Optimistic Gittins Indices

Vivek F. Farias
Sloan School of Management
Massachusetts Institute of Technology
email: vivekf@mit.edu

Eli Gutin
Operations Research Center
Massachusetts Institute of Technology
email: gutin@mit.edu

Abstract

Starting with the Thomspon Sampling algorithm, recent years have seen a resurgence of interest in Bayesian algorithms for the Multi-armed Bandit (MAB) problem. These algorithms seek to exploit prior information on arm biases and while several have been shown to be regret optimal, their design has not emerged from a principled approach. In contrast, if one cared about Bayesian regret discounted over an infinite horizon at a *fixed, pre-specified* 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 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 by offering substantially improved performance with little to no additional computational overhead. In addition, we prove that the simplest of these approximations yields frequentist regret that matches the Lai-Robbins lower bound, including achieving matching constants.

Keywords: multi-armed bandits; Gittins index; online learning;

1. Introduction

The Multi-Armed Bandit (MAB) problem is perhaps the simplest example of a learning problem that exposes the tension between exploration and exploitation. Recent years have seen a resurgence of interest in Bayesian MAB problems wherein we are endowed with a prior on arm rewards, and a number of index schemes that exploit this prior have been proposed and analyzed. These include Thompson Sampling (Thompson 1933), Bayes-UCB (Kaufmann et al. 2012), KL-UCB (Garivier 2011), and Information Directed Sampling (Russo and Van Roy 2014). The ultimate motivation for these algorithms appears to be two-fold: superior empirical performance and light computational burden. The strongest performance results available for these algorithms establish regret lower bounds that match the Lai-Robbins lower bound (Lai and Robbins 1985). Even among this set of recently proposed algorithms, there is a wide spread in empirically observed performance. Table 1 lists well-known algorithms in the literature and what is known about them.

Algorithm	Bayes/Frequentist	Regret Optimal	Framework
KL-UCB	Frequentist	Yes	Index-based
UCB	Frequentist	Unknown	Index-based
MOSS	Frequentist	Unknown	Index-based
Thompson Sampling	Bayes	Yes	Posterior Sampling
Bayes UCB	Bayes	Yes	Index-based
IDS	Bayes	Unknown	Mixed
Gittins Index	Bayes	No	Index-based

Table 1: Summary of some famous bandit policies and their properties.

Interestingly, the design of the index policies referenced above has been somewhat ad-hoc as opposed to having emerged from a principled analysis of the underlying Markov Decision Process. Now if in contrast to requiring 'small' regret for all sufficiently large time horizons, we cared about minimizing Bayesian regret over an infinite horizon, discounted at a fixed, pre-specified rate (or equivalently, maximizing discounted infinite horizon rewards), the celebrated Gittin's index theorem provides an *optimal, efficient* solution. Importing this celebrated result to the fundamental problem of designing algorithms that achieve low regret (either frequentist or Bayesian) simultaneously over all sufficiently large time horizons runs into two substantial challenges:
High-Dimensional State Space: Even minor 'tweaks' to the discounted infinite horizon objective appear to render the corresponding Markov Decision problem for the Bayesian MAB problem intractable. For instance, it is known that a Gittins-like index strategy is sub-optimal for a fixed, finite-horizon (Berry and Fristedt 1985). Moreover, the problem of minimizing regret simultaneously over all sufficiently large horizons is not well understood.

Computational Burden: Even in the context of the discounted infinite horizon problem, the computational burden of calculating a Gittins index is substantially larger than that required for any of the index schemes for the multi-armed bandit discussed thus far.

The present paper attempts to make progress on these challenges. Specifically, we make the following contributions:

- We propose a class of 'optimistic' approximations to the Gittins index that can be computed with significantly less effort. In fact, the computation of 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 nearest competitor, IDS.
- We establish that an arm selection rule that is greedy with respect to the simplest of these optimistic approximations achieves optimal regret in the sense of meeting the Lai-Robbins lower bound (including matching constants) provided the discount factor is increased at a certain rate.

- We show empirically that even the simplest optimistic approximation to the Gittins index proposed here *outperforms the state-of-the-art incumbent schemes discussed in this introduction by a non-trivial margin.* We view this as our primary contribution – the Bayesian MAB problem is fundamental making the performance improvements we demonstrate important.

1.1. Relevant Literature

Thompson Sampling Thompson (1933) was proposed as a heuristic to the MAB problem in 1933, but was largely ignored until the last decade. An empirical study by Chapelle and Li Chapelle and Li (2011) highlighted Thompson Sampling's superior performance and led to a series of strong theoretical guarantees for the algorithm being proved in Agrawal and Goyal (2012, 2013), Kaufmann et al. (2012) (for specific cases when Gaussian and Beta priors are used). Recently, these proofs were generalized to the 1D exponential family of distributions in Korda et al. (2013). A few decades after Thompson Sampling was introduced, Gittins Gittins (1979) showed that an index policy was optimal for the infinite horizon discounted MAB problem. Several different proofs for the optimality of Gittins index, were shown in Tsitsiklis (1994), Weber et al. (1992), Whittle (1980), Bertsimas and Niño-Mora (1996). Inspired by this breakthrough, Lai and Robbins (1985), Lai (1987), while ignoring the original MDP formulation, proved an asymptotic lower bound on achievable (non-discounted) regret and suggested policies that attained it.

Simple and efficient UCB algorithms were later developed in Agrawal (1995), Auer et al. (2002), Audibert and Bubeck (2010), with finite time regret bounds. These were followed by the KL-UCB (Garivier (2011)) and Bayes UCB (Kaufmann et al. (2012)) algorithms. The Bayes UCB paper drew attention to how well Bayesian algorithms performed in the frequentist setting. In that paper, the authors also demonstrated that a policy using indices similar to Gittins' had the lowest regret. The use of Bayesian techniques for bandits was explored further in Russo and Van Roy (2014) where the authors propose Information Directed Sampling, an algorithm that exploits complex information structures arising from the prior. There is also a very recent paper, Lattimore (2016), which also focuses on regret minimization using approximated Gittins Indices. However, in that paper, the time horizon is assumed to be known and fixed, which is different from the focus in this paper on finding a policy that has low regret over all sufficiently long horizons.

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. Next, we consider a more general version of the multi-armed bandit problem in Section 5, where multipled simultaneous “pulls" are allowed and propose heuristics for that setting, which are derived from the framework in this paper. Finally, Section 6 presents experiments showing how OGI achieves lower Bayesian regret than state of the art policies and is computationally efficient and, in Section 7 we state open questions that remain following this paper.

2. Model and Preliminaries

We consider a multi-armed bandit problem with a finite set of arms $\mathcal{A} = \{1, \ldots, A\}$. Arm $i \in \mathcal{A}$ if pulled at time t, generates a stochastic reward $X_{i,N_i(t)}$ where $N_i(t)$ denotes the cumulative number of pulls of arm i up to and including time t. $(X_{i,s}, s \in \mathbb{N})$ is an i.i.d. sequence of random variables, each distributed according to $p_{\theta_i}(\cdot)$ where $\theta_i \in \Theta$ is a parameter. Denote by θ the tuple of all θ_i. The expected reward from the i^{th} arm is denoted by $\mu_i(\theta_i) := \mathsf{E}[X_{i,1}|\theta_i]$. We denote by $\mu^*(\theta)$ the maximum expected reward across arms; $\mu^*(\theta) := \max_i \mu_i(\theta_i)$ and let i^* be an optimal arm. The present paper will focus on the Bayesian setting, and so we suppose that each θ_i is an independent draw from some prior distribution q over Θ. All random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define a policy, $\pi := (\pi_t, t \in \mathbb{N})$, to be a stochastic process taking values in $\mathcal{A}$. We require that π be adapted to the filtration $\mathcal{F}_t$ generated by the history of arm pulls and their corresponding rewards up to and including time $t-1$.

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

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

the reward accumulated up to time T when using policy π. We write $V(\pi, T) := \mathsf{E}V(\pi, T, \theta)$. The regret of a policy over T time periods, for a specific realization $\theta \in \Theta^A$, is the expected shortfall against always pulling the optimal arm, namely

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

In a seminal paper, Lai and Robbins (1985) established a lower bound on achievable regret. They considered the class of policies under which for *any* choice of θ and positive constant a, any policy in the class achieves $o(n^a)$ regret. They showed that for any policy π in this class, and any θ with a unique maximum, we must have

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

where d_{KL} is the Kullback-Liebler divergence. The Bayes' risk (or Bayesian regret) is simply the expected regret over draws of θ according to the prior q:

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

In yet another landmark paper, Lai and Robbins (1985) 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) were also Bayes optimal. Interestingly, however, this class of algorithms ignores information about the prior altogether. 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. (2012), KL-UCB Garivier (2011), and Information Directed Sampling Russo and Van Roy (2014).

The Bayesian setting endows us with the structure of a (high dimensional) Markov Decision process. An alternative objective to minimizing Bayes risk, is the maximization of the cumulative reward discounted over an infinite horizon. Specifically, for any positive discount factor $\gamma < 1$, define

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

The celebrated Gittin's index theorem provides an *optimal, efficient* solution to this problem that we will describe in greater detail shortly; unfortunately as alluded to earlier even a minor 'tweak' to the objective above – such as maximizing cumulative expected reward over a finite horizon renders the Gittins index sub-optimal Niño-Mora (2011).

As a final point of notation, every scheme we consider will maintain a posterior on the mean of an arm at every point in time. We denote by $q_{i,s}$ the posterior on the mean of the ith arm after $s-1$ pulls of that arm; $q_{i,1} := q$. Since our prior on θ_i will frequently be conjugate to the distribution of the reward X_i, $q_{i,s}$ will permit a succinct description via a sufficient statistic we will denote by $y_{i,s}$; denote the set of all such sufficient statistics $\mathcal{Y}$. We will thus use $q_{i,s}$ and $y_{i,s}$ interchangeably and refer to the latter as the 'state' of the ith arm after $s-1$ pulls.

3. Gittins Indices and Approximations

One way to compute the Gittins Index is via the so-called retirement value formulation Whittle (1980). The *Gittins Index* for arm i in state y is the value for λ that solves

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

We denote this quantity by $\nu_\gamma(y)$. 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 can thus succinctly be stated as follows: at time t, play an arm in the set $\text{argmax}_i\, v_\gamma(y_{i,N_i(t)})$. Ignoring computational considerations, we cannot hope for a scheme such as the one above to achieve acceptable regret or Bayes risk. Specifically, denoting the Gittins policy by $\pi^{G,\gamma}$, we have

Lemma 1. *There exists an instance of the multi armed bandit problem for which*

$$\text{Regret}\left(\pi^{G,\gamma}, T\right) = \Omega(T)$$

for any $\gamma \in (0,1)$.

Proof. Berry and Fristedt (1985) show that the Gittins index is, in general, sub-optimal for bandit problems without both geometric discounting and an infinite horizon. ■

The above result is expected. 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. Fixing this issue then requires that the discount factor employed increase over time.

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

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

and consider using the policy that at time t picks an arm from the set $\text{argmax}_i\, \nu_{\gamma_t}(y_{i,N_i(t)})$. Denote this policy by π^{D}. The following proposition shows that this 'doubling' policy achieves Bayes risk that is within a factor of $\log T$ of the optimal Bayes risk. Specifically, we have:

Proposition 1.

$$\text{Regret}(\pi^{\text{D}}, T) = O\left(\log^3 T\right).$$

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

The proof of this simple result (Appendix A) 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 a doubling trick.

While increasing discount factors does not appear to get us to the optimal Bayes risk (the achievable lower bound being $\log^2 T$; see Lai (1987)); we conjecture that in fact this is a deficiency in our analysis for Proposition 1. In any case, the policy π^{D} is not the primary subject of the paper but merely a motivation for the discount factor schedule proposed. Putting aside this issue, one is still left with the computational burden associated with π^{D} – which is clearly onerous relative to any of the incumbent index rules discussed in the introduction.

3.1. 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. Since the state space $\mathcal{Y}$ associated with this problem is typically at least countable, solving this stopping problem, although not necessarily intractable, is a non-trivial computational task. 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 drawn from the prior on expected arm reward specified by y. Define the retirement value $R_{\lambda,K}(s,y)$ according to

$$R_{\lambda,K}(s,y) = \begin{cases} \lambda, & \text{if } s < K+1 \\ \max\left(\lambda, R(y)\right), & \text{otherwise} \end{cases}$$

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

$$\frac{\lambda}{1-\gamma} = \sup_{1<\tau\leq K+1} \mathsf{E}\left[\sum_{s=1}^{\tau-1} \gamma^{s-1} X_{i,s} + \gamma^{\tau-1}\frac{R_{\lambda,K}(\tau, y_{i,\tau})}{1-\gamma} \mid y_{i,1} = y\right]. \tag{3}$$

We denote the solution to this equation by $v_\gamma^K(y)$. The problem above admits a simple, attractive interpretation: nature reveals the *true* mean reward for the arm at time $K+1$ should we choose to not retire prior to that time, which enables the decision maker to then instantaneously decide whether to retire at time $K+1$ or else, never retire. In this manner one is better off than in the stopping problem inherent to the definition of the Gittins index, so that we use the moniker optimistic. Since we need to look ahead at most K steps in solving the stopping problem implicit in the definition above, the computational burden in index computation is limited. The following Lemma formalizes this intuition

Lemma 2. *For all discount factors γ and states $y \in \mathcal{Y}$, we have*

$$v_\gamma^K(y) \geq v_\gamma(y) \quad \forall K.$$

Proof. See Appendix A.1 ■

It is instructive to consider the simplest version of the approximation proposed here, namely the case where $K = 1$. There, equation (3) simplifies to

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

where $\hat{\mu}(y) := \mathsf{E}[R(y)]$ is the mean reward under the prior given by y. The equation for λ above can also be viewed as an upper confidence bound to an arm's expected reward. Solving equation (4) is often simple in practice, and we list a few examples to illustrate this:

Example 1 (Beta). *In this case y is the pair (a, b), which specifices a Beta prior distribution. The 1-step Optimistic Gittins Index, is the value of λ that solves*

$$\lambda = \frac{a}{a+b} + \gamma \mathsf{E}[(\lambda - \mathrm{Beta}(a,b))^+] = \frac{a}{a+b}(1 - \gamma F^{\beta}_{a+1,b}(\lambda)) + \gamma\lambda(1 - F^{\beta}_{a,b}(\lambda))$$

where $F^{\beta}_{a,b}$ is the CDF of a Beta distribution with parameters a, b.

Example 2 (Gaussian). *Here $y = (\mu, \sigma^2)$, which specifices a Gaussian prior and the corresponding equation is*

$$\begin{aligned}\lambda &= \mu + \gamma \mathsf{E}[(\lambda - \mathcal{N}(\mu, \sigma^2))^+] \\ &= \mu + \gamma \left[(\lambda - \mu)\Phi\left(\frac{\mu - \lambda}{\sigma}\right) + \sigma\phi\left(\frac{\mu - \lambda}{\sigma}\right) \right]\end{aligned}$$

Notice that in both the Beta and Gaussian examples, the equations for λ are in terms of distribution functions. Therefore it's straightforward to compute a derivative for these equations (which would be in terms of the density and CDF of the prior) and makes finding a solution, using a method such as Newton-Raphson, simple and efficient.

We summarize the Optimistic Gittins Index (OGI) algorithm succinctly as follows.

Assume the state of arm i at time t is given by $y_{i,t}$, and let $\gamma_t = 1 - 1/t$. Play an arm

$$i^* \in \underset{i}{\operatorname{argmax}}\, v^K_{\gamma_t}(y_{i,t}),$$

and update the posterior on the arm based on the observed reward.

4. Analysis and Regret bounds

We establish a regret bound for Optimistic Gittins Indices when the algorithm is given the parameter $K = 1$, the prior distribution q is uniform and arm rewards are Bernoulli. The result shows that the algorithm, in that case, meets the Lai-Robbins lower bound and is thus asymptotically optimal, in both a frequentist and Bayesian sense. After stating the main theorem, we briefly discuss two generalizations to the algorithm.

In the sequel, whenever $x, y \in (0,1)$, we will simplify notation and let $d(x,y) := d_{\mathrm{KL}}(\mathrm{Ber}(x), \mathrm{Ber}(y))$. Also, we will refer to the Optimistic Gittins Index policy as π^{OG} and let the lookahead parameter K be any positive integer. Moreover, we denote the Optimistic Gittins Index of the i^{th} arm as $v_{i,t}^K := v_{1-1/t}^K(y_{i,t})$. Now we state the main result:

Theorem 1. *Let $\epsilon > 0$. For the multi-armed bandit problem with Bernoulli rewards and any parameter vector $\theta \subset [0,1]^A$, there exists $T^* = T^*(\epsilon, \theta)$ and $C = C(\epsilon, \theta)$ such that for all $T \geq T^*$,*

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

where $C(\epsilon, \theta)$ is a constant that is only determined by ϵ and the parameter θ.

Proof. Because we prove frequentist regret, the first few steps of the proof will be similar to that of UCB and Thompson Sampling.

Assume w.l.o.g that arm 1 is uniquely optimal, and therefore $\theta^* = \theta_1$. Fix an arbitrary suboptimal arm, which for convenience we will say is arm 2. Let $N_i(t)$ denote the number of pulls of the ith arm, by (but not including) time t. Finally, we let $S_i(t)$ be the corresponding total reward accumulated from the ith arm from the $N_i(t)$ pulls. That is,

$$N_i(t) = \sum_{t'=1}^{t=1} \mathbb{1}\left(\pi^{\mathrm{OG}} = i\right), \qquad S_i(t) = \sum_{t'=1}^{N_i(t)} X_{i,t'}.$$

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 $L(T) := \frac{\log T}{d(\eta_2, \eta_3)}$.

We upper bound the expected number of pulls of the second arm as follows,

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

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 describe at a high-level with the details in the Appendix.

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

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

Proof outline. The goal is to bound $\mathbb{P}\left(v_{1,t}^K < \eta\right)$ by an expression that decays fast enough in t so that the series converges. To prove this, we express the event $\{v_{1,t}^K < \eta\}$ in the form $\{W_t < 1/t\}$ for some sequence of random variables W_t. It turns out that for large enough t, $\mathbb{P}(W_t < 1/t) \leq \mathbb{P}\left(cU^{1/(1+h)} < 1/t\right)$ where U is a uniform random variable, $c, h > 0$ and therefore $\mathbb{P}\left(v_{1,t}^K < \eta\right) = O\left(\frac{1}{t^{1+h}}\right)$. The full proof is in Appendix B.1.

We remark that the core technique in the proof of Lemma 3 is the use of the Beta CDF. As such, our analysis can, in some sense, improve the result for Bayes UCB. In the main theorem of Kaufmann et al. (2012), the authors state that the quantile in their algorithm is required to be $1 - 1/(t \log^c T)$ for some parameter $c \geq 5$, however they show simulations with the quantile $1 - 1/t$ and suggest that, in practice, it should be used instead. By utilizing techniques in our analysis, it is possible to prove that the use of $1 - 1/t$, as a discount factor, in Bayes UCB 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 $T^* = T^*(\epsilon, \theta)$ sufficiently large and a constant $C_2 = C_2(\epsilon, \theta_1, \theta_2)$ so that for any $T \geq T^*$, we have*

$$\sum_{t=1}^{T} \mathbb{P}\left(\pi_t^{\text{OG}} = 2,\ v_{2,t}^K \geq \eta_3,\ N_2(t) \geq L(T)\right) \leq C_2.$$

Proof outline. This relies on a concentration of measure result and the assumption that the 2^{nd} arm was sampled at least $L(T)$ times. The full proof is given in Appendix B.2. ■

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

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

From this the regret bound follows. ■

4.1. Generalizations and a tuning parameter

There is an argument in Agrawal and Goyal Agrawal and Goyal (2012) which shows that any algorithm optimal for the Bernoulli bandit problem, can be modified to yield an algorithm that has $O(\log T)$ regret with general bounded stochastic rewards. Therefore Optimistic Gittins Indices is an effective and practical alternative to policies such as Thompson Sampling and UCB. We also suspect that the proof of Theorem 1 can be generalized to all lookahead values ($K > 1$) and to a general exponential family of distributions.

A slight modification to Optimistic Gittins Indices gives an algorithm that has $O(\log T)$ regret for the general stochastic bandit problem with bounded rewards. Specifically, when arms have arbitrary reward distributions, bounded in the interval $[a, b]$, the approach is to each time sample the reward $X_{i,t}$, then generate an 'artifical' Bernoulli reward with probability $X_{i,t}/(b-a)$ and provide that as input to the policy. The choice of arm pulls from the resulting algorithm leads to an $O(\log T)$ regret bound, as is shown in that paper. The constant in front of $\log T$, however, depends on the KL divergences of *Bernoulli* random variables as opposed to the actual underlying distributions.

Another important observation is that the discount factor for Optimistic Gittins Indices does not have to be exactly $1-1/t$. In fact, a tuning parameter can be added to make the discount factor $\gamma_{t+\alpha} = 1 - 1/(t+\alpha)$ instead. 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. Multiple simultaneous arm pulls

In this section we show how the approach in this paper can be applied to a more general Multi-Armed Bandit problem, where the decision maker is able to "pull" up to a certain number (say $m < A$) of the arms simultaneously. (Whittle 1988) considers a slightly more general version of the problem just discussed, where arms that are not pulled (idled) are able to change state and proposes an index scheme, Whittle's heuristic, for it. However, if arms that are idled are frozen in state Whittle's heuristic becomes equivalent to pulling arms with the m largest Gittins indices.

For the purposes of this section, we denote by the action space $\mathcal{A}$ the set of all binary vectors with K ones in them, and X_t to be a tuple of (potential) rewards from all A arms at time t. A policy π is then a non-anticipative sequence of such vectors and we define its regret over T periods to be

$$\text{Regret}\,(\pi, T) = T \cdot \max_{a \in A} a^\top \mathsf{E}[X_t] - \sum_{t=1}^{T} \mathsf{E}[\pi_t^\top X_t]$$

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

We give two examples of policies that empirically have low regret. Both rely on the use of an increasing discount factor, that is $\gamma_t = 1 - 1/t$, and on approximating solutions to a sequence of Markov Decision problems whose rewards are discounted by γ_t (if it's the t^{th} problem). The first of these heuristics involves pulling arms with m largest Optimistic Gittins Indices, which is essentially an approximation to Whittle's heuristic. The second is approximating the solution of the t^{th} Markov Decision problem by an using a Linear Programming relaxation of it (Bertsimas and Niño-Mora 1996).

6. Computational Experiments

Our goal is to benchmark Optimistic Gittins Indices (OGI) against state-of-the-art algorithms in Bayesian and frequentist setting. 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. We replicate some of the same experiments as in papers of Russo and Van Roy (2014), Kaufmann et al. (2012) and then add our own for evaulating the problem with multiple simultaneous arm pulls.

We consider the OGI algorithm for two values of the lookahead parameter K (1 and 3) , and in one experiment included for completeness, the case of exact Gittins indices ($K = \infty$). We used a common discount factor schedule in all experiments setting $\gamma_t = 1 - 1/(100 + t)$. The choice of $\alpha = 100$ is second order and our conclusions remain unchanged (and actually appear to improve in an absolute sense) with other choices (we show this in a second set of experiments).

A major consideration in running the 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 CDF over, at least, hundreds of iterations. By contrast, the version of OGI with $K = 1$ uses 10 iterations of the Newton-Raphson method. In the remainder of this section, we discuss the results.

6.1. Gaussian

This experiment (Table 2) replicates one in Russo and Van Roy (2014). Here 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. We simulate 1000 independent trials with 10 arms and 1000 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; see Powell and Ryzhov (2012), Chapter 6.1.3. We term the use of that approximation 'OGI(∞) Approx'. In addition to regret, we show the average CPU time taken, in seconds, to execute each trial.

Algorithm	**OGI(1)**	**OGI(∞) Approx.**	**IDS**	**TS**	**Bayes UCB**
Mean Regret	49.19	47.64	55.83	67.40	60.30
S.D.	51.07	50.59	65.88	47.38	45.35
1st quartile	17.49	16.88	18.61	37.46	31.41
Median	41.72	40.99	40.79	63.06	57.71
3rd quartile	73.24	72.26	78.76	94.52	86.40
CPU time (s)	0.02	0.01	11.18	0.01	0.02

Table 2: Gaussian experiment. OGI(1) denotes OGI with $K = 1$, while OGI Approx. uses the approximation to the Gaussian Gittins Index from Powell and Ryzhov (2012).

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 since it gives a closer approximation to the Gittins Index. At the same time, OGI is essentially as fast as Thomspon sampling, and three orders of magnitude faster than its nearest competitor (in terms of regret).

6.2. Bernoulli

In this experiment regret is simulated over 1000 periods, with 10 arms each having a uniformly distributed Bernoulli parameter, over 1000 independent trials (Table 3). We use the same setup as in Russo and Van Roy (2014) for consistency.

Algorithm	**OGI(1)**	**OGI(3)**	**OGI(∞)**	**IDS**	**TS**	**Bayes UCB**
Mean Regret	18.12	18.00	17.52	19.03	27.39	22.71
S.D.	20.71	20.37	21.40	21.42	18.19	17.27
1st quartile	6.26	5.60	4.45	5.85	14.62	10.09
Median	15.08	14.84	12.06	14.06	23.53	18.52
3rd quartile	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

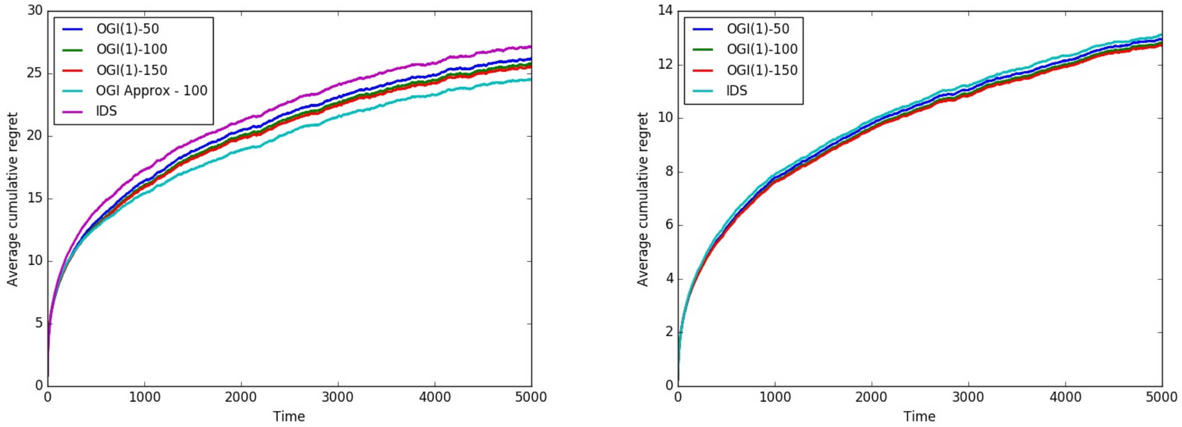
Table 3: Bernoulli experiment. OGI(K) denotes the OGI algorithm with a K step approximation and tuning parameter $\alpha = 100$. OGI(∞) is the algorithm that uses Gittins Indices.

Each version of OGI outperforms other algorithms and the one that uses (actual) Gittins Indices has the lowest mean regret. Perhaps, unsurprisingly, when OGI looks ahead 3 steps it performs marginally better than with a single step. Nevertheless, looking ahead 1 step is a reasonably close approximation to the Gittins Index in the Bernoulli problem. In fact the approximation error, when using an optimistic 1 step approximation, is around 15% and if K is increased to 3, the error drops to around 4%.

6.2.1. Longer horizon and robustness

To understand how Bayes' regret grows in the long run, we simulate Bernoulli and Gaussian bandit problems for a longer horizon of 5000 time steps with the results shown in Figure 1a.

In the Bernoulli experiment of this section, due to the computational cost, we are only able to simulate OGI with $K = 1$. In addition, to show robustness with respect to the choice of tuning parameter α, we show results for $\alpha = 50, 100, 150$. The message here is essentially the same as in the earlier experiments: the OGI scheme offers a non-trivial performance improvement at a tiny fraction of the computational effort required by its nearest competitor. We omit Thompson Sampling and Bayes UCB from the plots in order to more clearly see the difference between OGI and IDS.



(a) Gaussian experiment

(b) Bernoulli experiment

Figure 1: Bayesian regret. In the legend, OGI(K)-α is the format used to indicate parameters K and α. The OGI Appox policy uses the approximation to the Gittins index from Powell and Ryzhov (2012).

In the Gaussian experiment, we again see that OGI dominates other policies and the tuning parameter has the same effect of lowering regret as it's increased. Also, just as in Table 2, the OGI algorithm that uses the Gaussian-specific approximation has the best performance.

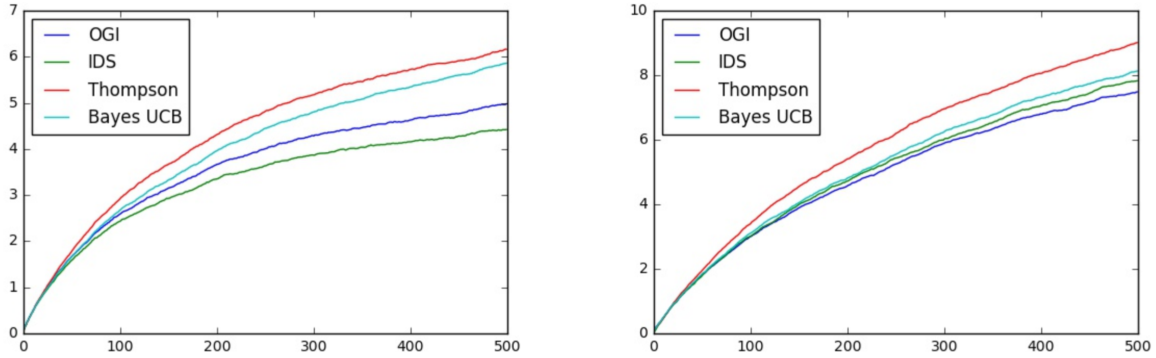
6.3. Bayes UCB, Kaufmann et al. (2012) experiment

For this section that replicates some of the experiments in Kaufmann et al. (2012), we simulate the Bernoulli bandit problem over 500 time steps and with 2 arms. The arms' parameters are fixed to the same values as in Kaufmann et al. (2012) and regret is averaged over 5,000 independent trials. We show the results in Figures 2a and 2b, and as in the Bayes' regret case we see that OGI offers notable performance improvements of Thompson Sampling and Bayes UCB for this modest horizon, but where the arm parameters are fixed rather than being drawn at random.

6.4. Bandits with multiple arm pulls

We evaluate the heuristics outlined in Section 5, for the case when m simulataneous arm pulls are allowed, and compare them against Thompson Sampling and IDS. The implementation of IDS in this setting was the same as for the linear bandits problem; see (Russo and Van Roy 2014) for details. Thompson Sampling was straightfoward in that it merely involved sampling from the prior of all arms and picking the arms with the largest m samples.

For the experiment we have $A = 6$ binary arms with uniformly distributed biases and fixed m to be 3. We simulate 1000 independent trials and show the results in Figure 3 and Table 4. There is a significant spread in the performance between the OGI-inspired heuristics and both Thompson Sampling and IDS. Again, despite the closest competitor being IDS, the computational cost incurred



(a) Bernoulli experiment with two arms and $\mu_1 = 0.1$, $\mu_2 = 0.2$

(b) Bernoulli experiment with two arms and $\mu_1 = 0.1$, $\mu_2 = 0.2$

Figure 2: Frequentist regret. The OGI policy is configured $K = 1$ and $\alpha = 100$.

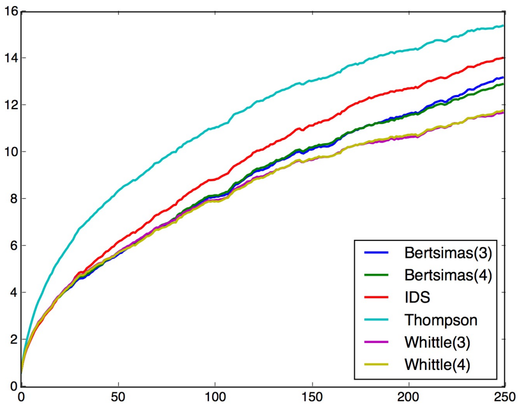


Figure 3: Results from restless experiment

in running IDS makes Whittle's heuristic, which has the some order of complexity as for the case $m = 1$, seem like a particularly attractive algorithm. In fact the implementation of IDS for this experiment required taking Monte-Carlo samples, which can be prohibitively expensive.

	Bertsimas(3)	Bertsimas(4)	IDS	Thompson	Whittle(3)	Whittle(4)
Mean	13.20	12.92	14.04	15.41	11.70	11.79
Std	19.42	19.07	19.17	13.39	14.96	15.19
Quartile1	0.76	0.06	1.07	6.21	1.50	1.27
Median	10.06	9.87	11.21	15.08	10.68	10.56
Quartile3	21.90	21.37	23.61	23.28	20.75	21.32

Table 4: Regret from the multiple arm pulls experiment. "Whittle(K)" refers to the Whittle heuristic-like policy described in Section 5, using K lookahead steps. "Bertsimas(K)" refers to the alternative heuristic from the same section using the same number of lookahead steps.

7. Conclusions

This paper proposed a novel way for designing Bayesian Multi-Armed Bandit algorithms by treating the problem of minimizing regret as a sequence of separate Markov Decision problems 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 sub-linear regret and, when applied to a binary bandit problem, using a simple and efficient algorithm with a flat Beta prior achieves the optimal rate of growth in regret.

There are many open questions following this work. First, it remains to be proven that using Gittins Indices, or equivalently playing an arm optimally for each MDP, using the increasing discount factor technique does produce an algorithm whose regret matches the Lai-Robbins lower bound. Secondly, it is worth exploring whether the idea of this framework can be extended to contextual bandit problems or those 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 solutions to bandit problems with dependent arms.

References

Agrawal R (1995) Sample mean based index policies with $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability* 1054–1078.

Agrawal S, Goyal N (2012) Analysis of thompson sampling for the multi-armed bandit problem. *COLT*.

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

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

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed 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)* (Springer).

Bertsimas D, Niño-Mora J (1996) Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Mathematics of Operations Research* 21(2):257–306.

Chapelle O, Li L (2011) An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 2249–2257.

Cover TM, Thomas JA (2012) *Elements of information theory* (John Wiley & Sons).

Garivier A (2011) The kl-ucb algorithm for bounded stochastic bandits and beyond. *COLT*.

Gittins JC (1979) Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* 148–177.

Jogdeo K, Samuels SM (1968) Monotone convergence of binomial probabilities and a generalization of ramanujan's equation. *The Annals of Mathematical Statistics* 1191–1195.

Kaufmann E, Korda N, Munos R (2012) Thompson sampling: An asymptotically optimal finite-time analysis. *Algorithmic Learning Theory*, 199–213 (Springer).

Korda N, Kaufmann E, Munos R (2013) Thompson sampling for 1-dimensional exponential family bandits. *Advances in Neural Information Processing Systems*, 1448–1456.

Lai TL (1987) Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics* 1091–1114.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

Lattimore T (2016) Regret analysis of the anytime optimally confident ucb algorithm. *arXiv preprint arXiv:1603.08661* .

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

Powell WB, Ryzhov IO (2012) *Optimal learning*, volume 841 (John Wiley & Sons).

Russo D, Van Roy B (2014) Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 1583–1591.

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

Tsitsiklis JN (1994) A short proof of the gittins index theorem. *The Annals of Applied Probability* 194–199.

Weber R, et al. (1992) On the gittins index for multiarmed bandits. *The Annals of Applied Probability* 2(4):1024–1033.

Whittle P (1980) Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* 143–149.

Whittle P (1988) Restless bandits: Activity allocation in a changing world. *Journal of applied probability* 287–298.

A. Proof of Proposition 1

Proof. First, letting $\gamma_n = 1 - 1/n$, we show that

$$\text{Regret}\left(\pi^{G,\gamma_n}, n\right) = O\left(\log^2(n)\right). \tag{7}$$

Let $H \sim \mathrm{Geo}(1/n)$ be an exogenous geometric random variable that is independent of θ and not observed by the agent. As an abbreviation, define $\mu^* = \mathsf{E}_q[\mu^*(\theta)]$. We then have

$$
\begin{align}
\sum_{t=1}^{\infty} \gamma^{t-1} \mathsf{E}\left[X_{\pi^{G,\gamma_n},t}\right] &= \mathsf{E}\left[\sum_{t=1}^{H} X_{\pi_t^{G,\gamma_n},t}\right] \notag \\
&= \mathsf{E}\left[H\mu^*(\theta) - \mathrm{Regret}\left(\pi^{G,\gamma_n}, H\right)\right] \tag{8} \\
&= n\mu^* - \mathsf{E}\left[\mathrm{Regret}\left(\pi^{G,\gamma_n}, H\right)\right] \notag \\
&\leq n\mu^* - \mathsf{E}\left[\mathrm{Regret}\left(\pi^{G,\gamma_n}, H\right) \mid H > n\right] \mathbb{P}\left(H > n\right) \notag \\
&\leq n\mu^* - \mathsf{E}\left[\mathrm{Regret}\left(\pi^{G,\gamma_n}, n\right)\right] (1 - 1/n)^n \notag \\
&= n\mu^* - \mathsf{E}\left[\mathrm{Regret}\left(\pi^{G,\gamma_n}, n\right)\right] (e^{-1} + o(1)). \tag{9}
\end{align}
$$

Let q, Q be the density and CDF, respectively, of the prior distribution. Now, by Theorem 3, part 1 of Lai (1987), there exists (an efficient) policy $\tilde{\pi}$, such that as n becomes sufficiently large

$$
\mathrm{Regret}\left(\tilde{\pi}, n\right) \sim \left(A(A-1)\int_{\Theta} q^2(\theta) Q^{A-2}(\theta)\, d\theta\right) \log^2 n.
$$

Therefore for some prior-dependent constant C_q, we have $\mathrm{Regret}\left(\tilde{\pi}, n\right) \leq C_q \log^2 n$. Let $\Delta(\theta)$ denote worst case single period regret under parameter θ, that is, $\Delta(\theta) = \max_i \mu(\theta^*) - \mu(\theta_i)$. Let Δ denote its expectation over θ, from which we obtain the lower bound,

$$
\begin{align}
\mathsf{E}\sum_{t=1}^{H} X_{\pi_t^{G,\gamma_n},t} &\geq \mathsf{E}\left[\sum_{t=1}^{H} X_{\tilde{\pi}_t,t}\right] \tag{10} \\
&= \mathsf{E}\left[H\mu^*(\theta) - \mathrm{Regret}\left(\tilde{\pi}, H\right)\right] \notag \\
&\geq \mathsf{E}\left[H\mu^*(\theta) - \mathrm{Regret}\left(\tilde{\pi}, H\right) \mathbb{1}\left(H \geq e\right) - 2\mathbb{1}\left(H < e\right)\Delta(\theta)\right] \notag \\
&\geq n\mu^* - C_q \mathsf{E}\left[(\log(H))^2 \mathbb{1}\left(H \geq 3\right)\right] - 2\Delta \notag \\
&\geq n\mu^* - C_q \mathsf{E}\left[(\log(H))^2 \mid H \geq 3\right] \mathbb{P}\left(H \geq 3\right) - 2\Delta \notag \\
&\geq n\mu^* - C_q \log^2(n+3) \mathbb{P}\left(H \geq 3\right) - 2\Delta \tag{11}
\end{align}
$$

where (10) holds by optimality of the Gittins Index. The bound (11) follows from the memoryless property of the Geometric distribution, from Jensen's inequality and the fact that function $\log^2 x$ is a concave function on $[e, +\infty)$. Thus, equation (7) is implied by the bounds (9) and (11).

Now, for any policy π, we define $\tilde{L}_\pi(m) := m\mu^* - \sum_{t=1}^{m} X_{\pi_t,t}$ to be the random m period shortfall against the expected Bayes' optimal arm and let $g_k = 1 - 1/2^{k-1}$. We break up the time horizon T

into geometrically growing epochs and bound, conservatively, the Bayes' risk in each one:

$$
\begin{aligned}
\text{Regret}\left(\pi^D, T\right) &\leq \text{Regret}\left(\pi^D, 2^{\lceil \log_2 T \rceil}\right) && (12)\\
&= \sum_{k=1}^{\lceil \log_2 T \rceil} \mathsf{E}\left[\mathsf{E}\left[\tilde{L}_{\pi^D}(2^{k-1}) \,\middle|\, \mathcal{F}_{2^{k-1}-1}\right]\right]\\
&= \sum_{k=1}^{\lceil \log_2 T \rceil} \mathsf{E}\left[\mathsf{E}\left[\tilde{L}_{\pi^{G,g_k}}(2^{k-1}) \,\middle|\, \mathcal{F}_{2^{k-1}-1}\right]\right]\\
&\leq \sum_{k=1}^{\lceil \log_2 T \rceil} \mathsf{E}\left[\mathsf{E}\left[\tilde{L}_{\pi^{G,g_k}}(2^{k-1}) \,\middle|\, \mathcal{F}_0\right]\right] && (13)\\
&= \sum_{k=1}^{\lceil \log_2 T \rceil} \text{Regret}\left(\pi^{G,g_k}, 2^{k-1}\right) = O\left(\sum_{k=1}^{\lceil \log_2 T \rceil} k^2\right) && (14)\\
&= O(\log^3 T)
\end{aligned}
$$

where (14) follows from equation (7) and (13) holds because regret increases if history is discarded. ■

A.1. Proof of Lemma 2

Proof. Fix $\lambda > 0$ and an arm i. Let $V_\lambda(y)$ be the value of the RHS of (2) with the per-period reward of λ, and define $\hat{V}_\lambda^K(y)$, similarly, for problem (3) (where y is, as before, the state of an arm). Notice that because rewards are generated according to an unknown parameter θ_i, which needs to be learned, that if we condition on a fixed θ_i, we have for any stopping time τ that

$$
\mathsf{E}\left[\sum_{t=1}^{\tau-1} \gamma^{t-1} X_{i,t} + \gamma^{\tau-1}\frac{\lambda}{1-\gamma} \,\middle|\, \theta_i\right] \leq \mathsf{E}\left[\sum_{t=1}^{\tau-1} \gamma^{t-1} \mu(\theta_i) + \gamma^{\tau-1}\frac{\lambda}{1-\gamma} \,\middle|\, \theta_i\right] \qquad (15)
$$

where the expectation is also taken over the agent's prior on θ_i. Simply put, the best performance in the bandit game can be achieved if the parameter governing expected rewards is known from the beginning by the agent. Now recall that $R(y_{i,t})$ is a random variable drawn from the prior on the arm's mean reward at time t. We also define the function

$$
r_{\lambda,K}(t,x) = \begin{cases} \lambda & t < K \\ \max(x,\lambda) & \text{otherwise} \end{cases}
$$

Let τ be the stopping time at which the agent retires and define $\tau_K = \tau \wedge (K+1)$. We then bound $V_\lambda(y)$,

$$\begin{aligned}
V_\lambda(y) &= \sup_{\tau>1} \mathsf{E}\left[\sum_{t=1}^{\tau-1} \gamma^{t-1} X_{i,t} + \gamma^{\tau-1}\frac{\lambda}{1-\gamma} \;\Big|\; y_{i,1}=y\right] \\
&= \sup_{\tau>1} \mathsf{E}\left[\mathsf{E}\left[\sum_{t=1}^{\tau-1} \gamma^{t-1} X_{i,t} + \gamma^{\tau-1}\frac{\lambda}{1-\gamma} \;\Big|\; \theta_i\right] \;\Big|\; y_{i,1}=y\right] \\
&= \sup_{\tau>1} \mathsf{E}\left[\sum_{t=1}^{\tau_K-1} \gamma^{t-1} X_{i,t} + \mathsf{E}\left[\sum_{t=\tau_K}^{\tau-1} \gamma^{t-1} X_{i,t} + \gamma^{\tau-1}\frac{\lambda}{1-\gamma} \;\Big|\; \theta_i\right] \;\Big|\; y_{i,1}=y\right] \\
&\le \sup_{\tau>1} \mathsf{E}\left[\sum_{t=1}^{\tau_K-1} \gamma^{t-1} X_{i,t} + \mathsf{E}\left[\sum_{t=\tau_K}^{\tau-1} \gamma^{t-1} \mu(\theta_i) + \gamma^{\tau-1}\frac{\lambda}{1-\gamma} \;\Big|\; \theta_i\right] \;\Big|\; y_{i,1}=y\right] \qquad (16) \\
&= \sup_{\tau>1} \mathsf{E}\left[\sum_{t=1}^{\tau_K-1} \gamma^{t-1} X_{i,t} + \mathsf{E}\left[\gamma^{\tau_K-1}\frac{r_{\lambda,K}(\tau_K,\mu(\theta_i))}{1-\gamma} \;\Big|\; \theta_i\right] \;\Big|\; y_{i,1}=y\right] \\
&= \sup_{\tau>1} \mathsf{E}\left[\sum_{t=1}^{\tau_K-1} \gamma^{t-1} X_{i,t} + \frac{\gamma^{\tau_K-1}}{1-\gamma}\mathsf{E}\left[\mathsf{E}\left[r_{\lambda,K}(\tau_K,\mu(\theta_i)) \;\Big|\; \theta_i\right] \;\Big|\; \mathcal{F}_{\tau_K-1}\right]\right] \\
&= \sup_{\tau>1} \mathsf{E}\left[\sum_{t=1}^{\tau_K-1} \gamma^{t-1} X_{i,t} + \frac{\gamma^{\tau_K-1}}{1-\gamma}\underbrace{\mathsf{E}\left[r_{\lambda,K}(\tau_K,R(y_{i,\tau_K}))\right]}_{=R_{\lambda,K}(\tau_K,y_{i,\tau_K})} \;\Big|\; y_{i,1}=y\right] \\
&= \sup_{1<\tau\le K+1} \mathsf{E}\left[\sum_{t=1}^{\tau} \gamma^{t-1} X_{i,t} + \gamma^{\tau-1}\frac{R_{\lambda,K}(\tau,y_{i,\tau})}{1-\gamma} \;\Big|\; y_{i,1}=y\right] = \hat{V}_\lambda^K(y).
\end{aligned}$$

The main step in the above is (16) where we bound on the inner conditional expection (in terms of θ_i) by applying (15). We also used the fact that $\mu(\theta_i) \mid \mathcal{F}_{t-1} \sim R(y_{i,t})$ for all t. Finally observe that both $\hat{V}_\lambda^K(y)$ and $V_\lambda(y)$ are increasing in λ for any fixed y. Therefore if $\lambda_1 = (1-\gamma)\hat{V}_{\lambda_1}^K(y)$ and $\lambda_2 = (1-\gamma)V_{\lambda_2}(y)$, then, because $V_\lambda(y) \le \hat{V}_\lambda^K(y)$ for any λ, it must be that $\lambda_1 \ge \lambda_2$. A simple argument shows this, which we omit. ■

A.2. Results for the frequentist regret bound proof

A.2.1. Definitions and properties of Binomial distributions.

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

Definition 1. *$F^B_{n,p}(.)$ is the CDF of the Binomial distribution with parameters n and p, and $F^\beta_{a,b}(.)$ is the CDF of the Beta distribution with parameters a and b.*

Lemma 5. *Let a and b be positive integers and $y \in [0,1]$,*

$$F^\beta_{a,b}(y) = 1 - F^B_{a+b-1,y}(a-1)$$

Proof. Proof is found in Agrawal and Goyal (2012). ■

Lemma 6. *The median of a Binomial(n, p) distribution is either $\lceil np \rceil$ or $\lfloor np \rfloor$.*

Proof. Proof is found in Jogdeo and Samuels (1968). ■

Corollary 1 (Corollary of Fact 6)**.** *Let n be a positive integer and $p \in (0, 1)$. For any nonnegative integer $s < np$*

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

Lemma 7. *Let n be a positive integer and $p \in [0, 1]$. Then for any $k \in \{0, \ldots, n\}$,*

$$(1 - p)F_{n-1,p}(k) \leq F_{n,p}(k) \leq F^B_{n-1,p}(k)$$

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

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

Now to prove $(1 - p)F_{n-1,p}(k) \leq F_{n,p}(k)$, it's enough to observe that $F_{n,p}(k) = pF_{n-1,p}(k - 1) + (1 - p)F_{n-1,p}(k)$. ■

A.2.2. Ratio of Binomial CDFs.

Lemma 8. *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*

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

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

$$\frac{e^{-nd(j/n,r)}}{(1 + n)^2} \leq F_{n,r}(j) \leq (n + 1)^2 e^{-nd(j/n,r)}. \tag{17}$$

Because $k < nq < np$, by applying (17) to both the numerator and denominator, we get

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

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,

$$\frac{F_{n,q}(k)}{F_{n,p}(k)} > \frac{e^{nd(q,p)}}{(n+1)^4} = \frac{e^{\frac{n}{2}d(q,p)}}{(n+1)^4}e^{\frac{n}{2}d(q,p)} \geq e^{\frac{n}{2}d(q,p)}. \tag{18}$$

The final lower bound in (18) follows from the assumption on n. ■

B. Optimistic Gittins Index results.

For this section it is useful to define the value of a stopping problem used in the calculation of the Optimistic Gittins Index. For any fixed arm i, we write

$$V_K(y;x,\gamma) := \sup_{1<\tau\leq K} \mathsf{E}\left[\sum_{t=1}^{\tau-1}(1-\gamma)X_{i,t} + \gamma^{\tau-1}R_{x,K}(\tau, y_{i,\tau}) \mid y_{i,1} = y\right].$$

Therefore the Optimistic Gittins Index, in state y with discount factor γ, is the solution in x to $V_K(y;x,\gamma) = x$. We show some key properties of V_K, which we exploit later on. For fixed y, K and γ, we will call $V(x) \triangleq V(y;x,\gamma)$.

Fact 1. *$V(x)$ is convex and differentiable.*

Proof. We prove this by induction. For $K = 1$, we have $\tau = 2$ almost surely and so

$$\begin{aligned} V(x) &= V_1(y;x,\gamma) \\ &= (1-\gamma_t)\mathsf{E}\left[X_{i,1} \mid y_{i,1=y}\right] + \gamma_t\mathsf{E}\left[R_{x,1}(2, y_{i,1}) \mid y_{i,1} = y\right] \\ &= (1-\gamma_t)\mathsf{E}\left[X_{i,1} \mid y_{i,1=y}\right] + \gamma_t\mathsf{E}\left[\max(x, \mathsf{E}X_{i,1}) \mid y_{i,1} = y\right] \end{aligned}$$

The function is convex because for any random variable Z the term $\max(g, Z)$ is convex and taking expectations preserves convexity. Also, we verify through the bounded convergence theorem that $V(x)$ is differentiable and the event $\{\mathsf{E}X_{i,1} = z \mid y_{i,1} = y\}$, at which V is not differentiable, has measure zero.

For $K > 1$, assume that V_{K-1} is convex and differentiable. By writing the Bellman equation

$$V_K(y;x,\gamma) = (1-\gamma_t)\mathsf{E}\left[X_{i,1} \mid y_{i,1} = y\right] + \mathsf{E}\left[\gamma_t \max(x, V_{K-1}(y_{i,t+1};x,\gamma)) \mid y_{i,1} = y\right]$$

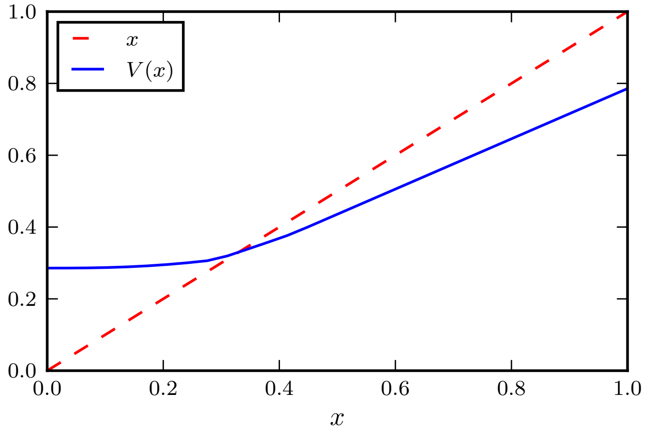


Figure 4: Visualization of Lemma 9's proof for a Beta-Bernoulli problem with parameters $(3, 7)$ and $\gamma = 0.7$. The intersection of the two lines represents the Optimistic Gittins Index.

we again notice a maximum of convex functions in x. This form for V_K implies that it is convex and differentiable. ■

Lemma 9. *Let* $\gamma \in (0, 1)$ *and*

$$\lambda = \sup\{x \in [0, 1] : V(x) \geq x\} \tag{19}$$

For all $x \in (0, 1)$*, the following equivalence holds*

$$\lambda < x \iff V(x) < x. \tag{20}$$

Proof. Figure 4 gives a visualization of the proof. In the plot, the point x^*, where $V(x^*)$ and x^* intersect, is the Optimistic Gittins Index. Notice how if $x > x^*$, $V(x)$ is below x and otherwise $V(x)$ is above x; this is the crux of the proof.

We also give a formal proof. Firstly let's assume that $\lambda < x$. If $x \leq V(x)$, then λ would not be the supremum over all $z \in [0, 1]$ such that $z \leq V(z)$. Therefore $V(x) < x$.

For the converse, assume $\lambda \geq x$. Recall that $V(z)$ is convex. Since V is continuous on $[0, 1]$, by the Intermediate Value Theorem, there exists a point λ at which $\lambda = V(\lambda)$. Therefore let $\epsilon < (1 - \lambda)/2$ and from the first direction of the proof, we have $\lambda + \epsilon > V(\lambda + \epsilon)$. Thus

$$V(\lambda + \epsilon) \geq V(\lambda) + \epsilon V'(\lambda) = \lambda + \epsilon V'(\lambda)$$

where the inequality follows from V being convex and differentiable. This implies that $V'(\lambda) < 1$

and, moreover, because V is also increasing, it follows that $V'(\lambda) \in (0,1)$, whence

$$\begin{aligned} V(x) &\geq V(\lambda) - (\lambda - x)V'(\lambda) \\ &= \lambda - (\lambda - x)V'(\lambda) \\ &= (1 - V'(\lambda))\lambda + V'(\lambda)x \\ &\geq \min(x, \lambda) = x. \end{aligned}$$

This completes the proof. ■

Corollary 2. *Let $v_{i,t}$ be the Optimistic Gittins Index of arm i at time t and let $x \in (0,1)$. The following equivalence holds*

$$\{v_{i,t}^K < x\} = \{V_K(y_{i,t}; x, \gamma_t) < x\}$$

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

Proof. By the definition in Equation (4), $v_{i,t}^K$ can be characterized with the relation

$$v_{i,t} = \sup\left\{x \in [0,1] : x \leq V_K(y_{i,t}; x, \gamma_t)\right\}.$$

The conclusion then follows from Lemma 9. ■

B.1. Proof of Lemma 3

Proof. This proof involves induction. We state what the base case is, in the following Lemma, but we defer its proof to the end of the section.

Lemma 10. *Suppose the OGI algorithm uses only one lookahead step, and so $K = 1$. Then we have*

$$\mathbb{P}\left(v_{1,t}^1 < \eta\right) = \mathcal{O}\left(\frac{1}{t^{1+h(\eta)}}\right)$$

where h is a positive yet decreasing function of η.

Now we show the induction step. Suppose that for $K \geq 2$, it holds that

$$\mathbb{P}\left(v_{1,t}^{K-1} < \eta\right) = \mathcal{O}\left(\frac{1}{t^{1+h(\eta)}}\right).$$

We show the same is true for $v^K_{1,t}$. Indeed, we have for $t > 1$

$$\begin{align}
\mathbb{P}\left(v^K_{1,t} < \eta\right) &= \mathbb{P}\left(V_K(y_{1,t};\eta,\gamma_t) < \eta\right) \tag{21}\\
&= \mathbb{P}\left((1-\gamma_t)\mathsf{E}\left[X_{1,t} \mid y_{1,t}\right] + \gamma_t\mathsf{E}\left[\max(\eta, V_{K-1}(y_{1,t+1};\eta,\gamma_t)) \mid y_{1,t}\right] < \eta\right) \notag\\
&\le \mathbb{P}\left(\frac{1}{t}\mathsf{E}\left[X_{1,t} \mid y_{1,t}\right] + \gamma_t \max(\eta, \mathsf{E}\left[V_{K-1}(y_{1,t+1};\eta,\gamma_t) \mid y_{1,t}\right]) < \eta\right) \tag{22}\\
&\le \mathbb{P}\left(\frac{1}{t}\mathsf{E}\left[X_{1,t} \mid y_{1,t}\right] + \left(1-\frac{1}{t}\right)\mathsf{E}\left[V_{K-1}(y_{1,t+1};\eta,\gamma_t) \mid y_{1,t}\right] < \eta\right) \notag\\
&\le \mathbb{P}\left(\frac{1}{t}\mathsf{E}\left[X_{1,t} \mid y_{1,t}\right] + \left(1-\frac{1}{t}\right)V_{K-1}(y_{1,t};\eta,\gamma_t) < \eta\right) \tag{23}\\
&\le \mathbb{P}\left(V_{K-1}(y_{1,t};\eta,\gamma_t) < \eta\left(\frac{t}{t-1}\right)\right) \notag\\
&\le \mathcal{O}\left(\frac{1}{t^{1+h(\eta t/(t-1))}}\right) = \mathcal{O}\left(\frac{1}{t^{1+h(\eta)}}\right) \notag
\end{align}$$

where (21) follows from Lemma 2 and (22) uses Jensen's inequality. [(23) is yet to be shown but I suspect it's 100% true from computer simulations]

We finish the proof by using the following asymptotic argument. Take M to be a large enough integer, then we have, using the result of the induction proof,

$$\sum_{t=1}^{\infty}\mathbb{P}\left(v^K_{1,t} < \eta\right) \le M + \sum_{t=M+1}^{\infty}\frac{C_1}{t^{1+h(\eta)}} \le M + C_2$$

where $C_2 = C_2(\eta)$ is the limit of the series and C_1 is a constant used in the definition of the big-Oh. ■

B.1.1. Proof of the base case in Lemma 10

Proof. For simplicity let's abbreviate $v^1_{1,t}$ as $v_{1,t}$. 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.

The proof consists of showing two claims

Claim 1: $\{v_{1,t} < \eta\} \subseteq \left\{F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t}\right\}$**:**

Let V_t ~Beta$(S_1(t)+1, N_1(t)-S_1(t)+1)$ be the agent's posterior on the optimal arm. Using Corollary 2 and the simplified form for $K = 1$

$$V_K((S_1(t)+1, N_1(t)-S_1(t)+1);\eta,\gamma_t) = \mathsf{E}V_t + \gamma_t\mathsf{E}\left[(\eta - V_t)^+\right]$$

we find that

$$\begin{aligned}
\{v_{1,t} < \eta\} &= \left\{\mathsf{E}\left[V_t\right] + \gamma_t \mathsf{E}\left[(\eta - V_t)^+\right] < \eta\right\} \\
&= \left\{(1-1/t)\mathsf{E}\left[(\eta - V_t)^+\right] < \mathsf{E}\left[\eta - V_t\right]\right\} \qquad (24) \\
&= \left\{\mathsf{E}\left[(\eta - V_t)^+\right] - \mathsf{E}\left[\eta - V_t\right] < \frac{1}{t}\mathsf{E}\left[(\eta - V_t)^+\right]\right\} \\
&= \left\{\mathsf{E}\left[(V_t - \eta)^+\right] < \frac{1}{t}\mathsf{E}\left[(\eta - V_t)^+\right]\right\} \\
&\subseteq \left\{\mathsf{E}\left[(V_t - \eta)^+\right] < \frac{1}{t}\right\} \qquad (25)
\end{aligned}$$

where (24) follows from the definition of γ_t and (25) is due to V_t, η both lying in the interval $[0,1]$. We approximate the conditional expectation in (25) with

$$\begin{aligned}
\mathsf{E}\left[(V_t - \eta)^+ \mid S_1(t), N_1(t)\right] &= \mathsf{E}\left[(V_t - \eta)\mathbb{1}\left(V_t \geq \eta\right)\right] \\
&= \mathsf{E}\left[(V_t - \eta)\mathbb{1}\left(\eta + \delta > V_t \geq \eta\right)\right] \\
&\quad + \mathsf{E}\left[(V_t - \eta)\mathbb{1}\left(V_t \geq \eta + \delta\right)\right] \\
&> \mathsf{E}\left[(V_t - \eta)\mathbb{1}\left(V_t \geq \eta + \delta\right)\right] \\
&\geq \delta \mathbb{P}\left(V_t \geq \eta'\right) \\
&= \delta(1 - F_{S_1(t)+1, N_1(t)-S_1(t)+1}(\eta')) = \delta F^B_{N_1(t)+1,\eta'}(S_1(t)) \qquad (26)
\end{aligned}$$

The last equality is due to Fact 5 and this proves the claim.

Claim 2: $\sum_{t=1}^{\infty} \mathbb{P}\left(F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t}\right) \leq C_1$ where C_1 is a constant:

Let us fix the sequence $f_t = -\frac{\log \delta t}{\log(1-\eta')} - 1 = O(\log t)$. We then have

$$\begin{aligned}
\mathbb{P}\left(F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t}\right) &= \mathbb{P}\left(F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t},\ N_1(t) > f_t\right) \\
&\quad + \mathbb{P}\left(F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t},\ N_1(t) \leq f_t\right). \qquad (27)
\end{aligned}$$

For the second term in the RHS of (27) we have the following bound,

$$\begin{aligned}
\mathbb{P}\left(F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t},\ N_1(t) \leq f_t\right) &\leq \mathbb{P}\left(F^B_{N_1(t)+1,\eta'}(0) < \frac{1}{\delta t},\ N_1(t) \leq f_t\right) \\
&= \mathbb{P}\left((1-\eta')^{N_1(t)+1} < \frac{1}{\delta t},\ N_1(t) \leq f_t\right) \\
&\leq \mathbb{P}\left((1-\eta')^{f_t+1} < \frac{1}{\delta t}\right) = 0. \qquad (28)
\end{aligned}$$

Now we use the following fact to bound the left term on the RHS of (27). Define the function

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

which is the inverse CDF. Then it is known that if $U \sim \text{Unif}(0,1)$, then $F_{n,p}^{-B}(U) \sim \text{Binomial}(n,p)$. Furthermore, $F_{n,p}^{B}(F_{n,p}^{-B}(U)) \geq U$ due to the definition of the inverse CDF.

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

1. M is such that $e^{d(\eta',\theta_1)f_M/2} > (f_M + 1)^4$
2. $M > \frac{4}{(1-\eta')\delta}$
3. $\lceil f_M \rceil > 0$ and $F^B_{\lceil f_M \rceil, \eta'}(f_M \eta') > 1/4$. Note that 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 $\{F^B_{N_1(t),\eta'}(S_1(t)) < \frac{1}{(1-\eta')\delta t},\ S_1(t) \geq N_1(t)\eta',\ N_1(t) > f_t\}$ has measure zero because of the assumptions made on M. Therefore if $t > M$, we have

$$\begin{aligned}
&\mathbb{P}\Big(F^B_{N_1(t)+1,\eta'}(S_1(t)) < \frac{1}{\delta t},\ N_1(t) > f_t\Big) \\
&\leq \mathbb{P}\left(F^B_{N_1(t),\eta'}(S_1(t)) < \frac{1}{(1-\eta')\delta t},\ N_1(t) > f_t\right) && (29)\\
&= \mathbb{P}\left(F^B_{N_1(t),\eta'}(S_1(t)) < \frac{1}{(1-\eta')\delta t},\ S_1(t) < N_1(t)\eta',\ N_1(t) > f_t\right) \\
&= \mathbb{P}\left(F^B_{N_1(t),\theta_1}(S_1(t))\frac{F^B_{N_1(t),\eta'}(S_1(t))}{F^B_{N_1(t),\theta_1}(S_1(t))} < \frac{1}{(1-\eta')\delta t},\ S_1(t) < N_1(t)\eta',\ N_1(t) > f_t\right) \\
&\leq \mathbb{P}\left(F^B_{N_1(t),\theta_1}(S_1(t))e^{N_1(t)D} < \frac{1}{(1-\eta')\delta t},\ N_1(t) > f_t\right) && (30)\\
&\leq \mathbb{P}\left(F^B_{N_1(t),\theta_1}(S_1(t))e^{f_t D} < \frac{1}{(1-\eta')\delta t}\right) \\
&= \mathbb{P}\left(F^B_{N_1(t),\theta_1}(F^{-B}_{N_1(t),\theta_1}(U)) < \frac{e^{-f_t D}}{(1-\eta')\delta t}\right) && (31)\\
&\leq \mathbb{P}\left(U < \frac{e^{-f_t D}}{(1-\eta')\delta t}\right) \\
&= \frac{e^{-f_t D}}{(1-\eta')\delta t} \\
&= O\left(\frac{1}{t^{1+Dc_{\eta'}}}\right) && (32)
\end{aligned}$$

where $D = d(\eta', \theta_1) > 0$ and $c_{\eta'} = -\log^{-1}(1-\eta') > 0$ are constant. The bound (29) holds due to Fact (7). Bound (30) follows from an application of Lemma 8 and the fact that $t > M$. Equation (31) follows from $S_1(t) \sim \text{Binomial}(N_1(t), \theta_1)$ and the inverse sampling technique. By combining bounds (32), (28) and (27), we get the big-Oh bound. ■

B.2. 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 $v_{i,t}^K \leq v_{i,t}^{K-1} \leq \ldots \leq v_{i,t}^1$, it will be sufficient to consider this proof only for $v_{2,t}^1$, which we also will abbreviate as $v_{2,t} \triangleq v_{2,t}^1$.

Firstly, by the law of total probability, we find that

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(v_{2,t} \geq \eta_3,\ N_2(t) \geq L,\ \pi_t^{\mathrm{OG}} = 2) \\
&\qquad = \sum_{t=1}^{T} \mathbb{P}\left(v_{2,t} \geq \eta_3,\ N_2(t) \geq L,\ S_2(t) < \lfloor N_2(t)\eta_1 \rfloor,\ \pi_t^{\mathrm{OG}} = 2\right) \\
&\qquad\quad + \sum_{t=1}^{T} \mathbb{P}\left(v_{2,t} \geq \eta_3,\ N_2(t) \geq L,\ S_2(t) \geq \lfloor N_2(t)\eta_1 \rfloor,\ \pi_t^{\mathrm{OG}} = 2\right) \\
&\qquad \leq \sum_{t=1}^{T} \mathbb{P}\left(v_{2,t} \geq \eta_3,\ N_2(t) \geq L,\ S_2(t) < \lfloor N_2(t)\eta_1 \rfloor\right) + \sum_{t=1}^{T} \mathbb{P}\left(\pi_t^{\mathrm{OG}} = 2,\ S_2(t) \geq \lfloor N_2(t)\eta_1 \rfloor\right) \qquad (33)
\end{aligned}$$

Let $V_t \sim \mathrm{Beta}(S_2(t)+1, N_2(t)-S_2(t)+1)$ denote the agent's posterior on the second arm at time t, then

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(v_{2,t} \geq \eta_3,\ N_2(t) \geq L,\ S_2(t) < \lfloor N_2(t)\eta_1 \rfloor) \\
&\qquad = \sum_{t=1}^{T} \mathbb{P}\left(\mathsf{E}\left[V_t\right] + \gamma_t \mathsf{E}\left[(\eta_3 - V_t)^+\right] \geq \eta_3,\ N_2(t) \geq L,\ S_2(t) < \lfloor N_2(t)\eta_1 \rfloor\right) \\
&\qquad = \sum_{t=1}^{T} \mathbb{P}\left(\frac{\mathsf{E}\left[(\eta_3 - V_t)^+\right]}{\mathsf{E}\left[(V_t - \eta_3)^+\right]} \leq t,\ N_2(t) \geq L,\ S_2(t) < \lfloor N_2(t)\eta_1 \rfloor\right) \qquad (34)
\end{aligned}$$

where the second equality follows from Corollary 2 in Appendix B. The following result lets us bound (34),

Lemma 11. *Let $0 < x < y < 1$. For any nonnegative integers s and k with $s < \lfloor kx \rfloor$, it holds that*

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

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

Proof. See Appendix B.2.1. ∎

Therefore, from equation (34) and Lemma 11, we find that whenever $T > \left(\frac{2}{\eta_3-\eta_1}\right)^{1/\epsilon} =: T^*(\epsilon,\theta)$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(v_{2,t} \geq \eta_3,\ N_2(t) \geq L,\ S_2(t) < \lfloor N_2(t)\eta_1 \rfloor) \\
&\qquad\qquad \leq \sum_{t=1}^{T} \mathbb{P}\left((\eta_3-\eta_1)\exp\{N_2(t)d(\eta_1,\eta_3)\} \leq 2t,\ N_2(t) \geq L\right) \\
&\qquad\qquad \leq \sum_{t=1}^{T} \mathbb{P}\left((\eta_3-\eta_1)\exp\{Ld(\eta_1,\eta_3)\} \leq 2t\right) \\
&\qquad\qquad = \sum_{t=1}^{T} \mathbb{P}\left((\eta_3-\eta_1)T^{1+\epsilon} \leq 2t\right) = 0 \qquad (35)
\end{aligned}$$

All that is left is to bound the second term in (33), and to do so we apply the following Lemma whose proof is in Appendix B.2.2

Lemma 12. *There exist positive constants $C = C(\theta_2, \eta_1)$ and $x' > \theta_2$ such that*

$$\sum_{t=1}^{T} \mathbb{P}\left(S_2(t) \geq \lfloor N_2(t)\eta_1 \rfloor,\ \pi_t^{\mathrm{OG}} = 2\right) \leq K + \frac{1}{1-e^{-d(x',\theta_2)}}$$

Combining Lemma 12, (35), (33) and (34) shows the claim. ■

B.2.1. Proof of Lemma 11.

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

$$\begin{aligned}
\mathsf{E}\left[(V-y)^+\right] &= \frac{1}{B(s+1,k-s+1)}\int_y^1 (t-y)t^s(1-t)^{k-s}\,dt \\
&= \frac{1}{B(s+1,k-s+1)}\int_y^1 t^{s+1}(1-t)^{k-s}dt - y\mathbb{P}\left(V \geq y\right) \\
&= \frac{B(s+2,k-s+1)}{B(s+1,j-s+1)}\left(\frac{1}{B(s+2,k-s+1)}\right)\int_y^1 t^{s+1}(1-t)^{k-s}dt - y\mathbb{P}\left(V \geq y\right) \\
&= \frac{s+1}{k+2}F^B_{k+2,y}(s+1) - y\mathbb{P}\left(V \geq y\right) \qquad (36) \\
&\leq \frac{s+1}{k+2}F^B_{k+2,y}(s+1) \leq F^B_{k,y}(kx) \leq \exp\left\{-kd(x,y)\right\} \qquad (37)
\end{aligned}$$

where we use Fact 5 and the definition of the Beta CDF to establish equation (36). The final bound in (37) is the result of the Chernoff-Hoeffding theorem and Fact 7. Let $\delta := y - x$, and note that

$s < kx \implies s \leq \lfloor (k+1)x \rfloor$ due to s being integer, whence

$$
\begin{aligned}
\mathsf{E}\left[(y-V)^+\right] &= \mathsf{E}\left[(y-V)\mathbb{1}\left(V \leq y\right) \mid s,k\right] \\
&= \mathsf{E}\left[(y-V)\mathbb{1}\left(y-\delta \leq V \leq y\right) \mid s,k\right] + \mathsf{E}\left[(y-V)\mathbb{1}\left(V < y-\delta\right) \mid s,k\right] \\
&> \mathsf{E}\left[(y-V)\mathbb{1}\left(V < y-\delta\right) \mid s,k\right] \\
&\geq \delta\mathsf{E}\left[\mathbb{1}\left(V < y-\delta\right) \mid s,k\right] \qquad (38) \\
&= \delta\mathbb{P}\left(V < x \mid s\right) \\
&= \delta\left(1 - F^B_{k+1,x}(s)\right) \qquad (39) \\
&\geq \delta/2 \qquad (40)
\end{aligned}
$$

where equation (39) relies on Fact 5. The bound (40) is justified from Fact 6 and $s \leq \lfloor (k+1)x \rfloor$. Thus using the inequalities for both the numerator and denominator, we obtain the desired bound. ■

B.2.2. Proof of Lemma 12.

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,

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}\left(S_2(t) \geq \lfloor N_2(t)\eta_1 \rfloor,\ \pi_t^{\text{OG}} = 2\right) &\leq \mathsf{E}\left[\sum_{\ell=1}^{T} \sum_{t=\tau_\ell}^{\tau_{\ell+1}-1} \mathbb{1}\left(S_2(\ell) \geq \lfloor N_1(\ell)\eta_1 \rfloor\right) \mathbb{1}\left(\pi_t^{\text{OG}} = 2\right)\right] \\
&= \mathsf{E}\left[\sum_{\ell=1}^{T} \mathbb{1}\left(S_2(\tau_\ell) \geq \lfloor (\ell-1)\eta_1 \rfloor\right) \sum_{t=\tau_\ell}^{\tau_{\ell+1}-1} \mathbb{1}\left(\pi_t^{\text{OG}} = 2\right)\right] \\
&= \mathsf{E}\left[\sum_{\ell=0}^{T-1} \mathbb{1}\left(S_2(\tau_{\ell+1}) \geq \lfloor \ell\eta_1 \rfloor\right)\right] \\
&\leq N + \sum_{\ell=N+1}^{T-1} \mathbb{P}\left(S_2(\tau_{\ell+1}) \geq \ell\eta_1 \frac{\lfloor \ell\eta_1 \rfloor}{\ell\eta_1}\right) \\
&\leq N + \sum_{\ell=N+1}^{T-1} \mathbb{P}\left(S_2(\tau_{\ell+1}) \geq \ell x'\right) \\
&\leq N + \sum_{\ell=1}^{\infty} \exp(-\ell d(x', \theta_2)) \qquad (41) \\
&= N + \frac{1}{1 - e^{-d(x',\theta_2)}}
\end{aligned}
$$

■

The bound (41) follows from the Chernoff-Hoeffding theorem and that $S_2(\tau_{\ell+1}) \sim \text{Binomial}(N_1(\ell+1), \theta_2) \sim \text{Binomial}(\ell, \theta_2)$.