JMLR: Workshop and Conference Proceedings vol 35:1–22, 2014

Stochastic Regret Minimization via Thompson Sampling[*]

Sudipto Guha SUDIPTO@CIS.UPENN.EDU
Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA.

Kamesh Munagala KAMESH@CS.DUKE.EDU
Department of Computer Science, Duke University, Durham, NC 27708-0129.

Abstract

The Thompson Sampling (TS) policy is a widely implemented algorithm for the stochastic multi-armed bandit (MAB) problem. Given a prior distribution over possible parameter settings of the underlying reward distributions of the arms, at each time instant, the policy plays an arm with probability equal to the probability that this arm has largest mean reward conditioned on the current posterior distributions of the arms. This policy generalizes the celebrated "probability matching" heuristic which has been experimentally and widely observed in human decision making. However, despite its ubiquity, the Thompson Sampling policy is poorly understood.

Our goal in this paper is to make progress towards understanding the empirical success of this policy. We proceed using the lens of approximation algorithms and problem definitions from stochastic optimization. We focus on an objective function termed *stochastic regret* that captures the expected number of times the policy plays an arm that is not the eventual best arm, where the expectation is over the prior distribution. Given such a definition, we show that TS is a 2–approximation to the optimal decision policy in two extreme but canonical scenarios. One such scenario is the two-armed bandit problem which is used as a calibration point in all bandit literature. The second scenario is stochastic optimization where the outcome of a random variable is revealed in a single play to a high or low deterministic value. We show that the 2 approximation is tight in both these scenarios. We provide an uniform analysis framework that in theory is capable of proving our conjecture that the TS policy is a 2–approximation to the optimal decision policy for minimizing stochastic regret, for any prior distribution and any time horizon.

Keywords: Thompson Sampling; Multi-armed bandits; Stochastic optimization

1. Introduction

In this paper, we consider the decision theoretic problem of iteratively choosing among competing options when faced with uncertainty about these. In a celebrated set of (widely replicated) human experiments surveyed in Vulkan (1992), it has been shown that humans follow a policy termed as *probability matching* when faced with competing options. More concretely, suppose there are K colors. At each step, one color i is chosen in an $i.i.d.$ fashion with probability p_i (so that $\sum_{i=1}^{K} p_i = 1$) and shown to a subject. These probabilities can be assumed to be known to the subject. The subject is supposed to guess the color before it is shown, and his goal is to *minimize expected number of mistakes made*. Clearly, from the viewpoint of stochastic optimization, this is a one-shot decision problem, and the optimal policy is to always guess color $i^* = \text{argmax}_i p_i$, and the expected number of mistakes per step is $1 - p_{i^*}$. However the human subjects appear to follow a

[*] Sudipto Guha is supported by NSF award CCF-1117216. Kamesh Munagala is supported by an award from Cisco, and by NSF grants CCF-0745761, IIS-0964560, and CCF-1348696.

different policy – at each step, they guess the color i with probability p_i. This "probability matching" policy matches the probability that an option is correct (p_i for option i) with the probability of choosing that option. Such a policy clearly does not optimize the expected number of mistakes. However, the expected number of mistakes is $\sum_i p_i(1-p_i) \le 2(1-p_{i^*})$, and this bound is tight. Therefore, regardless of K, probability matching is a 2–approximation to the expected number of mistakes. There are several reasons proposed for why humans employ this policy – however that is not the concern of this paper.

Our focus in this paper is a related decision policy termed *Thompson Sampling* (TS) that is widely implemented for the *multi-armed bandit* (MAB) problem. In the MAB problem, there are competing options (henceforth stated as arms) with underlying reward distributions, and the policy has a belief (or prior distribution) over the parameters of these distributions. At each step, the policy chooses (or plays) one arm, obtains its reward, and this play refines the prior via Bayes' rule. The policy executes over a horizon of T steps with the goal of maximizing expected reward. The Thompson sampling policy generalizes probability matching in the natural way – at each step t, p_{it} is the probability arm i has largest expected reward conditioned on the current prior, and the policy plays arm i with probability p_{it}.

Thompson Sampling is widely used in practical MAB settings Chapelle and Li (2011); Li et al. (2010); Dudík et al. (2011); Agrawal and Goyal (2012b,a); Scott (2010); Kaufmann et al. (2012); Bubeck and Liu (2013); Russo and Van-Roy (2013), particularly in the context of learning click-through rates in ad auctions and search results at almost every large technology company. There are several reasons for its wide popularity, most notably it has the *best* empirical performance for the stochastic MAB problem and its variants Chapelle and Li (2011) – understanding this aspect (in the Bayesian setting) is our goal in this paper.

Since Thompson Sampling is a policy for a stochastic decision problem (MAB with priors), and since it coincides with probability matching experiments Vulkan (1992) for $T = 1$ steps, this naturally brings up the questions: *Can we bring to bear the analysis techniques in stochastic optimization to the analysis of Thompson Sampling? Is an approximation result, such as the* 2*–approximation for probability matching discussed above, true for Thompson Sampling as well?* We show that this is indeed possible using the following roadmap:

Mistake-based Objective and Stochastic Optimization. Generalizing the notion of mistakes in probability matching, we define a *Stochastic Regret* objective that accounts for the expected number of times a policy plays an arm whose underlying expected reward is not the maximum, where the expectation is over the prior probabilities (as is standard in stochastic optimization). Our notion of stochastic regret is related to the notions of 0–1 regret Lai and Robbins (1985); Auer et al. (2002) and Bayes risk Lai (1987) in computational learning; however, there are subtle differences (see Section 1.3), and to the best of our knowledge it has not been studied before. Given this objective, we consider the stochastic optimization problem of minimizing it over a horizon of T steps for a given prior distribution. The optimum solution is a (likely exponential size) dynamic program, and we can compare the performance of any decision policy for the MAB problem against its value. The optimum and the candidate algorithm start with the same prior information – this reduces the algorithm design question to a tradeoff between efficiency and performance.

Approximation Analysis. Our main contribution is in the analysis of Thompson Sampling as an approximation algorithm for this stochastic optimization problem. Analogous to probability matching, we show that this policy is a 2-approximation over any time horizon T in two canonical cases – the two armed bandit problem (which is the standard setting in which MAB problems have been widely studied), and the setting where the uncertainty about an arm resolves to a high or low value in a single play (Bernoulli point priors). Despite its simplicity, the latter formulation is very useful in a variety of emerging systems optimization tasks Babu et al. (2009); Demberel et al. (2009); Herodotou and Babu (2011).

Though TS generalizes probability matching in a natural way, this does not extend to the analysis, which becomes non-trivial due to available information about the arms changing over time, in a different fashion for different policies. We intuitively define an "approximation preserving coupling" between two executions of a policy that differ only in their first play, and show that any policy that preserves a certain guarantee in this coupling is a 2–approximation. We show that TS preserves this property in the settings we focus on. Based on simulation studies, we finally conjecture that the 2–approximation holds for all priors that are independent across arms. We believe that the above described analysis avenue will be fruitful for both the design of new approximation algorithms as well as analysis of Thompson Sampling type algorithms in more general settings.

1.1. Preliminaries and Definitions

Problem 1 (BAYESIAN MAB GITTINS AND JONES (1972); SCOTT (2010); THOMPSON (1933)) *In this problem, there are n independent arms; for each arm i, the rewards are drawn $i.i.d.$ from distribution $D_i(\theta_i)$, where the parameter θ_i is unknown a priori. If arm i is played at time t, a reward r_{it} is drawn $i.i.d.$ from $D_i(\theta_i)$ and observed. As input, there is a n-dimensional prior distribution $\mathcal{D}$ over the possible values of the parameters $\boldsymbol{\theta}$.*[1]

A **decision policy** *P plays exactly one arm each step, and there is a horizon of T steps. At time t, the information available to the decision policy is the sequence of plays and corresponding observations until time t and the remaining horizon $T - t$; the policy is therefore a mapping from this information space to the set $\{1, 2, \ldots, n\}$ corresponding to which arm to play next. Let $R_P(\boldsymbol{\theta})$ be the expected reward of P over a horizon of T steps when the underlying parameters are $\boldsymbol{\theta}$. The goal is to design P to maximize $\mathbf{E}_{\boldsymbol{\theta}\sim\mathcal{D}}\left[R_P(\boldsymbol{\theta})\right]$.*

In a Bayesian sense, each play and corresponding observation refines the prior $\mathcal{D}$ to a corresponding posterior via Bayes rule. Let $\mathcal{D}_t$ denote the posterior at time t; the policy P is a mapping from this posterior and remaining horizon $T - t$ to an action (playing one of the n arms). Throughout this paper, we will assume $\mathcal{D}$ is a product distribution across arms, *i.e.*, there are independent priors $\mathcal{D}_i$ for the parameters θ_i. Playing arm i only updates $\mathcal{D}_i$; the priors of other arms stay the same since they are independent.

Definition 1 (THOMPSON SAMPLING, *henceforth denoted as* TS) *Let $\mu_i(\theta_i) = \mathbf{E}[D_i(\theta_i)]$ denote the expected reward from a play of arm i given that the underlying parameter is θ_i. Let $k(\boldsymbol{\theta}) = argmax_i \mu_i(\theta_i)$ denote the arm with the highest expected mean given parameter vector $\boldsymbol{\theta}$. The Thompson Sampling (TS) policy is defined at time step t as follows. Let $\mathcal{D}_t$ denote the current posterior density over $\boldsymbol{\theta}$. Define q_{it} as the probability of arm i having largest expected mean given this posterior,* i.e., *$q_{it} = Pr_{\boldsymbol{\theta}\sim\mathcal{D}_t}\left[k(\boldsymbol{\theta}) = i\right]$. The TS policy plays arm i with probability q_{it}. (We assume the probability that two arms have the same mean to be a measure zero event, so that the policy is well defined.)*

[1] We are assuming θ_i is scalar; this is without loss of generality.

1.1.1. THE STOCHASTIC OPTIMIZATION PROBLEM

As mentioned before, we extend the notion of *mistake* in probability matching in a natural fashion to define the notion of *stochastic regret*.

Definition 2 (STOCHASTIC REGRET) *Given a decision policy P for Problem 1, let $M(i, \boldsymbol{\theta})$ denote the expected number of times the policy plays arm i, conditioned on the underlying parameter vector being $\boldsymbol{\theta}$, where the expectation is over the random outcomes observed when the arms are played. Let $M(\boldsymbol{\theta}) = \sum_{i \neq k(\boldsymbol{\theta})} M(i, \boldsymbol{\theta})$ denote the expected number of times an arm other than $k(\boldsymbol{\theta})$ is played. The* **stochastic regret** *of policy P is the expected number of steps the policy makes a sub-optimal play relative to a policy that knew $\boldsymbol{\theta}$, where the expectation is over $\boldsymbol{\theta} \sim \mathcal{D}$. Therefore, the Stochastic Regret of policy P is $\mathbf{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[M(\boldsymbol{\theta})]$.*

Given any time horizon T and prior $\mathcal{D}$, there is an optimal policy $OPT(\mathcal{D}, T)$ that minimizes the stochastic regret, which can be computed by dynamic programming (see Section 2). In this paper, we try to resolve the following basic algorithmic conjecture:

Conjecture 3 *Thompson sampling is an anytime 2-approximation to the optimal stochastic regret policy $OPT(\mathcal{D}, T)$, for all priors $\mathcal{D}$* independent *across arms, horizons T, and number of arms n.*

In essence, we ask if the probability matching policy has the same approximation guarantee *even in the presence of uncertain information (priors) that is iteratively refined via Bayes' rule*. This statement, however, is *not* a naive extension of the probability matching case ($T = 1$). To see this, consider the $T = 1$ case where (as we have seen) the conjecture is true and the guarantee is in fact tight. The optimal policy in this case simply plays the arm that minimizes the current probability $(1 - q_{it})$, of making a mistake. However, when the horizon becomes larger ($T > 1$), it is easy to check that the myopic optimum is **not** an approximation algorithm.

1.2. Our Results and Techniques

We resolve Conjecture 3 for two canonical cases.

The Two-armed Bandit Problem. There are $n = 2$ arms with arbitrary priors $\mathcal{D}$. Observe that just the fact that there are two bandits does not give a 2 approximation.This is the canonical case of the MAB problem that has been studied extensively in literature, starting with the original work of Thompson Thompson (1933) and Robbins Robbins (1952). In fact, different policies diverge in their empirical behavior even in this case Garivier and Cappé (2011); Kaufmann et al. (2012); Chapelle and Li (2011), and most bandit heuristics are evaluated in this setting.

Bernoulli Point Priors. In this setting, we restrict attention to a specific class of priors. We assume $D_i(\theta_i) = \theta_i$ is a deterministic distribution. The prior $\mathcal{D}_i$ is over two values: $\theta_i = a_i$ with probability p_i, and $\theta_i = 0$ otherwise. We assume that $a_1 > a_2 > \cdots > a_n > 0$, and $p_n = 1$. (The case when $\mathcal{D}_i$ is a general two-valued distribution reduces to this case.) The first play of arm i resolves its reward θ_i to either a_i or 0, and subsequent plays yield exactly this reward. This is the simplest non-trivial setting where we observe a trade-off between exploration and exploitation; furthermore, classical frequentist analysis (see Section 1.3) do not yield relevant bounds for point priors. For this canonical case, we again show that TS is a 2 approximation for all n, T. We present a 4-approximation proof in Section 4, and show the tight 2-approximation proof in Section B.

The choice of these two canonical cases are motivated by the fact that two arm bandit problem has a rich history and that Bernoulli priors highlight the difference with respect to the existing frequentist literature which we discuss in Section 1.3.

1.2.1. Analysis Roadmap and Techniques

The analysis is surprisingly non-trivial because we are comparing a myopic policy TS against an exponential size dynamic programming (DP) optimum, which quite likely has no simple characterization even for two arms. Analyses for greedy policies in decision theoretic problems typically uses LP-based bounds (e.g., Goel et al. (2009); Guha et al. (2010); Dean et al. (2004)), and our problem does not seem to admit a tractable LP relaxation. In fact, one of our main contributions is in developing new analysis techniques that directly compare against the DP optimum.

Denote optimal decision policy P^* for the input $\mathcal{D}$ and T; call its value OPT. We wish to show $TS \leq 2OPT$. Note that as time progresses, the information sets of TS and OPT diverge since they play different arms, which complicates a direct analysis. In Section 2, we reduce the approximation guarantee of $(c+1)$ to showing the following property of TS: For any arm i, consider the policy $TS[i,T]$ that plays arm i at the first time step, and then executes TS for the remaining $T-1$ time steps. Let $TS[T]$ denote the stochastic regret of TS executed for T steps. Then, for all $T \geq 1$ and $1 \leq i \leq n$, we have:

$$TS[T] - TS[i,T] \leq c(1-q_i) \tag{1}$$

where $q_i = \Pr_{\boldsymbol{\theta} \sim \mathcal{D}}[k(\boldsymbol{\theta}) = i]$. To develop intuition, suppose $c = 1$ and arm i is the "best" arm. Then expected regret $TS[T]$ incurs in not playing arm i at the first step is roughly $1 - q_i$. The above precondition suggests that the only loss that $TS[T]$ incurs over $TS[i,T]$, and this includes the advantage $TS[i,T]$ obtains in knowing the value of arm i at future time steps.

The rest of the paper involves proving the precondition for $c = 1$, for the two cases outlined above. We establish this by coupling the executions of $TS[i,T]$ and $TS[T]$. However, just as with comparing TS with OPT directly, the immediate difficulty is that $TS[i,T]$ has different information at the second step compared to $TS[T]$ since they have played different arms at the first step. Despite this, we show an inductive coupling in Section 3 for the case of two arms ($n = 2$) via an interchange argument on the plays. For the case when $\mathcal{D}$ is drawn from the family of Bernoulli point priors (Section 4), we exhibit a careful coupling between the two executions over a subset of the arms. We show that these two executions have "almost" the same information in terms of arms they have played and observed. We combine this coupling with establishing the martingale structure of a suitably defined potential function, and finally use Doob's optional stopping theorem to establish that the difference in regret between the two executions is bounded. However, a basic argument only yields a 4 approximation. Showing a tight 2 approximation (Section B) requires a more careful analysis of the states introduced by the coupling. The analysis requires expressing the difference of $TS[T] - TS[i,T]$ into several sub-functions, each of which is a sub-martingale or super-martingale.

At a high level, the difficulty in our proofs (and in resolving Conjecture 3) stems from the relation of Precondition (1) to the *value of information*: Conditioned on playing arm i once and knowing its reward, does the regret of $TS[T]$ only decrease? Such a statement is clearly true of optimal policies via Jensen's inequality, but such statements need not hold for sub-optimal policies like TS. We are not aware of *any* prior work on proving such inequalities for sub-optimal policies.

1.3. Comparison with Frequentist Analysis

Despite its practical success, progress on understanding the theoretical properties of Thompson Sampling has only recently begun to be made. A sequence of recent papers Kaufmann et al. (2012); Agrawal and Goyal (2012b,a); Bubeck and Liu (2013); Russo and Van-Roy (2013); Gopalan et al. (2014) have proposed a *frequentist* explanation for why this policy works well in practice. In the frequentist setting, there is no prior distribution $\mathcal{D}$. Instead, we assume the parameters $\boldsymbol{\theta}$ are adversarially chosen. Recall that $k(\boldsymbol{\theta})$ is the arm with largest expected reward – if a policy knew $\boldsymbol{\theta}$, it would play this arm every time step. For any other arm $i \neq k(\boldsymbol{\theta})$, the *regret* of this arm is the number of times this arm is played by the constructed policy. zero for a policy that knows $\boldsymbol{\theta}$).

Culminating a line of research, the work of Gopalan et al. (2014) shows that as long as the priors satisfy certain properties, then regardless of the exact choice of prior, Thompson Sampling achieves frequentist regret that, as $T \to \infty$, matches an information theoretic asymptotic lower bound established by Lai and Robbins Lai and Robbins (1985) (see also Lai (1987); Burnetas and Katehakis (1996)). To compete against the the asymptotic Lai-Robbins bound, most frequentist algorithms proceed by constructing high-probability upper-confidence (Chernoff-type) bounds (UCB) on the mean rewards of the arms given the observations so far, and play the arm with the highest UCB. A UCB constructed by maximizing KL-divergence provides the optimal bound Lai and Robbins (1985); Burnetas and Katehakis (1996); Garivier and Cappé (2011); Gopalan et al. (2014), and the frequentist analyses show that TS mimics the behavior of such a policy.

Our approach is a significant deviation, and can be viewed as an alternative style of analysis. We view Thompson Sampling as a stochastic decision policy and ask whether it uses the given prior information efficiently, and what objective it is trying to approximate. We indeed find such an objective in stochastic regret – in contrast with the Lai-Robbins bound that only holds asymptotically in T Garivier and Cappé (2011), the dynamic program provides a benchmark for any T. We provide evidence of a different fundamental property of Thompson sampling in how it uses prior information in order to compete continuously with the optimal stochastic regret policy. In fact, we show such an analysis even for cases (point priors) where the classical frequentist approach Gopalan et al. (2014) does not yield relevant bounds. Finally, our analysis for the two-arm case provides (in hindsight) a much simpler argument than the frequentist analysis.

2. Dynamic Programming and Precondition (1)

Let $\mathcal{D}$ denote the prior distribution over the arms. Denote a generic action and reward at time t by $\sigma_t = (a_t, r_t)$, where $a_t \in \{1, 2, \ldots, n\}$ and r_t is the observed reward from this play. Let $\boldsymbol{\sigma_t} = \sigma_1\sigma_2 \cdots \sigma_{t-1}$ denote a sequence of actions and corresponding rewards till time t. At time t, any decision policy's state is encoded by some $\boldsymbol{\sigma_t}$. Define

$$q_i(\boldsymbol{\sigma}) = \Pr\left[k(\boldsymbol{\theta}) = i \mid \mathcal{D}, \boldsymbol{\sigma}\right]$$

as the probability arm i has the maximum mean reward given the state $\boldsymbol{\sigma}$, and the prior $\mathcal{D}$. This probability can be computed by updating the prior $\mathcal{D}$ to the posterior $\mathcal{D}(\boldsymbol{\sigma})$ using Bayes' rule, and computing the probability that i has the maximum mean when $\boldsymbol{\theta}$ is drawn from $\mathcal{D}(\boldsymbol{\sigma})$. Similarly, let $\mathcal{D}_i(\boldsymbol{\sigma})$ denote the posterior distribution over parameter θ_i given the state $\boldsymbol{\sigma}$.

Let $OPT[\boldsymbol{\sigma}, t]$ denote the regret of the optimal decision policy conditioned on having state $\boldsymbol{\sigma}$ of size $T - t$, with a horizon of t time steps to go. This policy has the choice of playing one of n

arms; if it plays arm i, the regret incurred by this play is $1 - q_i(\boldsymbol{\sigma})$ and the policy observes a reward r drawn from $D_i(\theta_i)$, where the parameter $\theta_i \sim \mathcal{D}_i(\boldsymbol{\sigma})$. For notational convenience, we denote this draw as $r \sim \mathcal{D}_i(\boldsymbol{\sigma})$. The optimal policy is now the solution to the following dynamic program.

$$OPT[\boldsymbol{\sigma}, t] = \min_{i=1}^{n} \left(1 - q_i(\boldsymbol{\sigma}) + \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[OPT[(\boldsymbol{\sigma} \cdot (i, r)), t-1]\right]\right) \tag{2}$$

The base case when $t = 1$ is simply: $OPT[\boldsymbol{\sigma}, 1] = \min_{i=1}^{n}(1 - q_i(\boldsymbol{\sigma}))$.

We first present an important property of q_i; the proof follows almost by definition.

Lemma 4 (Martingale Property) *For all $i, j \in \{1, 2, \ldots, n\}$ and all $\boldsymbol{\sigma}$ we have:*

$$q_j(\boldsymbol{\sigma}) = \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[q_j(\boldsymbol{\sigma} \cdot (i, r))\right]$$

Consider the Thompson Sampling policy. Faced with state $\boldsymbol{\sigma}$ and a remaining horizon of $t \geq 1$ steps, the policy plays arm i with probability $q_i(\boldsymbol{\sigma})$, and hence we have:

$$TS[\boldsymbol{\sigma}, t] = \sum_{i=1}^{n} q_i(\boldsymbol{\sigma}) \times \left(1 - q_i(\boldsymbol{\sigma}) + \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[TS[(\boldsymbol{\sigma} \cdot (i, r)), t-1]\right]\right)$$

with the base case being $TS[\boldsymbol{\sigma}, 1] = \sum_{i=1}^{n} q_i(\boldsymbol{\sigma}) \times (1 - q_i(\boldsymbol{\sigma}))$.

Let $TS[i, \boldsymbol{\sigma}, t]$ denote the regret of the policy that plays arm i at the first time step, and subsequently executes Thompson Sampling for the remaining $t - 1$ time steps. We have:

$$TS[i, \boldsymbol{\sigma}, t] = 1 - q_i(\boldsymbol{\sigma}) + \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[TS[(\boldsymbol{\sigma} \cdot (i, r)), t-1]\right]$$

so that we have: $TS[\boldsymbol{\sigma}, t] = \sum_{i=1}^{n} q_i(\boldsymbol{\sigma}) TS[i, \boldsymbol{\sigma}, t]$.

The next lemma reduces the approximation guarantee to a property of the function TS. The statement holds for any policy. The proof (in Appendix A) follows by induction on Equation (2).

Lemma 5 *Given a prior $\mathcal{D}$, horizon T, and a policy $\mathcal{P}$ with value function V, suppose that for all $T \geq t \geq 1$, all $\boldsymbol{\sigma}$ (of size $T - t$), and all $1 \leq i \leq n$ we have:*

$$V[\boldsymbol{\sigma}, t] \leq V[i, \boldsymbol{\sigma}, t] + c(1 - q_i(\boldsymbol{\sigma})) \tag{3}$$

Suppose further that $V[\boldsymbol{\sigma}, 1] \leq (c+1) OPT[\boldsymbol{\sigma}, 1]$. Then for all $t \leq T$ and $\boldsymbol{\sigma}$ of size $T - t$, we have $V[\boldsymbol{\sigma}, t] \leq (c+1) OPT[\boldsymbol{\sigma}, t]$.

For the Thompson Sampling policy, it is easy to show that $TS[\boldsymbol{\sigma}, 1] \leq 2 OPT[\boldsymbol{\sigma}, 1]$: Fix some $\boldsymbol{\sigma}$, and let $p_i = q_i(\boldsymbol{\sigma})$. Note that $\sum_i p_i = 1$. Let $i^* = \text{argmin}_i (1 - p_i)$. Then, $OPT = 1 - p_{i^*}$, and $TS = p_{i^*}(1 - p_{i^*}) + \sum_{j \neq i^*} p_j (1 - p_j) \leq 1 - p_{i^*} + \sum_{j \neq i^*} p_j = 2(1 - p_{i^*})$. Therefore, to show TS is a 2 approximation, it suffices to establish precondition (3) when $c = 1$.

3. Two-armed Bandits: A 2-Approximation Analysis

The proof of the precondition (3) for $n = 2$ arms and $c = 1$ uses induction over the remaining horizon. This will show that Thompson sampling is a 2 approximation for arbitrary priors $\mathcal{D}$, when there are $n = 2$ arms. Denote the two arms by $\{a, b\}$. The following lemma presents an equivalent characterization of precondition (3).

Lemma 6 *For the case of $n = 2$ arms denoted $\{a, b\}$, we have:*

$$\forall \boldsymbol{\sigma}, t : \qquad TS[\boldsymbol{\sigma}, t] \leq TS[x, \boldsymbol{\sigma}, t] + 1 - q_x(\boldsymbol{\sigma}) \ \forall x = \{a, b\} \iff |TS[a, \boldsymbol{\sigma}, t] - TS[b, \boldsymbol{\sigma}, t]| \leq 1$$

Proof Fix some $\boldsymbol{\sigma}, t$, and omit these from the notation. Suppose $TS \leq TS[a] + 1 - q_a$. Expanding $TS = q_a TS[a] + q_b TS[b]$, and observing that $q_a + q_b = 1$, we obtain $TS[b] - TS[a] \leq 1$. Conversely, if $TS[a] \leq 1 + TS[b]$, then $TS = q_a TS[a] + q_b TS[b] \leq TS[b] + 1 - q_b$. Reversing the roles of a and b completes the proof. ■

Assume w.l.o.g. that the current state corresponds to $\boldsymbol{\sigma} = \phi$; we omit this state in the notation when obvious. Let $\sigma_a(r)$ denote the state if arm a is played and r is observed; let $\sigma_{ab}(r, s)$ denote the state if arm a is played and r observed, followed by b played and s observed. By induction, assume precondition (3) (and its consequence via Lemma 6) is true for horizons less than t. Consider playing arm a first. Note that $q_a + q_b = 1$ for all states. We have:

$$\begin{aligned} TS[a, \phi, t] &= 1 - q_a + \mathbf{E}_{r \sim \mathcal{D}_a} [TS[\sigma_a(r), t-1] \\ &\leq 1 - q_a + \mathbf{E}_{r \sim D_a} [1 - q_b(\sigma_a(r)) + TS[b, \sigma_a(r), t-1]] \\ &= 1 - q_a + \mathbf{E}_{r \sim D_a} \left[2 q_a(\sigma_a(r)) + \mathbf{E}_{s \sim \mathcal{D}_b(\sigma_a(r))} \left[TS\left[\sigma_{ab}(r, s), t-2\right]\right]\right] \\ &= 1 + q_a + \mathbf{E}_{r \sim D_a} \left[\mathbf{E}_{s \sim \mathcal{D}_b(\sigma_a(r))} \left[TS\left[\sigma_{ab}(r, s), t-2\right]\right]\right] \end{aligned}$$

Here, the first inequality follows from the inductive hypothesis applied with $i = b$; the following equality applies because $q_a + q_b = 1$ for all states; and the final equality holds by the Martingale property of q_a (Lemma 4). Similarly, if arm b is played first (using the obvious change in notation):

$$\begin{aligned} TS[b, \phi, t] &= 1 - q_b + \mathbf{E}_{s \sim \mathcal{D}_b} [TS[\sigma_b(s), t-1]] \\ &= q_a + \mathbf{E}_{s \sim \mathcal{D}_b} [q_a(\sigma_b(s)) TS[a, \sigma_b(s), t-1] + q_b(\sigma_b(s)) TS[b, \sigma_b(s), t-1]] \\ &\geq q_a + \mathbf{E}_{s \sim \mathcal{D}_b} [TS[a, \sigma_b(s), t-1] - q_b(\sigma_b(s))] \\ &= q_a - q_b + \mathbf{E}_{s \sim \mathcal{D}_b} [TS[a, \sigma_b(s), t-1]] \\ &= q_a - q_b + \mathbf{E}_{s \sim \mathcal{D}_b} \left[1 - q_a(\sigma_a(s)) + \mathbf{E}_{r \sim \mathcal{D}_a(\sigma_b(s))} [TS[\sigma_{ba}(s, r), t-2]]\right] \\ &= q_a - q_b + q_b + \mathbf{E}_{s \sim \mathcal{D}_b} \left[\mathbf{E}_{r \sim \mathcal{D}_a(\sigma_b(s))} [TS[\sigma_{ba}(s, r), t-2]]\right] \\ &= q_a + \mathbf{E}_{r \sim D_a} \left[\mathbf{E}_{s \sim \mathcal{D}_b(\sigma_a(r))} \left[TS\left[\sigma_{ab}(r, s), t-2\right]\right]\right] \end{aligned}$$

Here, the first inequality follows by the inductive hypothesis combined with Lemma 6. The next equality and the penultimate one follow from the martingale property (Lemma 4), and the final equality follows since the plays of the arms are independent, so that the final states are statistically identical whether arm a is played first and then arm b, or the other way around. This shows $TS[a, \phi, t] - TS[b, \phi, t] \leq 1$. Switching the roles of a and b in the above argument shows that $|TS[a, \phi, t] - TS[b, \phi, t]| \leq 1$. Combining this with Lemma 6, precondition (3) follows.

We note that the above proof of 2 approximation does not need the prior $\mathcal{D}$ to be a product distribution over $\mathcal{D}_a$ and $\mathcal{D}_b$; it only needs that the plays are $i.i.d.$ for each arm.

4. Bernoulli Point Priors: A 4-Approximation Analysis

In this section, we consider the case when $\mathcal{D} = X_1 \times X_2 \times \cdots \times X_n$, where each X_i is a Bernoulli distribution that takes on value $a_i > 0$ with probability p_i, and 0 otherwise. The priors for different arms are independent. We will use the notation $X_i = B(a_i, p_i)$. In this setting, the distribution $D_i(\theta_i) = \theta_i$ is deterministic, and $\theta_i \sim X_i$. Therefore, the first time arm i is played, the reward of the play ($\theta_i = a_i$ or $\theta_i = 0$) resolves the arm, and subsequent plays yield exactly this reward.

At a high level, we first define a potential function in Section 4.1 that combined with the regret of TS, defines a martingale. Subsequently, in Section 4.2, we exhibit a coupling between the executions of $TS[t]$ and $TS[i, t+1]$. We argue about the change in potential functions via the coupling, and use Doob's optional stopping theorem to bound the difference in regret between the two executions. This will establish precondition (3) when $c = 3$, implying a 4-approximation. Using a more intricate analysis, we improve this to a 2-approximation in Section B. Some of the the proofs in this section are relegated to Appendix A.

In order to make Thompson Sampling well-defined, we will restrict attention to canonical Bernoulli distributions:

Definition 7 *Given Bernoulli distributions $X_1, X_2, \ldots, X_n$ define them to be* **canonical** *if the distributions are $X_i = B(p_i, a_i)$ where $a_1 > a_2 \cdots a_n > 0$ and $p_n = 1$.*

Definition 8 *Given a set of canonical distributions, let q_i be the probability that the variable X_i is the true maximum. Therefore, $q_i = \prod_{j<i}(1 - p_j)p_i$.*

Lemma 9 *Given any canonical set of distributions, we have $1 + \sum_i q_i^2 - 2\sum_i \frac{q_i^2}{p_i} = 0$. Therefore we have:* $\displaystyle\sum_{i \neq n} \left(\frac{q_i^2}{p_i} - q_i^2 \right) = \frac{1}{2} \left(1 - \sum_i q_i^2 \right)$.

4.1. The Potential Function

We now define a potential function, and argue that when combined with the regret of Thompson sampling till time t, it defines a martingale. Consider the execution of the Thompson sampling policy. At time t, the policy has some random state $\boldsymbol{\sigma}$; we omit explicitly mentioning the state since we are conditioning on reaching it. Let q_{it} denote the probability that arm i is the maximum. Define the quantity p_{it} as follows: If arm i has never been played, it is p_i; if arm i has been played and the observed outcome is a_i, then $p_{it} = 1$; else (if arm i has been played and the observed outcome is 0 then) $p_{it} = 0$. Let S_t denote the set of arms with $q_{it} > 0$; TS only plays one of the arms in S_t at time t. In the remaining discussion, we restrict to these arms unless otherwise stated.

Definition 10 *Define the potential function $\Phi[t]$ as $\Phi[t] = \sum_{i \in S_t} \frac{q_{it}}{p_{it}}$.*

Note that there can be only one arm in S_t with $p_{it} = 1$ – we term this the *backup arm* and denote it as b_t. It is the arm with lowest index that has been played and observed to have a value > 0, *i.e.*, $b_t = \max\{j | j \in S_t\}$. For any arm $i \in S_t \setminus \{b_t\}$, observe that this arm has not been played so far, so that $\frac{q_{it}}{p_{it}} = \frac{\sum_{j>i, j \in S_t} q_{jt}}{(1-p_{it})}$.

Definition 11 *For the Thompson sampling policy, let the random variable $\mathcal{R}_t$ denote the* **total** *regret incurred up to, but not including, the time step t. Therefore, $\mathcal{R}_1 = 0$.*

Lemma 12 *For the TS policy, let $\Delta_{it} = \Phi[t] - \mathbb{E}\left[\Phi[t+1] | \text{Arm } i \text{ is played at time } t\right]$. If $i \in S_t \setminus \{b_t\}$, then $\Delta_{it} = \frac{q_{it}(1-p_i)}{p_i}$, and if $i = b_t$ then $\Delta_{it} = 0$.*

Lemma 13 *For the Thompson sampling policy, $\mathcal{Q}_t = \mathcal{R}_t + 2\Phi[t]$ is a martingale sequence.*

Proof At time t, Thompson sampling plays $i \in S_t$ with probability q_{it}. Suppose it plays $i \neq b_t$. Then $p_{it} = p_i$, so that with p_i, the policy observes $X_i = a_i$ and $\mathcal{R}_{t+1} - \mathcal{R}_t = (1 - q_{i(t+1)})$ where $q_{i(t+1)} = q_{it}p_i$. With probability $(1-p_i)$, the policy observes $X_i = 0$, so that $\mathcal{R}_{t+1} - \mathcal{R}_t = 1$. Therefore the expected increase in $\mathcal{R}_t$ is $1 - q_{it}$. By Lemma 12, Φ_t decreases in expectation by $\frac{2q_{it}(1-p_i)}{p_i}$. Therefore:

$$\mathbb{E}\left[\mathcal{Q}_{(t+1)} - \mathcal{Q}_t | \mathcal{Q}_t, i \neq b_t \text{ played}\right] = 1 - q_{it} - 2\frac{q_{it}(1-p_i)}{p_i} = 1 + q_{it} - \frac{2q_{it}}{p_i} = 1 + q_{it} - \frac{2q_{it}}{p_{it}}$$

Now consider the case $i = b_t$ is played, so that $X_i = a_i$ is known even before the play. In this case $\mathcal{R}_{t+1} - \mathcal{R}_t = 1 - q_{it}$, and Φ is unchanged. Since $p_{it} = 1$, we have:

$$\mathbb{E}\left[\mathcal{Q}_{(t+1)} - \mathcal{Q}_t | \mathcal{Q}_t, i = b_t \text{ played}\right] = 1 - q_{it} = 1 + q_{it} - \frac{2q_{it}}{p_{it}}$$

Since Thompson sampling plays arm i with probability q_{it}, we have

$$\mathbb{E}\left[\mathcal{Q}_{(t+1)} - \mathcal{Q}_t | \mathcal{Q}_t\right] = \sum_i q_{it}\left(1 + q_{it} - \frac{2q_{it}}{p_{it}}\right) = 1 + \sum_i q_{it}^2 - 2\sum_i \frac{q_{it}^2}{p_{it}} = 0$$

where the last equality follows from Lemma 9. ∎

4.2. Coupling $TS[t]$ and $TS[i, t+1]$

We now proceed to establish precondition (3) when $c = 3$, which will show that Thompson sampling is a 4-approximation. Recall that $TS[t-1]$ is Thompson sampling with a horizon of $t-1$ steps; the random variable $\mathcal{R}_t$ is the regret on some sample path of this policy. Similarly, $TS[i,t]$ is the policy that plays arm i at the first step, and executes Thompson sampling for the next $t-1$ steps.

Definition 14 *Denote the regret of $TS[i,t]$ as $\mathcal{R}_{t+1}[i]$. Let $\mathcal{R}^{ex}_{t+1}[i]$ be the regret of $TS[i,t]$* **excluding** *the regret from step* 1.

To compare $TS[t+1] = \mathbb{E}\left[\mathcal{R}_{t+2}\right]$ and $TS[i, t+1] = \mathbb{E}\left[\mathcal{R}_{t+2}[i]\right]$ we follow a two step approach:

(I) We define a natural coupling that allows us to compare $\mathbb{E}\left[\mathcal{R}_{t+1}\right]$ (denoting the first t plays of $TS[t+1]$) and $\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right]$ which excludes the regret of the first play of $TS[i, t+1]$.

(II) We then relate the regret of the first play of $TS[i, t+1]$ to the last play of $TS[t+1]$.

Recall the definition of $\Phi[t']$ from Def. 10. Analogously define $\Phi[i, t']$ for the policy that plays arm i at the first step, and subsequently executes the Thompson sampling policy. The following ensues from Lemma 13 and Doob's optional stopping theorem.

Corollary 15 *(a)* $2\mathbb{E}\left[\Phi[t+1]\right] + \mathbb{E}\left[\mathcal{R}_{t+1}\right] = 2\Phi[1]$; *(b)* $2\mathbb{E}\left[\Phi[i, t+2]\right] + \mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right] = 2\mathbb{E}\left[\Phi[i, 2]\right]$.

Note $\mathbb{E}\left[\mathcal{R}_1\right] = 0$ and $\mathbb{E}\left[\mathcal{R}_2^{ex}[i]\right] = 0$. We now define a coupling between the information sets in steps $1, \ldots, t$ of $TS[t+1]$ and those of steps $2, \ldots, t+1$ of $TS[i, t+1]$. In particular, we show an execution of $TS[i, t+1]$ (call it $\tilde{TS}[i, t+1]$) at time $t'+1$ based on the execution of $TS[t+1]$ at time t'. This will define the coupling. Initially i is declared SPECIAL and the coupling is $(\emptyset, \emptyset)$. Define an event COLLAPSE to be initially false.

Invariant. At a general point in time t', suppose s is the SPECIAL arm. Then we maintain the invariant that $\tilde{TS}[i, t+1]$ and $TS[t]$ differ only in the knowledge of the value of arm s. $\tilde{TS}[i, t+1]$ would have played s and knows X_s. $TS[t]$ has not played s and does not know X_s. Therefore $p_{st'} = p_s$ below. If $X_{\mathrm{i}} = a_s$, then $TS[t]$ might have information about arms $j > s$.

The coupling at time t' is given below. We only perform this if COLLAPSE is false; otherwise the coupling will be trivial.

(a) Given a current states (σ, σ') for $TS[t]$ at time t' and $\tilde{TS}[i, t+1]$ at time $t'+1$, first execute the next play of $TS[t]$ conditioned on σ. Suppose this is for arm j. Therefore $\sigma \to \sigma \circ j$.

(b) If $j < s$ then set $\sigma' \to \sigma' \circ j$.

(c) If $j \geq s$ where s is SPECIAL:

  (1) If $X_s = a_s$ then $\tilde{TS}[i, t+1]$ plays s; that is $\sigma' \to \sigma' \circ s$.

  (2) If $X_s = 0$ and $j > s$ then $\tilde{TS}[i, t+1]$ plays j; that is $\sigma' \to \sigma' \circ j$.

  (3) If $X_s = 0$ and $j = s$ then let $\Delta_{st'} = \sum_{j' > s} q_{j't'}$ where $q_{j't'}$ is the probability $TS[t]$ would have played j'. Note that $\Delta_{st'} = \frac{q_{st'}}{p_s}(1 - p_s)$. Then $\tilde{TS}[i, t+1]$ plays a $j' > s$ with probability $\frac{q_{j't'}}{\Delta_{st'}}$; set $\sigma' \to \sigma' \circ j'$ and j' is now SPECIAL.

(d) If $TS[t]$ played an arm $j \leq s$ and observed $X_j = a_j$ then set COLLAPSE as true: $TS[t], \tilde{TS}[i, t+1]$ have the same information and identical executions subsequently.

Lemma 16 *The executions of $\tilde{TS}[i, t+1], TS[i, t+1]$ are statistically identical.*

Proof We will show that any state σ' is arrived with the same probability by $TS[i, t+1]$ and $\tilde{TS}[i, t+1]$. The proof is by induction on $t' = |\sigma'|$. The claim is trivially true for $|\sigma'| = t' = 0$. We now observe that $\tilde{TS}[i, t'+1]$ and $TS[t']$ differ only the knowledge of the state of the distribution with index s. Therefore the probability of playing an arm with index $j < s$ is identical for both processes, and this is identical to the probability that $TS[i, t+1]$ with state σ' plays this arm – this proves Step (b) is statistically identical.

Consider now the case where $TS[t]$ plays $j > s$. If $X_{\mathrm{i}} = a_s$ then $TS[i, t'+1]$ cannot plays j. But the probability of playing s by $TS[i, t'+1]$ is $\sum_{j<s} q_{jt'}$, and Step (c1) shows that $\tilde{TS}[i, t+1]$ plays s with the same probability.

If on the other hand, $X_{\mathrm{i}} = 0$, then the probability that $TS[i, t'+1]$ plays $j > s$ is $q_{jt'}/(1 - p_s)$. Now $TS[t']$ plays j with probability $q_{jt'}$ in Step (c2). However $TS[t']$ plays s with probability $q_{\mathrm{i}t'}$ in Step (c3); and therefore $\tilde{TS}[i, t+1]$ plays j with total probability

$$q_{jt'} + q_{st'} \frac{q_{jt'}}{\Delta_{st'}} = q_{jt'} \left(1 + \frac{q_{st'} p_s}{(1 - p_s)}\right) = \frac{q_{jt'}}{(1 - p_s)}$$

The probabilities of all plays are therefore identical between $TS[i, t+1]$ and $\tilde{TS}[i, t+1]$. ■

Lemma 17 *For any t, $\mathbb{E}\left[\Phi[t]\right] - \mathbb{E}\left[\Phi[i, t+1]\right] \geq 0$.*

Proof We will prove this via the coupling. Consider any sample path in $TS[t]$. In the event COLLAPSE is true, we have $\Phi[t] = \Phi[i, t+1]$. Else suppose s was the last special element; note that $TS[t]$ does not know the state of arm s, and we will take expectation over this unknown state. Both the expressions $\Phi[t]$ and $\Phi[i, t+1]$ agree on the contributions of all arms $j < s$.

Let the contribution from $\{j | j \leq s\}$ to $\Phi[i, t+1]$ be $\Delta[i]$ and to $\Phi[t]$ be Δ. If $X_s = a_s$, then $\Delta[i] = \frac{q_{st}}{p_s}$ since that is the probability that X_s is the maximum, and from $j < s$, the contribution is 0. Here, q_{it} refers to the probability that i is the maximum in $TS[t]$. If $X_s = 0$, then $\Delta[i] = \frac{1}{1-p_s}\left(\Delta - \frac{q_{st}}{p_s}\right)$ since the processes have the same information except for the state of X_s. Therefore $\mathbb{E}_{X_s}\left[\Delta[i]\right] = \Delta - \frac{q_{st}}{p_s} + q_{st} < \Delta$. Thus $\mathbb{E}\left[\Phi[t] - \Phi[i, t+1]\right] = \mathbb{E}_{X_s}\left[\Delta - \Delta[i]\right] \geq 0$. ■

Lemma 18 $\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+2}\right] + 3(1 - q_{i1}) \geq 0$.

Proof We first prove that$\left(\mathbb{E}\left[\mathcal{R}_{t+2}[i] - \mathcal{R}^{ex}_{t+2}[i]\right]\right) - \left(\mathbb{E}\left[\mathcal{R}_{t+2} - \mathcal{R}_{t+1}\right]\right) \geq -1 + q_{i1}$. To see this, consider two cases. If arm i is the maximum, then $TS[t]$ has $q_{it} \geq q_{i1}$ for all t, and plays i with probability at least q_{i1} each step. In this case, the first term is zero, and the second term is at most $1 - q_{i1}$. Otherwise, if arm i is not the maximum, the first term is 1 and dominates the second term. In either case, the inequality is true. Now from Corollary 15 and Lemma 17:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+1}\right] &= 2\mathbb{E}\left[\Phi[i,2]\right] - 2\Phi[1] + 2\mathbb{E}\left[\Phi[t+1]\right] - 2\mathbb{E}\left[\Phi[i, t+2]\right] \\ &\geq 2\mathbb{E}\left[\Phi[i,2]\right] - 2\Phi[1] = -\frac{2q_{i1}(1-p_i)}{p_i} \geq -2(1 - q_{i1})\end{aligned}$$

The lemma now follows from summing up the last two inequalities. ■

The next theorem is now immediate. An improved statement is presented in Appendix B.

Theorem 19 *Thompson sampling is a 4 approximation to the optimal stochastic regret for Bernoulli point priors.*

5. Conclusion

Resolving Conjecture 3 for any number of arms and arbitrary priors is a tantalizing open question. We posit the conjecture based on fairly extensive simulations of TS for $n > 2$ arms and general point priors. Analytically, we can show that if TS has increasing *value of information*, meaning that the regret of TS only decreases in expectation if provided more refined information about a prior (in a certain specific sense), then the conjecture is true for n arms and any priors. We omit the details, but note that this property of TS seems intuitively correct.

We highlight where our specific approaches break down. We can formulate an inductive hypothesis similar to that in Section 3 for n arms; however, this ends up being too weak to perform induction on. It is not clear what a stronger hypothesis should be. Similarly, though we can account for the regret as a martingale even for general (non-Bernoulli) point priors, the coupling method in Section 4 falls apart. (However, a martingale argument suffices to show a 2 approximation for the infinite horizon case with general point priors.) We therefore believe that resolving the general conjecture requires developing techniques that go well beyond what we have presented here.

References

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012a.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. *CoRR*, abs/1209.3353, 2012b.

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

S. Babu, N. Borisov, S. Duan, H. Herodotou, and V. Thummala. Automated experiment-driven management of (database) systems. *Proc. of HotOS*, 2009.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. *NIPS*, pages 638–646, 2013.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122 – 142, 1996.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

B. C. Dean, M. X. Goemans, and J. Vondrak. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 208–217, 2004.

A. Demberel, J. Chase, and S. Babu. Reflective control for an elastic cloud appliation: An automated experiment workbench. *Proc. of HotCloud*, 2009.

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. *CoRR*, abs/1106.2369, 2011.

A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Proc. COLT*, 2011.

J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in statistics (European Meeting of Statisticians)*, 1972.

A. Goel, S. Khanna, and B. Null. The ratio index for budgeted learning, with applications. In *SODA*, pages 18–27, 2009.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex bandit problems. In *ICML*, 2014.

S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *J. ACM*, 58(1), 2010.

H. Herodotou and S. Babu. Profiling, what-if analysis, and cost-based optimization of mapreduce programs. *Proc. of VLDB*, 2011.

Emilie Kaufmann, Nathaniel Korda, and Remi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. *Proceedings of ALT*, 2012.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

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

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670, 2010.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

Daniel Russo and Benjamin Van-Roy. Learning to optimize via posterior sampling. *CORR; http://arxiv.org/abs/1301.2609*, 2013.

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

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):pp. 285–294, 1933.

N. Vulkan. An economist's perspective on probability matching. *Journal of Economic Surveys*, 1992.

Appendix A. Omitted Proofs

A.1. Proof of Lemma 5

Proof We prove this by induction over the remaining horizon t; the base case follows by assumption. Suppose the claim is true for horizon $t-1$ and all $\boldsymbol{\sigma}$. Then for horizon t and $\boldsymbol{\sigma}$, we have:

$$\begin{aligned}
V[\boldsymbol{\sigma}, t] &\leq V[i, \boldsymbol{\sigma}, t] + c(1 - q_i(\boldsymbol{\sigma})) && \forall i \\
&= (c+1)(1 - q_i(\boldsymbol{\sigma})) + \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[V[(\boldsymbol{\sigma} \cdot (i,r)), t-1]\right] && \forall i \\
&\leq (c+1)\left((1 - q_i(\boldsymbol{\sigma})) + \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[OPT[(\boldsymbol{\sigma} \cdot (i,r)), t-1]\right]\right) && \forall i \\
&\leq (c+1)\min_{i=1}^{n}\left((1 - q_i(\boldsymbol{\sigma})) + \mathbf{E}_{r \sim \mathcal{D}_i(\boldsymbol{\sigma})}\left[OPT[(\boldsymbol{\sigma} \cdot (i,r)), t-1]\right]\right) \\
&= (c+1)OPT[\boldsymbol{\sigma}, t]
\end{aligned}$$

Here, the first inequality follows from assumption and the second inequality from the inductive hypothesis. This completes the proof. ■

A.2. Proof of Lemma 9

Proof We prove this by induction on n. This is trivially true for one distribution. It is easy to see that it holds for $n = 2$ since $q_1 = p_1, q_2 = (1 - p_1)$ and (note $p_2 = 1$):

$$1 + p_1^2 + (1 - p_1)^2 - 2p_1 - 2(1 - p_1)^2 = 1 + p_1^2 - (1 - p_1)^2 - 2p_1 = 0$$

Now assuming that the claim is true for any $n - 1$ distributions, consider combining the first 2 arms into a single arm with probability $p = p_1 + (1 - p_1)p_2$. This is another canonical distribution and the identity holds. Now observe that in the modified setting (using the primed notation) $q'_1 = p'_1 = p$ and so the contribution of this variable to the identity is:

$$\begin{aligned} (q'_1)^2 - 2\frac{(q'_1)^2}{p'_1} &= p^2 - 2p = p_1^2 + p^2 - 2p_1 p + p_1^2 - 2p_1 - 2[p - p_1 - p_1 p + p_1^2] \\ &= p_1^2 + (p - p_1)^2 - 2\frac{p_1^2}{p_1} - 2\frac{(p - p_1)^2}{\frac{p - p_1}{1 - p_1}} = q_1^2 + q_2^2 - 2\frac{q_1^2}{p_1} - 2\frac{q_2^2}{p_2} \end{aligned}$$

since $q_1 = p_1$ and $q_2 = p - p_1$. The lemma follows. ■

A.3. Proof of Lemma 12

Proof If $i = b_t$, then playing i does not change the state of any arm. Therefore $\Phi[t + 1] = \Phi[t]$. Otherwise, suppose arm $i \in S_t \setminus \{b_t\}$ is played. Observe that for $j < i$, $q_{j(t+1)} = q_{jt}$. Arm i is observed to have value a_i with probability p_i and 0 otherwise. In the former case, note that all $j > i$ drop out of S_{t+1}. Since $q_{i(t+1)} = q_{it}/p_i$ and $p_{i(t+1)} = 1$, we have

$$\mathbb{E}\left[\Phi[t + 1] | X_i = a_i\right] - \Phi[t] = -\sum_{j \geq i} \frac{q_{jt}}{p_{jt}} + \frac{q_{i(t+1)}}{1} = -\sum_{j \geq i} \frac{q_{jt}}{p_{jt}} + \frac{q_{it}}{p_i} = -\sum_{j > i} \frac{q_{jt}}{p_{jt}} \tag{4}$$

In the latter case ($X_i = 0$ is observed), i drops out of S_{t+1}. In this case for $j > i$, we have $q_{j(t+1)} = q_{jt}/(1 - p_i)$ and $p_{j(t+1)} = p_{jt}$. Therefore the change in this case is:

$$\mathbb{E}\left[\Phi[t + 1] | X_i = 0\right] - \Phi[t] = -\sum_{j \geq i} \frac{q_{jt}}{p_{jt}} + \sum_{j > i} \frac{q_{jt}}{p_{jt}(1 - p_i)} \tag{5}$$

Adding Equation 4 multiplied by p_i and Equation 5 by $1 - p_i$, the lemma follows. ■

Appendix B. Bernoulli Point Priors: A Tight 2-Approximation Analysis

In this section we improve Theorem 19 to 2-approximation result in Theorem 30. Note that this is tight as shown by the example in Section 1. We prove

$$\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+2}\right] + (1 - q_{i1}) \geq 0$$

which is precondition (1) in the notation of Section 4, instead of proving $\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+2}\right] + 3(1 - q_{i1}) \geq 0$ as in Lemma 18. Towards this end, observe that the notion of the coupling defined in Section 4.2 *is dependent on the special element s and the element i*. That facet of the dependence has to be reflected in any tight analysis. The simplistic potential $\Phi[t]$ does not suffice any more.

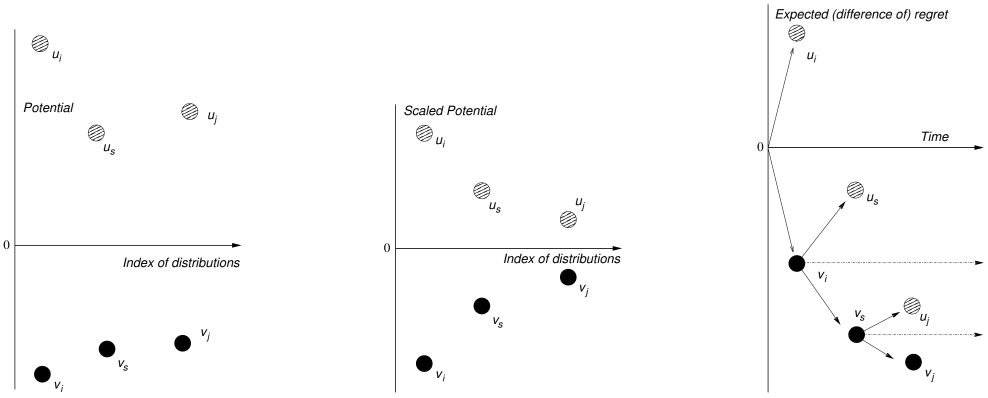
The overall approach: We define a submartingale DIFF with initial mean 0 (and therefore in expectation nondecreasing) that relates $\mathbb{E}\left[\mathcal{R}_{t+1}\right]$ and $\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right]$. The submartingale definition will now explicitly involve the element s – but note that we are only making the analysis dependent on s, the algorithm remains unchanged. Note that $\mathbb{E}\left[\mathcal{R}_{t+1}\right]$ corresponds to the expected regret of $TS[t]$ and $\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right]$ corresponds to the expected regret of $TS[i, t+1]$ excluding the first step.

- We use three different submartingale functions DIFF, LOW, and HIGH.

  The function LOW corresponds to the function DIFF when the special element is observed to have an outcome 0 (hence the name). The function HIGH corresponds to DIFF when the special element is observed to have an outcome a_s. The subprocess HIGH acts as an absorbing subprocess – when we switch to this mode then we remain in that mode. The submartingale DIFF switches to HIGH at most once. The subprocess LOW models the behavior before this switch (note that the special element as defined by the coupling keeps changing). Although out target is the function DIFF, the two functions LOW, HIGH simplify the proof considerably.

  These functions now depend on a 4-tuple (i, s, t, t') where i, s are as defined in the coupling, t is the current time and t' is the starting point where the pair (i, s) became relevant. Intuitively, we are analysing a branching process where from the state (i, s) of the coupling we can go to the state (i, j').

- As expected, with three functions we would need three different potentials.



(a) Potentials (b) Potential × probability (c) Overall process

Figure 1: The potential function of the subprocesses. The upward movement in (c) is the transition DIFF (or LOW, except the time 0) to HIGH. The point is that if the initial arm i had $X_i = 0$ then the process $TS[t] - TS[i, t+1]$ is negative in the first step itself and no subsequent step where $TS[i, t+1]$ knows the special element to be $X_s = a_s$ helps in recovery (of crossing the horizontal line). This assertion holds recursively, if $X_s = 0$ then the difference only increases.

$\boldsymbol{\Upsilon(t'+1,s)}$ **:** where s is a special element which is known to $TS[i,t+1]$ but not to $TS[t]$ when we are about to make the t'-th play for $TS[t]$. Initially $s=i$ but that changes as the process continues (see Definition 20).

$\mathbb{M}_{\boldsymbol{jt}}$ **:** which corresponds to the events where $TS[i,t+1]$ has outplayed $TS[t]$ when the maximum element is j (see Definition 21).

$\mathbb{W}_{\boldsymbol{jt'}}$ **:** which corresponds to the event that X_j became the unknown distribution **and** the previous unknown distribution X_s was observed to be 0 (see Definition 22).

In figure 1; the different subfigures correspond to the knowledge of the special element – $X_s=a_s$ is marked as shaded and $X_s=0$ is marked solid. The value of this $W_{jt'}$ is accounted appropriately such that it balances out the possible contribution of M_{jt}. Without further ado, we define the three potential functions.

Definition 20 *Given a coupling (σ,σ') as defined in Section 4.2, where the special element is s at time t corresponding to $TS[i,t+1]$ and $TS[t]$; at time t $TS[t]$ defines $\{q_{it}\},\{p_{it}\}$. Given such $\{q_{it}\},\{p_{it}\}$ if s has not been observed before then define*

$$\Upsilon(t+1,s)=\begin{cases}0 & \textit{if } p_{st'}=1 \textit{ at time } t' \textit{ when } s \textit{ was declared special}\\ \frac{q_{st}}{p_{st}}\frac{(1-p_s)}{p_s} & \textit{If } X_s=a_s\\ -\frac{q_{st}}{p_{st}} & \textit{Otherwise}\end{cases}$$

Definition 21 *Given a coupling (σ,σ') as defined in Section 4.2; define $\mathbb{M}_{st}$ to be the indicator event which is a conjunction of (i) at the $(t-1)^{th}$ step the special element is s; (ii) we play (j,s) on (σ,σ') where $j>s$ at the $(t-1)^{th}$ step; and (iii) X_s is the maximum. Observe that $\mathbb{M}_{st}\neq 0$ implies $X_s=a_s$. Note $\mathbb{M}_{\cdot 1}=0$ because we are not accounting for plays in 0^{th} step. Moreover condition (ii) implies that no* COLLAPSE *has happened.*

Definition 22 *Given a coupling (σ,σ') as defined in Section 4.2; let $\mathbb{W}_{jt}=\frac{q_{jt}}{p_{jt}}\frac{(1-p_j)}{p_j}$ when the conjunction of the following happens: (i) at the start of the $(t-1)^{th}$ step the special element is s; (ii) we play (s,j) on (σ,σ') where $j>s$ at the $(t-1)^{th}$ step; which implies j is the new special element and (iii) $X_j=a_j$ and j has not been played by σ before; that is; $p_{jt}=p_j$. Observe that $\mathbb{W}_{jt}\neq 0$ implies $X_s=0$. Note $\mathbb{W}_{\cdot 1}=0$ because we are not accounting for plays in 0^{th} step.*

Definition 23 *Define* $\text{HIGH}'(i,s,t,t')=\mathcal{R}^{ex}_{t+1}[i]-\mathcal{R}^{ex}_{t'}[i]-\mathcal{R}_t+\mathcal{R}_{t'}+\sum_j\sum_{t''=t'}^{t}\mathbb{M}_{jt''}$ *and* $\text{HIGH}(i,s,t,t')=\text{HIGH}'(i,s,t,t')+\Upsilon(t+1,s)$.

Lemma 24 *Conditioned on $X_s=a_s$, s becoming a special element at time t',* $\text{HIGH}'(i,s,t,t')$ *is a submartingale for $t\geq t'$ under the coupling (σ,σ'). Moreover $\mathbb{E}[\Upsilon(t+2,s)]=\Upsilon(t+1,s)$ and therefore* $\text{HIGH}(i,s,t,t')$ *is also a submartingale for $t\geq t'$ under the coupling (σ,σ').*

Proof Suppose the next play in σ,σ' are u,v respectively. There are two cases to consider and in each case we show that $\mathbb{E}_{(u,v)}[\text{HIGH}'(i,s,t+1,t')]\geq\text{HIGH}'(i,s,t,t')$. We then show that $\mathbb{E}[\Upsilon(s,t+2)]=\Upsilon(s,t+1)$ proving the second part of the lemma.

(1) Suppose $u < v$ and $v = s$. In this case $\mathcal{R}_{t+1} - \mathcal{R}_t = 1$. Observe that using the definition of $\mathbb{M}_{j(t+1)}$ as an indicator variable:

$$\left(\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] + \sum_j \mathbb{M}_{j(t+1)} \right) - (\mathcal{R}_{t+1} - \mathcal{R}_t) = 0$$

In this case $\mathbb{E}_{(b,c)}\left[\textsc{High}'(i, s, t+1, t')\right] \geq \textsc{High}'(i, s, t, t')$.

(2) Suppose $u = v$ and $v \leq s$. In this case $\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] = \mathcal{R}_{t+1} - \mathcal{R}_t$ and $\mathbb{M}_{\cdot(t+1)} = 0$ by definition. Recall that the definition only counts the events (j, s) with $j > s$. Therefore $\mathbb{E}_{(b,c)}\left[\textsc{High}'\right](i, s, t+1, t')] \geq \textsc{High}'(i, s, t, t')$ in this case as well.

Therefore the first part of the lemma follows. If $p_{st'} = 1$ when s was declared special then $\Upsilon(s) = 0$ throughout and we have nothing to prove. Now in case (1) we have $\Upsilon(t+2, s) = \Upsilon(t+1, s)$ since $u < s$. In case (2), if $u = v > s$ then observe that $\mathbb{E}\left[\Upsilon(t+2, s)\right] = \Upsilon(t+1, s)$. This follows because with probability p_v we have $\Upsilon(t+2, s) = 0$ and with probability $1 - p_v$ we have $\Upsilon(t+2, s) = \Upsilon(t+1, s)/(1 - p_v)$. Finally in case (2) is $b = c = s$ then observe that:

$$\Upsilon(t+2, s) = \frac{q_{s(t+1)}}{p_{s(t+1)}} \frac{(1 - p_s)}{p_s} = \frac{(q_{st}/p_{st})}{1} \frac{(1 - p_s)}{p_s} = \Upsilon(t+1, s)$$

since $q_{s(t+1)} = q_{st}/p_{st}$ and $p_{s(t+1)} = 1$. Therefore lemma follows. ■

Definition 25 *Let* $\textsc{Low}(i, s, t, t') = \mathcal{R}^{ex}_{t+1}[i] - \mathcal{R}^{ex}_{t'}[i] + \Upsilon(t+1, s) - \mathcal{R}_t + \mathcal{R}_{t'} - \sum_j \sum_{t''=t'}^{t} \mathbb{W}_{jt''}$.

Lemma 26 *Conditioned on* $X_s = 0$*, s being a special element at time* t'*,* $\textsc{Low}(i, s, t, t')$ *is a submartingale for* $t \geq t'$ *under the coupling* (σ, σ')*.*

Proof Suppose the next play in σ, σ' are u, v respectively. Again there are two cases to consider and in each we show that $\mathbb{E}_{(u,v)}\left[\textsc{Low}(i, s, t+1, t')\right] \geq \textsc{Low}(i, s, t, t')$.

(1) Suppose $u = v$. In this case $\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] = \mathcal{R}_{t+1} - \mathcal{R}_t$. If $u < s$ then we could not have played u before and so $p_{ut} = p_u$. If $X_u = 0$ then we have $\Upsilon(t+2, s) = \frac{\Upsilon(t+1,s)}{1-p_u}$. Otherwise, if $X_u \neq 0$ then $\Upsilon(t+2, s) = 0$ since $q_{s(t+1)} = 0$. Therefore $\mathbb{E}_{X_u}\left[\Upsilon(t+2, s)\right] = \Upsilon(t+1, s)$. If $u > s$ then $\Upsilon(t+2, s) = \Upsilon(t+1, s)$. Therefore irrespective of $u > s$ or $u < s$ it follows that $\mathbb{E}_{(b,c)}\left[\textsc{Low}(i, s, t+1, t')\right] \geq \textsc{Low}(i, s, t, t')$.

(2) The only other case is $u < v$ and $u = s$. Observe that since $X_s = 0$ we have $(\mathcal{R}_{t+1} - \mathcal{R}_t) = 1$. Now v will be the new special element at time $t + 1$. Now

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right] - \mathcal{R}^{ex}_{t+1}[i] + q_{v(t+1)} - (\mathcal{R}_{t+1} - \mathcal{R}_t) = 0$$

where $q_{v(t+1)} = q_{vt}/(1 - p_s)$ is the probability that X_v is the maximum since we already know $X_s = 0$. We now consider the two subcases that (i) X_v has not been observed before and (ii) X_v has been observed before.

Consider the subcase (i) that X_v was not observed before. If $X_v = 0$ then $\Upsilon(t+2, v) = -\frac{q_{v(t+1)}}{p_v}$ and $\mathbb{W}_{v(t+1)} = 0$. Note $\Upsilon(t+1, s) = -\frac{q_{st}}{p_s}$. Therefore;

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] + \Upsilon(t+2,v) - \mathbb{W}_{v(t+1)}\right] - (\mathcal{R}_{t+1} - \mathcal{R}_t) - \Upsilon(t+1,s) \\ = -\frac{q_{v(t+1)}}{p_v} + \frac{q_{st}}{p_s} - q_{v(t+1)} \qquad (6)$$

If $X_v = a_v$ then $\Upsilon(t+2,v) = \mathbb{W}_{v(t+1)} = \frac{q_{v(t+1)}}{p_v}\frac{(1-p_v)}{p_v}$ and so:

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] + \Upsilon(t+2,v) - \mathbb{W}_{t+1}\right] - (\mathcal{R}_{t+1} - \mathcal{R}_t) - \Upsilon(t+1,s) \\ = \frac{q_{st}}{p_s} - q_{v(t+1)} \qquad (7)$$

Therefore, taking the linear combination of the last two equations we get:

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] + \Upsilon(t+2,v) - \mathbb{W}_{v(t+1)}\right] - (\mathcal{R}_{t+1} - \mathcal{R}_t) - \Upsilon(t+1,s) \\ = -\frac{q_{v(t+1)}}{p_v}(1-p_v) + \frac{q_{st}}{p_s} - q_{v(t+1)} = -\frac{q_{v(t+1)}}{p_v} + \frac{q_{st}}{p_s} \qquad (8)$$

Now $\frac{q_{st}}{p_s} = \frac{1}{(1-p_s)}\sum_{j<s} q_{jt} = \sum_{j>s} q_{j(t+1)}$. And $\frac{q_{v(t+1)}}{p_v} = \sum_{j\geq v} q_{j(t+1)}$. Therefore the last equation rewrites to

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] + \Upsilon(t+2,v) - \mathbb{W}_{v(t+1)}\right] - (\mathcal{R}_{t+1} - \mathcal{R}_t) - \Upsilon(t+1,s) \\ = \sum_{v<j<s} q_{j(t+1)} \geq 0$$

Therefore the lemma is true for this subcase (i) because $\mathbb{W}_{j(t+1)} = 0$ for $j \neq v$.

Now consider the subcase (ii) that v had been observed before – this implies that $p_{v(t+1)} = 1$ and $X_v = a_v$. However $\Upsilon(t+2,v) = \mathbb{W}_{v(t+1)} = 0$ in this subcase and it follows that:

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}^{ex}_{t+1}[i] + \Upsilon(t+2,v) - \mathbb{W}_{v(t+1)}\right] - (\mathcal{R}_{t+1} - \mathcal{R}_t) - \Upsilon(t+1,s) = \frac{q_{st}}{p_s} \geq 0$$

Again $\mathbb{W}_{j(t+1)} = 0$ for $j \neq v$. Therefore $\mathbb{E}_{(u,v)}\left[\text{LOW}(i,s,t+1,t')\right] \geq \text{LOW}(i,s,t,t')$ in all of case (2) as well.

The lemma follows. ■

Based on the last two lemmas, we immediately conclude the following (we use a new variable $\bar{t}$ for notational convenience):

Corollary 27 *Let*

$$\text{DIFF}(i,s,\bar{t},t') = \mathcal{R}^{ex}_{\bar{t}+1}[i] - \mathcal{R}^{ex}_{t'}[i] + \Upsilon(\bar{t}+1,s) - \mathcal{R}_{\bar{t}} + \mathcal{R}_{t'} + \sum_j \sum_{t''=t'}^{\bar{t}} \left(\mathbb{M}_{jt''} - \mathbb{W}_{jt''}\right)$$

Then $\text{DIFF}(i,s,\bar{t},t')$ *is a submartingale and in particular* $\text{DIFF}(i,i,\bar{t},1)$ *is a submartingale.*

Proof Follows from the observation that(i) $X_s = a_s$ implies $\textsc{Diff}(i,s,\bar{t},t') = \textsc{High}(i,s,\bar{t},t')$ and all $\mathbb{W}_{jt''} = 0$ along with the conditioning in the statement of Lemma 24 and (ii) $X_s = 0$ implies $\textsc{Diff}(i,s,\bar{t},t') = \textsc{High}(i,s,\bar{t},t')$ and all $\mathbb{M}_{jt''} = 0$ along with the conditioning in Lemma 26. ■

Lemma 28 $\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+1}\right] + \left(\frac{1-p_i}{p_i}\right)(q_{i1} - \mathbb{P}r\left[\mathcal{Z}_i\right]) \geq 0$ *where $\mathcal{Z}_j$ is the event that j is the maximum and there has been no* COLLAPSE *over the entire horizon of t steps.*

Proof Using Corollary 27 and Doob's Optional Stopping Theorem on $\textsc{Diff}(i,i,\bar{t},1)$ for $\bar{t} = t+1$. Note that the stopping time is a fixed horizon $T = t+1$ and does not depend on the knowledge of the maximum arm. We immediately get:

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}_{t+1} + \Upsilon(t+2,s) + \sum_j \sum_{t'=1}^{t+1} \left(\mathbb{M}_{jt'} - \mathbb{W}_{jt'}\right)\right] \geq \mathbb{E}\left[\textsc{Diff}(i,i,1,1)\right] = 0$$

The last part follows from the fact $\mathbb{E}\left[\Upsilon(2,i)\right] = 0$. Now, using the same observation, we note that $\mathbb{E}\left[\Upsilon(t+2,s)\right] = 0$ where s is the final special element. Thus:

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}_{t+1} + \sum_j \sum_{t'=1}^{t+1} \left(\mathbb{M}_{jt'} - \mathbb{W}_{jt'}\right)\right] \geq 0 \tag{9}$$

For any j, let the first time for a play when j declared special be $t(j)$. Note $t(j)$ is a random variable except for $j = i$ when $t(i) = 1$. In the following we show that

$$\mathbb{E}\left[\sum_{t'=t(j)}^{t+1} \mathbb{M}_{jt'} \,\middle|\, t(j) \text{ is defined for } j\right] \leq \left(\frac{1-p_j}{p_j}\right)\left(q_{jt(j)} - \mathbb{P}r\left[\mathcal{Z}_j\right]\right)$$

Observe that $\mathbb{M}_{t'} \neq 0$ implies that we have not yet had a COLLAPSE. Let $g(j)$ be the expected number of plays of type (j',j) starting from $t(j)$ before COLLAPSE or end of horizon given that j is maximum. Therefore:

$$\mathbb{E}\left[\sum_{t'=t(j)}^{t+1} \mathbb{M}_{jt'}\right] = g(j) q_{jt(j)}$$

For any $t \geq t(j)$ the probability mass $\sum_{j'>j} q_{j't} = \frac{q_{jt}}{p_j}(1-p_j)$ till we either had a collapse or end of horizon. Therefore the number of times we see any $j' > j$ before we see j (conditioned on j being the maximum) is $\frac{1-p_j}{p_j}$. If $\mathcal{Z}_j$ is the event that s is the maximum and there has been no COLLAPSE over the entire horizon of t steps then;

$$\left(\frac{1-p_j}{p_j}\right) = g(j) + \mathbb{P}r\left[\mathcal{Z}_j | j \text{ is the maximum}\right]\left(\frac{1-p_j}{p_j}\right)$$

based on the infinite horizon – since the expected number of occurrences of j' before j is unchanged if there has been no collapse. As a consequence, multiplying by $q_{jt(j)}$ and removing the conditioning on j being maximum we have:

$$\mathbb{E}\left[\sum_{t'=t(j)}^{t+1} \mathbb{M}_{jt'}\right] = q_{jt(j)} \left(1 - \mathbb{P}r\left[\mathcal{Z}_j | j \text{ is the maximum}\right]\right) \left(\frac{1-p_j}{p_j}\right) \leq \left(q_{jt(j)} - \mathbb{P}r\left[\mathcal{Z}_j\right]\right) \left(\frac{1-p_j}{p_j}\right)$$

For $j \neq i$, observe that $\mathbb{E}\left[\mathbb{W}_{jt(j)}\right] = \frac{q_{jt(j)}}{p_j}\left(\frac{1-p_j}{p_j}\right) \cdot p_j$ from definition 21. Therefore

$$\mathbb{E}\left[\sum_{t'=t(j)}^{t+1} \mathbb{M}_{jt'}\right] - \mathbb{E}\left[W_{jt(j)}\right] \leq 0$$

and as a consequence $\mathbb{E}\left[\sum_{t'=1}^{t+1}\left(\mathbb{M}_{jt'} - \mathbb{W}_{jt'}\right)\right] \leq 0$ for any $j \neq i$. Therefore Equation!9 is transformed into:

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i] - \mathcal{R}_{t+1} + \sum_{t'=1}^{t+1}\left(\mathbb{M}_{it'} - \mathbb{W}_{it'}\right)\right] \geq 0 \tag{10}$$

At the same time since $\mathbb{W}_{it} = 0$ for all t and $t(i) = 1$ and we get

$$\mathbb{E}\left[\sum_{t'=1}^{t+1}\left(\mathbb{M}_{it'} - \mathbb{W}_{it'}\right)\right] = \left(q_{i1} - \mathbb{P}r\left[\mathcal{Z}_i\right]\right)\left(\frac{1-p_i}{p_i}\right)$$

which when used in Equation 10 proves the lemma. ■

Lemma 29 $\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+2}\right] + (1 - q_{i1}) \geq 0.$

Proof Observe that if i is not the maximum then conditioned on that event,

$$\left(\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right]\right) - \left(\mathbb{E}\left[\mathcal{R}_{t+2}\right] - \mathbb{E}\left[\mathcal{R}_{t+1}\right]\right) \geq 0 \tag{11}$$

If i is the maximum, which happens with probability q_{i1} then if there has been a COLLAPSE then $q_{i(t+1)} \geq q_{i1}/p_i$. Otherwise $q_{i(t+1)} \geq q_{i1}$. Recall (see Lemma 28) that $\mathcal{Z}_i$ is the event that i is the maximum and there has been no collapse. Thus conditioned on i being the maximum:

$$\begin{aligned}
\left(\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right]\right) \quad &- \quad \left(\mathbb{E}\left[\mathcal{R}_{t+2}\right] - \mathbb{E}\left[\mathcal{R}_{t+1}\right]\right) \\
&\geq \quad -1 + \frac{q_{i1}}{p_i}\left(1 - \mathbb{P}r\left[\mathcal{Z}_i | i \text{ is maximum}\right]\right) + q_{i1}\mathbb{P}r\left[\mathcal{Z}_i | i \text{ is maximum}\right] \\
&= \quad -1 + \frac{q_{i1}}{p_i} - \frac{q_{i1}(1-p_i)}{p_i}\mathbb{P}r\left[\mathcal{Z}_i | i \text{ is maximum}\right] \qquad (12)
\end{aligned}$$

Therefore removing the conditioning, using Equations 11 and 12, and the fact that the probability of i being the maximum is q_{i1} we get:

$$\left(\mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right]\right) - \left(\mathbb{E}\left[\mathcal{R}_{t+2}\right] - \mathbb{E}\left[\mathcal{R}_{t+1}\right]\right) \geq -q_{i1} + \frac{q_{i1}^2}{p_i} - \frac{q_{i1}(1-p_i)}{p_i}\mathbb{P}r\left[\mathcal{Z}_i\right] \tag{13}$$

From Lemma 28 we have

$$\mathbb{E}\left[\mathcal{R}^{ex}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+1}\right] \geq -\left(\frac{1-p_i}{p_i}\right)(q_{i1} - \mathbb{P}r\left[\mathcal{Z}_i\right])$$

Summing up we get

$$\begin{aligned} \mathbb{E}\left[\mathcal{R}_{t+2}[i]\right] - \mathbb{E}\left[\mathcal{R}_{t+2}\right] &\geq -\left(\frac{1-p_i}{p_i}\right) q_{i1} - q_{i1} + \frac{q_{i1}^2}{p_i} + \left(\frac{(1-p_i)\mathbb{P}r\left[\mathcal{Z}_i\right]}{p_i}\right)(1-q_{i1}) \\ &\geq -\frac{q_{i1}}{p_i}(1-q_{i1}) \geq -(1-q_{i1}) \end{aligned}$$

which proves the lemma. ■

Theorem 30 *Thompson sampling is a* 2 *approximation of the regret of the optimum finite horizon problem for Bernoulli point priors.*