Optimality of Thompson Sampling for Gaussian Bandits Depends on Priors

Junya Honda **Akimichi Takemura**
The University of Tokyo
{honda, takemura}@stat.t.u-tokyo.ac.jp

Abstract

In stochastic bandit problems, a Bayesian policy called Thompson sampling (TS) has recently attracted much attention for its excellent empirical performance. However, the theoretical analysis of this policy is difficult and its asymptotic optimality is only proved for one-parameter models. In this paper we discuss the optimality of TS for the model of normal distributions with unknown means and variances as one of the most fundamental examples of multiparameter models. First we prove that the expected regret of TS with the uniform prior achieves the theoretical bound, which is the first result to show that the asymptotic bound is achievable for the normal distribution model. Next we prove that TS with Jeffreys prior and reference prior cannot achieve the theoretical bound. Therefore choice of priors is important for TS and non-informative priors are sometimes risky in cases of multiparameter models.

1 INTRODUCTION

In reinforcement learning a tradeoff between exploration and exploitation of knowledge is considered. The multiarmed bandit problem is one formulation of the reinforcement learning and is a model of a gambler playing a slot machine with multiple arms. A dilemma for the gambler is that he cannot know whether the expectation of an arm is high or not without pulling it many times but he suffers a loss if he pulls suboptimal (i.e., not optimal) arms many times.

This problem was formulated by Robbins (1952) and its theoretical bound was derived by Lai and Robbins (1985) for one-parameter models, which was extended to multiparameter models by Burnetas and Katehakis (1996). These theoretical bounds show that any suboptimal arm has to be pulled at least logarithmic number of rounds and its coefficient is determined by the distributions of suboptimal arms and the expectation of the optimal arm.

Along with the asymptotic bound for this problem, achievability of the bound has also been considered in many models. Lai and Robbins (1985) proved the asymptotic optimality of a policy based on the notion of *upper confidence bound* (UCB) for Laplace distributions (which do not belong to exponential families) and some exponential families including normal distributions with *known* variances. The achievability of the bound was later extended to a subclass of one-parameter exponential families (Cappé et al., 2013).

On the other hand in multiparameter or nonparametric settings, Burnetas and Katehakis (1996) and Honda and Takemura (2010) proved the achievability for finite-support distributions and bounded-support distributions, respectively. However, the above two models are both compact and achievability of the bound is not known for non-compact multiparameter models, which include normal distributions with unknown means and variances. Since this model, which we simply call Gaussian model, is one of the most basic settings of stochastic bandits, many researches have been conducted for this model (Burnetas & Katehakis, 1996; Auer et al., 2002; Kaufmann et al., 2012a). However, to the authors' knowledge, only UCB-normal policy (Auer et al., 2002) theoretically assures a (non-optimal) logarithmic regret for this model[1].

This paper discusses the asymptotic optimality of Thompson sampling (TS) (Thompson, 1933) for the Gaussian model. TS is a Bayesian policy which

Appearing in Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

[1] The theoretical analysis of UCB-normal contains conjectures verified only numerically and the logarithmic regret is not assured in the strict sense.

chooses an arm randomly according to the posterior probability with which the arm is the optimal. This policy was recently rediscovered and is researched extensively because of its excellent empirical performance for many models (Chapelle & Li, 2012). The theoretical analysis of TS was first given for Bernoulli model (Agrawal & Goyal, 2012; Kaufmann et al., 2012b) and was later extended to general one-parameter exponential families (Korda et al., 2013).

The asymptotic optimality of TS under uniform prior is proved for Bernoulli model in Kaufmann et al. (2012b), whereas it is proved for a more general model, one-parameter exponential family, under Jeffreys prior in Korda et al. (2013). Therefore, TSs with uniform prior and Jeffreys prior are asymptotically equivalent at least for the Bernoulli model. Nevertheless, we prove for the Gaussian model that TS with uniform prior achieves the asymptotic bound whereas TS with Jeffreys prior and reference prior cannot. Furthermore, TS with Jeffreys prior cannot even achieve a logarithmic regret and suffers a polynomial regret in expectation. This result implies that TS may be more sensitive to the choice of priors than expected and non-informative priors are sometimes risky (in other words, too optimistic) for multiparameter models.

This paper is organized as follows. In Sect. 2, we formulate the bandit problem for the Gaussian model and introduce Thompson sampling. We give the main result on the optimality of TS in Sect. 3. The remaining sections are devoted to the proof of the main result. In Sect. 4, we derive inequalities for probabilities which appear in the Gaussian model. We prove the optimality of TS with conservative priors in Sect. 5 and prove the non-optimality of TS with optimistic priors in Sect. 6.

2 Preliminaries

We consider the K-armed stochastic bandit problem in the Gaussian model. The gambler pulls an arm $i \in \{1, 2, \cdots, K\}$ at each round and receives a reward independently distributed by $\mathcal{N}(\mu_i, \sigma_i^2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The gambler does not know the parameter $(\mu_i, \sigma_i^2) \in \mathbb{R} \times (0, \infty)$. The maximum expectation is denoted by $\mu^* = \max_{i \in \{1,2,\cdots,K\}} \mu_i$. Let $J(t)$ be the arm pulled at the t-th round. We define the number of times that the arm i is pulled before the t-th round by $N_i(t) = \sum_{m=1}^{t-1} \mathbb{1}[J(m) = i]$, where $\mathbb{1}[\cdot]$ denotes the indicator function. Then the regret of the gambler at the T-th round is given for $\Delta_i = \mu^* - \mu_i$ by

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \Delta_{J(t)} = \sum_i \Delta_i N_i(T+1) .$$

Let $X_{i,n}$ be the n-th reward from the arm i. We define

$$\bar{x}_{i,n} = \frac{1}{n} \sum_{m=1}^{n} X_{i,m} ,$$
$$S_{i,n} = \sum_{m=1}^{n} (X_{i,m} - \bar{x}_{i,n})^2 = \sum_{m=1}^{n} X_{i,m}^2 - n\bar{x}_{i,n}^2 ,$$

that is, $\bar{x}_{i,n}$ and $S_{i,n}$ denote the sample mean and the sum of squares from n samples from the arm i, respectively. We denote the sample mean and the sum of squares before the t-th round by $\bar{x}_i(t) = \bar{x}_{i,N_i(t)}$ and $S_i(t) = S_{i,N_i(t)}$. It is well known that

$$\bar{x}_{i,n} \sim \mathcal{N}(\mu_i, \sigma_i^2/n) , \qquad S_{i,n}/\sigma_i^2 \sim \chi^2_{n-1} , \tag{1}$$

where the chi-squared distribution χ^2_{n-1} with degree of freedom $n-1$ has the density

$$\chi^2_{n-1}(s) = \frac{s^{\frac{n-3}{2}} \mathrm{e}^{-\frac{s}{2}}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} .$$

2.1 Asymptotic Bound

It is shown in Burnetas and Katehakis (1996) that under any policy satisfying a mild regularity condition the expected regret satisfies

$$\liminf_{T\to\infty} \frac{\mathrm{E}[\mathrm{Regret}(T)]}{\log T} \ge \sum_{i:\Delta_i>0} \frac{\Delta_i}{\inf_{(\mu,\sigma):\mu>\mu^*} D(\mathcal{N}(\mu_i, \sigma_i^2) \| \mathcal{N}(\mu, \sigma^2))} , \tag{2}$$

where $D(\cdot\|\cdot)$ is the KL divergence. Since the KL divergence between normal distributions is

$$D(\mathcal{N}(\mu_a, \sigma_b^2) \| \mathcal{N}(\mu_a, \sigma_b^2)) = \frac{1}{2}\left(\log \frac{\sigma_b^2}{\sigma_a^2} + \frac{\sigma_a^2 + (\mu_b - \mu_a)^2}{\sigma_b^2} - 1\right) ,$$

the infimum in (2) is expressed for $\mu_i < \mu^*$ as $D_{\inf}(\Delta_i, \sigma_i^2)$ where

$$D_{\inf}(\Delta, \sigma^2) = \frac{1}{2} \log\left(1 + \frac{\Delta^2}{\sigma^2}\right) .$$

Thus we can rewrite the theoretical bound in (2) as

$$\liminf_{T\to\infty} \frac{\mathrm{E}[\mathrm{Regret}(T)]}{\log T} \ge \sum_{i:\Delta_i>0} \frac{\Delta_i}{D_{\inf}(\Delta_i, \sigma_i^2)} . \tag{3}$$

2.2 Bayesian Theory and Thompson Sampling

Thompson sampling is a policy based on the Bayesian viewpoint. We mainly consider the prior $\pi(\mu_i, \sigma_i^2) \propto$

Algorithm 1 Thompson Sampling

```
Parameter: $\alpha \in \mathbb{R}$, $\bar{n} \geq \max\{2, 3 - \lfloor 2\alpha \rfloor\}$.
Initialization: Pull each arm $\bar{n}$ times.
Loop:
    1. Sample $\tilde{\mu}_i(t)$ from the posterior $\pi_i(\mu_i | \hat{\theta}_i(t))$ under prior $\pi(\mu_i, \sigma_i^2) \propto (\sigma_i^2)^{-1-\alpha}$ for each arm $i$.
    2. Pull an arm $i$ maximizing $\tilde{\mu}_i(t)$.
```

$(\sigma_i^2)^{-1-\alpha}$, or equivalently, $\pi(\mu_i, \sigma_i) \propto \sigma_i^{-1-2\alpha}$. Since the density of the inverse gamma distribution is $(\beta^\alpha / \Gamma(\alpha)) \mathrm{e}^{-\beta/x} x^{-1-\alpha}$, the above prior for σ_i^2 corresponds to this distribution with parameters $(\alpha, \beta) = (\alpha, 0)$. Here the inverse gamma distribution is a conjugate prior for variances of normal distributions and is often used for this model (see, e.g. Robert (2001) for results on Bayesian theory given in this section). The cases $\alpha = -1, -1/2, 0, 1/2$ correspond to uniform for parameter σ_i^2, uniform for parameter σ_i, reference and Jeffreys priors, respectively. Under this prior, the posterior distribution on μ_i is given by

$$\sqrt{\frac{n(n + 2\alpha - 1)}{S_{i,n}}} (\mu_i - \bar{x}_{i,n}) \bigg| \hat{\theta}_{i,n} \sim f_{n+2\alpha-1} , \qquad (4)$$

where we write $\hat{\theta}_{i,n} = (\bar{x}_{i,n}, S_{i,n})$ and f_ν denotes t-distribution with degree of freedom ν, which has density

$$f_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} . \qquad (5)$$

Thompson sampling is the policy which chooses an arm randomly according to the probability with which the arm is the optimal when each μ_i is distributed independently by the posterior $\pi(\mu_i | \hat{\theta}_i(t))$ for $\hat{\theta}_i(t) = (\bar{x}_i(t), S_i(t))$. This policy is formulated as Algorithm 1. Note that we require $\max\{2, 2 - \lceil 2\alpha \rceil\}$ initial pulls to avoid improper posteriors. We use $\bar{n} \geq \max\{2, 3 - \lfloor 2\alpha \rfloor\}$ for simplicity of the analysis.

3 Regret of Thompson Sampling

In this section we give the main result of this paper. First we show that TS achieves the asymptotic bound for the prior $\pi(\mu_i, \sigma_i^2) \propto (\sigma_i^2)^{-1-\alpha}$ with $\alpha < 0$.

Theorem 1. *Let $\epsilon > 0$ be arbitrary and assume that there is a unique optimal arm. Under Thompson sampling with $\alpha < 0$, the expected regret is bounded as*

$$\mathrm{E}[\mathrm{Regret}(T)] \leq \sum_{i:\Delta_i > 0} \frac{\Delta_i \log T}{D_{\inf}(\Delta_i, \sigma_i^2)} + \mathrm{O}((\log T)^{4/5}) .$$

See Lemma 5 for the specific representation of the reminder term $\mathrm{O}((\log T)^{4/5})$. We see from this theorem that TS with $\alpha < 0$ is asymptotically optimal in view of (3).

Next we show that TS with $\alpha \geq 0$ cannot achieve the asymptotic bound. To simplify the analysis we consider a two-armed setting more advantageous to the gambler in which the full information on the arm 2 is known beforehand, that is, the prior on the arm 2 is the unit point mass measure $\delta_{\{(\mu_2, \sigma_2^2)\}}$ instead of $\pi(\mu_2, \sigma_2^2) \propto (\sigma_2^2)^{-1-\alpha}$.

Theorem 2. *Assume that there are $K = 2$ arms and $\mu_1 > \mu_2$. Then, under Thompson sampling such that $\tilde{\mu}_1(t) \sim \pi(\mu_1 | \hat{\theta}_1(t))$ with $\alpha \geq 0$ and $\tilde{\mu}_2(t) = \mu_2$, there exists a constant $\xi > 0$ independent of σ_2 such that*

$$\liminf_{T \to \infty} \frac{\mathrm{E}[\mathrm{Regret}(T)]}{\log T} \geq \xi . \qquad (6)$$

In particular, if $\alpha > 0$ then there exist $\xi' > 0$ such that

$$\liminf_{N \to \infty} \frac{\mathrm{E}[\mathrm{Regret}(T)]}{T^{\frac{2\alpha}{\bar{n}-1+2\alpha}}} \geq \xi' . \qquad (7)$$

Eq. (7) means that TS with $\alpha > 0$ suffers a polynomial regret in expectation. Also note that the asymptotic bound in (3) approaches zero for sufficiently small σ_2 in the above two-armed setting since $D_{\inf}(\Delta, \sigma^2) \to \infty$ as $\sigma \to 0$. Nevertheless, the LHS of (6) does not go to zero as $\sigma_2 \to 0$ because $\xi > 0$ is independent of σ_2. Therefore TS with $\alpha = 0$ also does not achieve the asymptotic bound at least for sufficiently small σ_2.

Recall that Jeffreys and reference priors correspond to $\alpha = 1/2$ and $\alpha = 0$, respectively. Therefore this theorem means that TS with these non-informative priors does not achieve the asymptotic bound.

Remark 1. For any $\mu > \bar{x}_i$, posterior probability of event $\mu_i > \mu$ becomes large when the prior has heavy weight at large σ_i^2, that is, when α is small. Therefore, as α decreases, TS becomes a "conservative" policy which chooses a seemingly suboptimal arm frequently. Theorems 1 and 2 mean that the prior should be conservative to some extent and non-informative priors are too optimistic.

Remark 2. Although TS with non-informative priors does not achieve the asymptotic bound in the sense of expectation, this fact does not necessarily mean that these priors are "bad" ones. As we can see from a close inspection of the proof of Theorem 2, the expected regret of TS with these priors becomes large because an enormously large regret arises with fairly small probability. Therefore this policy performs well except for the case arising with this small probability, and the authors think that TS with these priors also

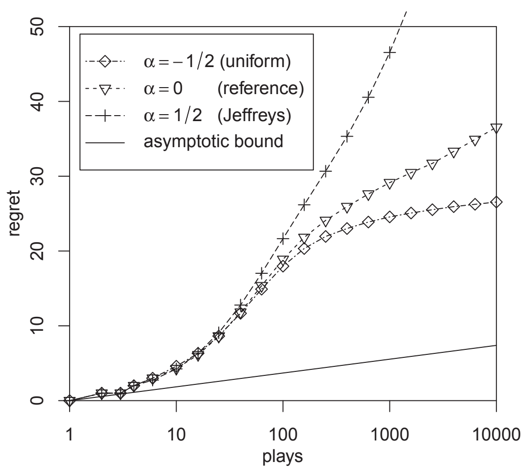

Figure 1: Regret of Thompson sampling with priors $\alpha = -1/2,\, 0,\, 1/2$.

becomes a good policy in the probably approximately correct (PAC) framework. In any case, we should be aware that these non-informative priors are "risky" in the sense of expectation.

3.1 Relation with Problem-independent Viewpoint

Regret bounds derived in this paper are given from a frequentist's viewpoint, that is, parameters (μ_i, σ_i) (and the number of arms K) are regarded as fixed constants.

From a Bayesian viewpoint, it is proved in Russo and Roy (2013) that TS achieves a Bayes risk $\mathrm{O}(\sqrt{KT\log T})$ with an arbitrary prior in the case of bounded rewords $X_{i,n} \in [0,1]$. As a similar result for bounded rewards, TS with a certain prior achieves a regret $\mathrm{O}(\sqrt{KT\log K})$ independent of distributions (Agrawal & Goyal, 2013). Theorem 2 implies that such bounds cannot be achieved by TS with prior $\alpha > (\bar{n}-1)/2$ in the case of Gaussian rewards since the regret is a polynomial larger than $\mathrm{O}(\sqrt{T})$.

3.2 Empirical Evaluation

Fig. 1 shows the simulation result of TS with $\pi(\mu_i, \sigma_i^2) \propto (\sigma_i^2)^{-1-\alpha}$ and $\bar{n} = 3 - \lfloor 2\alpha \rfloor$ for a two-armed setting such that $(\mu_1, \mu_2) = (1, 0)$ and $(\sigma_1, \sigma_2) = (3.0, 0.3)$. Each plot is an average over 20,000 trials. We see from the figure that TS with Jeffreys prior actually suffers a polynomial regret and TS with reference prior has a larger asymptotic slope than that of the uniform prior. We also report that in this simulation Jeffreys prior performed the best in most trials although the average regret was the worst because of large regrets in few trials.

4 Inequalities for Gaussian Model

In this section we derive fundamental inequalities for distributions appearing in Thompson sampling for the Gaussian model. We prove them in the supplementary material.

First we give large deviation probabilities (see, e.g., Dembo and Zeitouni (1998)) for empirical means and variances.

Lemma 3. *For any $\mu > \mu_i$*

$$\Pr[\bar{x}_{i,n} \geq \mu] \leq \mathrm{e}^{-\frac{n(\mu-\mu_i)^2}{2\sigma_i^2}} \tag{8}$$

and for any $\sigma^2 > \sigma_i^2$

$$\Pr[S_{i,n} \geq n\sigma^2] \leq \mathrm{e}^{-nh\left(\frac{\sigma^2}{\sigma_i^2}\right)} \tag{9}$$

where $h(x) = (x - 1 - \log x)/2 \geq 0$.

Remark 3. We can derive a similar bound to (9) by simply applying Chebycheff's inequality to a chi-squared random variable $S_{i,n}/\sigma_i^2$ which has moment generating function $\mathrm{E}[\mathrm{e}^{\lambda S_{i,n}/\sigma_i^2}] = (1-2\lambda)^{-(n-1)/2}$. Furthermore, it is well known that Mills' ratio (Kendall & Stuart, 1977, Chap. 5) gives a tighter bound for the tail probability of normal distributions, and a similar technique can also be applied to the tail weight of χ^2-distributions. However, we use bounds in Lemma 3 based on the large deviation principle for simplicity and convenience of analysis.

Next we evaluate the posterior distribution of the mean for Thompson sampling. Probability that the sample from the posterior is larger than or equal to μ, which we write

$$p_n(\mu|\hat{\theta}_{i,n}) = P^{\pi}(\mu_i \geq \mu|\hat{\theta}_{i,n})\,,$$

is bounded as follows.

Lemma 4. *If $\mu > \bar{x}_{i,n}$ and $n \geq \bar{n}$ then*

$$p_n(\mu|\hat{\theta}_{i,n}) \leq \frac{1}{\sqrt{n}}\frac{\sqrt{S_{i,n}}}{\mu - \bar{x}_{i,n}}\left(1 + \frac{n(\mu-\bar{x}_{i,n})^2}{S_{i,n}}\right)^{-\frac{n-2}{2}-\alpha} \tag{10}$$

and

$$p_n(\mu|\hat{\theta}_{i,n}) \geq A_{n,\alpha}\left(1 + \frac{n(\mu-\bar{x}_{i,n})^2}{S_{i,n}}\right)^{-\frac{n-1}{2}-\alpha}, \tag{11}$$

where

$$A_{n,\alpha} = \frac{1}{2\mathrm{e}^{1/6}\sqrt{\pi(\frac{n}{2}+\alpha)}}\,. \tag{12}$$

5 Analysis for Conservative Priors

In this section we show that Thompson sampling achieves the asymptotic bound if $\alpha < 0$. The main result of this section is given as follows.

Lemma 5. *Fix any $\alpha < 0$ and assume that $(\mu_1, \sigma_1^2) = (0,1)$ and the arm 1 is the unique optimal arm. Then, for any $\epsilon < \min_{i:\Delta_i>0} \Delta_i/2$,*

$$\begin{aligned}
&\mathrm{E}[\mathrm{Regret}(T)] \\
&\le \sum_{i:\Delta_i>0} \Delta_i \Biggl(\frac{\log T}{D_{\mathrm{inf}}(\Delta_i - 2\epsilon, \sigma_i^2+\epsilon)} + 2 - 2\alpha \\
&\qquad + \frac{\sqrt{\sigma_i^2+\epsilon}}{\Delta_i - 2\epsilon} + \frac{1}{1-\mathrm{e}^{-\frac{\epsilon}{2\sigma_i^2}}} + \frac{1}{1-\mathrm{e}^{-h(1+\frac{\epsilon}{\sigma_i^2})}} \Biggr) \\
&\quad + \Delta_{\max} \Biggl(\frac{1}{1-\mathrm{e}^{-\frac{\epsilon^2}{2}}} + \frac{1}{1-\mathrm{e}^{-h(2)}} + \frac{\mathrm{B}(1/2,-\alpha)}{(1-\mathrm{e}^{-\frac{\epsilon^2}{2}})^2} \\
&\qquad + \frac{2\sqrt{2}}{\epsilon} \frac{\left(1+\epsilon^2/8\right)^{1-\alpha}}{1-\left(1+\epsilon^2/8\right)^{-1/2}} \Biggr) \\
&= \sum_{i:\Delta_i>0} \frac{\Delta_i \log T}{D_{\mathrm{inf}}(\Delta_i - 2\epsilon, \sigma_i^2+\epsilon)} + \mathrm{O}(\epsilon^{-4}),
\end{aligned}$$

where $\Delta_{\max} = \max_i \Delta_i$ and $\mathrm{B}(\cdot,\cdot)$ is the beta function.

Corollary 6. *Under the same assumption as Lemma 5,*

$$\mathrm{E}[\mathrm{Regret}(T)] \le \sum_{i:\Delta_i>0} \frac{\Delta_i \log T}{D_{\mathrm{inf}}(\Delta_i, \sigma_i^2)} + \mathrm{O}((\log T)^{4/5}).$$

This corollary is straightforward from Lemma 5 with $\epsilon := \mathrm{O}((\log T)^{-1/5})$.

Note that $D_{\mathrm{inf}}(\mu^* - \mu_i, \sigma_2^2)$ is invariant under the location and scale transformation, that is,

$$D_{\mathrm{inf}}(\mu^* - \mu_i, \sigma_i^2) = D_{\mathrm{inf}}\left(\frac{\mu^*-a}{b} - \frac{\mu_i - a}{b}, \frac{\sigma_i^2}{b^2}\right).$$

Thus Theorem 1 easily follows from Corollary 6 by the transformation $((\mu_1,\sigma_1^2),(\mu_2,\sigma_2^2),\cdots,(\mu_K,\sigma_K^2)) \mapsto ((0,1),((\mu_2-\mu_1)/\sigma_1,\sigma_2^2/\sigma_1^2),\cdots,((\mu_K-\mu_1)/\sigma_1, \sigma_K^2/\sigma_1^2))$.

Proof of Lemma 5. Define events

$$\begin{aligned}
\mathcal{A}(t) &= \{\tilde{\mu}^*(t) \ge -\epsilon\}, \\
\mathcal{B}_i(t) &= \{\bar{x}_i(t) \le \mu_i + \epsilon,\ S_i(t) \le n(\sigma_i^2+\epsilon)\},
\end{aligned}$$

where $\tilde{\mu}^*(t) = \max_i \tilde{\mu}_i(t)$. Then the regret at the T-th round is bounded as

$$\begin{aligned}
&\mathrm{Regret}(T) \\
&= \sum_{t=1}^{T} \Delta_{J(t)} \\
&\le \Delta_{\max} \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \ne 1,\ \mathcal{A}^c(t)] \\
&\quad + \sum_{i=2}^{K} \Delta_i \left(\bar{n} + \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=i,\ \mathcal{A}(t)] \right) \\
&\le \Delta_{\max} \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \ne 1,\ \mathcal{A}^c(t)] \\
&\quad + \sum_{i=2}^{K} \Delta_i \Biggl(\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=i,\ \mathcal{A}(t),\ \mathcal{B}_i(t)] \\
&\qquad + \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=i,\ \mathcal{B}_i^c(l)] + \bar{n} \Biggr), \qquad (13)
\end{aligned}$$

where the superscript "c" denotes the complementary set. In the following Lemmas 7–9 we bound the expectation of the above three terms and the proof is completed. □

Lemma 7. *If $\alpha < 0$ then*

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \ne 1 \cup \mathcal{A}^c(t)] \right] \\
&\le \frac{1}{1-\mathrm{e}^{-\frac{\epsilon^2}{8}}} + \frac{1}{1-\mathrm{e}^{-h(2)}} + \frac{2\sqrt{2}}{\epsilon} \frac{\left(1+\frac{\epsilon^2}{8}\right)^{1-\alpha}}{1-\left(1+\frac{\epsilon^2}{8}\right)^{-1/2}} \\
&\quad + \frac{\mathrm{B}(1/2,-\alpha)}{\left(1-\mathrm{e}^{-\frac{\epsilon^2}{2}}\right)^2} \\
&= \mathrm{O}(\epsilon^{-4}).
\end{aligned}$$

Lemma 8. *For any $i \ne 1$,*

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=i,\ \mathcal{A}(t),\ \mathcal{B}_i(t)] \right] \\
&\le \frac{\log T}{D_{\mathrm{inf}}(\Delta_i - 2\epsilon, \sigma_i^2+\epsilon)} + 2 - 2\alpha + \frac{\sqrt{\sigma_i^2+\epsilon}}{\Delta_i - 2\epsilon} \\
&= \frac{\log T}{D_{\mathrm{inf}}(\Delta_i - 2\epsilon, \sigma_i^2+\epsilon)} + \mathrm{O}(1).
\end{aligned}$$

Lemma 9. *For any $i \ne 1$,*

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=i,\ \mathcal{B}_i^c(t)] \right] \\
&\le \frac{1}{1-\mathrm{e}^{-\frac{\epsilon^2}{2\sigma_i^2}}} + \frac{1}{1-\mathrm{e}^{-h\left(1+\frac{\epsilon}{\sigma_i^2}\right)}} = \mathrm{O}(\epsilon^{-2}).
\end{aligned}$$

We prove Lemmas 8 and 9 in the supplementary material and prove Lemma 7 in this section.

Whereas the second term of (13) becomes the main term of the regret, the evaluation of the first term is the most difficult point of the proof, which corresponds to Lemma 7. In fact, it is reported in Burnetas and Katehakis (1996) that they were not able to prove the asymptotic optimality of a policy for the Gaussian model because of difficulty of the evaluation corresponding to this term. Also note that this is the term which does not become a constant in the case $\alpha \geq 0$ and is considered in the proof of Theorem 2.

In this paper we evaluate this term by first bounding this term for a fixed statistic $\hat{\theta}_{i,n} = (\bar{x}_{i,n}, S_{i,n})$ and finally taking its expectation, whereas a probability on this statistic is first evaluated in Burnetas and Katehakis (1996). By leaving the evaluation on the distribution of $\hat{\theta}_{i,n}$ to the latter part, we can significantly simplify the integral by variable transformation.

Proof of Lemma 7. First we bound the summation as

$$\begin{aligned}
&\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \neq 1,\ \mathcal{A}^c(t)] \\
&= \sum_{n=\bar{n}}^{T} \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \neq 1,\ \mathcal{A}^c(t),\ N_1(t) = n] \\
&= \sum_{n=\bar{n}}^{T} \sum_{m=1}^{T} \\
&\qquad \mathbb{1}\left[m \leq \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \neq 1,\ \mathcal{A}^c(t),\ N_1(t) = n] \right] .
\end{aligned}$$

Note that

$$m \leq \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \neq 1,\ \mathcal{A}^c(t),\ N_1(t) = n]$$

implies that $\tilde{\mu}_1(t) \leq \tilde{\mu}^*(t) \leq -\epsilon$ occurred for the first m elements of $\{t : \mathcal{A}^c(t),\ N_1(t) = n\}$. Therefore,

$$\begin{aligned}
&\Pr\left[m \leq \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \neq 1,\ \mathcal{A}^c(t),\ N_1(t) = n] \right] \\
&\qquad\qquad \leq (1 - p_n(-\epsilon | \hat{\theta}_{1,n}))^m
\end{aligned}$$

and we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \neq 1 \cup \mathcal{A}^c(t)] \right] \\
&\leq \mathrm{E}\left[\sum_{n=\bar{n}}^{T} \sum_{m=1}^{T} (1 - p_n(-\epsilon | \hat{\theta}_{1,n}))^m \right] \\
&\leq \sum_{n=\bar{n}}^{T} \mathrm{E}\left[\frac{1 - p_n(-\epsilon | \hat{\theta}_{1,n})}{p_n(-\epsilon | \hat{\theta}_{1,n})} \right] . \qquad (14)
\end{aligned}$$

Since $1/p_n(-\epsilon | \hat{\theta}_{i,n}) \leq 2$ and $1 - p_n(-\epsilon | \hat{\theta}_{i,n}) \leq 1/2$ for $-\epsilon \leq \bar{x}_{i,n}$ from the symmetry of t-distribution, this expectation is partitioned into

$$\begin{aligned}
&\mathrm{E}\left[\frac{1 - p_n(-\epsilon | \hat{\theta}_{1,n})}{p_n(-\epsilon | \hat{\theta}_{1,n})} \right] \\
&\leq 2\mathrm{E}\left[\mathbb{1}[-\epsilon \leq \bar{x}_{1,n}] \, (1 - p_n(-\epsilon | \hat{\theta}_{1,n})) \right] \\
&\qquad + \mathrm{E}\left[\frac{\mathbb{1}[\bar{x}_{1,n} \leq -\epsilon]}{p_n(-\epsilon | \hat{\theta}_{1,n})} \right] \\
&\leq \Pr\left[-\epsilon < \bar{x}_{1,n} \leq -\epsilon/2 \right] \\
&\qquad + \Pr\left[-\epsilon/2 < \bar{x}_{1,n},\ S_{1,n} \geq 2n \right] \\
&\qquad + 2\mathrm{E}\big[\mathbb{1}[-\epsilon/2 < \bar{x}_{i,n},\ S_{1,n} \leq 2n] \, (1 - p_n(-\epsilon | \hat{\theta}_{i,n})) \big] \\
&\qquad + \mathrm{E}\left[\frac{\mathbb{1}[\bar{x}_{1,n} \leq -\epsilon]}{p_n(-\epsilon | \hat{\theta}_{1,n})} \right] . \qquad (15)
\end{aligned}$$

From Lemma 3, the first and second terms of (15) are bounded as

$$\begin{aligned}
\Pr\left[-\epsilon < \bar{x}_{1,n} \leq -\epsilon/2 \right] &\leq \mathrm{e}^{-\frac{n\epsilon^2}{8}} , \\
\Pr\left[-\epsilon/2 < \bar{x}_{1,n},\ S_{1,n} \geq 2n \right] &\leq \mathrm{e}^{-nh(2)} , \qquad (16)
\end{aligned}$$

respectively. Next, recall that $\hat{\theta}_{1,n} = (\bar{x}_{1,n}, S_{1,n})$. Then, from the symmetry of t-distribution

$$\begin{aligned}
1 - p_n(-\epsilon | \bar{x}_{1,n}, S_{1,n}) &= 1 - p_n(-\bar{x}_{1,n} - \epsilon | 0, S_{1,n}) \\
&= p_n(\bar{x}_{1,n} + \epsilon | 0, S_{1,n}) \\
&= p_n(2\bar{x}_{1,n} + \epsilon | \bar{x}_{1,n}, S_{1,n})
\end{aligned}$$

and the third term of (15) is bounded from (10) as

$$\begin{aligned}
&\mathrm{E}\big[\mathbb{1}[-\epsilon/2 < \bar{x}_{1,n},\ S_{1,n} \leq 2n] \, (1 - p_n(-\epsilon | \hat{\theta}_{i,n})) \big] \\
&= \mathrm{E}\big[\mathbb{1}[-\epsilon/2 < \bar{x}_{1,n},\ S_{1,n} \leq 2n] \, p_n(2\bar{x}_{1,n} - \epsilon | \hat{\theta}_{i,n}) \big] \\
&\leq \frac{2\sqrt{2}}{\epsilon} \left(1 + \frac{\epsilon^2}{8} \right)^{-\frac{n}{2} - \alpha + 1} . \qquad (17)
\end{aligned}$$

Finally we evaluate the fourth term of (15). From (1) and (11), we have

$$\begin{aligned}
&\mathrm{E}\left[\frac{\mathbb{1}[\bar{x}_{1,n} \leq -\epsilon]}{p_n(-\epsilon | \hat{\theta}_{1,n})} \right] \\
&\leq \frac{1}{A_{n,\alpha}} \int_{-\infty}^{-\epsilon} \int_{0}^{\infty} \left(1 + \frac{n(x+\epsilon)^2}{s} \right)^{\frac{n-1}{2} + \alpha} \\
&\qquad \cdot \sqrt{\frac{n}{2\pi}} \mathrm{e}^{-\frac{nx^2}{2}} \frac{s^{\frac{n-3}{2}} \mathrm{e}^{-\frac{s}{2}}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \mathrm{d}s \mathrm{d}x \\
&\leq \frac{\mathrm{e}^{-n\epsilon^2/2}}{A_{n,\alpha}} \int_{-\infty}^{-\epsilon} \int_{0}^{\infty} \left(1 + \frac{n(x+\epsilon)^2}{s} \right)^{\frac{n-1}{2} + \alpha} \\
&\qquad \cdot \sqrt{\frac{n}{2\pi}} \mathrm{e}^{-\frac{n(x+\epsilon)^2}{2}} \frac{s^{\frac{n-3}{2}} \mathrm{e}^{-\frac{s}{2}}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \mathrm{d}s \mathrm{d}x \qquad (18)
\end{aligned}$$

by $x^2 \ge (x+\epsilon)^2 + \epsilon^2$ for $x \le -\epsilon \le 0$. By letting

$$(x,s) = \left(-\epsilon - \sqrt{\frac{2zw}{n}},\, 2z(1-w)\right),$$

we have

$$\begin{aligned}\mathrm{d}x\mathrm{d}s &= \det\begin{pmatrix} -\sqrt{\frac{w}{2nz}} & 2(1-w) \\ -\sqrt{\frac{z}{2nw}} & -2z \end{pmatrix}\mathrm{d}z\mathrm{d}w \\ &= \sqrt{\frac{2z}{nw}}\mathrm{d}z\mathrm{d}w\end{aligned}$$

and (18) is rewritten as

$$\begin{aligned}
&\mathrm{E}\left[\frac{\mathbb{1}[\bar{x}_{i,n} \le -\epsilon]}{p_n(-\epsilon|\hat{\theta}_{i,n})}\right] \\
&\le \frac{\mathrm{e}^{-n\epsilon^2/2}}{A_{n,\alpha}} \int_0^1 \int_0^\infty \left(1 + \frac{w}{1-w}\right)^{\frac{n-1}{2}+\alpha} \\
&\qquad \cdot \sqrt{\frac{n}{2\pi}}\mathrm{e}^{-zw} \frac{(z(1-w))^{\frac{n-3}{2}}\mathrm{e}^{-z(1-w)}}{2\Gamma(\frac{n-1}{2})} \sqrt{\frac{2z}{nw}}\mathrm{d}z\mathrm{d}w \\
&= \frac{\mathrm{e}^{-n\epsilon^2/2}}{2\sqrt{\pi}A_{n,\alpha}\Gamma(\frac{n-1}{2})} \int_0^1 w^{-1/2}(1-w)^{-1-\alpha}\mathrm{d}w \\
&\qquad \cdot \int_0^\infty \mathrm{e}^{-z} z^{\frac{n}{2}-1}\mathrm{d}z \\
&= \frac{\mathrm{e}^{-n\epsilon^2/2}}{2\sqrt{\pi}A_{n,\alpha}\Gamma(\frac{n-1}{2})} \mathrm{B}(1/2,-\alpha)\Gamma(n/2) \\
&\le \frac{\mathrm{e}^{1/3}\sqrt{(n+2\alpha)(n-1)}}{2}\mathrm{e}^{-n\epsilon^2/2}\mathrm{B}(1/2,-\alpha) \qquad (19) \\
&\le n\mathrm{e}^{-n\epsilon^2/2}\mathrm{B}(1/2,-\alpha). \quad (\text{by } \mathrm{e}^{1/3} \le 2 \text{ and } \alpha < 0), \qquad (20)
\end{aligned}$$

where (19) follows from (12) and Lemma 10 in the supplementary material. By combining (14), (15), (16), (17) with (20), we obtain

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t) \ne 1,\ \mathcal{A}^c(t)]\right] \\
&\le \sum_{n=\bar{n}}^{T} \Bigg(\mathrm{e}^{-\frac{n\epsilon^2}{8}} + \mathrm{e}^{-nh(2)} + \frac{2\sqrt{2}}{\epsilon}\left(1+\frac{\epsilon^2}{8}\right)^{-\frac{n}{2}-\alpha+1} \\
&\qquad\qquad + n\mathrm{e}^{-n\epsilon^2/2}\mathrm{B}(1/2,-\alpha)\Bigg) \\
&\le \frac{1}{1-\mathrm{e}^{-\frac{\epsilon^2}{8}}} + \frac{1}{1-\mathrm{e}^{-h(2)}} + \frac{2\sqrt{2}}{\epsilon}\frac{\left(1+\frac{\epsilon^2}{8}\right)^{1-\alpha}}{1-\left(1+\frac{\epsilon^2}{8}\right)^{-1/2}} \\
&\quad + \frac{\mathrm{B}(1/2,-\alpha)}{\left(1-\mathrm{e}^{-\frac{\epsilon^2}{2}}\right)^2} \\
&= \mathrm{O}(\epsilon^{-2}) + \mathrm{O}(1) + \mathrm{O}(\epsilon^{-3}) + \mathrm{O}(\epsilon^{-4}) = \mathrm{O}(\epsilon^{-4}). \quad \square
\end{aligned}$$

6 Analysis for Optimistic Priors

In this section we prove Theorem 2. As mentioned before, the evaluation in the proof corresponds to Lemma 7, in which $\alpha < 0$ is required so that $\mathrm{B}(1/2,-\alpha)$ becomes finite. We show in the following proof that this requirement is actually necessary to achieve the asymptotic bound.

Proof of Theorem 2. We assume $(\mu_1,\sigma_1^2) = (0,1)$ without loss of generality. First we have

$$\begin{aligned}
\mathrm{E}[\mathrm{Regret}(T)] &= \Delta_i \sum_{t=1}^{T} \mathbb{1}[J(t)=2] \\
&\ge \Delta_i \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=2,\ N_1(t)=\bar{n}]\,. \qquad (21)
\end{aligned}$$

Note that $N_1(K\bar{n}+1) = \bar{n}$ holds and $\{J(t) \ne 2,\ N_1(t) = \bar{n}\}$ implies $N_1(t') > \bar{n}$ for any $t' > t$. Therefore, for any $t \ge \bar{n}$,

$$\begin{aligned}
\{J(t)=2,\ N_1(t)=\bar{n}\} &\Leftrightarrow \bigcup_{k=1}^{t-K\bar{n}} \{J(K\bar{n}+k)=2\} \\
&\Leftarrow \bigcup_{k=1}^{t-K\bar{n}} \{\tilde{\mu}_1(K\bar{n}+k) < \mu_2\}\,.
\end{aligned}$$

By defining $\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}) = 1 - p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})$ and $T' = T - K\bar{n}$ we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[J(t)=2,\ N_1(t)=\bar{n}]\right] \\
&\ge \mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}\left[\bigcup_{k=1}^{t-K\bar{n}} \{\tilde{\mu}_1(K\bar{n}+k) < \mu_2\}\right]\right] \\
&= \mathrm{E}\left[\sum_{t=K\bar{n}+1}^{T} (\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{t-K\bar{n}}\right] \\
&\ge \mathrm{E}\left[\sum_{m=1}^{T'} (\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{m}\right] \\
&= \mathrm{E}\left[\left(1-(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T'}\right)\frac{\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}{p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}\right] \\
&\ge \frac{1}{2}\mathrm{E}\left[\mathbb{1}\left[(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T'} \le 1/2\right]\frac{\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}{p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}\right] \\
&\ge \frac{1}{2}\mathrm{E}\left[\frac{\mathbb{1}\left[(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T'} \le 1/2\right]}{p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}\right] - \frac{1}{2}\,. \qquad (22) \\
&\qquad\qquad (\text{by } (1-p)/p = 1/p - 1)
\end{aligned}$$

Here we obtain from (11) that

$$
\begin{aligned}
&(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T'} \le 1/2 \\
&\Leftarrow \left(1+\frac{\bar{n}(\mu_2-\bar{x}_{1,\bar{n}})^2}{S_{1,\bar{n}}}\right)^{-\frac{\bar{n}-1}{2}-\alpha} \ge \frac{1-2^{-\frac{1}{T}}}{A_{\bar{n},\alpha}} \\
&\Leftarrow \left(1+\frac{\bar{n}(\mu_2-\bar{x}_{1,\bar{n}})^2}{S_{1,\bar{n}}}\right)^{-\frac{\bar{n}-1}{2}-\alpha} \ge \frac{\log 2}{A_{\bar{n},\alpha}T'} \\
&\qquad\qquad\qquad\qquad\qquad\qquad (\text{by } 2^x \ge 1 + x\log 2) \\
&\Leftrightarrow \frac{\bar{n}(\mu_2-\bar{x}_{1,\bar{n}})^2}{S_{1,\bar{n}}} \le \left(\frac{A_{\bar{n},\alpha}T'}{\log 2}\right)^{\frac{1}{\frac{\bar{n}-1}{2}+\alpha}} - 1\,.
\end{aligned}
$$

Therefore, by letting

$$
\begin{aligned}
C_T &:= \left(\frac{A_{\bar{n},\alpha}T'}{\log 2}\right)^{\frac{1}{\frac{\bar{n}-1}{2}+\alpha}} - 1 \\
&= \left(\frac{A_{\bar{n},\alpha}(T-K\bar{n})}{\log 2}\right)^{\frac{2}{\bar{n}-1+2\alpha}} - 1 \\
&= \mathrm{O}\left(T^{\frac{2}{\bar{n}-1+2\alpha}}\right)\,, \qquad (23)
\end{aligned}
$$

we can bound the expectation in (22) from (10) as

$$
\begin{aligned}
&\mathrm{E}\left[\frac{\mathbb{1}\left[(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T'} \le 1/2\right]}{p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}\right] \\
&\ge \iint_{\substack{x\le\mu_2,\, s\ge 0,\\ \frac{\bar{n}(\mu_2-x)^2}{s}\le C_T}} \frac{\sqrt{\bar{n}}(\mu_2-x)}{\sqrt{s}}\left(1+\frac{\bar{n}(\mu_2-x)^2}{s}\right)^{\frac{\bar{n}}{2}+\alpha-1} \\
&\qquad\qquad \cdot\sqrt{\frac{\bar{n}}{2\pi}}\mathrm{e}^{-\frac{\bar{n}x^2}{2}}\frac{s^{\frac{\bar{n}-3}{2}}\mathrm{e}^{-\frac{s}{2}}}{2^{\frac{\bar{n}-1}{2}}\Gamma(\frac{\bar{n}-1}{2})}\mathrm{d}x\mathrm{d}s \\
&= \frac{\bar{n}\mathrm{e}^{-\bar{n}\mu_2^2/2}}{\sqrt{\pi}2^{\frac{\bar{n}}{2}}\Gamma(\frac{\bar{n}-1}{2})} \iint_{\substack{x\le\mu_2,\, s\ge 0,\\ \frac{\bar{n}(\mu_2-x)^2}{s}\le C_T}} \mathrm{e}^{-\frac{\bar{n}(\mu_2-x)^2}{2}+\mu_2(\mu_2-x)} \\
&\quad \cdot(\mu_2-x)\left(1+\frac{\bar{n}(\mu_2-x)^2}{s}\right)^{\frac{\bar{n}}{2}+\alpha-1} s^{\frac{\bar{n}}{2}-2}\mathrm{e}^{-\frac{s}{2}}\mathrm{d}x\mathrm{d}s\,.
\end{aligned}
$$

By letting

$$
(x,s) = \left(\mu_2 - \sqrt{\frac{2zw}{\bar{n}}},\, 2z(1-w)\right)\,,
$$

we obtain in a similar way to (20) that

$$
\begin{aligned}
&\mathrm{E}\left[\frac{\mathbb{1}\left[(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T} \le 1/2\right]}{p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}\right] \\
&\ge \frac{\mathrm{e}^{-\bar{n}\mu_2^2/2}}{2\sqrt{\pi}\Gamma(\frac{\bar{n}-1}{2})}\int_0^{\frac{1}{1+\frac{1}{C_T}}}\int_0^\infty \mathrm{e}^{-z}\mathrm{e}^{\mu_2\sqrt{\frac{2zw}{\bar{n}}}} \\
&\qquad \cdot z^{\frac{\bar{n}}{2}-1}(1-w)^{-1-\alpha}\mathrm{d}z\mathrm{d}w \\
&\ge \frac{\mathrm{e}^{-\bar{n}\mu_2^2/2}}{2\sqrt{\pi}\Gamma(\frac{\bar{n}-1}{2})}\int_0^\infty \mathrm{e}^{\mu_2\sqrt{\frac{2z}{\bar{n}}}}\mathrm{e}^{-z}z^{\frac{\bar{n}}{2}-1}\mathrm{d}z \\
&\quad \cdot\int_0^{\frac{1}{1+\frac{1}{C_T}}}(1-w)^{-1-\alpha}\mathrm{d}w\,. \quad (\text{by } \mu_2 < \mu_1 = 0)
\end{aligned}
$$

Here note that

$$
\int_0^{\frac{1}{1+\frac{1}{C_T}}}(1-w)^{-1-\alpha}\mathrm{d}w = \begin{cases}\log(1+C_T), & \alpha = 0,\\ \frac{(1+C_T)^\alpha-1}{\alpha}, & \alpha > 0.\end{cases}
$$

Then there exists a constant $B_{\bar{n},\alpha,\mu_2}$ such that

$$
\begin{aligned}
&\mathrm{E}\left[\frac{\mathbb{1}\left[(\bar{p}_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}}))^{T'} \le 1/2\right]}{p_{\bar{n}}(\mu_2|\hat{\theta}_{1,\bar{n}})}\right] \\
&\ge \begin{cases} B_{\bar{n},\alpha,\mu_2}\log(1+C_T), & \alpha = 0,\\ B_{\bar{n},\alpha,\mu_2}((1+C_T)^\alpha - 1), & \alpha > 0.\end{cases} \qquad (24)
\end{aligned}
$$

Finally by putting (21), (22) and (24) together we obtain for $\alpha = 0$ that

$$
\mathrm{E}[\mathrm{Regret}(T)] \ge (\Delta_2/2)(B_{n,\alpha,\mu_2}\log(1+C_T)-1)\,.
$$

Eq. (6) easily follows from (23). Eq. (7) for $\alpha > 0$ is obtained in the same way. □

7 Conclusion

We considered the stochastic multiarmed bandit problem such that each reward follows a normal distribution with an unknown mean and variance. We proved that Thompson sampling with prior $\pi(\mu_i, \sigma_i^2) \propto (\sigma_i^2)^{-1-\alpha}$ achieves the asymptotic bound if $\alpha < 0$ but cannot if $\alpha \ge 0$, which includes reference prior $\alpha = 0$ and Jeffreys prior $\alpha = 1/2$.

A future work is to examine whether TS with noninformative priors is risky or not for other multiparameter models where TS is used without theoretical analysis (see e.g., Chapelle and Li (2012)). Since the analysis of this paper heavily depends on the specific form of normal distributions, it is currently unknown whether the technique of this paper can be applied to other models and this generalization remains as an important open problem.

References

Agrawal, S., & Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. *Proceedings of COLT 2012*, *23*, 39.1–39.26.

Agrawal, S., & Goyal, N. (2013). Further optimal regret bounds for Thompson sampling. *Proceedings of Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)* (pp. 99–107).

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

Burnetas, A. N., & Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, *17*, 122–142.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., & Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, *41*, 1516–1541.

Chapelle, O., & Li, L. (2012). An empirical evaluation of Thompson sampling. *Proceedings of NIPS 2011* (pp. 1252–1260). Granada, Spain.

Dembo, A., & Zeitouni, O. (1998). *Large deviations techniques and applications*, vol. 38 of *Applications of Mathematics*. New York: Springer-Verlag. Second edition.

Honda, J., & Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. *Proceedings of COLT 2010* (pp. 67–79). Haifa, Israel.

Kaufmann, E., Cappé, O., & Garivier, A. (2012a). On Bayesian upper confidence bounds for bandit problems. *Proceedings of Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS2012)* (pp. 592–600).

Kaufmann, E., Korda, N., & Munos, R. (2012b). Thompson sampling: an asymptotically optimal finite-time analysis. *Proceedings of the 23rd international conference on Algorithmic Learning Theory (ALT'12)* (pp. 199–213). Berlin, Heidelberg: Springer-Verlag.

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics*, vol. 1: Distribution theory. C. Griffin London. 4th edition.

Korda, N., Kaufmann, E., & Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. *Proceedings of NIPS 2013*. Lake Tahoe, NV, USA.

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

Olver, F. W., Lozier, D. W., Boisvert, R. F., & Clark, C. W. (2010). *NIST handbook of mathematical functions*. New York, NY, USA: Cambridge University Press. 1st edition.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–35.

Robert, C. P. (2001). *The Bayesian choice*. New York: Springer. 2nd edition.

Russo, D., & Roy, B. V. (2013). Learning to optimize via posterior sampling. arXiv:1301.2609.

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