

# A Broader View of Thompson Sampling

Yanlin Qu

Columbia Business School, qu.yanlin@columbia.edu

Hongseok Namkoong

Columbia Business School, namkoong@gsb.columbia.edu

Assaf Zeevi

Columbia Business School, assaf@gsb.columbia.edu

Thompson Sampling is one of the most widely used and studied bandit algorithms, known for its simple structure, low regret performance, and solid theoretical guarantees. Yet, in stark contrast to most other families of bandit algorithms, the exact mechanism through which posterior sampling (as introduced by Thompson) is able to “properly” balance exploration and exploitation, remains a mystery. In this paper we show that the core insight to address this question stems from recasting Thompson Sampling as an online optimization algorithm. To distill this, a key conceptual tool is introduced, which we refer to as “faithful” stationarization of the regret formulation. Essentially, the finite horizon dynamic optimization problem is converted into a stationary counterpart which “closely resembles” the original objective (in contrast, the classical infinite horizon discounted formulation, that leads to the Gittins index, alters the problem and objective in too significant a manner). The newly crafted time invariant objective can be studied using Bellman’s principle which leads to a time invariant optimal policy. When viewed through this lens, Thompson Sampling admits a simple online optimization form that mimics the structure of the Bellman-optimal policy, and where greediness is regularized by a measure of residual uncertainty based on point-biserial correlation. This answers the question of how Thompson Sampling balances exploration-exploitation, and moreover, provides a principled framework to study and further improve Thompson’s original idea.

*Key words:* Multi-armed Bandit, Bellman Equation, Thompson Sampling, Online Optimization

---

## 1. Introduction

**Background and motivation.** Thompson Sampling is a heuristic algorithm, introduced by [Thompson \(1933\)](#) in the context of solving treatment allocation in medical trials; the objective is to maximize patient outcomes while simultaneously learning the best treatment. This motivating application has since been abstracted to what we recognize today as the multi-armed bandit (MAB) problem. The algorithm proceeds in each round to sample from the posterior distribution, the updated belief over problem parameters, and then select the treatment (arm) that is perceived to be optimal in the sampled environment.

While Thompson Sampling remained obscure throughout the 20th century, the MAB problem has attracted significant attention ever since it was formalized by [Robbins \(1952\)](#). In addition to formalizing the problem, that paper made a foundational observation about the tension between exploration and exploitation: any procedure aiming to maximize long-run average reward must

explore all arms infinitely often. In continuation of this principle, a landmark paper by [Lai and Robbins \(1985\)](#) introduced the notion of *regret*, the loss incurred by a policy relative to an oracle that knows the identity of the best arm, and proposed a policy that carefully assigns (infinitely many) pulls to each arm to achieve the minimal possible growth rate of regret. This policy was later simplified to the Upper Confidence Bound (UCB) algorithm, popularized by [Auer et al. \(2002a\)](#).

Close to a decade after the [Auer et al. \(2002a\)](#) paper, Thompson Sampling was finally resurrected, triggered by several studies that indicated remarkably strong empirical performance (e.g., [Scott 2010](#), [Chapelle and Li 2011](#)), often rivaling or even surpassing that of UCB. Since then, practitioners have applied Thompson Sampling across a wide range of domains, including online advertising (e.g., [Agarwal 2013](#)), recommendation systems (e.g., [Kawale et al. 2015](#)), and website optimization (e.g., [Hill et al. 2017](#)). Meanwhile, a substantial body of theoretical work has been developed to bound the regret of Thompson Sampling, essentially showing that it achieves the goal of long-term regret minimization; see the frequentist regret bounds in [Agrawal and Goyal \(2012, 2013\)](#) and Bayesian regret bounds in [Russo and Van Roy \(2014b, 2016\)](#).

The aforementioned theory introduced several innovative ideas and technical tools that extend beyond Thompson Sampling. However, in contrast to upper confidence bound policies (in particular the simplified version and proofs in [Auer et al. \(2002a\)](#)), and variants thereof such as explore-then-commit, epsilon-greedy and the like (see [Lattimore and Szepesvári \(2020\)](#)), said theory falls short of elucidating the key optimization principle or at least the explicit exploration-exploitation tradeoffs that guide Thompson Sampling.

To that end, it is worth noting that neither Thompson Sampling nor the UCB family are derived from first principles such as dynamic programming ([Bellman 1957](#)). A key illustration of the latter is the Gittins index policy ([Gittins 1979](#)), which formulates the Bayesian version of the MAB problem as a Markov decision process (MDP), and derives the optimal policy that maximizes expected cumulative discounted reward. While Discounting simplifies the problem by making it stationary, in contrast to the traditional finite horizon regret setting, it also results in a significant deviation from the intuitive principle laid out by [Robbins \(1952\)](#). Specifically, the Gittins index policy may pull the optimal arm only finitely many times, and hence fail to “identify” it, resulting in performance dramatically inferior to that of Thompson Sampling (and UCB) over longer horizons.

In this paper, akin to Gittins, we aim to harness Bellman’s more principled approach to shed further light on the optimization considerations underlying Thompson Sampling. But toward that end, and to remain within the traditional finite horizon regret formulation, where the success of Thompson Sampling was established and validated, we depart from Gittins’ infinite horizon discounted reward formulation. In lieu of that, we propose a different form of stationarization, which is more “faithful” to Robbin’s original principle, and show that through this lens, Thompson

Sampling takes the form of an online optimization algorithm that at each step balances between greediness and a measure of residual uncertainty which serves as a *regularizer*. Beyond addressing the core question of what Thompson Sampling optimizes, it also provides a principled framework for further study and improvement to this important class of posterior sampling algorithms.

**Main contributions and overview of key ideas.** We first describe our proposed notion of “faithful” stationarization of the long-term regret minimization problem, as this holds the key to explaining Thompson Sampling through an optimization lens. For simplicity of exposition, and to stay true to Thompson’s original 1933 setup, we consider a two-armed bandit with independent arms, each generating random rewards when pulled. (The principles and key ideas carry over to the  $K$ -armed case, as discussed before Theorem 1.) The learner’s goal is maximizing expected cumulative reward, or equivalently minimizing expected cumulative regret

$$\mathcal{R}_T(Q; \pi_0) = \mathbb{E}_{\pi_0} \left[ \sum_{t=0}^{T-1} (\max(\theta_1, \theta_2) - \theta_{A_t}) \right], \quad (1)$$

where  $Q$  is a policy,  $T$  is a finite time horizon,  $\theta_k$  is the mean reward of arm  $k$ , and  $A_t$  is the arm chosen at time  $t$ . In the Bayesian setting, the expectation is taken over the randomness of interaction (rewards observed and arms pulled) and environment, as the unknown parameter  $\theta = (\theta_1, \theta_2)$  is drawn from a prior distribution  $\pi_0$  before the game begins.

As noted by Gittins (1979), this bandit problem becomes a Markov decision process (MDP) when posterior distributions  $\pi_1, \pi_2, \dots$  are viewed as states. For MDPs, perhaps the most principled framework for optimizing performance is dynamic programming, typically expressed through Bellman equations. In particular, stationary Bellman equations (e.g., infinite horizon with discounting) are typically more tractable than their non-stationary counterparts (e.g., finite horizon). For example, maximizing expected cumulative discounted reward

$$\mathbb{E}_{\pi_0} \left[ \sum_{t=0}^{\infty} \gamma^t \theta_{A_t} \right], \quad \gamma \in (0, 1) \quad (2)$$

leads to the elegant optimal policy known as the Gittins index. However, as mentioned earlier, discounted (2) and non-discounted (1) are fundamentally different objectives. This discrepancy has significant consequences. In fact, the Gittins index policy, despite maximizing (2), can suffer linear regret, i.e., (1) grows linearly in  $T$ ; see, e.g., Rothschild (1974). To obtain a stationary Bellman equation that is faithful to minimizing (1), we consider minimizing expected cumulative *squared regret*

$$\mathcal{R}^2(Q; \pi_0) = \mathbb{E}_{\pi_0} \left[ \sum_{t=0}^{\infty} r^2(q_t; \pi_t) \right], \quad (3)$$

where  $q_t$  is the distribution of  $A_t|\pi_t$  under policy  $Q$ , and  $r(q_t; \pi_t) = \mathbb{E}_{\pi_t} [\max(\theta_1, \theta_2) - \theta_{A_t}]$  is the expected next-round regret. This new objective is aligned with the original one in the sense that minimizing  $\mathcal{R}^2(Q; \pi_0)$  minimizes the following regret bound

$$\mathcal{R}_T(Q; \pi_0) \leq \sqrt{\mathcal{R}^2(Q; \pi_0) \cdot T}. \quad (4)$$

The  $\mathcal{R}^2$ -optimal policy, characterized by the corresponding stationary Bellman equation, turns out to have an online optimization form (derived later in the paper), expressed as follows

$$x^*(\pi) = \operatorname{argmin}_x [\bar{r}^2(x; \pi) + \nu(\pi)x],$$

where  $\pi$  is the current belief (with the time subscript omitted), and the decision variable  $x = q \cdot \mathbb{E}_\pi \theta$  is the expected next-round reward. Since there are only two arms, selecting  $x$  within the interval between  $\mathbb{E}_\pi \theta_1$  and  $\mathbb{E}_\pi \theta_2$  amounts to selecting the probability of pulling arm 1. The function  $\bar{r}^2(x; \pi) = (\mathbb{E}_\pi \max(\theta_1, \theta_2) - x)^2 = (\mathbb{E}_\pi \max(\theta_1, \theta_2) - q \cdot \mathbb{E}_\pi \theta)^2 = r^2(q; \pi)$  is the square of the expected next-round regret, and its minimization is regularized by the linear term  $\nu(\pi)x$ , where the regularizer  $\nu(\pi)$  is determined by the solution to the stationary Bellman equation (i.e., the optimal value function). Intuitively,  $\nu(\pi)$  should measure the remaining uncertainty about which arm is better, in order to adaptively regularize the greediness that would result from minimizing  $\bar{r}^2(x; \pi)$  alone. The greater the uncertainty, the stronger the incentive to explore. In addition,  $\nu(\pi)$  has the same unit as the reward, keeping the online objective dimensionally homogeneous.

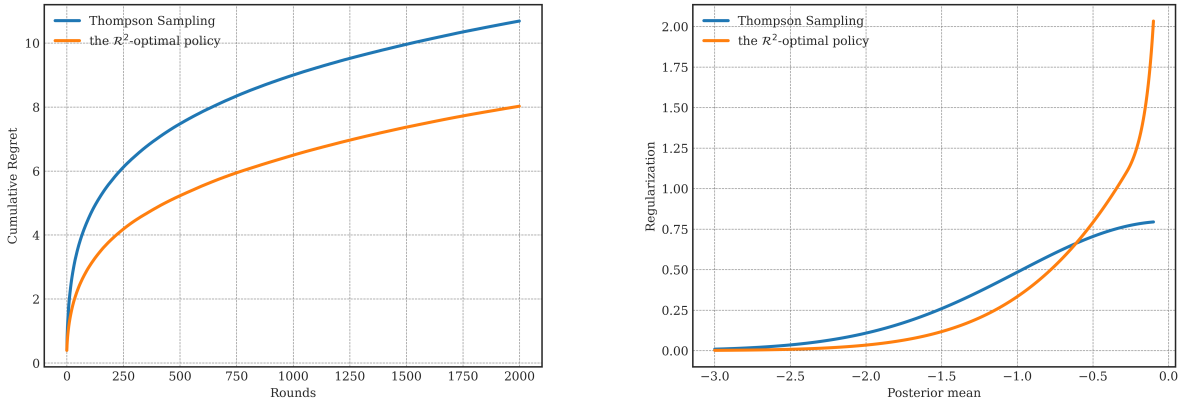
With the online optimization form of the  $\mathcal{R}^2$ -optimal policy in hand, Thompson Sampling can be expressed in similar form

$$x^{\text{TS}}(\pi) = \operatorname{argmin}_x [\bar{r}^2(x; \pi) + \tilde{\nu}(\pi)x],$$

where  $\tilde{\nu}(\pi) = \text{Cov}_\pi(\theta_1 - \theta_2, \text{sign}(\theta_1 - \theta_2))$ . The regularizer of Thompson Sampling turns out to be the covariance between *the reward gap* and *the identity of the optimal arm*. The study of the relationship between a metric (continuous) variable and a dichotomous (binary) variable dates back to [Pearson \(1909\)](#), and the standard formula for the point-biserial correlation makes explicit how Thompson Sampling measures the remaining uncertainty (about which arm is better) in the same unit as the reward.

Thompson Sampling can now be viewed as a member of the family of  $\mathcal{R}^2$ -driven online optimization algorithms, characterized by its distinctive regularizer based on the “biserial” covariance. It is natural to compare it with the  $\mathcal{R}^2$ -optimal policy, characterized by its oracle regularizer based on the stationary Bellman equation. The left panel of [Figure 1](#) compares their cumulative regret, revealing a concrete gap between what Thompson Sampling achieves versus its optimally designed

counterpart. This gap also confirms that minimizing  $\mathcal{R}^2(Q; \pi_0)$  is indeed a faithful surrogate for minimizing the sequence  $\{\mathcal{R}_T(Q; \pi_0) : T \geq 1\}$ . What, then, is to blame for the suboptimality of Thompson Sampling? Thanks to the online optimization form shared by the two policies, this question can be addressed by comparing their regularizers  $\nu$  and  $\tilde{\nu}$ . The right panel of Figure 1 compares the two regularizers as the reward gap shrinks (arm 1 has slightly lower posterior mean but much higher posterior variance than arm 2), showing that the regularizer of Thompson Sampling remains conservative even as exploring arm 1 (with probability 1) becomes clearly the right thing to do. In fact, as the benefit of having a principled framework with a well-defined benchmark, we can not only identify but also address such issues with Thompson Sampling, as will be illustrated at the end of this paper.



**Figure 1** Thompson Sampling and the  $\mathcal{R}^2$ -optimal policy play a Gaussian bandit with reward variance 1. Left: comparing their cumulative regret  $\mathcal{R}_T(Q^{\text{TS}}; \pi_0)$  vs.  $\mathcal{R}_T(Q^*; \pi_0)$  where  $\pi_0 = N(0, 1) \times N(0, 0)$  (20K trials). Right: comparing the two regularizers  $\tilde{\nu}(N(\mu, 1) \times N(0, 0))$  vs.  $\nu(N(\mu, 1) \times N(0, 0))$  where  $\mu$  approaches 0 from below.

The rest of the paper is organized as follows: In Section 2, we review the Bayesian MAB problem and related algorithms. In Section 3, we introduce a “faithful” stationarization of the long-term regret minimization problem. In Section 4, we rediscover Thompson Sampling as an online optimization algorithm addressing the stationarized problem. In Section 5, we illustrate how the regularizer of Thompson Sampling measures uncertainty and guides exploration. In Section 6, we compare Thompson Sampling with the optimal policy that solves the stationarized problem.

## 2. Preliminaries

### 2.1. Bayesian Stochastic Bandit as an MDP

To begin, let us recall the mechanism of a two-armed Bayesian stochastic bandit. The two arms are labeled with 1 and 2. Their joint reward distribution  $P_\theta$  depends on an (unknown) environment

parameter  $\theta \in \Theta$ . Before the game begins,  $\theta$  is drawn from a prior distribution  $\pi_0$  and remains fixed throughout the game. At each round, a potential reward vector is drawn independently from  $P_\theta$ , but only the entry corresponding to the pulled arm is observed. Conditional on  $\theta$ , these potential reward vectors form an independent and identically distributed (iid) sequence

$$R_1|\theta, R_2|\theta, \dots \stackrel{\text{iid}}{\sim} P_\theta, \theta \sim \pi_0.$$

After  $t$  rounds, each involving a partial observation of a potential reward vector, the posterior distribution  $\pi_t$  of  $\theta$  is obtained by updating  $\pi_0$  according to Bayes' rule. For simplicity of exposition, we take the environment parameter to be the mean reward vector

$$\mathbb{E}[R_1|\theta] = \mathbb{E}[(R_{1,1}, R_{2,1})|\theta] = (\theta_1, \theta_2) = \theta.$$

As noted by [Gittins \(1979\)](#), the Bayesian stochastic bandit can be viewed as a Markov decision process (MDP); see page 31 of [Ghavamzadeh et al. \(2015\)](#) for an illustrative example. The MDP formulation is as follows:

- State: the current belief  $\pi_t$ .
- Action: selecting one of the two arms to pull, i.e.,  $A_t = 1$  or  $2$ .
- Transition: updating  $\pi_t$  to  $\pi_{t+1}$  after observing the  $A_t$ -th entry of

$$R_{t+1} \sim P_{\theta''}, \theta'' \sim \pi_t.$$

- Reward (in the MDP): the expected reward of arm  $A_t$  under  $\pi_t$ .

Note that the next potential reward vector is drawn from the posterior predictive distribution, so the system can evolve forward as a Markov chain without knowing which  $\theta$  was drawn and fixed at the very beginning. This MDP formulation provides a natural framework for analyzing Bayesian bandit algorithms directly, without resorting to frequentist analysis followed by integration over the prior. We adopt this formulation and focus on stationary Markov policies, where the current state  $\pi_t$  determines the distribution of the next action  $A_t$  in a time-invariant manner.

## 2.2. Thompson Sampling and the Gittins Index Policy

We provide a brief review of Thompson Sampling ([Thompson 1933](#)) and the Gittins index policy ([Gittins 1979](#)), both examples of stationary Markov policies. These two policies offer a sharp contrast: Thompson Sampling is heuristic but surprisingly effective, while the Gittins index policy is optimal by design but not in the “usual” sense.

**Thompson Sampling.** Recall that  $\pi_t$  is the posterior distribution of  $\theta$  after  $t$  rounds. For the next round, Thompson Sampling draws  $\theta'$  from  $\pi_t$  and selects arm  $A_t = \text{argmax}(\theta'_1, \theta'_2)$  as if  $\theta'$  were

the true mean reward vector. Despite having access to the posterior distribution that encapsulates all available information about the environment, Thompson Sampling merely draws a single sample and acts greedily with respect to it. No objective is defined, and no optimization is performed. Decades later, long-term regret minimization (minimizing (1) for large  $T$ ) became the standard goal for bandit algorithms. Decades later still, Thompson Sampling was shown to achieve  $O(\sqrt{T})$  regret (Russo and Van Roy 2016). Throughout this paper, we use the term “regret” to refer to the Bayesian regret defined in (1), unless otherwise specified.

Thanks to the MDP formulation, by sampling the next reward from the posterior predictive distribution, Thompson Sampling runs forward as a Markov chain; see Algorithm 1 (Gaussian rewards with Gaussian prior and posterior) and Algorithm 2 (Bernoulli rewards with Beta prior and posterior).

---

**Algorithm 1** Thompson Sampling (Gaussian)

---

**Initialize:**  $N(\mu_1, \sigma_1^2)$ ,  $N(\mu_2, \sigma_2^2)$ ,  $\tau^2$ ,  $T$

**for**  $t = 1, 2, \dots, T$  **do**

    Sample

$$(\theta'_1, \theta'_2) \sim N(\mu_1, \sigma_1^2) \times N(\mu_2, \sigma_2^2)$$

    Select  $A = \operatorname{argmax}(\theta'_1, \theta'_2)$

    Observe  $R \sim N(\mu_A, \sigma_A^2 + \tau^2)$

    Update

$$\mu_A \leftarrow (\mu_A / \sigma_A^2 + R / \tau^2) / (1 / \sigma_A^2 + 1 / \tau^2)$$

$$\sigma_A^2 \leftarrow 1 / (1 / \sigma_A^2 + 1 / \tau^2)$$

**end for**

---



---

**Algorithm 2** Thompson Sampling (Bernoulli)

---

**Initialize:**  $\operatorname{Beta}(\alpha_1, \beta_1)$ ,  $\operatorname{Beta}(\alpha_2, \beta_2)$ ,  $T$

**for**  $t = 1, 2, \dots, T$  **do**

    Sample

$$(\theta'_1, \theta'_2) \sim \operatorname{Beta}(\alpha_1, \beta_1) \times \operatorname{Beta}(\alpha_2, \beta_2)$$

    Select  $A = \operatorname{argmax}(\theta'_1, \theta'_2)$

    Observe  $R \sim \operatorname{Ber}(\alpha_A / (\alpha_A + \beta_A))$

    Update

$$\alpha_A \leftarrow \alpha_A + R$$

$$\beta_A \leftarrow \beta_A + (1 - R)$$

**end for**

---

**Gittins index policy.** While Thompson Sampling earned justification decades after its invention, the Gittins index policy was designed to be optimal from the start, but for an objective that differs from the now-standard one. Assuming the two arms are independent (i.e.,  $\pi_t = \pi_{1,t} \times \pi_{2,t}$ ), the expected cumulative discounted reward defined in (2) is maximized by always selecting the arm with the highest Gittins index

$$G_k(\pi_{k,t}) = \sup_{\tau \geq 1} \frac{\mathbb{E}_{\pi_{k,t}} \left[ \sum_{s=0}^{\tau-1} \gamma^s r(\pi_{k,t+s}) \right]}{\mathbb{E}_{\pi_{k,t}} \left[ \sum_{s=0}^{\tau-1} \gamma^s \right]}, \quad k = 1, 2$$

where  $r(\pi_{k,t+s}) = \mathbb{E}_{\pi_{k,t+s}} \theta_k$  is the posterior mean reward, and the supremum is taken over stopping times. Given its optimality, the Gittins index policy has to coincide with the solution to the Bellman equation associated with maximizing (2)

$$\bar{V}(\pi_t) = \max_{q_t} [q_t \cdot \mathbb{E}_{\pi_t} \theta + \gamma \mathbb{E}_{\pi_t, q_t} \bar{V}(\pi_{t+1})], \quad (5)$$

where  $q_t$  is the distribution of the next action, and

$$\mathbb{E}_{\pi_t, q_t} \bar{V}(\pi_{t+1}) = q_t \cdot \mathbb{E}_{\pi_t} [\bar{V}(\pi_{t+1}) | A_t = \cdot].$$

This Bellman equation is stationary in the sense that the optimal action depends on  $\pi_t$  but not  $t$ . Although the equation itself does not reveal the Gittins index policy, its stationarity is what makes the existence of such an elegant optimal policy possible. However, as discussed in the introduction, discounting is not a “faithful” way to stationarize the problem of *long-term* regret minimization.

### 3. Faithful Stationarization

#### 3.1. A Notion of Squared Regret

Long-term regret minimization means minimizing the sequence  $\{\mathcal{R}_T(Q; \pi_0) : T \geq 1\}$  introduced in (1). To stationarize this problem “faithfully”, we seek a single quantity that aggregates the sequence in such a way that its minimization implies the minimization of the sequence as a whole (e.g., via a regret bound). The quantity we choose is the squared regret  $\mathcal{R}^2(Q; \pi_0)$  defined in (3). To motivate this choice, we now briefly derive the regret bound  $\mathcal{R}_T(Q; \pi_0) \leq \sqrt{\mathcal{R}^2(Q; \pi_0) \cdot T}$  stated in (4). By conditioning the  $t$ -th term of (1) on  $\pi_t$ , we have

$$\mathcal{R}_T(Q; \pi_0) = \mathbb{E}_{\pi_0} \left[ \sum_{t=0}^{T-1} r(q_t; \pi_t) \right],$$

where  $q_t$  is the distribution of  $A_t | \pi_t$  under policy  $Q$ , and  $r(q_t; \pi_t) = \mathbb{E}_{\pi_t} [\max(\theta_1, \theta_2) - \theta_{A_t}]$  is the expected next-round regret conditional on  $\pi_t$ . By the Cauchy–Schwarz inequality followed by Jensen’s inequality, we have

$$\begin{aligned} \mathcal{R}_T(Q; \pi_0) &\leq \mathbb{E}_{\pi_0} \left[ \left( \sum_{t=0}^{T-1} 1 \right)^{1/2} \left( \sum_{t=0}^{T-1} r^2(q_t; \pi_t) \right)^{1/2} \right] \\ &\leq \sqrt{T} \cdot \left( \mathbb{E}_{\pi_0} \left[ \sum_{t=0}^{T-1} r^2(q_t; \pi_t) \right] \right)^{1/2} \\ &\leq \sqrt{\mathcal{R}^2(Q; \pi_0) \cdot T}. \end{aligned}$$

As a direct corollary of the information-theoretic analysis in [Russo and Van Roy \(2016\)](#), Thompson Sampling achieves finite squared regret and therefore enjoys the  $O(\sqrt{T})$  regret bound above.



**PROPOSITION 1 ( $\mathcal{R}^2$ -finiteness of Thompson Sampling).** *If there exists a finite constant  $\sigma > 0$  such that the posterior predictive distribution of the reward ( $R_{t+1} \sim P_{\theta''}$ ,  $\theta'' \sim \pi_t$ ) is always  $\sigma$ -sub-Gaussian, then Thompson Sampling satisfies  $\mathcal{R}^2(Q^{\text{TS}}; \pi_0) < \infty$ , and hence  $\mathcal{R}_T(Q^{\text{TS}}; \pi_0) = O(\sqrt{T})$ .*

We now have several reasons to believe that squared regret leads to a faithful stationarization of the long-term regret minimization problem.

- Reasonable algorithms such as Thompson Sampling achieve finite squared regret.
- $\mathcal{R}^2(Q; \pi_0)$  directly controls the growth rate of  $\mathcal{R}_T(Q; \pi_0)$  via the  $O(\sqrt{T})$  regret bound.
- $O(\sqrt{T})$  is minimax optimal (Auer et al. 2002b), which corresponds to  $\mathcal{R}^2$  but not  $\mathcal{R}^{1.9}$  or  $\mathcal{R}^{2.1}$ .

These observations make squared regret a natural and meaningful objective to minimize.

### 3.2. Another Stationary Bellman Equation

To minimize squared regret, we derive the corresponding Bellman equation. Let  $V(\pi_t) = \mathcal{R}^2(Q^*; \pi_t)$  be the minimal squared regret incurred from  $\pi_t$  onward, achieved by the  $\mathcal{R}^2$ -optimal policy  $Q^*$ . The corresponding Bellman equation is

$$V(\pi_t) = \min_{q_t} [r^2(q_t; \pi_t) + \mathbb{E}_{\pi_t, q_t} V(\pi_{t+1})], \quad (6)$$

where  $q_t$  is the distribution of the next action, and

$$\mathbb{E}_{\pi_t, q_t} V(\pi_{t+1}) = q_t \cdot \mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = \cdot].$$

This Bellman equation is stationary in the sense that the optimal action depends on  $\pi_t$  but not  $t$ . Next, we briefly compare the two stationary Bellman equations corresponding to the Gittins index policy and the  $\mathcal{R}^2$ -optimal policy, given by (5) and (6), respectively.

**Both are finite.** The solution to (5) is the maximal expected cumulative discounted reward, which is finite due to *extrinsic* geometric discounting. In contrast, the solution to (6) is the minimal expected cumulative squared regret, which is finite due to *intrinsic* regret decay.

**The discounted one is indexable.** Note that the function being maximized in (5) is linear in  $q_{1,t}$ , so the maximizer is either 1 or 0, determined by which arm has the highest Gittins index. In contrast, the function being minimized in (6) is quadratic in  $q_{1,t}$ , so the minimizer can be in  $(0, 1)$ , i.e., the  $\mathcal{R}^2$ -optimality cannot be achieved by any deterministic index policy.

**The squared one is faithful.** As discussed in the introduction, the Gittins index policy can lead to linear regret, i.e.,  $\mathcal{R}_T(Q; \pi_0) = \Theta(T)$ , in certain settings; see, e.g., Rothschild (1974). In contrast, the  $\mathcal{R}^2$ -optimal policy satisfies the regret bound  $\mathcal{R}_T(Q; \pi_0) \leq \sqrt{\mathcal{R}^2(Q; \pi_0) \cdot T}$  with the best possible constant  $\sqrt{V(\pi_0)}$ . How can this faithful stationarization deepen our understanding of Thompson Sampling and extend our insights beyond it?

**REMARK 1.** In this paper, we focus on the  $\mathcal{R}^2$ -stationarization. There are other ways to stationarize the problem. Whether the  $\mathcal{R}^2$ -stationarization is the best in some sense is left for future research.

## 4. Online Optimization Form

### 4.1. The $\mathcal{R}^2$ -optimal Policy

The stationary Bellman equation (6) not only gives rise to the  $\mathcal{R}^2$ -optimal policy but also grants it an online optimization form, which we now derive. We choose the expected next-round reward

$$x_t = q_t \cdot \mathbb{E}_{\pi_t} \theta = q_{1,t} \mathbb{E}_{\pi_t} \theta_1 + q_{2,t} \mathbb{E}_{\pi_t} \theta_2$$

as the decision variable, as choosing  $q_{1,t}$  or  $q_{2,t}$  would break the symmetry between the two arms. When  $\mathbb{E}_{\pi_t} \theta_1 \neq \mathbb{E}_{\pi_t} \theta_2$ , the possible values of  $x_t$  span an interval, and each point in this interval corresponds to a unique pair

$$q_{1,t} = \frac{x_t - \mathbb{E}_{\pi_t} \theta_2}{\mathbb{E}_{\pi_t} \theta_1 - \mathbb{E}_{\pi_t} \theta_2}, \quad q_{2,t} = \frac{\mathbb{E}_{\pi_t} \theta_1 - x_t}{\mathbb{E}_{\pi_t} \theta_1 - \mathbb{E}_{\pi_t} \theta_2}.$$

By a change of variables in (6), the  $\mathcal{R}^2$ -optimal policy

$$q_t^* = \underset{q_t}{\operatorname{argmin}} \left[ r^2(q_t; \pi_t) + q_t \cdot \mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = \cdot] \right] \quad (7)$$

becomes

$$x_t^* = \underset{x_t}{\operatorname{argmin}} \left[ (\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - x_t)^2 + \nu(\pi_t) x_t \right], \quad (8)$$

where  $\nu(\pi_t)$  is given by

$$\frac{\mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = 1] - \mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = 2]}{\mathbb{E}_{\pi_t} \theta_1 - \mathbb{E}_{\pi_t} \theta_2}. \quad (9)$$

In the online optimization form (8), the objective consists of two terms: an instantaneous regret term for exploitation and a linear regularization term for exploration. The greediness that would result from minimizing the first term alone is regularized by the second term when  $\nu > 0$ . We call  $\nu$  the regularizer. According to (9), the regularizer is positive when there is clear tension between exploration and exploitation: pulling one arm yields higher immediate mean reward (e.g.,  $\mathbb{E}_{\pi_t} \theta_1 > \mathbb{E}_{\pi_t} \theta_2$ ), which favors exploitation, but pulling the other yields lower future squared regret (e.g.,  $\mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = 2] < \mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = 1]$ ), which favors exploration. The  $\mathcal{R}^2$ -optimal policy quantifies this tension as an *exploration-exploitation ratio* (9) and incorporates it as the regularizer in (8).

Next, we consider the case where  $\mathbb{E}_{\pi_t} \theta_1 = \mathbb{E}_{\pi_t} \theta_2$ . Since the two arms have the same posterior mean reward, the instantaneous squared regret  $r^2(q_t; \pi_t)$  in (7) becomes constant with respect to  $q_t$ . As a result, the  $\mathcal{R}^2$ -optimal policy  $q_t^*$  places all its probability mass on the arm that yields lower future squared regret  $\mathbb{E}_{\pi_t} [V(\pi_{t+1}) | A_t = \cdot]$ . This is clearly the right thing to do: when the two arms appear equally rewarding on average, we should pull the more uncertain one to learn more about it. The more we learn, the less we regret.

REMARK 2. When  $\mathbb{E}_{\pi_t} \theta_1 = \mathbb{E}_{\pi_t} \theta_2$ , the range of  $x_t$  collapses to a single point, but we can still recover  $q_t^*$  from  $x_t^*$  by imagining an infinitesimal difference between the two posterior mean rewards.

## 4.2. Thompson Sampling

Instead of the  $\mathcal{R}^2$ -optimal policy itself, perhaps its online optimization form in (8) is more valuable. It reveals what a reasonable bandit algorithm should look like when the MAB problem is viewed through the lens of online optimization: minimizing the instantaneous squared regret with some linear regularization. We now show that Thompson Sampling also takes this form, focusing on the two-armed case

$$q_t^{\text{TS}} = (P_{\pi_t}(\theta_1 > \theta_2), P_{\pi_t}(\theta_1 \leq \theta_2)),$$

as the  $K$ -armed case can be viewed as repeating the two-armed case  $K$  times (to determine the  $K$  pulling probabilities)

$$q_{1,t}^{\text{TS}} = P_{\pi_t}(\theta_1 > \theta_2, \dots, \theta_K) = P_{\pi_t}(\theta_1 > \theta_{-1}),$$

where  $\theta_{-1} = \max\{\theta_2, \dots, \theta_K\}$  can be viewed as a single competing arm against  $\theta_1$ . All proofs are deferred to Section 7.

**THEOREM 1 (Online optimization).** *The online optimization form of Thompson Sampling is*

$$x_t^{\text{TS}} = \underset{x_t}{\operatorname{argmin}} [\bar{r}^2(x_t; \pi_t) + \tilde{\nu}(\pi_t)x_t],$$

where  $x_t^{\text{TS}} = q_t^{\text{TS}} \cdot \mathbb{E}_{\pi_t} \theta$ ,  $\bar{r}(x_t; \pi_t) = \mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - x_t$ , and  $\tilde{\nu}(\pi_t) = \operatorname{Cov}_{\pi_t}(\theta_1 - \theta_2, \operatorname{sign}(\theta_1 - \theta_2))$ .

Note that the regularizer of Thompson Sampling is the covariance between the following two fundamental quantities

$\Delta = \theta_1 - \theta_2$  the reward gap between the two arms

$\Lambda = \operatorname{sign}(\theta_1 - \theta_2)$  the identity of the optimal arm.

The study of the relationship between a metric variable and a dichotomous variable dates back to [Pearson \(1909\)](#), and the “biserial” covariance has a well-known expression; see, e.g., [Lev \(1949\)](#).

**PROPOSITION 2 (Covariance factorization).** *If  $\operatorname{Var}_{\pi_t} \Lambda = 0$ , then  $\operatorname{Cov}_{\pi_t}(\Delta, \Lambda) = 0$ . Otherwise,*

$$\frac{\operatorname{Cov}_{\pi_t}(\Delta, \Lambda)}{\operatorname{Var}_{\pi_t} \Lambda} = \frac{\mathbb{E}_{\pi_t}[\Delta | \Delta > 0] - \mathbb{E}_{\pi_t}[\Delta | \Delta \leq 0]}{2}.$$

Recall that the regularizer in the online optimization form (8) should measure the remaining uncertainty (about which arm is better) in the same unit as the reward, so that it can adaptively regularize greediness while keeping the online objective dimensionally homogeneous. This requirement is met by the regularizer of Thompson Sampling. In the factorization of  $\operatorname{Cov}_{\pi_t}(\Delta, \Lambda)$ , the unit-less variance  $\operatorname{Var}_{\pi_t} \Lambda$  captures the uncertainty in identifying the optimal arm, and it is converted into the reward scale by an interesting notion of regret, namely the average of two terms:

the expected regret from pulling arm 2 conditional on arm 1 being better, and the expected regret from pulling arm 1 conditional on arm 2 being better.

Typically, an online objective consists of three components: a loss term, a regularization term, and a parameter that balances the two (Lagrange multiplier). In Theorem 1, the multiplier is 1. A natural question is whether Thompson Sampling can be improved by adjusting this multiplier. The answer is no.

**PROPOSITION 3 (Incomplete learning).** *For each  $\lambda \neq 1$ , there exists a prior under which the policy*

$$x_t^\lambda = \underset{x_t}{\operatorname{argmin}} [\bar{r}^2(x_t; \pi_t) + \lambda \tilde{\nu}(\pi_t) x_t]$$

*suffers from incomplete learning, i.e., it fully commits to one arm while the other still has a chance of being better.*

To conclude this section, we present a new description of Thompson Sampling in the language of online optimization. At each round, Thompson Sampling minimizes the instantaneous squared regret adaptively regularized by the biserial covariance.

- The loss term (squared regret) corresponds to the faithful stationarization.
- The linear regularization format is determined by the stationary Bellman equation.
- The regularizer  $\tilde{\nu}$  measures the remaining uncertainty in the same unit as the reward.
- The Lagrange multiplier must be 1 to avoid incomplete learning.

**REMARK 3.** Note that Information-Directed Sampling (IDS) (Russo and Van Roy 2014a) is also  $\mathcal{R}^2$ -driven and follows the online optimization form (8)

$$\begin{aligned} x_t^{\text{IDS}} &= \underset{x_t}{\operatorname{argmin}} \frac{\bar{r}^2(x_t; \pi_t)}{\mathcal{I}(x_t; \pi_t)} \\ &= \underset{x_t}{\operatorname{argmin}} [\bar{r}^2(x_t; \pi_t) + \bar{\lambda}(\pi_t) \bar{\nu}(\pi_t) x_t]. \end{aligned}$$

Here,  $\mathcal{I}(x_t; \pi_t)$  is the “information gain” from executing  $x_t$ . The multiplier  $\bar{\lambda}(\pi_t)$  is the minimized information ratio. The regularizer

$$\bar{\nu}(\pi_t) = \frac{\mathcal{I}(\mathbb{E}_{\pi_t} \theta_2; \pi_t) - \mathcal{I}(\mathbb{E}_{\pi_t} \theta_1; \pi_t)}{\mathbb{E}_{\pi_t} \theta_1 - \mathbb{E}_{\pi_t} \theta_2}$$

is positive when one arm gives more reward while the other gives more information.

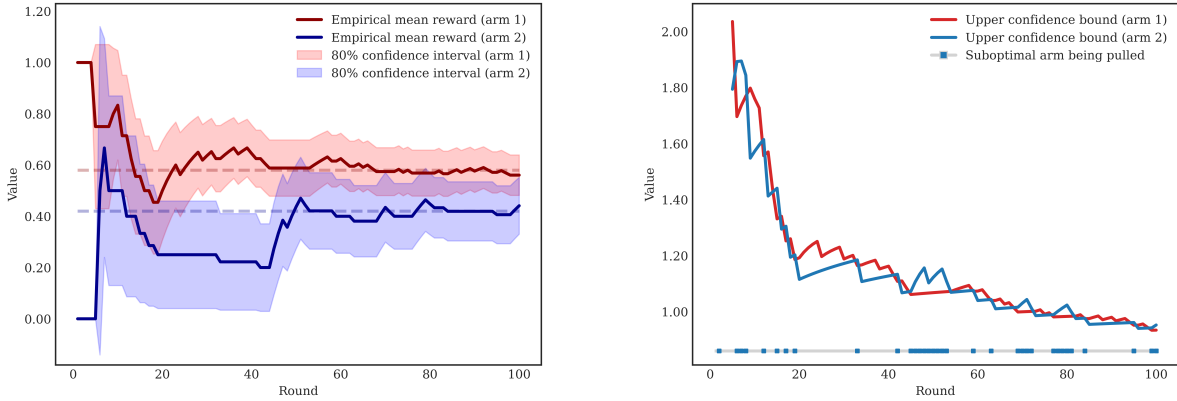
## 5. Uncertainty, Exploration, and Regularizer

The online optimization form of Thompson Sampling reveals that it measures uncertainty through the biserial covariance to guide exploration. Since the Upper Confidence Bound (UCB) algorithm (Auer et al. 2002a) is well known to measure uncertainty via confidence intervals, we can now draw a side-by-side comparison of their exploration philosophies.

Recall that UCB is given by

$$A_t = \operatorname{argmax}_{k \in \{1,2\}} \left( \hat{\mu}_{k,t} + \sqrt{\frac{2 \log t}{N_{k,t}}} \right),$$

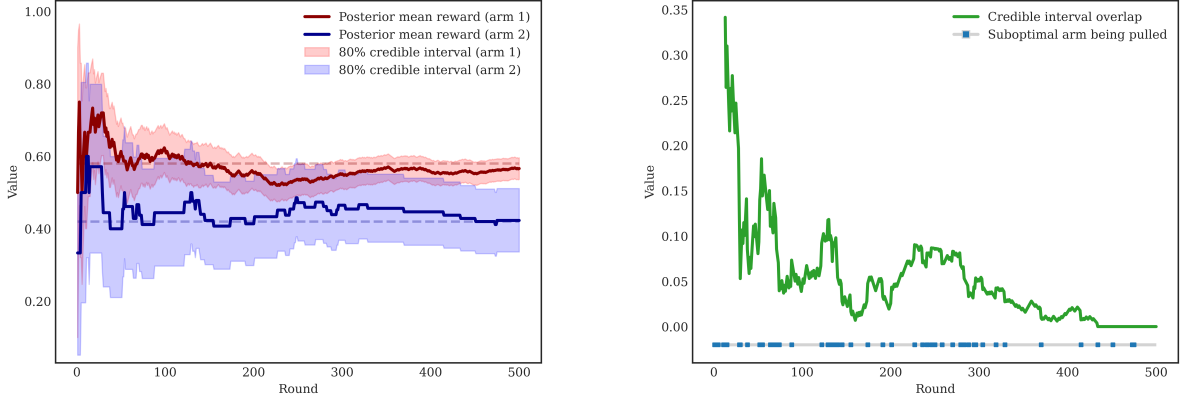
where  $N_{k,t}$  is the number of times arm  $k$  has been pulled up to time  $t$ , and  $\hat{\mu}_{k,t}$  is the corresponding empirical mean reward. As suggested by the central limit theorem, the uncertainty of  $\hat{\mu}_{k,t}$  can be represented by a confidence interval (left plot of Figure 2), the width of which is of order  $1/\sqrt{N_{k,t}}$ . The two confidence intervals are then scaled by  $\sqrt{2 \log t}$ , creating a catch-up game between the two upper confidence bounds (right plot of Figure 2), which guides the exploration of UCB.



**Figure 2** UCB plays a two-armed Bernoulli bandit. Left: confidence intervals around empirical means. Right: upper confidence bounds. The suboptimal arm (arm 2) is pulled whenever the corresponding upper confidence bound is higher.

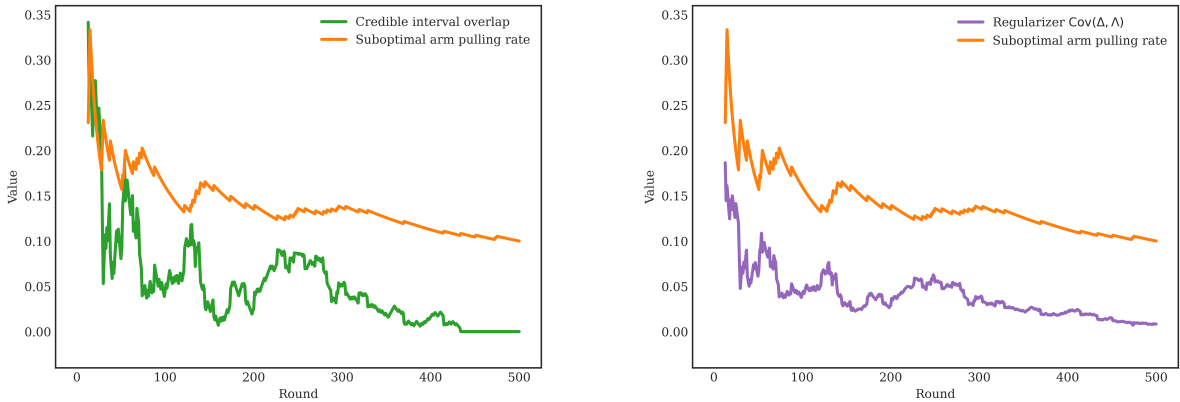
In contrast, the exploration of Thompson Sampling is guided by a single draw from the posterior distribution, with the sampling procedure implicitly accounting for uncertainty. The higher the probability that the current leader is not truly optimal, the more frequently the other arm is sampled. This probability is intuitively reflected by the overlap of credible intervals (left plot of Figure 3), but not entirely, since the credible intervals eventually detach while the exploration continues (right plot of Figure 3). This issue is resolved by the biserial covariance  $\text{Cov}_{\pi_t}(\Delta, \Lambda)$ , the regularizer of Thompson Sampling. (The 80% credible interval of a posterior distribution is  $(F^{-1}(0.1), F^{-1}(0.9))$  where  $F$  denotes the posterior CDF.)

As Thompson Sampling plays a two-armed Bernoulli bandit, we compare the rate at which the suboptimal arm is pulled (i.e.,  $N_{2,t}/t$ ), first with the overlap of credible intervals (left plot of Figure 4), and then with the regularizer (right plot of Figure 4). On the left, we see that the overlap, as an intuitive proxy for uncertainty, does guide the exploration to some extent, until the overlap vanishes. The larger the overlap, the more Thompson Sampling allocates pulls to the suboptimal



**Figure 3** Thompson Sampling plays a two-armed Bernoulli bandit. Left: credible intervals around posterior means. Right: the overlap of credible intervals. The overlap, when present, reflects the frequency of pulling the suboptimal arm (arm2).

arm in order to resolve the uncertainty there. On the right, we see that the regularizer also captures the behavior of the “exploration rate”. This connection is not a coincidence: the regularizer (a formal notion of uncertainty) maintains a strong temporal correlation (Pearson coefficient 0.995) with the overlap (an informal notion of uncertainty), until the overlap eventually vanishes.



**Figure 4** Thompson Sampling plays a two-armed Bernoulli bandit. Left: the overlap of credible intervals vs. the pulling rate of the suboptimal arm. Right: the regularizer of Thompson Sampling vs. the pulling rate of the suboptimal arm.

In summary, UCB measures uncertainty through a pair of confidence intervals that guides exploration via the catch-up game with logarithmic scaling, whereas Thompson Sampling measures uncertainty through the covariance-based regularizer that guides exploration via the online optimization form in Theorem 1. This optimization perspective enables us to understand Thompson Sampling in a more principled and less heuristic manner, much like how we understand UCB.

## 6. A Closer Look at the $\mathcal{R}^2$ -optimal Policy

After placing Thompson Sampling within the family of  $\mathcal{R}^2$ -finite policies, a natural question is how far Thompson Sampling is from  $\mathcal{R}^2$ -optimality. To address this, we study the  $\mathcal{R}^2$ -optimal policy to benchmark Thompson Sampling, presenting a closed-form solution in the one-armed case and an approximate implementation in the two-armed case. The  $\mathcal{R}^2$ -optimal policy achieves substantially lower cumulative regret than Thompson Sampling, confirming that  $\mathcal{R}^2$  is indeed a faithful surrogate for the  $\mathcal{R}_T$ -sequence. More importantly, comparing their regularizers allows us to clearly identify and address the issues of Thompson Sampling, underscoring the appeal of a principled framework with a well-defined benchmark.

### 6.1. One Arm

In the one-armed case, where only one of the two arms is unknown (e.g.,  $\pi_0 = N(0, 1) \times N(0, 0)$ ), the  $\mathcal{R}^2$ -optimal policy is fully tractable. Without loss of generality, we take arm 2 to be the known arm with  $\theta_2 \equiv 0$ . The stationary Bellman equation (6) becomes

$$\begin{aligned} 0 &= \min_{q_t} \left[ (\mathbb{E}_{\pi_t}[(\theta_1)_+]) - q_{1,t} \mathbb{E}_{\pi_t} \theta_1)^2 \right. \\ &\quad + q_{1,t} (\mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t = 1] - V(\pi_t)) \\ &\quad \left. + q_{2,t} (\mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t = 2] - V(\pi_t)) \right] \\ &= \min_{q_{1,t}} \left[ (\mathbb{E}_{\pi_t}[(\theta_1)_+]) - q_{1,t} \mathbb{E}_{\pi_t} \theta_1)^2 \right. \\ &\quad \left. - q_{1,t} (V(\pi_t) - \mathbb{E}_{\pi_t}^1 V(\pi_{t+1})) \right], \end{aligned}$$

where  $(\theta_1)_+ = \max(\theta_1, 0)$ ,  $\mathbb{E}_{\pi_t}^1 V(\pi_{t+1}) = \mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t = 1]$ , and  $\mathbb{E}_{\pi_t}[V(\pi_{t+1})|A_t = 2] = V(\pi_t)$  as pulling the known arm (arm 2) brings no new information ( $\pi_{t+1} = \pi_t$ ). In contrast, pulling the unknown arm (arm 1) reduces uncertainty and hence future regret, yielding  $\mathbb{E}_{\pi_t}^1 V(\pi_{t+1}) < V(\pi_t)$ . As a result, the above minimization is equivalent to

$$q_{1,t}^* = \operatorname{argmin}_{q_{1,t}} \left[ \frac{(\mathbb{E}_{\pi_t}[(\theta_1)_+]) - q_{1,t} \mathbb{E}_{\pi_t} \theta_1)^2}{q_{1,t}} \right] \quad (10)$$

with minimum  $V(\pi_t) - \mathbb{E}_{\pi_t}^1 V(\pi_{t+1}) > 0$ . Since the objective no longer contains  $V$ , the  $\mathcal{R}^2$ -optimal policy is fully tractable.

**PROPOSITION 4 (Closed-form solution).** *When  $\theta_2 \equiv 0$  and  $\mathbb{E}_{\pi_t} \theta_1 \neq 0$ , the  $\mathcal{R}^2$ -optimal policy pulls arm 1 with probability*

$$q_{1,t}^* = \min \left( \frac{\mathbb{E}_{\pi_t}[(\theta_1)_+]}{|\mathbb{E}_{\pi_t} \theta_1|}, 1 \right),$$

*and its regularizer (9) becomes*

$$\nu(\pi_t) = 4\mathbb{E}_{\pi_t}[(\theta_1)_+] - \frac{(\mathbb{E}_{\pi_t}[(\theta_1)_+] + \mathbb{E}_{\pi_t} \theta_1)_+^2}{\mathbb{E}_{\pi_t} \theta_1}.$$

Note that  $q_{1,t}^*$  (and likewise  $\nu(\pi_t)$ ) exhibits a “phase change” at  $\mathbb{E}_{\pi_t}[(\theta_1)_+] + \mathbb{E}_{\pi_t}\theta_1 = 0$ . When  $\mathbb{E}_{\pi_t}\theta_1 \geq 0$ , we must have  $q_{1,t}^* = 1$  as  $\mathbb{E}_{\pi_t}[(\theta_1)_+] \geq \mathbb{E}_{\pi_t}\theta_1$ . This is clearly the right thing to do: we should keep pulling the unknown arm as long as its posterior mean is non-negative (i.e., no worse than the known arm). When  $\pi_t$  is concentrated far to the left,  $\mathbb{E}_{\pi_t}[(\theta_1)_+]$  is small while  $|\mathbb{E}_{\pi_t}\theta_1|$  is large, so their ratio  $q_{1,t}^*$  is correspondingly small. Between these two extremes, the phase change occurs when  $\mathbb{E}_{\pi_t}\theta_1 < 0$  but the associated loss exactly offsets the exploratory benefit of pulling the unknown arm with probability 1. In the Gaussian case, the phase change can be characterized explicitly. Let  $\Phi$  and  $\phi$  be the CDF and PDF of  $N(0,1)$ , respectively.

**PROPOSITION 5 (Phase change).** *When  $\theta_2 \equiv 0$  and  $\theta_1 \sim N(\mu_t, \sigma_t^2)$  under  $\pi_t$ ,*

$$q_{1,t}^* = 1 \Leftrightarrow \mu_t/\sigma_t \geq \bar{x},$$

where  $\bar{x} \approx -0.276$  is the unique root of the increasing function  $x\Phi(x) + \phi(x) + x$ .

We may interpret 0.276 as the (relative) “fair price” to pay for the exploratory benefit of pulling the unknown arm. Whenever the “current price” falls below this threshold, the  $\mathcal{R}^2$ -optimal policy pulls the unknown arm with probability 1 to maximize “arbitrage”.

In Figure 1 (at the end of the introduction), Thompson Sampling and the  $\mathcal{R}^2$ -optimal policy play a Gaussian bandit with reward variance 1. Starting from  $\pi_0 = N(0,1) \times N(0,0)$ , where  $\theta_2 \equiv 0$ , the  $\mathcal{R}^2$ -optimal policy achieves substantially lower cumulative regret than Thompson Sampling (left plot of Figure 1). The victory of the  $\mathcal{R}^2$ -optimal policy illustrates that optimizing  $\mathcal{R}^2(Q; \pi_0)$  does correspond to lowering  $\{\mathcal{R}_T(Q; \pi_0) : T \geq 1\}$ . Why does Thompson Sampling lose? One reason is that the covariance-based regularizer is too “conservative” in certain scenarios. When  $\pi_0 = N(\mu, 1) \times N(0,0)$  and  $\mu$  approaches 0 from below, the regularizer of the  $\mathcal{R}^2$ -optimal policy  $\nu(\pi_0)$  diverges to infinity, whereas the regularizer of Thompson Sampling  $\tilde{\nu}(\pi_0)$  converges to  $\sqrt{2/\pi}$  (right plot of Figure 1). When  $\mu \approx 0$ , the tension between exploration and exploitation vanishes. Therefore, we should explore the more uncertain arm 1 with probability 1, but the regularizer of Thompson Sampling does not grow fast enough to encourage such pure exploration. In fact, Thompson Sampling never prioritizes the more uncertain arm when the two arms have the same posterior mean.

**REMARK 4.** If we look closely at the right plot of Figure 1, we can spot the phase change of the regularizer of the  $\mathcal{R}^2$ -optimal policy at  $-0.276$  (Proposition 5), where the curve becomes slightly less smooth than elsewhere.



## 6.2. Two Arms

As shown in the one-armed case, comparing regularizers provides clear insight into the behavior of Thompson Sampling relative to the  $\mathcal{R}^2$ -optimal policy. More can be learned in the two-armed case, but the  $\mathcal{R}^2$ -optimal policy is no longer tractable there. In what follows, we show how the  $\mathcal{R}^2$ -optimal policy can nevertheless be approximately implemented in the two-armed case.

We consider a two-armed Bernoulli bandit with prior  $\pi_0 = \text{Beta}(\alpha_1, \beta_1) \times \text{Beta}(\alpha_2, \beta_2)$ . The stationary Bellman equation (6) becomes

$$\begin{aligned} & V_{\alpha_1, \beta_1, \alpha_2, \beta_2} \\ &= \min_{p, q} \left[ (E_{\alpha_1, \beta_1, \alpha_2, \beta_2} - (pE_{\alpha_1, \beta_1} + qE_{\alpha_2, \beta_2}))^2 \right. \\ & \quad + p(E_{\alpha_1, \beta_1} V_{\alpha'_1, \beta_1, \alpha_2, \beta_2} + \bar{E}_{\alpha_1, \beta_1} V_{\alpha_1, \beta'_1, \alpha_2, \beta_2}) \\ & \quad \left. + q(E_{\alpha_2, \beta_2} V_{\alpha_1, \beta_1, \alpha'_2, \beta_2} + \bar{E}_{\alpha_2, \beta_2} V_{\alpha_1, \beta_1, \alpha_2, \beta'_2}) \right], \end{aligned}$$

where  $p + q = 1$ ,  $\alpha'_1 = \alpha_1 + 1$ ,  $E_{\alpha_1, \beta_1} = \alpha_1 / (\alpha_1 + \beta_1)$ ,  $\bar{E}_{\alpha_1, \beta_1} = 1 - E_{\alpha_1, \beta_1}$ , and

$$E_{\alpha_1, \beta_1, \alpha_2, \beta_2} = \mathbb{E}[\max(\text{Beta}(\alpha_1, \beta_1), \text{Beta}(\alpha_2, \beta_2))].$$

Let

$$\begin{aligned} & V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} \\ &= V_{\alpha_1, \beta_1, \alpha_2, \beta_2} - E_{\alpha_1, \beta_1} V_{\alpha'_1, \beta_1, \alpha_2, \beta_2} - \bar{E}_{\alpha_1, \beta_1} V_{\alpha_1, \beta'_1, \alpha_2, \beta_2} \end{aligned}$$

be the benefit of pulling arm 1. Then the benefit of pulling arm 2 is simply  $V'_{\alpha_2, \beta_2, \alpha_1, \beta_1}$  ( $V$  is symmetric). This benefit function satisfies a backward recursion as well as two boundary conditions.

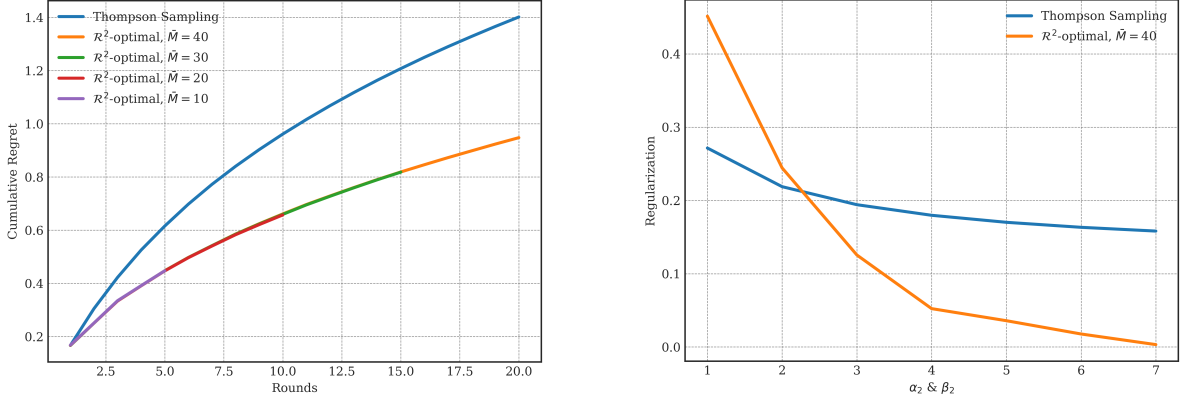
**PROPOSITION 6 (Benefit function).** *The function  $V'$  characterizes the  $\mathcal{R}^2$ -optimal policy. When  $\alpha_1 + \beta_1 < \infty$  and  $\alpha_2 + \beta_2 < \infty$ ,  $V'_{\alpha_1, \beta_1, \alpha_2, \beta_2}$  and  $V'_{\alpha_2, \beta_2, \alpha_1, \beta_1}$  can be computed from  $V'_{\alpha_1, \beta_1, \alpha'_2, \beta_2}$ ,  $V'_{\alpha_1, \beta_1, \alpha_2, \beta'_2}$ ,  $V'_{\alpha_2, \beta_2, \alpha'_1, \beta_1}$ ,  $V'_{\alpha_2, \beta_2, \alpha_1, \beta'_1}$ . When  $\alpha_1 + \beta_1 = \infty$ ,  $V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} = 0$ . When  $\alpha_2 + \beta_2 = \infty$ ,*

$$\begin{aligned} & V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} \\ &= \min_{p, q} \left[ \frac{(E_{\alpha_1, \beta_1, \alpha_2, \beta_2} - (pE_{\alpha_1, \beta_1} + qE_{\alpha_2, \beta_2}))^2}{p} \right]. \end{aligned}$$

A natural way to approximately implement the  $\mathcal{R}^2$ -optimal policy is to impose the two boundary conditions on  $\{(\alpha_1, \beta_1, \alpha_2, \beta_2) : \alpha_1 + \beta_1 = \bar{M} \text{ or } \alpha_2 + \beta_2 = \bar{M}\}$  where  $\bar{M}$  is finite (i.e., an arm is regarded as fully known after  $\bar{M}$  pulls), and then propagate the values of  $V'$  inward to  $\{(\alpha_1, \beta_1, \alpha_2, \beta_2) : \alpha_1 + \beta_1 < \bar{M} \text{ and } \alpha_2 + \beta_2 < \bar{M}\}$ .

In Figure 5, Thompson Sampling and the  $\mathcal{R}^2$ -optimal policy (with different values of  $\bar{M}$ ) play a Bernoulli bandit. For each value of  $\bar{M}$ , let the corresponding policy  $Q^{\bar{M}}$  play  $\bar{M}/2$  rounds (left plot of Figure 5). The resulting regret curves are nearly indistinguishable, indicating that these values of  $\bar{M}$  are already enough to reveal what the  $\mathcal{R}^2$ -optimal policy does in the first 20 rounds.

After 20 rounds, the  $\mathcal{R}^2$ -optimal policy achieves a reduction of more than 30% in cumulative regret compared to Thompson Sampling. Again, the reason behind this large gap can be understood through a comparison of their regularizers.



**Figure 5** Thompson Sampling and the  $\mathcal{R}^2$ -optimal policy (with different values of  $\bar{M}$ ) play a Bernoulli bandit. Left: comparing their cumulative regret  $\mathcal{R}_T(Q^{\text{TS}}; \pi_0)$  vs.  $\mathcal{R}_T(Q^{\bar{M}}; \pi_0)$  where  $\pi_0 = \text{Beta}(1, 1) \times \text{Beta}(1, 1)$  (200K trials). Right: comparing the two regularizers  $\tilde{\nu}(\text{Beta}(5, 4) \times \text{Beta}(k, k))$  vs.  $\nu^{\bar{M}}(\text{Beta}(5, 4) \times \text{Beta}(k, k))$  where  $\bar{M} = 40$  and  $k = 1, \dots, 7$ .

Let  $\nu^{\bar{M}}$  be the regularizer of  $Q^{\bar{M}}$ , which approximates the regularizer of the  $\mathcal{R}^2$ -optimal policy. When  $\pi_0 = \text{Beta}(5, 4) \times \text{Beta}(k, k)$  and  $k$  increases from 1 to 7, the approximate regularizer of the  $\mathcal{R}^2$ -optimal policy  $\nu^{\bar{M}}(\pi_0)$  drops sharply from above 0.4 to nearly 0, whereas the regularizer of Thompson Sampling  $\tilde{\nu}(\pi_0)$  drops gradually from above 0.2 to below 0.2 (right plot of Figure 5). After 9 pulls of arm 1, its posterior mean of 5/9 is slightly better than that of a fair coin. As  $k$  increases, we become increasingly certain that arm 2 with posterior mean  $k/(2k)$  behaves like a fair coin, which is worse than arm 1. After 14 pulls of arm 2 ( $k = 7$ ), the seemingly better arm 1 becomes relatively underexplored. Therefore, we should explore (and exploit) arm 1 with probability 1, but the regularizer of Thompson Sampling does not drop fast enough to abandon arm 2 (i.e., to set  $q_{2,t} = 0$  temporarily). In fact, Thompson Sampling never abandons any arm unless the optimal one is known with certainty, but it is entirely reasonable to abandon one arm when the other is good for both exploration and exploitation.

**A simple fix.** Thanks to the shared online optimization form, we can address the shortcoming of Thompson Sampling in a principled way: by adjusting its regularizer ( $\tilde{\nu}$ ) to better align with that of the  $\mathcal{R}^2$ -optimal policy ( $\nu$ ). As discussed above,  $\tilde{\nu}$  does not drop as fast as  $\nu$  to abandon the runner-up arm when the leading arm becomes relatively underexplored. A simple fix for this issue is shutting down regularization when there is no tension between exploration and exploitation

$$\nu^{\text{fix}}(\pi_t) = (1 - s(\pi_t))\tilde{\nu}(\pi_t).$$

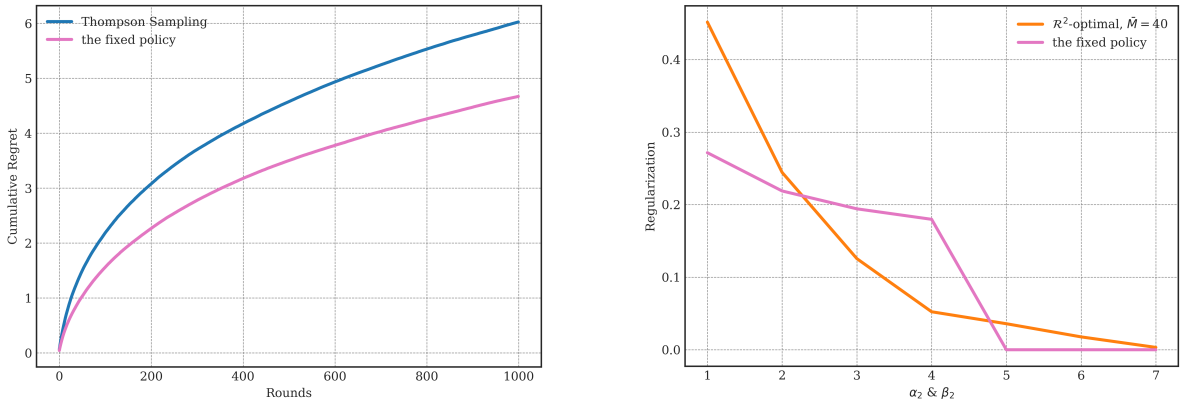
Here, the shutdown criterion is

$$s(\pi_t) = I(\mathbb{E}_{\pi_t} \theta_1 > \mathbb{E}_{\pi_t} \theta_2) I(\mathcal{V}_1(\pi_t) > \mathcal{V}_2(\pi_t)) \\ + I(\mathbb{E}_{\pi_t} \theta_2 > \mathbb{E}_{\pi_t} \theta_1) I(\mathcal{V}_2(\pi_t) > \mathcal{V}_1(\pi_t)),$$

where  $\mathcal{V}_k(\pi_t) = \text{Var}_{\pi_t} \mathbb{E}_{\pi_t}(\theta_k | \Lambda)$  is the variance-based “information gain” from pulling arm  $k$  (Russo and Van Roy 2014a). When  $s(\pi_t) = 1$ , the arm giving more reward also gives more information (hence no tension between exploration and exploitation), rendering regularization unnecessary.

**PROPOSITION 7 ( $\mathcal{R}^2$ -finiteness of the fixed policy).** *If the reward distribution is sub-Gaussian, as in Proposition 1, then the fixed policy is  $\mathcal{R}^2$ -finite.*

On the right of Figure 6, when  $\pi_0 = \text{Beta}(5, 4) \times \text{Beta}(k, k)$  and  $k$  increases from 1 to 7, the fixed regularizer  $\nu^{\text{fix}}(\pi_0)$  vanishes once arm 2 receives more pulls than arm 1 ( $k \geq 5$ ), better aligned with the  $\mathcal{R}^2$ -optimal regularizer  $\nu(\pi_0)$ . On the left of Figure 6, starting from  $\pi_0 = \text{Beta}(5, 4) \times \text{Beta}(500, 500)$ , where arm 1 is good for both exploration and exploitation, we observe that abandoning arm 2 significantly reduces regret.



**Figure 6** Thompson Sampling and the fixed policy play a Bernoulli bandit. Left: comparing their cumulative regret  $\mathcal{R}_T(Q^{\text{TS}}; \pi_0)$  vs.  $\mathcal{R}_T(Q^{\text{fix}}; \pi_0)$  where  $\pi_0 = \text{Beta}(5, 4) \times \text{Beta}(500, 500)$  (2K trials). Right: comparing the fixed regularizer with the optimal one  $\nu(\text{Beta}(5, 4) \times \text{Beta}(k, k))$  vs.  $\nu^{\text{fix}}(\text{Beta}(5, 4) \times \text{Beta}(k, k))$  where  $k = 1, \dots, 7$ .

## 7. Proofs

*Proof of Theorem 1* Recall that  $\Delta = \theta_1 - \theta_2$ ,  $\Lambda = \text{sign}(\theta_1 - \theta_2)$ , and

$$\begin{aligned}
\frac{\text{Cov}_{\pi_t}(\Delta, \Lambda)}{2} &= \mathbb{E}_{\pi_t} \Delta I(\Delta > 0) - P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta \\
&= P_{\pi_t}(\Delta \leq 0) \mathbb{E}_{\pi_t} \Delta I(\Delta > 0) \\
&\quad + P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta I(\Delta > 0) \\
&\quad - P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta I(\Delta > 0) \\
&\quad - P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta I(\Delta \leq 0) \\
&= P_{\pi_t}(\Delta \leq 0) \mathbb{E}_{\pi_t} \Delta I(\Delta > 0) \\
&\quad - P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta I(\Delta \leq 0).
\end{aligned}$$

By differentiation, the minimizer of the quadratic function is

$$\begin{aligned}
&\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - \frac{\text{Cov}_{\pi_t}(\Delta, \Lambda)}{2} \\
&= P_{\pi_t}(\Delta \leq 0) (\mathbb{E}_{\pi_t} \theta_2 + \mathbb{E}_{\pi_t} \Delta I(\Delta > 0)) \\
&\quad + P_{\pi_t}(\Delta > 0) (\mathbb{E}_{\pi_t} \theta_1 - \mathbb{E}_{\pi_t} \Delta I(\Delta \leq 0)) \\
&\quad - P_{\pi_t}(\Delta \leq 0) \mathbb{E}_{\pi_t} \Delta I(\Delta > 0) \\
&\quad + P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta I(\Delta \leq 0) \\
&= P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \theta_1 + P_{\pi_t}(\Delta \leq 0) \mathbb{E}_{\pi_t} \theta_2,
\end{aligned}$$

which is the expected next-round reward of Thompson Sampling  $x_t^{\text{TS}}$ .

*Proof of Proposition 2* When  $\text{Var}_{\pi_t} \Lambda > 0$ , we have

$$\begin{aligned}
\frac{\text{Cov}_{\pi_t}(\Delta, \Lambda)}{\text{Var}_{\pi_t} \Lambda} &= \frac{2P_{\pi_t}(\Delta \leq 0) \mathbb{E}_{\pi_t} \Delta I(\Delta > 0)}{4P_{\pi_t}(\Delta > 0)P_{\pi_t}(\Delta \leq 0)} \\
&\quad - \frac{2P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \Delta I(\Delta \leq 0)}{4P_{\pi_t}(\Delta > 0)P_{\pi_t}(\Delta \leq 0)} \\
&= \frac{\mathbb{E}_{\pi_t}[\Delta | \Delta > 0] - \mathbb{E}_{\pi_t}[\Delta | \Delta \leq 0]}{2}.
\end{aligned}$$

*Proof of Proposition 3* For  $\lambda \neq 1$ , the minimizer of the corresponding quadratic function is

$$\begin{aligned}
&\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - \frac{\lambda \text{Cov}_{\pi_t}(\Delta, \Lambda)}{2} \\
&= \lambda \left( \mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - \frac{\text{Cov}_{\pi_t}(\Delta, \Lambda)}{2} \right) \\
&\quad + (1 - \lambda) \mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) \\
&= \lambda (P_{\pi_t}(\Delta > 0) \mathbb{E}_{\pi_t} \theta_1 + P_{\pi_t}(\Delta \leq 0) \mathbb{E}_{\pi_t} \theta_2) \\
&\quad + (1 - \lambda) \mathbb{E}_{\pi_t} \max(\theta_1, \theta_2).
\end{aligned}$$

Let  $\bar{x}_t^\lambda$  be this minimizer. By clipping  $\bar{x}_t^\lambda$  to be between  $\mathbb{E}_{\pi_t}\theta_1$  and  $\mathbb{E}_{\pi_t}\theta_2$ , we obtain  $x_t^\lambda$ . When  $\pi_t = N(1 - \lambda, 0) \times N(0, \sigma^2)$ , we have

$$\begin{aligned}\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) &= \mathbb{E} \max(1 - \lambda, N(0, \sigma^2)) \\ &= \sigma \mathbb{E} \max((1 - \lambda)/\sigma, N(0, 1)) \\ &\rightarrow \infty\end{aligned}$$

as  $\sigma \rightarrow \infty$ . When  $\sigma$  is large enough, we have

$$\begin{aligned}\lambda < 1 &\Rightarrow \bar{x}_t^\lambda > \mathbb{E}_{\pi_t}\theta_1 > \mathbb{E}_{\pi_t}\theta_2 \Rightarrow x_t^\lambda = \mathbb{E}_{\pi_t}\theta_1, \\ \lambda > 1 &\Rightarrow \bar{x}_t^\lambda < \mathbb{E}_{\pi_t}\theta_1 < \mathbb{E}_{\pi_t}\theta_2 \Rightarrow x_t^\lambda = \mathbb{E}_{\pi_t}\theta_1.\end{aligned}$$

In either case, arm 1 is pulled with probability 1. Since pulling the known arm produces no posterior update, this choice persists indefinitely. Consequently, the policy keeps pulling arm 1 while arm 2 may be better (incomplete learning).

*Proof of Proposition 4* The minimizer of

$$\begin{aligned}&\frac{(\mathbb{E}_{\pi_t}[(\theta_1)_+] - q_{1,t}\mathbb{E}_{\pi_t}\theta_1)^2}{q_{1,t}} \\ &= q_{1,t}(\mathbb{E}_{\pi_t}\theta_1)^2 + \frac{\mathbb{E}_{\pi_t}[(\theta_1)_+]^2}{q_{1,t}} - 2\mathbb{E}_{\pi_t}[(\theta_1)_+]\mathbb{E}_{\pi_t}\theta_1\end{aligned}$$

in  $[0, 1]$  is clearly

$$q_{1,t}^* = \min\left(\frac{\mathbb{E}_{\pi_t}[(\theta_1)_+]}{|\mathbb{E}_{\pi_t}\theta_1|}, 1\right).$$

For the regularizer (9), the denominator is  $\mathbb{E}_{\pi_t}\theta_1$  while the numerator is

$$\begin{aligned}&\mathbb{E}_{\pi_t}^1 V(\pi_{t+1}) - V(\pi_t) \\ &= -\min_{q_{1,t}} \left[ \frac{(\mathbb{E}_{\pi_t}[(\theta_1)_+] - q_{1,t}\mathbb{E}_{\pi_t}\theta_1)^2}{q_{1,t}} \right] \\ &= -(\mathbb{E}_{\pi_t}[(\theta_1)_+] - \mathbb{E}_{\pi_t}\theta_1)^2 I(q_{1,t}^* = 1) \\ &\quad + 4\mathbb{E}_{\pi_t}[(\theta_1)_+]\mathbb{E}_{\pi_t}\theta_1(1 - I(q_{1,t}^* = 1)) \\ &= -(\mathbb{E}_{\pi_t}[(\theta_1)_+] + \mathbb{E}_{\pi_t}\theta_1)^2 I(q_{1,t}^* = 1) \\ &\quad + 4\mathbb{E}_{\pi_t}[(\theta_1)_+]\mathbb{E}_{\pi_t}\theta_1 \\ &= 4\mathbb{E}_{\pi_t}[(\theta_1)_+]\mathbb{E}_{\pi_t}\theta_1 - (\mathbb{E}_{\pi_t}[(\theta_1)_+] + \mathbb{E}_{\pi_t}\theta_1)_+^2,\end{aligned}$$

where the last line is because

$$\begin{aligned}q_{1,t}^* = 1 &\Leftrightarrow \mathbb{E}_{\pi_t}[(\theta_1)_+] \geq |\mathbb{E}_{\pi_t}\theta_1| \\ &\Leftrightarrow \mathbb{E}_{\pi_t}[(\theta_1)_+] \geq -\mathbb{E}_{\pi_t}\theta_1 \\ &\Leftrightarrow \mathbb{E}_{\pi_t}[(\theta_1)_+] + \mathbb{E}_{\pi_t}\theta_1 \geq 0.\end{aligned}$$

*Proof of Proposition 5* When  $\theta_1 \sim N(\mu_t, \sigma_t^2)$  under  $\pi_t$ , we have

$$\begin{aligned}
q_{1,t}^* = 1 &\Leftrightarrow \mathbb{E}_{\pi_t}[(\theta_1)_+] + \mathbb{E}_{\pi_t}\theta_1 \geq 0 \\
&\Leftrightarrow \mu_t \Phi\left(\frac{\mu_t}{\sigma_t}\right) + \sigma_t \phi\left(\frac{\mu_t}{\sigma_t}\right) + \mu_t \geq 0 \\
&\Leftrightarrow \frac{\mu_t}{\sigma_t} \Phi\left(\frac{\mu_t}{\sigma_t}\right) + \phi\left(\frac{\mu_t}{\sigma_t}\right) + \frac{\mu_t}{\sigma_t} \geq 0 \\
&\Leftrightarrow \frac{\mu_t}{\sigma_t} \geq \bar{x},
\end{aligned}$$

where  $\bar{x} \approx -0.276$  is the unique root of the increasing function  $x\Phi(x) + \phi(x) + x$ .

*Proof of Proposition 6* The function  $V'$  characterizes the  $\mathcal{R}^2$ -optimal policy as the stationary Bellman equation becomes

$$\begin{aligned}
&V'_{\alpha_2, \beta_2, \alpha_1, \beta_1} \\
&= \min_{p, q} \left[ (E_{\alpha_1, \beta_1, \alpha_2, \beta_2} - (pE_{\alpha_1, \beta_1} + qE_{\alpha_2, \beta_2}))^2 \right. \\
&\quad \left. - p(V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} - V'_{\alpha_2, \beta_2, \alpha_1, \beta_1}) \right].
\end{aligned}$$

When  $\alpha_1 + \beta_1 < \infty$  and  $\alpha_2 + \beta_2 < \infty$ , the backward recursion is given by the above equation and

$$\begin{aligned}
&V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} - E_{\alpha_2, \beta_2} V'_{\alpha_1, \beta_1, \alpha'_2, \beta_2} - \bar{E}_{\alpha_2, \beta_2} V'_{\alpha_1, \beta_1, \alpha_2, \beta'_2} \\
&= V'_{\alpha_2, \beta_2, \alpha_1, \beta_1} - E_{\alpha_1, \beta_1} V'_{\alpha_2, \beta_2, \alpha'_1, \beta_1} - \bar{E}_{\alpha_1, \beta_1} V'_{\alpha_2, \beta_2, \alpha_1, \beta'_1}
\end{aligned}$$

as both sides equal to

$$\begin{aligned}
&V_{\alpha_1, \beta_1, \alpha_2, \beta_2} - E_{\alpha_1, \beta_1} V_{\alpha'_1, \beta_1, \alpha_2, \beta_2} - \bar{E}_{\alpha_1, \beta_1} V_{\alpha_1, \beta'_1, \alpha_2, \beta_2} \\
&- E_{\alpha_2, \beta_2} V_{\alpha_1, \beta_1, \alpha'_2, \beta_2} - \bar{E}_{\alpha_2, \beta_2} V_{\alpha_1, \beta_1, \alpha_2, \beta'_2} \\
&+ E_{\alpha_1, \beta_1} E_{\alpha_2, \beta_2} V_{\alpha'_1, \beta_1, \alpha'_2, \beta_2} + \bar{E}_{\alpha_1, \beta_1} \bar{E}_{\alpha_2, \beta_2} V_{\alpha_1, \beta'_1, \alpha_2, \beta'_2} \\
&+ \bar{E}_{\alpha_1, \beta_1} E_{\alpha_2, \beta_2} V_{\alpha_1, \beta'_1, \alpha'_2, \beta_2} + E_{\alpha_1, \beta_1} \bar{E}_{\alpha_2, \beta_2} V_{\alpha'_1, \beta_1, \alpha_2, \beta'_2},
\end{aligned}$$

which remains unchanged when subscripts 1 and 2 are swapped ( $V$  is symmetric). When  $\alpha_1 + \beta_1 = \infty$  (arm 1 is fully known), we have  $V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} = 0$ , as pulling arm 1 brings no further benefit. When  $\alpha_2 + \beta_2 = \infty$  (arm 2 is fully known), we have

$$\begin{aligned}
&V'_{\alpha_1, \beta_1, \alpha_2, \beta_2} \\
&= \min_{p, q} \left[ \frac{(E_{\alpha_1, \beta_1, \alpha_2, \beta_2} - (pE_{\alpha_1, \beta_1} + qE_{\alpha_2, \beta_2}))^2}{p} \right],
\end{aligned}$$

as the benefit is computable in the one-armed case.

*Proof of Proposition 7* It suffices to show that the information ratio of  $Q^{\text{fix}}$  is bounded by that of  $Q^{\text{TS}}$ . When  $\mathbb{E}_{\pi_t}\theta_1 > \mathbb{E}_{\pi_t}\theta_2$ ,  $\mathcal{V}_1(\pi_t) > \mathcal{V}_2(\pi_t)$ , and  $q_{1,t}^{\text{fix}} = 1$ , we clearly have

$$\begin{aligned}
&\frac{(\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - \mathbb{E}_{\pi_t} \theta_1)^2}{\mathcal{V}_1(\pi_t)} \\
&\leq \frac{(\mathbb{E}_{\pi_t} \max(\theta_1, \theta_2) - (q_{1,t}^{\text{TS}} \mathbb{E}_{\pi_t} \theta_1 + q_{2,t}^{\text{TS}} \mathbb{E}_{\pi_t} \theta_2))^2}{q_{1,t}^{\text{TS}} \mathcal{V}_1(\pi_t) + q_{2,t}^{\text{TS}} \mathcal{V}_2(\pi_t)}.
\end{aligned}$$

## References

- Agarwal D (2013) Computational advertising: the LinkedIn way. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 1585–1586.
- Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *Conference on Learning Theory*, 39–1 (JMLR Workshop and Conference Proceedings).
- Agrawal S, Goyal N (2013) Further optimal regret bounds for Thompson sampling. *Artificial Intelligence and Statistics*, 99–107 (PMLR).
- Auer P, Cesa-Bianchi N, Fischer P (2002a) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47:235–256.
- Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002b) The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1):48–77.
- Bellman R (1957) *Dynamic Programming* (Princeton University Press).
- Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems* 24.
- Ghavamzadeh M, Mannor S, Pineau J, Tamar A, et al. (2015) Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning* 8(5-6):359–483.
- Gittins JC (1979) Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 41(2):148–164.
- Hill DN, Nassif H, Liu Y, Iyer A, Vishwanathan S (2017) An efficient bandit algorithm for realtime multi-variate optimization. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1813–1821.
- Kawale J, Bui HH, Kveton B, Tran-Thanh L, Chawla S (2015) Efficient Thompson sampling for online matrix-factorization recommendation. *Advances in Neural Information Processing Systems* 28.
- Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4–22.
- Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press).
- Lev J (1949) The point biserial coefficient of correlation. *The Annals of Mathematical Statistics* 20(1):125–126.
- Pearson K (1909) On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* 7(1/2):96–105.
- Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

- Rothschild M (1974) A two-armed bandit theory of market pricing. *Journal of Economic Theory* 9(2):185–202.
- Russo D, Van Roy B (2014a) Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems* 27.
- Russo D, Van Roy B (2014b) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243.
- Russo D, Van Roy B (2016) An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research* 17(68):1–30.
- Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26(6):639–658.
- Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.