Operations Research

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

Optimistic Gittins Indices

Vivek F. Farias, Eli Gutin

To cite this article:
Vivek F. Farias, Eli Gutin (2022) Optimistic Gittins Indices. Operations Research 70(6):3432-3456. https://doi.org/10.1287/opre.2021.2207

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 6, November–December 2022, pp. 3432–3456
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Optimistic Gittins Indices

Vivek F. Farias,[a] **Eli Gutin**[b,*]

[a] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [b] Uber Technologies, Inc., San Francisco, California 94518
*Corresponding author
Contact: vivekf@mit.edu (VFF); gutin@uber.com, https://orcid.org/0000-0002-4211-0756 (EG)

Received: January 10, 2019
Revised: July 23, 2020; June 29, 2021
Accepted: August 10, 2021
Published Online in Articles in Advance: March 1, 2022

Area of Review: Decision Analysis

https://doi.org/10.1287/opre.2021.2207

Abstract. Recent years have seen a resurgence of interest in Bayesian algorithms for the multiarmed bandit (MAB) problem, such as Thompson sampling. These algorithms seek to exploit prior information on arm biases. The empirically observed performance of these algorithms makes them a compelling alternative to their frequentist counterparts. Nonetheless, there appears to be a wide range in empirical performance among such Bayesian algorithms. These algorithms also vary substantially in their design (as opposed to being variations on a theme). In contrast, if one cared about Bayesian regret discounted over an infinite horizon at a *fixed, prespecified* rate, the celebrated Gittins index theorem offers an *optimal* algorithm. Unfortunately, the Gittins analysis does not appear to carry over to minimizing Bayesian regret over all sufficiently large horizons and computing a Gittins index is onerous relative to essentially any incumbent index scheme for the Bayesian MAB problem. The present paper proposes a tightening sequence of optimistic approximations to the Gittins index. We show that the use of these approximations in concert with the use of an increasing discount factor appears to offer a compelling alternative to state-of-the-art index schemes proposed for the Bayesian MAB problem in recent years. We prove that these optimistic indices constitute a regret optimal algorithm, in the sense of meeting the Lai-Robbins lower bound, including matching constants. Perhaps more interestingly, the use of even the loosest of these approximations appears to offer substantial performance improvements over state-of-the-art alternatives (including Thompson sampling, information direct sampling, and the Bayes UCB algorithm) while incurring little to no additional computational overhead relative to the simplest of these alternatives.

Funding: Both authors were partially supported by NSG Grant CMMI 1727239.

Keywords: multiarmed bandits • Gittins index • online learning

1. Introduction

The multiarmed bandit (MAB) problem is perhaps the simplest example of a learning problem that exposes the tension between exploration and exploitation. In its simplest form, we are given a collection of random variables or arms. By adaptively sampling these random variables, we seek to eventually sample consistently from the random variable with the highest mean. This is typically formalized by asking that we minimize cumulative regre'; a notion we make precise in a later section.

Recent years have seen a resurgence of interest in *Bayesian* algorithms for the MAB problem. In this variant of the MAB problem, we are endowed with a prior on arm means, and a number of algorithms that exploit this prior have been proposed and analyzed. These include Thompson sampling (Thompson 1933), Bayes-upper confidence bound (UCB) (Kaufmann et al. 2012b), Kullback-Leibler (KL)-UCB (Garivier and Cappé 2011), and information directed sampling (IDS) (Russo and Van Roy 2018). The ultimate motivation for these algorithms appears to be the empirical performance they offer. Specifically, these Bayesian algorithms appear to incur smaller regret than their frequentist counterparts such as the UCB algorithm proposed by Auer et al. (2002), even when regret is measured in a frequentist sense. This empirical evidence has, very recently, been reinforced by theoretical performance guarantees. For instance, it has been shown that both Thompson sampling and Bayes UCB enjoy upper bounds on frequentist regret that match the Lai-Robbins lower bound (Lai and Robbins 1985). Interestingly, even amongst the various Bayesian algorithm proposed there appears to be a wide range in empirical performance. For instance, empirical evidence presented in Russo and Van Roy (2018) suggests that the IDS algorithm offer a substantial improvement in frequentist regret over Thompson sampling and the Bayes UCB algorithm, among others. The former algorithm does not, however, enjoy the optimal data dependent frequentist regret

bounds that the latter two do. Perhaps more importantly, these algorithms also vary substantially in their design (as opposed to being variations on a theme).

Now a prior on arm means endows us with the structure of a Markov decision process (MDP), and none of the Bayesian algorithms alluded to previously exploit this structure. This is especially surprising in light of the celebrated Gittins index theorem. That breakthrough result proved the optimality of a certain index policy for a *horizon dependent* variant of the Bayesian MAB. Specifically, imagine that we cared about the expected (Bayes) regret incurred over an exponentially distributed horizon, where the mean horizon length is known to the algorithm designer. This problem is nominally a high dimensional MDP. Gittins, however, proved that a simple to compute index rule was optimal for this task resolving a problem that had remained open for several decades (Gittins 1979). Why does the Gittins index theorem not immediately help resolve the design of an optimal algorithm for the variant of the Bayesian MAB problem that is the subject of the approaches discussed in the preceding paragraph? As we will discuss more carefully in our literature review, this is certainly not from lack of research effort (Lattimore 2016). In fact, one must deal with several substantial challenges:

1. Dependence on Horizon: The notion of regret optimality as popularized by Lai and Robbins (1985) is 'anytime'. Colloquially, this can be thought of as follows: we desire an algorithm that performs well for *any* time horizon. This fact is fundamentally at odds with Gittins' variant of the MAB problem that (via a discount factor) effectively specifies a (exponentially distributed) horizon. Gittins' result is intimately connected to this choice of horizon; even seemingly minor changes appear to render the problem intractable. For instance, it is known that a Gittins-like index strategy is suboptimal for a fixed, finite-horizon (Berry and Fristedt 1985). Algorithms for other notions of optimality that one may reasonably conjecture are better aligned with 'anytime' regret optimality (such as, say, Cesaro-overtaking optimality) are similarly elusive (Katehakis and Rothblum 1996).

2. Computation: Separate from the issues made in the previous point, consider the task of computing a Gittins index at every point in time. The computation of a Gittins index can be reduced to the solution of a certain infinite horizon stopping problem. For the Bayesian MAB, the state space for this problem must describe all possible posteriors one may encounter on a given arm. Assuming conjugate priors, one may hope for a finite dimensional state space, but tractable computation will typically call for some form of state-space truncation. This computation is far more onerous than any of the aforementioned indices. Furthermore, it is reasonable to conjecture that as time progresses one may require increasingly more accurate estimates of the Gittins index, which further complicates computation, and calls into question the correctness of a naive state-space truncation scheme.

Against this backdrop, the present paper makes the following contribution. We show that picking arms according to a certain tractable approximation to their Gittins index, computed for a time dependent discount factor we characterize precisely, constitutes a regret optimal bandit policy. The resulting index rule is both simple to compute and in computational experiments appears to outperform state-of-the-art bandit algorithms by a material margin.

In greater detail, we outline our contributions as follows:

1. Optimistic Approximations: We propose a sequence of 'optimistic' approximations to the Gittins index. These optimistic approximations can be interpreted as providing a tightening sequence of upper bounds on the optimal stopping problem defining a Gittins index, yielding the index itself in the limit. The computation associated with the simplest of these approximations is no more burdensome than the computation of indices for the Bayes UCB algorithm, and several orders of magnitude faster than the best performing alternative from an empirical perspective (the IDS algorithm).

2. Regret Optimality: We establish that an arm selection rule that is greedy with respect to any optimistic approximation to the Gittins index achieves optimal regret in the sense of meeting the Lai-Robbins lower bound (including matching constants) for the canonical case of Beta-Bernoulli bandits. A crucial ingredient required for this scheme to work is that as time progresses, the discount factor employed in computing the index must be increased at a certain rate which we characterize precisely. This implicitly resolves the challenge of horizon dependence.

3. Empirical Performance: We show empirically that even the simplest optimistic approximation to the Gittins index outperforms the state-of-the-art incumbent schemes discussed in this introduction by a nontrivial margin. Our empirical study is careful to recreate several ensembles of problem instances considered by previous authors (including a particularly computationally intensive study by Chapelle and Li (2011) that prompted the reexamination of the Thompson sampling algorithm in recent years). The margin of improvement we demonstrate increases further as one employs successfully tighter optimistic approximations, at the cost of computational effort.

In summary, we propose a new index rule for the Baysian MAB problem that uses Gittins indices in a novel way. This new index rule enjoys the strongest possible data-dependent regret guarantees while also offering excellent empirical performance.

1.1. Relevant Literature

We organize our literature review around the primary topics that this paper touches on. The study of exploration-exploitation problems is vast, even if it is restricted to a problem with a finite number of arms. Consequently, our review will be focused on stochastic, noncontextual, versions of the MAB problem. Even with this restriction, the literature remains vast, and therefore we focus on papers that are either seminal in nature or particularly relevant to our own work; this review is by no means comprehensive with respect to the MAB problem.

1.1.1. Regret Optimality and the Bandit Problem. Robbins (1952) motivated the study of the MAB problem and left open questions on how to design effective policies. Since then, Lai and Robbins (1985) proved a cornerstone result, namely an asymptotic lower bound on regret that any consistent strategy incurs. The same paper proposes an upper-confidence bound (UCB) algorithm that asymptotically achieves the lower bound. Computationally efficient UCB algorithms were developed by Agrawal (1995) and Katehakis and Robbins (1995). Later, Auer et al. (2002) and Audibert and Bubeck (2010) proved finite time regret bounds for UCB algorithms and demonstrated ways to tune them to improve performance. Garivier and Cappé (2011) and Maillard et al. (2011) have proposed other UCB-type algorithms where the confidence bounds are calculated using the KL-divergence function. Those authors provide a finite-time analysis, and their algorithms are shown to achieve the Lai-Robbins bound.

1.1.2. Bayesian Bandit Algorithms. Another powerful approach to bandit problems is to work with a Bayesian prior to model one's uncertainty about an arm's expected reward. Lai (1987) proves an asymptotic lower bound on Bayes' risk and develops a horizon-dependent algorithm that achieves it. Thompson sampling (Thompson 1933), one of the earliest algorithms proposed for the MAB problem, is in fact a Bayesian one. Empirical studies by Chapelle and Li (2011) and Scott (2010) highlight Thompson sampling's hugely superior performance over some UCB algorithms even when the prior is mismatched. A series of tight regret bounds for Thompson sampling have been established by Agrawal and Goyal (2012, 2013) and Kaufmann et al. (2012b). These authors have shown Thompson sampling to be regret optimal for the canonical Beta-Bernoulli bandit. Recently, Korda et al. (2013) generalized the aforementioned results to bandit problems where the arm distributions belong to a one-dimensional exponential family. Interestingly enough, Robbins (1952) seems to have been unaware of Thompson sampling and its effectiveness in the non-Bayesian setting.

Several other Bayesian algorithms exist. Kaufmann et al. (2012b) propose Bayes UCB, which they show is competitive with Thompson sampling. The main idea behind Bayes UCB is to treat quantiles of the arm's prior as an upper confidence bound and let the quantile grows at some prespecified rate. Russo and Van Roy (2018) propose information directed sampling (IDS), an algorithm that exploits information theoretic quantities arising from the prior distributions over the arms. In simulations, IDS is shown to dominate many of the aforementioned algorithms, including Thompson sampling, Bayes UCB, and KL-UCB. In our empirical investigation, we will see that IDS is the closest competitor to the approach we propose here (we recreate the experiments from Russo and Van Roy (2018)).

1.1.3. Gittins Index and Its Approximations. There is another stream of literature that models the MAB problem as an MDP. For the case of two arms, where one arm's reward is deterministic, Bradt et al. (1956) showed that for this one-dimensional MDP, an index rule is an optimal strategy. When the objective is to maximize the infinite sum of expected *discounted* rewards Gittins (1979) famously showed the optimality of an index policy. The Gittins index is similar to that proposed by Bradt et al. (1956) but takes discounting into account. Several alternative proofs of Gittins' result are available (Whittle 1980, Weber et al. 1992, Tsitsiklis 1994) and (Bertsimas and Niño-Mora 1996). These alternative proofs also provide illuminating alternative interpretations of the Gittins index.

Computing the Gittins index can be an onerous task, especially when the state space corresponding to posterior sufficient statistics is large or high-dimensional. As such, approximations to the index have been proposed by Yao et al. (2006), Katehakis and Veinott (1987), and Varaiya et al. (1985) (see Chakravorty and Mahajan 2013 for a survey). Our approach allows us to transparently leverage approaches developed in the literature to quickly compute Gittins indices. For instance, whereas the experiments in the present paper leverage approximations to the Gittins index proposed in Brezzi and Lai (2002) and Powell and Ryzhov (2012), parametric linear programming-based approaches may also be used as an attractive alternative for finite state bandits (Niño-Mora 2007).

It is also worth noting that asymptotic links between Gittins indices and UCB methods have been recognized by Fang and Lai (1987). That paper considers a diffusion approximation to the stopping problem associated with the computation of a Gittins index and shows that the analog to the Gittins index in the context of that problem enjoys an asymptotic expansion that resembles an upper confidence bound. Subsequent to our work here, Russo (2021) establishes an explicit relationship between Gittins indices and Bayesian UCBs.

This paper also relies on Gittins index approximations, and we develop simple general ones that enable our algorithm to be regret optimal.

Bayesian bandit algorithms (such as Bayes UCB, Thompson sampling, or IDS) all admit natural extensions to settings substantially more complex that the independent arm case that is the focus of this paper. It is also the case that the Gittins index theorem has inspired index calculations for more sophisticated bandit problems (an example is Whittle's heuristic; Whittle 1988). Our hope is that together with the use of an increasing discount factor schedule analyzed in this paper, such index policies can provide a starting point for algorithm design in more complex bandit problems.

Finally, we note that others have (contemporaneous with an earlier version of the present work) attempted to leverage the Gittins index in the construction of a Bayesian MAB algorithm. For instance, Kaufmann (2018) considers a variety of heuristics based on a finite horizon version of the Gittins index (essentially, the index proposed by Bradt et al. (1956)) and shows promising empirical results. Lattimore (2016) analyzes the regret under a similar index and shows it to be logarithmic for a *fixed* horizon. Unfortunately, the index policies studied in both Kaufmann (2018) and Lattimore (2016) require a priori knowledge of a horizon. As such, this does not yield an index rule that works for any sufficiently large horizon but rather one that only works for a fixed prespecified horizon. In fact, such schemes cannot be expected to work well for time horizons other than the prespecified horizon determining the index. In contrast, we seek to provide a compelling alternative to the host of state-of-the-art "anytime" regret optimal index rules discussed heretofore.

Finally, in trying to extend the contemporaneous Lattimore (2016) policy to one that does not require the horizon to be prespecified, one may rely on the so-called doubling trick. The doubling trick involves breaking up the MAB problem into a sequence of episodes, each one doubling in length, and analyzing an algorithm that systematically increases the amount of exploration in each episode (see Section 3.2 for details). In using the doubling trick, one incurs a multiplicative increase in regret by a $\log T$ factor yielding a suboptimal regret bound. A recent paper (Besson and Kaufmann 2018) apparently motivated by the same issue, shows that *exponential* doubling tricks can reduce this multiplicative term to a constant (that constant is precisely two) for the Lattimore (2016) approach. Although much better, this bound remains slightly suboptimal. More interestingly, however, the resulting algorithm performs very poorly, by up to almost an order of magnitude worse than vanilla UCB (Besson and Kaufmann 2018).

1.2. Structure of the Paper

The remainder of this paper is organized as follows: in the next section, we state our notation, objectives of interest and key results such as the Lai-Robbins lower bound. The third section focuses on the Gittins index and explains how it fails to minimize regret in a sense that is made clear later. At the end, we address another issue, namely the computational cost of calculating the Gittins index, which inspires us to develop the optimistic Gittins index (OGI) policy. Section 4 establishes an optimal regret bound for the algorithm; namely, one that matches the Lai-Robbins lower bound. Following that, Section 5 presents experiments showing how OGI achieves lower Bayesian regret than state of the art policies and is computationally efficient. In addition to the problem studied in earlier sections of the paper, we also demonstrate computationally the algorithm's effectiveness in a more general setting where it is possible to pull several arms at once in every iteration. Finally, in Section 6 we state open questions that remain following this work.

2. Model and Preliminaries

The multiarmed bandit problem is described via a handful of primitives. These include the notion of an arm, the concept of an arm selection rule or policy and the notion of regret. This section seeks to formalize each of these notions.

2.1. Arms

We consider a multiarmed bandit problem with $A > 1$ arms. We index arms by i and denote by $\mathcal{A}$ the set of all arm indices, $\{1, \ldots, A\}$. At each point in time, $t \in \mathbb{N}$, we are permitted to select or pull a single arm. We denote by $N_i(t)$ the cumulative number of pulls of arm i up to and including time t. If arm i were pulled at time t, we collect a reward $X_{i,N_i(t)} \in \mathbb{R}$.

All random variables are generated on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For a given arm i, $(X_{i,s}, s \in \mathbb{N})$ is assumed to be an independent and identically distributed (i.i.d.) sequence of random variables, each distributed according to a distribution $p_{\theta_i}(\cdot)$. Denote by $\mu(\theta_i)$ the mean of this distribution. Thus, θ_i is a parameter specifying the reward distribution for arm i and we denote by $\mathbf{\Theta}$ the set of all possible values of θ_i. We let

$$\boldsymbol{\theta} \triangleq (\theta_1, \theta_2, \ldots, \theta_A)$$

denote a tuple of the parameters defining the reward distributions for all of the arms. $(X_{i,s}, i \in \mathcal{A}, s \in \mathbb{N})$ is itself assumed to be an independent sequence of random variables so that the arms are independent.

2.2. Policies

At every point in time, we choose an arm to pull according to some history dependent policy π. Formally, any policy π is specified by an $\mathcal{A}$-valued stochastic process $(\pi_t, t \in \mathbb{N})$. Denote by $\mathcal{F}_t$ the filtration generated by the sequence of indices of the first t arms pulled, as well as their corresponding rewards

$$\mathcal{F}_t \triangleq \sigma\left(\left(\pi_s, X_{\pi_s, N_{\pi_s}(s)}\right), s = 1, 2, \ldots, t\right).$$

We require that the process π_t be $\mathcal{F}_t$-predictable[1] and denote by Π the space of all such policies.

2.3. Frequentist Regret and Regret Optimality

Over time, the agent accumulates rewards, and we denote by

$$V(\pi, T, \boldsymbol{\theta}) := \mathsf{E}\left[\sum_{t=1}^{T} X_{\pi_t, N_{\pi_t}(t)} \,\middle|\, \boldsymbol{\theta}\right]$$

the reward accumulated up to time T when using policy π. Denote by $\mu^*(\boldsymbol{\theta})$ the maximum expected reward across arms for a given $\boldsymbol{\theta}$: $\mu^*(\boldsymbol{\theta}) \triangleq \max_i \mu(\theta_i)$. The frequentist regret of a policy over T time periods, for a given $\boldsymbol{\theta} \in \Theta^A$, is the expected shortfall against always pulling the optimal arm for that $\boldsymbol{\theta}$, namely

$$\text{Regret}(\pi, T, \boldsymbol{\theta}) := T\mu^*(\boldsymbol{\theta}) - V(\pi, T, \boldsymbol{\theta}).$$

In a seminal paper, Lai and Robbins (1985) established a lower bound on achievable regret. They showed that for any policy $\pi \in \Pi$, and any $\boldsymbol{\theta}$ such that the set of arms with expected reward $\mu^*(\boldsymbol{\theta})$ is a singleton, we must have

$$\liminf_T \frac{\text{Regret}(\pi, T, \boldsymbol{\theta})}{\log T} \geq \sum_i \frac{\mu^*(\boldsymbol{\theta}) - \mu(\theta_i)}{d_{\text{KL}}(p_{\theta_i}, p_{\theta_{i^*}})}, \tag{1}$$

where d_{KL} is the Kullback-Liebler divergence. A policy π' that achieves this lower bound is considered *regret optimal*. Specifically, π' is regret optimal if and only if

$$\limsup_T \frac{\text{Regret}(\pi', T, \boldsymbol{\theta})}{\log T} \leq \sum_i \frac{\mu^*(\boldsymbol{\theta}) - \mu(\theta_i)}{d_{\text{KL}}(p_{\theta_i}, p_{\theta_{i^*}})}.$$

2.4. Bayesian Bandits

A *Bayesian* MAB problem is endowed with additional structure: we are given a prior on $\boldsymbol{\theta}$. Specifically, we suppose that each θ_i is, in fact, an independent draw according to some prior distribution q that is supported on $\boldsymbol{\Theta}$. We assume that q is conjugate to p_{θ_i} and that $\mathsf{E}[|\mu(\theta_i)|] < \infty$. With a minor abuse of notation, we denote by y the sufficient statistic specifying q and by $\mathcal{Y}$ the set of all possible values of y.

An algorithm that leverages knowledge of q will frequently maintain a posterior distribution on θ_i given observations from that arm. To that end, denote by $q_{i,s}$ the posterior distribution on θ_i given the first s rewards from that arm, $X_{i,1}, X_{i,2}, \ldots, X_{i,s}$. Denote by $y_{i,s}$ the corresponding values of the sufficient statistic describing the posterior. Of course, $q_{i,0} \triangleq q$.

Now, one can define a notion of regret that depends on the prior q. Specifically, the Bayes risk (or Bayesian regret) for any policy π is simply the expected regret over draws of $\boldsymbol{\theta}$ according to the prior q:

$$\text{Regret}(\pi, T) := \int_{\Theta} \text{Regret}(\pi, T, \boldsymbol{\theta}) q^A(d\boldsymbol{\theta}).$$

In yet another landmark paper, Lai (1987) showed that for a restricted class of priors q, a similar class of algorithms to those found to be regret optimal in Lai and Robbins (1985) was also Bayes optimal in the sense that they achieved a lower bound on the Bayes risk (also established in Lai 1987). Interestingly, however, this class of algorithms ignores information about the prior altogether; that is, they do not require knowledge of q. However, this class of algorithms is not anytime and does require knowledge of the problem horizon. A number of algorithms that *do* exploit prior information have in recent years received a good deal of attention; these include Thompson sampling (Thompson 1933), Bayes UCB (Kaufmann et al. 2012b), KL-UCB (Garivier and Cappé 2011), and IDS (Russo and Van Roy 2018). All these algorithms maintain a posterior on the mean of an arm but leverage this posterior in different ways. It has been empirically observed that these approaches offer excellent performance, even in a frequentist sense. In fact, Thompson sampling, Bayes UCB, and KL-UCB have each been shown to be regret optimal in the sense of meeting the lower bound (1).

2.5. Discounted Infinite Horizon Objective

Assuming the structure afforded by the Bayesian setting, that is, the prior q, one may consider a distinct objective to Bayesian regret. Specifically, given some fixed discount factor $\gamma < 1$, one could consider the problem of maximizing discounted infinite horizon rewards. Assume we start with a prior q on the mean of any arm and as before denote by y the sufficient statistic corresponding to this prior. In the parlance of MDPs, we might refer to this as starting with every arm in state y. For a given policy π, we define the expected discounted infinite horizon reward under that policy according to

$$V_\gamma^\pi(\mathbf{y}) = \mathsf{E}_{\mathbf{y}}\left[\sum_{t=1}^{\infty} \gamma^{t-1} X_{\pi_t, N_{\pi_t}(t)}\right],$$

where $\mathbf{y}$ is an A-tuple with every entry equal to y. The subscript on the expectation indicates that θ_i is drawn according to a prior with sufficient statistic y for each arm i. An optimal such policy must solve the problem

$$V_\gamma^*(\mathbf{y}) \triangleq \max_{\pi \in \Pi} V_\gamma^\pi(\mathbf{y}).$$

This is, of course a challenging MDP in that it has a high dimensional state space ($\mathcal{Y}^A$). The celebrated Gittins index theorem (which we present in the next section) provides an approach to computing an optimal policy by instead simply solving a dynamic program on the state space $\mathcal{Y}$.

3. Optimistic Gittins Index Algorithm

This section introduces the notion of an optimistic Gittins index and presents an algorithm for the MAB problem that we will subsequently show is optimal in that it achieves the Lai-Robbins lower bound. We will begin with reviewing the Gittins index theorem for the discounted infinite horizon bandit problem and show that one cannot expect the use of the index from that well known result to yield a regret optimal policy for the MAB problem. We then show that the use of the Gittins index in concert with an increasing discount factor yields polylogarithmic Bayesian regret. This coarse result motivates the discount factor schedule we eventually propose. Finally, we present a series of optimistic approximations to the Gittins index with the view of minimizing the computational burden of index computation. Putting these ingredients together yields the optimistic Gittins index algorithm that is the subject of our paper. The regret optimality of the optimistic Gittins index, for Beta-Bernoulli bandits, is proved in Section 4 (Theorem 1). That is our main theoretical result.

3.1. Gittins Index and Regret

The Gittins index theorem presents a surprisingly simple solution to the problem of computing an optimal policy for the discounted infinite horizon bandit problem. Specifically, the theorem defines for each arm state $y \in \mathcal{Y}$, an index we denote $v_\gamma(y)$; we define this index shortly. The theorem shows that an arm selection rule that at every time selects the arm with the highest index is optimal. The result is powerful in that the computation of the index for a given arm requires the solution of an MDP on the state space $\mathcal{Y}$, as opposed to solving an MDP on the considerably larger state space $\mathcal{Y}^A$.

One way to compute the Gittins index $v_\gamma(y)$ for an arm in state y is via the so-called retirement value formulation (Whittle 1980). Specifically, $v_\gamma(y)$ is defined as the value of λ that solves

$$\frac{\lambda}{1-\gamma} = \sup_{\tau>0} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \gamma^\tau \frac{\lambda}{1-\gamma}\right], \tag{2}$$

where the subscript on the expectation indicates that the prior on the (say, ith) arm's mean at time $t = 1$, $y_{i,0}$, equals y. If one thought of the notion of retiring as receiving a deterministic reward λ in every period, then the value of λ that solves the above equation could be interpreted as the per-period retirement reward that makes us indifferent between retiring immediately, and the option of continuing to play arm i with the potential of retiring at some future time. The Gittins index policy itself, which we denote by $\pi^{G,\gamma}$, can succinctly be stated as follows:

At time t, play an arm in the set $\arg\max_i v_\gamma(y_{i,N_i(t-1)})$,

where $N_i(0) \equiv 0$ and $y_{i,0}$ *is* understood to be the sufficient statistic corresponding to the prior on that arm. Ignoring computational considerations, we cannot hope for the policy $\pi^{G,\gamma}$ to be regret optimal. In fact, as the result below indicates, one cannot even hope for such a policy to be consistent (i.e., have sublinear regret) in the sense of Lai and Robbins (1985):

Lemma 1. *For any $\gamma > 0$, there exists an instance of the multi armed bandit problem with a prior q^A and vector $\boldsymbol{\theta}$* in the support of q^{A}, for which

$$\mathrm{Regret}(\pi^{G,\gamma}, T, \boldsymbol{\theta}) = \Omega(T).$$

The proof, given in Appendix A, rests on the simple fact that for any fixed discount factor, if the posterior means on the two arms are sufficiently apart, the Gittins index policy will pick the arm with the larger posterior mean. The threshold beyond which the Gittins policy exploits depends on the discount factor and with a fixed discount factor there is a positive probability that the superior arm is never explored sufficiently so as to establish that it is, in fact, the superior arm.

3.2. Increasing Discount Factors Yield Sublinear Bayesian Regret

Lemma 1 tells us that we cannot hope for sublinear regret by applying the Gittins index policy with a constant discount factor. One may naturally wonder whether an increasing discount factor might fix this issue. Now observe that any schedule of increasing discount factors effectively implies a change in the tradeoff between exploration and exploitation. With a fixed discount factor, we have already seen that once the priors between two arms are sufficiently far apart, the Gittins policy will not explore, thereby leading to the possibility of linear regret. As the discount factor increases, the gap between priors that exploration is not justified goes up over time. If we increase this gap too fast, we might incur too much exploration; too slow, and we might incur too little exploration. As such, the schedule at which we increase the discount factor is likely to play a significant role in determining the regret of the resulting policy.

Now notice that the Gittins index policy for a discount factor γ can be viewed as optimal for a *random* finite horizon, distributed geometrically with parameter $1-\gamma$. As γ approaches one, this may be thought of as a near optimal policy for the fixed finite horizon $1/(1-\gamma)$. Now consider for a moment that we had access to a policy that has optimal T period expected regret (assuming T is known in advance). One way to convert such a policy into a policy that has low regret for any T is to use the so-called doubling trick: Apply the optimal policy for the horizon T for T steps, then

the optimal policy for horizon $2T$ for the following $2T$ steps, followed by the optimal policy for $4T$ for the next $4T$ steps, and so forth. Intuitively, such doubling tricks leverage finite time horizon results to get so-called anytime results by picking an update schedule such that the choice of horizon at any point is roughly consistent with the elapsed time up to that point; Besson and Kaufmann (2018) provide an insightful analysis of this trick and its broad applicability. This doubling trick is akin to using a discount factor that increases at roughly the rate $1-1/t$ and will be near-optimal for any horizon. The remainder of this section makes this intuition precise but can be skipped without loss of continuity.

Consider using discount factors that increase at roughly the rate $1-1/t$; specifically, consider setting

$$\gamma_t = 1 - \frac{1}{2^{\lfloor \log_2 t \rfloor}},$$

and consider using the policy that at time t picks an arm from the set $\arg\max_i \nu_{\gamma_t}(y_{i,N_i(t-1)})$. Denote this policy by π^{D}. The following proposition shows that this doubling policy achieves Bayes risk that is within a factor of $\log^2 T$ of the optimal Bayes risk. Specifically, we have the following.

Proposition 1. *Assume that q satisfies the requirements of* theorem 3 *in Lai (1987). Then,*

$$\mathrm{Regret}(\pi^{\mathrm{D}}, T) = O(\log^4 T),$$

where the constant in the big-Oh term depends on the prior q and A.

The proof of this result (Appendix B) relies on showing that the finite horizon regret achieved by using a Gittins index with an appropriate fixed discount factor is within a constant factor of the optimal finite horizon regret. The second ingredient is the doubling trick described above. The previous coarse analysis illustrates that the use of the Gittins index policy together with an increasing discount factor does indeed yield an algorithm with sublinear Bayesian regret. It is worth noting that the previous result does not show that such a policy achieves *optimal* Bayesian regret (the achievable lower bound being on the order of $\log^2 T$; Lai 1987). The analysis does, however, suggest a candidate discount rate schedule that we will eventually show to yield a regret optimal policy (in a frequentist sense).

3.3. Optimistic Approximations to the Gittins Index

The retirement value formulation makes clear that computing a Gittins index is equivalent to solving a discounted, infinite horizon stopping problem. Solving this problem requires substantially more computational effort than, say, Thompson sampling or the Bayes UCB algorithm. In fact, this computation can even be rendered intractable should the prior on the mean of an arm be specified by a high-dimensional parameter vector, that is, should $\mathcal{Y}$ be high-dimensional. This motivates an approximation to the Gittins index that is the subject of this section. Specifically, we introduce a sequence of optimistic approximations to the Gittins index that will alleviate computational burden.

Consider the following alternative stopping problem that requires as input the parameters λ (which has the same interpretation as it did before), and K, an integer limiting the number of steps that we need to look ahead. For an arm in state y (recall that the state specifies sufficient statistics for the current prior on the arm reward), let $R(y)$ be a random variable distributed as the prior on expected arm reward specified by y. Define the retirement value $R_{\lambda,K}(s,y)$ according to

$$R_{\lambda,K}(s,y) = \begin{cases} \lambda, & \text{if } s < K \\ \max(\lambda, R(y)), & \text{otherwise.} \end{cases}$$

For a given K, the *optimistic Gittins index* for arm i in state y is now defined as the value for λ that solves

$$\frac{\lambda}{1-\gamma} = \sup_{1\le\tau\le K} \mathsf{E}_y\left[\sum_{s=1}^{\tau} \gamma^{s-1} X_{i,s} + \gamma^{\tau} \frac{R_{\lambda,K}(\tau, y_{i,\tau-1})}{1-\gamma}\right], \quad (3)$$

where we recall that the subscript on the expectation indicates that $y_{i,0} = y$. We denote the solution to this equation by $v_\gamma^K(y)$.

Let us interpret the previous stopping problem. Assume that after $\tau-1$ pulls of an arm, we choose to pull that arm only once more and then subsequently retire. If τ were less than K, we then receive a reward λ per period, over the rest of time, discounted at the rate γ. This is no different from what happens in the stopping problem defining the usual Gittins index (2). In contrast, unlike that formulation, we are forced to retire after the Kth arm pull if we have not done so already. Should we retire at that time, nature reveals the true mean reward of the arm, and we receive the greater of that quantity and λ as our per period retirement payoff. In this manner, one is better off than in the stopping problem inherent to the definition of the Gittins index (2), so that we use the moniker optimistic. The following lemma formalizes this intuition.

Lemma 2. *The function $v_\gamma^K(y)$ is nonincreasing in K for all discount factors γ and states $y \in \mathcal{Y}$. Moreover, $v_\gamma^K(y) \to v_\gamma(y)$ as $K \to \infty$.*

Proof. See Appendix C.1. □

Now, because we need to look ahead at most K steps in solving the stopping problem implicit in the previous definition, the computational burden in index computation is limited. In fact, we will see in a

subsequent section that *even the choice of* $K = 1$ will make for a compelling policy.

3.4. Optimistic Gittins Index Algorithm

The discussion thus far suggests a simple class of bandit algorithms we dub the optimistic Gittins index (OGI) algorithm. The algorithm itself requires as input a prior on arm means (as does any Bayesian algorithm for the MAB), and a parameter K.

The OGI algorithm may be summarized succinctly as follows:

At time t *play an arm in the set* $\arg\max_i v^K_{\gamma_t}(y_{i,N_i(t-1)})$,

where $\gamma_t = 1 - \frac{1}{t}$.

The following section will establish that the previous algorithm achieves the Lai-Robbins lower bound (and thus is regret optimal) for *any* finite K. We will establish this result for Beta priors and Bernoulli rewards. Although we do not state this result formally until the next section (see Theorem 1), it is worth pausing to reflect on the implications of such a result:

1. As K grows large the optimistic Gittins index approaches the Gittins index. The result thus establishes that the use of a set of arbitrarily close approximations to the Gittins index with the discount factor schedule $\gamma_t = 1 - 1/t$ is a regret optimal algorithm. This is a simple, surprising result that bridges two very different flavors of the multiarmed bandit problem. It also suggests the natural conjecture that the use of the Gittins index itself with the discount factor schedule $\gamma_t = 1 - 1/t$ is a regret optimal algorithm.

2. At the other end, because the result establishes regret optimality for any finite K, we have regret optimality for $K = 1$. Computing the optimistic Gittins index in this case is a particularly trivial task and offers the specter of a computationally practical algorithm. In fact, in Section 5, we shall see precisely this: the choice of $K = 1$ yields an index that materially outperforms a host of state-of-the-art alternatives while requiring little to no computational overhead relative to even the simplest schemes.

We end this section with some brief commentary on computation. For concreteness, let us focus on the case of a Beta-Bernoulli bandit. First, we note that solving the stopping problem implicit in the definition of $v^K_\gamma(y)$ for any given value of the retirement subsidy λ requires the solution of a relatively simple dynamic program with just $O(K)$ states. This dynamic program can be solved exactly in $O(K^2)$ time. The optimal value of λ can be found by bisection. For small values of K this is substantially less effort than computing a Gittins index. The case of $K = 1$ is particularly appealing. There, Equation (3) simplifies to

$$\lambda = \mathsf{E}[R(y)] + \gamma \mathsf{E}[(\lambda - R(y))^+]. \tag{4}$$

This equation is easily solved via a method such as Newton-Raphson.[2] In fact, the gradients required for the use of the Newton-Raphson approach are often readily available in closed form. To wit, in the case of a Beta prior with sufficient statistics (a, b), (4) reduces to

$$\lambda = \frac{a}{a+b}\left(1 - \gamma F^\beta_{a+1,b}(\lambda)\right) + \gamma\lambda F^\beta_{a,b}(\lambda) \triangleq g_{a,b}(\lambda),$$

wherein we see that $\frac{\partial}{\partial\lambda} g_{a,b}(\lambda)$ can be computed in closed form. This makes the use of the Newton-Raphson method for the solution of the equation $\lambda = g_{a,b}(\lambda)$ particularly simple. In our computational experiments, we will see that the choice of $K = 1$ already provides a material improvement in empirical performance over state-of-the-art alternatives.

4. Analysis and Regret Bounds

We establish a regret bound for the OGI algorithm that applies when the prior distribution q is Beta(1, 1) (i.e., uniform) and arm rewards are Bernoulli. The choice of uniform prior is natural since it represents the lack of prior knowledge about arm rewards; the analysis here can potentially be extended to certain other priors. The result shows that the algorithm, in that case, meets the Lai-Robbins lower bound and is thus asymptotically optimal in a frequentist sense. After stating the main theorem, we briefly discuss a generalization to the algorithm.

In the sequel, we will simplify notation and let $d(x, y) := d_{\mathrm{KL}}(\mathrm{Ber}(x), \mathrm{Ber}(y))$ denote the KL divergence between Bernoulli random variables with parameters x and y. We will also refer to the OGI policy, which uses a look-ahead parameter of K, as $\pi^{\mathrm{OG},K}$ and will write the OGI of the ith arm at time t as $v^K_{i,t} \triangleq v^K_{1-1/t}(y_{i,N_i(t-1)})$. That way, for the sake of brevity, we will suppress the index's dependence on $y_{i,N_i(t-1)}$. We are now ready to state the main result.

Theorem 1. *Let* $\epsilon > 0$ *and consider an OGI policy configured with a parameter* $K \in \mathbb{N}$ *and that assumes Beta*(1, 1) *priors. For the multiarmed bandit problem with Bernoulli rewards and any parameter vector* $\boldsymbol{\theta} \in (0,1)^A$, *there exists* $T^* = T^*(\epsilon, \boldsymbol{\theta}, K)$ *and* $C = C(\epsilon, \boldsymbol{\theta}, K)$ *such that for all* $T \geq T^*$,

$$\mathrm{Regret}\left(\pi^{\mathrm{OG},K}, T, \boldsymbol{\theta}\right) \leq \sum_{\substack{i=1,\ldots,A \\ i \neq i^*}} \frac{(1+\epsilon)^2(\theta^* - \theta_i)}{d(\theta_i, \theta^*)} \log T + C(\epsilon, \boldsymbol{\theta}, K), \tag{5}$$

where $C(\epsilon, \boldsymbol{\theta}, K)$ *is a constant that is determined by* ϵ, *the parameter* $\boldsymbol{\theta}$, *and* K.

Proof. Assume, without loss of generality, that the first arm is uniquely optimal so that $\theta^* = \theta_1$. Fix an arbitrary suboptimal arm, which for convenience we will say is the second arm. We will strategically fix

three constants in between the expected rewards of the first and second arms, namely θ_1 and θ_2. In particular, we let $\eta_1, \eta_2, \eta_3 \in (\theta_2, \theta_1)$ be chosen such that $\eta_1 < \eta_2 < \eta_3$, $d(\eta_1, \eta_3) = \frac{d(\theta_2, \theta_1)}{1+\epsilon}$ and $d(\eta_2, \eta_3) = \frac{d(\eta_1, \eta_3)}{1+\epsilon}$. Next, we define the constant $L(T) := \frac{\log T}{d(\eta_2, \eta_3)}$ to be, intuitively, the optimal length of the exploration period.

The main step in this proof will be to upper bound the expected number of pulls of the second arm, as follows:

$$\begin{aligned}
\mathsf{E}[N_2(T)] &\le L(T) + \sum_{t=\lfloor L(T)\rfloor+1}^{T} \mathbb{P}\left(\pi_t^{\mathrm{OG},K} = 2,\, N_2(t-1) \ge L(T)\right) \\
&\le L(T) + \sum_{t=1}^{T} \mathbb{P}\left(v_{1,t}^K < \eta_3\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(\pi_t^{\mathrm{OG},K} = 2,\, v_{1,t}^K \ge \eta_3,\, N_2(t-1) \ge L(T)\right) \\
&\le L(T) + \sum_{t=1}^{T} \mathbb{P}\left(v_{1,t}^K < \eta_3\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(\pi_t^{\mathrm{OG},K} = 2,\, v_{2,t}^K \ge \eta_3,\, N_2(t-1) \ge L(T)\right) \\
&\le \frac{(1+\epsilon)^2 \log T}{d(\theta_2, \theta_1)} + \underbrace{\sum_{t=1}^{\infty} \mathbb{P}\left(v_{1,t}^K < \eta_3\right)}_{A} \\
&\quad + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(\pi_t^{\mathrm{OG},K} = 2,\, v_{2,t}^K \ge \eta_3,\, N_2(t-1) \ge L(T)\right)}_{B},
\end{aligned} \tag{6}$$

where the first step is the same as in the analysis of Auer et al. (2002) and applies to any bandit policy. All that remains is to show that terms A and B are bounded by constants. These bounds are given in Lemmas 3 and 4 whose proofs we will now describe at a high-level and defer the full details to the Appendix D.2 and D.3.

Lemma 3 (Bound on Term A). *For any $\eta < \theta_1$, the following bounds holds for some constant $C_1 = C_1(\epsilon, \theta_1, K)$*

$$\sum_{t=1}^{\infty} \mathbb{P}\left(v_{1,t}^K < \eta\right) \le C_1.$$

Proof Outline. The goal is to bound $\mathbb{P}(v_{1,t}^K < \eta)$ by an expression that decays fast enough in t so that the series converges. This demonstrates that the algorithm encourages enough exploration such that the optimal arm is never underestimated for too long, in expectation. Specifically, we show that there exists a positive constant h so that $\mathbb{P}(v_{1,t}^K < \eta) = O(\frac{1}{t^{1+h}})$ using an induction argument. Proving the base case requires using properties specific to Beta and Bernoulli random variables, whereas the inductive step is more general. The full proof is contained in Appendix D.2.

We remark that the core steps in the proof of Lemma 3, at least in the base case of the induction, rely on properties of the Beta and Bernoulli variables. Because of this, we suspect our analysis can strengthen a similar theoretical result for the Bayes UCB algorithm. In particular, the main theorem of Kaufmann et al. (2012b) states that the quantile parameter in the Bayes UCB algorithm should be $1 - 1/(t\log^c T)$ for some constant $c \ge 5$. However, what is perplexing is that their simulation experiments suggest that using a simpler sequence of quantiles, namely $1 - 1/t$, results empirically in a lower mean regret. By using techniques in our analysis, it is possible to prove that the use of the quantiles $1 - 1/t$ would lead to the same optimal regret bound. Therefore, the scaling by $\log^c T$ is unnecessary. □

Lemma 4 (Bound on Term B). *There exists $T^* = T^*(\epsilon, \boldsymbol{\theta})$ sufficiently large and a constant $C_2 = C_2(\epsilon, \theta_1, \theta_2)$ so that for any $T \ge T^*$, we have*

$$\sum_{t=1}^{T} \mathbb{P}\left(\pi_t^{\mathrm{OG},K} = 2,\, v_{2,t}^K \ge \eta_3,\, N_2(t-1) \ge L(T)\right) \le C_2.$$

Proof Outline. This relies on a concentration of measure result and the assumption that the second arm was sampled at least $L(T)$ times. Because our index is nonincreasing in K, from Lemma 2, it is enough to only consider the simplest case when $K = 1$. The full proof is given in Appendix D.3. □

Lemmas 3 and 4, together with (6), imply that

$$\mathsf{E}[N_2(T)] \le \frac{(1+\epsilon)^2 \log T}{d(\theta_2, \theta_1)} + C_1 + C_2,$$

and from this, the regret bound follows. □

4.1. Generalizations and a Tuning Parameter

As we have shown, the OGI algorithm is regret optimal for the Bernoulli bandit problem. Moreover, it is possible to generalize our algorithm to problems with any bounded reward distribution and prove a weaker $O(\log T)$ regret bound. We see this immediately from the discussion in Agrawal and Goyal (2012), where it is shown that any bandit algorithm that is regret optimal for the Bernoulli bandit problem can be modified to yield an algorithm that has $O(\log T)$ regret in a general setting with (bounded) stochastic rewards. Informally, this would require emulating a Bernoulli bandit problem and assuming Beta(1, 1) priors as before.

Yet another key observation is that the discount factor for OGIs does not need to be exactly $1 - 1/t$. In fact, a tuning parameter can be included to make the discount factor $\gamma_{t+\alpha} = 1 - 1/(t + \alpha)$ instead. Intuitively, this would encourage a greater degree of exploration over the arms. An inspection of the proofs of Lemmas 3 and 4 shows that the result in Theorem 1 would still hold were one to use such a tuning parameter. In

practice, performance is remarkably robust to our choice of K and α.

5. Computational Experiments

Our goal is to benchmark OGIs against state-of-the-art Bayesian algorithms. Specifically, we compare ourselves against Thomson sampling, Bayes UCB, and IDS. Each of these algorithms has in turn been shown to substantially dominate other extant schemes. Our experimental setup closely follows that of Russo and Van Roy (2018), Kaufmann et al. (2012a), and Chapelle and Li (2011). The experiment from Kaufmann et al. (2012a) is deferred to Appendix E.1 because it is brief and sends a similar message to the rest of this section. We conclude with a novel experiment to test the problem with multiple simultaneous arm pulls.

For most experiments, we configure the OGI algorithm with $K = 1$ to keep the computational burden under control (there we simply use the Newton-Raphson method to compute the index directly). In one experiment, included for completeness, we test OGI with $K = 3$ and $K = \infty$, where the latter is equivalent to using Gittins indices. There we use direct dynamic programming for the $K = 3$ case and rely on an approximation because of Powell and Ryzhov (2012) for the $K = \infty$ case. The purpose of those experiments is to show the (limited) value of a higher lookahead in the OGI algorithm.

We use a common discount factor schedule in all experiments setting $\gamma_t = 1 - 1/(100 + t)$. The choice of $\alpha = 100$ is second order; our conclusions remain unchanged and actually appear to improve in an absolute sense with other choices of α. In addition, in one experiment we examine the regret of OGI relative to its competitors up to a horizon of 10^6 epochs, so that this choice of α does not represent an attempt to tune the performance of OGI for a specific time horizon.

5.1. Smaller-Scale Experiments with IDS

This section considers a series of smaller scale experiments (10 arms, 1,000 time periods) drawn from the paper introducing the IDS algorithm (Russo and Van Roy 2018). A major consideration in running these experiments is that the CPU time required to execute IDS, the closest competitor, based on the current suggested implementation is orders of magnitudes greater than that of the index schemes or Thompson sampling. The main bottleneck is that IDS uses numerical integration, requiring the calculation of a Cumulative Distribution Function (CDF) over, at least, hundreds of iterations. By contrast, the version of OGI with $K = 1$ uses 10 iterations of the Newton-Raphson method.

5.1.1. Gaussian. We replicate the Gaussian experiments from Russo and Van Roy (2018). In the first experiment (Table 1), the arms generate Gaussian rewards $X_{i,t} \sim \mathcal{N}(\theta_i, 1)$, where each θ_i is independently drawn from a standard Gaussian distribution. That is, the prior q on each arm's reward is a stand Gaussian prior. We simulate 1,000 independent trials with 10 arms and 1,000 time periods. The implementation of OGI in this experiment uses $K = 1$. It is difficult to compute exact Gittins indices in this setting, but a classical approximation for Gaussian bandits does exist (Powell and Ryzhov 2012, chapter 6.1.3). We term the use of that approximation OGI(∞) Approx. It is shown in Powell and Ryzhov (2012) that the Gittins index here can be expressed in terms of a univariate function, which cannot be computed analytically but can be approximated reasonably well with a piecewise closed-form function. As mentioned in that book, the approximation is more accurate for smaller values of the posterior variance; that is, as we play an arm more times, we expect its Gittins index approximation to improve.

Table 1. Gaussian Experiment

	OGI(1)	OGI(∞) Approx.	IDS	TS	Bayes UCB
Mean	49.19	47.64	55.83	67.40	60.30
Standard error	1.61	1.6	2.08	1.5	1.43
25%	17.49	16.88	18.61	37.46	31.41
50%	41.72	40.99	40.79	63.06	57.71
75%	73.24	72.26	78.76	94.52	86.40
CPU time (s)	0.02	0.01	11.18	0.01	0.02

Note. OGI(1) denotes OGI with $K = 1$, whereas OGI Approx. uses the approximation to the Gaussian Gittins index from Powell and Ryzhov (2012).

In addition to regret, we show the average CPU time taken, in seconds, to execute each trial.

The key feature of the results here is that OGI offers an approximately 10% improvement in regret over its nearest competitor IDS and larger improvements (20% and 40%, respectively) over Bayes UCB and Thompson sampling. The best performing policy is OGI with the specialized Gaussian approximation because it gives a closer approximation to the Gittins index. At the same time, OGI is essentially as fast as Thompson sampling and three orders of magnitude faster than its nearest competitor (in terms of regret).

5.1.2. Bernoulli. We next replicate the Beta-Bernoulli experiments from Russo and Van Roy (2018). In this experiment, regret is simulated over 1,000 periods, with 10 arms each having a uniformly distributed Bernoulli parameter. We compute approximations to the exact Gittins index, that is, OGI(∞), via value iteration.[3] We simulate 1,000 independent trials, and Table 2 summarizes the results.

Each version of OGI outperforms other algorithms, and the one that uses exact Gittins indices shows the lowest mean regret. Perhaps, unsurprisingly, when OGI looks ahead three steps (or when the lookahead

Table 2. Bernoulli Experiment

	OGI(1)	OGI(3)	OGI(∞)	IDS	TS	Bayes UCB
Mean	18.12	18.00	17.52	19.03	27.39	22.71
Standard error	0.65	0.64	0.68	0.67	0.57	0.56
25%	6.26	5.60	4.45	5.85	14.62	10.09
50%	15.08	14.84	12.06	14.06	23.53	18.52
75%	27.63	27.74	24.93	26.48	36.11	30.58
CPU time (s)	0.19	0.89	(?) hours	8.11	0.01	0.05

Notes. OGI(K) denotes the OGI algorithm with a K step approximation and tuning parameter $\alpha = 100$. OGI(∞) is the algorithm that uses Gittins indices.

is not limited), it performs better than with a single step. It is, however, apparent that in each of these cases the improvement over simply setting $K = 1$ is marginal. Indeed, looking ahead one step is a reasonably close approximation to the Gittins index in the Bernoulli problem. In the Appendix E.3, we report the approximation error in approximating the Gittins index for various choice of K. When using an optimistic one-step approximation, the error is around 15%, and if K is increased to three, the error drops to around 4% (Tables E.3 and E.4 in the Appendix).

As an aside, we note that the regret we computed for the IDS algorithm is slightly different from that reported by Russo and Van Roy (2018). Specifically, we obtain slightly lower regret for IDS than they report in the Gaussian experiments and slightly higher values for the Beta-Bernoulli case; we include a link to the code we used to implement the algorithms[4] as a reference.

5.2. Large-Scale Experiment

This experiment replicates a large-scale synthetic experiment in Chapelle and Li (2011). Here the arms' rewards are Bernoulli, and their means are independently sampled from a uniform prior. Every algorithm that we will test assumes this same prior over the arms' mean rewards. The key feature here is that we simulate a longer horizon of $T = 10^6$ and include a large number of arms; particularly, we let $A = 100$. This is an order of magnitude greater than in the majority of synthetic bandit experiments we are aware of. Our goal is to see how the algorithms scale both computationally and in terms of performance. Such a setup is practically relevant because in applications such as e-commerce or online advertising, the problems of interest are typically modeled with many arms relative to the horizon, where each arm could represent a product or ad.

Because all the methods we test in our numerical experiments are regret optimal, any relative difference in regret must shrink after a sufficiently large number of time periods. The length of time for this 'burn in' period intuitively depends on the number of arms in the problem. In particular, we can think of the horizon as giving us a rough budget on the number of trials per arm via the ratio T/A. The idea is that with more trials per arm we should expect a smaller relative difference between the algorithms (and indeed the theoretical guarantees for the algorithms require this to happen). We will see that even when the ratio T/A and A itself are large, there is a substantial difference between OGI and the competing benchmarks in both a relative and absolute sense.

As this experiment requires an order of magnitude more iterations than the earlier ones, we are only able to simulate the fastest algorithms, which are OGI with $K = 1$, Thompson sampling, and Bayes UCB. It was not possible to include IDS because its performance is hindered by the fact that each arm pull decision requires time that is quadratic in the number of arms to compute.

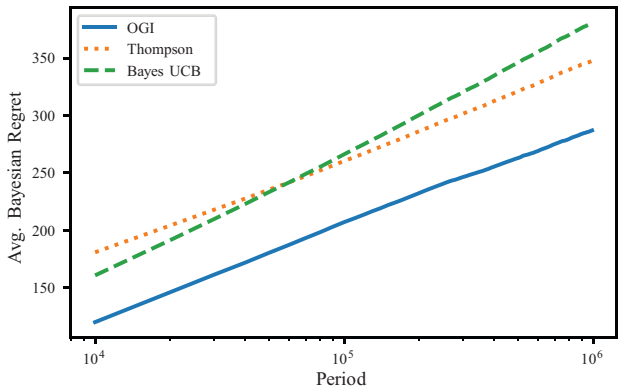
We show the algorithms' regret averaged over 5,000 trials in Figure 1 and Table 3.

As before, the OGI scheme consistently dominates the other two. What is particularly interesting is that despite going out to a horizon of 10^6 time periods, the relative improvement in regret over these algorithms remains substantial. For instance, going from a horizon length of 2×10^5 (corresponding to a heuristic budget of $T/A = 2{,}000$ pulls per arm) to a horizon length of 10^6 (corresponding to a heuristic budget of $T/A = 10{,}000$ pulls per arm) resulted in the relative improvement offered by OGI shrining only marginally, from 18.9% to 17.4%.

5.3. Bandits with Multiple Arm Pulls

In this section, we consider a somewhat exploratory experiment; we seek to adapt OGI to a more complex bandit problem (here we allow for multiple simultaneous arm pulls). Again, in the discounted infinite

Figure 1. (Color online) Cumulative Regret in the Large-Scale Problem of This Section Averaged over 5,000 Independent Trials



Note. We plot the number of periods T on a logarithmic scale.

Table 3. Regret in the Large-Scale Experiment from OGI, Thompson Sampling, and Bayes UCB

T/A	OGI	Thompson	Bayes UCB	Relative improvement (%)	Absolute improvement
2,000	230.5	284.4	297.9	18.9	53.9
4,000	254.7	311.6	333.5	18.3	57.0
6,000	268.6	327.4	354.5	18.0	58.8
8,000	279.1	339.2	369.6	17.7	60.1
10,000	287.1	347.7	380.7	17.4	60.6

Note. The last two columns show the relative and absolute difference from Thompson sampling, which is the closest competitor to OGI.

horizon setting, a number of heuristic approaches have been proposed to adapt the Gittins index to more complex settings; a good example is the so-called Whittle relaxation for restless bandits. One might consider applying those same heuristic strategies to the OGI.

For this experiment, we consider a more general MAB problem, where the agent is able to pull up to a certain number (say $m < A$) of the arms simultaneously. To describe the problem, we recall that A is the total number of arms and define $\mathcal{D}_m := \{d \in \{0,1\}^A : \sum_i d_i \le m\}$ to be the set of all A-dimensional binary vectors with up to m ones in them, which we take to be the action space. Let $X_t = (X_{1,t}, \ldots, X_{A,t})$ be a tuple of (potential) rewards from the A arms at time t, where the definition of $X_{i,t}$ for any arm i is the same as in Section 2. Given a decision $d \in \mathcal{D}_m$, which encodes the subset of arms pulled, the reward $d^\top X_t$ is earned and an arm j's reward $X_{j,t}$ is observed if and only if that arm is pulled, that is, $d_j = 1$. We can then define a policy $(\pi_t, t \in \mathbb{N})$ to be a $\mathcal{D}_m$-valued stochastic process adapted to an information set generated by past actions and observed feedback. A policy π's regret is given by the following equation:

$$\text{Regret}(\pi, T) = T \cdot \mathsf{E}\left[\max_{d \in \mathcal{D}_m} \sum_{i=1}^{A} d_i \mu(\theta_i)\right] - \sum_{t=1}^{T} \mathsf{E}[\pi_t^\top X_t],$$

where the expectation is over the randomness in the rewards, the prior, and the policy's actions.

We propose a heuristic to this problem using our scheme, which is to compute the OGI of every arm, at time t, using a discount factor of $1 - 1/t$ (just as before). However, for this problem, we pick m arms with the largest indices. This is essentially Whittle's heuristic (Whittle 1988), which was originally given for the restless bandit problem but can be described as picking several arms with the largest Gittins indices. We break ties randomly, and it would be interesting to explore smarter tiebreaking schemes that could improve performance (Brown and Smith 2020).

To test our policy, we simulate $A = 6$ binary arms with uniformly distributed biases and fix $m = 3$. We benchmark our heuristic against Thompson sampling and IDS. Because the arms give independent Bernoulli rewards, we will use a flat Beta prior for all the algorithms. We implement the version of IDS designed for the linear bandit problem because this experiment is a special case of a linear bandit. Our implementation of IDS also uses 10,000 Monte Carlo samples per iteration.

The results, produced from 1,000 independent trials, are summarized in Figure 2 and Table 4. We notice a significant spread in the performance between OGI and both Thompson sampling and IDS. Just like for our main algorithm, the primary computational bottleneck in using OGI comes from solving the stopping problem, and this can be onerous if K is large. However, as the results suggest, the policy works well even for low to moderate look-ahead parameters. The experiment here sets the stage for an exploration of the appropriate extensions to the OGI algorithm for more complex bandit problems (such as contextual bandits), which we leave for future work.

The results, produced from 2,000 independent trials, are summarized in Figure 2 and Table 4. The horizon is limited to 250 time periods because of the increased computational effort required to execute a single trial of both the IDS algorithm and Whittle's heuristic, when $K > 1$. This extra time is on the order of minutes for these algorithms. For the sake of simplicity, we dub this algorithm as exactly Whittle's heuristic for the remainder of this section.

We notice a significant spread in performance between Whittle's heuristic and Thompson sampling. Meanwhile, IDS and Whittle's heuristic show similar performance; however, the computational cost of running the latter algorithm is up to 25 times lower (IDS

Figure 2. (Color online) Regret for Bandits with Multiple Simultaneous Arm Pulls

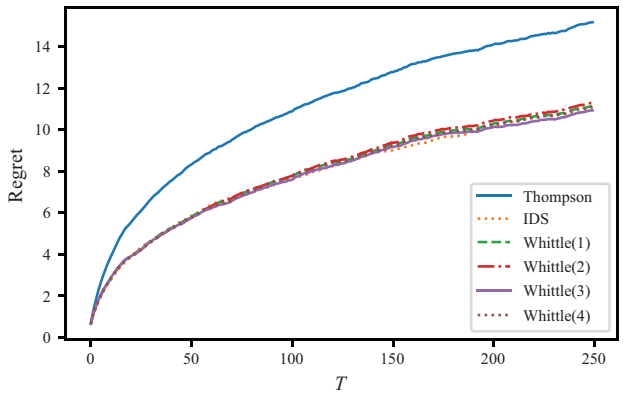


Table 4. Regret from the Multiple Arm Pulls Experiment

	IDS	Thompson	Whittle(1)	Whittle(2)	Whittle(3)	Whittle(4)
Mean	10.97	15.23	11.13	11.29	10.93	11.07
Standard error	0.21	0.13	0.14	0.15	0.15	0.15
25%	1.18	6.60	1.66	1.57	1.20	1.39
50%	10.84	14.75	10.34	9.96	9.91	9.74
75%	24.60	23.52	19.62	19.41	19.27	19.13
CPU time (s)	54.45	2.07	14.20	1,122.12	2,196.83	4,106.89

Note. "Whittle(K)" refers to the Whittle heuristic policy, where K look-ahead steps are used in computing the OGI.

requires generating 10,000 Monte Carlo samples on each iteration). Just like for our main algorithm, the primary computational bottleneck in using Whittle's heuristic comes from solving the stopping problem, and this can be onerous if K is large. However, as the results suggest, the policy works well even for low to moderate look-ahead parameters but nonetheless improves slightly when K increases.

6. Conclusions

This paper proposed a novel way for designing Bayesian MAB algorithms by treating the problem of minimizing regret as a sequence of separate MDPs where the discount factor increases from one problem to the next, according to a carefully chosen rate. We showed that the fundamental idea of using such a heuristic results in sublinear regret and, when applied to a binary bandit problem, that a simple and efficient algorithm with a flat Beta prior achieves the optimal rate of growth in regret.

There are some open questions following this work. First, it remains to be proven that playing arms with maximum (exact) Gittins indices together with the increasing discount factor schedule does produce an algorithm whose regret matches the Lai-Robbins lower bound. We have a strong reason to suspect this because of the findings in our numerical experiments. Second, it is worth exploring whether the idea of this framework can be extended to contextual bandit problems where dependencies between arms exist. In our setting, the fact that arms were independent allowed us to exploit the Gittins index, but there could be other ways to approximate optimal solutions to bandit problems with dependent arms.

Appendix A. Proof of Lemma 1

Proof. Consider an instance of the MAB with $A = 2$ arms and Bernoulli rewards. We assume that the prior on arm 1 is degenerate with mean $\lambda = 1/2$, whereas arm 2 has a Beta$(\alpha, \alpha+1)$ prior where α is a parameter we set later. Furthermore, we assume that the true parameter for arm 2 is equal to $\theta \in (1/2, 1)$, which represents a draw from the prior distribution on that arm. Now, the continuation value from pulling arm 1 at this stage is lower bounded by $\frac{1/2}{1-\gamma}$, whereas the continuation value from pulling arm 2 is upper bounded by

$$\frac{\alpha}{1+2\alpha} + \frac{\gamma \mathsf{E}\left[\max\left(R(y_{2,0}), 1/2\right) \mid y_{2,0} = (\alpha, \alpha+1)\right]}{1-\gamma}.$$

It follows that the Gittins index policy must pull arm 1 if

$$\frac{1/2}{1-\gamma} > \frac{\alpha}{1+2\alpha} + \frac{\gamma \mathsf{E}\left[\max\left(R(y_{2,0}), 1/2\right) \mid y_{2,0} = (\alpha, \alpha+1)\right]}{1-\gamma},$$

an inequality that in turn is satisfied if

$$1/2 > \frac{(1-\gamma)\alpha}{1+2\alpha} + \gamma\left(\mathbb{P}\left(R(y_{2,0}) > 1/2 \mid y_{2,0} = (\alpha, \alpha+1)\right) + 1/2\right).$$

However, the right-hand side of the previous expression goes to $\gamma/2 < 1/2$ as $\alpha \to 0$. Consequently, we can choose an α such that the Gittins index policy chooses to pull the first arm; let α^* be the largest such α. Because the state of the first arm does not change (the prior on that arm was assumed degenerate), the same condition must hold at subsequent iterations. Consequently, $\pi^{G,\gamma}$ must incur a T-period frequentist regret lower bounded by $T(\theta - 1/2)$. The result follows. □

Appendix B. Proof of Proposition 1

We first establish notation useful in the proof. Specifically, we let $h_t^{t'}$ denote the sequence of arms pulled and the corresponding rewards earned between periods t and t' ($h_t^{t'} = \emptyset$ if $t' < t$). Recall that $\mathbf{y}$ denotes the A-tuple of sufficient statistics of priors on the arm means of each of the A arms. We denote by $g(\mathbf{y}, h_t^{t'})$ the A-tuple of sufficient statistics of posteriors on the arm means of each of the A arms, obtained starting with the prior $\mathbf{y}$ and observing the sequence of arm pulls $h_t^{t'}$.

We begin by establishing a simple lemma concerning the allocation rule proposed in Lai (1987); we note that the allocation rule there is specified for a *fixed* time horizon and the following lemma extends it to a policy that specifies a choice of arm for every epoch.

Lemma B.1. *Let the prior with sufficient statistic* $\mathbf{y}$ *satisfy the requirements of Theorem 3 in Lai (1987). Then, there exists a policy* $\tilde{\pi}_{\mathbf{y}}$*, and a constant* $C_{\mathbf{y}}$ *for which*

$$\text{Regret}(\tilde{\pi}_{\mathbf{y}}, T) \triangleq \mathsf{E}_{\mathbf{y}}\left[\text{Regret}(\tilde{\pi}_{\mathbf{y}}, T, \theta)\right] \le C_{\mathbf{y}} \log^3 T,$$

for all T.

Proof. By theorem 3 in Lai (1987), we know that there exists a constant $\hat{C}_{\mathbf{y}}$ and a sequence of policies, $\tilde{\pi}_{\mathbf{y},T}$[5] such that

$$\lim_{T\to\infty} \frac{\mathsf{E}_{\mathbf{y}}[\text{Regret}(\tilde{\pi}_{\mathbf{y},T}, T, \boldsymbol{\theta})]}{\log^2 T} = \hat{C}_{\mathbf{y}}.$$

Consequently, there exists a constant $\bar{C}_{\mathbf{y}}$, so that for any T,

$$\mathsf{E}_{\mathbf{y}}[\text{Regret}(\tilde{\pi}_{\mathbf{y},T}, T, \boldsymbol{\theta})] \le \bar{C}_{\mathbf{y}} \log^2 T.$$

Now consider the doubling policy $\tilde{\pi}_{\mathbf{y}}$, which at time t selects arms according to the policy $\tilde{\pi}_{\mathbf{y},2^k}$ applied at the state $g\left(\mathbf{y}, h_{2^{k(t)}-1}^{t-1}\right)$, where $k(t) = \lfloor \log_2(t+1) \rfloor$. In words, this is the policy obtained wherein (a) time is divided into epochs such that the kth epoch extends from time $2^k - 1$ to $2^{k+1} - 2$, and (b) at the start of the kth epoch, we forget everything learned up until that time and subsequently use the policy $\tilde{\pi}_{\mathbf{y},2^k}$ over the course of that epoch. Thus,

$$\begin{aligned}
\mathsf{E}_{\mathbf{y}}[\text{Regret}(\tilde{\pi}_{\mathbf{y}}, T, \boldsymbol{\theta})] &\le \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}_{\mathbf{y}}\left[\mathsf{E}_{\mathbf{y}_{2^k-2}}\left[\text{Regret}\left(\tilde{\pi}_{\mathbf{y},2^k}, 2^k, \boldsymbol{\theta}\right)\right]\right] \\
&= \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}_{\mathbf{y}}\left[\text{Regret}\left(\tilde{\pi}_{\mathbf{y},2^k}, 2^k, \boldsymbol{\theta}\right)\right] \\
&\le \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \bar{C}_{\mathbf{y}} \log^2 2^k \\
&\le \bar{C}_{\mathbf{y}} C \log^3 T,
\end{aligned}$$

where the equality follows from the tower property and where C is some absolute constant. □

Next, we establish a simple result related to the Gittins index policy that relates the finite horizon performance of the policy to the (discounted) infinite horizon performance.

Lemma B.2. *For any $\hat{\mathbf{y}} \in \mathcal{Y}$, and horizon $T' \ge 2$, we have*

$$\mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, T', \boldsymbol{\theta}\right)\right] \le 4\mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, H_{T'}, \boldsymbol{\theta}\right)\right],$$

where $H_{T'}$ is an independent Geometrically distributed random variable with mean T'.

Proof. We have

$$\begin{aligned}
&\mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, T', \boldsymbol{\theta}\right)\right] \\
&= \mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, T', \boldsymbol{\theta}\right)\right] \frac{\mathbb{P}(H_{T'} > T')}{(1-1/T')^{T'}} \\
&\le \mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, H_{T'}, \boldsymbol{\theta}\right) \mid H_{T'} > T'\right] \frac{\mathbb{P}(H_{T'} > T')}{(1-1/T')^{T'}} \\
&\le (1-1/T')^{-T'} \mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, H_{T'}, \boldsymbol{\theta}\right)\right] \\
&\le 4\mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, H_{T'}, \boldsymbol{\theta}\right)\right],
\end{aligned}$$

where the first inequality follows from the fact that $\mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, n, \boldsymbol{\theta}\right)\right]$ is nondecreasing in n, and the second inequality follows from the fact that regret is nonnegative. □

We can now proceed with the proof of the proposition. First, because the Gittins policy with discount factor $1 - 1/T'$ is optimal for a geometrically distributed horizon with mean T', we must have for any $\hat{\mathbf{y}} \in \mathcal{Y}$:

$$\mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\pi^{G,1-1/T'}, H_{T'}, \boldsymbol{\theta}\right)\right] \le \mathsf{E}_{\hat{\mathbf{y}}}\left[\text{Regret}\left(\tilde{\pi}_{\mathbf{y}}, H_{T'}, \boldsymbol{\theta}\right)\right]. \quad \text{(B.1)}$$

However, we have

$$\begin{aligned}
\mathsf{E}_{\mathbf{y}}[\text{Regret}(\pi^D, T, \boldsymbol{\theta})] &\le \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}_{\mathbf{y}}\left[\mathsf{E}_{\hat{\mathbf{y}}_{2^k-2}}\left[\text{Regret}\left(\pi^{G,1-1/2^k}, 2^k, \boldsymbol{\theta}\right)\right]\right] \\
&\le \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}_{\mathbf{y}}\left[4\mathsf{E}_{\hat{\mathbf{y}}_{2^k-2}}\left[\text{Regret}\left(\pi^{G,1-1/2^k}, H_{2^k}, \boldsymbol{\theta}\right)\right]\right] \\
&\le \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}_{\mathbf{y}}\left[4\mathsf{E}_{\hat{\mathbf{y}}_{2^k-2}}\left[\text{Regret}\left(\tilde{\pi}_{\mathbf{y}}, H_{2^k}, \boldsymbol{\theta}\right)\right]\right] \\
&= 4 \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}_{\mathbf{y}}[\text{Regret}(\tilde{\pi}_{\mathbf{y}}, H_{2^k}, \boldsymbol{\theta})] \\
&\le 4C_{\mathbf{y}} \sum_{k=1}^{\lfloor \log(T+2) \rfloor} \mathsf{E}[\log^3 H_{2^k}] \\
&\le 4C_{\mathbf{y}} \sum_{k=1}^{\lfloor \log(T+2) \rfloor} k^3,
\end{aligned}$$

where the first inequality follows simply from the definition of π^D, the second inequality follows from Lemma B.2, the third inequality follows from the aforementioned optimality of the Gittins policy (namely (B.1)), the first equality follows from the tower property, the fourth inequality follows from Lemma B.1, and the fifth and final inequality is simply Jensen's inequality.

Appendix C. Properties of the OGI

This section gives proofs for a few properties of the OGI that are used throughout the paper and particularly in the proof of Theorem 1. It shall be useful, in what follows, to define the continuation value for the Vittles's retirement problem (Whittle 1980) as

$$V_\gamma(y, \lambda) \triangleq \sup_{\tau > 0} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \gamma^\tau \frac{\lambda}{1-\gamma}\right],$$

so that the Gittins index is then the solution in λ to $\lambda/(1-\gamma) = V_\gamma(y, \lambda)$. In an analogous fashion, we define the optimistic continuation value, for parameters K and λ, to be

$$V_\gamma^K(y, \lambda) \triangleq \sup_{1 \le \tau \le K} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \gamma^\tau \frac{R_{\lambda,K}(\tau, y_{i,\tau-1})}{1-\gamma}\right].$$

From this definition, it follows that the solution for λ to the equation $\lambda/(1-\gamma) = V_\gamma^K(y, \lambda)$ is the OGI.

Throughout this section, we will sometimes discuss the value of the index at some particular time t during the execution of the algorithm, which depends on the statistic gathered about the arm using information up to but strictly *not including* time t. As such, we will define the number of pulls of arm i up to time $t-1$ as

$$P_i(t) \triangleq N_i(t-1),$$

where we recall $N_i(t)$ is the counter for the number of total pulls up to and including t. From the $P_i(t)$ pulls of the

arm, the total reward accumulated is defined as

$$S_i(t) \triangleq \sum_{s=1}^{P_i(t)} X_{i,s}.$$

We begin by investigating the effect of the parameter λ, which gives the deterministic payoff in (3), on the continuation value $V_\gamma^K(y,\lambda)$ and use that to find out how close an approximation $v_\gamma^K(y)$ is to the Gittins index.

Fact 1. For any state $y \in \mathcal{Y}$, discount factor γ and parameter K, the function $V_\gamma^K(y,\lambda)$ is convex in λ. Moreover, the function $V_\gamma^K(y,\lambda)$ is Lipschitz continuous in λ with a Lipshitz constant of $\gamma/(1-\gamma)$.

Proof. Fix an arbitrary state y and discount factor $\gamma \in (0,1)$. Our proof is by induction on the parameter K. For $K = 1$, recall from Section 3 that

$$V_\gamma^1(y,\lambda) = \mathsf{E}_y[X_{i,1}] + \frac{\gamma}{1-\gamma}\mathsf{E}_y[\max(\lambda, R(y_{i,0}))].$$

Thus, the function is convex because it is an expectation over a convex piecewise linear function of random variables $X_{i,1}$ and $R(y_{i,0})$. To prove Lipschitz continuity, it's enough to note that for any $\lambda_1, \lambda_2 \in \mathbb{R}$, that

$$|V_\gamma^1(y,\lambda_1) - V_\gamma^1(y,\lambda_2)| = \frac{\gamma}{1-\gamma}|\mathsf{E}_y[\max(\lambda_1, R(y_{i,0})) - \max(\lambda_2, R(y_{i,0}))]|,$$

$$\le \frac{\gamma}{1-\gamma}|\lambda_1 - \lambda_2|.$$

Now we prove the inductive step. For any $K > 1$, assume that $V_\gamma^{K-1}(y,\lambda)$ is convex and Lipshitz continuous. By writing the Bellman equation,

$$V_\gamma^K(y,\lambda) = \mathsf{E}_y[X_{i,1}] + \gamma\mathsf{E}_y\left[\max\left(\lambda, V_\gamma^{K-1}(y_{i,1},\lambda)\right)\right],$$

we again notice an expectation over a maximum of convex functions in λ. This form for $V_\gamma^K(y,\lambda)$ implies that it is also convex in λ. Finally, to prove Lipschitz continuity, we will use the fact that the pointwise maximum of two Lipschitz functions, having respective constants L_1 and L_2, is also Lipshitz with a constant of $\max(L_1, L_2)$. Because of this, by fixing $\lambda_1, \lambda_2 \in \mathbb{R}$, we have that

$$\begin{aligned}
&|V_\gamma^K(y,\lambda_1) - V_\gamma^K(y,\lambda_2)| \\
&\quad = \gamma\left|\mathsf{E}_y[\max(\lambda_1, V_\gamma^{K-1}(y_{i,1},\lambda_1)) - \max(\lambda_2, V_\gamma^{K-1}(y_{i,1},\lambda_2))]\right|, \\
&\quad \le \frac{\gamma^2}{1-\gamma}|\lambda_1 - \lambda_2|, \\
&\quad \le \frac{\gamma}{1-\gamma}|\lambda_1 - \lambda_2|,
\end{aligned}$$

where the second-to-last inequality follows from the induction hypothesis that $V_\gamma^{K-1}(y_{i,1},\lambda))$ is Lipschitz continuous in λ with a constant of $\gamma/(1-\gamma)$ and also from the fact that the identity function for λ (within the maximum expression) is trivially Lipschitz continuous. □

Lemma C.1. *Suppose that arm rewards are bounded. That is, there exists a constant $B \in \mathbb{R}_+$ such that $X_{i,t} \in [0,B]$ for every arm i and time t. Further assume that $X_{i,t}$ is not almost surely equal to B.*

Let $v_{i,t}^K$ be the OGI of arm i at time t and let η be a scalar, then the following equivalence holds:

$$\{v_{i,t}^K < \eta\} = \{(1-\gamma_t)V_{\gamma_t}^K(y_{i,P_i(t)},\eta) < \eta\},$$

where $y_{i,P_i(t)}$ is the sufficient statistic for estimating the ith arm's parameter θ_i at time t.

Proof. Fix any state y and discount factor γ. At $\lambda = 0$, we have

$$(1-\gamma)V_\gamma^K(y,0) \ge 0$$

because $V_\gamma^K(0,y)$ is the expectation of a sum of non-negative terms. Also, in the other extreme case when $\lambda = B$, the function in question evaluates to

$$V_\gamma^K(y,B) = \mathsf{E}_y[X_{i,1}] + \frac{\gamma B}{(1-\gamma)} < \frac{B}{(1-\gamma)}.$$

Therefore, $(1-\gamma)V_\gamma^K(y,B) < B$.

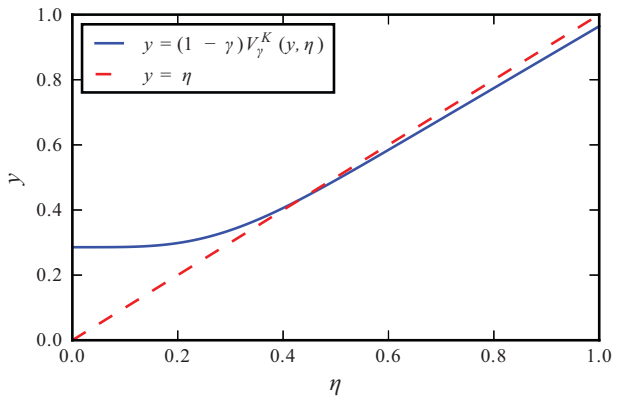
Next we prove that $V_\gamma^K(y,\lambda)$ is monotonically increasing in λ. To show this, pick any $\lambda' < \lambda''$ and let $\tau^*(\lambda')$ and $\tau^*(\lambda'')$ denote the two optimal stopping times under λ', λ'', respectively. It then follows, from $R_{\lambda,\tau}(.)$ being an increasing function of λ on every sample path, that

$$\begin{aligned}
V_\gamma^K(y,\lambda') &= \mathsf{E}_y\left[\sum_{t=1}^{\tau^*(\lambda')} \gamma^{t-1}X_{i,t} + \gamma^{\tau^*(\lambda')}\frac{R_{\lambda',K}(\tau, y_{i,\tau^*(\lambda')-1})}{1-\gamma}\right], \\
&\le \mathsf{E}_y\left[\sum_{t=1}^{\tau^*(\lambda')} \gamma^{t-1}X_{i,t} + \gamma^{\tau^*(\lambda')}\frac{R_{\lambda'',K}(\tau, y_{i,\tau^*(\lambda')-1})}{1-\gamma}\right], \\
&\le \mathsf{E}_y\left[\sum_{t=1}^{\tau^*(\lambda'')} \gamma^{t-1}X_{i,t} + \gamma^{\tau^*(\lambda'')}\frac{R_{\lambda'',K}(\tau, y_{i,\tau^*(\lambda'')-1})}{1-\gamma}\right], \\
&= V_\gamma^K(y,\lambda'').
\end{aligned}$$

Let's put together these observations:

- The inequality $(1-\gamma)V_\gamma^K(y,\lambda) \ge \lambda$ at $\lambda = 0$.
- The inequality $(1-\gamma)V_\gamma^K(y,\lambda) < \lambda$ at $\lambda = B$
- The function $(1-\gamma)V_\gamma^K(y,\lambda)$ is monotonically increasing in λ.

Figure C.1. (Color online) Visualization of Lemma C.1's Proof for an Instance of the Problem with a Beta Prior Corresponding to the Pair $y = (4,5)$, a Discount Factor of $\gamma = 0.95$, and $K = 2$



Note. The intersection of the two lines represents the OGI.

These together with Fact 1 show that the univariate function $(1-\gamma)V_\gamma^K(y,\lambda)-\lambda$ is continuous and decreasing in λ. Moreover, this function is non-negative for any $\lambda \le v_\gamma^K(y)$ (because $v_\gamma^K(y)$ is the root of the function) and is also negative for $\lambda > v_\gamma^K(y)$. This proves the result in question. Figure C.1 also provides a visualization. □

C.1. Proof of Lemma 2

Proof. Let $K < M$ be two look-ahead parameters used in the definition of OGI. We will show that $V_\gamma^K(y,\lambda) \le V_\gamma^M(y,\lambda)$ where we recall the definitions of these functions from the beginning of Appendix C.

We begin with a fundamental step. Let τ_1 and τ_2 be any predictable stopping times (i.e., $\mathcal{F}_{t-1}$-measurable random times) such that τ_1 precedes τ_2 almost surely, that is, $\tau_1 < \tau_2$. Recall that the expected reward of the ith arm satisfies $\mathsf{E}[X_{i,t} \mid \theta_i] = \mu(\theta_i)$ for all t. Let $\hat{\theta}_i \in \Theta$ denote a realization of the random variable θ_i and let $\zeta(\hat{\theta}_i)$ be a real-valued, measurable function of $\hat{\theta}_i$. In this case, we have that

$$\begin{aligned}
&\mathsf{E}\left[\sum_{t=\tau_1+1}^{\tau_2} \gamma^{t-1} X_{i,t} + \gamma^{\tau_2}\frac{\zeta(\hat{\theta}_i)}{1-\gamma}\,\middle|\,\theta_i=\hat{\theta}_i\right]\\
&\quad= \mu(\hat{\theta}_i)\mathsf{E}\left[\sum_{t=\tau_1+1}^{\tau_2}\gamma^{t-1}\,\middle|\,\theta_i=\hat{\theta}_i\right] + \mathsf{E}\left[\frac{\gamma^{\tau_2}}{1-\gamma}\,\middle|\,\theta_i=\hat{\theta}_i\right]\zeta(\hat{\theta}_i),\\
&\quad\le \mathsf{E}\left[\gamma^{\tau_1}\,\middle|\,\theta_i=\hat{\theta}_i\right]\frac{\max(\zeta(\hat{\theta}_i),\mu(\hat{\theta}_i))}{1-\gamma}.
\end{aligned}$$

Thus, we conclude, because $\hat{\theta}_i$ was arbitrary, that almost surely,

$$\mathsf{E}\left[\sum_{t=\tau_1+1}^{\tau_2}\gamma^{t-1}X_{i,t}+\gamma^{\tau_2}\frac{\zeta(\theta_i)}{1-\gamma}\,\middle|\,\theta_i\right] \le \mathsf{E}[\gamma^{\tau_1}\mid\theta_i]\frac{\max(\zeta(\theta_i),\mu(\theta_i))}{1-\gamma}. \tag{C.1}$$

Let $\tau^\star$ be a stopping time that achieves the supremum in $V_\gamma^M(y,\lambda)$ and define the predictable stopping time $\tau_K^\star \triangleq K \wedge \tau^\star$. Consider the (conditional) cumulative rewards in the definition of $V_\gamma^M(y)$, from time $\tau_K^\star+1$ onward, given the sufficient statistic observed at time $\tau_K^\star$. That is,

$$\mathsf{E}\left[\sum_{t=\tau_K^\star+1}^{\tau^\star}\gamma^{t-1}X_{i,t}+\gamma^{\tau^\star}R_{\lambda,M}(\tau^\star,y_{i,\tau^\star-1})/(1-\gamma)\,\middle|\,y_{i,\tau_K^\star-1}\right].$$

We upper bound this random variable as follows. First, we note that, at any time s and for any statistic $\hat{y} \in \mathcal{Y}$, the following statement holds:

$$\mathbb{P}(R(\hat{y}) \le r) = \mathbb{P}(\mu(\theta_i) \le r \mid y_{i,s} = \hat{y}), \qquad \forall\, r \in \mathbb{R}, \tag{C.2}$$

meaning that the posterior distribution of the arm's expected reward $R(y_{i,s})$ is the same as $\mu(\theta_i)$ *conditioned* on having observed statistic $\hat{y}$ about the arm. This holds by definition of the random variable $R(y)$. Because of this observation, we have that the following inequality holds almost surely,

$$\begin{aligned}
&\gamma^{\tau^\star}\frac{R_{\lambda,M}(\tau^\star,y_{i,\tau^\star-1})}{1-\gamma},\\
&\quad= \gamma^{\tau^\star}\left(\mathbb{1}(\tau^\star=M)\frac{\max(\lambda,R(y_{i,\tau^\star-1}))}{1-\gamma}+\mathbb{1}(\tau^\star<M)\frac{\lambda}{1-\gamma}\right),\\
&\quad= \mathbb{1}(\tau^\star=M)\gamma^M\frac{\max(\lambda,R(y_{i,M-1}))}{1-\gamma}+\mathbb{1}(\tau^\star<M)\gamma^{\tau^\star}\frac{\lambda}{1-\gamma},\\
&\quad\overset{(*)}{=} \mathbb{1}(\tau^\star=M)\gamma^M\frac{\mathsf{E}[\max(\lambda,R(y_{i,M-1}))\mid y_{i,M-1}]}{1-\gamma}\\
&\qquad+\mathbb{1}(\tau^\star<M)\gamma^{\tau^\star}\frac{\lambda}{1-\gamma},\\
&\quad\overset{(\dagger)}{=} \mathbb{1}(\tau^\star=M)\gamma^M\frac{\mathsf{E}[\max(\lambda,\mu(\theta_i))\mid y_{i,M-1}]}{1-\gamma}\\
&\qquad+\mathbb{1}(\tau^\star<M)\gamma^{\tau^\star}\frac{\lambda}{1-\gamma},\\
&\quad\overset{(**)}{\le} \frac{\mathsf{E}[\gamma^{\tau^\star}\max(\lambda,\mu(\theta_i))\mid y_{i,\tau^\star-1}]}{1-\gamma},
\end{aligned}$$

where $(*)$ and $(**)$ both use the fact that for any t, $\tau^\star \le t$ is measurable with respect to the σ-algebra generated by $y_{i,t-1}$, namely $\mathcal{F}_{t-1}$. Equation $(\dagger)$ follows from (C.2). Therefore, immediately using the previous inequality and conditioning on the event $\tau^\star > K$, we have that

$$\begin{aligned}
&\mathsf{E}\left[\sum_{t=\tau_K^\star+1}^{\tau^\star}\gamma^{t-1}X_{i,t}+\gamma^{\tau^\star}\frac{R_{\lambda,M}(\tau^\star,y_{i,\tau^\star-1})}{1-\gamma}\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right],\\
&\le \mathsf{E}\left[\sum_{t=\tau_K^\star+1}^{\tau^\star}\gamma^{t-1}X_{i,t}+\mathsf{E}\left[\gamma^{\tau^\star}\frac{\max(\lambda,\mu(\theta_i)))}{1-\gamma}\,\middle|\,y_{i,\tau^\star-1}\right]\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right],\\
&= \mathsf{E}\left[\sum_{t=\tau_K^\star+1}^{\tau^\star}\gamma^{t-1}X_{i,t}+\gamma^{\tau^\star}\frac{\max(\lambda,\mu(\theta_i))}{1-\gamma}\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right], &\text{(C.3)}\\
&= \mathsf{E}\left[\mathsf{E}\left[\sum_{t=K+1}^{\tau^\star}\gamma^{t-1}X_{i,t}+\gamma^{\tau^\star}\frac{\max(\lambda,\mu(\theta_i))}{1-\gamma}\,\middle|\,\theta_i\right]\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right], &\text{(C.4)}\\
&\le \mathsf{E}\left[\gamma^{\tau_K^\star}\frac{\max(\mu(\theta_i),\lambda)}{1-\gamma}\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right], &\text{(C.5)}\\
&= \mathsf{E}\left[\gamma^{\tau_K^\star}\frac{\max(R(y_{i,\tau_K^\star-1}),\lambda)}{1-\gamma}\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right], &\text{(C.6)}\\
&= \mathsf{E}\left[\frac{\gamma^{\tau_K^\star}R_{\lambda,K}(\tau_K^\star,y_{i,\tau_K^\star-1})}{1-\gamma}\,\middle|\,\tau^\star>K,\,y_{i,\tau_K^\star-1}\right], &\text{(C.7)}
\end{aligned}$$

where (C.3) and (C.4) use the tower property, and (C.5) follows from the bound in (C.1) because $\tau_K^\star < \tau^\star$, almost surely. Equation (C.6) follows from Statement (C.2) and that the event $\tau^\star > K$ is $\mathcal{F}_{K-1}$-measurable (we can decide whether to pull arm i or retire based on information up to and including time $K-1$). Finally Equation (C.7) is derived by substituting in the definition of $R_{\lambda,K}$ (as given in Section 3) and noting that $\tau_K^\star = K$ under the previous conditioning.

We now condition on the complement of the previous event we considered, namely, $\tau^\star \le K$. Under that event, $\tau^\star$

occurred early enough before time $K+1$ and thus $\tau_K^\star = \tau^\star$. Therefore, it follows from this observation that

$$\begin{aligned}
&\mathsf{E}\left[\sum_{t=\tau_K^\star+1}^{\tau^\star} \gamma^{t-1} X_{i,t} + \gamma^{\tau^\star} \frac{R_{\lambda,M}(\tau^\star, y_{i,\tau^\star-1})}{1-\gamma} \,\middle|\, \tau^\star \le K, y_{i,\tau_K^\star-1}\right] \\
&= \mathsf{E}\left[\gamma^{\tau^\star} \frac{\lambda}{1-\gamma} \,\middle|\, \tau^\star \le K, y_{i,\tau_K^\star-1}\right] \\
&\le \mathsf{E}\left[\gamma^{\tau_K^\star} \frac{R_{\lambda,K}(\tau_K^\star, y_{i,\tau_K^\star-1})}{1-\gamma} \,\middle|\, \tau^\star \le K, y_{i,\tau_K^\star-1}\right], \qquad \text{(C.8)}
\end{aligned}$$

where (C.8) is obtained by noting that $R_{\lambda,K}(\tau, y) \ge \lambda$ for any choice of τ, K and y. Thus, by the law of total expectation and (C.7) and (C.8), we establish that

$$\begin{aligned}
&\mathsf{E}\left[\sum_{t=\tau_K^\star+1}^{\tau^\star} \gamma^{t-1} X_{i,t} + \gamma^{\tau^\star} \frac{R_{\lambda,M}(\tau^\star, y_{i,\tau^\star-1})}{1-\gamma} \,\middle|\, y_{i,\tau_K^\star-1}\right] \\
&\le \mathsf{E}\left[\gamma^{\tau_K^\star} \frac{R_{\lambda,K}(\tau_K^\star, y_{i,\tau_K^\star-1})}{1-\gamma} \,\middle|\, y_{i,\tau_K^\star-1}\right]. \qquad \text{(C.9)}
\end{aligned}$$

We are ready to complete our main argument in this proof by using the previous bound and breaking up the $V_\gamma^M(y,\lambda)$ into rewards from times before $\tau_K^\star$ and after (and bounding the latter terms). More precisely, we obtain that

$$\begin{aligned}
V_\gamma^M(y,\lambda) &= \mathsf{E}_y\left[\sum_{t=1}^{\tau^\star} \gamma^{t-1} X_{i,t} + \gamma^{\tau^\star} \frac{R_{\lambda,M}(\tau^\star, y_{i,\tau^\star-1})}{1-\gamma}\right], \qquad \text{(C.10)} \\
&= \mathsf{E}_y\left[\sum_{t=1}^{\tau_K^\star} \gamma^{t-1} X_{i,t} + \sum_{t'=\tau_K^\star+1}^{\tau^\star} \gamma^{t'-1} X_{i,t'} + \gamma^{\tau^\star} \frac{R_{\lambda,M}(\tau^\star, y_{i,\tau^\star-1})}{1-\gamma}\right], \\
&= \mathsf{E}_y\left[\sum_{t=1}^{\tau_K^\star} \gamma^{t-1} X_{i,t} + \mathsf{E}\left[\sum_{t'=\tau_K^\star+1}^{\tau^\star} \gamma^{t'-1} X_{i,t'} + \gamma^{\tau^\star} \frac{R_{\lambda,M}(\tau^\star, y_{i,\tau^\star-1})}{1-\gamma} \,\middle|\, y_{i,\tau_K^\star-1}\right]\right], \qquad \text{(C.11)} \\
&\le \mathsf{E}_y\left[\sum_{t=1}^{\tau_K^\star} \gamma^{t-1} X_{i,t} + \mathsf{E}\left[\gamma^{\tau_K^\star} \frac{R_{\lambda,K}(\tau_K^\star, y_{i,\tau_K^\star-1})}{1-\gamma} \,\middle|\, y_{i,\tau_K^\star-1}\right]\right], \qquad \text{(C.12)} \\
&= \mathsf{E}_y\left[\sum_{t=1}^{\tau_K^\star} \gamma^{t-1} X_{i,t} + \gamma^{\tau_K^\star} \frac{R_{\lambda,K}(\tau_K^\star, y_{i,\tau_K^\star-1})}{1-\gamma}\right], \qquad \text{(C.13)} \\
&\le \sup_{1\le\tau\le K} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \gamma^{\tau} \frac{R_{\lambda,K}(\tau, y_{i,\tau-1})}{1-\gamma}\right], \\
&= V_\gamma^K(y,\lambda), \qquad \text{(C.14)}
\end{aligned}$$

where Equations (C.11) and (C.13) use the tower property, and (C.12) is immediately derived by using the bound of (C.9). Finally, an almost identical proof can be given to show that $V_\gamma^K(y,\lambda) \ge V_\gamma(y,\lambda)$ where the lower bound is the continuation value used to compute the Gittins index.

We have shown that for any λ and y, that $V_\gamma^K(y,\lambda)$ is nonincreasing in K and that $V_\gamma(y,\lambda)$ is a lower bound to this sequence. We make use of these facts to now prove that $v_\gamma^K(y)$ is also nonincreasing in K. To this end, let us suppose for contradiction that there exist two integers $K_1 \le K_2$ and $v_\gamma^{K_1}(y) < v_\gamma^{K_2}(y)$. From Lemma C.1 we know that

$$V_\gamma^{K_2}(y, v_\gamma^K(y)) > v_\gamma^K(y)/(1-\gamma) = V_\gamma^K(y, v_\gamma^K(y)), \qquad \text{(C.15)}$$

which contradicts the claim just shown. Therefore, $v_\gamma^K(y)$ must also be a nonincreasing sequence in K. The same argument can be used to further show that $v_\gamma^K(y) \ge v_\gamma(y)$.

We now turn our attention to proving the convergence property stated in the lemma. The first step will be to prove that for all $y \in \mathcal{Y}$ and $\lambda \in \mathbb{R}_+$, that

$$\lim_{K\to\infty} V_\gamma^K(y,\lambda) = V_\gamma(y,\lambda). \qquad \text{(C.16)}$$

Indeed, we upper bound the optimistic continuation value for a fixed parameter M as follows:

$$\begin{aligned}
&V_\gamma^M(y,\lambda) \\
&= \sup_{1\le\tau\le M} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \frac{\gamma^\tau R_{\lambda,M}(\tau, y_{i,\tau-1})}{1-\gamma}\right] \\
&= \sup_{1\le\tau\le M} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \frac{\gamma^\tau \lambda}{1-\gamma} + \frac{\gamma^\tau R_{\lambda,M}(\tau, y_{i,\tau-1})}{1-\gamma} - \frac{\gamma^\tau \lambda}{1-\gamma}\right] \\
&\le \sup_{\tau\ge 1} \mathsf{E}_y\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \frac{\gamma^\tau \lambda}{1-\gamma}\right] \\
&\quad + \sup_{1\le\tau\le M} \mathsf{E}_y\left[\frac{\gamma^\tau R_{\lambda,M}(\tau, y_{i,\tau-1})}{1-\gamma} - \frac{\gamma^{\tau-1}\lambda}{1-\gamma}\right] \\
&= V_\gamma(y,\lambda) + \sup_{1\le\tau\le M} \mathsf{E}_y\left[\frac{\gamma^\tau [R_{\lambda,M}(\tau, y_{i,\tau-1}) - \lambda]}{1-\gamma}\right] \\
&\le V_\gamma(y,\lambda) + \gamma^M \mathsf{E}_y\left[\frac{R_{\lambda,M}(M, y_{i,M-1}) - \lambda}{1-\gamma}\right] \\
&= V_\gamma(y,\lambda) + \gamma^M \mathsf{E}_y\left[\frac{(R(y_{i,M-1}) - \lambda)^+}{1-\gamma}\right] \\
&\le V_\gamma(y,\lambda) + \gamma^M \mathsf{E}_y\left[\frac{|R(y_{i,M-1})|}{1-\gamma}\right] \\
&= V_\gamma(y,\lambda) + \gamma^M \mathsf{E}_y\left[\frac{|\mu(\theta_i)|}{1-\gamma}\right], \qquad \text{(C.17)}
\end{aligned}$$

where Equation (C.17) follows from the definition of the random variable $R(.)$ and the law of iterated expectation. Now because $0 < \gamma < 1$ and $\mathsf{E}_y\left[|\mu(\theta_i)|\right] < \infty$, the right-hand side above converges to $V_\gamma(y,\lambda)$. Finally, notice that $V_\gamma^M(y,\lambda) \ge V_\gamma(y,\lambda)$, and from this, Equation (C.16) follows.

To finish the proof, we consider the sequence of fixed points of the equations $\lambda = V_\gamma^K(y,\lambda)$, $\{v_\gamma^K(y)\}$. Because this sequence is monotone (established in the first part of this proof) and bounded, we know that this sequence, has a limit; $v_\gamma^K(y) \to \hat{v}_\gamma(y)$.

It remains to show that $\hat{v}_\gamma(y) = V_\gamma(y, \hat{v}_\gamma(y))$. For this, it suffices to show that $v_\gamma^K(y) \to V_\gamma(y, \hat{v}_\gamma(y))$, which we establish as follows:

$$\begin{aligned}
|v_\gamma^K(y) - V_\gamma(y, \hat{v}_\gamma(y))| &= |V_\gamma^K(y, v_\gamma^K(y)) - V_\gamma(y, \hat{v}_\gamma(y))|, \\
&\le \underbrace{|V_\gamma^K(y, v_\gamma^K(y)) - V_\gamma^K(y, \hat{v}_\gamma(y))|}_{=:a_k} \\
&\quad + \underbrace{|V_\gamma^K(y, \hat{v}_\gamma(y)) - V_\gamma(y, \hat{v}_\gamma(y))|}_{=:b_k}.
\end{aligned}$$

We already proved (C.16) and therefore know that $b_k \to 0$ as $k \to \infty$. As for the a_k sequence, we have

$$\begin{aligned}
a_k &= |V_\gamma^K(y, v_\gamma^K(y)) - V_\gamma^K(y, \hat{v}_\gamma(y))| \\
&\le |v_\gamma^K(y) - \hat{v}_\gamma(y)| \\
&\to 0, \qquad \text{(C.18)}
\end{aligned}$$

where (C.18) follows from the Lipschitz continuity of $V_\gamma^K(\cdot,\cdot)$ in its second argument as shown in Fact 1. Therefore, $v_\gamma^K(y) \to V_\gamma(y, \hat{v}_\gamma(y))$, which completes the proof. □

The next lemma will be the final property of the function V_γ^K that we prove. This will subsequently be used in the proof of Lemma 3.

Lemma C.2. *Let i be any arm. For any look-ahead parameter $K \in \mathbb{Z}_+$, discount factor γ and any constant η, we have*

$$\mathsf{E}_y\left[V_\gamma^K(y_{i,1},\eta)\right] \geq V_\gamma^K(y,\eta),$$

where we recall that $y_{i,1}$ is the summary statistic corresponding to the posterior obtained from pulling arm i once.

Proof. For any $\hat{y} \in \mathcal{Y}$, let $\tau^\star(\hat{y})$ be the (predictable) optimal stopping time for the problem (involving computing V_γ^K) whose initial state is $y_{i,0} = \hat{y}$. With this notation in hand, we conclude that

$$\mathsf{E}_y\left[V_\gamma^K(y_{i,1},\eta)\right] = \mathsf{E}_y\left[\mathsf{E}_{y_{i,1}}\left[\sum_{s=1}^{\tau^\star(y_{i,1})}\gamma^{s-1}X_{i,s} + \frac{\gamma^{\tau^\star(y_{i,1})}R_{\eta,K}(\tau, y_{i,\tau^\star(y_{i,1})-1})}{1-\gamma}\right]\right], \tag{C.19}$$

$$\geq \mathsf{E}_y\left[\mathsf{E}_{y_{i,2}}\left[\sum_{s=1}^{\tau^\star(y)}\gamma^{s-1}X_{i,s} + \frac{\gamma^{\tau^\star(y)}R_{\eta,K}(\tau, y_{i,\tau^\star(y)-1})}{1-\gamma}\right]\right], \tag{C.20}$$

$$= \mathsf{E}_y\left[\sum_{s=1}^{\tau^\star(y)}\gamma^{s-1}X_{i,s} + \frac{\gamma^{\tau^\star(y)}R_{\eta,K}(\tau, y_{i,\tau^\star(y)-1})}{1-\gamma}\right], \tag{C.21}$$

$$= V_\gamma^K(y,\eta),$$

where (C.19) and (C.21) both follow from the tower property, and (C.20) is because of the suboptimality of the stopping rule $\tau^\star(y)$ when the actual starting state is $y_{i,1}$. Intuitively, we lose out revenue by throwing away information about the arm. □

Appendix D. Results for the Frequentist Regret Bound

This section contains proofs of results required to show Theorem 1. It is helpful to go over the definitions and some general properties of the OGI given in Appendix C when reading this.

D.1. Definitions and Properties of Binomial Distributions

We list notation and facts related to Beta and Binomial distributions, which are used through this section.

Definition D.1. The function $F_{n,p}^B(.)$ is the CDF of the Binomial distribution with parameters n and p, and $F_{a,b}^\beta(.)$ is the CDF of the Beta distribution with parameters a and b.

Lemma D.1. *Let a and b be positive integers and $y \in [0,1]$,*

$$F_{a,b}^\beta(y) = 1 - F_{a+b-1,y}^B(a-1).$$

Proof. The proof is found in Agrawal and Goyal (2012). □

Lemma D.2. *The median of a Binomial(n, p) distribution is either $\lceil np \rceil$ or $\lfloor np \rfloor$.*

Proof. A proof of this fact can be found in Jogdeo and Samuels (1968). □

Corollary D.1 (Corollary of Fact 10). *Let n be a positive integer and $p \in (0,1)$. For any non-negative integer $s < np$,*

$$F_{n,p}^B(s) \leq 1/2.$$

Lemma D.3. *Let n be a positive integer and $p \in [0,1]$. Then for any $k \in \{0,\ldots,n\}$,*

$$(1-p)F_{n-1,p}^B(k) \leq F_{n,p}^B(k) \leq F_{n-1,p}^B(k).$$

Proof. To prove $F_{n,p}^B(k) \leq F_{n-1,p}^B(k)$, we let $X_1,\ldots,X_n$ be i.i.d. samples from a Bernoulli(p) distribution. We then have

$$F_{n,p}^B(k) = \mathbb{P}\left(\sum_{i=1}^n X_i \leq k\right) \leq \mathbb{P}\left(\sum_{i=1}^{n-1} X_i \leq k\right) = F_{n-1,p}^B(k).$$

Now to prove $(1-p)F_{n-1,p}^B(k) \leq F_{n,p}^B(k)$, it is enough to observe that $F_{n,p}^B(k) = pF_{n-1,p}^B(k-1) + (1-p)F_{n-1,p}^B(k)$. □

D.1.1. Ratio of Binomial CDFs.

Lemma D.4. *Let $0 < q < p < 1$. Let n be a positive integer such that $e^{\frac{n}{2}d(q,p)} \geq (n+1)^4$ and let k be a nonnegative integer such that $k < nq$. It then follows that*

$$F_{n,q}^B(k)/F_{n,p}^B(k) > e^{\frac{n}{2}d(q,p)}.$$

Proof. From the method of types (Cover and Thomas 2012), we have for any $r \in (0,1)$ and $j < nr$,

$$\frac{e^{-nd(j/n,r)}}{(1+n)^2} \leq F_{n,r}^B(j) \leq (n+1)^2 e^{-nd(j/n,r)}. \tag{D.1}$$

Because $k < nq < np$, by applying (D.1) to both the numerator and denominator, we get

$$\frac{F_{n,q}^B(k)}{F_{n,p}^B(k)} \geq \frac{e^{-nd(k/n,q)}}{(n+1)^4 e^{-nd(k/n,p)}} = \frac{e^{n(d(k/n,p)-d(k/n,q))}}{(n+1)^4}.$$

Examining the exponent, we find

$$\begin{aligned} d(k/n,p) - d(k/n,q) &= \frac{k}{n}\log\frac{q}{p} + \left(1-\frac{k}{n}\right)\log\frac{1-q}{1-p}, \\ &> q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}, \\ &= d(q,p), \end{aligned}$$

where the bound holds because the expression is decreasing in k, and $k < nq$. Therefore,

$$\frac{F_{n,q}^B(k)}{F_{n,p}^B(k)} > \frac{e^{nd(q,p)}}{(n+1)^4} = \frac{e^{\frac{n}{2}d(q,p)}}{(n+1)^4} e^{\frac{n}{2}d(q,p)} \geq e^{\frac{n}{2}d(q,p)}. \tag{D.2}$$

The final lower bound in (D.2) follows from the assumption on n. □

D.2. Proof of Lemma 3

Proof. The proof hinges on showing that for any K, which is the number of look-ahead steps used to compute the OGI, that

$$\mathbb{P}\left(v^K_{1,t} < \eta\right) = O\left(\frac{1}{t^{1+h_\eta}}\right), \tag{D.3}$$

where $h_\eta > 0$ is some constant that depends on η. After showing the previous statement, the result would follow because of convergence of the series $\sum_{t=1}^{\infty}\mathbb{P}(v^K_{1,t} < \eta)$. The first step will be to show that for any $K \geq 1$ and any $\zeta \geq 0$ that there exists $h'_\eta > 0$, such that

$$\mathbb{P}\left((1-\gamma_t)V^K_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \zeta/t\right) = O_{\eta,\zeta}\left(\frac{1}{t^{1+h'_\eta}}\right), \tag{D.4}$$

where $V^K_{\gamma_t}$ is the continuation value defined in Appendix C, and $O_{\eta,\zeta}$ means that the constant in front the big-Oh depends on both ζ and η. After showing the previous claim, Lemma C.1 would imply Equation (D.3) because we know from that result that

$$\begin{aligned}\mathbb{P}\left(v^K_{1,t} < \eta\right) &= \mathbb{P}\left((1-\gamma_t)V^K_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta\right),\\ &= O\left(\frac{1}{t^{1+h_\eta}}\right),\end{aligned}$$

for some $h_\eta > 0$. The second equation is just a special case of (D.4) when $\zeta = 0$.

Ultimately, showing Equation (D.4), and thus proving the lemma, is an induction over the parameter K, and we begin with the base case, which requires some work using properties of the Beta and Binomial distributions.

Proof of the Base Case. Let us fix $\zeta \geq 0$. We prove that when the algorithm uses a look-ahead parameter of $K = 1$, that there exists a positive constant h_η such that

$$\mathbb{P}\left((1-\gamma_t)V^1_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \zeta/t\right) = O_{\eta,\zeta}\left(\frac{1}{t^{1+h_\eta}}\right). \tag{D.5}$$

First, we define $\delta := (\theta_1 - \eta)/2$ and $\eta' := \eta + \delta$. In other words, δ is half the distance between η and θ_1; η' is the point half-way. Recall that $P_i(t)$ refers to the counting process for the number of pulls of arm i up to *but not including* time t and that $S_i(t)$ is the corresponding total reward (or number of successes from all the Bernoulli trials). Showing this base case consists of showing two claims:

Claim 1: $\left\{(1-\gamma_t)V^1_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \zeta/t\right\} \subseteq \left\{F^B_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t}\right\}$.

Let $V_t \sim \text{Beta}(S_1(t)+1, P_1(t) - S_1(t) + 1)$ be the agent's posterior on the expected reward from the optimal arm (notice that $y_{1,P_1(t)} = (S_1(t)+1, P_1(t) - S_1(t) + 1)$ in this case). Using the simplified equation for the continuation value when $K = 1$, namely $V^1_{\gamma_t}$ (see Equation (4)),

$$\begin{aligned}&(1-\gamma_t)V^1_{\gamma_t}((S_1(t)+1, P_1(t) - S_1(t) + 1),\eta)\\ &\quad = \mathsf{E}[V_t] + \gamma_t \mathsf{E}\left[(\eta - V_t)^+\right],\end{aligned}$$

we find that

$$\begin{aligned}&\left\{(1-\gamma_t)V^1_{\gamma_t}(y_{1,N_1(t)},\eta) < \eta + \frac{\zeta}{t}\right\},\\ &\quad = \left\{\mathsf{E}[V_t] + \gamma_t \mathsf{E}\left[(\eta - V_t)^+\right] < \eta + \frac{\zeta}{t}\right\},\\ &\quad = \left\{(1-1/t)\mathsf{E}\left[(\eta - V_t)^+\right] < \mathsf{E}[\eta - V_t] + \frac{\zeta}{t}\right\}, \qquad \text{(D.6)}\\ &\quad = \left\{\mathsf{E}\left[(\eta - V_t)^+\right] - \mathsf{E}[\eta - V_t] < \frac{1}{t}\mathsf{E}\left[(\eta - V_t)^+\right] + \frac{\zeta}{t}\right\},\\ &\quad = \left\{\mathsf{E}\left[(V_t - \eta)^+\right] < \frac{1}{t}\mathsf{E}\left[(\eta - V_t)^+\right] + \frac{\zeta}{t}\right\},\\ &\quad \subseteq \left\{\mathsf{E}\left[(V_t - \eta)^+\right] < \frac{\zeta+1}{t}\right\}, \qquad \text{(D.7)}\end{aligned}$$

where (D.6) follows from the definition of γ_t, and (D.7) is because of V_t, η both lying in the interval $[0,1]$. We approximate the conditional expectation in (D.7) with the following bound:

$$\begin{aligned}\mathsf{E}\left[(V_t - \eta)^+\right] &= \mathsf{E}\left[(V_t - \eta)\mathbb{1}(V_t \geq \eta)\right]\\ &= \mathsf{E}\left[(V_t - \eta)\mathbb{1}(\eta + \delta > V_t \geq \eta)\right]\\ &\quad + \mathsf{E}\left[(V_t - \eta)\mathbb{1}(V_t \geq \eta + \delta)\right]\\ &> \mathsf{E}\left[(V_t - \eta)\mathbb{1}(V_t \geq \eta + \delta)\right]\\ &\geq \delta\mathbb{P}(V_t \geq \eta')\\ &= \delta(1 - F^\beta_{S_1(t)+1,P_1(t)-S_1(t)+1}(\eta'))\\ &= \delta F^B_{P_1(t)+1,\eta'}(S_1(t)), \qquad \text{(D.8)}\end{aligned}$$

where the final equality is because of Fact 9. The claim then follows from the above bound and (D.7). We proceed with the second part of the base case's proof.

Claim 2: $\mathbb{P}\left(F^B_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t}\right) = O\left(\frac{1}{t^{1+h_\eta}}\right)$ for some $h_\eta > 0$.

Let us fix the sequence $f_t \triangleq -\frac{\log(\delta t/(\zeta+1))}{\log(1-\eta')} - 1 = O(\log t)$. We then have by a straightforward decomposition that

$$\begin{aligned}&\mathbb{P}\left(F^B_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t}\right)\\ &= \mathbb{P}\left(F^B_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t}, P_1(t) > f_t\right)\\ &\quad + \mathbb{P}\left(F^B_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t}, P_1(t) \leq f_t\right). \qquad \text{(D.9)}\end{aligned}$$

Then notice that for the second term in the right-hand side of (D.9), we have the following bound:

$$\begin{aligned}&\mathbb{P}\left(F^B_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t}, P_1(t) \leq f_t\right)\\ &\leq \mathbb{P}\left(F^B_{P_1(t)+1,\eta'}(0) < \frac{\zeta+1}{\delta t}, P_1(t) \leq f_t\right)\\ &= \mathbb{P}\left((1-\eta')^{P_1(t)+1} < \frac{\zeta+1}{\delta t}, P_1(t) \leq f_t\right)\\ &\leq \mathbb{P}\left((1-\eta')^{f_t+1} < \frac{\zeta+1}{\delta t}\right)\\ &= 0. \qquad \text{(D.10)}\end{aligned}$$

Now we use the following fact to correspondingly bound the left term on the right-hand side of (D.9). Define the function

$$F^{-B}_{n,p}(u) := \inf\{x : F^{B}_{n,p}(x) \geq u\},$$

which is the inverse CDF. Then it is known that if $U \sim \text{Uniform}(0,1)$, then $F^{-B}_{n,p}(U) \sim \text{Binomial}(n,p)$. Furthermore, the event $F^{B}_{n,p}(F^{-B}_{n,p}(U)) \geq U$ occurs with a probability of one because of the definition of the inverse CDF.

Now let us only consider large t, in particular $t > M = M(\theta_1, \eta')$, where

1. The constant M is such that $e^{d(\eta',\theta_1)f_M/2} > (f_M+1)^4$ (we need this condition when we use Lemma D.4).
2. We have that $M > \frac{4(\zeta+1)}{(1-\eta')\delta}$.
3. We have that $\lceil f_M \rceil > 0$ and $F^{B}_{t',\eta'}(t'\eta') > 1/4$ for all $t' > \lceil f_M \rceil$. There is a large enough integer for this because $F^{B}_{\lceil f_t \rceil,\eta'}(f_t\eta') \to \frac{1}{2}$ as $t \to \infty$.

Suppose that $t > M$. It then follows that the event that

$$\left\{F^{B}_{P_1(t),\eta'}(S_1(t)) < \frac{\zeta+1}{(1-\eta')\delta t},\ S_1(t) \geq P_1(t)\eta',\ P_1(t) > f_t\right\}$$

has measure zero because of the assumptions made on M. Therefore, if $t > M$, we have

$$\begin{aligned}
&\mathbb{P}\left(F^{B}_{P_1(t)+1,\eta'}(S_1(t)) < \frac{\zeta+1}{\delta t},\ P_1(t) > f_t\right),\\
&\leq \mathbb{P}\left(F^{B}_{P_1(t),\eta'}(S_1(t)) < \frac{\zeta+1}{(1-\eta')\delta t},\ P_1(t) > f_t\right), \qquad \text{(D.11)}\\
&= \mathbb{P}\left(F^{B}_{P_1(t),\eta'}(S_1(t)) < \frac{\zeta+1}{(1-\eta')\delta t},\ S_1(t) < P_1(t)\eta',\ P_1(t) > f_t\right),\\
&= \mathbb{P}\left(F^{B}_{P_1(t),\theta_1}(S_1(t))\frac{F^{B}_{P_1(t),\eta'}(S_1(t))}{F^{B}_{P_1(t),\theta_1}(S_1(t))} < \frac{\zeta+1}{(1-\eta')\delta t},\right.\\
&\qquad \left. S_1(t) < P_1(t)\eta',\ P_1(t) > f_t\right),\\
&\leq \mathbb{P}\left(F^{B}_{P_1(t),\theta_1}(S_1(t))e^{P_1(t)D} < \frac{\zeta+1}{(1-\eta')\delta t},\ P_1(t) > f_t\right),\\
&\leq \mathbb{P}\left(F^{B}_{P_1(t),\theta_1}(S_1(t))e^{f_tD} < \frac{\zeta+1}{(1-\eta')\delta t}\right), \qquad \text{(D.12)}\\
&= \mathbb{P}\left(F^{B}_{P_1(t),\theta_1}(F^{-B}_{P_1(t),\theta_1}(U)) < \frac{e^{-f_tD}(\zeta+1)}{(1-\eta')\delta t}\right), \qquad \text{(D.13)}\\
&\leq \mathbb{P}\left(U < \frac{e^{-f_tD}(\zeta+1)}{(1-\eta')\delta t}\right),\\
&= \frac{e^{-f_tD}(\zeta+1)}{(1-\eta')\delta t},\\
&= \mathcal{O}_{\eta,\zeta}\left(\frac{1}{t^{1+Dc_{\eta'}}}\right), \qquad \text{(D.14)}
\end{aligned}$$

where $D = d(\eta', \theta_1) > 0$ and $c_{\eta'} = -\log^{-1}(1-\eta') > 0$ are constant. The bound (D.11) holds because of Fact 11. Bound (D.12) follows from an application of Lemma D.4 and the fact that $t > M$. Equation (D.13) follows from $S_1(t) \sim \text{Binomial}(P_1(t), \theta_1)$ and the inverse sampling technique. By combining bounds (D.14), (D.10), and (D.9), we finally obtain the result for the base case by taking $h_\eta = Dc_{\eta'}$.

Proof of the Inductive Step. Now, suppose that for $K-1 \geq 1$ and any $\zeta \geq 0$, the following induction hypothesis holds

$$\mathbb{P}\left((1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \frac{\zeta}{t}\right) = O_{\eta,\zeta}\left(\frac{1}{t^{1+h_\eta}}\right)$$

for some $h_\eta > 0$. We prove the same result for the next integer K. Observe that when t is large enough, using the Bellman equation for V^K_γ, we have

$$\begin{aligned}
&\mathbb{P}\left((1-\gamma_t)V^{K}_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \frac{\zeta}{t}\right),\\
&= \mathbb{P}\left((1-\gamma_t)\mathsf{E}[X_{1,t} \mid y_{1,P_1(t)}]\right.,\\
&\qquad \left. + \gamma_t\mathsf{E}\left[\max(\eta, (1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)+1},\eta)) \mid y_{1,P_1(t)}\right] < \eta + \frac{\zeta}{t}\right), \qquad \text{(D.15)}\\
&\leq \mathbb{P}\left(\left(1-\frac{1}{t}\right)\mathsf{E}\left[(1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)+1},\eta) \mid y_{1,P_1(t)}\right] < \eta + \frac{\zeta}{t}\right),\\
&\leq \mathbb{P}\left(\left(1-\frac{1}{t}\right)(1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \frac{\zeta}{t}\right), \qquad \text{(D.16)}\\
&\leq \mathbb{P}\left((1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)},\eta) < \frac{t}{t-1}\left(\eta + \frac{\zeta}{t}\right)\right),\\
&\leq \mathbb{P}\left((1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \frac{\eta}{t-1} + \frac{\zeta}{t-1}\right),\\
&\leq \mathbb{P}\left((1-\gamma_t)V^{K-1}_{\gamma_t}(y_{1,P_1(t)},\eta) < \eta + \frac{\zeta+1}{t}\right),\\
&= O_{\eta,\zeta}\left(\frac{1}{t^{1+h_\eta}}\right), \qquad \text{(D.17)}
\end{aligned}$$

where the final inequality holds when t is large enough because $\eta < 1$, Equation (D.15) results from an expansion of Bellman's equation, and Bound (D.16) follows from Lemma C.2. Finally, Equation (D.17) follows from the induction hypothesis. □

D.3. Proof of Lemma 4

Proof. See the main proof of Theorem 1 to recall the definition of constants η_1, η_3, and their relationship with θ_2 and θ_1. As an abbreviation, we let $L = L(T)$. Moreover, because for any arm i $v^K_{i,t} \leq v^{K-1}_{i,t} \leq \ldots \leq v^1_{i,t}$ (Lemma 2), it will be sufficient to consider this proof only for $v^1_{2,t}$, which we also will abbreviate as $v_{2,t} \triangleq v^1_{2,t}$. Similarly, we will abbreviate the notation for the OGI policy as π^{OG} and suppress the parameter K.

First, by the law of total probability and the definition of $P_i(t)$ in Appendix C, we find that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(v_{2,t}\ge\eta_3, N_2(t-1)\ge L, \pi_t^{\mathrm{OG}}=2)\\
&\quad=\sum_{t=1}^{T}\mathbb{P}\Big(v_{2,t}\ge\eta_3, P_2(t)\ge L, S_2(t)<\lfloor P_2(t)\eta_1\rfloor, \pi_t^{\mathrm{OG}}=2\Big)\\
&\qquad+\sum_{t=1}^{T}\mathbb{P}\Big(v_{2,t}\ge\eta_3, P_2(t)\ge L, S_2(t)\ge\lfloor P_2(t)\eta_1\rfloor, \pi_t^{\mathrm{OG}}=2\Big)\\
&\quad\le\sum_{t=1}^{T}\mathbb{P}(v_{2,t}\ge\eta_3, P_2(t)\ge L, S_2(t)<\lfloor P_2(t)\eta_1\rfloor)\\
&\qquad+\sum_{t=1}^{T}\mathbb{P}\Big(\pi_t^{\mathrm{OG}}=2, S_2(t)\ge\lfloor P_2(t)\eta_1\rfloor\Big), \qquad \text{(D.18)}
\end{aligned}$$

where $S_2(t)$ is also defined in Appendix C as the total reward from the second arm observed up to time $t-1$. Let $V_t\sim\mathrm{Beta}(S_2(t)+1, P_2(t)-S_2(t)+1)$ denote the agent's posterior on the second arm at time t, then

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(v_{2,t}\ge\eta_3,\ P_2(t)\ge L, S_2(t)<\lfloor P_2(t)\eta_1\rfloor)\\
&=\sum_{t=1}^{T}\mathbb{P}\Big(\mathsf{E}[V_t]+\gamma_t\mathsf{E}\big[(\eta_3-V_t)^+\big]\ge\eta_3, P_2(t)\ge L, S_2(t)<\lfloor P_2(t)\eta_1\rfloor\Big)\\
&=\sum_{t=1}^{T}\mathbb{P}\left(\frac{\mathsf{E}\big[(\eta_3-V_t)^+\big]}{\mathsf{E}\big[(V_t-\eta_3)^+\big]}\le t, P_2(t)\ge L, S_2(t)<\lfloor P_2(t)\eta_1\rfloor\right), \qquad \text{(D.19)}
\end{aligned}$$

where the first equality follows from Lemma C.1 and the simplified form of the continuation value (defined in Appendix C) when $K=1$. The following result lets us bound (D.19).

Lemma D.5. *Let $0<x<y<1$. For any nonnegative integers s and k with $s<\lfloor kx\rfloor$, it holds that*

$$\frac{\mathsf{E}\big[(y-V)^+\big]}{\mathsf{E}\big[(V-y)^+\big]}\ge\frac{(y-x)\exp(kd(x,y))}{2},$$

where $V\sim\mathrm{Beta}(s+1,k-s+1)$.

Proof. See Appendix D.3.1. □

Therefore, from Equation (D.19) and Lemma D.5, we find that whenever $T>\left(\frac{2}{\eta_3-\eta_1}\right)^{1/\epsilon}=:T^*(\epsilon,\boldsymbol{\theta})$,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(v_{2,t}\ge\eta_3,\ P_2(t)\ge L, S_2(t)<\lfloor P_2(t)\eta_1\rfloor)\\
&\le\sum_{t=1}^{T}\mathbb{P}((\eta_3-\eta_1)\exp\{P_2(t)d(\eta_1,\eta_3)\}\le 2t, P_2(t)\ge L)\\
&\le\sum_{t=1}^{T}\mathbb{P}((\eta_3-\eta_1)\exp\{Ld(\eta_1,\eta_3)\}\le 2t)\\
&=\sum_{t=1}^{T}\mathbb{P}\Big((\eta_3-\eta_1)T^{1+\epsilon}\le 2t\Big)\\
&=0. \qquad \text{(D.20)}
\end{aligned}$$

All that is left is to bound the second term in (D.18), and to do so, we apply the following lemma whose proof is in Appendix D.3.2.

Lemma D.6. *There exist positive constants $C=C(\theta_2,\eta_1)$ and $x'>\theta_2$ such that*

$$\sum_{t=1}^{T}\mathbb{P}\Big(S_2(t)\ge\lfloor P_2(t)\eta_1\rfloor, \pi_t^{\mathrm{OG}}=2\Big)\le K+\frac{1}{1-e^{-d(x',\theta_2)}}.$$

Combining Lemma D.6, (D.20), (D.18), and (D.19) shows the claim. □

D.3.1. Proof of Lemma D.5.

Proof. We upper bound the denominator as follows. Given that $s<\lfloor kx\rfloor$, we have $s\le kx-1$. Let $B(a,b)$ denote the Beta function for parameters $a,b>0$, that is

$$B(a,b)\triangleq\int_0^1 t^{a-1}(1-t)^{b-1}\,dt,$$

which is used in the definition of the Beta CDF. We can derive an upper bound on the denominator in the following way. Namely, we have

$$\begin{aligned}
\mathsf{E}\big[(V-y)^+\big]&=\frac{1}{B(s+1,k-s+1)}\int_y^1(t-y)t^s(1-t)^{k-s}\,dt,\\
&=\frac{1}{B(s+1,k-s+1)}\int_y^1 t^{s+1}(1-t)^{k-s}\,dt-y\mathbb{P}(V\ge y),\\
&=\frac{B(s+2,k-s+1)}{B(s+1,j-s+1)}\left(\frac{1}{B(s+2,k-s+1)}\right)\\
&\qquad\int_y^1 t^{s+1}(1-t)^{k-s}\,dt-y\mathbb{P}(V\ge y),\\
&=\frac{s+1}{k+2}F^B_{k+2,y}(s+1)-y\mathbb{P}(V\ge y), \qquad \text{(D.21)}\\
&\le\frac{s+1}{k+2}F^B_{k+2,y}(s+1),\\
&\le F^B_{k,y}(kx),\\
&\le\exp\{-kd(x,y)\}, \qquad \text{(D.22)}
\end{aligned}$$

where we use Fact 9 and the definition of the Beta CDF to establish Equation (D.21). The final bound in (D.22) is the result of the Chernoff-Hoeffding theorem and Fact 11. Let $\delta:=y-x$, and note that $s<kx\Rightarrow s\le\lfloor(k+1)x\rfloor$ because of s being integer, whence

$$\begin{aligned}
\mathsf{E}\big[(y-V)^+\big]&=\mathsf{E}\big[(y-V)\mathbb{1}(V\le y)\big],\\
&=\mathsf{E}\big[(y-V)\mathbb{1}(y-\delta\le V\le y)\big]\\
&\quad+\mathsf{E}\big[(y-V)\mathbb{1}(V<y-\delta)\big],\\
&>\mathsf{E}\big[(y-V)\mathbb{1}(V<y-\delta)\big],\\
&\ge\delta\mathsf{E}\big[\mathbb{1}(V<y-\delta)\big], \qquad \text{(D.23)}\\
&=\delta\mathbb{P}(V<x),\\
&=\delta\Big(1-F^B_{k+1,x}(s)\Big), \qquad \text{(D.24)}\\
&\ge\delta/2, \qquad \text{(D.25)}
\end{aligned}$$

where Equation (D.24) relies on Fact 9. Bound (D.25) is justified from Fact 10 and $s\le\lfloor(k+1)x\rfloor$. Thus, using the inequalities for both the numerator and denominator, we obtain the desired bound. □

D.3.2. Proof of Lemma D.6.

Proof. The steps in this proof follow a similar one in Agrawal and Goyal (2013), but we show them for completeness. We bound the number of times the suboptimal arm's mean is overestimated. Let τ_ℓ be the time step in which the suboptimal arm is sampled for the ℓ th time. Because for any x, $\lim_{n\to\infty} \frac{\lfloor nx \rfloor}{nx} = 1$, we can let N be a large enough integer so that if $\ell \geq N$, then $\eta_1 \frac{\lfloor \ell\eta_1 \rfloor}{\ell\eta_1} > x' := (\theta_2 + \eta_1)/2 > \theta_2$. In that case,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left(S_2(t) \geq \lfloor P_2(t)\eta_1 \rfloor, \pi_t^{\mathrm{OG}} = 2\right) \\
&\leq \mathsf{E}\left[\sum_{\ell=1}^{T} \sum_{t=\tau_\ell}^{\tau_{\ell+1}-1} \mathbb{1}\left(S_2(\ell) \geq \lfloor P_2(\ell)\eta_1 \rfloor\right) \mathbb{1}\left(\pi_t^{\mathrm{OG}} = 2\right)\right] \\
&= \mathsf{E}\left[\sum_{\ell=1}^{T} \mathbb{1}\left(S_2(\tau_\ell) \geq \lfloor (\ell-1)\eta_1 \rfloor\right) \sum_{t=\tau_\ell}^{\tau_{\ell+1}-1} \mathbb{1}\left(\pi_t^{\mathrm{OG}} = 2\right)\right] \\
&= \mathsf{E}\left[\sum_{\ell=0}^{T-1} \mathbb{1}\left(S_2(\tau_{\ell+1}) \geq \lfloor \ell\eta_1 \rfloor\right)\right] \\
&\leq N + \sum_{\ell=N+1}^{T-1} \mathbb{P}\left(S_2(\tau_{\ell+1}) \geq \ell\eta_1 \frac{\lfloor \ell\eta_1 \rfloor}{\ell\eta_1}\right) \\
&\leq N + \sum_{\ell=N+1}^{T-1} \mathbb{P}\left(S_2(\tau_{\ell+1}) \geq \ell x'\right) \\
&\leq N + \sum_{\ell=1}^{\infty} \exp\left(-\ell d(x', \theta_2)\right) \\
&= N + \frac{1}{1 - e^{-d(x',\theta_2)}}, \qquad \text{(D.26)}
\end{aligned}$$

where (D.26) follows from the Chernoff-Hoeffding theorem and the fact that $S_2(\tau_{\ell+1})$ is drawn from a Binomial$(P_2(\ell+1), \theta_2) \equiv$ Binomial(ℓ, θ_2) distribution. □

Appendix E. Further Experimental Results

E.1. Bayes UCB Experiment

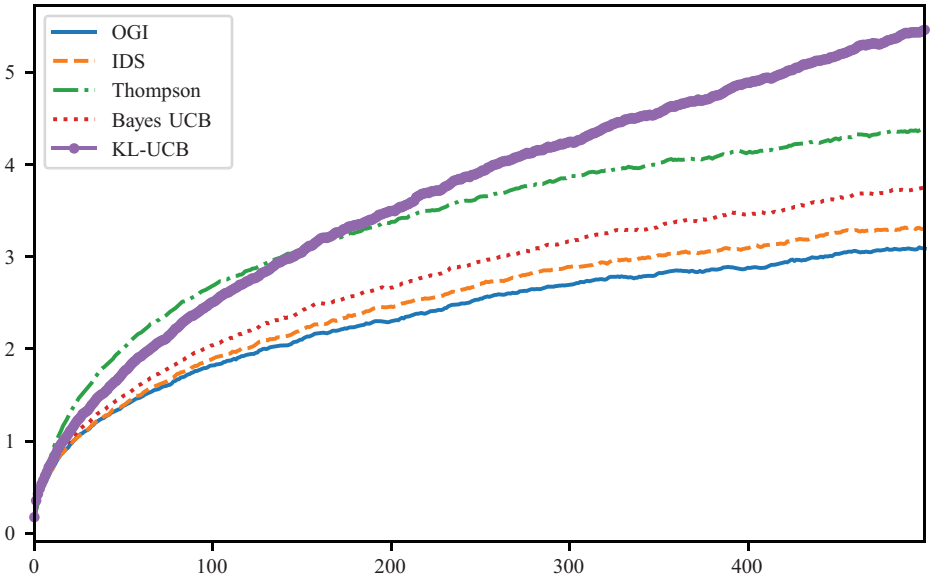
This experiment is motivated by Kaufmann et al. (2012a), and in it, we simulate the Bernoulli bandit problem with $T = 500$ and two arms. Because we are interested in measuring expected regret over the prior, we draw the arms' mean rewards at random from the uniform distribution. There are 5,000 independent trials, and we show the results in Figure E.1. OGI offers notable performance improvements over both Thompson sampling and IDS for this modest horizon.

E.2. Additional Benchmark Algorithms

In this section, we simulate a few additional algorithms to understand the importance of the varying discount factor, and to try out a different approximation of the Gittins index. We also simulate the greedy policy to see the inherent value of exploration in our benchmark problems. Specifically, the algorithms we experiment with are as follows:

- OGI with a one-step lookahead and a fixed discount factor of γ, which we will refer to throughout as FOGI($1/(1-\gamma)$). The quantity $1/(1-\gamma)$ can be interpreted as a rough horizon over which this policy is optimal.
- OGI in which the Gittins index approximation equals the closed-form expression given in Brezzi and Lai (2002). We will refer to this policy as BL-OGI.

Figure E.1. (Color online) Frequentist Regret



Note. The OGI policy is configured with $K = 1$ and $\alpha = 100$.

Table E.1. Comparison Against Some of OGI's Simpler Variants in the Gaussian Setup

	BL-OGI	Greedy	OGI(1)	FOGI(T)	FOGI($T/10$)	FOGI($10T$)
Mean	58.54	167.16	49.19	49.61	60.72	59.09
Standard error	2.14	9.74	1.61	1.90	4.25	1.47
25%	45.83	102.75	17.49	39.28	34.85	50.94
50%	56.87	156.63	41.72	47.29	52.19	57.84
75%	67.67	216.77	73.24	60.04	87.63	67.59

Table E.2. Comparison Against Some of OGI's Simpler Variants in the Bernoulli Setup

	BL-OGI	Greedy	OGI(1)	FOGI(T)	FOGI($T/10$)	FOGI($10T$)
Mean	23.50	56.32	18.12	16.52	18.69	20.14
Standard error	0.63	2.36	0.65	0.62	0.82	0.58
25%	18.74	37.61	6.26	12.36	12.70	16.36
50%	22.50	55.16	15.08	15.55	17.14	19.43
75%	28.18	74.64	27.63	19.72	23.29	23.62

• The greedy policy, which plays the arm in $\arg\max_i \mathsf{E}_{y_{i,N_i(t-1)}}[X_{i,t}]$. Effectively it is equivalent to FOGI(1) and completely disregards the value of future exploration. We will simply call this policy Greedy in our tables and plots.

Recall that the Gittins index policy is optimal for a geometrically distributed horizon with mean T. Because FOGI(T) is precisely an approximation for that policy, we would expect it to perform well in our experiments when the horizon is T (although it really should be geometrically distributed).

We reuse the two main experimental setups from Section 5: the Gaussian bandit with 10 independent arms and the Bernoulli equivalent. Notice from Table E.1, in the Gaussian setup, that there is value in knowing the true horizon T because FOGI(T) is the dominant policy.[6] We also see that either over- or underestimating the horizon leads to worse performance as demonstrated by the regret from FOGI($T/10$), FOGI($10T$), and Greedy. Interestingly, we also see that BL-OGI shows larger regret than OGI (Table 1), suggesting that there is perhaps value in using our optimistic approximation for this particular problem. The comparison against FOGI, BL-OGI, and Greedy in the Bernoulli case, presented in Table E.2, tells a similar story as in Table E.1.

E.3. Additional Tables for Section 5

Table E.3. Optimistic and Exact Gittins Indices When $\gamma = 0.9$ for Different Beta-Bernoulli Parameters

α	β	OGI(1)	OGI(3)	OGI(5)	Gittins
1	1	0.760	0.721	0.712	0.703
1	2	0.571	0.522	0.511	0.500
1	3	0.452	0.401	0.389	0.380
1	4	0.374	0.321	0.312	0.302
2	1	0.853	0.818	0.809	0.800
2	2	0.702	0.657	0.646	0.635
2	3	0.591	0.543	0.530	0.516
2	4	0.508	0.458	0.445	0.434
3	1	0.893	0.864	0.855	0.845
3	2	0.771	0.729	0.719	0.707
3	3	0.671	0.626	0.613	0.601
3	4	0.592	0.545	0.532	0.518
4	1	0.916	0.890	0.882	0.872
4	2	0.813	0.776	0.765	0.754
4	3	0.724	0.682	0.670	0.658
4	4	0.651	0.607	0.593	0.581

Table E.4. Optimistic and Exact Gittins Indices When $\gamma = 0.95$ for Different Beta-Bernoulli Parameters

α	β	OGI(1)	OGI(3)	OGI(5)	Gittins
1.0	1.0	0.817	0.784	0.774	0.761
1.0	2.0	0.637	0.590	0.577	0.560
1.0	3.0	0.514	0.463	0.449	0.433
1.0	4.0	0.430	0.376	0.364	0.348
2.0	1.0	0.890	0.860	0.851	0.838
2.0	2.0	0.752	0.710	0.698	0.681
2.0	3.0	0.643	0.596	0.581	0.562
2.0	4.0	0.558	0.509	0.494	0.475
3.0	1.0	0.921	0.896	0.887	0.874
3.0	2.0	0.811	0.773	0.762	0.744
3.0	3.0	0.715	0.672	0.658	0.639
3.0	4.0	0.637	0.591	0.575	0.556
4.0	1.0	0.938	0.916	0.908	0.895
4.0	2.0	0.847	0.812	0.801	0.784
4.0	3.0	0.763	0.722	0.709	0.690
4.0	4.0	0.691	0.648	0.633	0.613

Endnotes

[1] To capture the possibility of a randomized policy, $\mathcal{F}_t$ must also contain the realization of a random variable describing the randomization, but we ignore this here for notational brevity.

[2] Intuitively, we solve a degenerate stopping problem, where one is forced to stop at time 1. Such a problem involves no recourse decisions (or learning), and nature immediately reveals the true mean reward after pulling the arm once. This is why the index is especially easy to compute when $K = 1$.

[3] We use the termination condition in proposition 2.2.1 of Bertsekas (2011). That is, we iterate the value iteration operator k times so that $\bar{c}_k - \underline{c}_k$ (as in proposition 2.2.1) is at most some predetermined fixed tolerance.

[4] See https://github.com/gutin/FastGittins.

[5] This entails a minor abuse of our definition of a policy: $\tilde{\pi}_{\mathbf{y},T}$ is not specified for $t > T$.

[6] Knowing the horizon T, in the context of this paper, should be viewed as a form of cheating because we are interested in anytime policies.

References

Agrawal R (1995) Sample mean based index policies with $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem. *Adv. Appl. Probabilities* 27(4):1054–1078.

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. Mannor S, Srebro N, Williamson RC, eds. *Proc. Conf. on Learning Theory, Edinburgh, Scotland* (JMLR), 39.1–39.26. http://proceedings.mlr.press/v23/agrawal12/agrawal12.pdf.

Agrawal S, Goyal N (2013) Further optimal regret bounds for Thompson sampling. Mannor S, Srebro N, Williamson RC, eds. *Proc. 16th Internat. Conf. on Artificial Intelligence and Statist., Scottsdale* (JMLR), 99–107. http://proceedings.mlr.press/v31/agrawal13a.pdf.

Audibert JY, Bubeck S (2010) Regret bounds and minimax policies under partial monitoring. *J. Machine Learning Res.* 11(Oct):2785–2836.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multi-armed bandit problem. *Machine Learning* 47(2-3):235–256.

Berry DA, Fristedt B (1985) *Bandit Problems: Sequential Allocation of Experiments.* Monographs on Statistics and Applied Probability (Chapman and Hall, London).

Bertsekas DP (2011) *Dynamic Programming and Optimal Control*, vol. ii, 3rd ed. (Athena Scientific, Belmont, MA).

Bertsimas D, Niño-Mora J (1996) Conservation laws, extended polymatroids and Multi-armed Bandit problems: A polyhedral approach to indexable systems. *Math. Oper. Res.* 21(2):257–306.

Besson L, Kaufmann E (2018) What doubling tricks can and can't do for multi-armed bandits. Preprint, submitted March 19, https://arxiv.org/abs/1803.06971.

Bradt RN, Johnson S, Karlin S (1956) On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* 27(4):1060–1074.

Brezzi M, Lai TL (2002) Optimal learning and experimentation in bandit problems. *J. Econom. Dynamic Control* 27(1):87–108.

Brown DB, Smith JE (2020) Index policies and performance bounds for dynamic selection problems. *Management Sci.* 66(7):3029–3050.

Chakravorty J, Mahajan A (2013) Multi-armed bandits, Gittins index, and its calculation. *Methods Appl. Statist. Clinical Trials: Planning, Anal., Inferential Methods* 2:416–435.

Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. *Adv. Neural Inform. Processing Systems* 24:2249–2257.

Cover TM, Thomas JA (2012) *Elements of Information Theory* (John Wiley & Sons, Hoboken, NJ).

Fang LS, Lai CC (1987) Characterization and purification of biliverdin reductase from the liver of eel, *Anguilla japonica*. *Comparative Biochemistry Physiology Part B: Comparative Biochemistry* 88(4):1151–1155.

Garivier A, Cappé O (2011) The KL-UCB algorithm for bounded stochastic bandits and beyond. Kakade S, von Luxburg U, eds. *Proc. 24th Annual Conf. on Learning Theory, Budapest, Hungary* (JMLR), 359–376. http://proceedings.mlr.press/v19/garivier11a/garivier11a.pdf

Gittins JC (1979) Bandit processes and dynamic allocation indices. *J. Royal Statist. Soc. B* 41(2):148–177.

Jogdeo K, Samuels SM (1968) Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *Ann. Math. Statist.* 39(4):1191–1195.

Katehakis MN, Robbins H (1995) Sequential choice from several populations. *Proc. National. Acad. Sci. USA* 92(19):8584.

Katehakis MN, Veinott AF Jr (1987) The multi-armed bandit problem: Decomposition and computation. *Math. Oper. Res.* 12(2):262–268.

Katehakis MN, Rothblum UG (1996) Finite state multi-armed bandit problems: Sensitive-discount, average-reward and average-overtaking optimality. *Ann. Appl. Probabilities* 6(3):1024–1034.

Kaufmann E (2018) On Bayesian index policies for sequential resource allocation. *Ann. Statist.* 46(2):842–865.

Kaufmann E, Cappé O, Garivier A (2012a) On Bayesian upper confidence bounds for bandit problems. Lawrence N, ed. *15th Internat. Conf. Artificial Intelligence Statist. (AISTATS), Palma, Canary Islands* (PLMR), 592–600.

Kaufmann E, Korda N, Munos R (2012b) *Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. Algorithmic Learning Theory* (Springer, Berlin).

Korda N, Kaufmann E, Munos R (2013) Thompson sampling for 1-dimensional exponential family bandits. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems, Lake Tahoe, California (NeurIPS).*

Lai TL (1987) Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* 15(3):1091–1114.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.

Lattimore T (2016) Regret analysis of the finite-horizon Gittins index strategy for Multi-armed Bandits. Feldman V, Rakhlin A, Shamir O, eds. *Proc. Conf. on Learning Theory, New York* (JMLR), 1–32. http://proceedings.mlr.press/v49/lattimore16.pdf.

Maillard OA, Munos R, Stoltz G (2011) A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. Kakade S, von Luxburg U, eds. *Proc. Conf. on Learning Theory, Budapest, Hungary* (JMLR), 497–514.

Niño-Mora J (2007) A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov Chain. *INFORMS J. Comput.* 19(4):596–606.

Powell WB, Ryzhov IO (2012) *Optimal Learning*, vol. 841 (John Wiley & Sons, Hoboken, NJ).

Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. (Nova Scotia)* 58(5):527–535.

Russo D (2021) A note on the equivalence of upper confidence bounds and gittins indices for patient agents. *Oper. Res.* 69(1):273–278.

Russo D, Van Roy B (2018) Learning to optimize via information-directed sampling. *Oper. Res.* 66(1):230–252.

Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Appl. Stochastic Models Bus. Industry* 26(6):639–658.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3):285–294.

Tsitsiklis JN (1994) A short proof of the Gittins index theorem. *Ann. Appl. Probabilities* 4(1):194–199.

Varaiya P, Walrand J, Buyukkoc C (1985) Extensions of the multi-armed bandit problem: The discounted case. *IEEE Trans. Automated Control* 30(5):426–439.

Weber R, et al (1992) On the Gittins index for multi-armed bandits. *Ann. Appl. Probabilities* 2(4):1024–1033.

Whittle P (1980) Multi-armed bandits and the Gittins index. *J. Royal Statist. Soc. B.* 42(2):143–149.

Whittle P (1988) Restless bandits: Activity allocation in a changing world. *J. Appl. Probabilities* 25(1):287–298.

Yao YC, et al. (2006) *Some Results on the Gittins Index for a Normal Reward Process. Time Series and Related Topics* (Institute of Mathematical Statistics).

Vivek F. Farias is the Patrick J. McGovern Professor at the MIT Sloan School, affiliated with the operations management group and the Operations Research Center.

Eli Gutin is a senior machine learning engineer at Uber Technologies, Inc.