Bayesian Bandits, Secretaries, and Vanishing Computational Regret

Ashish Goel\(^*\)  Sudipto Guha\(^†\)  Kamesh Munagala\(^‡\)

Abstract

We consider the finite-horizon multi-armed bandit problem under the standard stochastic assumption of independent priors over the reward distributions of the arms. We define a new notion of computational regret against the Bayesian optimum solution instead of worst-case against the true underlying distributions. We show that when the priors of the arms satisfy a log-concavity condition, there is a simple index-type policy that achieves per-step computational regret $O \left( \frac{\log \log T}{\log T} \right)$ regardless of how the number $n$ of arms relates to the time horizon $T$. This shows that the regret vanishes as $T \to \infty$. As a corollary, we present an additive PTAS for this problem. This complements existing literature on worst-case regret bounds that only hold for the case when $n$ is much smaller than $T$. The log-concavity condition is widely used and is satisfied by Beta, Gaussian, and Uniform priors, which are the most common in these settings.

Our policy is far simpler to implement than the well-known Gittins index policy, which is also not optimal for the finite horizon case. We also give evidence that the log-concavity condition is necessary for the type of regret bounds we show.

Finally, we show that our results also extend to the related “budgeted learning” and the secretary problems.

\(^*\)Departments of Management Science and Engineering and (by courtesy) Computer Science, Stanford University. Email: ashishg@stanford.edu. Research supported by an NSF ITR grant, the Stanford-KAUST alliance for academic excellence, and gifts from Google, Microsoft, and Cisco.

\(^†\)Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia PA 19104-6389. Email: sudipto@cis.upenn.edu. Research supported in part by an Alfred P. Sloan Research Fellowship, an NSF CAREER Award, and NSF Award CCF-0644119.

\(^‡\)Department of Computer Science, Duke University, Durham NC 27708-0129. Research supported by an Alfred P. Sloan Research Fellowship, and by NSF via a CAREER award and grant CNS-0540347. Email: kamesh@cs.duke.edu.
1 Introduction

The celebrated multi-armed bandit problem (MAB) in decision theory models the central trade-off between exploration and exploitation, or in other words between learning about the state of a system and utilizing the system. In the simplest version of this problem, there are $n$ competing independent options, referred to as “arms,” yielding unknown rewards $\{r_i\} \in [0,1]$ drawn from a fixed underlying distribution $D_i$. Playing an arm yields a reward drawn from an underlying distribution, and the information from the reward observed partially resolves its distribution. The goal is to sequentially play the arms in order to maximize reward obtained over a time horizon $T$.

Depending on the kind of information available to the player, there are two settings in which this problem has been typically studied:

**Bayesian Setting:** In this setting, a prior distribution $D_i$ is specified over possible distributions $D_i$. As the arms are played, the observations resolve the prior into a sharper posterior distribution via Bayesian updates. The goal is to design a policy that specifies which arm to play given the current posterior distributions such that we maximize the expected reward over the horizon $T$. This is the *stochastic multi-armed bandit problem* [1, 6, 17, 21]. Standard dynamic programming solves this problem; however, the computation time and policy description are exponential in the number of arms, $n$. The key issue here is computational: Do succinct policies (with poly-size specification) with good performance exist? Can they be found in polynomial time?

**Adversarial Setting:** In this setting, the distributions $D_i$ are assumed to be arbitrary and unknown a priori. The goal is to design a polynomial time computable and executable policy that uses only the observed rewards to make the decision about which arm to play next. As before, the goal is to maximize the expected reward in the time horizon $T$. The key issue here is informational: How much does it hurt to not know the underlying distributions?

For MAB problems, it is typical to study performance in terms of *informational regret* [4, 2, 3]. This is the regret accrued by a (poly-bounded) strategy that is not aware of the underlying distribution or its prior (and hence works only with the observed rewards). In the adversarial setting, this is defined as the difference between the time-average per-step reward of the policy that knows the underlying distributions $D_i$ (and hence always plays the arm with largest $p_i = \mathbb{E}[D_i]$), and the time-average reward of the policy that does not have this information. A closely related notion is the *Bayes risk* [13], which is simply the expected informational regret, where the expectation is over the prior distributions of the arms.

In their celebrated result, Auer, Cesa-Bianchi, and Fisher [3] build on the work of Lai and Robbins [13] to show several intuitive greedy algorithms that achieve (near-tight) time-average informational regret of $O\left(\frac{n \log T}{\Delta T}\right)$, where $\Delta = p^{(1)} - p^{(2)}$ is the difference between the expected rewards of the best and second best arms. If $\Delta$ is arbitrarily small, this bound becomes $O\left(\sqrt{\frac{n \log T}{T}}\right)$. Similar (near-tight) bounds can be shown for the Bayes risk [13]. We note that for any fixed $n$ and any set of distributions $\{D_i\}$, the informational regret (resp. Bayes risk) vanishes as $T \to \infty$.

Informational regret (resp. Bayes risk) is the right concept when $n$ is held fixed, so that $T \gg n$. In several modern settings such as keyword search auctions, where the arms correspond to keywords that an advertiser can bid on [16], and web crawling applications [19], where the arms are possible web pages that can be refreshed, the number of arms $n$ and the time-horizon $T$ are
quite often comparable. In fact, there are often insufficient resources to play all the arms even once, so it is quite likely that $T \leq n$. Such problems have received a lot of attention recently; see for instance [12, 16, 18, 19]. In such scenarios, it is often not possible to have low regret against an algorithm that knows the underlying distributions upfront; in other words, the informational regret is necessarily large, and hence has little meaning. However, in many such settings, it is indeed possible to develop coarse prior models for the underlying reward processes. The relevant question in this context is whether there is a notion of performance measure against the best possible policy (or the Bayesian optimal solution), since even the best such policy will not have full knowledge and hindsight. In the Bayesian setting, the bottleneck is in computing the optimal policy in poly-time, and hence we ask: Is there a notion of computational regret in this setting, and does it have any additional benefit over existing measures?

Index Policies and related techniques used in practice. The Bayesian MAB problem has been widely studied in the discounted reward setting [6, 8, 20], where for a fixed parameter $\beta \in (0, 1)$, the objective of the policy to maximize the discounted reward, which is $\sum_{t=0}^{\infty} \beta^t R_t$, where $R_t$ is the expected reward $t$ steps in the future. For this version of the problem, the celebrated Gittins index policy [8] is optimal. This policy computes a number called the index separately for each possible posterior distribution of each arm, and at each step, plays the arm with the highest index. This has the advantage of having running time polynomial in $n$. However, the computation of the Gittins index for an arm with given prior requires solving an infinite horizon discounted dynamic program over the evolution of the prior of that arm. Therefore, a key issue with the Gittins index is that it not clear how to perform the index computation efficiently even for reasonable priors such as Gaussian densities, and typically, functional approximations are used in this context.

A natural approach to the finite horizon problem is to set $\beta = 1 - \frac{1}{T}$, and use the Gittins index policy; it is well-known [13] that though this policy is not optimal for the finite horizon version, it has vanishing Bayes risk when $n \ll T$ and $T \to \infty$. A natural question in this regard is the performance of the Gittins Index when $n \not\ll T$.

1.1 Our Results

Our contributions lie on three related directions: (i) We introduce a new conceptual comparison measure and show that even for $n \not\ll T$ this notion allows us to develop meaningful (and surprisingly good) algorithms for a broad class of priors; (ii) We show that small (but important) modification of well-known policies relate to this new algorithm and thereby provide the first analysis of these policies; and (iii) We show that this new analysis scheme provides the first additive PTASs for these problems (for this class of priors), improving upon the best known constant factor results.

Comparison Measures: Our main modeling contribution is in defining Computational Regret, which is the regret against the best policy that uses the same prior information as the algorithm, but that is not restricted by computational power. In the Bayesian setting, we define this as the difference between the expected time-average per-step reward of the optimal policy (or decision tree) and the time-average reward achievable by a policy with computation time and description complexity polynomial in the input size (i.e., the descriptions of $D_t$). In this definition, the expectation is over the priors $D_t$; in that sense, both the optimal policy and the constructed policy use the same prior information $D_t$, and the regret arises purely due to comparing an exponential size decision tree with one of polynomial complexity, and is different from informational regret that compares against the policy that knows the underlying distributions $D_t$. See Section 2 for a more
formal definition. A special case is the regret of a policy that is unaware of the prior, and we term this the *oblivious computational* regret.

Our main focus is to study this new notion of computational regret. We show that computational regret is indeed a relevant concept in several scenarios, including the example above where \( n \ll T \).

Our main result is the following surprisingly strong guarantee for the computational regret:

**Theorem 1.1** (Proved in section 3). Suppose the reward distributions \( D_i \in [0,1] \) for arm \( i \) are drawn from a prior \( D_i \). Let \( Q_i \) denote the resulting prior distribution of \( p_i = E[D_i] \) (which is also a random variable). If all \( Q_i \) have non-decreasing hazard rate (see Condition 2.1), then there is a policy with computational regret \( O \left( \frac{\log \log T}{\log T} \right) \). In other words, the computational regret is \( o(1) \).

In other words, the above results show that the computational regret with respect to the Bayesian optimal policy is \( o(1) \), and vanishes as \( n, T \to \infty \) regardless of how \( n \) scales with \( T \), and (for the most part), regardless of the priors of the arms. We note that the above is a non-trivial statement, since the priors themselves can depend on \( T \), and furthermore, our policy itself runs in linear time, so we are not “cheating” by scaling computational power.

We will overload terminology and term priors that satisfy the condition in the above theorem as *log-concave priors*. We note that log-concave distributions capture a wide class of priors \([5]\), which include Gaussians, Uniform, and Beta(\( a, b \)) priors\(^1\) (for \( a, b \geq 1 \)) that are widely used in these settings. For a general characterization, refer to Appendix B. Furthermore, the non-parametric maximum likelihood estimator of the density from \( n \) i.i.d. samples from a log-concave density converges to the true density fairly rapidly \([15, 22]\). Our main lemma (Lemma 3.1) presents a new probabilistic result for log-concave densities and hence, complements this line of work.

When the prior distributions are i.i.d., our policy is in fact oblivious to the priors!

**Corollary 1.2.** Under the conditions of Theorem 1.1, if the priors \( D_i \) are i.i.d., then there is an oblivious policy (i.e. one that ignores the priors) that has computational regret of \( O \left( \frac{\log \log T}{\log T} \right) \).

We also give evidence that our positive results are unlikely to generalize to all priors:

**Theorem 1.3** (Proved in appendix D). There exist \( D_i \) that are i.i.d. (but not necessarily log-concave), for which any oblivious policy has computational regret of \( \Omega(1) \).

---

**Allocational Policy**

\( Q_i \in [0,1] = \) Prior distribution on mean reward \( p_i \) of arm \( i \). \( F_i(x) = \Pr[Q_i \geq x] \).

**Define** the overflow index of arm \( i \) as: \( \gamma_i = F_i^{-1} \left( \frac{1}{T} \right) \).

1. Sort the arms in decreasing order of \( \gamma_i \).
2. Choose \( R \) to be the first \( \min(n, \frac{T}{\log T}) \) arms in this sorted order.
3. Explore each arm in \( R \) for \( \log^3 T \) time steps, for a total of \( \frac{T}{\log T} \) time steps.
4. Exploit the arm \( i^* \) with the highest observed sample average reward for the remaining time steps.

**Figure 1:** The Allocational Policy.

\(^1\)Informally, the Beta(\( a, b \)) prior corresponds to having observed \( a - 1 \) successes and \( b - 1 \) failures in \( a + b - 2 \) experiments.
Our Policy and its connection to Gittins Index: Our policy is allocational, and has three separate phases shown in Figure 1. This has structure much like an index policy, where the index for an arm is simply \( \gamma_i = F_i^{-1} \left( \frac{1}{T} \right) \). We term this the overflow index. Unlike the Gittins index, the overflow index is no harder to compute than drawing a random sample from the distribution over the expected reward, and can hence be computed efficiently for most well-behaved distributions. We further show in Section 3.3 that for log-concave priors and a discount factor \( \beta = 1 - \frac{1}{T} \), the Gittins index of the prior is within an additive \( O \left( \frac{\log \log T}{\log T} \right) \) of the overflow index, and hence, the Gittins index can be used instead of the overflow index in Figure 1.

We note that the prior distributions are only used in the subset selection phase, that too only if these priors are not i.i.d. We next note that a low informational regret scheme such as UCB1 can be used to improve the regret bound in Steps (2) and (3) however, a simpler scheme suffices in our case since the bottleneck in the regret bound turns out to be Step (1). Therefore, an alternate view of our results (especially Theorem 1.1) is a new Bayesian analysis of low adversarial regret policies such as UCB1 when \( n \) is comparable to \( T \).

Though the allocational policies are intuitive, there is no a priori reason to expect they will have vanishing computational regret. In fact, Theorem 4.1 shows that such policies do not always have vanishing computational regret against the Bayesian optimum even for general i.i.d. priors.

Other Applications: A closely related problem is the budgeted learning problem [10, 9], where the goal is to optimize the expected reward of the arm chosen at the \((T+1)\)st step, instead of the sum of the rewards in the first \( T \) steps. We can define computational regret analogously for this problem. As a simple corollary of Theorem 1.1, we show the following:

Corollary 1.4. Under the conditions on the priors in Theorem 1.1, there is a policy with computational regret \( O \left( \frac{\log \log T}{\log T} \right) \) for the budgeted learning problem.

Building on the work of Guha and Munagala [10] for budgeted learning, Goel, Khanna, and Null [9] show that the Gittins Index policy with discount factor \( \beta = 1 - \frac{1}{T} \), is also a constant factor approximation to the optimal reward in the finite horizon version regardless of \( n, T \). These are in fact, the best possible when there are side constraints, such as costs of switching between arms [11]. However, a \( O(1) \) guarantee does not imply vanishing regret against the Bayesian optimum. In contrast, our regret indeed vanishes as \( T \to \infty \) regardless of how \( n \) scales with \( T \), provided the underlying means follow log-concave prior densities and there are no side-constraints.

We also note that the exploration and exploitation phases can be interleaved to obtain a computational regret of \( O \left( \frac{\log \log T}{\log T} \right) \) simultaneously for all \( T \) (i.e. without advance knowledge of \( T \)).

Secretary Problem: Mahdian et al. [14] show that the hazard rate condition also implies a vanishing small sample size for the optimal aspiration strategy in the secretary problem. Our probability lemmas complement this result by showing fairly directly (Appendix A) that in addition, the optimal aspiration strategy has vanishing gap (or regret) against the optimal omniscient strategy.

2 Preliminaries

Formally, the problem we consider is the following. There is a bandit with \( n \) independent arms. The reward on playing arm \( i \) follows unknown distribution \( D_i \in [0,1] \). There is a prior distribution\(^2\)

\[ \text{Using UCB1 to replace our steps (2) and (3) will allow us to choose } \Theta(T/\log^3 T) \text{ arms as opposed to } \Theta(T/\log^5 T) \text{ as we currently do; however this does not improve our overall regret guarantee.} \]
\( D_i \) from which \( D_i \) is drawn; the Bayesian decision policy is aware of this prior. Let \( p_i = E[D_i] \); note that this is a random variable whose distribution \( Q_i \) can be derived from \( D_i \).

Any policy is a decision tree of depth \( T \), where the decision at time \( t \) about which arm to play depends on the observed rewards so far. Given the priors, the expected reward \( V(\mathcal{P}) \) of the policy \( \mathcal{P} \) is the expected total rewards from the plays in the decision tree, where the expectation is over the priors \( D_i \) and the distributions \( D_i \) drawn from these priors. Given the priors, each policy \( \mathcal{P} \) is associated with a unique expected reward. Let \( \mathcal{P}^* \) be the optimal reward policy, which can have exponential size in \( n, T \). Then, the computational regret of a poly-time computable and representable policy \( \mathcal{P} \) is simply \( \frac{1}{T} (V(\mathcal{P}^*) - V(\mathcal{P})) \).

Note that this is different from the minimum Bayes risk, which is \( E[\max_i p_i] - \frac{V(\mathcal{P}^*)}{T} \), i.e., the expected informational regret between the policy that knows the \( D_i \) and the Bayesian optimal policy \( \mathcal{P}^* \) that only knows the priors and not the \( D_i \). The Bayes risk does not vanish if \( n \gg T \) [3]. Our main contribution is to show that the computational regret indeed vanishes.

**Example:** Suppose \( T = 1 \). Then, the optimal policy is to choose the arm with the highest \( E[p_i] \) and play once. Therefore, \( V(\mathcal{P}^*) = \max_i E[p_i] \). Since this policy is poly-time computable, the computational regret is zero. On the other hand, the policy that knows the underlying \( D_i \) chooses the arm with highest \( \max_i E[D_i] = \max_i p_i \), and hence, the expected reward over the prior is \( E[\max_i p_i] \). The Bayes risk is therefore \( E[\max_i p_i] - \max_i E[p_i] \).

Our positive results holds for the following class of priors. We note that several well-known priors such as Uniform, Gaussian, Gamma, and Beta (Beta\((a, b)\) for \( a, b \geq 1 \)) satisfy this condition [5]. We show a general characterization of priors satisfying this condition in Appendix [3].

**Condition 2.1 (Non-decreasing Hazard Rate (IHR)).** For any prior \( D_i \) on distributions \( D_i \) over \([0, 1] \), the distribution \( Q_i \) that it defines over \( p_i = E[D_i] \). Let \( f(x) \) denote the density function of \( Q \), and \( F(x) = \Pr[Q \geq x] \). The condition we need is that \( F(x) \) has monotone hazard rate (IHR), i.e., \( \frac{f(x)}{F(x)} \) be monotonically non-decreasing in \( x \). This corresponds to \( \log F(x) \) being a concave function.

As is typical with most regret bounds, our results will crucially use Hoeffding’s inequality.

**Lemma 2.1 (Hoeffding’s Inequality).** Let \( X_1, X_2, \ldots, X_n \in [0, 1] \) be independent random variables, and let \( X = \sum_{i=1}^{n} X_i \). Then for any \( \delta > 0 \), we have: \( \Pr[|X - E[X]| > \delta] \leq 2e^{-2\delta^2/n} \).

### 3 Computational Regret: Upper Bounds for IHR Priors

Given priors \( D_i \) over possible distributions \( D_i \) for the arms \( i = 1, 2, \ldots, n \), let \( Q_i \) denote the distribution of the mean reward \( p_i = E[D_i] \) that is induced by \( D_i \); this satisfies IHR. We will prove Theorem 1.1, i.e., show that the allocational policy in Fig. 1 has computational regret \( O \left( \frac{\log \log T}{\log T} \right) \).

Let \( U \) denote the set of arms with \(|U| = n\), and let \( OPT \) denote the value (per step reward) of the optimal policy. Let \( F_i(x) = \Pr[Q_i \geq x] \). We define the index of the arm as \( \gamma_i = F_i^{-1} \left( \frac{i}{n} \right) \).

#### 3.1 Properties of IHR Densities

Our main result in this section is the following general probability lemma on the behavior of \( F_i^{-1}(y) \).
Lemma 3.1. For $Q$ satisfying IHR, let $F(x) = \Pr[Q \geq x]$. For any $0 \leq p < q \leq 1$, we have:

$$
\frac{F^{-1}(p) - F^{-1}(q)}{\log(1/p) - \log(1/q)} \leq \frac{F^{-1}(q) - F^{-1}(0)}{\log(1/q)}
$$

The bound is tight for $Q$ being the exponential distribution.

Proof. Let $x_1 = F^{-1}(p)$, and $x_2 = F^{-1}(q)$. We need to bound $\Delta = x_1 - x_2$. Let $h(x) = -\log F(x)$; this is a monotonically non-decreasing function. Further, the IHR condition implies this function is convex. Since $h(x_1) - h(x_2) = \log(1/p) - \log(1/q)$, and by the convexity of $h$:

$$
h(x_1) - h(x_2) \geq \Delta h'(x_2) \implies h'(x_2) \leq \frac{\log(1/p) - \log(1/q)}{\Delta}
$$

However, by the definition of $x_2$, we have $h(x_2) = \log(1/q)$. But, since $h$ is convex with $h(0) = 0$, the following completes the proof:

$$
h(x_2) \leq h(0) + x_2 h'(x_2) = F^{-1}(q) h'(x_2) \leq F^{-1}(q) \frac{\log(1/p) - \log(1/q)}{\Delta}
$$

$$
\implies \Delta = F^{-1}(p) - F^{-1}(q) \leq F^{-1}(q) \frac{\log(1/p) - \log(1/q)}{\log(1/q)}
$$

The lower bound follows by using $F^{-1}(x) = -\frac{\log x}{\lambda}$ i.e., the exponential distribution.

Corollary 3.2. Under the conditions of the above lemma, for $Q \in [0, 1]$, and for constants $r, s \geq 1$:

$$
F^{-1} \left( \frac{1}{T \log^s T} \right) - F^{-1} \left( \frac{\log^r T}{T} \right) \leq (r + s + o(1)) \frac{\log \log T}{\log T}
$$

Furthermore, there is a distribution $Q$ for which the above bound is tight to within a factor of $e$.

Proof. We use $p = \frac{1}{T \log^s T}$ and $q = \frac{\log^r T}{T}$ in the above lemma to show the upper bound. To show a matching lower bound, let $Q = \text{Beta}(1, \log T)$. We have: $F(x) = (1 - x)^{\log T}$. Therefore:

$$
F^{-1} \left( \frac{1}{T \log^s T} \right) - F^{-1} \left( \frac{\log^r T}{T} \right) = e^{-1+s \frac{\log \log T}{\log T}} - e^{-1-r \frac{\log \log T}{\log T}}
$$

$$
= \frac{r + s \log \log T}{e \log T} + O \left( \left( \frac{\log \log T}{\log T} \right)^2 \right) \geq \frac{r + s}{e + o(1)} \frac{\log \log T}{\log T}
$$

Here, the second equality follows from a Taylor series expansion of $e^{x-1}$ around $x = 0$. □

3.2 Analysis

Consider the policy with optimal expected reward (where the expectation is over $D_i$ drawn from prior distributions $D_i$) that is restricted to playing a set $S \subseteq U$ of arms. Let $J(S)$ denote the optimal per-step reward for this policy. We note that $OPT = J(U)$.

Consider the subset $R$ constructed in Figure[1]. Add an arm with deterministic reward $\gamma_m = \min_{j \in R} \gamma_j$ to this set, and call this new set $R^*$. We first have:

\[ \text{For densities such as Uniform}[a, b], \text{ the function } F(x) \text{ may not be differentiable at finitely many points; in this case, we can use any sub-gradient of } h(x) \text{ instead of } h'(x) \text{ without changing the proof.} \]
Lemma 3.3. $J(R^*) \geq J(U) - O\left(\frac{\log \log T}{T}\right) = OPT - O\left(\frac{\log \log T}{T}\right)$.

Proof. Let $L = U \setminus R$. The optimal policy can only play $T$ arms from set $L$ on any decision path, and these are independent. Furthermore, for any arm $j \in L$, we have $F_j^{-1}\left(\frac{1}{T}\right) \leq \gamma_m$. For some decision path $r$, let $S_r \subseteq L$ denote the subset of these arms that are played. For each of these arms $j \in S_r$, $\Pr\left[p_j \geq F_j^{-1}\left(\frac{1}{T \log T}\right)\right] = \frac{1}{T \log T}$. By union bounds, we have:

$$\Pr\left[\max_{i \in S_r} p_i \geq \max_{j \in L} F_j^{-1}\left(\frac{1}{T \log T}\right)\right] \leq \frac{1}{\log T}$$

Combining this with Corollary 3.2 and the condition that $\gamma_m \geq F_j^{-1}\left(\frac{1}{T}\right)$ for all $j \in S_r$, we have:

$$\Pr\left[\max_{i \in S_r} p_i \leq \gamma_m + O\left(\frac{\log \log T}{T \log T}\right)\right] \geq 1 - \frac{1}{\log T}$$

where the second inequality follows from Corollary 3.2. Since the arms are independent, the above inequality is true regardless of the decision path $r$ taken. Therefore, if the policy simply plays the deterministic arm with reward $\gamma_m$ whenever it had originally decided to play an arm from $L$, the per-step loss in reward will be at most $O\left(\frac{\log \log T}{T \log T}\right) = O\left(\frac{\log \log T}{T}\right)$. The new policy never plays an arm in $L$, and is therefore a policy over $R^*$, which completes the proof.

Definition 1. For any subset $S$ of arms, let $\Phi(S) = E[\max_{i \in S} p_i]$.

Lemma 3.4. For any subset $S$ of arms, $J(S) \leq \Phi(S)$.

Proof. For each realization of $p_i$ drawn according to $Q_i$, the maximum possible expected reward of the omniscient policy is $\max_{i \in S} p_i$. Therefore, $J(S) \leq E[\max_{i \in S} p_i] = \Phi(S)$.

Therefore, combining the above two lemmas, we have:

$$OPT \leq \Phi(R^*) + O\left(\frac{\log \log T}{T}\right) \quad (1)$$

The allocational policy in Figure 1 uses the set $R$ which is $R^*$ without the deterministic arm with reward $\gamma_m$. We finally have:

Theorem 3.5. The computational regret of the allocational policy in Figure 1 is $O\left(\frac{\log \log T}{T}\right)$ when the prior distribution over the mean reward satisfies the IHR condition.

Proof. We first show that the computational regret of the allocational policy is at most $OPT - \Phi(R) + O\left(\frac{1}{T}\right)$. In any realization of the underlying means, suppose arm $i \in R$ has mean $p_i$ drawn from distribution $Q_i$.

Let $\mu_i$ denote the sample average at the end of the explore phase for arm $i$. The policy chooses the arm $i^* = \arg\max_{i \in R} \mu_i$ for the exploit phase. Note that all these are random variables that depend on the outcome of the exploration phase. In the subsequent discussion, all expectations are over the prior on the $\{p_i\}$, and the outcome of the explore phase.
The arm $i^*$ is played $T \left(1 - \frac{1}{\log T}\right)$ steps in the exploit phase. Since $p_i \leq 1$, the regret is:

$$\text{Regret} = OPT - \left(1 - \frac{1}{\log T}\right) E[p_{i^*}] \leq OPT - E[p_{i^*}] + O\left(\frac{1}{\log T}\right)$$

Since $\mu_i$ is the average of $\log^4 T$ plays of arm $i$, using Hoeffding’s inequality, we have:

$$\Pr \left[|p_i - \mu_i| \geq \frac{1}{\log T}\right] \leq \frac{1}{T^3} \Rightarrow \Pr \left[\max_{i \in R} p_i \geq p_{i^*} + \frac{2}{\log T}\right] \leq \frac{1}{T} \Rightarrow E[p_{i^*}] \geq \Phi(R) - O\left(\frac{1}{\log T}\right)$$

The first two inequalities hold for any choice of $\{p_i\}$; the probability is over the outcomes of the exploration given these $\{p_i\}$. The final inequality holds since all $p_i \leq 1$ and since $\Phi(R) = E[\max_{i \in R} p_i]$. This shows that the regret is at most $OPT - \Phi(R) + O\left(\frac{1}{\log T}\right)$.

If $R = U$, then $\Phi(R) = \Phi(U) \geq OPT$, which completes the proof. Otherwise, $|R| \geq \frac{T}{\log^6 T}$.

For any $j \in R$, we have $\Pr \left[p_j < F_j^{-1}\left(\frac{\log^6 T}{T}\right)\right] = 1 - \frac{\log^6 T}{T}$. Since the $j \in R$ are independent, we have:

$$\Pr \left[\max_{j \in R} p_j < \min_{j \in R} F_j^{-1}\left(\frac{\log^6 T}{T}\right)\right] = \left(1 - \frac{\log^6 T}{T}\right)^{T \log^6 T} \leq e^{-\log T} = \frac{1}{T}$$

Using $\min_{j \in R} F_j^{-1}\left(\frac{1}{T}\right) = \gamma_m$, and combining this with Corollary 3.2, we have

$$\Pr \left[\max_{j \in R} p_j < \gamma_m - O\left(\frac{\log \log T}{\log T}\right)\right] = O\left(\frac{1}{T}\right)$$

Conditioned on the above event not happening, we have $\max_{j \in R} p_j \geq \max_{j \in R^*} p_j - O\left(\frac{\log \log T}{\log T}\right)$, since $R^*$ only has the extra deterministic arm with reward $\gamma_m$. Therefore, we have:

$$\Phi(R) \geq \Phi(R^*) - O\left(\frac{\log \log T}{\log T}\right) \geq OPT - O\left(\frac{\log \log T}{\log T}\right)$$

The final inequality follows from Equation (1). Since the computational regret of the allocational policy is $OPT - \Phi(R) + O\left(\frac{1}{\log T}\right)$, the above completes the proof.

**Related Results.** We show in Appendix C that our regret analysis is tight given our upper bound on $OPT$. The following corollaries are also immediate from the above theorem:

**Corollary 3.6.** If the priors of the arms are i.i.d. and the distribution of the mean satisfies IHR, the allocational policy is oblivious to the priors (i.e., same for all priors), and yields regret $O\left(\frac{\log \log T}{\log T}\right)$.

**Corollary 3.7.** The finite horizon MAB problem with IHR priors admits to an additive PTAS.

### 3.3 Connection to Gittins Index

The Gittins index [8][20] is known to be optimal for the infinite horizon discounted reward version of the problem. For discount factor $\beta \in [0,1)$, the discounted reward of a policy is defined as $\sum_{t=0}^{\infty} \beta^t R_t$, where $R_t$ is the expected reward at time $t$. The Gittins index is defined for a given arm $i$, an initial prior on the distribution of the reward, and discount factor $\beta$ as follows: Consider
the policy that only plays this arm; at any point of time, the policy has two choices: (1) Play the arm, gain the reward, but pay penalty $G$; or (2) Stop. Let $R(G)$ denote the value of that policy which maximizes the infinite horizon discounted reward minus penalty. As $G$ increases, $R(G)$ is non-increasing. The Gittins index is the largest $G$ for which $R(G) > 0$. There are other equivalent definitions, but we will find this one the most convenient to use.

We note that this index depends on the prior as well as the discount factor $\beta$. We also note that by definition, $G \geq \mu$ for all $\beta \in [0,1)$, where $\mu$ is the expected mean of the prior. We now show that if the prior distribution of the underlying mean reward follows IHR, the Gittins index with discount factor $\beta = 1 - 1/T$ is well-approximated by the overflow index, $F^{-1}(\frac{1}{T})$.

The theorem below makes a tight connection between the overflow index we define, and the widely used Gittins index in this context. Since the overflow index is much simpler to compute than the Gittins index, this makes a case for preferring the overflow index to the Gittins index (with discount factor $\beta = 1 - \frac{1}{T}$) when the time horizon $T$ and the number of arms $n$ are large.

**Theorem 3.8.** For an arm with prior having IHR mean distribution $Q \in [0,1]$, let $F(x) = \Pr[Q \geq x]$, and $\mu = \mathbb{E}[Q]$. Let $G$ be the Gittins index of the arm for discount factor $\beta = 1 - 1/T$. Then,

$$|G - \max \left(\mu, F^{-1}\left(\frac{1}{T}\right)\right)| = O\left(\frac{\log \log T}{\log T}\right)$$

**Proof.** We will first show that $G \leq \max \left(\mu + \frac{1}{\log T}, F^{-1}\left(\frac{1}{T \log T}\right)\right)$. In the (hypothetical) best case, the distribution of the arm’s reward resolves completely with one play (for which it yields net reward $\mu - G$). In this case, to maximize net reward, the arm is subsequently played only if $Q \geq G$. If $G = F^{-1}\left(\frac{1}{T \log T}\right)$, then this case happens w.p. $\frac{1}{T \log T}$. Further, in this case, the maximum possibly per step net reward is 1, which yields a maximum discounted reward of $1 + \beta + \beta^2 + \cdots = T$. Therefore, the maximum possible net reward for the setting $G = F^{-1}\left(\frac{1}{T \log T}\right)$ is:

$$\text{Reward} = \mu - G + \frac{T}{T \log T} = \mu + \frac{1}{\log T} - G$$

If the Gittins index is at least $F^{-1}\left(\frac{1}{T \log T}\right)$, then the above quantity is positive. But this implies $G \leq \mu + \frac{1}{\log T}$. Therefore, $G \leq \max \left(\mu + \frac{1}{\log T}, F^{-1}\left(\frac{1}{T \log T}\right)\right)$.

Next, it is straightforward from the definition of $G$ that $G \geq \mu$. We will next show that $G \geq F^{-1}\left(\frac{\log^8 T}{T}\right) - \frac{1}{\log T}$.

To see this, set $G = F^{-1}\left(\frac{\log^8 T}{T}\right) - \frac{1}{\log T}$. We will demonstrate a policy with positive reward. Consider the following two-phase policy: In the first phase, play the arm for $T \log T$ steps. The true mean is at least $F^{-1}\left(\frac{\log^8 T}{T}\right)$ w.p. $\frac{\log^7 T}{T}$. Therefore, by Hoeffding’s inequality, the sample mean is at least $F^{-1}\left(\frac{\log^8 T}{T}\right) - \frac{1}{\log^2 T}$ with probability $\frac{\log^7 T}{T}$. In this case, play the arm forever in the second phase; else it stops playing. In the former case, the net reward of each play in the second phase is at least:

$$Q - G \geq \left(F^{-1}\left(\frac{\log^8 T}{T}\right) - \frac{1}{\log^2 T}\right) - \left(F^{-1}\left(\frac{\log^8 T}{T}\right) - \frac{1}{\log T}\right) \geq \frac{1}{2 \log T}$$

Since this reward is discounted starting at $t = \log^5 T$, the total discounted reward for $\beta = 1 - \frac{1}{T}$ in the second phase is at least: $\frac{1}{2 \log T} \cdot \frac{T}{2}$. Since this event happens w.p. $\frac{\log^7 T}{T}$, the expected discounted
reward of the second phase is at least \( \log^6 T \). The expected net reward of the first phase is at least \((\mu - G) \log^5 T \geq -\log^5 T\). Therefore, the expected net reward of both phases is at least:

\[
\text{Reward} \geq (\mu - G) \log^5 T + \frac{\log^7 T}{T} \frac{1}{2\log^2 T} \geq \frac{\log^6 T}{4} - \log^5 T \geq 0
\]

The above holds for \( T = \omega(1) \). Therefore, \( G \geq \max \left( \mu, F^{-1} \left( \log^{8/2} \frac{T}{T} \right) - \frac{1}{\log^2 T} \right) \). We now combine these bounds with Corollary 3.2 to complete the proof.

One consequence of the above theorem is the following, whose proof is simple and hence omitted:

**Corollary 3.9.** Consider the allocational policy in Figure 1, where in Step (1), the arms are sorted in decreasing order of \( G_i \), the Gittins index with discount factor \( 1 - 1/T \) (instead of by \( \gamma_i = F_i^{-1} \left( \frac{T}{T} \right) \)). This policy also has computational regret \( O \left( \frac{\log \log^2 T}{\log T} \right) \).

### 3.4 Budgeted Learning

In this problem [10, 9], which is also defined over a time horizon \( T \), the policy has two options at any point in time: (1) Pick an arm and play it; and (2) Stop and choose an arm. The policy only obtains a reward on executing Step (2). In this case, suppose the arm chosen by the policy has posterior distribution \( \tilde{Q}_i \) on the mean; then the reward obtained is \( E[\tilde{Q}_i] \). The goal is to design a policy whose expected reward is maximized. An alternate way of viewing this problem is that the goal is to maximize the expected posterior reward at the \((T + 1)^{st}\) step, instead of the total reward over the first \( T \) steps.

We note that the regret bound of \( O \left( \frac{\log \log^2 T}{\log T} \right) \) as well as the additive PTAS extend in a straightforward fashion to the budgeted learning problem. The only difference is that since the goal is to maximize the expected posterior reward at the \((T + 1)^{st}\) step, the policy in Figure 1 is modified to omit the exploit phase; the explore phase uses \( \frac{T}{\log^4 T} \) arms for \( \log^4 T \) steps each; and the arm chosen at the \((T + 1)^{th}\) step is the arm with the highest sample mean. The proof follows the same outline as for the finite horizon bandit problem, and details are omitted.

### 4 Lower Bounds for Non-IHR Priors

We first show an example of priors not satisfying the IHR condition for which any allocational policy (Fig. 1 for any choice of \( R \) and any choice of the number of plays per arm in Step (2)) does not have vanishing regret even when the policy is allowed to use the prior. The proof of the following theorem is deferred to appendix D.

**Theorem 4.1.** Suppose the underlying distributions are \( D_i \sim \text{Bernoulli}(1, p_i) \), where \( p_i \sim Q \). Then there exists \( Q \) for which any allocational policy has regret \( \Omega(1) \) against the Bayesian optimum.

**Bound on Oblivious Computational Regret:** We now give further evidence that our results are specific to distributions satisfying IHR. In particular, we prove the following theorem in appendix D.

**Theorem 4.2 (Theorem 1.3).** There exists a prior distribution \( D \) for which any oblivious policy (one that does not use prior information) has computational regret \( \Omega(1) \) when the distributions \( D_i \) are drawn i.i.d. from \( D \).
References


A Secretary Problem: Comparison to the Omniscient Strategy

In the secretary problem [7], n values are drawn i.i.d. from an unknown distribution \( Q \in [0,1] \) and presented sequentially. For each value presented, the policy can choose that value and stop, or reject that value and continue. The goal is to find the policy that maximizes the expected value chosen.

An aspiration strategy works as follows: In the first phase, observe the first \( n_0 \) values and find their maximum (say \( x \)). In the second phase, choose the first value presented beyond this point whose value exceeds \( x \); if the values are exhausted, choose the last value and stop.

It is well-known that if the distribution \( Q \) is arbitrary, the optimal aspiration strategy is to set \( n_0 = n/e \), which yields a probability \( 1/e \) of choosing the maximum value [7]. It was shown recently [14] that if \( Q \) satisfies IHR (see Condition 2.1), setting \( n_0 = n/\log n \) maximizes the expected value among the class of aspiration strategy. However, this work does not address the question of how good are aspiration strategies compared to the optimal, possibly omniscient, strategy? The best known result in this context is an \( e \)-approximation [7] by choosing \( n_0 = n/e \).

We show that the aspiration strategy with \( n_0 = n/\log^r n \) for some constant \( r \geq 1 \) has expected value at most \( O \left( \log \log n \log n \right) \) lower than the optimal omniscient strategy. This omniscient strategy can know \( Q \) and can choose the best value with the foreknowledge of all the values that will be presented. Let the value of this strategy be \( OPT \).

**Theorem A.1.** Let \( OPT \) denote the value of the optimal omniscient strategy. Under the hazard rate condition (see Condition 2.1), the aspiration strategy with \( n_0 = n/\log^r n \) for some constant \( r \geq 1 \) has expected value at least \( OPT - O \left( \log \log n \log n \right) \).

**Proof.** Let \( X_1, X_2, \ldots, X_n \) denote the random variables corresponding to the values presented. Let \( \Phi(k) = \mathbb{E}[\max_{i=1}^k X_i] \). Then, \( OPT \leq \Phi(n) \). Now let \( F(x) = \Pr[Q \geq x] \). Then,

\[
\Pr \left[ X_i \geq F^{-1} \left( \frac{1}{n \log n} \right) \right] = \frac{1}{n \log n}
\]

Therefore, by union bounds, w.p. \( 1 - \frac{1}{\log n} \), we have \( \max_{i=1}^n X_i \leq F^{-1} \left( \frac{1}{n \log n} \right) \).
Now consider the aspiration strategy for \( r = 1 \). With probability at least \( 1 - \frac{1}{\log n} \), the strategy finds a value in the second phase which is larger than the maximum observed in the first phase. Furthermore, using the same proof as Theorem 3.5 it is easy to see that with probability \( 1 - \frac{1}{T} \), we have \( \max_{i=1}^{n} X_i \geq F^{-1}\left(\frac{\log n}{n}\right) \).

Therefore, by union bounds, with probability \( 1 - \frac{2}{\log n} \), the value chosen by the aspiration strategy is at least \( F^{-1}\left(\frac{\log n}{n}\right) \). From above, with probability \( 1 - \frac{1}{\log n} \), the optimal value is at most \( F^{-1}\left(\frac{\log n}{\log n}\right) \). Again using union bounds and combining with Corollary 3.2, we have that the difference between the optimal value and the value chosen by the aspiration strategy is at most \( O\left(\frac{\log \log n}{\log n}\right) \). With the remaining probability, the difference is at most 1; this completes the proof for \( r = 1 \). The proof for any constant \( r \geq 1 \) is similar.

\[ \square \]

### B Exponential Families and the IHR Condition

For any random variable \( X \) with continuous density function \( q(x) \), let \( Q(x) = \Pr[X \geq x] \). Then, \( X \) is said to have a log-concave density if \( \log q(x) \) is a concave function of \( x \).

**Lemma B.1** (Refer [5]). If \( X \) has a log-concave density, then \( Q(x) \) is log-concave as well.

#### B.1 Exponential Families and Conjugate Priors

In the context of Bayesian optimization, it is typical to assume the underlying distribution of an arm follows a parametrized family, in particular, the single-parameter exponential family. This is the context in which the regret bounds for the bandit problem were originally formulated by Lai and Robbins [13]. The single-parameter exponential family corresponds to the class of density functions \( f(x; \theta) \), one for each parameter \( \theta \geq 0 \), where:

\[
f(x; \theta) = h(x)e^{\theta x - A(\theta)} \quad \text{where} \quad A(\theta) = \log \int_{x=1}^{1} h(x)e^{\theta x}dx
\]

Note that an exponential family parametrized by \( \theta \) is completely specified by the choice of the function \( h(x) \) in its density. We note that several well-known distributions, Bernoulli, Binomial, Poisson, Gaussian, and Gamma belong to the single-parameter exponential family, where the parameters are fixed functions of the mean of the distribution. For instance, for Bernoulli(1, \( p \)), we have \( \theta = \log \frac{p}{1-p} \), and for Gaussian(\( \mu, \sigma^2 \)) with known \( \sigma \), we have \( \theta = \mu \).

Let \( X_\theta \) denote the underlying random variable, and let \( G(\theta) = \int_{x=0}^{1} xf(x; \theta)dx \) denote the mean of the density. Let \( M_k(\theta) \) denote the \( k^{th} \) cumulant of \( X_\theta \), where the first cumulant is the mean, the second the variance, and so on. The following properties of these densities follow by differentiation:

1. \( A(\theta) \) is a convex function of \( \theta \). In particular, the \( k^{th} \) derivative of \( A(\theta) \) is precisely \( M_k(\theta) \).
2. The previous property implies \( X_\theta \) has a log-concave density. In other words, \( \log f(x; \theta) \) is concave in \( x \) for all \( \theta \geq 0 \).
3. \( G(\theta) \) is a non-decreasing function of \( \theta \). To see this, observe that \( G'(\theta) = M_2(\theta) \geq 0 \).
Conjugate Priors. The reason exponential families are widely studied in Bayesian statistics is that a *conjugate* prior can be defined over the parameter \( \theta \). This prior over \( \theta \) is specified by two parameters \( a \) and \( b \). The density \( r_{a,b}(\theta) \) is given by:

\[
r_{a,b}(\theta) \propto e^{a\theta - bA(\theta)} \quad \text{where} \quad b > 0
\]

The nice aspect of a conjugate prior is that the posterior on the parameter \( \theta \) after a sequence of observations from the underlying distribution \( X_\theta \) also belongs to the same class of priors, albeit with different parameters \( a',b' \).

B.2 Distribution of the Mean

Our goal will be to derive conditions on the exponential family and the conjugate so that the density of the random variable \( Q = \mathbb{E}[X_\theta] \) is log-concave, and hence via Lemma B.1 satisfies Condition 2.1. This will imply vanishing computational regret for the allocational policy in Figure 1. Note that the experiment corresponding to choosing \( Q \) is as follows: Draw \( \theta \) according to \( r_{a,b}(\theta) \), and set \( Q = G(\theta) \).

We assume that \( G(\theta) \) is a strictly increasing and smooth function of \( \theta \). In that case the function \( G^{-1} = H \) is well-defined and smooth. We will denote \( H(\mu) = \theta \) if \( G(\theta) = \mu \). We have \( H'(\mu) = \frac{1}{\sigma^2(\theta)} = \frac{1}{M^2(\theta)} \).

The distribution of the mean \( \mu \) follows the density:

\[
s_{a,b}(\mu) = r_{a,b}(H(\mu)) H'(\mu) = \frac{r_{a,b}(\theta)}{M^2(\theta)} \quad \text{where} \quad \theta = H(\mu)
\]

Recall that an *exponential family* parametrized by \( \theta \) is completely specified by the choice of the function \( h(x) \) in its density. For several well-known choices of \( h(x) \), the density \( s_{a,b}(\mu) \) is indeed log-concave and hence satisfies Condition 2.1 via Lemma B.1. We classify based on the underlying density; note that scaling and restricting the density to be on \([0,1] \) does not change log-concavity):

**Gaussian**\( (\mu,\sigma^2) \) with known \( \sigma \) and unknown \( \mu \), the conjugate prior yields a Gaussian distribution on \( \mu \), which is log-concave.

**Bernoulli**\( (1,p) \) and **Binomial**\( (n,p) \): The conjugate prior yields density Beta\( (a,b) \) on the mean \( p \), which is log-concave for \( a, b \geq 1 \).

**Poisson**\( (\lambda) \): The conjugate prior yields density Gamma\( (\alpha,\beta) \) on the mean \( \lambda \), which is log-concave for \( \alpha \geq 1 \).

B.2.1 Some General Characterizations

We are interested in conditions on the exponential family under which the distribution of the mean under the prior is log-concave, so that by Lemma B.1 the family satisfies Condition 2.1.

We now show sufficient conditions for this to happen. The first condition holds for Gaussian densities with unknown mean \( \theta \) and known variance, where the conjugate prior is also Gaussian; and for Exponential densities where the conjugate prior is also exponential. Here, \( G(\theta) = \theta \).

**Lemma B.2.** If \( H \) is concave, and \( H' \) is log-concave, then \( s_{a,b}(\mu) \) is log-concave. In particular, if \( G \) (resp. \( H \)) is linear, then \( s_{a,b}(\mu) \) is log-concave.
Proof. First, $G$ (resp. $H$) is an increasing function. Next, since $A(\theta)$ is convex, the function $r_{a,b}(\theta)$ is log-concave for any $b > 0$. This implies that for any increasing concave function $H$, the function $r_{a,b}(H(\mu))$ is log-concave \cite{5}. Next, note that $H'$ is log-concave. Since the product of log-concave functions is log-concave, this implies $s_{a,b}(\mu) = r_{a,b}(H(\mu))H'(\mu)$ is log-concave. Note that if $H$ is linear, $H'$ is a constant, and is satisfies the conditions implied by the lemma.

We now show a more involved sufficient condition based on the cumulants of $X_\theta$. Let $\mu, M_2, M_3$ denote the mean, variance, and third cumulant of $X_\theta$.

**Theorem B.3.** The prior density $s_{a,b}(\mu)$ is log-concave (and hence the exponential family satisfies Condition 2.1) if for the random variable $X_\theta$, the function:

$$C(\mu) = \frac{a - b\mu}{M_2} - \frac{M_3}{M_2^2}$$

is a non-increasing function of $\mu$ (resp. $\theta$).

Proof. We have:

$$\log s_{a,b}(\mu) = \log r_{a,b}(H(\mu)) + \log H'(\mu) = aH(\mu) - bA(H(\mu)) + \log H'(\mu)$$

Note now that:

$$H'(\mu) = 1/G'(\theta) = 1/M_2 \quad \text{and} \quad \frac{d}{d\mu}A(H(\mu)) = A'(H(\mu))H'(\mu) = \frac{\mu}{M_2}$$

Further,

$$\frac{H''(\mu)}{H'(\mu)} = G'(\theta) \frac{d}{d\theta} \frac{1}{G'(\theta)} \frac{d\theta}{d\mu} = -G'(\theta) \frac{G''(\theta)}{(G'(\theta))^2} H'(\mu) = - \frac{G''(\theta)}{(G'(\theta))^2} = - \frac{M_3}{M_2^2}$$

Therefore, the derivative of $\log s_{a,b}(\mu)$ w.r.t $\mu$ is $C(\mu) = \frac{a - b\mu}{M_2} - \frac{M_3}{M_2^2}$. If $\log s_{a,b}(\mu)$ is concave, the latter must be decreasing as a function of $\mu$. Since $\mu$ is monotone in $\theta$, this is equivalent to insisting it is a monotonically decreasing function of $\theta$. 

We note that for the density Bernoulli(1, $p$), we have $\mu = p$, $M_2 = p(1 - p)$ and $M_3 = p(1 - p)(1 - 2p)$. Therefore, $C(p) = \frac{a - 1}{p} - \frac{b - a - 1}{1 - p}$, so that density of the mean $p$ according to the conjugate prior is log-concave when $a \geq 1$ and $c = b - a \geq 1$, which corresponds to the prior Beta($a, c$) for $a, c \geq 1$.

Next note that for the Poisson($\lambda$) density, we have $\mu = M_2 = M_3 = \lambda$. Therefore: $C(\lambda) = \frac{a - 1}{\lambda} - b$, which is decreasing in $\lambda$ for $a \geq 1$. This corresponds to the Gamma($a$, $\beta$) prior on $\lambda$, which is log-concave for $a \geq 1$.

### C Tightness of Analysis in Section 3

We now give evidence that our analytic method in Section 3 is tight in the sense that against the upper bound on $OPT$ that we use, our regret bound is likely best possible for any policy. Therefore, improving the regret bound would need a different upper bound on $OPT$. 

15
Consider the case of i.i.d. priors. Let $Q_i$ denote the distribution of the mean of arm $i$. Let $\Phi(k) = E[\max_{i=1}^k Q_i]$. In this case, the upper bound on $OPT$ that our analysis uses is $OPT = \min(\Phi(n), \Phi(T))$. Suppose $n \geq T$. Suppose $Q_i \sim \text{Beta}(1, c)$ for $c \geq 2$ to be chosen later. In this case, denoting the Beta function by $B$, we have:

\[
\Phi(T) = \int_{x=0}^{1} \left(1 - (1 - (1 - x)^c)^T\right) dx = 1 - \frac{1}{c} \int_{y=0}^{1} (1 - y)^T y^{1/c - 1} dy
\]

\[
= 1 - \frac{1}{c} B\left(T + 1, \frac{1}{c}\right)
\]

The regret we incur is at least $\Phi(T) - \Phi\left(\frac{T}{\log T}\right)$. From above, this is equal to:

\[
\text{Regret} \geq \frac{1}{c} \left(B\left(T, \frac{1}{c}\right) - B\left(T + 1, \frac{1}{c}\right)\right)
\]

Ignoring low order terms, we now use the Stirling approximation to the Beta function as $B(n, x) \approx \Gamma(x)n^{-x}$, where $x \leq 1$, $n$ is large, and $\Gamma$ is the Gamma function. We have for $c = \log T$:

\[
\text{Regret} \geq \frac{1}{c} \Gamma\left(\frac{1}{c}\right) \left(e^{-\frac{\log T - \log \log T}{\log T}} - e^{-\frac{\log T}{\log T}}\right)
\]

\[
= \frac{1}{e} \Gamma\left(1 + \frac{1}{c}\right) \left(e^{-\frac{\log T}{\log T}} - 1\right) = \Omega\left(\frac{\log T}{\log T}\right)
\]

In the above, we have used that $\Gamma(x) \geq 1/2$ for $x \geq 0$, and that $e^x - 1 = x + o(x)$ for $x = o(1)$.

### D Proofs from Section 4

**Proof of Theorem 4.1:** The prior distribution $Q$ over the $p_i$ is as follows: $\Pr[Q = 1] = \frac{(\log \log T)^2}{T}$; $\Pr[Q = 1 - \epsilon] = \frac{1}{\sqrt{T}}$, where $\epsilon > 0$ is a constant; and $Q = 0$ with the remaining probability.

In the discussion below, we will use “high probability” to mean probability $1 - o(1)$. The following non-allocational policy has reward $1 - o(1)$. Choose a set $S$ of $\frac{T}{\log \log T}$ of these arms. With high probability, there is at least one arm with $p_i = 1$, and the number of arms with $p_i = 1 - \epsilon$ is between $\frac{\sqrt{T}}{\log T}$ and $\sqrt{T} \log T$. Play each arm in $S$ once. Discard all arms with observed reward 0. With high probability, there are at most $\sqrt{T} \log T$ arms that remain. Play these for $\log^3 T$ steps each; with high probability, all arms with $p_i = 1 - \epsilon$ will have at least one observed reward 0, and can be discarded. Play any of the remaining arms (which will have $p_i = 1$ with high probability) for the remaining $T \left(1 - O\left(\frac{1}{\log \log T}\right)\right)$ time steps.

Now, any allocational policy needs to play at least $\frac{T}{(\log \log T)^3}$ arms in the exploration phase to have a constant probability of choosing any arm with $p_i = 1$. Such a policy can therefore play the arms at most $(\log \log T)^3$ time steps in the exploration phase. For any arm with $p_i = 1 - \epsilon$, the probability of observing $(\log \log T)^3$ consecutive non-zero rewards is $\epsilon^{(\log \log T)^3}$, so that with high probability, $\Omega(T^{1/4})$ of these arms are indistinguishable from the $O((\log \log T)^3)$ arms with $p_i = 1$. Therefore, with high probability, the arm chosen for exploitation must have $p_i = 1 - \epsilon$, which implies a constant regret against the non-allocational policy described above. □
Proof of Theorem 1.3: For small constant $\epsilon > 0$, let $q = \epsilon T$. The prior $\mathcal{D}$ is defined over $2q$ reward distributions $X_1, X_2, \ldots, X_q$ and $Y_1, Y_2, \ldots, Y_q$. The distributions are defined over $q$ values $0 < a_1 < a_2 < \cdots < a_q = O(\epsilon^{-T})$. Reward distribution $X_i$ takes value $a_i$ w.p. $1 - \epsilon$ and 1 w.p. $\epsilon$, whereas distribution $Y_i$ takes value $a_i$ w.p. 1. There are $q$ possible priors $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_q$. Prior $\mathcal{D}_i$ has $\Pr[Y_1] = \Pr[Y_2] = \cdots = \Pr[Y_{i-1}] = \Pr[Y_{i+1}] = \cdots = \Pr[Y_q] = \frac{1}{q}$, and $\Pr[X_i] = \frac{1}{q}$. The adversary chooses any one of these priors uniformly at random, and the reward distributions of the $n$ arms are drawn i.i.d. from this prior.

In the ensuing discussion, we use w.h.p. to mean probability $1 - o(1)$. Consider the Bayesian optimum policy that knows the prior $\mathcal{D}_i$ from which the reward distributions have been drawn. This policy will play each arm once, until it finds an arm whose distribution is $X_i$ (this arm yields reward 1 or $a_i$), and plays this for the remaining time steps. Since $\Pr[X_i] = \frac{1}{q}$, the time $\eta$ to find this arm follows distribution Geometric($\frac{1}{q}$). Since $\mathbb{E}[X_i] = \epsilon$, this implies a per-step reward of:

$$\epsilon \left(1 - \mathbb{E}\left[\frac{\min(\eta, T)}{T}\right]\right) = \epsilon(1 - O(\epsilon))$$

Now, an oblivious policy is unaware of the prior $\mathcal{D}_i$. Even if it is given $q = \epsilon T$ arms with distinct reward distributions, it is unaware of the distinction between $X_i$ and $Y_i$ for any $i$. Therefore, such a policy is forced to play all these arms in a round-robin fashion until it observes a 1, and then stick to this arm. Since any arm with distribution $X_i$ yields a 1 w.p. $\epsilon$ and the $i \in \{1, 2, \ldots, q\}$ for which $X_i$ instead of $Y_i$ is used as reward distribution is chosen at random, the expected time $\gamma$ to observe a 1 follows distribution Geometric($\frac{1}{T}$). Therefore, the expected reward of such a policy is:

$$\epsilon \left(1 - \mathbb{E}\left[\frac{\min(\gamma, T)}{T}\right]\right) \leq \frac{\epsilon}{e}$$

Therefore, the computational regret of any oblivious strategy is $\Omega(\epsilon) = \Omega(1)$. \qed

17