This article was downloaded by: [171.66.161.73] On: 14 March 2022, At: 07:57
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Mathematics of Operations Research

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

Satisficing in Time-Sensitive Bandit Learning

Daniel Russo, Benjamin Van Roy

To cite this article:
Daniel Russo, Benjamin Van Roy (2022) Satisficing in Time-Sensitive Bandit Learning. Mathematics of Operations Research

Published online in Articles in Advance 14 Mar 2022

. https://doi.org/10.1287/moor.2021.1229

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–25
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Satisficing in Time-Sensitive Bandit Learning

Daniel Russo,[a,*] **Benjamin Van Roy**[b]

[a] Columbia Business School, Columbia University, New York, New York 10027; [b] Department of Electrical Engineering and Department of Management Science and Engineering, Stanford University, Stanford, California 94305
*Corresponding author

Contact: djr2174@columbia.edu, https://orcid.org/0000-0001-5926-8624 (DR); bvr@stanford.edu (BVR)

Received: March 5, 2018
Revised: December 15, 2019
Accepted: December 2, 2020
Published Online in Articles in Advance: March 14, 2022

MSC2020 Subject Classification: Primary: 68T05; secondary: 62C10

https://doi.org/10.1287/moor.2021.1229

Copyright: © 2022 INFORMS

Abstract. Much of the recent literature on bandit learning focuses on algorithms that aim to converge on an optimal action. One shortcoming is that this orientation does not account for time sensitivity, which can play a crucial role when learning an optimal action requires much more information than near-optimal ones. Indeed, popular approaches, such as upper-confidence-bound methods and Thompson sampling, can fare poorly in such situations. We consider instead learning a *satisficing action*, which is near-optimal while requiring less information, and propose *satisficing Thompson sampling*, an algorithm that serves this purpose. We establish a general bound on expected discounted regret and study the application of satisficing Thompson sampling to linear and infinite-armed bandits, demonstrating arbitrarily large benefits over Thompson sampling. We also discuss the relation between the notion of satisficing and the theory of rate distortion, which offers guidance on the selection of satisficing actions.

Funding: B. Van Roy was generously supported by a research grant from Boeing and a Marketing Research Award from Adobe.

Keywords: online optimization • bandit learning • Thompson sampling • satisficing • information theory • rate-distortion theory

1. Introduction

As noted by Herbert Simon (Simon [25], p. 498) when receiving his Nobel Prize, "decision makers can satisfice either by finding optimum solutions for a simplified world, or by finding satisfactory solutions for a more realistic world." Oriented around the former, much of the bandit learning literature has focused on algorithms that aim to converge on an optimal action. As this area progresses, researchers study increasingly complex models, often with very large or infinite action sets. For some such models, convergence to optimality can be a very slow process. It is natural to ask whether a satisficing action can be identified much more quickly. If so, that may be preferable, especially considering that the model itself is a simplified approximation to reality and thus even an exact optimal action is not truly optimal. The following example illustrates the issue.

Example 1 (Many-Armed Deterministic Bandit). Consider an action set $\mathcal{A} = \{1, \ldots, K\}$. Each action $a \in \mathcal{A}$ results in reward θ_a. We refer to this as a *deterministic bandit* because the realized reward is determined by the action and not distorted by noise. The agent begins with a prior over each θ_a that is independent and uniform over $[0,1]$ and sequentially applies actions $A_0, A_1, A_2, \ldots$, selected by an algorithm that adapts decisions as rewards are observed. As K grows, it takes longer to identify the optimal action $A^* = \arg\max_{a \in \mathcal{A}} \theta_a$. Indeed, for any algorithm, $\mathbb{P}(A^* \in \{A_0, \ldots A_{t-1}\}) \leq t/K$. Therefore, no algorithm can expect to select A^* within time $t \ll K$. On the other hand, by simply selecting actions in order, with $A_0 = 1, A_1 = 2, A_2 = 3, \ldots$, the agent can expect to identify an ϵ-optimal action within $t = 1/\epsilon$ time periods, independent of K. □

It is disconcerting that popular algorithms perform poorly when specialized to this simple problem. Thompson sampling (TS) (Thompson [26]), for example, is likely to sample a new action in each time period so long as $t \ll K$. The underlying issue is most pronounced in the asymptotic regime of $K \to \infty$, for which TS never repeats any action because, at any point in time, there will be actions better than those previously selected. A surprisingly simple modification offers dramatic improvement: settle for the first action a for which $\theta_a \geq 1 - \epsilon$. This alternative can be thought of as a variation of TS that aims to learn a *satisficing* action $\tilde{A} = \min\{a : \theta_a \geq 1 - \epsilon\}$. We will refer to an algorithm that samples from the posterior distribution of a satisficing action instead of the optimal action, as *satisficing Thompson sampling* (STS).

Although stylized, Example 1 captures the essence of a basic dilemma faced in all decision problems and not adequately addressed by popular algorithms. The underlying issue is time preference. In particular, if an agent

is only concerned about performance over an asymptotically long time horizon, it is reasonable to aim at learning A^*, whereas this can be a bad idea if shorter term performance matters and a satisficing action $\tilde{A}$ can be learned more quickly. To model time preference and formalize benefits of STS, we will assess performance in terms of expected discounted regret, which for the many-armed deterministic bandit can be written as $\mathbb{E}[\sum_{t=0}^{\infty}\alpha^t(\theta_{A^*}-\theta_{A_t})]$. The constant $\alpha \in [0,1]$ is a discount factor that conveys time preference. It is easy to show, as is done through Theorem A.1 in Appendix A, that in the asymptotic regime of $K \to \infty$, TS experiences expected discounted regret of $1/2(1-\alpha)$, whereas that of STS is bounded above by $1/\sqrt{1-\alpha}$. For α close to one, we have $1/\sqrt{1-\alpha} \ll 1/(1-\alpha)$; therefore, STS vastly outperforms TS. In fact, as α approaches one, the ratio between expected discounted regret of TS and that of STS goes to infinity. This stylized example demonstrates potential advantages of STS.

Of course, satisficing can also play a critical role in more complicated decision problems. The following example provides one simple illustration by adding a structured correlation pattern to the infinite-armed deterministic bandit treated in Example 1.

Example 2 (Infinite-Armed Deterministic Bandit with Hierarchical Structure). A website seeks the best advertisement to display out of an enormous number of alternatives. Each particular ad $a \in \mathcal{A}$ is associated with a known binary feature vector $\phi(a) \in \{0,1\}^d$, where individual features may indicate whether an ad pertains to a particular category of product, contains an image of a particular celebrity, includes bright colors, etc. When displayed, advertisement a generates revenue governed by the linear mixed model

$$R_a = \sum_{j=1}^{d} \phi(a)_j \theta_{0j} + \theta_{1a},$$

where $(\theta_{01}, \cdots, \theta_{0d})$ encodes the effect of each feature, whereas θ_{1a} encodes an ad-specific effect. Any given feature vector $\phi(a)$ for $a \in \mathcal{A}$ is shared by an infinite number of other ads, that is, $|\{a' \in \mathcal{A} | \phi(a') = \phi(a)\}| = \infty$. This serves to model settings where the number of ads with any given feature vector exceeds the problems time horizon. The agent begins with a prior over $\theta = ((\theta_{0j})_{j=1,\cdots,d}, (\theta_{1a})_{a\in\mathcal{A}})$ under which each variable is independent with marginal distributions $\theta_{0j} \sim N(0, \sigma_0^2)$ and $\theta_{1a} \sim \text{Uniform}(0,1)$. □

Similar to the infinite armed bandit with independent arms, identifying an exactly optimal arm is hopeless, though it may be possible to quickly identify a satisficing arm. The search for a satisficing arm can be further accelerated because the linear mixed model enables some generalization across arms. In particular, data gathered from trying some arms inform estimates for other arms, helping to guide the search.

In order to avoid trying each individual arm, one might be tempted to approximate this model via a linear bandit, ignoring the ad-specific effect θ_{1a} altogether. In particular, one could assume that rewards are generated according to $\sum_{j=1}^{d} \phi(a)_j \theta_{0j} + W_t$, where W_t is independent and identically distributed (iid) zero-mean noise. But ignoring consistent ad-specific effects in this way can lead popular algorithms like Thompson sampling and linear upper-confidence bound approaches to converge on poorly performing arms. We will revisit this example in Section 8, where we demonstrate how our approach can efficiently identify satisficing actions while modeling ad-specific effects.

1.1. Our Contributions

This paper develops a general framework for studying satisficing in sequential learning. Satisficing algorithms aim to learn a satisficing action $\tilde{A}$. Building on the work of Russo and Van Roy [21], we will establish a general information-theoretic regret bound, showing that any algorithm's expected discounted regret relative to a satisficing action is bounded in terms of the mutual information $I(\theta;\tilde{A})$ between the model parameters θ and the satisficing action and a newly defined notion of information ratio, which measures the cost of information acquired about the satisficing action. The mutual information $I(\theta;\tilde{A})$ can be thought of as the number of bits of information about θ required to identify $\tilde{A}$; the fact that the bound depends on this quantity instead of the entropy of A^*, as does the bound of Russo and Van Roy [21], allows it to capture the reduction of discounted regret made possible by settling for the satisficing action.

A natural and deep question concerns the choice of satisficing action $\tilde{A}$ and the limits of performance attainable via satisficing. An exploration of this question yields novel connections between sequential learning and rate-distortion theory. In Section 5, we define a natural rate-distortion function for Bayesian decision making, which captures the minimal information about θ a decision maker must acquire in order to reach an ϵ-optimal decision. Combining this rate-distortion function with our general regret bound leads to new results and insights. As an example, although previous information-theoretic regret bounds for the linear bandit problem

become vacuous in contexts with infinite action spaces, our rate-distortion function leads to a strong bound on expected discounted regret.

We will also study the infinite-armed bandit problem with noisy rewards. Here, we will consider a satisficing action $\tilde{A} = \min\{a : \theta_a \geq 1 - \epsilon\}$. Simple numerical experiments demonstrate the benefits of STS over TS. We instantiate our general regret analysis in the infinite-armed bandit problem by bounding the mutual information and information ratio in that problem. This yields a bound on expected discounted regret that formalizes the benefits of STS over TS. We complement this upper bound by establishing a matching lower bound on the regret of any algorithm for infinite armed bandit.

1.2. Alternative Approaches

Many papers (Bubeck et al.[6], Kleinberg et al. [14], Rusmevichientong and Tsitsiklis [19]) have studied bandit problems with continuous action spaces, where it is also necessary to learn only approximately optimal actions. However, because these papers focus on the asymptotic growth rate of regret, they implicitly emphasize later stages of learning, where the algorithm has already identified extremely high performing actions but exploration is needed to identify even better actions. Our discounted framework instead focuses on the initial cost of learning to attain good, but not perfect, performance. Recent papers (Francetich and Kreps [11], Francetich and Kreps [12]) study several heuristics for a discounted objective, though without an orientation toward formal regret analysis. The knowledge gradient algorithm of Ryzhov et al. [23] also takes time horizon into account and can learn suboptimal actions when it's not worthwhile to identify the optimal action. This algorithm tries to directly approximate the optimal Bayesian policy using a one-step lookahead heuristic, but there are no performance guarantees for this method. Deshpande and Montanari [10] consider a linear bandit problem with dimension that is too large relative to the desired horizon. They propose an algorithm specifically for that problem that limits its exploration and learns something useful within this short time frame.

Berry et al. [2], Wang et al. [29], and Bonald and Proutiere [3] study an infinite-armed bandit problem in which it is impossible to identify an optimal action and propose algorithms to minimizes the asymptotic growth rate of regret. Their strategies are carefully designed but appear to be difficult to adapt to more complex problems. For example, one algorithm in Berry et al. [2] discards an arm as soon as it produces a reward of zero in some period, which is sensible only for infinite armed bandits with uniform prior and binary rewards. A brief section describes extensions to nonuniform priors, but still these cannot extend beyond binary feedback. Bonald and Proutiere [3] design a procedure specifically for the infinite-armed bandit with uniform prior and binary rewards under which the first-order term in an asymptotic expansion of regret is minimized. The algorithm of Wang et al. [29] discards all but k arms at the start of the problem then applies a standard algorithm for k-armed bandits. The main contribution of that work is to carefully analyze regret as a function of the prior, allowing them to choose k to minimize regret upper bounds. Although we will instantiate our general regret bound for STS on the infinite-armed bandit problem, we view this example as a very simple and stylized special case that we provide to illustrate basic concepts. The flexibility of STS and our analysis framework allow this work to be applied to more complicated time-sensitive learning problems.

Our approach is built on Thompson sampling. One simple reason is that Thompson sampling is enjoying wide practical use because of its ease of use, ability to incorporate complex prior information, and resilience to delayed feedback (Chapelle and Li [8], Russo et al. [22], Scott [24]). Given this, we see value in broadening the class of problem to which Thompson sampling approaches may be applied. It is also worth emphasizing that formulations of problems like the infinite armed-bandit in Example 1 are inherently Bayesian. Arms are modeled as independent and yet the decision maker is required to make inferences about the quality of arms for which no data are available. By contrast, frequentist algorithms like KL-UCB (Cappe et al. [7]) usually require an initial phase in which all arms are tested at least once. Beyond simple infinite-armed bandits, developing satisficing variants of Thompson sampling appears to be a natural way to approach more complex problems like the hierarchical bandit in Example 2 ,which requires both satisficing and generalizing across arms. A separate motivation for us is to improve the information theoretic analysis of complex Bayesian bandit problems. This analysis is centered on Thompson sampling; but beyond any interest in the algorithm, it has been an effective tool for establishing regret upper bounds, including for bandit convex optimization (Bubeck and Eldan [5]), for partial monitoring (Lattimore and Szepesvári [16]), for bandits with graph-structured feedback (Tossou et al. [27]), and for reinforcement learning (Lu and Van Roy [18]).

2. Problem Formulation

An agent sequentially chooses actions $(A_t)_{t\in\mathbb{N}_0}$ from the action set $\mathcal{A}$ and observes the corresponding outcomes $(Y_t)_{t\in\mathbb{N}_0} \subset \mathcal{Y}$. The agent associates a reward $R(y)$ with each outcome $y \in \mathcal{Y}$. Let $R_t \equiv R(Y_t)$ denote the reward

corresponding to outcome Y_t. The outcome Y_t in period t depends on the chosen action A_t, idiosyncratic randomness associated with that time step, and a random variable θ that is fixed over time. Formally, there is a known system function g and an iid sequence of disturbances $(W_t)_{t\in\mathbb{N}_0}$ such that

$$Y_t = g(A_t, \theta, W_t).$$

The disturbances W_t are independent of θ and have a known distribution. This is without loss of generality, as uncertainty about g and the distribution of W_t could be included in the definition of θ. From this, we can define

$$\mu(a, \theta) = \mathbb{E}[R(g(a, \theta, W_t))|\theta],$$

to be the expected reward of an action a under parameter θ. Ours can be thought of as a Bayesian formulation, in which the distribution of θ represents the agent's prior uncertainty about the true characteristics of the system, and conditioned on θ; the remaining randomness in Y_t represents idiosyncratic noise in observed outcomes.

The history available when selecting action A_t is $\mathcal{H}_{t-1} = (A_0, Y_0, \ldots, A_{t-1}, Y_{t-1})$. The agent selects actions according to a policy, which is a sequence of functions $\psi = (\psi_t)_{t\in\mathbb{N}_0}$, each mapping a history and an exogenous random variable ξ to an action, with $A_t = \psi_t(\mathcal{H}_{t-1}, \xi)$ for each t. Throughout the paper, we use ξ to denote some random variable that is independent of θ and the disturbances $(W_t)_{t\in\mathbb{N}_0}$.

Let $R^* = \sup_{a\in\mathcal{A}}\mu(a, \theta)$ denote the supremal reward; and let $A^* \in \arg\max_{\mathrm{a}\in\mathcal{A}} \mu(\mathrm{a}, \theta)$ denote the true optimal action, when this maximum exists. As a performance metric, we consider *expected discounted regret* of a policy ψ is defined by

$$\mathrm{Regret}(\alpha, \psi) = \mathbb{E}^{\psi}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - R_t)\right],$$

which measures a discounted sum of the expected performance gap between an omniscient policy, which always chooses the optimal action A^*, and the policy ψ, which selects the actions $(A_t)_{t\in\mathbb{N}_0}$. This deviates from the typical notion of expected regret in its dependence on a discount factor $\alpha \in [0, 1]$. Regular expected regret corresponds to the case of $\alpha = 1$. Smaller values of α convey time preference by weighting gaps in nearer-term performance higher than gaps in longer-term performance.

The definition above compares regret relative to the optimal action A^* and corresponding reward R^*. It is useful to also consider performance loss relative to a less stringent benchmark. We define the satisficing regret at level $D \geq 0$ to be

$$\mathrm{SRegret}(\alpha, \psi, D) = \mathbb{E}^{\psi}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - D - R_t)\right].$$

This measures regret relative to an action that is near optimal in the sense that it yields expected reward $R^* - D$, which is within D of optimal. This notation was chosen because of the connection we develop with rate-distortion theory, where D typically denotes a tolerable level of "distortion" in a lossy compression scheme. Of course, for all $D \geq 0$,

$$\mathrm{Regret}(\alpha, \psi) = \mathrm{SRegret}(\alpha, \psi, D) + \frac{D}{1-\alpha}; \tag{1}$$

and so one can easily translate between bounds on regret and bounds on satisficing regret. However, directly studying satisficing regret helps focus our attention on the design of algorithms that purposefully avoid the search for exactly optimal behavior in order to limit exploration costs.

2.1. Additional Notation

Before beginning, let us first introduce some additional notation. We denote the entropy of a random variable X by $H(X)$, the Kullback-Leibler divergence between probability distributions P and Q by $D(P\|Q)$, and the mutual information between two random variables X and Y by $I(X; Y)$. We will frequently be interested in the conditional mutual information $I(X; Y|\mathcal{H}_{t-1})$.

We sometimes denote by $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{H}_{t-1}]$ the expectation operator conditioned on the history up to time t and similarly define $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|\mathcal{H}_{t-1})$. The definitions of entropy and mutual information depend on a base measure. We use $H_t(\cdot)$ and $I_t(\cdot, \cdot)$ to denote entropy and mutual information when the base measure is

the posterior distribution $\mathbb{P}_t$. For example, if X is and Z are discrete random variables taking values in sets $\mathcal{X}$ and $\mathcal{Z}$,

$$I_t(X;Z) = \sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} \mathbb{P}_t(X=x, Z=z)\log\left(\frac{\mathbb{P}_t(X=x, Z=z)}{\mathbb{P}_t(X=x)\mathbb{P}_t(Z=z)}\right).$$

Because of its dependence on the realized history, $\mathcal{H}_{t-1}$, $I_t(X;Z)$ is a random variable. The standard definition of conditional mutual information integrates over this randomness and, in particular, $\mathbb{E}[I_t(X;Z)] = I(X;Z|\mathcal{H}_{t-1})$.

3. Satisficing Actions

We will consider learning a satisficing action $\tilde{A}$ instead of an optimal action A^*. The idea is to target a satisficing action that is near optimal yet easy to learn. The information about θ required to learn an action $\tilde{A}$ is captured by $I(\theta;\tilde{A})$, whereas the performance loss is $\mathbb{E}[R^* - \tilde{R}]$, where $\tilde{R} = \mu(\tilde{A},\theta)$. For $\tilde{A}$ to be easy to learn relative to A^*, we want $I(\theta;\tilde{A}) \ll I(\theta;A^*)$. We will motivate this abstract notion through several examples.

Our first example addresses the infinite-armed deterministic bandit, as discussed in Section 1.

Example 3 (First Satisfactory Arm). Consider the infinite-armed deterministic bandit of Section 1. For this problem, the prior of A^* is uniformly distributed across a large number K of actions, and $I(\theta;A^*) = H(A^*) = \log K$. Consider a satisficing action $\tilde{A} = \min\{k \mid \theta_k \geq R^* - \epsilon\}$, which represents the first action that attains reward within ϵ of the optimum $R^* = \max_k \theta_k$. As $K \to \infty$, $I(\theta;A^*) \to \infty$. But $I(\theta;\tilde{A})$ remains finite, as in this limit $\tilde{A}$ converges weakly to a geometric random variable, with $I(\theta;\tilde{A}) = H(\tilde{A}) = -((1-\epsilon)\log(1-\epsilon) + \epsilon\log(\epsilon))/\epsilon$. □

The next example addresses the infinite-armed bandit with hierarchical structure from Section 1. Naturally, the satisficing action is itself defined in a hierarchical way, first defining a subset of arms with the most attractive features and then selecting the first arm among those with a satisfactory arm-specific effect.

Example 4 (Hierarchical Satisficing). Consider a hierarchical infinite-armed deterministic bandit of Section 1. The set of arms is $\mathcal{A} = \{1,2,3,\ldots\}$. The reward generated by an arm $a \in \mathcal{A}$ can be written as $\mu(a,\theta) = \langle\phi(a), \theta_0\rangle + \theta_{1a}$, where $\phi(a) \in \{0,1\}^d$ is a known feature vector associated with the arm, $\theta_0 \in \mathbb{R}^d$ is drawn from some prior, and the arm-specific effects $\theta_{1,a} \sim \text{Unif}[0,1]$ are drawn independently across arms $a \in \mathcal{A}$. For each action a, we assume there is an infinite collection of other actions $\{a' \mid \phi(a') = \phi(a)\}$ that share the same feature vector. (One should have in mind problems where the features encode relatively coarse categories.) We take $\mathcal{A}(\theta_0) = \{a \in \mathcal{A} | \langle\phi(a),\theta_0\rangle \geq \langle\phi(a'),\theta_0\rangle \quad \forall a' \in \mathcal{A}\}$ to be the subset of actions that have optimal features. We take a satisficing action to be the first such action that has a sufficiently large ad-specific effect:

$$\tilde{A} = \arg\min_{a\in\mathcal{A}(\theta_0)} \{a \mid \theta'_{1a} \geq 1-\epsilon\}. \tag{2}$$

In this case, although there are an infinite number of arms, the entropy of the satisficing action is bounded as

$$I(\tilde{A};\theta) = I(\tilde{A};\theta_0) + I(\tilde{A};\theta_1 \mid \theta_0) \leq d\log(2) - ((1-\epsilon)\log(1-\epsilon) + \epsilon\log(\epsilon))/\epsilon.$$

where we have used that θ_0 is supported on 2^d possible values and that, conditioned on θ_0, $\tilde{A}$ follows a geometric distribution with survival probability $1-\epsilon$. □

The next example involves reducing the granularity of a discretized action space.

Example 5 (Discretization). Consider a linear bandit with $\mathcal{A} \subset \mathbb{R}^p$ and $\mu(a,\theta) = a^\top\theta$ for an unknown vector θ. Suppose that $\theta \sim N(0,I)$ and $\mathcal{A}$ consists of K vectors spread out uniformly along boundary of the p-dimensional unit sphere $\{a \in \mathbb{R}^p : \|a\|_2 = 1\}$. The optimal action $A^* = \arg\max_{a\in\mathcal{A}} a^\top\theta$ is then uniformly distributed over $\mathcal{A}$, and therefore $I(\theta;A^*) = H(A^*) = \log K$. As $K \to \infty$, it takes an enormous amount of information to exactly identify A^*. The results of Russo and Van Roy [20] become vacuous in this limit. Consider a satisficing action $\tilde{A}$ that represents a coarser version of A^*. In particular, for $M \ll K$, let $\tilde{\mathcal{A}}$ consist of M vectors spread out uniformly along the boundary of the p-dimensional unit sphere, with M is chosen such that for each element of $\mathcal{A}$, there is a close approximation in $\tilde{\mathcal{A}}$. Let $\tilde{A} = \arg\max_{a\in\tilde{\mathcal{A}}} \theta^\top a$. This can be viewed as a form of lossy-compression, for which $H(\tilde{A}) \ll H(A^*)$, whereas $\mathbb{E}[R^* - \tilde{R}]$ remains small. □

In the previous examples, θ determined $\tilde{A}$, and therefore $I(\theta;\tilde{A}) = H(\tilde{A})$. We now consider an example in which $I(\theta;\tilde{A})$ is controlled by randomly perturbing the satisficing action. Here, $I(\theta;\tilde{A})$ can be small even though $H(\tilde{A})$ is large.

Example 6 (Random Perturbation). Consider again the linear bandit from the previous example. An alternative satisficing action $\tilde{A}$ results from optimizing a perturbed objective $\tilde{A} \in \arg\max_{a \in \mathcal{A}} a^T(\theta + Z)$ where $Z \sim N(0, (\epsilon/p)^2 I)$. Because Z is not observable, it is not possible in this case to literally learn $\tilde{A}$. Instead, we consider learning to behave in a manner that is statistically indistinguishable from $\tilde{A}$. The variance of Z is chosen such that $\mathbb{E}[R^* - \tilde{R}] = \epsilon$. Moreover, it can be shown that $I(\tilde{A}; \theta) \le I(\theta + Z; \theta) = p\log(1 + p^2/\epsilon^2)$; therefore, the information required about θ is bounded independently of the number of actions. □

4. A General Regret Bound

This section provides a general discounted regret bound and a new information-theoretic analysis technique. The first subsection introduces an alternative to the information ratio of Russo and Van Roy [21], which is more appropriate for time-sensitive online learning problems. The following subsection establishes a general discounted regret bound in terms of this information ratio.

4.1. A New Information Ratio

First, we make a simplification to the information ratio $(\mathbb{E}_t[R^* - R_t])^2 / I_t(A^*; (A_t, Y_t))$ defined by Russo and Van Roy [21]. That expression depends on the history $\mathcal{H}_{t-1}$ and hence is a random variable. In this paper, we observe that this can be avoided and instead take as a starting point a simplified form of the information ratio that integrates out all randomness. In particular, we study

$$\frac{(\mathbb{E}[R^* - R_t])^2}{I(A^*; (A_t, Y_t) | \mathcal{H}_{t-1})}. \tag{3}$$

Uniform bounds on the information ratio of the type established in past work (Bubeck and Eldan [5], Liu et al. 17], Russo and Van Roy [21]) imply those on (3). Precisely, if $(\mathbb{E}_t[R^* - R_t])^2 / I_t(A^*; (A_t, Y_t))$ is bounded by $\lambda \in \mathbb{R}$ almost surely (i.e., for any history $\mathcal{H}_{t-1}$), then (3) is bounded by λ because

$$(\mathbb{E}[R^* - R_t])^2 \le \mathbb{E}[(\mathbb{E}_t[R^* - R_t])^2] \le \mathbb{E}[\lambda I_t(A^*; (A_t, Y_t))] = \lambda I(A^*; (A_t, Y_t) | \mathcal{H}_{t-1}).$$

A more important change comes from measuring information about a benchmark action $\tilde{A}$, which could be defined as in the examples in the previous section rather than with respect to the optimal action A^*. For a benchmark action $\tilde{A}$ we consider the single period information ratio

$$\frac{(\mathbb{E}[\tilde{R} - R_t])^2}{I(\tilde{A}; (A_t, Y_t) | \mathcal{H}_{t-1})},$$

where $\tilde{R} = \mu(\tilde{A}, \theta)$. This ratio relates the current shortfall in performance relative to the benchmark action $\tilde{A}$ to the amount of information acquired about the benchmark action. We study the discounted average of these single period information ratios, defined for any policy ψ as

$$\Gamma(\tilde{A}, \psi) = (1 - \alpha^2) \sum_{t=0}^{\infty} \alpha^{2t} \left(\frac{(\mathbb{E}[\tilde{R} - R_t])^2}{I(\tilde{A}; (A_t, Y_t) | \mathcal{H}_{t-1})} \right), \tag{4}$$

where the actions $(A_t)_{t \in \mathbb{N}_0}$ are chosen under ψ. The square in the discount factor α is consistent with the problem's original discount rate because $(\mathbb{E}[\alpha^t(\tilde{R} - R_t)])^2 = \alpha^{2t}(\mathbb{E}[\tilde{R} - R_t])^2$.

4.2. General Regret Bound

The following theorem bounds the expected discounted regret of any algorithm, or policy, ψ in terms of the information ratio (4).

Theorem 1. *For any policy ψ, any $D \ge 0$, and any $\tilde{A} = f(\theta, \xi)$ where ξ is independent of the disturbances $(W_t)_{t \in \mathbb{N}_0}$, if $\mathbb{E}[\mu(\tilde{A}, \theta)] \ge R^* - D$, then*

$$\mathrm{SRegret}(\alpha, \psi, D) \le \sqrt{\frac{\Gamma(\tilde{A}, \psi) I(\tilde{A}; \theta)}{1 - \alpha^2}}.$$

Proof. We first show that the mutual information between $\tilde{A}$ and θ bounds the expected accumulation of mutual information between $\tilde{A}$ and observations $(A_t, Y_t)_{t\in\mathbb{N}_0}$. By the chain rule for mutual information, for any T,

$$\begin{aligned}\sum_{t=0}^{T} I(\tilde{A}; (A_t, Y_t) \mid \mathcal{H}_{t-1}) &= \sum_{t=0}^{T} I(\tilde{A}; (A_t, Y_t) \mid A_0, Y_0, \ldots, A_{t-1}, Y_{t-1}) \\ &= I(\tilde{A}; \mathcal{H}_T) \\ &\le I(\tilde{A}; (\theta, \mathcal{H}_T)) \\ &= I(\tilde{A}; \theta) + I(\tilde{A}; \mathcal{H}_T \mid \theta) \\ &= I(\tilde{A}; \theta),\end{aligned}$$

where the final inequality uses that, conditioned on θ, $\tilde{A}$ is independent of $\mathcal{H}_T$. Taking the limit as $T \to \infty$ implies

$$\sum_{t=0}^{\infty} I(\tilde{A}; (A_t, Y_t) \mid \mathcal{H}_{t-1}) \le I(\tilde{A}; \theta),$$

where the infinite series is assured to converge by the nonnegativity of mutual information. Now, let

$$\Gamma_t \equiv \frac{(\mathbb{E}[\tilde{R} - R_t])^2}{I(\tilde{A}; (A_t, Y_t) \mid \mathcal{H}_{t-1})}$$

denote the information ratio at time t under the benchmark action $\tilde{A}$ and actions $(A_t)_{t\in\mathbb{N}_0}$ chosen according to ψ. Then

$$\begin{aligned}\mathrm{SRegret}(\alpha, \psi, D) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - D - R_t)\right] &= \sum_{t=0}^{\infty} \alpha^t \mathbb{E}[\tilde{R} - R_t] \\ &= \sum_{t=0}^{\infty} \sqrt{\alpha^{2t}\Gamma_t} \sqrt{I(\tilde{A}; (A_t, Y_t) \mid \mathcal{H}_{t-1})} \\ &\le \sqrt{\sum_{t=0}^{\infty} \alpha^{2t}\Gamma_t} \sqrt{\sum_{t=0}^{\infty} I(\tilde{A}; (A_t, Y_t) \mid \mathcal{H}_{t-1})} \\ &\le \sqrt{\sum_{t=0}^{\infty} \alpha^{2t}\Gamma_t} \sqrt{I(\tilde{A}; \theta)} \\ &= \sqrt{\frac{\Gamma(\tilde{A}, \psi) I(\tilde{A}; \theta)}{1 - \alpha^2}},\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second was established earlier in this proof. □

An immediate consequence of this bound on satisficing regret is the discounted regret bound

$$\mathrm{Regret}(\alpha, \psi) \le \frac{\mathbb{E}[R^* - \tilde{R}]}{1 - \alpha} + \sqrt{\frac{\Gamma(\tilde{A}, \psi) I(\tilde{A}; \theta)}{1 - \alpha^2}}. \tag{5}$$

This bound decomposes regret into the sum of two terms: one that captures the discounted performance shortfall of the benchmark action $\tilde{A}$ relative to A^* and one that bounds the additional regret incurred while learning to identify $\tilde{A}$. Breaking things down further, the mutual information $I(\theta; \tilde{A})$ measures how much information the decision maker must acquire in order to implement the action $\tilde{A}$, and the information ratio measures the regret incurred in gathering this information. It is worth highlighting that for any given action process, this bound holds simultaneously for all possible choices of $\tilde{A}$; in particular, it holds for the $\tilde{A}$ minimizing the right-hand side of (5).

5. Connections with Rate Distortion Theory

This section considers the optimal choice of satisfactory action $\tilde{A}$ and develops connections with the theory of rate distortion in information theory. We construct a natural rate-distortion function for Bayesian decision

making in the next subsection. Subsection 5.2 then develops a general regret bound that depends on this rate-distortion function.

5.1. A Rate-Distortion Function for Bayesian Decision Making

In information theory, the entropy of a source characterizes the length of an optimal lossless encoding. The celebrated rate-distortion theory (Cover and Thomas [9, chapter 10]) characterizes the number of bits required for an encoding to be close in some loss metric. This theory resolves when it is possible to derive a satisfactory lossy compression scheme while transmitting far less information than required for a lossless compression. The rate-distortion function for a random variable X taking values in $\mathcal{X}$ with respect to a loss function $\ell : \hat{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ is

$$\begin{aligned} \mathcal{R}(D) = \min\ & I(\hat{X}; X) \\ & s.t.\ \mathbb{E}[\ell(\hat{X}, X)] \le D, \end{aligned} \tag{6}$$

where the minimum is taken over the choice of random variables $\hat{X}$ taking values in $\hat{\mathcal{X}}$ and $I(X;\hat{X})$ denotes the mutual information between X and $\hat{X}$. One can view this optimization problem as specifying a conditional distribution $P(\hat{X} \in \cdot \mid X)$ that minimizes the information $\hat{X}$ uses about X among all choices incurring an average loss less than D.

We will explore a powerful link with sequential Bayesian decision making, where the rate-distortion function characterizes the minimal amount of new information the decision maker must gather in order make a satisfactory decision. Typically (6) is applied in the context of representing X as closely as possible by $\hat{X}$, and the loss function is taken to be something like the squared distance or total variation distance between the two. For our purposes, we replace X with θ and $\hat{X}$ with a benchmark action $\tilde{A}$. The interpretation is that $\tilde{A} = f(\theta;\xi)$ is a function of the unknown parameter θ and some exogenous randomness ξ that offers a similar reward to playing A^* but hopefully can be identified using much less information about θ. We specify a loss function $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$ measuring the single period regret from playing a under θ:

$$\ell(a, \theta) = \max_{a' \in \mathcal{A}} \mu(a', \theta) - \mu(a, \theta).$$

As a result

$$\mathbb{E}[\ell(\tilde{A}, \theta)] = \mathbb{E}[R^* - \mu(\tilde{A}, \theta)].$$

We come to the rate-distortion function

$$\begin{aligned} \mathcal{R}(D) := \min\ & I(\tilde{A}; \theta) \\ & s.t.\ \mathbb{E}[R^* - \mu(\tilde{A}, \theta)] \le D. \end{aligned} \tag{7}$$

As before, the minimum above is taken over the choice of random variable $\tilde{A}$ taking values in $\mathcal{A}$. That is, the minimum is taken over all conditional probability distributions $\mathbb{P}(\tilde{A} \in \cdot \mid \theta)$ specifying a distribution over actions as a function of θ. Because the choice $\tilde{A} = A^*$ is always feasible, for all $D > 0$

$$\mathcal{R}(D) \le I(A^*; \theta) = H(A^*),$$

where $H(A^*)$ denotes the entropy of the optimal action. Rate distortion is never larger than entropy, but it may be small even if the entropy of A^* is infinite.

The following, somewhat artificial, example explicitly links communication with decision making and may help clarify the role of the rate-distortion function $\mathcal{R}(D)$.

Example 7. A military command center waits to hear from an outpost before issuing orders. The outpost, stationed close to the conflict, determines its message based on a wealth of nuanced information – at the level of readouts from weather sensors and full transcripts of intercepted enemy communication. The command center could make very complicated decisions as a function of the detailed information it receives, with the possibility of specifying commands at the level of individual troops and equipment. How much must decision quality degrade if decisions are based only on coarser information? At an intuitive level, the outpost only needs to communicate surprising revelations that are important to reaching a satisfactory decision. As a result, our answer can depend in a complicated way on the extent to which the outpost's observations are predictable a priori and the extent to which decision quality is reliant on this information. The rate-distortion function precisely quantifies these effects.

To map this problem onto our formulation of the rate-distortion function, take θ to consist of all information observed by the outpost, $\tilde{A}$ to be the order issued by the command center, and the rewards to indicate whether

the orders led to a successful outcome. The mutual information $I(\tilde{A};\theta)$ captures the average amount of information the outpost must send in order for $\tilde{A}$ to be implemented. The goal is to develop a plan for placing orders that requires minimal communication from the outpost among all plans that degrade the chance of success by no more than D. □

5.2. Uniformly Bounded Information Ratios

The general regret bound in Theorem 1 has a superficial relationship to the rate-distortion function $\mathcal{R}(D)$ through its dependence on the mutual information $I(\tilde{A};\theta)$ between the benchmark action and the true parameter θ. Indeed, for a benchmark action $\tilde{A}$ attaining the rate-distortion limit, $I(\tilde{A};\theta) = \mathcal{R}(D)$; we attain a regret bound that depends explicitly on the rate-distortion level. However, the information ratio $\Gamma(\tilde{A},\psi)$ also depends on the choice of benchmark action and may be infinite for a poor choice.

This second dependence on $\tilde{A}$ does not appear in rate-distortion theory and reflects a fundamental distinction between communication problems and sequential learning problems. Indeed, a key feature enabling the sharp results of rate distortion theory is that no bit of information is more costly to send and receive than others; the question is to what extent useful communication is possible while transmitting many fewer bits of information on average. By contrast, sequential learning agents must explore to uncover information and the cost per unit of information uncovered may vary widely depending on which pieces of information are sought. This is accounted for by the information ratio $\Gamma(\tilde{A},\psi)$, which roughly captures the expected cost, in terms of regret incurred, per bit of information acquired about the benchmark action.

Despite this, regret bounds in terms of rate distortion apply in many important cases. Theorem 2, which is an immediate consequence of Theorem 1, provides a general bound of this form. Roughly, the uniform information ratio Γ_U in the theorem reflects something about the quality of the feedback the agent receives when exploring; it means that for *any* choice of benchmark action $\tilde{A}$, there is a sequential learning strategy that learns about $\tilde{A}$ with the cost per bit of information less than Γ_U. The next section applies this result to online linear optimization, where several possible uniform information ratio bounds are possible depending on the problem's precise feedback structure.

Theorem 2. *Suppose there is a uniform bound on the achievable information ratio*

$$\Gamma_U := \sup_{\tilde{A}} \inf_{\psi} \Gamma(\tilde{A},\psi) < \infty.$$

Then, for any $D \geq 0$,

$$\inf_{\psi} \operatorname{SRegret}(\alpha,\psi,D) \leq \sqrt{\frac{\Gamma_U \mathcal{R}(D)}{1-\alpha^2}}.$$

6. Application to Online Linear Optimization

Consider a special case of our formulation in which the decision maker learns to solve a linear optimization problem. Precisely, suppose expected rewards follow the linear model $\mathbb{E}[R_t \mid \theta, A_t] = \theta^\top A_t$ where $A_t \in \mathcal{A} \subset \mathbb{R}^p$, $\theta \in \mathbb{R}^p$ and $R_t \in \left[-\frac{1}{2},\frac{1}{2}\right]$ almost surely. We will consider several natural forms of a feedback Y_t the decision maker may receive.

In each case, uniform bounds on the information ratio hold for satisficing Thompson sampling. More precisely, for any $\tilde{A}$, let $\psi_{\tilde{A}}^{\mathrm{STS}}$ denote the strategy that randomly samples an action at each time t by probability matching with respect to $\tilde{A}$, that is, $\mathbb{P}(A_t \in \cdot \mid \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{A} \in \cdot \mid \mathcal{H}_{t-1})$. Applying the same proofs as in Russo and Van Roy [21] yields bounds of the form $\Gamma(\tilde{A};\psi_{\tilde{A}}^{\mathrm{STS}}) \leq \lambda$, where λ depends on the problem's feedback structure but not the choice of benchmark action. Throughout this section, we assume for simplicity that the minimum in (7) is attained[1] and let $\tilde{A}$ be a minimizer, so $I(\tilde{A};\theta) = \mathcal{R}(D)$ and $\mathbb{E}[R^* - \mu(\tilde{A};\theta)] \leq D$. We denote by ψ_D^{STS} satisficing Thompson sampling with respect to this choice of satisfactory action.

Now, let us choose $\tilde{A}$ to attain the rate distortion limit (7), so $I(\tilde{A};\theta) = \mathcal{R}(D)$ and $\mathbb{E}[R^* - \mu(\tilde{A};\theta)] \leq D$. We denote by ψ_D^{STS} satisficing Thompson sampling with respect to this choice of satisfactory action.

6.1. Full Information

Suppose $R_t = A_t^\top Z_t$ for a random vector Z_t with $\mathbb{E}[Z_t \mid \theta, \mathcal{H}_{t-1}] = \theta$. This is an extreme point of our formulation, where all information is revealed without active exploration. For all $\tilde{A}$, the information ratio is bounded as $\Gamma(\psi_{\tilde{A}}^{\mathrm{STS}}; \tilde{A}) \le 1/2$ and hence

$$\mathrm{SRegret}(\alpha, \psi_D^{\mathrm{STS}}, D) \le \sqrt{\frac{\mathcal{R}(D)}{2(1-\alpha^2)}}.$$

6.2. Bandit Feedback

Suppose the agent only observes the reward the action he or she chooses ($Y_t = R_t$). This is the so-called linear bandit problem. For all $\tilde{A}$, the information ratio is bounded as $\Gamma(\psi_{\tilde{A}}^{\mathrm{STS}}; \tilde{A}) \le p/2$. This gives the regret bound

$$\mathrm{SRegret}(\alpha, \psi_D^{\mathrm{STS}}, D) \le \sqrt{\frac{\mathcal{R}(D)p}{2(1-\alpha^2)}}.$$

6.3. Semi-Bandit Feedback

Assume again that $R_t = A_t^\top Z_t$ for all a. Take the action set $\mathcal{A} \subset \{0,1\}^p$ to consist of binary vectors where $\sum_{i=1}^p a_i \le m$ for all $a \in \mathcal{A}$. Upon playing action $A_t = a$, the agent observes $Z_{t,i}$ for every component $\{i \in \{1,\dots,m\} : a_i = 1\}$ that was active in a. We make the additional assumption that the components of Z_t are independent conditioned on $\mathcal{H}_{t-1}$. Then, for all $\tilde{A}$, the information ratio is bounded as $\Gamma(\psi_{\tilde{A}}^{\mathrm{STS}}; \tilde{A}) \le p/2m$ and hence

$$\mathrm{SRegret}(\alpha, \psi_D^{\mathrm{STS}}, D) \le \sqrt{\frac{\mathcal{R}(D)(p/m)}{2(1-\alpha^2)}}.$$

By following the appendix of Russo and Van Roy [21], each of these results can be extended gracefully to settings where noise distributions are *sub-Gaussian*. For example, suppose θ follows a multivariate Gaussian distribution and the reward at time t is $R_t = \theta^\top A_t + W_t$ where W_t is a zero mean Gaussian random variable. Then, if the variance of rewards $\mathbb{E}[(\theta^\top a + W - \mathbb{E}[\theta]^\top a)^2] \le \sigma^2$ is bounded by some σ^2 for all a, the previous bounds on the information ratio scale by a factor of σ.

It is worth noting that these results immediately reduce to bounds on (nonsatisficing) regret when the action space is finite. As mentioned in Subsection 5.1, for problems with a finite action set $\mathcal{R}(D) \le H(A^*) \le \log|\mathcal{A}|$ for all $D \ge 0$. For example, with a linear bandit with finite action set, (1) gives the bound

$$\mathrm{Regret}(\alpha, \psi_D^{\mathrm{STS}}) \le \sqrt{\frac{H(A^*)p}{2(1-\alpha^2)}} + \frac{D}{1-\alpha},$$

so STS with a small satisficing level attains desirable regret bounds when the action set is not too large. A special case of this problem is the classical k-armed bandit problem, in which case $p = k$. We can reach similar conclusions in other cases by arguing $\mathcal{R}(D)$ grows slowly as $D \to 0$. We illustrate this idea for a Gaussian linear bandit below.

The next theorem considers an explicit choice of satisfactory action $\tilde{A}$. This yields a computationally efficient version of STS as well as explicit upper bounds on the rate-distortion function. Consider the case where $\theta \sim N(\mu, \Sigma)$ follows a multivariate Gaussian prior and reward noise is Gaussian. We study the optimizer $\tilde{A} = \arg\max_{a \in \mathcal{A}} \langle a, \theta + \xi \rangle$ of a randomly perturbed objective. The small perturbation controls the mutual information between $\tilde{A}$ and θ without substantially degrading decision quality. It is easy to implement probability matching with respect to $\tilde{A}$ whenever linear optimization problems over $\tilde{A}$ are efficiently solvable. In particular, if $\mu_t = \mathbb{E}[\theta \mid \mathcal{H}_{t-1}]$ and $\Sigma_t = \mathbb{E}[(\theta - \mu_t)(\theta - \mu_t)^\top \mid \mathcal{H}_{t-1}]$ denote the posterior mean and covariance matrix, which are efficiently computable using Kalman filtering, then by sampling $\hat{\theta}_t \sim N(\mu_t, \Sigma_t)$ and $\hat{\xi}_t \sim N(0, \Sigma_\xi)$ and setting $A_t = \arg\max_{a \in \mathcal{A}} \langle a, \hat{\theta}_t + \hat{\xi}_t \rangle$, one has $\mathbb{P}(A_t \in \cdot \mid \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{A} \in \cdot \mid \mathcal{H}_{t-1})$.

The result in the next theorem assumes the action set is contained within an ellipsoid $\{x \in \mathbb{R}^p : x^\top Q^{-1} x \le 1\}$, and the resulting bound displays a logarithmic dependence on the eigenvalues of Q. Precisely, note that the trace of the matrix Q, or sum of its eigenvalues, provides one natural measure of the size of the ellipsoid. Our result also depends on the covariance matrix Σ through $\mathrm{Trace}(Q\Sigma)$. To understand this, consider applying similarity transforms to the parameter and action vectors so that $\theta' = \Sigma^{-1/2}\theta$ is isotropic and the set of action vectors is

$\mathcal{A}' = \{\Sigma^{1/2}a : a \in \mathcal{A}\}$. This transformed action space is contained in the ellipsoid $\{x : x^\top Q'^{-1}x \le 1\}$, where $Q' = \Sigma^{1/2}Q\Sigma^{1/2}$. Then $\text{Trace}(Q') = \text{Trace}(Q\Sigma)$ provides a measure of the size of this ellipsoid.

Theorem 3. *Suppose $\mathcal{A}$ is a compact subset of the ellipsoid $\mathcal{A} \subset \{x \in \mathbb{R}^p : x^\top Q^{-1}x \le 1\}$ for some symmetric positive definite matrix Q and suppose $\theta \sim N(\mu, \Sigma)$ follows a p-dimensional multivariate Gaussian distribution. Set*

$$\tilde{A} = \arg\max_{a\in\mathcal{A}} \langle a, \theta + \xi\rangle,$$

where ξ is independent of θ and $\xi \sim N(0, \beta^2\Sigma)$. When $\beta = D/\sqrt{\text{Trace}(Q\Sigma)}$,

$$\mathbb{E}[\langle \theta, \tilde{A}\rangle] \ge \mathbb{E}[\langle \theta, A^*\rangle] - D$$

and

$$\mathcal{R}(D) \le I(\tilde{A};\theta) \le \frac{p}{2}\log\left(1 + \frac{\text{Trace}(Q\Sigma)}{D^2}\right).$$

Proof. By Jensen's inequality,

$$\mathbb{E}[\langle \tilde{A}, \theta + \xi\rangle] = \mathbb{E}\left[\max_{a\in\mathcal{A}} \langle a, \theta + \xi\rangle\right] \ge \mathbb{E}\left[\max_{a\in\mathcal{A}} \langle a, \theta\rangle\right] = \mathbb{E}[\langle A^*, \theta\rangle].$$

This implies

$$\mathbb{E}[\langle A^*, \theta\rangle] - \mathbb{E}[\langle \tilde{A}, \theta\rangle] \le \mathbb{E}[\langle \tilde{A}, \xi\rangle] \le \mathbb{E}\left[\max_{a\in\mathcal{A}} \langle a, \xi\rangle\right] \le \mathbb{E}\left[\max_{x:\|x\|_{Q^{-1}}\le 1} \langle x, \xi\rangle\right] = \mathbb{E}[\|\xi\|_Q],$$

where the final equality uses the explicit formula for the maximum of a linear function over an ellipsoid. Now,

$$\mathbb{E}[\|\xi\|_Q] \le \sqrt{\mathbb{E}[\xi^\top Q\xi]} = \sqrt{\mathbb{E}[\text{Trace}(\xi^\top Q\xi)]} = \sqrt{\text{Trace}(Q\mathbb{E}[\xi\xi^\top])} = \sqrt{\beta^2\text{Trace}(Q\Sigma)} = D.$$

Next we derive the bound on mutual information. We have

$$\begin{aligned} I(\tilde{A};\theta) \le I(\theta + \xi;\theta) &= H(\theta + \xi) - H(\theta + \xi|\theta) \\ &= H(\theta + \xi) - H(\xi) \\ &= \frac{1}{2}\log\left(\frac{\det(\Sigma + \beta^2\Sigma)}{\det(\beta^2\Sigma)}\right) \\ &= \frac{1}{2}\log\left(\frac{\det((1+\beta^2)\Sigma)}{\det(\beta^2\Sigma)}\right) \\ &= \frac{p}{2}\log\left(1 + \frac{1}{\beta^2}\right) \\ &= \frac{p}{2}\log\left(1 + \frac{\text{Trace}(Q\Sigma)}{D^2}\right) \end{aligned}$$

where $\det(\cdot)$ denotes the determinant of a matrix. Here the first inequality uses the data processing inequality, the third equality uses the explicit form the entropy of a multivariate Gaussian ($H(\theta) = \frac{1}{2}\log(2\pi e\ \det(\Sigma))$), and the penultimate equality uses that $\det(c\Sigma) = c^p\det(\Sigma)$ for any scalar c. □

7. Application to the Infinite-Armed Bandit

This section considers a generalization of the deterministic infinite-armed bandit problem in the introduction that allows for noisy observations and nonuniform priors. The action space is $\mathcal{A} = \{1, 2, \dots\}$. We assume $R_t \in [0,1]$ almost surely and $Y_t = R_t$, meaning the agent only observes rewards. The mean reward of action a is $\mu(\theta, a) = \theta_a \in [0,1]$ where the prior distribution of θ_a is independent across $a \in \mathcal{A}$. Let R^* denote the supremal value in the support of θ_a, so $\sup_{a\in\mathcal{A}}\theta_a = R^*$ almost surely.

7.1. STS for the Infinite-Armed Bandit Problem

We consider the simple satisficing action defined in the introduction: $\tilde{A} = \min\{a \in \mathcal{A} : \theta_a \ge R^* - D\}$. Rather than continue to explore until identifying the optimal action A^*, we will settle for the *first*[2] action known to yield reward within D of optimal.

We study satisficing Thompson sampling where actions are selected by probability matching with respect to $\tilde{A}$. Note that an algorithm for this problem must decide whether to sample a previously tested action—and if so which one to sample—or whether to try out an entirely new action. Let $\mathcal{A}_t = \{A_0, \dots, A_{t-1}\}$

denote the set of previously sampled actions. STS may sample an untested action $A_t \notin \mathcal{A}$ and does so with probability

$$\mathbb{P}(A_t \notin \mathcal{A}_t | \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{A} \notin \mathcal{A}_t | \mathcal{H}_{t-1}),$$

that is, with probability equal to the posterior probability no satisfactory action has yet been sampled. The remainder of the action probabilities are allocated among previously tested actions, with

$$\mathbb{P}(A_t = a | \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{A} = a | \mathcal{H}_{t-1}) \qquad \forall a \in \mathcal{A}_t.$$

There is a simple algorithmic implementation of STS that mirrors computationally efficient implementations of Thompson sampling. At time t, TS selects a random action A_t via probability matching with respect to A^*. Algorithmically, this is usually accomplished by first sampling $\hat{\theta}_t \sim \mathbb{P}(\theta \in \cdot | \mathcal{H}_{t-1})$ and solving for $A_t \in \arg\max_{a \in \mathcal{A}} \mu(a, \hat{\theta}_t)$. Similarly, we can implement STS by sampling and approximately optimizing a posterior sample. Over each tth period, STS selects an action A_t as follows:

1. For each $a \in \mathcal{A}_t$, sample $\hat{\theta}_a \sim \mathbb{P}(\theta_a \in \cdot | \mathcal{H}_{t-1})$.
2. Let $\hat{\tau} = \min\{\tau \in \{1, \ldots, t-1\} : \mu(A_\tau, \hat{\theta}_t) \geq R^* - D\}$.
3. If $\hat{\tau}$ is not null, set $A_t = A_{\hat{\tau}}$. Otherwise, play an untested action $A_t = |\mathcal{A}_t| + 1$.

In the third step, it not important which previously untested action is selected. We specify a choice only for convenience. Note that $D \geq 0$ is supplied to the algorithm as a tolerance parameter. When $D = 0$, STS is equivalent to TS. Otherwise, STS attributes preference to selecting previously selected actions, which can yield substantial benefit in the face of time preference.

This definition can be generalized to treat problems with a large, but finite, number of independent arms. Define the satisficing action $\tilde{A} = \min\{a \in \mathcal{A} | \theta_a \geq \sup_{a' \in \mathcal{A}} \mu(\theta, a') - D\}$. For the infinite-armed bandit, $\sup_{a' \in \mathcal{A}} \mu(\theta, a') = R^*$ with probability 1; but for problems with a finite number of arms, $\max_{a' \in \mathcal{A}} \mu(\theta, a')$ is not known a priori. One can efficiently sample from this satisficing action by modifying step 2 above with the alternative definition

$$\hat{\tau} = \min\left\{\tau \in \{1, \ldots, t-1\} : \mu(A_t, \hat{\theta}_t) \geq \sup_{a \in \mathcal{A}} \mu(a, \hat{\theta}_t) - D\right\}.$$

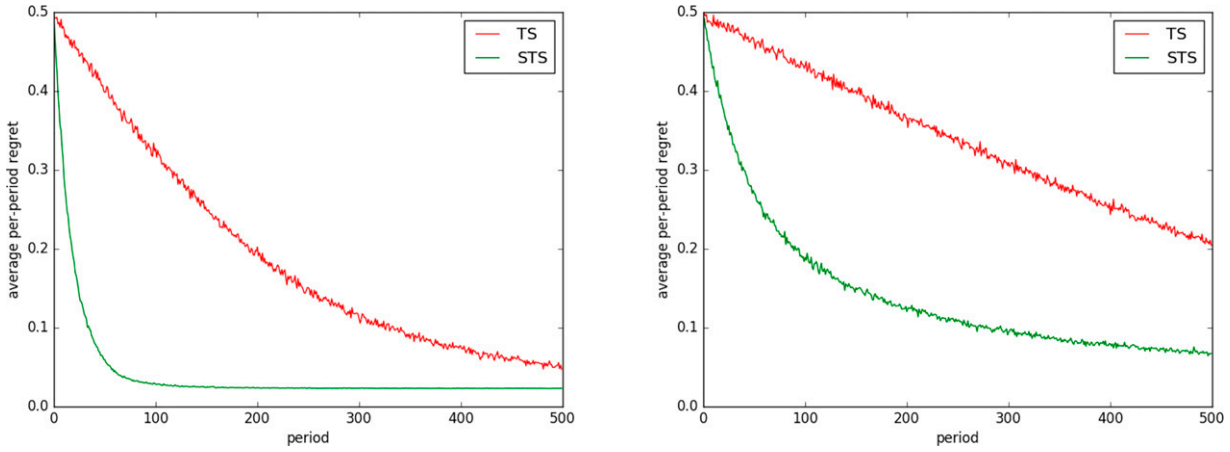
This version is applied in the numerical experiments in the next subsection.

7.2. Computational Comparison of STS and TS

We begin with a simple computational illustration of the potential benefits afforded by STS. We consider two many-armed bandit problems and demonstrate that per-period regret of STS diminishes much more rapidly than that of TS over early time periods.

We consider problems with 250 actions, where the mean reward θ_a associated with each action $a \in \{1, \ldots, 250\}$ is independently sampled uniformly from $[0, 1]$. We first consider the many-armed deterministic bandit, for which there is no observation noise. Figure 1(a) presents per-period regret of TS and STS over 500 time periods, averaged over 5,000 simulations, each with an independently sampled problem instance. STS is applied with

Figure 1. (Color online) TS versus STS for the (a) many-armed deterministic bandit and (b) many-armed bandit with observation noise.



tolerance parameter 0.05. We next consider incorporating observation noise. In particular, instead of observing θ_a, after selecting an action a, we observe a binary reward that is one with probability θ_a. Figure 1(b) displays the results of this experiment.

7.3. Information Ratio Analysis of the Infinite-Armed Bandit

The following theorem provides a discounted regret bound for STS in the infinite-armed bandit. The result follows from bounding the problem's information ratio and the mutual information $I(\tilde{A};\theta)$ and applying the general regret bound of Theorem 1. This requires substantial additional analysis, the details of which are delayed until Section B.1 in Appendix B.

Theorem 4. *Consider the infinite-armed bandit with noisy observations, and let* $\tilde{A} = \min\{a \in \mathcal{A} : \theta_a \geq R^* - D\}$. *Denote the STS policy with respect to* $\tilde{A}$ *by* ψ_D^{STS}. *Then,*

$$I(\tilde{A};\theta) \leq 1 + \log(1/\delta) \quad \textit{and} \quad \Gamma(\tilde{A}, \psi_D^{\mathrm{STS}}) \leq 2 + 2/\delta + (2/\delta)\log\left(\frac{1}{1-\alpha^2}\right),$$

where $\delta = \mathbb{P}(\theta_a \geq R^* - D)$. *Together with Theorem 1 this implies*

$$\begin{aligned}\mathrm{SRegret}(\alpha, \psi_D^{\mathrm{STS}}, D) &\leq \min\left\{\sqrt{\frac{\left(2 + 4/\delta + (2/\delta)\log\left(\frac{1}{1-\alpha^2}\right)\right)(1+\log(1/\delta))}{1-\alpha^2}}, \frac{1}{1-\alpha}\right\} \\ &= \tilde{O}\left(\min\left\{\sqrt{\frac{1/\delta}{1-\alpha}}, \frac{1}{1-\alpha}\right\}\right). \end{aligned} \tag{8}$$

The final expression (8) uses that $1-\alpha \leq 1-\alpha^2 \leq 2(1-\alpha)$. The upper bound of $1/(1-\alpha)$ is naive and applies to all algorithms. The bound of order $\sqrt{\frac{1/\delta}{1-\alpha}}$ requires an intelligent balance of exploration and exploitation; it is much stronger when the prior probability δ an arm is a satisficing arm is not too small.

7.4. A Lower Bound for the Infinite-Armed Bandit

This section establishes a lower bound on regret that matches the scaling of the upper bound in Theorem 4. In this problem, ϵ is the fraction of arms that are exactly optimal. For our purposes, the interesting regime is where $\epsilon \ll 1-\alpha$ but $\delta \gg 1-\alpha$, in which case identifying an optimal arm is hopeless; but it is worthwhile to search for a satisficing arm. In that regime, Theorem 4 shows a bound on satisficing regret of $\tilde{O}\left(\sqrt{\frac{1/\delta}{1-\alpha}}\right)$ and Theorem 5 shows this is unimprovable in general. The specific threshold on ϵ in the theorem statement was chosen for analytical convenience and could likely be tightened. The full proof of this result is provided in Section B.2 in Appendix B.

Theorem 5. *Fix any* $\alpha \in (0,1)$, $\delta \in (0,1/2)$ *and* $D \in (0,1/4)$. *Consider an instance of the infinite-armed bandit problem in which, for all* $a \in \mathcal{A}$,

$$\mathbb{P}\left(\theta_a = \frac{1}{2} - \Delta\right) = 1-\delta \qquad \mathbb{P}\left(\theta_a = \frac{1}{2}\right) = \delta - \epsilon \qquad \mathbb{P}\left(\theta_a = \frac{1}{2} + D\right) = \epsilon, \tag{9}$$

for $\Delta = \min\left\{\frac{1}{4}, \frac{1}{4\sqrt{2}} \cdot \sqrt{\frac{1-\alpha}{\delta}}\right\}$ *and* $\epsilon \leq \frac{1}{64} \cdot \min\left\{(1-\alpha)^2, \frac{(1-\alpha)^3}{4\delta}\right\}$. *Suppose* $R_t \in \{0,1\}$ *with* $\mathbb{P}(R_t = 1 | \theta, A_t, \mathcal{H}_{t-1}) = \theta_{A_t}$. *Then,*

$$\inf_{\psi} \mathrm{SRegret}(\alpha, \psi, D) \geq \frac{1}{32} \cdot \min\left\{\sqrt{\frac{1/4\delta}{1-\alpha}}, \frac{1}{1-\alpha}\right\},$$

where the infimum is over all adaptive policies.

7.5. Open Question: Gap Dependent Analysis of STS

It should be noted that Theorem 5 is a worst-case construction. It shows that, for a given discount factor α, satisficing level D and fraction of satisficing arms δ, there exists a hard instance of an infinite armed bandit problem in which the scaling of (8) is unavoidable. Stronger performance guarantees are likely possible for more benign problems, however. Here, we highlight one open problem in this direction.

As shown in Berry et al. [2] and Wang et al. [29], when the agent begins with a uniform prior over each θ_a, it is possible to attain undiscounted regret that scales as $\mathbb{E}[\sum_{t=1}^{T}(R^* - R_t)] = O(\sqrt{T})$. Although the worst-case

construction in Theorem 5 shows that Theorem 4 is tight without additional assumptions, it seems to yield an overly conservative regret bound of $O(T^{2/3})$ for this specific problem. To understand this, note that if the agent begins with a uniform prior over each θ_a, we find $\delta = D$. Choosing $D \approx (1-\alpha)^{1/3}$ to minimize the regret upper bound from combining Theorem 4 with Equation (5), we find $\text{Regret}(\alpha, \psi^D) \le \tilde{O}\left(\frac{1}{(1-\alpha)^{2/3}}\right)$, where ψ^D denotes satisficing Thompson sampling applied with parameter D. Because $\frac{1}{1-\alpha}$ is the effective time horizon in the problem, this roughly corresponds to a regret bound of $\tilde{O}(T^{2/3})$.

Simulations suggest that such a bound is conservative, and STS actually attains the optimal $\Theta(\sqrt{T})$ regret scaling in this problem. To test this, we applied STS over a range of horizons $T \in \{e^5, e^{5.5}, \cdots, e^{10.5}\}$. For each choice of horizon, we used the satisficing parameter $D = 3/\sqrt{T}$ and ran 500 independent trials. The numerical constant 3 was selected in a somewhat ad hoc manner and may be further optimized by tuning it in simulation. Figure 2 seems to suggest $\mathbb{E}[\sum_{t=1}^{T}(R^* - R_t)] = \Theta(\sqrt{T})$ because the logarithm of regret scales linearly as $\log(T)/2$. An analogous experiment in the discounted setting suggests $\text{Regret}(\alpha, \psi_{\text{D}}) = \Theta\left(\frac{1}{\sqrt{1-\alpha}}\right)$ as $\alpha \to 1$.

To understand the two different scalings of regret, it may be helpful to draw an analogy to the standard k-armed bandit problem. When there is a fixed gap of $\Delta > 0$ between the best and second best arm, it is possible to provide regret bounds of $O(\log(T)/\Delta)$, which scale very slowly with the time horizon but degrade as Δ shrinks. In worst-case instances, $\Delta \approx 1/\sqrt{T}$ is small relative to the horizon and regret bounds of $O(\sqrt{T})$, which are completely independent of the gap, are the best possible. Our analysis in Theorem 1 was effectively gap independent, as it did not make assumptions about the separation between satisficing actions a nonsatisficing actions. By imposing the additional assumption that each θ_a is drawn from a uniform prior, we ensure that most arms are easily distinguished from the satisficing arms—especially as the satisficing threshold D tends to zero—which is what makes the $O(\sqrt{T})$ bound possible. One may be able to leverage finite time gap-dependent analyses of Thompson sampling (Agrawal and Goyal [1]) to show an $O(\sqrt{T})$ regret bound for STS under a restricted class of priors, but we leave this as an open question that is beyond the scope of this work.

8. Application of STS to the Hierarchical Infinite-Armed bandit

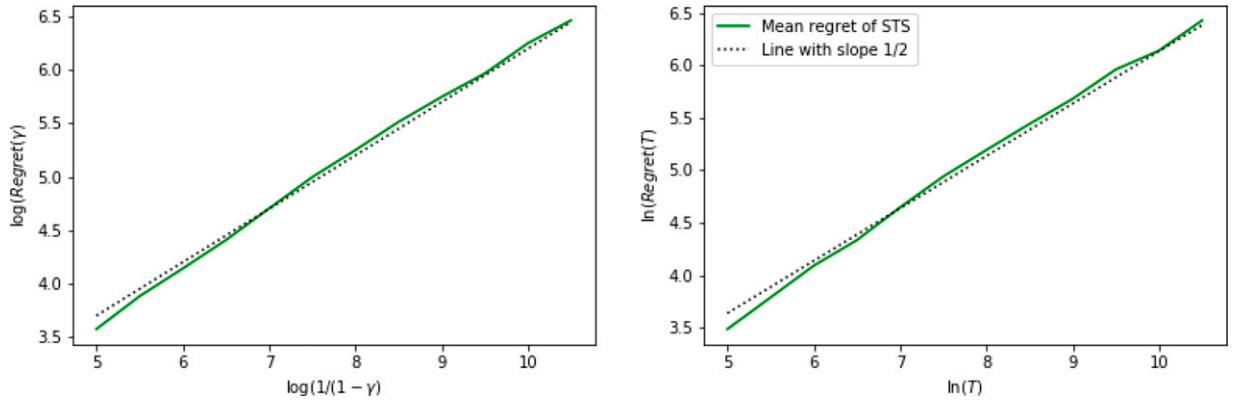
This section offers a preliminary study of satisficing in the hierarchical infinite armed bandit as described in Example 2 of Section 1. Algorithms that succeed on this example must simultaneously leverage prior beliefs, generalize across arms, and satisfice.

Example 4 of Section 3 describes a satisficing action $\tilde{A}$ for this problem. Precisely, according to Equation (2), $\tilde{A} = \min\{a \in \mathcal{A}(\theta_0) | \theta_{1a} \ge 1 - D\}$, where $\mathcal{A}(\theta_0) = \arg\max_{a' \in \mathcal{A}} \langle \phi(a'),\ \theta_0 \rangle\}$. Satisficing Thompson sampling performs probability matching with respect to $\tilde{A}$, setting $\mathbb{P}(A_t = a | \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{A} = a | \mathcal{H}_{t-1})$ for each $a \in \mathcal{A}$.

Let us describe how to implement probability matching in a manner that mirrors the description of STS for the infinite-armed bandit. Let $\mathcal{A}_t = \{A_0, \ldots A_{t-1}\}$ denote the set of previously sampled actions and $\mathcal{A}_t(\theta_0) = \mathcal{A}_t \cap \mathcal{A}(\theta_0)$ denote previously sampled actions whose feature vectors are optimal under parameter $\theta_0 \in \mathbb{R}^d$. Let $\theta_{1\mathcal{A}_t(\theta_0)} = \{\theta_{1a} : a \in \mathcal{A}_t(\theta_0)\}$ denote the corresponding components of θ_1. Over each tth period, STS selects an action A_t as follows:

1. Sample $\hat{\theta}_0 \sim \mathbb{P}(\theta_0 \in \cdot | \mathcal{H}_{t-1})$.

Figure 2. (Color online) Scaling of regret in an infinite-armed bandit with uniform prior for (a) a discounted infinite horizon and (b) an undiscounted finite horizon problem.



2. Sample $\hat{\theta}_{1\mathcal{A}_t(\hat{\theta}_0)} \sim \mathbb{P}(\theta_{1\mathcal{A}_t(\theta_0)} \in \cdot \mid \theta_0 = \hat{\theta}_0, \mathcal{H}_{t-1})$.
3. Let $\hat{\tau} = \min\{\tau \in \{1, \ldots, t-1\} : \hat{\theta}_{1A_\tau} \geq 1 - D\}$.
4. If $\hat{\tau}$ is not null, set $A_t = A_{\hat{\tau}}$. Otherwise, play an untested action $A_t \in \mathcal{A}(\hat{\theta}_0) \setminus \mathcal{A}_t$.

Steps 2–4 correspond to satisficing Thompson sampling for the infinite-armed bandit as described in Section 7 but applied conditioned on $\theta_0 = \hat{\theta}_0$. Using the ideas described in Subsection 7.1, a simple modification of this applies to problems with a large but finite number of arms and a prior that is not necessarily uniform. We run STS with the satisficing action

$$\tilde{A} = \arg\min_{a' \in \mathcal{A}(\theta_0)} \left\{ a' \mid \theta_{1a'} \geq \max_{a \in \mathcal{A}(\theta_0)} \theta_{1a} - D \right\}. \tag{10}$$

As in Subsection 7.1, we can sample from this action by applying the steps above with a modified definition of $\hat{\tau}$.

We run a simple numerical experiment to demonstrate the importance of satisficing and generalization to this problem. To simplify the implementation, we focus on the case where there is a normally distributed prior over θ_0 and each θ_{1a}, allowing the posterior distribution to be expressed in closed form. The experiment treats a problem with a large but finite number of arms, highlighting that satisficing is just as important in such settings.

Results are displayed in Figure 3. The simulation focuses on performance over the first $T = 50$ periods. The dimension of the linear model is $d = 2$. There is no observation noise, so the reward at time t is $\langle \phi(A_t), \theta_0 \rangle + \theta_{A_t}$. There are $k = 400$ actions, each of which has a feature vector $\phi(a) \in \{0,1\}^2$ that is drawn randomly. Parameters are drawn randomly under a prior with $\theta_0 \sim N(0, d^{-1/2} I_d)$ and $\theta_{1,a} \sim N(0,1)$ independently across arms a and from θ_0. We run STS with the satisficing action in (10).

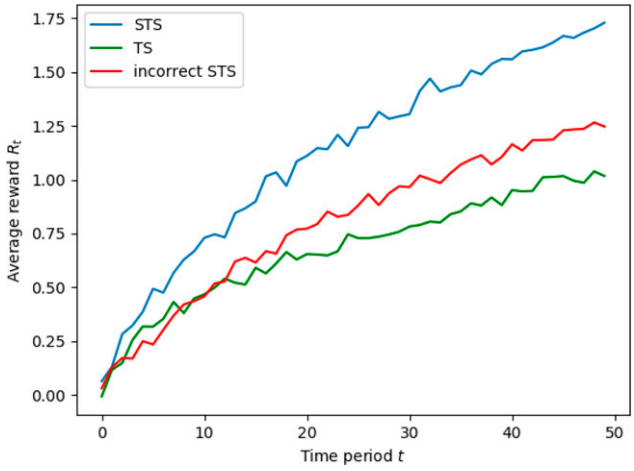
We compare the performance of STS against Thompson sampling and an incorrect version of STS that does not model the linear component of the model, instead treating the problem as an many-armed bandit with independent arms and normally distributed prior. We choose $D = 1$ for both variants of STS. Figure 3 shows the average reward earned by each algorithm in each of the first 50 periods. Results are averaged across 2,000 trials.[3] We see that STS earns much higher rewards in early periods, earning an average reward across the 50 time periods of 1.21, compared with 0.79 for incorrect STS and 0.67 for standard Thompson sampling. In this case, Thompson sampling generalizes across arms but does not satisfice, preventing it from settling on arms with large arm-specific effect. The incorrect variant of STS satisfices but does not generalize across arms. STS can simultaneously generalize and satisfice, leading to large improvement in performance.

9. Closing Remarks

We have put forth a way of thinking about satisficing in bandit learning. The per-period regret of an algorithm that learns a satisficing action $\tilde{A}$ does not converge to zero, but the time required can often be far less than what it would be to learn an optimal action A^*. Intuitively, this advantage stems from the fact that the mutual information $I(\theta; \tilde{A})$, which can be thought of as the number of bits of information about the model required to learn $\tilde{A}$, can be far less than $I(\theta; A^*)$.

Satisficing plays a particularly important role when there is a satisficing action $\tilde{A}$ for which $I(\theta; \tilde{A}) \ll I(\theta; A^*)$ and the agent exhibits time preference, valuing near-term over long-term rewards. To express this in terms of a formal objective, we considered expected discounted regret. We also introduced satisficing Thompson sampling

Figure 3. (Color online) Average reward earned in early periods of the hierarchical bandit problem.



and established results pertaining to infinite-armed and linear bandits demonstrating that this variant of Thompson sampling captures benefits of targeting a satisficing action.

We believe this paper puts forth a useful conceptual framework and that satisficing Thompson sampling could serve as a useful design principle for time-sensitive learning problems. But we also feel this paper takes only a very preliminary step toward understanding satisficing in modern bandit learning. Future work might try to analyze the rate-distortion function and information ratio for broader classes of problems than addressed in this paper. We suspect models like the hierarchical bandit in Example 2 may be quite useful in practice and warrant more complete study. Satisficing is even more important in reinforcement learning than in bandit problems. Most ideas in this paper extend gracefully to contextual bandits, but extensions to reinforcement seem difficult and are a fascinating direction for future work.

Acknowledgments

A special thanks is owed to David Tse, who played an important role in the early stages of this work. It was David who first emphasized that bounds based on entropy can be vacuous and pointed us to references on rate-distortion theory. The authors also thank Tor Lattimore for thoughtful comments on an early draft of this work.

Appendix A. Analysis of the Infinite-Armed Deterministic Bandit

Theorem A.1. *For all $\alpha \in [0,1]$, under Thompson sampling in the infinite-armed deterministic bandit,*

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - \theta_{A_t})\right] = \frac{1}{2(1-\alpha)}.$$

Under satisficing Thompson sampling with tolerance $\epsilon = \sqrt{1-\alpha}$ in the infinite-armed deterministic bandit,

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - \theta_{A_t})\right] \le \frac{1}{\sqrt{1-\alpha}}.$$

Proof. In every period t, TS samples a new action $A_t \notin \{A_1, \ldots, A_{t-1}\}$, which generates expected reward $\mathbb{E}[\theta_{A_t}] = \mathbb{E}[\theta_1] = 1/2$. The optimal expected reward is one, and therefore the expected discounted-regret of TS is

$$\sum_{t=0}^{\infty} \alpha^t (1 - 1/2) = \frac{1}{2(1-\alpha)}.$$

Now, let us analyze satisficing Thompson sampling. Let $\tau = \min\{t : \theta_{A_t} \ge 1 - \epsilon\}$ denote the first time it sampled a ϵ–optimal action. Then,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - \theta_{A_t})\right] &= \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (1 - \theta_{A_t})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t (1 - \theta_{A_t}) + \sum_{t=\tau}^{\infty} \alpha^t (1 - \theta_{A_t}) \middle| \tau\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t \mathbb{E}[1 - \theta_a | \theta_a \le 1 - \epsilon] + \sum_{t=\tau}^{\infty} \alpha^t \mathbb{E}[1 - \theta_a | \theta_a \ge 1 - \epsilon] \middle| \tau\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t (1/2 + \epsilon/2) + \sum_{t=\tau}^{\infty} \alpha^t (\epsilon/2) \middle| \tau\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t (1/2) + \sum_{t=0}^{\infty} \alpha^t (\epsilon/2) \middle| \tau\right]\right] \\
&= \mathbb{E}\left[\frac{(1-\alpha^\tau)}{2(1-\alpha)} + \frac{\epsilon}{2(1-\alpha)}\right] \\
&\le \left(\frac{1}{2\epsilon} + \frac{\epsilon}{2(1-\alpha)}\right). \qquad \text{(A.1)}
\end{aligned}$$

The inequality (Equation (A.1)) above follows from the calculation

$$\mathbb{E}[1 - \alpha^\tau] = 1 - \sum_{t=0}^{\infty} \epsilon (1-\epsilon)^t \alpha^t = 1 - \frac{\epsilon}{1 - \alpha(1-\epsilon)} = \frac{1 - \alpha(1-\epsilon) - \epsilon}{1 - \alpha(1-\epsilon)} \le \frac{1-\alpha}{\epsilon}.$$

The final bound of $\sqrt{1/(1-\alpha)}$ follows by choosing the minimizer $\epsilon^* = \sqrt{1-\alpha}$ of Equation (A.1). □

Appendix B. Proofs of Infinite Armed Bandit Results

B.1. Proof of the Upper Bound: Theorem 4

Our proof will use the following fact, which is a consequence of Pinsker's inequality and is stated as fact 9 in Russo and Van Roy [21].

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

$$\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)] \le \sqrt{\frac{1}{2} D(P \| Q)},$$

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

Beginning now, all lemmas in this section are implicitly understood to apply under the setting of Theorem 4 and, in particular, to $\tilde{A} = \min\{a \in \mathcal{A} : \theta_a \ge R^* - D\}$ and actions selected according to the probability matching property $\mathbb{P}_t(A_t = a) = \mathbb{P}_t(\tilde{A} = a)$ for each $a \in \mathcal{A}_t$ and $\mathbb{P}_t(A_t = |\mathcal{A}_t| + 1) = \mathbb{P}_t(\tilde{A} \notin \mathcal{A}_t)$. We begin by showing the mutual information bound stated as part of Theorem 4.

Lemma B.1 (Mutual Information Bound). *Let* $\delta = \mathbb{P}(\theta_a \ge 1 - D)$. *Then*

$$I(\tilde{A}; \theta) \le 1 + \log(1/\delta).$$

Proof. Because $\tilde{A} = \min\{a \in \mathcal{A} : \theta_a \ge 1 - D\}$ is a deterministic function of θ, we have $I(\tilde{A}; \theta) = H(\tilde{A}) = H(N)$, where $N \sim \text{Geom}(\delta)$ is a geometric random variable. This implies

$$\begin{aligned}
I(\tilde{A}; \theta) &= H(N) \\
&= -\sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} \log(\delta(1-\delta)^{k-1}) \\
&= -\sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} \log(\delta) - \sum_{k=1}^{\infty} \delta(1-\delta)^{k-1} \log((1-\delta)^{k-1}) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(N = k) \log(1/\delta) - \log(1-\delta) \sum_{k=1}^{\infty} \delta(1-\delta)^{k-1}(k-1) \\
&= \log(1/\delta) + \log\left(\frac{1}{1-\delta}\right)(\mathbb{E}[N] - 1) \\
&= \log(1/\delta) + \log\left(1 + \frac{\delta}{1-\delta}\right)\left(\frac{1-\delta}{\delta}\right) \\
&\le \log(1/\delta) + \left(\frac{\delta}{1-\delta}\right)\left(\frac{1-\delta}{\delta}\right) \\
&= 1 + \log(1/\delta). \quad \square
\end{aligned}$$

Throughout the remainder of the proof, we use the notation

$$\Gamma_t = \frac{(\mathbb{E}_t[\theta_{\tilde{A}} - \theta_{A_t}])^2}{I_t(\tilde{A}; (A_t, Y_t))}.$$

This represents the expected one-step information ratio in period t under the posterior measure $\mathbb{P}(\cdot | \mathcal{H}_{t-1})$. The next lemma shows that the cumulative information ratio can be bounded by the expected discounted average of these one-step information ratios.

Lemma B.2 (Relating the Information Ratio to the One-Step-Information-Ratio). *The information ratio is bounded as*

$$\Gamma(\tilde{A}, \psi_D^{\text{STS}}) \le (1 - \alpha^2) \sum_{t=0}^{\infty} \alpha^{2t} \mathbb{E}[\Gamma_t].$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[\tilde{R} - R_t] &= \mathbb{E}[\theta_{\tilde{A}} - \theta_{A_t}] \\
&\le \mathbb{E}\left[\sqrt{\Gamma_t} \sqrt{I_t(\tilde{A}; (A_t, Y_t))}\right] \\
&\le \sqrt{\mathbb{E}[\Gamma_t]\, \mathbb{E}[I_t(\tilde{A}; (A_t, Y_t))]} \\
&= \sqrt{\mathbb{E}[\Gamma_t]\, I(\tilde{A}; (A_t, Y_t) | \mathcal{H}_{t-1})}.
\end{aligned}$$

Then, by the definition of the information ratio,

$$\Gamma(\tilde{A},\psi) = (1-\alpha^2)\sum_{t=0}^{\infty}\alpha^{2t}\left(\frac{(\mathbb{E}[\tilde{R}-R_t])^2}{I(\tilde{A};(A_t,Y_t)|\mathcal{H}_{t-1})}\right) \le (1-\alpha^2)\sum_{t=0}^{\infty}\alpha^{2t}\mathbb{E}[\Gamma_t]. \quad \square$$

The next lemma provides a bound on the one-step information ratio. Recall that $\mathcal{A}_t = \cup_0^{t-1}\{A_s\}$ is the set of actions that were sampled before period t and $\delta \equiv \mathbb{P}(\theta_i \ge R^* - D)$ is the prior probability an arm is D optimal.

Lemma B.3. *For any $t \in \{0,1,2\cdots\}$, with probability 1,*

$$\Gamma_t \le 2|\mathcal{A}_t| + 2/\delta.$$

Proof. Define

$$L \equiv \mathbb{E}[\theta_i|\theta_i \ge R^* - D] - \mathbb{E}[\theta_i].$$

and

$$\delta \equiv \mathbb{P}(\theta_i \ge R^* - D).$$

Here δ is the probability an unsampled arm is D optimal, and L is the difference between the expected reward of a D-optimal arm and that of an arm sampled uniformly at random. In the case where $\theta_i \sim \text{Unif}(0,1)$, $\delta = D$ and $L = (1-D)/2$.

We can write expected regret as

$$\begin{aligned}
\mathbb{E}_t[\theta_{\tilde{A}} - \theta_{A_t}] &= \sum_{a\in\mathcal{A}}\mathbb{P}_t(\tilde{A}=a)\mathbb{E}_t[\theta_a|\tilde{A}=a] - \sum_{a\in\mathcal{A}}\mathbb{P}_t(A_t=a)\mathbb{E}_t[\theta_a] \\
&= \sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)(\mathbb{E}_t[\theta_a|\tilde{A}=a] - \mathbb{E}_t[\theta_a]) \\
&+ \sum_{a\notin\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)\mathbb{E}_t[\theta_a|\tilde{A}=a] - \sum_{a\notin\mathcal{A}_t}\mathbb{P}_t(A_t=a)\mathbb{E}_t[\theta_a] \\
&= \sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)(\mathbb{E}_t[\theta_a|\tilde{A}=a] - \mathbb{E}_t[\theta_a]) + \mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)(\mathbb{E}[\theta_a|\theta_a \ge 1-D] - \mathbb{E}[\theta_a]) \\
&= \underbrace{\sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)(\mathbb{E}_t[\theta_a|\tilde{A}=a] - \mathbb{E}_t[\theta_a])}_{\Delta_{t,1}} + \underbrace{\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)L}_{\Delta_{t,2}}.
\end{aligned}$$

This decomposes regret into the sum of two terms: one that captures the regret due to suboptimal selection within the set of previously sampled actions $\mathcal{A}_t$ and one due to the remaining possibility that none of the sampled actions are D optimal. The proof develops a similar decomposition for mutual information and then bounds both terms.

Let $Y_{t,a} = g(a,\theta,W_t)$ denote the outcome that would have been realized from a choice of action a at time t. We can express mutual information as follows:

$$\begin{aligned}
I_t(\tilde{A};(A_t,Y_t)) &= \sum_{a\in\mathcal{A}}\mathbb{P}_t(A_t=a)I_t(\tilde{A};Y_t) \\
&= \sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)I_t(\tilde{A};Y_{t,a}) + \sum_{a\notin\mathcal{A}_t}\mathbb{P}_t(A_t=a)I_t(\tilde{A};Y_{t,a})
\end{aligned}$$

Let us focus on the second sum, which captures the information acquired because of sampling previously untested actions $a \notin \mathcal{A}_t$. Such an action provides information about $\tilde{A}$, because if $\tilde{A} \notin \mathcal{A}_t$ and $\theta_a \ge 1-D$, then a is the first sampled action to be sufficiently close to optimal and $\tilde{A} = a$. Using the shorthand $P_t(X) = \mathbb{P}_t(X \in \cdot)$ to denote the posterior distribution of a random variable X, we have that for an untested action $a \notin \mathcal{A}_t$,

$$\begin{aligned}
I_t(\tilde{A};Y_{t,a}) &= \sum_{\tilde{a}\in\mathcal{A}}\mathbb{P}_t(\tilde{A}=\tilde{a})D(P_t(Y_{t,a}\mid\tilde{A}=\tilde{a})\,\|\,P_t(Y_{t,a})) \\
&\ge \mathbb{P}_t(\tilde{A}=a)D(P_t(Y_{t,a}\mid\tilde{A}=a)\,\|\,P_t(Y_{t,a})) \\
&= \mathbb{P}_t(\tilde{A}=a\mid A_t=a)D(P_t(Y_{t,a}\mid\theta_a\ge 1-D)\,\|\,P_t(Y_{t,a})) \\
&\ge 2\mathbb{P}_t(\tilde{A}=a)(\mathbb{E}_t[\theta_a\mid\theta_a\ge 1-D] - \mathbb{E}_t[\theta_a])^2 \\
&= 2\mathbb{P}_t(\tilde{A}=a)L^2 \\
&= 2\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\mathbb{P}_t(\tilde{A}=a\mid\tilde{A}\notin\mathcal{A}_t)L^2 \\
&= 2\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\delta L^2.
\end{aligned}$$

The second inequality uses Fact B.1. This implies

$$\sum_{a\notin\mathcal{A}_t}\mathbb{P}_t(A_t=a)I_t(\tilde{A};Y_{t,a}) \ge 2\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\mathbb{P}_t(A_t\notin\mathcal{A}_t)\delta L^2 = 2\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)^2\delta L^2.$$

Next, following the proof of proposition 3 of Russo and Van Roy [21] shows

$$\begin{aligned}\sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)I_t(\tilde{A};Y_{t,a}) &= \sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)\sum_{\tilde{a}\in\mathcal{A}}D(P_t(Y_{t,a}|\tilde{A}=\tilde{a})\|P_t(Y_{t,a}))\mathbb{P}_t(\tilde{A}=\tilde{a})\\ &\geq \sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)^2 D(P_t(Y_{t,a}|\tilde{A}=a)\|P_t(Y_{t,a}))\\ &\geq 2\sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)^2(\mathbb{E}_t[\theta_a|\tilde{A}=a]-\mathbb{E}_t[\theta_a])^2\\ &\geq \frac{2}{|\mathcal{A}_t|}\left(\sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)(\mathbb{E}_t[\theta_a|\tilde{A}=a]-\mathbb{E}_t[\theta_a])\right)^2,\end{aligned}$$

where the second inequality uses Fact B.1. Therefore,

$$I_t(\tilde{A};(A_t,Y_t)) \geq \underbrace{\frac{2}{|\mathcal{A}_t|}\Big(\sum_{a\in\mathcal{A}_t}\mathbb{P}_t(\tilde{A}=a)(\mathbb{E}_t[\theta_a|\tilde{A}=a]-\mathbb{E}_t[\theta_a])\Big)^2}_{G_{t,1}} + \underbrace{2\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)^2\delta L^2}_{G_{t,2}}$$

is lower bounded by the sum of two terms: one that captures the information gain due to refining knowledge about previously sampled actions and one that captures the expected information gathered about previously unexplored actions.

To bound the information ratio, we'll separately consider two cases. If $\Delta_{t,1}\geq\Delta_{t,2}$, then

$$\frac{(\mathbb{E}_t[\theta_{\tilde{A}}-\theta_{A_t}])^2}{I_t(\tilde{A};(A_t,Y_t))} \leq \frac{(2\Delta_{t,1})^2}{G_{t,1}+G_{t,2}} \leq \frac{4(\Delta_{t,1})^2}{G_{t,1}} = 2|\mathcal{A}_t|.$$

If instead $\Delta_{t,1}<\Delta_{t,2}$, then

$$\frac{(\mathbb{E}_t[\theta_{\tilde{A}}-\theta_{A_t}])^2}{I_t(\tilde{A};(A_t,Y_t))} \leq \frac{(2\Delta_{t,2})^2}{G_{t,1}+G_{t,2}} \leq \frac{4(\Delta_{t,2})^2}{G_{t,2}} = \frac{2}{\delta}.$$

This shows

$$\frac{(\mathbb{E}_t[\theta_{\tilde{A}}-\theta_{A_t}])^2}{I_t(\tilde{A};(A_t,Y_t))} \leq 2|\mathcal{A}_t|+2/\delta. \quad \square$$

Combining this result with Lemma B.2 gives the bound

$$\begin{aligned}\Gamma(\tilde{A},\psi_D^{\mathrm{STS}}) &\leq (1-\alpha^2)\sum_{t=0}^{\infty}\alpha^{2t}\mathbb{E}[\Gamma_t]\\ &\leq 2/\delta + 2(1-\alpha^2)\sum_{t=0}^{\infty}\alpha^{2t}\mathbb{E}[|\mathcal{A}_t|].\end{aligned}$$

To use this result, we begin by bounding $\mathbb{E}[|\mathcal{A}_t|]$.

Lemma B.4. *We have* $|\mathcal{A}_0|=0$ *and for each* $T\in\{1,2,\dots\}$, $\mathbb{E}[|\mathcal{A}_T|]\leq 1+\log(T)/\delta$.

Proof. First, using the tower property of conditional expectation and the probability matching property of A_t gives

$$\mathbb{E}[|\mathcal{A}_T|] = \mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{1}(A_t\notin\mathcal{A}_t)\right] = \mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{P}_t(A_t\notin\mathcal{A}_t)\right] = \mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\right] = \sum_{t=0}^{T-1}\alpha_t,$$

where $\alpha_t\equiv\mathbb{P}(\tilde{A}\notin\mathcal{A}_t)$. The next step of the proof shows that α_t satisfies the recursive inequality $\alpha_{t+1}\leq\alpha_t-\alpha_t^2\delta$. Write

$$\begin{aligned}\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_{t+1}) = \mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\mathbb{P}_t(\tilde{A}\neq A_t|\tilde{A}\notin\mathcal{A}_t) &= \mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)[1-\mathbb{P}_t(\tilde{A}=A_t|\tilde{A}\notin\mathcal{A}_t)]\\ &= \mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)[1-\mathbb{P}_t(A_t\notin\mathcal{A}_t)\delta]\\ &= \mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)[1-\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\delta],\end{aligned}$$

where the penultimate inequality uses that the conditional probability of sampling a satisficing action, given that none has been tested previously, is δ times the probability the algorithm tries a previously untested action. Taking expectations and applying the tower property of conditional expectation together with Jensen's inequality gives

$$\alpha_{t+1} = \mathbb{E}\big[\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_{t+1})\big] = \mathbb{E}\big[\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)\big] - \delta\mathbb{E}\Big[\mathbb{P}_t(\tilde{A}\notin\mathcal{A}_t)^2\Big] \leq \alpha_t-\delta\alpha_t^2. \tag{B.1}$$

Analyzing this recursive relationship is similar to analyzing an associated ordinary differential equation.[4] We apply a change of variables, studying $f(\alpha_t)$ for the function $f(x)=1/x$. By the mean value theorem, there exists some $\alpha\in[\alpha_{t+1},\alpha_t]$

such that

$$f(\alpha_{t+1}) - f(\alpha_t) = f'(\alpha)(\alpha_{t+1} - \alpha_t) = \frac{\alpha_t - \alpha_{t+1}}{\alpha^2} \geq \frac{\alpha_t - \alpha_{t+1}}{\alpha_t^2} \geq \delta, \tag{B.2}$$

where the final step uses (B.1). Summing (B.2) over $t \in \{0, \cdots, s-1\}$ gives that

$$f(\alpha_s) - f(\alpha_0) \geq \delta s \Rightarrow \alpha_s \leq \frac{1}{1+\delta s}.$$

We complete the proof by bounding the generalized harmonic sum as

$$\sum_{s=0}^{T-1} \alpha_s \leq \sum_{s=0}^{T-1} \frac{1}{1+\delta s} \leq 1 + \int_0^{T-1} \frac{1}{1+\delta s} ds = 1 + \frac{\log(1+\delta(T-1))}{\delta} \leq 1 + \log(T)/\delta. \quad \square$$

The next technical lemma shows $\sum_{t=1}^{\infty} \gamma^{-t} \log(t) = \tilde{O}\left(\frac{1}{1-\gamma}\right)$. The proof is given in Appendix C.

Lemma B.5. *For any $\gamma \in (0,1)$,*

$$\sum_{t=1}^{\infty} \gamma^{-t} \log(t) \leq \frac{1}{1-\gamma}\left[1 + \log\left(\frac{1}{1-\gamma}\right)\right].$$

Finally, we can conclude with the proof of Theorem 4. As shown before,

$$\begin{aligned}\Gamma\left(\tilde{A}, \psi_D^{\mathrm{STS}}\right) &\leq (1-\alpha^2) \sum_{t=0}^{\infty} \alpha^{2t} \mathbb{E}[\Gamma_t] \\ &\leq 2/\delta + 2(1-\alpha^2) \sum_{t=0}^{\infty} \alpha^{2t} \mathbb{E}[|\mathcal{A}_t|].\end{aligned}$$

By Lemma B.4 and Lemma B.5, we find

$$\begin{aligned}(1-\alpha^2) \sum_{t=0}^{\infty} \alpha^{2t} \mathbb{E}[|\mathcal{A}_t|] &\leq (1-\alpha^2) \sum_{t=1}^{\infty} \alpha^{2t} (1 + \log(t)/\delta) \\ &\leq 1 + (1/\delta)(1-\alpha^2) \sum_{t=1}^{\infty} \alpha^{2t} \log(t) \\ &\leq 1 + (1/\delta)\left[1 + \log\left(\frac{1}{1-\alpha^2}\right)\right].\end{aligned}$$

This implies

$$\Gamma\left(\tilde{A}, \psi_D^{\mathrm{STS}}\right) \leq 2 + 4/\delta + (2/\delta)\log\left(\frac{1}{1-\alpha^2}\right) = O\left((1/\delta)\log\left(\frac{1}{1-\alpha^2}\right)\right).$$

and concludes the proof of Theorem 4.

B.2. Proof of the Lower Bound: Theorem 5

The lower bound analysis leverages many of the standard information theoretic techniques for establishing minimax lower bounds (see, e.g., Tsybakov [28]). We first give a reduction to a hypothesis testing problem in which the goal is simply to identify a satisficing action. We show identifying a satisficing action is hard by upper bounding a certain Kullback–Leibler (KL) divergence through the data-processing inequality and the chain rule. These techniques are based on a classical change of measure argument by Lai and Robbins [15] as well as other bandit lower bounds (Bubeck and Cesa-Bianchi [4], Kaufmann et al. [13]). Our proof resembles most closely a proof of Bubeck and Cesa-Bianchi [4] that the minimax regret for k-armed undiscounted stochastic bandits is lower bounded by $\frac{1}{20}\sqrt{kT}$, where T is the number of time periods. There is some novelty to our lower bound analysis, however. The most significant change is that our problem involves an infinite number of arms and an independent prior, so there are many satisficing arms and we need to argue it is difficult to consistently play any arm from that set. The construction in Bubeck and Cesa-Bianchi [4] involves a problem instance with a dependent prior, under which it is difficult to identify the single arm that differs from the other $k-1$ arms. We also show how to carry out lower bound analyses of discounted problems by analyzing the distribution of A_τ, the action chosen at some randomly selected time $\tau \sim \mathrm{Geom}(1-\alpha)$.

Proof of Theorem 5.

Step 1. Express discounted sums in terms of random times.

Let $\tau \sim \text{Geom}(1-\alpha)$ be distributed independently of all other random variables. In several places, we use that, by the explicit form of the probability mass function $\mathbb{P}(\tau = t) = \alpha^t(1-\alpha)$,

$$\mathbb{E}\sum_{t=0}^{\infty} \alpha^t f(A_t) = (1-\alpha)^{-1}\mathbb{E}[f(A_\tau)]$$

for an arbitrary function $f : \mathbb{N} \to \mathbb{R}$. This allows us to compactly compress functions of the entire sequence of interactions into expectations with respect to A_τ alone.

Step 2. Regret lower bound in terms of the probability of satisficing.

In this problem, the optimal expected reward is $R^* = \frac{1}{2} + D$. We put $S_\theta = \{i \in \mathbb{N} : \theta_i \geq \frac{1}{2}\}$ to be the set of satisficing actions and $\text{OPT}_\theta = \{i \in \mathbb{N} : \theta_i = R^*\}$ to be the set of optimal actions. These sets are random because of their dependence on θ. We have

$$\begin{aligned}
\text{SRegret}(\alpha, \psi, D) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - R_t - D)\right] &= \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \mathbb{E}[R^* - R_t - D | \mathcal{H}_{t-1}, \theta]\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \mathbb{E}\left[R^* - \theta_{A_t} - D | \mathcal{H}_{t-1}, \theta\right]\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t (R^* - \theta_{A_t} - D)\right] \\
&= (1-\alpha)^{-1}\mathbb{E}\left[R^* - \theta_{A_\tau} - D\right] \\
&= (1-\alpha)^{-1}\left[\Delta\mathbb{P}\left(\theta_{A_\tau} = \frac{1}{2} - D\right) - D\mathbb{P}\left(\theta_{A_\tau} = \frac{1}{2} + D\right)\right] \\
&= (1-\alpha)^{-1}\left[\Delta\mathbb{P}(A_\tau \notin S_\theta) - D\mathbb{P}(\theta_{A_\tau} \in \text{OPT}_\theta)\right]
\end{aligned}$$

We now upper bound $\mathbb{P}(A_\tau \in \text{OPT}_\theta)$ in terms of ϵ. Recall the definition $\mathcal{A}_t := \{A_0, \ldots, A_{t-1}\}$. Proceeding recursively, we have $\mathbb{P}(\theta_{A_0} < R^*) = 1 - \epsilon$, because A_0 is chosen independently of θ. Next, we have

$$\mathbb{P}(\theta_{A_0} < R^* \wedge \theta_{A_1} < R^*) = (1-\epsilon)\mathbb{P}(\theta_{A_1} < R^* | \theta_{A_0} < R^*) \geq (1-\epsilon)^2.$$

To understand the final inequality, note that $\mathbb{P}(\theta_{A_1} < R^* | \theta_{A_0} < R^*, A_0, A_1)$ is equal to one if $A_1 = A_0$ and is equal to $1-\epsilon$ otherwise. Hence, $\mathbb{P}(\theta_{A_1} < R^* | \theta_{A_0} < R^*, A_0, A_1) \geq 1 - \epsilon$ almost surely. Repeating this process inductively gives

$$\mathbb{P}(\theta_{A_0} < R^* \wedge \cdots \wedge \theta_{A_t} < R^*) \geq (1-\epsilon)^{t+1}.$$

This gives the bound

$$\begin{aligned}
\mathbb{P}(A_\tau \in \text{OPT}_\theta) \leq 1 - \mathbb{P}(\theta_{A_0} < R^* \wedge \cdots \wedge \theta_{A_\tau} < R^*) &= 1 - \mathbb{E}\left[\mathbb{P}(\theta_{A_0} < R^* \wedge \cdots \wedge \theta_{A_\tau} < R^* | \tau)\right] \\
&\leq 1 - \mathbb{E}\left[(1-\epsilon)^{\tau+1}\right] \\
&= 1 - \sum_{t=0}^{\infty} \alpha^t (1-\alpha)(1-\epsilon)^{t+1} \\
&= 1 - \frac{(1-\alpha)(1-\epsilon)}{1-\alpha(1-\epsilon)} \\
&= \frac{\epsilon}{1-\alpha(1-\epsilon)} \\
&\leq \frac{\epsilon}{1-\alpha}.
\end{aligned}$$

We've reached our desired result, which lower bounds satisficing regret in terms of the probability of playing a satisficing arm at the random time τ:

$$\text{SRegret}(\alpha, \psi, D) \geq \Delta \cdot \left(\frac{\mathbb{P}(A_\tau \notin S_\theta)}{1-\alpha}\right) - D \cdot \left(\frac{\epsilon}{(1-\alpha)^2}\right). \tag{B.3}$$

Step 3. Identifying a satisficing action is hard.

Our goal is to lower bound $\mathbb{P}(A_\tau \notin S_\theta)$. To do this, we consider two alternative infinite armed bandit models. Each induces an a probability measure as follows:

1. The probability measure $\mathbb{P}(\cdot)$ corresponds to the infinite-armed bandit model as described in the theorem statement. The collection $\theta \equiv (\theta_a)_{a \in \mathbb{N}}$ is drawn randomly according to the prior probabilities in (9). The random seed ξ is drawn uniformly from $[0,1]$. For each period t, the action $A_t = \psi(\mathcal{H}_{t-1}, \xi)$ is prescribed by the policy ψ. Then $\mathbb{P}(R_t = 1 | A_t, \theta, \mathcal{H}_{t-1}, \xi) = \theta_{A_t}$.

2. We consider an alternative model, which is identical except that rewards are always drawn from a Bernoulli distribution with mean $1/2 - \Delta$. Precisely, we let $\mathbb{Q}$ be an alternative probability measure with the following properties: As before, the

random seed ξ is drawn uniformly from $[0,1]$, $A_t = \psi(\mathcal{H}_{t-1}, \xi)$ for each period t, and θ is drawn from according to the prior probabilities in (9). However, rewards are now independent from θ, with $\mathbb{Q}(R_t = 1|A_t, \theta, \mathcal{H}_{t-1}, \xi) = 1/2 - \Delta$.

The idea of this construction is that $D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_{t-1} = \cdot)\|\mathbb{P}(\mathcal{H}_{t-1} = \cdot))$ will reduce to considering the divergence in reward distributions because this is the only source of discrepancy between the probability distributions. We continue to let $\tau \sim \mathrm{Geom}(1-\alpha)$ denote a geometric random variable that is mutually independent from θ and $(\mathcal{H}_t)_{t\in\mathbb{N}}$ under both $\mathbb{P}$ and $\mathbb{Q}$. We take $\mathbb{E}_{\mathbb{Q}}[\cdot]$ to denote the expectation under the probability measure $\mathbb{Q}$. Define the binary KL divergence function $d : [0,1] \times [0,1] \to \mathbb{R}$ by $d(p\|q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$.

Under $\mathbb{Q}$, the algorithm's observations are independent of θ, so for each t,

$$\mathbb{Q}(A_t \in S_\theta) = \mathbb{E}_{\mathbb{Q}}[\mathbb{Q}(A_t \in S_\theta|A_t)] = \mathbb{E}_{\mathbb{Q}}[\delta] = \delta$$

$$\mathbb{Q}(A_t \in \mathrm{OPT}_\theta) = \mathbb{E}_{\mathbb{Q}}[\mathbb{Q}(A_t \in \mathrm{OPT}_\theta|A_t)] = \mathbb{E}_{\mathbb{Q}}[\epsilon] = \epsilon,$$

(This property relies critically on the fact that the algorithm cannot adapt to θ under $\mathbb{Q}$ and is the key to the proof.) Applying this gives

$$\mathbb{Q}(A_\tau \in S_\theta) = \mathbb{E}_{\mathbb{Q}}[\mathbb{Q}(A_\tau \in S_\theta|\tau)] = \delta.$$

Then, by Pinsker's inequality,

$$\begin{aligned}\mathbb{P}(A_\tau \in S_\theta) &\le \mathbb{Q}(A_\tau \in S_\theta) + \sqrt{\frac{1}{2}d(\mathbb{Q}(A_\tau \in S_\theta)\|\mathbb{P}(A_\tau \in S_\theta))} \\ &= \delta + \sqrt{\frac{1}{2}d(\mathbb{Q}(A_\tau \in S_\theta)\|\mathbb{P}(A_\tau \in S_\theta))}. \end{aligned} \tag{B.4}$$

We upper bound and expand the KL divergence through repeated use of the data-processing inequality and the chain rule. We have

$$\begin{aligned} d(\mathbb{Q}(A_\tau \in S_\theta)\|\mathbb{P}(A_\tau \in S_\theta)) &\le D_{\mathrm{KL}}(\mathbb{Q}(A_\tau = \cdot)\|\mathbb{P}(A_\tau = \cdot)) \\ &\le D_{\mathrm{KL}}(\mathbb{Q}((\tau,\theta,\mathcal{H}_\tau,\xi) = \cdot)\|\mathbb{P}((\tau,\theta,\mathcal{H}_\tau,\xi) = \cdot)) \\ &= D_{\mathrm{KL}}(\mathbb{Q}((\tau,\theta,\xi) = \cdot)\|\mathbb{P}((\tau,\theta,\xi) = \cdot)) + D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_\tau = \cdot|\tau,\theta,\xi)\|\mathbb{P}(\mathcal{H}_\tau = \cdot|\tau,\theta,\xi)) \\ &= D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_\tau = \cdot|\tau,\theta,\xi)\|\mathbb{P}(\mathcal{H}_\tau = \cdot|\tau,\theta,\xi)) \\ &= \sum_{t=0}^{\infty}\mathbb{Q}(\tau = t)D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_\tau = \cdot|\tau = t,\theta,\xi)\|\mathbb{P}(\mathcal{H}_\tau = \cdot|\tau = t,\theta,\xi)) \\ &= \sum_{t=0}^{\infty}\mathbb{Q}(\tau = t)D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_t = \cdot|\theta,\xi)\|\mathbb{P}(\mathcal{H}_t = \cdot|\theta,\xi)) \end{aligned}$$

where the final equality uses the independence of τ and $(\theta, (\mathcal{H}_t)_{t\in\mathbb{N}}, \xi)$. Now, by the chain rule and the relation $\mathcal{H}_t = (\mathcal{H}_{t-1}, A_t, R_t)$, we have

$$\begin{aligned} &D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_t = \cdot|\theta,\xi)\|\mathbb{P}(\mathcal{H}_t = \cdot|\theta,\xi)) \\ &\quad= D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_{t-1} = \cdot|\theta,\xi)\|\mathbb{P}(\mathcal{H}_{t-1} = \cdot|\theta,\xi)) \\ &\qquad+ D_{\mathrm{KL}}(\mathbb{Q}(A_t = \cdot|\theta,\mathcal{H}_{t-1},\xi)\|\mathbb{P}(A_t = \cdot|\theta,\mathcal{H}_{t-1},\xi)) \\ &\qquad+ D_{\mathrm{KL}}(\mathbb{Q}(R_t = \cdot|\theta,\mathcal{H}_{t-1},A_t,\xi)\|\mathbb{P}(R_t = \cdot|\theta,\mathcal{H}_{t-1},A_t,\xi)) \\ &\quad= D_{\mathrm{KL}}(\mathbb{Q}(\mathcal{H}_{t-1} = \cdot|\theta,\xi)\|\mathbb{P}(\mathcal{H}_{t-1} = \cdot|\theta,\xi)) + \mathbb{E}_{\mathbb{Q}}\left[d(1/2-\Delta\|\theta_{A_t},)\right] \\ &\quad=\cdots \\ &\quad= \mathbb{E}_{\mathbb{Q}}\left[\sum_{\ell=0}^{t} d(1/2-\Delta\|\theta_{A_\ell})\right] \\ &\quad= \sum_{\ell=0}^{t}\left[\mathbb{Q}(\theta_{A_t} = 1/2)d(1/2-\Delta\|1/2) + \mathbb{Q}(\theta_{A_t} = 1/2+D)d(1/2-\Delta\|1/2+D)\right] \\ &\quad\le \sum_{\ell=0}^{t}\left[\mathbb{Q}(\theta_{A_t} = 1/2)d(1/2-\Delta\|1/2) + \mathbb{Q}(\theta_{A_t} = 1/2+D)d(1/4\|3/4)\right] \\ &\quad\le \sum_{\ell=0}^{t}\mathbb{Q}(A_t \in S_\theta)d(1/2-\Delta\|1/2) + \sum_{\ell=0}^{t}\mathbb{Q}(A_t \in \mathrm{OPt}_\theta)d(1/2-\Delta\|3/4). \end{aligned}$$

Here we used that $D_{\mathrm{KL}}(\mathbb{Q}(A_t = \cdot|\theta,\mathcal{H}_{t-1},\xi)\|\mathbb{P}(A_t = \cdot|\theta,\mathcal{H}_{t-1},\xi)) = 0$ because, conditioned on $(\mathcal{H}_{t-1},\xi)$, A_t is almost surely equal to $\psi_t(\mathcal{H}_{t-1},\xi)$ under $\mathbb{Q}(\cdot)$ or $\mathbb{P}(\cdot)$. The first inequality uses that $1/2-\Delta \ge 1/4$ and $\theta_a \le 3/4$ by hypothesis. The second

inequality uses that $\mathbb{Q}(A_t \in S_\theta) \geq \mathbb{Q}(\theta_{A_t} = 1/2)$ together with the nonnegativity of the KL divergence. Plugging this in above, we find

$$\begin{aligned}
&d(\mathbb{Q}(A_\tau \in S_\theta) \| \mathbb{P}(A_\tau \in S_\theta)) \\
&\leq \sum_{t=0}^{\infty} \mathbb{Q}(\tau = t) \sum_{\ell=0}^{t} [\mathbb{Q}(A_\ell \in S_\theta) d(1/2 - \Delta \| 1/2) + \mathbb{Q}(A_\ell \in \mathrm{OPt}_\theta) d(1/4 \| 3/4)] \\
&= \sum_{t=0}^{\infty} \mathbb{Q}(\tau \geq t) [\mathbb{Q}(A_t \in S_\theta) d(1/2 - \Delta \| 1/2) + \mathbb{Q}(A_t \in \mathrm{OPt}_\theta) d(1/4 \| 3/4)] \\
&= \sum_{t=0}^{\infty} \alpha^t \delta d(1/2 - \Delta \| 1/2) + \sum_{t=0}^{\infty} \alpha^t \epsilon d(1/4 \| 3/4) \\
&= \frac{\delta \cdot d(1/2 - \Delta \| 1/2)}{1-\alpha} + \frac{\epsilon d(1/4 \| 3/4)}{1-\alpha} \\
&= \frac{\delta \cdot \left[(1/2 - \Delta)\log(1 - 2\Delta) + (1/2 + \Delta)\log(1 + 2\Delta)\right]}{1-\alpha} + \frac{\epsilon(0.549\cdots)}{1-\alpha} \\
&\leq \frac{\delta \cdot \Delta \cdot \log\left(\frac{1+2\Delta}{1-2\Delta}\right)}{1-\alpha} + \frac{\epsilon(0.549\cdots)}{1-\alpha} \\
&\leq \frac{8\delta \cdot \Delta^2}{1-\alpha} + \frac{\epsilon}{1-\alpha}.
\end{aligned}$$

To conclude, plugging the above into (B.4) and using the concavity of the square root, we have shown

$$\mathbb{P}(A_\tau \in S_\theta) \leq \delta + 2\Delta\sqrt{\frac{\delta}{1-\alpha}} + \sqrt{\frac{\epsilon/2}{1-\alpha}}. \tag{B.5}$$

Step 4. Conclusion by plugging in for D and ϵ.
Combining (B.3) and (B.5), we have

$$\begin{aligned}
\mathrm{SRegret}(\alpha, \psi, D) &\geq \left(\frac{\Delta \cdot (1-\delta) - 2\Delta^2 \cdot \sqrt{\frac{\delta}{1-\alpha}}}{1-\alpha}\right) - \sqrt{\epsilon} \cdot \frac{\Delta/\sqrt{2}}{(1-\alpha)^{3/2}} - \epsilon \cdot \left(\frac{D}{(1-\alpha)^2}\right) \\
&\geq \left(\frac{\Delta/2 - 2\Delta^2 \cdot \sqrt{\frac{\delta}{1-\alpha}}}{1-\alpha}\right) - \sqrt{\epsilon} \cdot \frac{1/(4\sqrt{2})}{(1-\alpha)^{3/2}} - \epsilon \cdot \left(\frac{1/4}{(1-\alpha)^2}\right) \\
&\geq \left(\frac{\Delta/2 - 2\Delta^2 \cdot \sqrt{\frac{\delta}{1-\alpha}}}{1-\alpha}\right) - \cdot \frac{\sqrt{\epsilon} \cdot (1/2)}{(1-\alpha)^2} \\
&:= f(\Delta) - g(\epsilon),
\end{aligned}$$

where in the final step we will require $\epsilon \leq 1$.

We now focus on the first term and will eventually pick ϵ so the remaining term is sufficiently small. Set

$$f(\Delta) = \frac{1}{1-\alpha} \cdot \left[\frac{\Delta}{2} - 2\Delta^2 \cdot \sqrt{\frac{\delta}{1-\alpha}}\right].$$

This is a quadratic function with global minimum at

$$\Delta^* = \arg\min_{\Delta \in \mathbb{R}} \mathrm{f}(\Delta) = \frac{1}{8} \cdot \sqrt{\frac{1-\alpha}{\delta}}.$$

For

$$\Delta_0 \equiv \min\left\{\frac{1}{4}, \Delta^*\right\} = \begin{cases} 1/4 & \text{for } 4\delta \leq 1-\alpha, \\ \Delta^* & \text{for } 4\delta \geq 1-\alpha. \end{cases} \tag{B.6}$$

we have

$$f(\Delta_0) \geq \left\{ \begin{array}{ll} \frac{1}{16} \cdot \frac{1}{1-\alpha} & \text{for } 4\delta \geq (1-\alpha) \\ \frac{1}{32} \cdot \sqrt{\frac{1/\delta}{1-\alpha}} & \text{for } 4\delta \leq (1-\alpha) \end{array} \right\} \geq \frac{1}{16} \cdot \min\left\{ \frac{1}{1-\alpha}, \sqrt{\frac{1/4\delta}{1-\alpha}} \right\}$$

Now, we pick $\epsilon_0 \leq 1$ so that

$$g(\epsilon_0) \leq f(\Delta)/2.$$

We need

$$\frac{\sqrt{\epsilon_0} \cdot (1/2)}{(1-\alpha)^2} \leq \frac{1}{16} \cdot \min\left\{ \frac{1}{1-\alpha}, \sqrt{\frac{1/4\delta}{1-\alpha}} \right\}.$$

This is satisfied with equality for

$$\epsilon_0 = \frac{1}{64} \cdot \min\left\{ (1-\alpha)^2, \frac{(1-\alpha)^3}{4\delta} \right\}. \tag{B.7}$$

This shows that for a choice of $\Delta = \Delta_0$ and $\epsilon \leq \epsilon_0$, as in the theorem statement, we have

$$\text{SRegret}(\alpha, \psi, D) \geq \frac{1}{32} \cdot \min\left\{ \frac{1}{1-\alpha}, \sqrt{\frac{1/4\delta}{1-\alpha}} \right\}. \quad \square$$

Appendix C. Proof of Lemma B.5

$$\begin{aligned} \sum_{t=1}^{\infty} \gamma^{-t} \log(t) &\leq \sum_{t=1}^{\infty} e^{-(1-\gamma)t} \log(t) \\ &= \sum_{t=2}^{\infty} e^{-(1-\gamma)t} \log(t) \\ &\overset{*}{\leq} \int_1^{\infty} e^{-(1-\gamma)x} \log(x+1) dx \\ &= \frac{1}{1-\gamma} \int_1^{\infty} e^{-u} \log\left(\frac{u}{1-\gamma} + 1 \right) du \\ &\leq \frac{1}{1-\gamma} \left(\left[1 + \log\left(\frac{1}{1-\gamma} \right) \right] \int_1^{\infty} e^{-u} du + \int_1^{\infty} e^{-u} \log(u) du \right) \\ &= \frac{1}{1-\gamma} \left(\left[1 + \log\left(\frac{1}{1-\gamma} \right) \right] (1/e) + \int_1^{\infty} e^{-u} \log(u) du \right) \\ &\leq \frac{1}{1-\gamma} \left[1 + \log\left(\frac{1}{1-\gamma} \right) \right], \end{aligned}$$

where the last step uses a numerical approximation to the indefinite integral

$$\int_1^{\infty} e^{-u} \log(u) du \approx .22$$

along with the fact that $1/e + .22 \approx .57 < 1$.

The inequality (*) uses that for any $t \geq 2$,

$$e^{-(1-\gamma)t} \log(t) \leq \int_{t-1}^{t} e^{-(1-\gamma)x} \log(x+1)$$

because $e^{-(1-\gamma)x}$ is decreasing in x and $\log(x)$ is increasing in x. $\square$

Endnotes

[1] If the minimum is not attained, all arguments can be easily modified by taking $\tilde{A}$ to be an ϵ–minimizer for some arbitrarily small ϵ.

[2] Of course, there is nothing crucial about this ordering on actions. We can equivalently construct a randomized order in which actions are sampled; for each realization of the random variable ξ, let $\pi_\xi : \mathbb{N}_0 \to \mathbb{N}_0$ be a permutation and take $\tilde{A} = \min\{\pi_\xi(a) : \theta_a \geq 1 - D\}$.

[3] Standard errors are smaller than .01 for each algorithm and time period, so confidence intervals are omitted.

[4] In particular, consider the relationship $\alpha'(t) = -\delta\alpha(t)^2 t$. Rearranging, this is $\frac{\alpha'(t)}{\alpha(t)^2} = -\delta$. For $f(x) = 1/x$, this gives $-\frac{d}{dt}f(\alpha(t)) = -\delta$. Integrating both sides gives $f(T) = f(0) + \delta T$ and rearranging terms gives a bound on $\alpha(t)$.

References

[1] Agrawal S, Goyal N (2013) Further optimal regret bounds for Thompson sampling. *Proc. Sixteenth Internat. Conf. Artificial Intelligence Statist.*, 99–107.
[2] Berry DA, Chen RW, Zame A, Heath DC, Shepp LA (1997) Bandit problems with infinitely many arms. *Ann. Statist.* 25(5):2103–2116.
[3] Bonald T, Proutiere A (2013) Two-target algorithms for infinite-armed bandits with Bernoulli rewards. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates, Inc., Red Hook, NY), 2184–2192.
[4] Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.
[5] Bubeck S, Eldan R (2016) Multi-scale exploration of convex functions and bandit convex optimization. *Conf. Learn. Theory* (PMLR), 583–589.
[6] Bubeck S, Munos R, Stoltz G, Szepesvári C (2011) $\mathcal{X}$-armed bandits. *J. Machine Learn. Res.* 12:1655–1695.
[7] Cappé O, Garivier A, Maillard OA, Munos R, Stoltz G (2013) Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* 41(3):1516–1541.
[8] Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. Shawe-Taylor J, Zemel RS, Bartlett PL, eds. *Proc. 24th Internat. Conf. Neural Inform. Processing System* (Curran Associates Inc., Red Hook, NY), 2249–2257.
[9] Cover T, Thomas J (2012) *Elements of Information Theory* (John Wiley & Sons, New York).
[10] Deshpande Y, Montanari A (2012) Linear bandits in high dimension and recommendation systems. *50th Annual Allerton Conf. Comm. Control Comput.* (IEEE, Piscataway, NJ), 1750–1754.
[11] Francetich A, Kreps D (2020) Choosing a good toolkit, i: Prior-free heuristics. *J. Econom. Dynam. Control* 111:103813.
[12] Francetich A, Kreps D (2020) Choosing a good toolkit, ii: Bayes-rule based heuristics. *J. Econom. Dynam. Control* 111:103814.
[13] Kaufmann E, Cappé O, Garivier A (2016) On the complexity of best-arm identification in multi-armed bandit models. *J. Machine Learn. Res.* 17(1):1–42.
[14] Kleinberg R, Slivkins A, Upfal E (2008) Multi-armed bandits in metric spaces. *Proc. 40th ACM Sympos. Theory Comput.* (ACM, New York).
[15] Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances Appl. Math.* 6(1):4–22.
[16] Lattimore T, Szepesvári C (2019) An information-theoretic approach to minimax regret in partial monitoring. Beygelzimer A, Hsu D, eds. *Proc. Thirty-Second Conf. Learn. Theory*, vol. 99 (PMLR), 2111–2139.
[17] Liu F, Buccapatnam S, Shroff N (2018) Information directed sampling for stochastic bandits with graph feedback. *Proc. AAAI Conf. Artificial Intelligence*, vol. 32 (AAAI, Palo Alto CA).
[18] Lu X, Van Roy B (2019) Information-theoretic confidence bounds for reinforcement learning. *Advances in Neural Information Processing Systems (Neurips) 32* (Curran Associates, Inc., Red Hook, NY), 2458–2466.
[19] Rusmevichientong P, Tsitsiklis J (2010) Linearly parameterized bandits. *Math. Oper. Res.* 35(2):395–411.
[20] Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.
[21] Russo D, Van Roy B (2016) An information-theoretic analysis of Thompson sampling. *J. Machine Learn. Res.* 17(68):1–30.
[22] Russo DJ, Van Roy B, Kazerouni A, Osband I, Wen Z (2018) A tutorial on Thompson sampling. *Foundations Trends Machine Learn.* 11(1):1–96.
[23] Ryzhov I, Powell W, Frazier P (2012) The knowledge gradient algorithm for a general class of online learning problems. *Oper. Res.* 60(1):180–195.
[24] Scott S (2010) A modern Bayesian look at the multi-armed bandit. *Appl. Stochastic Models Bus. Indust.* 26(6):639–658.
[25] Simon HA (1979) Rational decision making in business organizations. *Amer. Econom. Rev.* 69(4):493–513.
[26] Thompson W (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.
[27] Tossou AC, Dimitrakakis C, Dubhashi DP (2017) Thompson sampling for stochastic bandits with graph feedback. *Proc. 31st AAAI Conf. Artificial Intelligence* (AAAI, Palo Alto, CA), 2660–2666.
[28] Tsybakov AB (2008) *Introduction to Nonparametric Estimation* (Springer Science & Business Media, Berlin).
[29] Wang Y, Audibert JY, Munos R (2009) Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems (NIPS)*, (Curran Associates, Inc., Red Hook, NY), 1729–1736.