendix A.1 shows that our theoretical guarantees extend to pure exploration problems.

The knowledge gradient (KG) algorithm uses a different measure of information to guide action selection: the algorithm computes the impact of a single observation on the quality of the decision made by a *greedy* algorithm, which simply selects the action with highest posterior expected reward. This measure was proposed by Mockus et al. [40] and studied further by Frazier et al. [23] and

Ryzhov et al. [48]. KG seems natural since it explicitly seeks information that improves decision quality. Computational studies suggest that for problems with Gaussian priors, Gaussian rewards, and relatively short time horizons, KG performs very well. However, even in some simple settings, KG may not converge to optimality. In fact, it may select a suboptimal action in *every* period, even as the time horizon tends to infinity. IDS also measures the information provided by a single observation, but our results imply it converges to optimality. In Appendix C, we define KG more formally, and provide some insight into why it can fail to converge to optimality.

Our work also connects to a much larger literature on Bayesian experimental design (see [17] for a review). Recent work has demonstrated the effectiveness of *greedy* or *myopic* policies that always maximize a measure of the information gain from the next sample. Jedynak et al. [32] and Waeber et al. [55] consider problem settings in which this greedy policy is optimal. Another recent line of work [25, 26] shows that measures of information gain sometimes satisfy a decreasing returns property known as adaptive sub-modularity, implying the greedy policy is competitive with the optimal policy. Our algorithm also only considers the information gain due to the *next sample*, even though the goal is to acquire information over many periods. Our results establish that the manner in which IDS encourages information gain leads to an effective algorithm, even for the different objective of maximizing cumulative reward.

3 Problem formulation

We consider a general probabilistic, or Bayesian, formulation in which uncertain quantities are modeled as random variables. The decision–maker sequentially chooses actions $(A_t)_{t\in\mathbb{N}}$ from a finite action set $\mathcal{A}$ and observes the corresponding outcomes $(Y_t(A_t))_{t\in\mathbb{N}}$. 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 $a_1, .., a_t \in \mathcal{A}$, $\mathbb{P}\left(Y_1(a_1) \in \tilde{\mathcal{Y}}_1, ..., Y_t(a_t) \in \tilde{\mathcal{Y}}_t | p^*\right) = \prod_{k=1}^{t} p^*_{a_k}(\tilde{\mathcal{Y}}_k)$. The true outcome distribution p^* is itself randomly drawn from the family $\mathcal{P}$ of distributions.

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,

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

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

$$\mathbb{E}\left[\text{Regret}(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left[R(Y_t(A^*)) - R(Y_t(A_t))\right]\right], \tag{2}$$

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*.

Filtrations and randomized policies. 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 $(\xi_t)_{t\in\mathbb{N}}$ that are jointly independent of the outcomes $\{Y_t(a)\}_{t\in\mathbb{N},a\in\mathcal{A}}$ and p^*. We define all random variables with respect to a probability space $(\Omega,\mathcal{F},\mathbb{P})$. Fix the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ where $\mathcal{F}_{t-1}\subset\mathcal{F}$ is the sigma–algebra generated by $(\xi_1,A_1,Y_1(A_1),...,\xi_{t-1},A_{t-1},Y_{t-1}(A_{t-1}))$. The action A_t is measurable with respect to the sigma–algebra generated by $(\mathcal{F}_{t-1},\xi_t)$. That is, given the history of past observations, A_t is random only through its dependence on ξ_t.

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* π, 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 π, letting $\mathbb{E}\left[\text{Regret}(T,\pi)\right]$ denote the expected value given by (2) when the actions $(A_1,..,A_T)$ are chosen according to π.

Further notation. We set $\alpha_t(a)=\mathbb{P}\left(A^*=a|\mathcal{F}_{t-1}\right)$ 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

$$D_{\text{KL}}(P||Q)=\int_{\mathcal{Y}}\log\left(\frac{dP}{dQ}\right)dP \tag{3}$$

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 $H(P)=-\sum_{x\in\mathcal{X}}P(x)\log\left(P(x)\right)$. The *mutual information* under the posterior distribution between two random variables $X_1:\Omega\to\mathcal{X}_1$, and $X_2:\Omega\to\mathcal{X}_2$, denoted by

$$I_t(X_1;X_2):=D_{\text{KL}}\left(\mathbb{P}\left((X_1,X_2)\in\cdot|\mathcal{F}_{t-1}\right)\;||\;\mathbb{P}\left(X_1\in\cdot|\mathcal{F}_{t-1}\right)\mathbb{P}\left(X_2\in\cdot|\mathcal{F}_{t-1}\right)\right), \tag{4}$$

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 $\mathbb{P}\left(\cdot|\mathcal{F}_{t-1}\right)$.

To reduce notation, we define the *information gain* from an action a to be $g_t(a):=I_t(A^*;Y_t(a))$. As shown for example in Lemma 5.5.6 of Gray [29], this is equal to the expected reduction in entropy of the posterior distribution of A^* due to observing $Y_t(a)$:

$$g_t(a)=\mathbb{E}\left[H(\alpha_t)-H(\alpha_{t+1})|\mathcal{F}_{t-1},A_t=a\right], \tag{5}$$

which plays a crucial role in our results. Let $\Delta_t(a):=\mathbb{E}\left[R_t(Y_t(A^*))-R(Y_t(a))|\mathcal{F}_{t-1}\right]$ denote the expected instantaneous regret of action a at time t.

We use overloaded notation for $g_t(\cdot)$ and $\Delta_t(\cdot)$. For an action sampling distribution $\pi\in\mathcal{D}(\mathcal{A})$, $g_t(\pi):=\sum_{a\in\mathcal{A}}\pi(a)g_t(a)$ denotes the expected information gain when actions are selected according to π_t, and $\Delta_t(\pi)=\sum_{a\in\mathcal{A}}\pi(a)\Delta_t(a)$ is defined analogously.

4 Information-directed sampling

IDS explicitly balances between having low expected regret in the current period and acquiring new information about which action is optimal. It does this by minimizing over all action sampling distributions $\pi\in\mathcal{D}(\mathcal{A})$ the ratio between the square of expected regret $\Delta_t(\pi)^2$ and information

gain $g_t(\pi)$ about the optimal action A^*. In particular, the policy $\pi^{\text{IDS}} = \left(\pi_1^{\text{IDS}}, \pi_2^{\text{IDS}}, ...\right)$ is defined by:

$$\pi_t^{\text{IDS}} \in \arg\min_{\pi \in \mathcal{D}(\mathcal{A})} \left\{ \Psi_t(\pi) := \frac{\Delta_t(\pi)^2}{g_t(\pi)} \right\}. \tag{6}$$

We call $\Psi_t(\pi)$ the *information ratio* of an action sampling distribution π and $\Psi_t^* = \min_\pi \Psi_t(\pi) = \Psi_t(\pi_t^{\text{IDS}})$ the *minimal information ratio*. Our analysis will provide a number of bounds on the minimal information ratio, which is key to information-directed sampling's theoretical guarantees. For example, we show that when rewards are bounded in $[0, 1]$, $\Psi_t^* \leq |\mathcal{A}|/2$. Bounds of this form show that, in any period, the algorithm's regret can only be large if it's expected to acquire a lot of information about which action is optimal. Equivalently, it shows that the "cost" per bit of information acquired cannot be too large.

Note that the solution to (6) may be a randomized policy, and we will show in Section 7 that this is essential to our results. However, as we will show in the next subsection, there is always an action sampling distribution minimizing (6) that has support over at most two actions.

4.1 Optimization

We now investigate the structure of the optimization problem (6) and present an efficient algorithm for solving it. Suppose that there are $K = |\mathcal{A}|$ actions, and that the posterior expected regret and information gain are stored in the vectors $\Delta \in \mathbb{R}_+^K$ and $g \in \mathbb{R}_+^K$. This subsection treats Δ and g as known, but otherwise arbitrary non-negative vectors. We will discuss methods for computing these quantities in later sections. Assume $g \neq 0$, as otherwise the optimal action is known with certainty. We will focus on the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \Psi(\pi) := \frac{\left(\pi^T \Delta\right)^2}{\pi^T g} \\ \text{subject to} \quad & \pi^T e = 1 \\ & \pi \geq 0 \end{aligned} \tag{7}$$

Here $e = (1, 1, ..., 1) \in \mathbb{R}^K$ and the set of feasible solutions is the probability simplex in $\mathbb{R}^K$.

The following result establishes that (7) is a convex optimization problem, and surprisingly, has an optimal solution with at most two non-zero components. Therefore, while IDS is a randomized policy, it randomizes over at most two actions.

Proposition 1. *The function* $\Psi : \pi \mapsto \left(\pi^T \Delta\right)^2 / \pi^T g$ *is convex on* $\left\{\pi \in \mathbb{R}^K | \pi^T g > 0\right\}$. *Moreover, there is an optimal solution* π^* *to* (7) *with* $|\{i : \pi_i^* > 0\}| \leq 2$.

Algorithm 1 uses Proposition 1 to choose an action in a manner that minimizes (6). For a problem with $|\mathcal{A}| = K$ actions, the algorithm requires inputs $\Delta \in \mathbb{R}_+^K$ and $g \in \mathbb{R}_+^K$ specifying respectively the expected regret and information gain of each action. The sampling distribution that minimizes (6) is computed by looping over all pairs of actions $(i, j) \in \mathcal{A} \times \mathcal{A}$ and finding the optimal probability of playing i instead of j. Finding this probability is particularly efficient because the objective function is convex. Golden section search, for example, provides a very efficient method for optimizing a convex function over $[0, 1]$. In addition, in this case, any solution to $\frac{d}{dq'} \left[q'\Delta_i + (1-q)\Delta_j\right]^2 / \left[q' g_i + (1-q') g_j\right] = 0$ is given by the solution to a quadratic equation, and therefore can be expressed in closed form.

Algorithm 1 chooseAction($\Delta \in \mathbb{R}_+^K$, $g \in \mathbb{R}_+^K$)

Initialize opt $\leftarrow \infty$
Calculate Optimal Sampling Distribution
for $j \in \{1, .., K-1\}$ **do**
    for $i \in \{j+1, .., K\}$ **do**
        $q \leftarrow \arg\min_{q' \in [0,1]} \left[q'\Delta_i + (1-q)\Delta_j\right]^2 / \left[q'g_i + (1-q')g_j\right]$
        objectiveValue $\leftarrow \left[q\Delta_i + (1-q)\Delta_j\right]^2 / \left[qg_i + (1-q)g_j\right]$
        if objectiveValue $<$ opt **then**
            $(i^*, j^*, q^*) \leftarrow (i, j, q)$
            opt $\leftarrow$ objectiveValue
        end if
    end for
end for

Select Action:
Sample $U \sim$ Uniform $([0,1])$
if $U < q^*$ **then**
    Play i^*
else
    Play j^*
end if

4.2 Mean-based information-directed sampling

Here we introduce an approximate form of IDS that is suitable for some problems with bandit feedback, satisfies our regret bounds for such problems, and can sometimes facilitate design of more efficient numerical methods. We will derive this policy by investigating the structure of the mutual information $g_t(a) = I_t\left(A^*, Y_t(a)\right)$, and considering a modified information measure.

Let $p_{t,a} = \mathbb{P}\left(Y_t(a) \in \cdot | \mathcal{F}_{t-1}\right)$ denote the posterior predictive distribution at an action a, and let $p_{t,a}\left(\cdot|a^*\right) = \mathbb{P}\left(Y_t(a) \in \cdot | \mathcal{F}_{t-1}, A^* = a^*\right)$ denote the posterior predictive distribution conditional on the event that a^* is the optimal action. Crucial to our results is the following fact, which is a consequence of standard properties of mutual information[1]:

$$g_t(a) = \mathop{\mathbb{E}}_{a^* \sim \alpha_t}\left[D_{\mathrm{KL}}\left(p_{t,a}(\cdot|a^*) \,||\, p_{t,a}\right)\right]. \tag{8}$$

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 $p_{t,a}$,

$$\begin{aligned} D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a^*) \,||\, p_{t,a}\right)^2 &:= \left(\mathop{\mathbb{E}}_{y \sim p_{t,a}(\cdot|a^*)}\left[R(y)\right] - \mathop{\mathbb{E}}_{y \sim p_{t,a}}\left[R(y)\right]\right)^2 \\ &= \left(\mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}, A^* = a^*\right] - \mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}\right]\right)^2. \end{aligned}$$

[1]Some details related to the derivation of this fact when $Y_t(a)$ is a general random variable can be found in the appendix of Russo and Van Roy [47].

Define

$$g_t^{\text{ME}}(a) = \mathop{\mathbb{E}}_{a^* \sim \alpha_t} \left[D_{\text{ME}} \left(p_{t,a}(\cdot|a^*) \,||\, p_{t,a} \right)^2 \right],$$

which replaces the Kullback-Leibler divergence in (8) with squared divergence in mean. We introduce the policy $\pi^{\text{IDS}_{\text{ME}}} = \left(\pi_1^{\text{IDS}_{\text{ME}}}, \pi_2^{\text{IDS}_{\text{ME}}}, ...\right)$ where

$$\pi_t^{\text{IDS}_{\text{ME}}} \in \mathop{\arg\min}_{\pi \in \mathcal{D}(\mathcal{A})} \frac{\Delta_t(\pi)^2}{g_t^{\text{ME}}(\pi)}. \tag{9}$$

5 General regret bounds

This section establishes regret bounds for information-directed sampling that scale with the entropy of the optimal action distribution. Recall that we have defined the information ratio of an action sampling distribution to be $\Psi_t(\pi) := \Delta_t(\pi)^2 / g_t(\pi)$. The *minimal* information ratio is

$$\Psi_t^* := \min_{\pi \in \mathcal{D}(\mathcal{A})} \frac{\Delta_t(\pi)^2}{g_t(\pi)} = \frac{\Delta_t(\pi_t^{\text{IDS}})^2}{g_t(\pi_t^{\text{IDS}})},$$

which is the smallest possible ratio between squared expected regret and expected information gain. The next proposition shows that bounds on a policy's information ratio imply bounds on expected regret.

Proposition 2. *Fix a deterministic $\lambda \in \mathbb{R}$ and a policy $\pi = (\pi_1, \pi_2, ...)$ such that $\Psi_t(\pi_t) \leq \lambda$ almost surely for each $t \in \{1, .., T\}$. Then,*

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

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 later in this section. The next proposition establishes regret bounds that depend on the average value of Ψ_t^* instead of a worst case bound.

Proposition 3. *For any $T \in \mathbb{N}$,*

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

6 Bounds on the minimal information ratio

We now establish upper bounds on the minimal information ratio Ψ_t^* in several important settings, which yields explicit regret bounds when combined with Proposition 2. These bounds show that, in any period, the algorithm's expected regret can only be large if it's expected to acquire a lot of information about which action is optimal. In this sense, it effectively balances between exploration and exploitation in *every* period.

The bounds on the minimal information ratio Ψ_t^* also help to clarify the role it 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 IDS is able to automatically exploit this structure.

The proofs of 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. In particular, since $\Psi_t^* \leq \Psi_t(\pi^{\text{TS}})$ where π^{TS} is the Thompson sampling policy, it is enough to bound $\Psi_t(\pi^{\text{TS}})$. Several such bounds were provided in Russo and Van Roy [47].[2] While the analysis is similar in the cases considered here, IDS outperforms Thompson sampling in simulation, and, as we will highlight in the next section, is sometimes provably much more informationally efficient.

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 $\inf_\pi \mathbb{E}\left[\text{Regret}(T, \pi)\right]$ in our framework.

To simplify the exposition, our results are stated under the assumption that rewards are uniformly bounded. This effectively controls the worst-case variance of the reward distribution, and as shown in the appendix of Russo and Van Roy [47], our results can be extended to the case where reward distributions are sub-Gaussian.

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

6.1 Worst case bound

The next proposition shows that Ψ_t^* is never larger than $|\mathcal{A}|/2$. That is, there is always an action sampling distribution $\pi \in \mathcal{D}(\mathcal{A})$ such that $\Delta_t(\pi)^2 \leq (|\mathcal{A}|/2) g_t(\pi)$. 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 $t \in \mathbb{N}$, $\Psi_t^* \leq \Psi_t\left(\pi_t^{\text{IDS}_{\text{ME}}}\right) \leq |\mathcal{A}|/2$ almost surely.*

Combining Proposition 4 with Proposition 2 shows that $\mathbb{E}\left[\text{Regret}\left(T, \pi^{\text{IDS}}\right)\right] \leq \sqrt{\frac{1}{2}|\mathcal{A}| H(\alpha_1) T}$ and $\mathbb{E}\left[\text{Regret}\left(T, \pi^{\text{IDS}_{\text{ME}}}\right)\right] \leq \sqrt{\frac{1}{2}|\mathcal{A}| H(\alpha_1) T}$.

6.2 Full information

Our focus in this paper is on problems with *partial feedback*. For such problems, what the decision maker observes depends on the actions selected, which leads to a tension between exploration and exploitation. Problems with full information arise as an extreme point of our formulation where the outcome $Y_t(a)$ is perfectly revealed by observing $Y_t(\tilde{a})$ for some $\tilde{a} \neq a$; what is learned does not depend on the selected action. The next proposition shows that under full information, the minimal information ratio is bounded by 1/2.

Proposition 5. *Suppose for each $t \in \mathbb{N}$ there is a random variable $Z_t : \Omega \to \mathcal{Z}$ such that for each $a \in \mathcal{A}$, $Y_t(a) = (a, Z_t)$. Then for all $t \in \mathbb{N}$, $\Psi_t^* \leq \frac{1}{2}$ almost surely.*

[2] $\Psi_t(\pi^{\text{TS}})$ is exactly equal to the term Γ_t^2 that is bounded in Russo and Van Roy [47].

Combining this result with Proposition 2 shows $\mathbb{E}\left[\text{Regret}(T, \pi^{\text{IDS}})\right] \leq \sqrt{\frac{1}{2}H(\alpha_1)T}$. Further, a worst–case bound on the entropy of α_1 shows that $\mathbb{E}\left[\text{Regret}(T, \pi^{\text{IDS}})\right] \leq \sqrt{\frac{1}{2}\log(|\mathcal{A}|)T}$. Dani et al. [19] show this bound is order optimal, in the sense that for any time horizon T and number of actions $|\mathcal{A}|$ there exists a prior distribution over p^* under which $\inf_\pi \mathbb{E}\left[\text{Regret}(T, \pi)\right] \geq c_0\sqrt{\log(|\mathcal{A}|)T}$ where c_0 is a numerical constant that does not depend on $|\mathcal{A}|$ or 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.

6.3 Linear optimization under bandit feedback

The stochastic linear bandit problem has been widely studied (e.g. [1, 20, 44]) 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 $a \in \mathcal{A}$*

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

then for all $t \in \mathbb{N}$, $\Psi_t^* \leq \Psi_t\left(\pi_t^{\text{IDS}_{\text{ME}}}\right) \leq d/2$ almost surely.

This result shows that $\mathbb{E}\left[\text{Regret}(T, \pi^{\text{IDS}})\right] \leq \sqrt{\frac{1}{2}H(\alpha_1)dT} \leq \sqrt{\frac{1}{2}\log(|\mathcal{A}|)dT}$ for linear bandit problems. Again, Dani et al. [19] 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 $\inf_\pi \mathbb{E}\left[\text{Regret}(T, \pi)\right] \geq c_0\sqrt{\log(|\mathcal{A}|)dT}$ where c_0 is a constant the 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 in informative.

6.4 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 $a \in \mathcal{A} \subset \{a' \subset \{0, 1, ..., d\} : |a'| \leq m\}$ is $m^{-1}\sum_{i\in\mathcal{A}}\theta_{t,i}$. 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 [8].

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.

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

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

Assume that the random variables $\{\theta_{t,i} : i \in \{1, ..., d\}\}$ *are independent conditioned on* $\mathcal{F}_{t-1}$ *and* $\theta_{t,i} \in [\frac{-1}{2}, \frac{1}{2}]$ *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 $\binom{d}{m}$ actions, but because IDS 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 $\mathbb{E}\left[\text{Regret}(T, \pi^{\text{IDS}})\right] \leq \frac{1}{m}\sqrt{\frac{d}{2}H(\alpha_1)T}$. Since $H(\alpha_1) \leq \log|\mathcal{A}| = O(m\log(\frac{d}{m}))$ this also yields a bound of order $\sqrt{\frac{d}{m}\log\left(\frac{d}{m}\right)T}$. As shown by Audibert et al. [8], the lower bound[3] 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.

7 The need for randomization

IDS is a *randomized policy*, and it's natural to wonder whether this randomization plays a fundamental role. Here, we will show through a simple example that no non-randomized policy can have a uniformly bounded information ratio. For this reason, randomization is essential for our theoretical results.

IDS is also a *stationary policy*, because action probabilities are selected based only on the current posterior distribution and not the current time period. More broadly, randomization seems to be crucial to the algorithm's convergence, as most stationary deterministic policies would eventually settle on playing only a single action.

Example 1. *Consider a problem with two actions* $\mathcal{A} = \{a_1, a_2\}$*. Rewards from* a_1 *follow a Bernoulli distribution with mean 1/2 and this is known apriori. The distribution of rewards from* a_2 *is Bernoulli(3/4) with probability* p *and is Bernoulli(1/4) with probability* $1 - p$*. Consider the limit as* $p \to 0$*. Then, it is almost certain that action 1 is optimal. The expected regret from choosing action 1 is therefore almost zero, but the information gain from choosing action 1 is exactly zero. Meanwhile, the regret from choosing action 2 is almost 1/4, but the information gain is very small (certainly no larger than* $H(\alpha_1)$*). Hence, for neither choice of action allows for a bounded information ratio* $\Psi_t(a)$*. Nevertheless, by mixing between the two actions IDS is able to attain low expected regret, positive expected information gain, and a bounded ratio between the two.*

8 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 [5, 16, 33, 38, 39] and satisfy strong regret bounds for some problems with dependent arms [15, 20, 21, 27, 44, 46, 51].

[3] 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 [14]), 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 $m\sqrt{\frac{d}{m}T} = \sqrt{mdT}$.

Both classes of algorithms experiment with all actions that could still *plausibly* be optimal given the observed data. This guarantees actions are not prematurely discarded, but also that samples are not "wasted" on clearly suboptimal actions. While this is enough to guarantee strong performance is some settings, we will show that these algorithms can perform very poorly when faced with more complex information structures. We demonstrate this through several examples - each of which is designed to be simple and transparent. To set the stage for our discussion, we now introduce UCB algorithms and Thompson sampling.

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 $\pi_t^{\text{TS}} = \alpha_t$. By definition, this means that for each $a \in \mathcal{A}$, $\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) = \mathbb{P}(A^* = a|\mathcal{F}_{t-1}) = \alpha_t(a)$. This algorithm is sometimes called *probability matching* because the action selection distribution is *matched* to the posterior distribution of the optimal action. Note that Thompson sampling draws actions only from the support of the posterior distribution of A^*. That is, it never selects an action a if $\mathbb{P}(A^* = a) = 0$. Put differently, this implies that it only selects actions that are optimal under some $p \in \mathcal{P}$.

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 $a \in \mathcal{A}$ an upper confidence bound $B_t(a)$ is constructed. Then, the algorithm selects an action $A_t \in \arg\max_{a \in \mathcal{A}} B_t(a)$ with maximal upper confidence bound. The upper confidence bound $B_t(a)$ represents the greatest mean reward value that is statistically plausible. In particular, $B_t(a)$ is typically constructed so that

$$\mathop{\mathbb{E}}_{y \sim p_a^*}[R(y)] \leq B_t(a)$$

with high probability, but that $B_t(a) \to \mathop{\mathbb{E}}_{y \sim p_a^*}[R(y)]$ as data about action a accumulates.

Like Thompson sampling, many UCB algorithms only select actions that are optimal under some $p \in \mathcal{P}$. For example, consider a UCB algorithm that constructs at each time t a confidence set $\mathcal{P}_t \subset \mathcal{P}$ containing the set of distributions that are statistically plausible given observed data. Upper confidence bounds are defined to be

$$B_t(a) = \max_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{y \sim p_a}[R(y)],$$

which is the highest expected reward attainable under one of the plausible distributions. Many optimistic algorithms take this form (see for example [20, 21, 31]). Any action $A_t \in \arg\max_{a \in \mathcal{A}} B_t(a)$ must be optimal under one of the outcome distributions $p \in \mathcal{P}_t$. An alternative method involves choosing $B_t(a)$ to be a particular quantile of the posterior distribution of the action's mean reward under p^* [34]. In each of the examples we construct, such an algorithm always chooses actions from the support of A^* unless the quantiles are so low that $\max_{a \in \mathcal{A}} B_t(a) < \mathbb{E}[R(Y_t(A^*))]$.

8.1 Example: 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

$$R(Y_t(a_j)) = \begin{cases} 1 & j = i \\ 0 & j \neq i,\, j \neq 0 \\ \frac{1}{2i} & j = 0 \end{cases}$$

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 $\{a_1, ..., a_K\}$, 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 K.

Information-directed sampling 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 immediately identifies the optimal action. IDS will select this action, learn which action is optimal, and select that action in all future periods. Its regret is independent of K.

8.2 Example: 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 $a^T\theta^*$. The true parameter θ^* is known to be 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: $\mathcal{A} = \left\{\frac{x}{\|x\|_1} : x \in \{0,1\}^d, x \neq 0\right\}$. 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 IDS 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 Θ_t denote the set of all parameters in Θ that are consistent with the rewards observed up to time t and let $\mathcal{I}_t = \{i \in \{1, ..., d\} : \theta_i = 1, \theta \in \Theta_t\}$ denote the corresponding set of possible positive components.

For this problem, $A^* = \theta^*$. That is, if θ^* were known, choosing the action θ^* would yield the highest possible reward. Thompson sampling and UCB algorithms only choose actions from the support of A^* and therefore will only sample actions $a \in \mathcal{A}$ that, like A^*, have only a single positive component. Unless that is also the positive component of θ^*, the algorithm will observe a reward of zero and rule out only one possible value for θ^*. 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 $i \in \mathcal{I}_t$ 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 θ^* is identified.

To see why this is the case, first note that all parameters in Θ_t are equally likely and hence the expected reward of an action a is $\frac{1}{|\mathcal{I}_t|}\sum_{i \in I_t} a_i$. Since $a_i \geq 0$ and $\sum_i a_i = 1$ for each $a \in \mathcal{A}$, every action whose positive components are in I_t yields the highest possible expected reward of $1/|I_t|$. 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 Θ_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 IDS.

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. We have

$$\Psi_1(\pi_1^{\text{TS}}) = \frac{(d-1)^2/d^2}{\frac{\log(d)}{d} + \frac{d-1}{d}\log\left(\frac{d}{d-1}\right)} \sim \frac{d}{\log(d)} \qquad \Psi_1^* = \Psi_1(\pi_1^{\text{IDS}}) = \frac{(d-1)^2/d^2}{\log(d/2)} \sim \frac{1}{\log(d)}$$

where $h(d) \sim f(d)$ if $h(d)/f(d) \to 1$ as $d \to \infty$. When the dimension d is large, $\Psi_1(\pi^{\text{IDS}})$ is much smaller.

8.3 Example: 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 $a \in \mathcal{A} = C^m$ of m products offered is characterized by the vector of product types $a = (c_1, .., c_m)$. 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 ith product utility $U_{ci}^{(t)}(a) = \beta \mathbf{1}_{\{a_i = c\}} + W_{ci}^{(t)}$, where W_{ci}^t 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

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

When an assortment a is offered at time t, the customer makes a choice $I_t = \arg\max_i U_{ci}^{(t)}(a)$ and leaves a review $U_{cI_t}^{(t)}(a)$ 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 $(\frac{1}{\beta})U_{cI_t}^{(t)}(a)$.

If the type c^* of the customer were known, then the optimal recommendation would be $A^* = (c^*, c^*, ..., c^*)$, which consists only of products targeted at the customer's 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 customer's true type. IDS 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 $\mathcal{C}$ 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 $U_{cI_t}^{(t)}(a)$ tends to $\mathbf{1}_{\{a_{I_t}=c^*\}}$, an indicator for whether the chosen product had type c^*. The initial assortment offered by IDS 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 $U_{cI_t}^{(t)}(a) \approx 1$ is received. With extremely high probability, this takes at most $\lceil n/m \rceil$ time periods. By diversifying the 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 IDS. The information ratio of IDS is more

that m times smaller:

$$\Psi_1(\pi_1^{\text{TS}}) = \frac{\left(1-\frac{1}{n}\right)^2}{\frac{\log(n)}{n} + \frac{n-1}{n}\log\left(\frac{n}{n-1}\right)} \sim \frac{n}{\log(n)} \qquad \Psi_1(\pi_1^{\text{IDS}}) = \frac{\left(1-\frac{1}{m}\right)^2}{\frac{m}{n}\log(n) + \frac{n-m}{n}\log\left(\frac{n}{n-m}\right)} \leq \frac{n}{m\log(n)}.$$

9 Computational experiments

This section presents computational experiments evaluating the effectiveness of information-directed sampling, and comparing its performance to other algorithms. In Section 8, we showed that popular approaches like UCB algorithms and Thompson sampling can perform very poorly when faced with complicated information structures, and for this reason are sometimes dramatically outperformed by IDS. In this section, we focus instead on simpler settings where current approaches are extremely effective. We find that even for these simple and widely studied settings, information-directed sampling displays performance exceeding state of the art. For each experiment, the algorithm used to implement IDS is presented in Appendix B.

9.1 Beta-Bernoulli experiment

Our first experiment involves a multi-armed bandit problem with independent arms and binary rewards. The mean reward of each arm is drawn from $\text{Beta}(1,1)$, which is the uniform distribution, and the means of separate arms are independent. Figure 1a and Table 1 present the results of 1000 independent trials of an experiment with 10 arms and a time horizon of 1000. We compared the performance of IDS to that of six other algorithms, and found that it had the lowest average regret of 18.16.

The famous UCB1 algorithm of Auer et al. [9] selects the action a which maximizes the upper confidence bound $\hat{\theta}_t(a) + \sqrt{\frac{2\log(t)}{N_t(a)}}$ where $\hat{\theta}_t(a)$ is the empirical average reward from samples of action a and $N_t(a)$ is the number of samples of action a up to time t. The average regret of this algorithm is 131.3, which is dramatically larger than that of IDS. For this reason UCB1 is omitted from Figure 1a.

The confidence bounds of UCB1 are constructed to facilitate theoretical analysis. For practical performance Auer et al. [9] proposed using an algorithm called UCB-Tuned. This algorithm selects the action a which maximizes the upper confidence bound $\hat{\theta}_t(a) + \sqrt{\frac{\min\{1/4\,,\,\overline{V}_t(a)\}\log(t)}{N_t(a)}}$, where $\overline{V}_t(a)$ is an upper bound on the variance of the reward distribution at action a. While this method dramatically outperforms UCB1, it is still outperformed by IDS. The MOSS algorithm of Audibert and Bubeck [7] is similar to UCB1 and UCB–Tuned, but uses slightly different confidence bounds. It is known to satisfy regret bounds for this problem that are minimax optimal up to a numerical constant factor.

In previous numerical experiments [18, 33, 34, 50], Thompson sampling and Bayes UCB exhibited state-of-the-art performance for this problem. Each also satisfies strong theoretical guarantees, and is known to be asymptotically optimal in the sense defined by Lai and Robbins [39]. Unsurprisingly, they are the closest competitors to IDS. The Bayes UCB algorithm, studied in Kaufmann et al. [34], constructs upper confidence bounds based on the quantiles of the posterior distribution: at time step t the upper confidence bound at an action is the $1-\frac{1}{t}$ quantile of the posterior distribution of that action[4].

[4]Their theoretical guarantees require choosing a somewhat higher quantile, but the authors suggest choosing this quantile, and use it in their own numerical experiments.

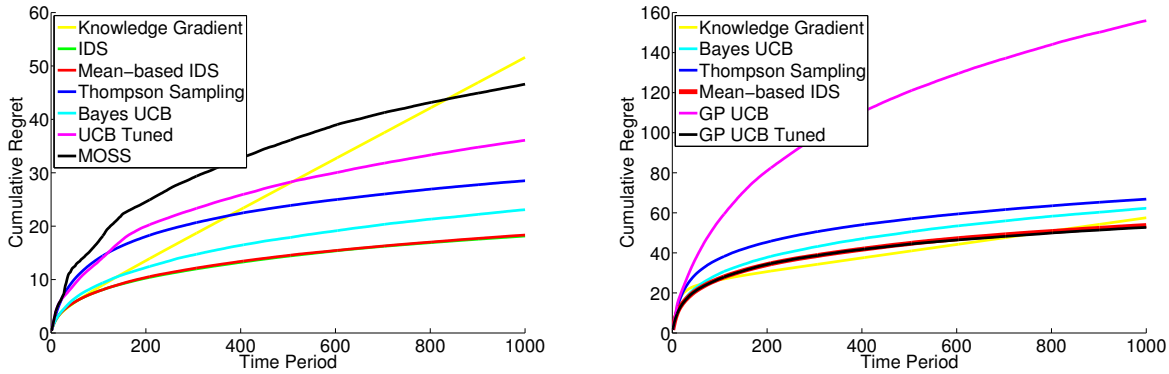
Algorithm	KG	IDS	Thompson	Bayes UCB	UCB1	UCB Tuned	MOSS
Mean Regret	51.595	18.16	28.504	23.086	131.3	36.081	46.592
Standard error	2.1782	0.5523	0.4647	0.4305	0.6124	0.3297	0.3231
quantile .1	0.62316	4.4098	13.584	8.8376	104.76	24.08	36.377
quantile .25	2.7314	8.0054	18.243	12.825	118.66	29.184	39.968
quantile .5	13.593	13.537	25.537	20.548	132.62	35.119	45.111
quantile .75	83.932	21.673	35.222	30.737	145.22	41.629	50.752
quantile .9	161.97	36.728	46.969	40.909	154.91	48.424	57.373

Table 1: Realized regret over 1000 trials in Bernoulli experiment

A somewhat different approach is the knowledge gradient (KG) policy of Powell and Ryzhov [43], which uses a one-step lookahead approximation to the value of information to guide experimentation. For reasons described in Appendix C, KG does not explore sufficiently to identify the optimal arm in this problem, and therefore its regret grows linearly with time. Because KG explores very little, its realized regret is highly variable, as depicted in Table 1. In 100 out of the 1000 trials, the regret of KG was lower than .62, reflecting that the best arm was almost always chosen. In the worst 100 out of the 1000 trials, the regret of KG was larger than 161.971. It should be noted that KG is particularly poorly suited to problems with discrete observations and long time horizons. It can perform very well in other types of experiments.

Finally, as displayed in Figure 1a, our results indicate that the the variation of IDS $\pi^{\text{IDS}_{\text{ME}}}$ presented in Subsection 4.2 has extremely similar performance to standard IDS for this problem.

It's worth pointing out that, although Gittins' indices characterize the Bayes optimal policy for infinite horizon discounted problems, the finite horizon formulation considered here is computationally intractable [24]. A similar index policy [41] designed for finite horizon problems could be applied as a heuristic in this setting. However, a specialized algorithm is required to compute these indices, and this procedure is extremely computationally intensive in our setting due to the long time horizon. Since the goal of this paper is not to develop the best possible policy for bandit problems with independent arms, we have opted to only compare IDS against a variety of simple and widely studied benchmarks.



(a) Binary rewards

(b) Gaussian rewards

Figure 1: Average cumulative regret over 1000 trials

Algorithm	KG	Bayes UCB	Thomp-son	IDS (Mean)	GPUCB	GPUCB–Tuned
Mean Regret	57.53	62.271	66.836	53.66	155.95	52.711
Standard error	3.4395	0.96969	0.96614	1.6605	1.2142	1.5906
quantile .1	16.199	34.282	37.005	23.356	106.47	23.387
quantile .25	20.431	42.589	47.483	28.471	130.28	29.529
quantile .5	25.474	56.828	61.082	38.051	155.53	40.922
quantile .75	35.465	73.431	79.867	55.198	180.8	59.563
quantile .9	120.53	98.383	99.93	96.904	205.49	85.585
quantile .95	279.03	111.34	120.66	148.30	220.48	124.99

Table 2: Realized regret over 1000 trials in independent Gaussian experiment

Time Horizon	10	25	50	75	100	250	500	750	1000
Regret of IDS	9.79	15.71	21.10	24.56	27.13	36.52	44.65	49.85	53.66
Regret of KG	9.08	15.26	19.94	23.79	25.52	34.62	44.32	50.63	62.87

Table 3: Average cumulative regret over 1000 trials

9.2 Independent Gaussian experiment

Our second experiment treats a different multi-armed bandit problem with independent arms. The reward value at each action a follows a Gaussian distribution $N(\theta_a, 1)$. The mean $\theta_a \sim N(0, 1)$ is drawn from a Gaussian prior, and the means of different reward distributions are drawn independently. We ran 1000 simulation trials of a problem with 10 arms. The results are displayed in Figure 1b and Table 2.

For this problem, we again compare Thompson sampling, Bayes UCB, KG, and IDS. We simulate the mean-based variant information-directed sampling $\pi^{\text{IDS}_{\text{ME}}}$ presented in Subsection 4.2, as it offers some computational advantages.

We also simulated the GPUCB of Srinivas et al. [51]. This algorithm maximizes the upper confidence bound $\mu_t(a) + \sqrt{\beta_t}\sigma_t(a)$ where $\mu_t(a)$ and $\sigma_t(a)$ are the posterior mean and standard deviation of θ_a. They provide regret bounds that hold with probability at least $1-\delta$ when $\beta_t = 2\log\left(|\mathcal{A}|t^2\pi^2/6\delta\right)$. This value of β_t is far too large for practical performance, at least in this problem setting. The average regret of GPUCB[5] is 156.6, which is roughly three times that of information-directed sampling. For this reason, we considered a tuned version of GPUCB that sets $\beta_t = c\log(t)$. We ran 1000 trials of many different values of c to find the value $c = .9$ with the lowest average regret for this problem. This tuned version of GPUCB had average regret of 52.7, which is

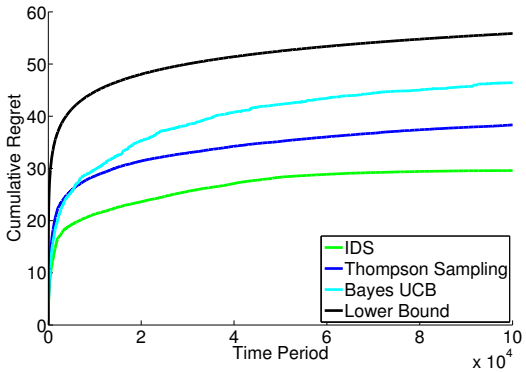


Figure 2: Cumulative regret over 200 trials.

[5]We set $\delta = 0$ in the definition of β_t, as this choice leads to a lower value of β_t and stronger performance.

roughly equivalent to IDS.

The work on knowledge gradient (KG) focuses almost entirely on problems with Gaussian reward distributions and Gaussian priors. We find KG performs better in this experiment than it did in the Bernoulli setting, and its average regret is competitive with that of IDS.

As in the Bernoulli setting, KG's realized regret is highly variable. The median regret of KG is the lowest of any algorithm, but in 50 of the 1000 trials its regret exceeded 279 – seemingly reflecting that the algorithm did not explore enough to identify the best action.

KG may is particularly effective over short time spans. Unlike information-directed sampling, KG takes the time horizon T as an input, and explores less aggressively when there are fewer time periods remaining. Table 3 compares the regret of KG and IDS over different time horizons. Even though IDS does not know the time horizon, it is competitive with KG even over short horizons. We are hopeful that IDS could be modified to exploit fixed and known time horizons even more effectively.

9.3 Asymptotic optimality

The previous subsections present numerical examples in which information-directed sampling outperforms Bayes UCB and Thompson sampling for some problems with independent arms. This is surprising since each of these algorithms is known, in a sense we will soon formalize, to be asymptotically optimal for these problems. This section presents simulation results over a much longer time horizon that suggest IDS scales in the same asymptotically optimal way.

We consider again a problem with binary rewards and independent actions. The action $a_i \in \{a_1, ..., a_K\}$ yields in each time period a reward that is 1 with probability θ_i and 0 otherwise. The seminal work of Lai and Robbins [39] provides the following asymptotic lower bound on regret of any policy π:

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\text{Regret}(T,\pi)|\theta\right]}{\log T} \geq \frac{\sum_{a\neq A^*}(\theta_{A^*} - \theta_a)}{\sum_{a\neq A^*} D_{\text{KL}}(\theta_{A^*} \,||\, \theta_a)} := c(\theta)$$

Note that we have conditioned on the parameter vector θ, indicating that this is a frequentist lower bound. Nevertheless, when applied with an independent uniform prior over θ, both Bayes UCB and Thompson sampling are known to attain this lower bound [33, 34].

Our next numerical experiment fixes a problem with three actions and with $\theta = (.3, .2, .1)$. We compare algorithms over a 10,000 time periods. Due to the expense of running this experiment, we were only able to execute 200 independent trials. Each algorithm uses a uniform prior over θ. Our results, along with the asymptotic lower bound of $c(\theta)\log(T)$, are presented in Figure 2.

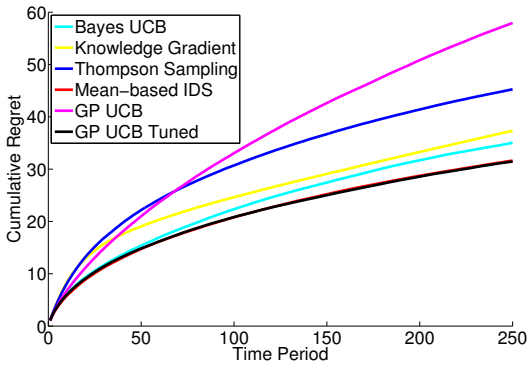


Figure 3: Regret in linear–Gaussian model.

9.4 Linear bandit problems

Our final numerical experiment treats a linear bandit problem. Each action $a \in \mathbb{R}^5$ is defined by a 5 dimensional feature vector. The reward of

Time Horizon	10	25	50	75	100	250
Regret of IDS	5.81	10.10	14.75	18.07	20.80	31.63
Regret of KG	5.72	9.99	14.33	18.47	22.09	37.37

Table 4: Average cumulative regret over 1000 trials in linear bandit experiment

action a at time t is $a^T\theta + \epsilon_t$ where $\theta \sim N(0, 10I)$ is drawn from a multivariate Gaussian prior distribution, and $\epsilon_t \sim N(0, 1)$ is independent Gaussian noise. In each period, only the reward of the selected action is observed. In our experiment, the action set $\mathcal{A}$ contains 30 actions, each with features drawn uniformly at random from $[-1/\sqrt{5}, 1/\sqrt{5}]$. The results displayed in Figure 3 are averaged over 1000 independent trials.

We compare the regret of six algorithms. Three of these - GP-UCB, Thompson sampling, and IDS - satisfy strong regret bounds for this problem[6]. Both GP-UCB and Thompson sampling are significantly outperformed by IDS.

We also include Bayes UCB [34] and a version of GP-UCB that was tuned, as in Subsection 9.2, to minimize its average regret. Each of these displays performance that is competitive with that of IDS. These algorithms are heuristics, in the sense that the way their confidence bounds are constructed differ significantly from those of linear UCB algorithms that are known to satisfy theoretical guarantees.

As discussed in Subsection 9.2, unlike IDS, KG takes the time horizon T as an input, and explores less aggressively when there are fewer time periods remaining. Table 4 compares IDS to KG over several different time horzions. Even though IDS does not exploit knowledge of the time horizon, it is competitve with KG over short time horizons.

10 Conclusion

This paper has proposed information-directed sampling – a new algorithm for online optimization problems in which a decision maker must learn from partial feedback. We establish a general regret bound for the algorithm, and specialize this bound to several widely studied problem classes. We show that it sometimes greatly outperforms other popular approaches, which don't carefully measure the information provided by sampling actions. Finally, for some simple and widely studied classes of multi-armed bandit problems we demonstrate simulation performance surpassing popular approaches.

Many important open questions remain, however. IDS solves a single-period optimization problem as a proxy to an intractable multi-period problem. Solution of this single-period problem can itself be computationally demanding, especially in cases where the number of actions is enormous or mutual information is difficult to evaluate. An important direction for future research concerns the development of computationally elegant procedures to implement IDS in important cases. Even when the algorithm cannot be directly implemented, however, one may hope to develop simple algorithms that capture its main benefits. Proposition 2 shows that any algorithm with small information ratio satisfies strong regret bounds. Thompson sampling is a very tractable algorithm that, we conjecture, sometimes has nearly minimal information ratio. Perhaps simple schemes with small information ratio could be developed for other important problem classes, like the sparse linear bandit problem.

In addition to computational considerations, a number of statistical questions remain open. One question raised is whether IDS attains the lower bound of Lai and Robbins [39] for some

[6]Regret analysis of GP-UCB can be found in [51]. Regret bounds for Thompson sampling can be found in [6, 46, 47]

bandit problems with independent arms. Beyond the empirical evidence presented in Subsection 9.3, there are some theoretical reasons to conjecture this is true. Next, a more precise understanding of problem's *information complexity* remains an important open question for the field. Our regret bound depends on the problem's information complexity through a term we call the information ratio, but it's unclear if or when this is the right measure. Finally, it may be possible to derive lower bounds using the same information theoretic style of argument used in the derivation of our upper bounds.

A Extensions

This section presents a number of ways in which the results and ideas discussed throughout this paper can be extended. We will consider the use of algorithms like information-directed sampling for pure–exploration problems, a form of information-directed sampling that aims to acquire information about p^* instead of A^*, and a version of information directed-sampling that uses a tuning parameter to control how aggressively the algorithm explores. In each case, new theoretical guarantees can be easily established by leveraging our analysis of information-directed sampling.

A.1 Pure exploration problems

Consider the problem of adaptively gathering observations $(A_1, Y_1(A_1), ..., A_{T-1}, Y_{T-1}(A_{T-1}))$ so as to minimize the expected loss of the best decision at time T,

$$\mathbb{E}\left[\min_{a\in\mathcal{A}} \Delta_T(a)\right]. \tag{10}$$

Recall that we have defined $\Delta_t(a) := \mathbb{E}\left[R\left(Y_t(A^*)\right) - R\left(Y_t(a)\right)|\mathcal{F}_{t-1}\right]$ to be the expected regret of action a at time t. This is a "pure exploration problem," in the sense that one is interested only in the terminal regret (10) and not in the algorithm's cumulative regret. However, the next proposition shows that bounds on the algorithm's cumulative expected regret imply bounds on $\mathbb{E}\left[\min_{a\in\mathcal{A}} \Delta_T(a)\right]$.

Proposition 8. *If actions are selected according to a policy π, then*

$$\mathbb{E}\left[\min_{a\in\mathcal{A}} \Delta_T(a)\right] \leq \frac{\mathbb{E}\left[\text{Regret}\left(T, \pi\right)\right]}{T}.$$

Proof. By the tower property of conditional expectation, $\mathbb{E}\left[\Delta_{t+1}(a)|\mathcal{F}_{t-1}\right] = \Delta_t(a)$. Therefore, Jensen's inequality shows $\mathbb{E}\left[\min_{a\in\mathcal{A}} \Delta_{t+1}(a)|\mathcal{F}_{t-1}\right] \leq \min_{a\in\mathcal{A}} \Delta_t(a) \leq \Delta_t(\pi_t)$. Taking expectations and iterating this relation shows that

$$\mathbb{E}\left[\min_{a\in\mathcal{A}} \Delta_T(a)\right] \leq \mathbb{E}\left[\min_{a\in\mathcal{A}} \Delta_t(a)\right] \leq \mathbb{E}\left[\Delta_t(\pi_t)\right] \quad \forall t \in \{1, ..., T\}. \tag{11}$$

The result follows by summing both sides of (11) over $t \in \{1, ..., T\}$ and dividing each by T. □

Information-directed sampling is designed to have low cumulative regret, and therefore balances between acquiring information and taking actions with low expected regret. For pure exploration problems, it's natural instead to consider an algorithm that always acquires as much information about A^* as possible. The next proposition provides a theoretical guarantee for an algorithm of this form. The proof of this result combines our analysis of information-directed sampling with the analysis used to prove Proposition 8, and is provided in Appendix G.

Proposition 9. *If actions are selected so that*

$$A_t \in \arg\max_{a \in \mathcal{A}} g_t(a),$$

and $\Psi_t^* \leq \lambda$ *almost surely for each* $t \in \{1, ..., T\}$*, then*

$$\mathbb{E}\left[\min_{a \in \mathcal{A}} \Delta_T(a)\right] \leq \sqrt{\frac{\lambda H(\alpha_1)}{T}}.$$

A.2 Using information gain about p^*

Information-directed sampling optimizes a single-period objective that balances earning high immediate reward and acquiring information. Information is quantified using the mutual information between the true optimal action A^* and the algorithm's next observation $Y_t(a)$. In this subsection, we will consider an algorithm that instead quantifies the amount learned through selecting an action a using the mutual information $I_t\left(p^*; Y_t(a)\right)$ between the algorithm's next observation and the true outcome distribution p^*. Such an algorithm could actively seek information that is irrelevant to future decisions. However, in some cases, such an algorithm can be computationally simple while offering statistically efficiency.

We introduce a modified form of the information ratio

$$\Psi_t^{p^*}(\pi) := \frac{\Delta(\pi)^2}{\sum_{a \in \mathcal{A}} \pi(a) I_t\left(p^*; Y_t(a)\right)} \tag{12}$$

which replaces the expected information gain about A^*, $g_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a) I_t\left(A^*; Y_t(a)\right)$, with the expected information gain about p^*.

Proposition 10. *For any action sampling distribution* $\tilde{\pi} \in \mathcal{D}(\mathcal{A})$*,*

$$\Psi_t^{p^*}(\tilde{\pi}) \leq \Psi_t(\tilde{\pi}). \tag{13}$$

Furthermore, if $\mathcal{P}$ *is finite, and there is some* $\lambda \in \mathbb{R}$ *and policy* $\pi = (\pi_1, \pi_2, ...)$ *satisfying* $\Psi_t^{p^*}(\pi_t) \leq \lambda$ *almost surely, then*

$$\mathbb{E}\left[\text{Regret}(T, \pi)\right] \leq \sqrt{\lambda H(p^*) T}. \tag{14}$$

Equation (13) relies on the inequality $I_t\left(A^*; Y_t(a)\right) \leq I_t\left(p^*; Y_t(a)\right)$, which itself follows from the data processing inequality of mutual information because A^* is a function of p^*. The proof of the second part of the proposition is almost identical to the proof of Proposition 2, and is omitted.

We have provided several bounds on the information ratio of π^{IDS} of the form $\Psi_t(\pi_t^{\text{IDS}}) \leq \lambda$. By this proposition, such bounds imply that if $\pi = (\pi_1, \pi_2, ...)$ satisfies

$$\pi_t \in \arg\min_{\pi \in \mathcal{D}(\mathcal{A})} \Psi_t^{p^*}(\pi)$$

then, $\Psi_t^{p^*}(\pi_t) \leq \Psi_t^{p^*}(\pi_t^{\text{IDS}}) \leq \Psi_t(\pi_t^{\text{IDS}}) \leq \lambda$, and the regret bound (14) applies.

A.3 A tunable version of information-directed sampling

In this section, we present an alternative form of information-directed sampling that depends on a tuning parameter $\lambda \in \mathbb{R}$. As λ varies, the algorithm strikes a different balanace between exploration and exploration. The following proposition provides regret bounds for this algorithm provided λ is sufficiently large.

Proposition 11. *Fix any $\lambda \in \mathbb{R}$ such that $\Psi_t^* \leq \lambda$ almost surely for each $t \in \{1, ..., T\}$. If $\pi = (\pi_1, \pi_2, ..)$ is defined so that*

$$\pi_t \in \operatorname*{arg\,min}_{\pi \in \mathcal{D}(\mathcal{A})} \left\{ \rho(\pi) := \Delta_t(\pi)^2 - \lambda g_t(\pi) \right\}, \tag{15}$$

then

$$\mathbb{E}\left[\text{Regret}(T, \pi)\right] \leq \sqrt{\lambda H(\alpha) T}.$$

Proof. We have that

$$\rho(\pi_t) \overset{(a)}{\leq} \rho(\pi_t^{\text{IDS}}) \overset{(b)}{\leq} 0,$$

where (a) follows since π_t^{IDS} is feasible for the optimization problem (15), and (b) follows since

$$0 = \Delta_t(\pi_t^{\text{IDS}})^2 - \Psi_t^* g_t(\pi_t^{\text{IDS}}) \geq \Delta_t(\pi_t^{\text{IDS}})^2 - \lambda g_t(\pi_t^{\text{IDS}}).$$

Since $\rho_t(\pi_t) \leq 0$, it must be the case that $\lambda \geq \Delta_t(\pi_t)^2 / g_t(\pi_t) \overset{\text{Def}}{=} \Psi_t(\pi_t)$. The result then follows by applying Proposition 2. □

B Implementations of IDS used in computational experiments

B.1 Beta–Bernoulli

Here we will present an implementation of information-directed sampling for the setting described in Subsection 9.1. Consider a multi-armed bandit problem with binary rewards and K actions denoted by $\mathcal{A} = \{a_1, ..., a_K\}$. The mean reward X_i of each arm a_i is drawn from a Beta prior distribution, and the means of separate arms are drawn independently.

Because the Beta distribution is a conjugate prior for the Bernoulli distribution, the posterior distribution of each X_i is a Beta distribution. The parameters (β_i^1, β_i^2) of this distribution can be updated easily. Let $f_i(x)$ and $F_i(x)$ denote respectively the PDF and CDF of the posterior distribution of X_i. The posterior probability that $A^* = a_i$ can be written as

$$\mathbb{P}\left(\bigcap_{j \neq i} \{X_j \leq X_i\}\right) = \int_0^1 f_i(x) \mathbb{P}\left(\bigcap_{j \neq i} \{X_j \leq x\} \middle| X_i = x\right) dx = \int_0^1 f_i(x) \left(\prod_{j \neq i} F_j(x)\right) dx = \int_0^1 \left[\frac{f_i(x)}{F_i(x)}\right] \overline{F}(x) dx$$

where $\overline{F} : x \mapsto \prod_{i=1}^K F_i(x)$.

Algorithm 2 uses this expression to compute the posterior probability α_i that an action a_i is optimal. To compute the information gain g_j of action j, we use equation (8). Let $M_{i,j} := \mathbb{E}\left[X_j | X_k \leq X_i \ \forall k\right]$ denote the expected value of X_j given that action i is optimal. Step 18 computes the information gain g_j of action a_j as the expected Kullback Leibler divergence between a Bernoulli distribution with mean $M_{i,j}$ and the posterior distribution at action j, which is Bernoulli with parameter $\beta_j^1 / (\beta_j^1 + \beta_j^2)$.

Finally, the algorithm computes the expected reward of the optimal action $\rho^* = \mathbb{E}\left[\max_j X_j\right]$ and uses that to compute the expected regret of action j:

$$\Delta_i = \mathbb{E}\left[\max_j X_j - X_i\right] = \rho^* - \frac{\beta_i^1}{(\beta_i^1 + \beta_i^2)}.$$

Practical implementations of this algorithm can approximate each definite integral by evaluating the integrand at a discrete grid of points in $\{x^1, ..., x^n\} \subset [0, 1]$. The values of $f_i(x), F_i(x), Q_i(x)$

and $\overline{F}(x)$ can be computed and stored for each value of x in this grid. In each period, the posterior distribution of only a single action is updated, and hence these values need to be updated for only one action each period.

Steps 14-21 are the most computationally intensive part of the algorithm. The computational cost of these steps scales as $K^2 n$ where K is the number of actions and n is the number of points used in the discretization of [0,1].

B.2 Independent Gaussian

Algorithm 3 presents an implementation of mean-based information-directed sampling for the setting described in Subsection 9.2. Consider a multi-armed bandit problem with K independent arms denoted by $\mathcal{A} = \{a_1, ..., a_K\}$. The mean reward X_i of each arm a_i is drawn from a Gaussian prior distribution, and the means of separate arms are drawn independently. Rewards are corrupted by zero mean Guassian noise with known variance.

Algorithm 3 is extremely similar to Algorithm 2. One clear difference is that this form of information-directed sampling uses the mean-based variation of IDS described in Subsection 4.2. In addition, the Guassian distribution has a special structure that simplifies the computation of

$$M_{i,j} = \mathbb{E}\left[X_j | X_i \geq \max_k X_k\right].$$

In particular, the computation of $M_{i,j}$ in the algorithm uses the following closed form expression for the expected value of a truncated normal distribution:

$$\mathbb{E}\left[X_j | X_j \leq x\right] = \mu_j - \sigma_j \phi\left(\frac{x - \mu_j}{\sigma_j}\right) / \Phi\left(\frac{x - \mu_j}{\sigma_j}\right) = \mu_j - \sigma_j^2 f_j(x)/F_j(x),$$

if $X_j \sim N(\mu_j, \sigma_j^2)$.

B.3 Linear (mean-based) information-directed sampling

This section provides an implementation of mean-based information-directed sampling for the problem of linear optimization under bandit feedback. Consider a problem where the action set $\mathcal{A}$ is a finite subset of $\mathbb{R}^d$, and 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 $a^T \theta^*$.

In Subsection 4.2 we introduced the term

$$g_t^{\text{ME}}(a) = \mathop{\mathbb{E}}_{a^* \sim \alpha_t}\left[D_{\text{ME}}\left(p_{t,a}(\cdot|a^*) \,||\, p_{t,a}\right)^2\right] \tag{16}$$

where

$$\begin{aligned} D_{\text{ME}}\left(p_{t,a}(\cdot|a^*) \,||\, p_{t,a}\right)^2 &:= \left(\mathop{\mathbb{E}}_{y \sim p_{t,a}(\cdot|a^*)}\left[R(y)\right] - \mathop{\mathbb{E}}_{y \sim p_{t,a}}\left[R(y)\right]\right)^2 \\ &= \left(\mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}, A^* = a^*\right] - \mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}\right]\right)^2. \end{aligned}$$

We will show that for this problem $g_t^{\text{ME}}(a)$ takes on a particularly simple form, and will present an algorithm that leverages this. Write $\mu_t = \mathbb{E}\left[\theta^*|\mathcal{F}_{t-1}\right]$ and write $\mu_t^{(a^*)} = \mathbb{E}\left[\theta^*|\mathcal{F}_{t-1}, A^* = a^*\right]$. Define

$$\text{Cov}_t(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X|\mathcal{F}_{t-1}\right]\right)\left(X - \mathbb{E}\left[X|\mathcal{F}_{t-1}\right]\right)^T |\mathcal{F}_{t-1}\right]$$

Algorithm 2 Beta-Bernoulli IDS

Initialize: Input posterior parameters: $(\beta^1 \in \mathbb{R}^K,\, \beta^2 \in R^K)$
$f_i(x) := \text{Beta.pdf}(x|\beta_i^1, \beta_i^2)$ for $i \in \{1, .., K\}$
$F_i(x) := \text{Beta.cdf}(x|\beta_i^1, \beta_i^2)$ for $i \in \{1, .., K\}$
$\overline{F}(x) := \prod_{i=1}^K F_i(x)$
$Q_i(x) := \int_0^x y f_i(y) dy$ for $i \in \{1, .., K\}$
$\text{KL}(p_1||p_2) := p_1 \log\left(\frac{p_1}{p_2}\right) + (1-p_1)\log\left(\frac{1-p_1}{1-p_2}\right)$

Calculate optimal action probabilities:
for $i \in \{1, .., K\}$ **do**
    $\alpha_i \leftarrow \int_0^1 \left[\frac{f_i(x)}{F_i(x)}\right] \overline{F}(x) dx$
end for

Calculate Information Gain
for $(i, j) \in \{1, .., K\} \times \{1, .., K\}$ **do**
    if $(i == j)$ **then**
        $M_{i,i} \leftarrow \frac{1}{\alpha_i} \int_0^1 \left[\frac{x f_i(x)}{F_i(x)}\right] \overline{F}(x) dx$
    else
        $M_{i,j} \leftarrow \frac{1}{\alpha_i} \int_0^1 \left[\frac{f_i(x)\overline{F}(x)}{F_i(x)F_j(x)}\right] Q_j(x) dx$
    end if
end for

Fill in problem data
$\rho^* \leftarrow \sum_{i=1}^K \alpha_i M_{i,i}$
for $i \in \{1, ..., K\}$ **do**
    $\Delta_i \leftarrow \rho^* - \frac{\beta_i^1}{\beta_i^1 + \beta_i^2}$ for $i \in \{1, .., K\}$
    $g_i \leftarrow \sum_{j=1}^K \alpha_j \text{KL}\left(M_{j,i} || \frac{\beta_i^1}{\beta_i^1 + \beta_i^2}\right)$
end for

chooseAction(Δ, g)

Algorithm 3 Independent Normal IDS (Mean)

Initialize: Input posterior parameters: $(\mu \in \mathbb{R}^K,\, \sigma \in R^K)$
$f_i(x) := \text{Normal.pdf}(x|\mu_i, \sigma_i^2)$.
$F_i(x) := \text{Normal.cdf}(x|\mu_i, \sigma_i^2)$
$\overline{F}(x) := \prod_{i=1}^K F_i(x)$

Calculate optimal action probabilities:
for $i \in \{1, .., K\}$ **do**
    $\alpha_i \leftarrow \int_{-\infty}^{\infty} \left[\frac{f_i(x)}{F_i(x)}\right] \overline{F}(x) dx$
end for

Calculate Information Gain
for $(i, j) \in \{1, .., K\} \times \{1, .., K\}$ **do**
    if $(i == j)$ **then**
        $M_{i,i} \leftarrow \frac{1}{\alpha_i} \int_{-\infty}^{\infty} \left[\frac{x f_i(x)}{F_i(x)}\right] \overline{F}(x) dx$
    else
        $M_{i,j} \leftarrow \mu_j - \frac{\sigma_j}{\alpha_i} \int_{\mathbb{R}} \left[\frac{f_i(x) f_j(x)}{F_i(x) F_j(x)}\right] \overline{F}(x) dx$
    end if
end for

Fill in problem data
$\rho^* \leftarrow \sum_{i=1}^K \alpha_i M_{i,i}$
for $i \in \{1, ..., K\}$ **do**
    $g_i \leftarrow \sum_{j=1}^K \alpha_j \left(M_{j,i} - \mu_i\right)^2$
    $\Delta_i \leftarrow \rho^* - \mu_i$
end for

chooseAction(Δ, g)

to be the posterior covariance of a random variable $X : \Omega \to \mathbb{R}^d$. Then,

$$D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a^*) \,||\, p_{t,a}\right)^2 = a^T \left([\mu_t^{(a^*)} - \mu_t][\mu_t^{(a^*)} - \mu_t]^T \right) a$$

and therefore

$$g_t^{\mathrm{ME}}(a) = a^T L_t a$$

where

$$L_t = \mathop{\mathbb{E}}_{a^* \sim \alpha_t} [\mu_t^{(a^*)} - \mu_t][\mu_t^{(a^*)} - \mu_t]^T = \mathrm{Cov}_t\left(\mu_t^{(A^*)}\right). \tag{17}$$

is exactly the posterior covariance matrix of $\mu_t^{(A^*)}$.

Algorithm 4 presents a simulation based procedure that computes $g_t^{\mathrm{ME}}(a)$ and $\Delta_t(a)$ and selects an action according to the distribution $\pi_t^{\mathrm{IDS}_{\mathrm{ME}}}$. 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 θ^*, which is denoted by $P(\cdot)$ in the algorithm. The action set $\mathcal{A} = \{a_1, ..., a_K\}$ is represented by a matrix $A \in \mathbb{R}^{K \times d}$ where the ith 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 (17). It does this by sampling parameters from the posterior distribution of θ^*, 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^*.

C On the non-convergence of the knowledge gradient algorithm

Like IDS, the *knowledge–gradient* policy [48] selects actions by optimizing a single period objective that encourages earning high expected reward and acquiring a lot of information. Define

$$V_t = \max_{a' \in \mathcal{A}} \mathbb{E}\left[R\left(Y_t(a')\right) | \mathcal{F}_{t-1}\right]$$

to be the expected reward earned by selecting the best estimated action at time t. Define the "KG factor"

$$v_{t,a}^{KG} := \mathbb{E}\left[V_{t+1} | \mathcal{F}_{t-1}, A_t = a\right] - V_t$$

to be the expected improvement in decision quality due to sampling action a and observing $Y_t(a)$. For a problem with time horizon T, the knowledge gradient (KG) policy selects an action in time period t by solving the maximization problem:

$$\max_{a \in \mathcal{A}} \left\{ \mathbb{E}\left[R\left(Y_t(a)\right) | \mathcal{F}_{t-1}\right] + (T-t) v_{t,a}^{KG} \right\}.$$

The measure of information $v_{t,a}^{KG}$ used by knowledge gradient is extremely natural. Information is valuable only if it improves the quality of the agent's decisions. Unfortunately, as highlighted by the next example, algorithms based on this measure may fail even in extremely simple cases.

Example 2. *Consider a problem with two actions $\mathcal{A} = \{a_1, a_2\}$ and binary rewards. Action 1 is known to yield a reward of* 1 *with probability* .51 *and a reward of zero otherwise. Action* 2 *yields a reward of* 1 *with unknown probability $\theta \sim \mathrm{Beta}(1, F)$. Recall that the mean of a random variable with distribution $\mathrm{Beta}(S, F)$ is $S/(S+F)$. If $F \geq 2$, then $v_{t,a_2}^{KG} = 0$, since after a single sample of action* 2*, the posterior mean could never be higher than $2/(2+F) < .51$. In particular, a single sample could never be influential enough to change which action has the highest posterior mean.*

Algorithm 4 Linear Information-Directed Sampling

Initialize: Input $A \in \mathbb{R}^{K \times d}$, $M \in \mathbb{N}$ and posterior distribution $P(\theta)$.
for $i \in \{1, .., K\}$ **do**
    $s^{(i)} \leftarrow 0 \in \mathbb{R}^d$
    $n_i \leftarrow 0 \in \mathbb{R}$
end for

Perform Monte Carlo:
for $m \in \{1, .., M\}$ **do**
    Sample $\theta \sim P(\cdot)$
    $I \leftarrow \arg\max_i \{(A\theta)_i\}$
    $n_I \mathrel{+}= 1$
    $s^{(I)} \mathrel{+}= \theta$
end for

Calculate Problem Data From Monte Carlo Totals
$\mu \leftarrow \frac{1}{M} \sum_{i=1}^{K} s^{(i)}$
$R \leftarrow A\mu \in \mathbb{R}^K$
for $i \in \{1, .., K\}$ **do**
    $\mu^{(i)} \leftarrow s^{(i)}/n_i$
    $\alpha_i \leftarrow n_i/M$
end for
$L \leftarrow \sum_{i=1}^{K} \alpha_i \left(\mu^{(i)} - \mu\right) \left(\mu^{(i)} - \mu\right)^T \in \mathbb{R}^{d \times d}$
$\rho^* \leftarrow \sum_{i=1}^{K} \alpha_i [a_i^T \mu^{(i)}] \in \mathbb{R}$
for $i \in \{1, .., K\}$ **do**
    $g_i \leftarrow a_i^T L a_i \in \mathbb{R}$
    $\Delta_i \leftarrow \rho^* - a_i^T \mu \in \mathbb{R}$
end for

chooseAction(Δ, g)

For this reason, the KG decision rule selects a_1 in the first period. Since nothing is learned from the resulting observation, it will continue selecting action 1 in all subsequent periods. Even as the time horizon T tends to infinity, the KG policy would never select action 2. Its cumulative regret over T time periods is therefore equal to $T\left(\mathbb{E}\left[\max\{.51, \theta\}\right] - .51\right)$, which grows linearly with T.

Many step lookahead. As noted by Frazier and Powell [22], the algorithm's poor performance in this case is due to the increasing returns to information. A single sample of action a_2 provides no value, even though sampling the action several times could be quite valuable. To address the possibility of increasing returns to information Frazier and Powell [22] propose a modified form of KG that considers the value of sampling a single alternative many times. Unfortunately, this approach also does not work in general. As shown the next example, even for problems with independent arms, the value of perfectly observing a single action could be exactly zero, even if there is value to combining information from multiple actions.

Example 3. *Consider a problem with two actions. The reward of each action $i \in \{1, 2\}$ is θ_i, but the parameters θ_1 and θ_2 are unknown. They are distributed independently according to a prior distribution with*

$$\begin{aligned} \mathbb{P}\left(\theta_1 = .6\right) = 1 - \mathbb{P}\left(\theta_1 = .4\right) \quad &\implies \quad \mathbb{E}[\theta_1] = .5 \\ \mathbb{P}\left(\theta_2 = .7\right) = 1 - \mathbb{P}\left(\theta_2 = .5\right) \quad &\implies \quad \mathbb{E}[\theta_2] = .6 \end{aligned}$$

The value of observing either θ_1 or θ_2 alone is zero, since choosing action 2 is (weakly) optimal regardless of what is observed. No realization of θ_1 could exceed $\mathbb{E}[\theta_2]$ and θ_2 is never less than $\mathbb{E}[\theta_1]$. Nevertheless, $I(A^*; \theta_1) > 0$ and $I(A^*; \theta_2) > 0$, so sampling either action provides information about the optimum.

D Proof of Proposition 1

Proof. First, we show the function $\Psi : \pi \mapsto \left(\pi^T \Delta\right)^2 / \pi^T g$ is convex on $\left\{\pi \in \mathbb{R}^K | \pi^T g > 0\right\}$. As shown in Chapter 3 of Boyd and Vandenberghe [11], $f : (x, y) \mapsto x^2/y$ is convex over $\{(x, y) \in \mathbb{R}^2 : y > 0\}$. The function $h : \pi \mapsto (\pi^T \Delta, \pi^T g) \in \mathbb{R}^2$ is affine. Since convexity is preserved under composition with an affine function, the function $\Psi = g \circ h$ is convex.

We now prove the second claim. Consider the optimization problems

$$\text{minimize } \Psi(\pi) \text{ subject to } \pi^T e = 1, \pi \geq 0 \tag{18}$$

$$\text{minimize } \rho(\pi) \text{ subject to } \pi^T e = 1, \pi \geq 0 \tag{19}$$

where

$$\rho(\pi) := \left(\pi^T \Delta\right)^2 - \Psi^* \left(\pi^T g\right),$$

and $\Psi^* \in \mathbb{R}$ denotes the optimal objective value for the minimization problem (18). The set of optimal solutions to (18) and (19) correspond. Note that

$$\Psi(\pi) = \Psi^* \implies \rho(\pi) = 0$$

but for any feasible π, $\rho(\pi) \geq 0$ since $\Delta(\pi)^2 \geq \Psi^* g(\pi)$. Therefore, any optimal solution π_0 to (18) is an optimal solutions to (19) and satisfies $\rho(\pi_0) = 0$. Similarly, if $\rho(\pi) = 0$ then simple algebra shows that $\Psi(\pi) = \Psi^*$ and hence that π is an optimal solution to (18)

We will now show that there is a minimizer of $\rho(\cdot)$ with at most two nonzero components, which implies the same is true of $\Psi(\cdot)$. Fix a minimizer π^* of $\rho(\cdot)$. Differentiating of $\rho(\pi)$ with respect to π at $\pi = \pi^*$ yields

$$\begin{aligned}\frac{\partial}{\partial \pi}\rho(\pi^*) &= 2\left(\Delta^T\pi^*\right)\Delta - \Psi^* g \\ &= 2L^*\Delta - \Psi^* g\end{aligned}$$

where $L^* = \Delta^T\pi^*$ is the expected instantaneous regret of the sampling distribution π^*. Let $d^* = \min_i \frac{\partial}{\partial \pi_i}\rho(\pi^*)$ denote the smallest partial derivative of ρ at π^*. 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

$$\pi_i^* > 0 \implies g_i = \frac{-d^*}{\Psi^*} + \frac{2L^*}{\Psi^*}\Delta_i. \tag{20}$$

Let $i_1, .., i_m$ be the indices such that $\pi_{i_k}^* > 0$ ordered so that $g_{i_1} \geq g_{i_2} \geq ... \geq g_{i_m}$. Then we can choose a $\beta \in [0, 1]$ so that

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

By equation (20), this implies as well that $\sum_{k=1}^m \pi_{i_k}^* \Delta_{i_k} = \beta\Delta_{i_1} + (1-\beta)\Delta_{i_m}$, and hence that the sampling distribution that plays a_{i_1} with probability β and a_{i_m} otherwise has the same instantaneous expected regret and the same expected information gain as π^*. That is, starting with a general sampling distribution π^* 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 $\rho(\pi)$. □

E Proof of Proposition 2 and Proposition 3

The following fact expresses the mutual information between A^* and $Y_t(a)$ as the as the expected reduction in the entropy of A^* due to observing $Y_t(a)$.

Fact 1. *(Lemma 5.5.6 of Gray [29])*

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

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

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

Proof.

$$\mathbb{E}\sum_{t=1}^{T} g_t\left(\pi_t\right) = \mathbb{E}\sum_{t=1}^{T}\mathbb{E}\left[H(\alpha_t) - H(\alpha_{t+1})|\mathcal{F}_{t-1}\right] = \mathbb{E}\sum_{t=1}^{T}\left(H(\alpha_t) - H(\alpha_{t+1})\right) = H(\alpha_1) - H(\alpha_{T+1}) \leq H(\alpha_1),$$

where the first equality relies on Fact 1 and the tower property of conditional expectation and the final inequality follows from the non-negativity of entropy. □

E.1 Proof of Proposition 2

Proof. By definition, if $\Psi_t(\pi_t) \leq \lambda$, then $\Delta_t(\pi_t) \leq \sqrt{\lambda}\sqrt{g_t(\pi_t)}$. Therefore,

$$\mathbb{E}\left[\mathrm{Regret}(T,\pi)\right] = \mathbb{E}\sum_{t=1}^{T}\Delta_t(\pi_t) \leq \sqrt{\lambda}\mathbb{E}\sum_{t=1}^{T}\sqrt{g_t(\pi_t)} \overset{(a)}{\leq} \sqrt{\lambda T}\sqrt{\mathbb{E}\sum_{t=1}^{T} g_t(\pi_t)} \overset{(b)}{\leq} \sqrt{\lambda H(\alpha_1) T}.$$

Inequality (a) follows from Hölder's inequality and (b) follows from Lemma 1. □

E.2 Proof of Proposition 3

Proof. By the definition of π_t^{IDS} given in equation (6) and the definition of Ψ_t^* given in equation (10), $\Delta_t\left(\pi_t^{\mathrm{IDS}}\right)^2 = \Psi_t^* g_t\left(\pi_t^{\mathrm{IDS}}\right)$. This implies

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Regret}\left(T,\pi^{\mathrm{IDS}}\right)\right] = \mathbb{E}\sum_{t=1}^{T}\Delta_t\left(\pi_t^{\mathrm{IDS}}\right) &= \mathbb{E}\sum_{t=1}^{T}\sqrt{\Psi_t^*}\sqrt{g_t\left(\pi_t^{\mathrm{IDS}}\right)} \\
&\overset{(a)}{\leq} \sqrt{\mathbb{E}\sum_{t=1}^{T}\Psi_t^*}\sqrt{\mathbb{E}\sum_{t=1}^{T} g_t\left(\pi_t^{\mathrm{IDS}}\right)} \\
&\overset{(b)}{\leq} \sqrt{H(\alpha_1)}\sqrt{\mathbb{E}\sum_{t=1}^{T}\Psi_t^*} \\
&= \sqrt{\left(\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\Psi_t^*\right) H(\alpha_1) T},
\end{aligned}$$

where (a) follows from Holder's inequality and (b) follows from Lemma 1. □

F Proof of bounds on the information ratio

Here we leverage the tools of our very recent analysis of Thompson sampling [47] to provide bounds on the information ratio of IDS. Since $\Psi_t^* = \min_\pi \Psi_t(\pi) \leq \Psi_t(\pi_t^{\mathrm{TS}})$, the bounds on (π_t^{TS}) provided by Russo and Van Roy [47] immediately yield bounds on the minimal information ratio. The bounds for problems with bandit feedback also apply to mean-based IDS, and some new analysis is required to show this.

F.1 Preliminaries

The following fact lower bounds the Kullback–Leibler divergence between two bounded random variables in terms of the difference between their means. It follows easily from an application of Pinsker's inequality.

Fact 2. *For any distributions P and Q such that that P is absolutely continuous with respect to Q, and any random variable $X : \Omega \to \mathbb{R}$ such that* $\sup X - \inf X \leq 1$*,*

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P||Q)},$$

where $\mathbb{E}_P$ and $\mathbb{E}_Q$ denote the expectation operators under P and Q.

A proof of this result can be found in the appendix of Russo and Van Roy [47]. Because of Assumption 1, this fact shows

$$D_{\mathrm{ME}}(p_{t,a}(\cdot|a^*)\,,\,p_{t,a}) \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(p_{t,a}(\cdot|a^*)\,||p_{t,a}\right)}.$$

The following corollary of this fact allows us to bound the information ratio by considering the information gain in mean $g_t^{\mathrm{ME}}(\pi)$ instead of $g(\pi)$.

Corollary 1. *For any action sampling distribution $\pi \in \mathcal{D}(\mathcal{A})$,*

$$\Psi_t(\pi) \overset{\mathrm{Def}}{=} \frac{\Delta_t(\pi)^2}{g_t(\pi)} \leq \frac{1}{2}\frac{\Delta_t(\pi)^2}{g_t^{\mathrm{ME}}(\pi)}.$$

By Corollary 1, bounds on the information ratio of π^{IDS} and $\pi^{\mathrm{IDS}_{\mathrm{ME}}}$ follow by bounding $(\Delta_t(\pi_t^{\mathrm{TS}})^2)/g_t^{\mathrm{ME}}(\pi_t^{\mathrm{TS}})$ since

$$\Psi_t\left(\pi_t^{\mathrm{IDS}}\right) \leq \Psi_t\left(\pi_t^{\mathrm{IDS}_{\mathrm{ME}}}\right) \leq \frac{1}{2}\frac{\Delta_t(\pi_t^{\mathrm{IDS}_{\mathrm{ME}}})^2}{g_t^{\mathrm{ME}}(\pi_t^{\mathrm{IDS}_{\mathrm{ME}}})} \leq \frac{1}{2}\frac{\Delta_t(\pi_t^{\mathrm{TS}})^2}{g_t^{\mathrm{ME}}(\pi_t^{\mathrm{TS}})}. \tag{21}$$

The following proposition presents a simplified form of the right hand side of this equation, which we will use in our subsequent analysis.

Proposition 12.

$$\frac{\Delta_t(\pi_t^{\mathrm{TS}})^2}{g_t^{\mathrm{ME}}(\pi_t^{\mathrm{TS}})} = \frac{\left(\mathop{\mathbb{E}}_{a\sim\alpha_t}\left[D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a)\,,\,p_{t,a}\right)\right]\right)^2}{\mathop{\mathbb{E}}_{\substack{a\sim\alpha_t\\ a^*\sim\alpha_t}}\left[D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a^*)\,,\,p_{t,a}\right)^2\right]} \tag{22}$$

Proof. The proof essentially follows by plugging in the definitions of $g_t^{\mathrm{ME}}(\cdot)$ and $D_{\mathrm{ME}}(\cdot,\cdot)$ in Subsection 4.2, the definition $\alpha_t(a) = \mathbb{P}\left(A^* = a|\mathcal{F}_{t-1}\right)$, and the definition $\pi_t^{\mathrm{TS}} = \alpha_t$ of Thompson sampling. First,

$$g_t^{\mathrm{ME}}(\pi_t^{\mathrm{TS}}) = \mathop{\mathbb{E}}_{a\sim\pi_t^{\mathrm{TS}}}\left[g_t^{\mathrm{ME}}(a)\right] = \mathop{\mathbb{E}}_{a\sim\alpha_t}\left[g_t^{\mathrm{ME}}(a)\right] = \mathop{\mathbb{E}}_{\substack{a\sim\alpha_t\\ a^*\sim\alpha_t}}\left[D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a^*)\,,\,p_{t,a}\right)^2\right].$$

Then,

$$\begin{aligned}
\Delta_t(\pi_t^{\mathrm{TS}}) = \mathop{\mathbb{E}}_{a\sim\alpha_t}\left[\Delta_t(a)\right] &= \mathbb{E}\left[R(Y_t(A^*))|\mathcal{F}_{t-1}\right] - \mathop{\mathbb{E}}_{a\sim\alpha_t}\left[\mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}\right]\right] \\
&\overset{(a)}{=} \mathop{\mathbb{E}}_{a\sim\alpha_t}\left[\mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}, A^* = a\right] - \mathbb{E}\left[R(Y_t(a))|\mathcal{F}_{t-1}\right]\right] \\
&= \mathop{\mathbb{E}}_{a\sim\alpha_t}\left[D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a)\,,\,p_{t,a}\right)\right]
\end{aligned}$$

where (a) follows by the tower property of expectation. □

F.2 Proof of Proposition 4

Proof. We bound the numerator of (22) by $|\mathcal{A}|$ times its denominator. This is sufficient because of (21). By the Cauchy–Schwarz inequality,

$$\begin{aligned}
\sum_{a\in\mathcal{A}}\alpha_t(a)D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a)\,||\,p_{t,a}\right) &\leq \sqrt{|\mathcal{A}|}\sqrt{\sum_{a\in\mathcal{A}}\alpha_t(a)^2 D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a)\,||\,p_{t,a}\right)^2}. \\
&\leq \sqrt{|\mathcal{A}|}\sqrt{\sum_{a,a^*\in\mathcal{A}}\alpha_t(a)\alpha_t(a')D_{\mathrm{ME}}\left(p_{t,a}(\cdot|a)\,||\,p_{t,a}\right)^2}.
\end{aligned}$$

The result then follows by squaring both sides. □

F.3 Proof of Proposition 5

Proof. The result is implied by Proposition 4 in [47], which shows $\Psi_t(\pi_t^{\text{TS}}) \leq 1/2$. The term Γ_t^2 in that paper is exactly $\Psi_t(\pi_t^{\text{TS}})$. □

F.4 Proof of Proposition 6

The proof of this result is the generalization of the proof of Proposition 4. That result followed through an appropriate application of the Cauchy–Schwartz inequality. The next fact provides an analogous result for matrices. For any rank r matrix $M \in \mathbb{R}^{n\times n}$ with singular values $\sigma_1, ..., \sigma_r$, let

$$\|M\|_* := \sum_{i=1}^{r} \sigma_i, \qquad \|M\|_F := \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n} M_{i,j}^2} = \sqrt{\sum_{i=1}^{r}\sigma_i^2}, \qquad \text{Trace}(M) := \sum_{i=1}^{n} M_{ii}, \quad (23)$$

denote respectively the Nuclear norm, Frobenius norm and trace of M.

Fact 3. *For any matrix $M \in \mathbb{R}^{k\times k}$,*

$$\text{Trace}\,(M) \leq \sqrt{\text{Rank}(M)}\|M\|_{\text{F}}.$$

A proof of this fact can be found in the appendix of Russo and Van Roy [47]. We now prove Proposition 6.

Proof. Write $\mathcal{A} = \{a_1, ..., a_K\}$ and define $M \in \mathbb{R}^{K\times K}$ by $M_{i,j} = \sqrt{\alpha_t(a_i)\alpha_t(a_j)}D_{\text{ME}}\left(p_{t,a_i}(\cdot|a_j)\,,\, p_{t,a_i}\right)$ for each $i, j \in \{1, .., K\}$. Then

$$\mathop{\mathbb{E}}_{a\sim\alpha_t}\left[D_{\text{ME}}(p_{t,a}(\cdot|a)\,,\, p_{t,a})\right] = \text{Trace(M)},$$

and

$$\mathop{\mathbb{E}}_{\substack{a\sim\alpha_t\\ a^*\sim\alpha_t}}\left[D_{\text{ME}}(p_{t,a}(\cdot|a^*)\,,\, p_{t,a})^2\right] = \|M\|_{\text{F}}^2.$$

This shows, by Fact 3 and Proposition 12 that

$$\frac{\Delta_t(\pi_t^{\text{TS}})^2}{g_t^{\text{ME}}(\pi_t^{\text{TS}})} = \frac{\text{Trace(M)}^2}{\|M\|_{\text{F}}^2} \leq \text{Rank}(M).$$

We now show $\text{Rank}(M) \leq d$. Define

$$\mu = \mathbb{E}\left[\theta_{p^*}|\mathcal{F}_{t-1}\right] \quad (24)$$

$$\mu^j = \mathbb{E}\left[\theta_{p^*}|\mathcal{F}_{t-1}, A^* = a_j\right]. \quad (25)$$

Then, by the linearity of the expectation operator, $D_{\text{ME}}\left(p_{t,a_i}(\cdot|a_j)\,,\, p_{t,a_i}\right) = (\mu - \mu^j)^T a_i$. Therefore, $M_{i,j} = \sqrt{\alpha_t(a_i)\alpha_t(a_j)}((\mu_j - \mu)^T a_i)$ and

$$M = \begin{bmatrix} \sqrt{\alpha_t(a_1)}\left(\mu^1 - \mu\right)^T \\ \vdots \\ \vdots \\ \sqrt{\alpha_t(a_K))}\left(\mu^k - \mu\right)^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha_t(a_1)}a_1 & \cdots & \cdots & \sqrt{\alpha_t(a_K)}a_K \end{bmatrix}.$$

Since M is the product of a K by d matrix and a d by K matrix, it has rank at most d. □

F.4.1 Proof of Proposition 7

Proof. Proposition 6 of Russo and Van Roy [47], shows

$$\Psi_t(\pi_t^{\mathrm{TS}}) \leq \frac{d}{2m^2}$$

almost surely for any $t \in \mathbb{N}$. Note that the term Γ_t^2 in that paper is exactly $\Psi_t(\pi_t^{\mathrm{TS}})^2$. The result then follows since

$$\Psi_t^* \overset{\mathrm{Def}}{=} \min_{\pi \in \mathcal{D}(\mathcal{A})} \Psi_t(\pi) \leq \Psi_t(\pi_t^{\mathrm{TS}}).$$

□

G Proof of Extensions

G.1 Proof of Proposition 9

Proof. With probability 1,

$$\min_{a \in \mathcal{A}} \Delta_t(a) \leq \Delta_t(\pi_t^{\mathrm{IDS}}) = \sqrt{\Psi_t^* g_t\left(\pi_t^{\mathrm{IDS}}\right)} \leq \sqrt{\Psi_t^*(\max_{a \in \mathcal{A}} g_t(a))} \leq \sqrt{\lambda(\max_{a \in \mathcal{A}} g_t(a))}.$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\min_{a \in \mathcal{A}} \Delta_T(a)\right] \overset{(a)}{\leq} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \min_{a \in \mathcal{A}} \Delta_t(a)\right] &\leq \frac{\sqrt{\lambda}}{T}\mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\max_{a \in \mathcal{A}} g_t(a)}\right] \\
&\overset{(b)}{\leq} \sqrt{\frac{\lambda}{T}}\sqrt{\mathbb{E}\sum_{t=1}^{T} \max_{a \in \mathcal{A}} g_t(a)} \\
&\overset{(c)}{\leq} \sqrt{\frac{\lambda H(\alpha_1)}{T}},
\end{aligned}$$

where (a) follows by Equation (11) in the Proof of Proposition 8, (b) follows by the Cauchy–Schwartz inequality, and (c) follows by Lemma 1. □

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] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[3] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(3), 1989.

[4] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space. *IEEE Transactions on Automatic Control*, 34 (12):1249–1259, 1989.

[5] 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.

[6] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of The 30th International Conference on Machine Learning*, pages 127–135, 2013.

[7] J.-Y. Audibert and S. Bubeck. Minimax policies for bandits games. *COLT 2009*, 2009.
[8] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 2013.
[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
[10] A. Bai, F. Wu, and X. Chen. Bayesian mixture modelling and inference based Thompson sampling in monte-carlo tree search. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013.
[11] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
[12] E. Brochu, V.M. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-23, Department of Computer Science, University of British Columbia, November 2009.
[13] J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.
[14] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and trends in machine learning*, 5(1):1–122, 2012.
[15] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, June 2011.
[16] 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.
[17] K. Chaloner, I. Verdinelli, et al. Bayesian experimental design: A review. *Statistical Science*, 10(3): 273–304, 1995.
[18] O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems (NIPS)*, 2011.
[19] 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.
[20] 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.
[21] 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.
[22] P.I. Frazier and W.B. Powell. Paradoxes in learning and the marginal value of information. *Decision Analysis*, 7(4):378–403, 2010.
[23] P.I. Frazier, W.B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.
[24] J. Gittins, K. Glazebrook, and R. Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, Ltd, 2011. ISBN 9780470980033.
[25] D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42(1):427–486, 2011.
[26] D. Golovin, A. Krause, and D. Ray. Near-optimal Bayesian active learning with noisy observations. In *NIPS*, volume 10, pages 766–774, 2010.
[27] A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *Proceedings of The 31st International Conference on Machine Learning*, pages 100–108, 2014.
[28] T.L. Graves and T.L. Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.
[29] R.M. Gray. *Entropy and information theory*. Springer, 2011.
[30] P. Hennig and C.J. Schuler. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 98888(1):1809–1837, 2012.
[31] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.
[32] B. Jedynak, P.I. Frazier, R. Sznitman, et al. Twenty questions with noise: Bayes optimal policies for entropy loss. *Journal of Applied Probability*, 49(1):114–136, 2012.
[33] E. Kauffmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite time analysis. In *International Conference on Algorithmic Learning Theory*, 2012.

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

[35] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008.

[36] L. Kocsis and Cs. Szepesvári. Bandit based Monte-Carlo planning. In *ECML*, 2006.

[37] H.J. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.

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

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

[40] J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2, 1978.

[41] J. Niño-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

[42] I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013.

[43] W.B. Powell and I.O. Ryzhov. *Optimal learning*, volume 841. John Wiley & Sons, 2012.

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

[45] P. Rusmevichientong, Z.-J. M. Shen, and D.B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

[46] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 2014. URL `http://dx.doi.org/10.1287/moor.2014.0650`.

[47] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *arXiv preprint arXiv:1403.5341*, 2014.

[48] I.O. Ryzhov, W.B. Powell, and P.I. Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

[49] D. Sauré and A. Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

[50] S.L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

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

[52] M. Valko, A. Carpentier, and R. Munos. Stochastic simultaneous optimistic optimization. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 19–27. JMLR Workshop and Conference Proceedings, May 2013.

[53] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

[54] J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

[55] R. Waeber, P.I. Frazier, and S.G. Henderson. Bisection search with noisy responses. *SIAM Journal on Control and Optimization*, 51(3):2261–2279, 2013.