Optimality of Thompson Sampling with Noninformative Priors for Pareto Bandits

Jongyeong Lee [1 2] **Junya Honda** [3 2] **Chao-Kai Chiang** [1] **Masashi Sugiyama** [2 1]

Abstract

In the stochastic multi-armed bandit problem, a randomized probability matching policy called Thompson sampling (TS) has shown excellent performance in various reward models. In addition to the empirical performance, TS has been shown to achieve asymptotic problem-dependent lower bounds in several models. However, its optimality has been mainly addressed under light-tailed or one-parameter models that belong to exponential families. In this paper, we consider the optimality of TS for the Pareto model that has a heavy tail and is parameterized by two unknown parameters. Specifically, we discuss the optimality of TS with probability matching priors that include the Jeffreys prior and the reference priors. We first prove that TS with certain probability matching priors can achieve the optimal regret bound. Then, we show the suboptimality of TS with other priors, including the Jeffreys and the reference priors. Nevertheless, we find that TS with the Jeffreys and reference priors can achieve the asymptotic lower bound if one uses a truncation procedure. These results suggest carefully choosing noninformative priors to avoid suboptimality and show the effectiveness of truncation procedures in TS-based policies.

1. Introduction

In the multi-armed bandit (MAB) problem, an agent plays an arm and observes a reward only from the played arm, which is partial feedback (Thompson, 1933; Robbins, 1952). The rewards are further assumed to be generated from the distribution of the corresponding arm in the stochastic MAB problem (Bubeck et al., 2012). Since only partial observations are available, the agent has to estimate unknown distributions to guess which arm is optimal while avoiding playing suboptimal arms that induce loss of resources. Thus, the agent has to cope with the dilemma of exploration and exploitation.

In this problem, Thompson sampling (TS), a randomized Bayesian policy that plays an arm according to the posterior probability of being optimal, has been widely adopted because of its outstanding empirical performance (Chapelle & Li, 2011; Russo et al., 2018). Following its empirical success, theoretical analysis of TS has been conducted for several reward models such as Bernoulli models (Agrawal & Goyal, 2012; Kaufmann et al., 2012), one-dimensional exponential families (Korda et al., 2013), Gaussian models (Honda & Takemura, 2014), and bounded support models (Riou & Honda, 2020; Baudry et al., 2021) where asymptotic optimality of TS was established. Here, an algorithm is said to be asymptotically optimal if it can achieve the theoretical problem-dependent lower bound derived by Lai et al. (1985) for one-parameter models and Burnetas & Katehakis (1996) for multiparameter or nonparametric models. Note that the performance of any reasonable algorithms cannot be better than these lower bounds.

Apart from the problem-dependent regret analysis, several works studied the problem-independent or prior-independent bounds of TS (Bubeck & Liu, 2013; Russo & Van Roy, 2016; Agrawal & Goyal, 2017). In this paper, we study how the choice of noninformative priors affects the performance of TS for any given problem instance. In other words, we focus on the asymptotic optimality of TS depending on the choice of noninformative priors.

The asymptotic optimality of TS has been mainly considered in the one-parameter model, while its optimality under the multiparameter model has not been well-studied. To the best of our knowledge, the asymptotic optimality of TS in the noncompact multiparameter model is only known for the Gaussian bandits (Honda & Takemura, 2014) where both the mean and variance are unknown. They showed that TS with the uniform prior is optimal while TS with the Jeffreys prior and reference prior cannot achieve the lower bound. The success of the uniform prior is due to its frequent playing

[1]Department of Computer Science, University of Tokyo, Tokyo, Japan [2]RIKEN AIP, Tokyo, Japan [3]Department of Systems Science, Kyoto University, Kyoto, Japan. Correspondence to: Jongyeong Lee <lee@ms.k.u-tokyo.ac.jp>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

of seemingly suboptimal arms. Its conservativeness comes from a moderate overestimation of the posterior probability that current suboptimal arms might be optimal.

In this paper, we consider the two-parameter Pareto models where the tail function is heavy-tailed. We first derive the closed form of the problem-dependent constant that appears in the theoretical lower bound in Pareto models, which is not trivial, unlike those for exponential families. Based on this result, we show that TS with some probability-matching priors achieves the optimal bound, which is the first result for two-parameter Pareto bandit models, to our knowledge.

We further show that TS with different choices of probability matching priors, called optimistic priors, suffers a polynomial regret in expectation. Therefore, being conservative would be better when one chooses noninformative priors to avoid suboptimality in view of expectation. Nevertheless, we show that TS with the Jeffreys prior or the reference prior can achieve the optimal regret bound if we add a truncation procedure on the shape parameter. Our contributions are summarized as follows:

- We prove the asymptotic optimality/suboptimality of TS under different choices of priors, which shows the importance of the choice of noninformative priors in cases of two-parameter Pareto models.
- We provide another option to achieve optimality: adding a truncation procedure to the parameter space of the posterior distribution instead of finding an optimal prior.

This paper is organized as follows. In Section 2, we formulate the stochastic MAB problems under the Pareto distribution and derive its regret lower bound. Based on the choice of noninformative priors and their corresponding posteriors, we formulate TS for the Pareto models and propose another TS-based algorithm to solve the suboptimality problem of the Jeffreys prior and the reference prior in Section 3. In Section 4, we provide the main results on the optimality of TS and TS with a truncation procedure, whose proof outline is given in Section 6. Numerical results that support our theoretical analysis are provided in Section 5.

2. Preliminaries

In this section, we formulate the stochastic MAB problem. We derive the exact form of the problem-dependent constant that appears in the lower bound of the expected regret in Pareto bandits.

2.1. Notations

We consider the stochastic K-armed bandit problem where the rewards are generated from Pareto distributions with fixed parameters. An agent chooses an arm a in $[K] := \{1, \dots, K\}$ at each round $t \in \mathbb{N}$ and observes an independent and identically distributed reward from $\mathrm{Pa}(\kappa_a, \alpha_a)$, where $\mathrm{Pa}(\kappa, \alpha)$ denotes the Pareto distribution parameterized by scale $\kappa > 0$ and shape $\alpha > 0$. This has the density function of the form

$$f^{\mathrm{Pa}}_{\kappa,\alpha}(x) = \frac{\alpha \kappa^{\alpha}}{x^{\alpha+1}} \mathbb{1}[x \geq \kappa], \tag{1}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function. We consider a bandit model where parameters $\theta_a = (\kappa_a, \alpha_a) \in \mathbb{R}_+ \times (1, \infty)$ are unknown to the agent. We denote the mean of a random variable following $\mathrm{Pa}(\theta_a)$ by $\mu_a = \mu(\theta_a) := \frac{\kappa_a \alpha_a}{\alpha_a - 1}$. Note that $\alpha > 1$ is a necessary condition of an arm to have a finite mean, which is required to define the sub-optimality gap $\Delta_a := \max_{i \in [K]} \mu_i - \mu_a$. We assume without loss of generality that the arm 1 has the maximum mean for simplicity, i.e., $\mu_1 = \max_{i \in [K]} \mu_i$. Let $j(t)$ be the arm played at round $t \in \mathbb{N}$ and $N_a(t) = \sum_{s=1}^{t-1} \mathbb{1}[j(s) = a]$ denote the number of rounds the arm a is played until round t. Then, the regret at round T is given as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \Delta_{j(t)} = \sum_{a=2}^{K} \Delta_a N_a(T+1).$$

Let $r_{a,n}$ be the n-th reward generated from the arm a. In the Pareto distribution, the maximum likelihood estimators (MLEs) of κ, α for arm a given n rewards and their distributions are given as follows (Malik, 1970):

$$\begin{aligned} \hat{\kappa}_a(n) &= \min_{s \in [n]} r_{a,s} \sim \mathrm{Pa}(\kappa_a, n\alpha_a), \\ \hat{\alpha}_a(n) &= \frac{n}{\sum_{s=1}^{n} \log(r_{a,s}) - n \log \hat{\kappa}_a(n)} \\ &\sim \mathrm{IG}(n-1, n\alpha_a), \end{aligned} \tag{2}$$

where $\mathrm{IG}(n, \alpha)$ denotes the inverse-gamma distribution with shape $n > 0$ and scale $\alpha > 0$. Note that Malik (1970) further showed the stochastic independence of $\hat{\alpha}(n)$ and $\hat{\kappa}(n)$.

2.2. Asymptotic lower bound

Burnetas & Katehakis (1996) provided a problem-dependent lower bound of the expected regret such that any uniformly fast convergent policy, which is a policy satisfying $\mathrm{Reg}(T) = o(T^{\alpha})$ for all $\alpha \in (0, 1)$, must satisfy

$$\begin{aligned} &\liminf_{T \to \infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \\ &\qquad \geq \sum_{a=2}^{K} \frac{\Delta_a}{\inf_{\theta : \mu(\theta) > \mu_1} \mathrm{KL}(\mathrm{Pa}(\kappa_a, \alpha_a), \mathrm{Pa}(\theta))}, \end{aligned} \tag{3}$$

where $\mathrm{KL}(\cdot, \cdot)$ denotes the Kullback-Leibler (KL) divergence. Notice that the bandit model $(\theta_a)_{a \in [K]}$ is considered as a fixed constant in the problem-dependent analysis.

The KL divergence between Pareto distributions is given as

$$\mathrm{KL}(\mathrm{Pa}(\kappa_1,\alpha_1),\mathrm{Pa}(\kappa_2,\alpha_2)) = \begin{cases} \log\left(\frac{\alpha_1}{\alpha_2}\right)+\alpha_2\log\left(\frac{\kappa_1}{\kappa_2}\right)+\frac{\alpha_2}{\alpha_1}-1 & \text{if } \kappa_2\le\kappa_1,\\ \infty & \text{otherwise.}\end{cases}$$

Here the divergence sometimes becomes infinity since the scale parameter κ determines the support of the Pareto distribution. We denote the numerator in (3) for $a\neq 1$ by

$$\begin{aligned}\mathrm{KL}_{\inf}(a) &:= \inf_{\theta:\mu(\theta)>\mu_1}\mathrm{KL}(\mathrm{Pa}(\kappa_a,\alpha_a),\mathrm{Pa}(\theta))\\ &= \inf_{\theta\in\Theta_a}\log\frac{\alpha_a}{\alpha}+\alpha\log\frac{\kappa_a}{\kappa}+\frac{\alpha}{\alpha_a}-1,\end{aligned}$$

where

$$\Theta_a=\{(\kappa,\alpha)\in(0,\kappa_a]\times(0,\infty):\mu(\kappa,\alpha)>\mu_1\}. \quad (4)$$

Notice that Θ_a allows parameters whose expected rewards are infinite ($\alpha\in(0,1]$), although we consider a bandit model with $\alpha_a>1$ for all $a\in[K]$ so that the sub-optimality gap Δ_a becomes finite. This implies that $\mathrm{KL}_{\inf}(a)$ does not depend on whether the agent considers the possibility that an arm has the infinite expected reward or not. Then, we can simply rewrite the lower bound in (3) as

$$\liminf_{T\to\infty}\frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T}\ge\sum_{a=2}^{K}\frac{\Delta_a}{\mathrm{KL}_{\inf}(a)}.$$

The following lemma shows the closed form of this infimum, whose proof is given in Appendix B.

Lemma 1. *For any arm $a\neq 1$, it holds that*

$$\mathrm{KL}_{\inf}(a)=\log\left(\alpha_a\frac{\mu_1-\kappa_a}{\mu_1}\right)+\frac{1}{\alpha_a}\frac{\mu_1}{\mu_1-\kappa_a}-1.$$

2.3. Relation with bounded moment models

In MAB literature, several algorithms based on the upper confidence bound (UCB) were proposed to tackle heavy-tailed models with infinite variance under additional assumptions on moments (Bubeck et al., 2013). One major assumption is that the moment of any arm a satisfies $\mathbb{E}[|r_{a,n}|^\gamma]\le v$ for some fixed $\gamma\in[1,2)$ and known $v<\infty$ (Bubeck et al., 2013). Note that the γ-th raw moment of the density function of X following $\mathrm{Pa}(\kappa,\alpha)$ is given as

$$\mathbb{E}[X^\gamma]=\begin{cases}\infty & \alpha\le\gamma,\\ \frac{\alpha\kappa^\gamma}{\alpha-\gamma} & \alpha>\gamma,\end{cases} \quad (5)$$

which implies that the Pareto models and the bounded moment models are not a subset of each other.

Recently, Agrawal et al. (2021) proposed an asymptotically optimal KL-UCB based algorithm that requires solving the optimization problem at every round. Since the bounded moment model only covers certain Pareto distributions in general, the known optimality result of KL-UCB does not necessarily imply the optimality in the sense of (3).

3. Thompson sampling and probability matching priors

TS is a policy from the Bayesian viewpoint, where the choices of priors are important. Although one can utilize prior knowledge on parameters when choosing the prior, such information would not always be available in practice. To deal with such scenarios, we consider noninformative priors based on the Fisher information (FI) matrix, which does not assume any information on unknown parameters.

For a random variable X with density $f(\cdot|\boldsymbol{\theta})$, FI is defined as the variance of the score, a partial derivative of $\log f$ with respect to $\boldsymbol{\theta}$, which is given as follows (Cover & Thomas, 2006):

$$\begin{aligned}[I(\boldsymbol{\theta})]_{ij}&=I_{ij}\\ &=\mathbb{E}_X\left[\left(\frac{\partial}{\partial\theta_i}\log f(X|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f(X|\boldsymbol{\theta})\right)\middle|\boldsymbol{\theta}\right]. \quad (6)\end{aligned}$$

It is known that the FI matrix in (6) coincides with the negative expected value of the Hessian matrix of $\log f(X|\boldsymbol{\theta})$ if the model satisfies the FI regular condition (Schervish, 2012). However, $\mathrm{Pa}(\kappa,\alpha)$ does not satisfy this condition since it is a parametric-support family. Therefore, for X with density function in (1), one can obtain the FI matrix of $\mathrm{Pa}(\kappa,\alpha)$ based on (6) as follows (Li et al., 2022):

$$I(\kappa,\alpha)=\begin{bmatrix}\frac{\alpha^2}{\kappa^2} & 0\\ 0 & \frac{1}{\alpha^2}\end{bmatrix}=\begin{bmatrix}I_{11}(\kappa)I_{11}(\alpha) & 0\\ 0 & I_{22}(\alpha)\end{bmatrix}, \quad (7)$$

where $I_{11}(\kappa)=\frac{1}{\kappa^2}$, $I_{11}(\alpha)=\alpha^2$, and $I_{22}(\alpha)=\frac{1}{\alpha^2}$. Note that I_{11} differs from $-\mathbb{E}\left[\frac{\partial^2}{\partial\kappa^2}\log f^{\mathrm{Pa}}_{\kappa,\alpha}(X;\theta)\middle|\theta\right]=\frac{\alpha}{\kappa^2}$.

Based on (7), the Jeffreys prior and the reference prior are given as $\pi_{\mathrm{J}}(\kappa,\alpha)\propto\sqrt{\det(I)}=\frac{1}{\kappa}$ and $\pi_{\mathrm{R}}(\kappa,\alpha)\propto\sqrt{I_{11}(\kappa)I_{22}(\alpha)}=\frac{1}{\kappa\alpha}$, respectively. Here, the reverse reference prior is the same as the reference prior from the orthogonality of parameters (Datta & Ghosh, 1995; Datta, 1996).

From the orthogonality of parameters, the probability matching prior when κ is of interest and α is the nuisance parameter is given as

$$\pi_{\mathrm{P}}(\kappa,\alpha)\propto\sqrt{I_{11}}g_1(\alpha)=\frac{\alpha}{\kappa}g_1(\alpha)$$

for arbitrary $g_1(\alpha)>0$ (Tibshirani, 1989). In this paper, we consider the prior $\pi(\kappa,\alpha)\propto\frac{\alpha^{-k}}{\kappa}$ for $k\in\mathbb{Z}$ since the cases $k=0,1$ correspond to the Jeffreys prior and the (reverse) reference prior, respectively.

Remark 1. The Pareto distribution discussed in this paper is sometimes called the Pareto type 1 distribution (Arnold, 2008). On the other hand, Kim et al. (2009) derived several noninformative priors for a special case of the Pareto

Algorithm 1 STS / STS-T

Parameter: $k \in \mathbb{Z}$, $\bar{n} = \max\{2, k+1\}$.
Initialization: Select each arm $\bar{n}$ times.
Loop:
Sample $\tilde{\alpha}_a(t) \sim \text{Erlang}\left(N_a(t) - k, \frac{N_a(t)}{\hat{\alpha}_a(N_a(t))}\right)$.
$\bar{\alpha}_a(N_a(t)) \leftarrow \min(N_a(t), \hat{\alpha}_a(N_a(t)))$.
Sample $\tilde{\alpha}_a(t) \sim \text{Erlang}\left(N_a(t) - k, \frac{N_a(t)}{\bar{\alpha}_a(N_a(t))}\right)$.
if $\{a \in [K] : \tilde{\alpha}_a(t) \leq 1\} \neq \emptyset$ **then**
  Select $j(t) = \arg\min_{a \in [K]} \tilde{\alpha}_a(t)$.
else
  Sample $u_a \sim U(0,1)$ for every $a \in [K]$.
  $\tilde{\kappa}_a(t) = \hat{\kappa}_a(N_a(t)) u_a^{1/(N_a(t)\tilde{\alpha}_a(t))}$.
  Play $j(t) = \arg\max_{a \in [K]} \frac{\tilde{\kappa}_a(t)\tilde{\alpha}_a(t)}{\tilde{\alpha}_a(t)-1}$
    $= \arg\max_{a \in [K]} \tilde{\mu}_a(t)$.
end if

type 2 distribution called the Lomax distribution (Lomax, 1954), where the FI matrix can be written using the negative Hessian.

In the multiparameter cases, the Jeffreys prior is known to suffer from many problems (Datta & Ghosh, 1996; Ghosh, 2011). For example, it is known that the Jeffreys prior leads to inconsistent estimators for the error variance in the Neyman-Scott problem (see Berger & Bernardo, 1992, Example 3.). This might be a possible reason why TS with Jeffreys prior suffers a polynomial expected regret in a multiparameter distribution setting. More details on the probability matching prior and the Jeffreys prior can be found in Appendix E.

3.1. Sampling procedure

Let $\mathcal{F}_t := (j(s), r_{j(s),N_{j(s)}(s)})_{s=1}^{t-1}$ be the history until round t. Under the prior $\frac{\alpha^{-k}}{\kappa}$ with $k \in \mathbb{Z}$, the marginalized posterior distribution of the shape parameter of arm a is given as

$$\alpha_a \mid \mathcal{F}_t \sim \text{Erlang}\left(N_a(t) - k, \frac{N_a(t)}{\hat{\alpha}_a(N_a(t))}\right), \quad (8)$$

where $\text{Erlang}(s, \beta)$ denotes the Erlang distribution with shape s and rate β. Note that we require $\bar{n} \geq \max\{2, k+1\}$ initial plays to avoid improper posteriors and MLE of α. When the shape parameter α_a is given as β, the cumulative distribution function (CDF) of the conditional posterior of κ_a is given as

$$\mathbb{P}\left[\kappa_a \leq x \mid \mathcal{F}_t, \alpha_a = \beta\right] = \left(\frac{x}{\hat{\kappa}_a(N_a(t))}\right)^{\beta N_a(t)}, \quad (9)$$

if $0 < x \leq \hat{\kappa}_a(N_a(t))$. Since one can derive the posteriors following the same steps as Sun et al. (2020), the detailed derivation is postponed to Appendix E.3. At round t, we denote the sampled scale and shape parameters of arm a by $\tilde{\kappa}_a(t)$ and $\tilde{\alpha}_a(t)$, respectively, and the corresponding mean reward by $\tilde{\mu}_a(t) := \mu(\tilde{\kappa}_a(t), \tilde{\alpha}_a(t))$. We first sample the shape parameter from the marginalized posterior in (8). Then, we sample the scale parameter given the sampled shape parameter from the CDF of the conditional posterior in (9) by using inverse transform sampling. TS based on this sequential procedure, which we call Sequential Thompson Sampling (STS), can be formulated as Algorithm 1.

In Theorem 3 given in the next section, STS with the Jeffreys prior and the reference prior turns out to be suboptimal in view of the lower bound in (3). Their suboptimality is mainly due to the behavior of the posterior in (8) when $\hat{\alpha}_1(n)$ is overestimated for small $N_1(t) = n$. To overcome such issues, we propose the STS-T policy, a variant of STS with truncation, where we replace $\hat{\alpha}(n)$ with $\bar{\alpha}(n) := \min(n, \hat{\alpha}(n))$ in (8). Note that such a truncation procedure is especially considered in the posterior sampling by (8) and (9). We show that STS-T with the Jeffreys prior and the reference prior can achieve the optimal regret bound in Theorem 4.

3.2. Interpretation of the prior parameter k

The Erlang distribution is a special case of the Gamma distribution, where the shape parameter is a positive integer. If a random variable X follows $\text{Erlang}(s, \beta)$, then it has the density of form

$$f^{\text{Er}}_{s,\beta}(x) = \frac{\beta^s}{\Gamma(s)} x^{s-1} e^{-\beta x} \mathbb{1}[x \in \mathbb{R}_+], \quad (10)$$

where $s \in \mathbb{N}$ and $\beta > 0$ denote the shape and rate parameter, respectively. Then, the CDF evaluated at $x > 0$ is given as

$$F^{\text{Er}}_{s,\beta}(x) = \frac{\int_0^{\beta x} t^{s-1} e^{-t} \mathrm{d}t}{\Gamma(s)} = \frac{\gamma(s, \beta x)}{\Gamma(s)}, \quad (11)$$

where $\gamma(\cdot, \cdot)$ denotes the lower incomplete gamma function. Since $\gamma(s+1, x) = s\gamma(s, x) - x^s e^{-x}$ holds, one can observe that for any $x > 0$

$$F^{\text{Er}}_{s,\beta}(x) \geq F^{\text{Er}}_{s+1,\beta}(x). \quad (12)$$

From the sampling procedure of STS and STS-T, $\tilde{\mu}$ depends on $\tilde{\kappa}$ only when $\tilde{\alpha} > 1$ holds since $\tilde{\alpha} \leq 1$ results in $\mu(\cdot, \tilde{\alpha}) = \infty$. Therefore, for any $\beta > 1$ in (9), $\tilde{\kappa}$ will concentrate on $\hat{\kappa}$ for sufficiently large $N_a(t) = n$. Thus, $\tilde{\mu}$ will be mainly determined by $\tilde{\alpha}$ and $\hat{\kappa}$, where the choice of k affects the sampling of $\tilde{\alpha}$ by (8). From (12), one could see that the probability of sampling small $\tilde{\alpha}$ increases as shape $n - k$ decreases. Therefore, $\tilde{\mu}$ of suboptimal arms would increase as k increases for the same n. In other words, the probability of sampling large $\tilde{\mu}$ becomes large as k increases. Therefore, TS with large k becomes a conservative policy that could frequently play currently suboptimal arms. In contrast, priors with small k yield an optimistic policy that focuses on playing the current best arm.

4. Main results

In this section, we provide regret bounds of STS and STS-T with different choices of $k \in \mathbb{Z}$. At first, we show the asymptotic optimality of STS for priors $\pi(\kappa, \alpha) \propto \frac{\alpha^{-k}}{\kappa}$ with $k \in \mathbb{Z}_{\geq 2}$.

Theorem 2. *Assume that arm 1 is the unique optimal arm with a finite mean. For every $a \in [K]$, let $\varepsilon_a = \min\left\{\frac{\kappa_a}{\alpha_a(\kappa_a+1)}, \frac{\kappa_a\delta_a}{\mu_a(\mu_a+\delta_a-\kappa_a)+\kappa_a\delta_a}, \frac{\kappa_a\delta_a}{\mu_a(1+\mu_a+\delta_a)}\right\}$ where $\delta_a = \frac{\Delta_a}{2}$ for $a \neq 1$ and $\delta_1 = \min_{a\neq 1}\delta_a$. Given arbitrary $\epsilon \in (0, \min_{a\in[K]}\varepsilon_a)$, the expected regret of* STS *with $k \in \mathbb{Z}_{\geq 2}$ is bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{a=2}^{K} \frac{\Delta_a \log T}{D_{a,k}(\epsilon)} + \mathcal{O}\left(\epsilon^{-2}\right).$$

Here, for $b_{a,k}(\epsilon) = (1 + (\max(0,k)+1)\alpha_a\epsilon)^{-1}$,

$$D_{a,k}(\epsilon) = \inf_{\theta:\mu(\theta)>\mu_1-\epsilon} \mathrm{KL}(\mathrm{Pa}(\kappa_a+\epsilon, \alpha_a b_{a,k}(\epsilon)), \mathrm{Pa}(\theta)),$$

which satisfies $\lim_{\epsilon\to 0} D_{a,k}(\epsilon) = \mathrm{KL}_{\inf}(a)$ for any fixed $k \in \mathbb{Z}$.

By letting $\epsilon = o(1)$ in Theorem 2, we see that STS with $k \in \mathbb{Z}_{\geq 2}$ satisfies

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \leq \sum_{a=2}^{K} \frac{\Delta_a}{\mathrm{KL}_{\inf}(a)},$$

which shows the asymptotic optimality of STS in terms of the lower bound in (3).

Next, we show that STS with $k \in \mathbb{Z}_{\leq 1}$ cannot achieve the asymptotic bound in the theorem below. Following the proofs for Gaussian bandits (Honda & Takemura, 2014), we consider two-armed bandit problems where the full information on the suboptimal arm is given to simplify the analysis. We further assume that two arms have the same scale parameter $\kappa_1 = \kappa_2$. A similar result for the case $\kappa_1 < \kappa_2$ can be found in Appendix D.

Theorem 3. *Consider a two-armed bandit problem where $\kappa_1 = \kappa_2$ and $1 < \alpha_1 < \alpha_2$. When $\tilde{\alpha}_1(t)$ and $\tilde{\kappa}_1(t)$ are sampled from the posteriors in (8) and (9) with prior parameter $k \in \mathbb{Z}_{\leq 1}$, respectively and $\tilde{\mu}_2(t) = \mu_2$ holds, there exists a constant $C(\alpha_1, \alpha_2) > 0$ satisfying*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geq C(\alpha_1, \alpha_2),$$

where $C(\alpha_1, \alpha_2) > \frac{\Delta_2}{\mathrm{KL}_{\inf}(2)}$ holds for some instances. In particular, for $k \in \mathbb{Z}_{\leq 0}$, there exists $C'(\alpha_1, \alpha_2) > 0$ satisfying

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\sqrt{T}} \geq C'(\alpha_1, \alpha_2).$$

From Theorems 2 and 3, we find that the prior should be conservative to some extent when one considers maximizing rewards in *expectation*.

Although STS with the Jeffreys prior ($k = 0$) and reference prior ($k = 1$) were shown to be suboptimal, we show that a modified algorithm, STS-T, can achieve the optimal regret bound with $k \in \mathbb{Z}_{\geq 0}$.

Theorem 4. *With the same notation as Theorem 2, the expected regret of* STS-T *with $k \in \mathbb{Z}_{\geq 0}$ is bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{a=2}^{K} \frac{\Delta_a \log T}{D_{a,k}(\epsilon)} + \mathcal{O}(\epsilon^{-m}),$$

where $m = \max(2, 3-k)$.

From Theorems 2 and 4, we have two choices to achieve the lower bound in (3): use either the conservative priors with MLEs or moderately optimistic priors with truncated samples. Since initialization steps require playing every arm $\max(2, k+1)$ times, if the number of arms K is large, the Jeffreys priors or the reference prior with the truncated estimator would be a better choice. On the other hand, if the model can contain arms with large α, where the truncation might be problematic for small n, it would be better to use STS with conservative priors.

5. Experiments

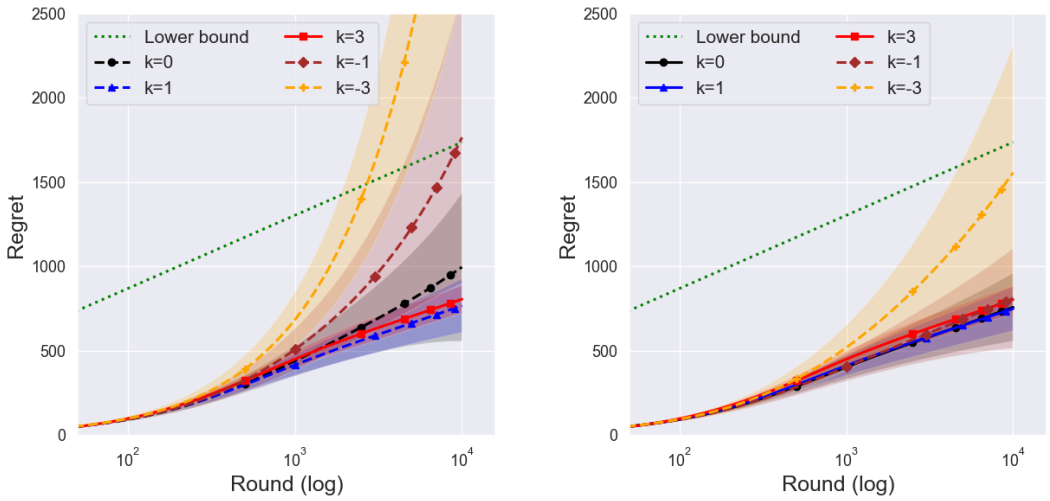
In this section, we present numerical results to demonstrate the performance of STS and STS-T, which supports our theoretical analysis. We consider the 4-armed bandit model $\boldsymbol{\theta}_4$ with parameters given in Table 1 as an example where suboptimal arms have smaller, equal, and larger κ compared with the optimal arm. $\boldsymbol{\theta}_4$ has $\boldsymbol{\mu} = (4.55, 3.2, 2.74, 3)$ and infinite variance. Further experimental results can be found in Appendix H.

Table 1. 4-armed bandit model $\boldsymbol{\theta}_4$.

	ARM 1	ARM 2	ARM 3	ARM 4
κ	1.3	1.2	1.3	1.5
α	1.4	1.6	1.9	2.0

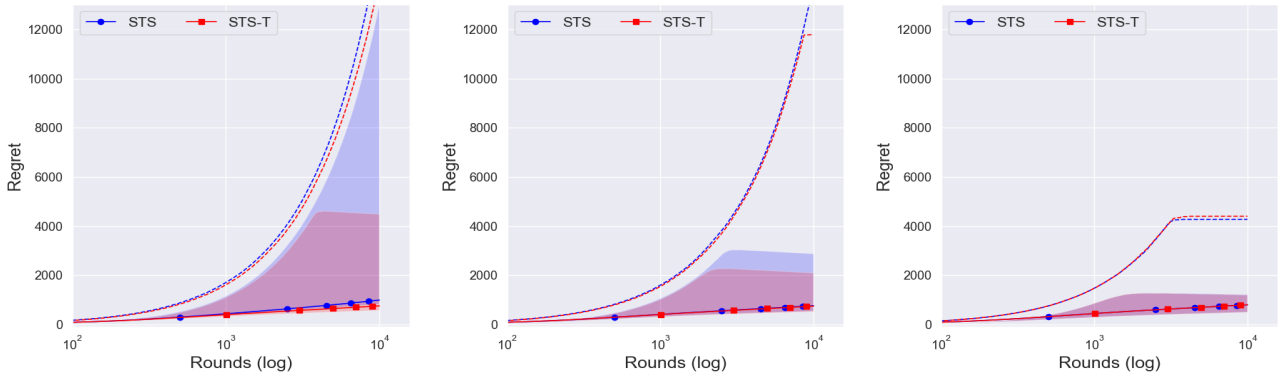
Figure 1 shows the cumulative regret for the proposed policies with various choices of parameters k on the prior. The solid lines denote the averaged cumulative regret over 100,000 independent runs of priors that can achieve the optimal lower bound in (3), whereas the dashed lines denote that of priors that cannot. The green dotted line denotes the problem-dependent lower bound and shaded regions denote a quarter standard deviation.

In Figures 2 and 3, we investigate the difference between STS and STS-T with the same k. The solid lines denote the averaged cumulative regret over 100,000 independent runs. The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of regret.



(a) Cumulative regret of STS with various k (b) Cumulative regret of STS-T with various k

Figure 1. The solid lines denote the averaged cumulative regret over 100,000 independent runs of priors that can achieve the optimal lower bound in (3). The dashed lines denote that of priors that cannot achieve the optimal lower bound in (3). The shaded regions show a quarter standard deviation. The green dotted line denotes the problem-dependent lower bound based on Lemma 1.



(a) The Jeffreys prior $k = 0$ (b) The reference prior $k = 1$ (c) Prior with $k = 3$

Figure 2. The solid lines denote an averaged regret over independent 100,000 runs. The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of the regret, respectively.

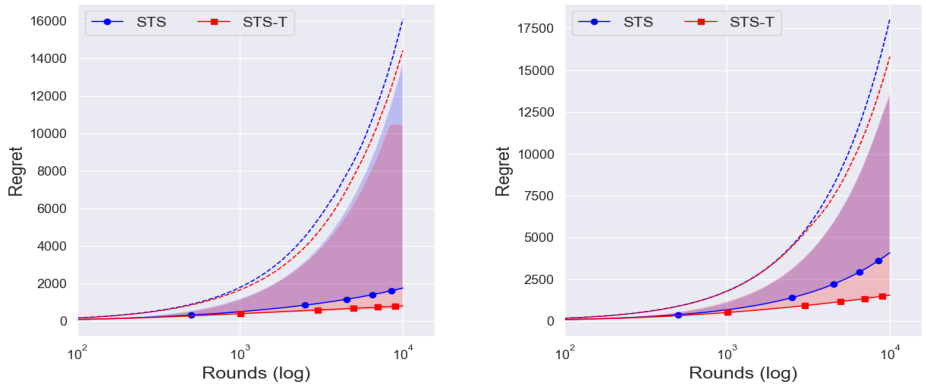
The Jeffreys prior ($k = 0$) In Figure 1(a), the Jeffreys prior seems to have a larger order of the regret compared with priors $k = 1, 3$, which performed the best in this setting. As Theorem 4 states, its performance improves under STS-T, which shows a similar performance to that of $k = 1, 3$.

Figure 2(a) illustrates the possible reason for the improvements, where the central 99% interval of the regret noticeably shrank under STS-T. Since the suboptimality of STS with the Jeffreys prior ($k = 0$) is due to an extreme case that induces a polynomial regret with small probability, this kind of shrink contributes to decreasing the expected regret of STS-T with the Jeffreys prior.

The reference prior ($k = 1$) The reference prior showed a similar performance to the asymptotically optimal prior $k = 3$, although it was shown to be suboptimal for some instances under STS in Theorem 3. Similarly to the Jeffreys prior ($k = 0$), the reference prior ($k = 1$) under STS-T has a smaller central 99% interval of the regret than that under STS as shown in Figure 2(b), although its decrement is comparably smaller than that of the Jeffreys prior. This would imply that the reference prior is more conservative than the Jeffreys prior.

The conservative prior ($k = 3$) Interestingly, Figure 2(c) showed that a truncated procedure does not affect the central 99% interval of the regret and even degrade the performance in upper 0.05%. Notice that the upper 0.05% of the regret of $k = 3$ is much lower than that of $k = 0, 1$, which shows the stability of the conservative prior in Figure 2.

Since a truncation procedure was adopted to prevent an



(a) Prior with $k = -1$ (b) Prior with $k = -3$

Figure 3. The solid lines denote an averaged regret over independent 100,000 runs. The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of the regret, respectively.

extreme case that was a problem for $k \in \mathbb{Z}_{\leq 1}$, it is natural to see that there is no difference between STS and STS-T with $k = 3$. This would imply that $k = 3$ is sufficiently conservative, and so the truncated procedure does not affect the overall performance.

Optimistic priors ($k < 0$) In Figure 1(a), one can see that the averaged regret of $k = -1$ and $k = -3$ increases much faster than that of $k = 0, 1, 3$ under the STS policy, which illustrates the suboptimality of STS with priors $k \in \mathbb{Z}_{<0}$.

As the optimistic priors ($k < 0$) showed better performance under STS-T in Figure 1, we can check the effectiveness of a truncation procedure in the posterior sampling with optimistic priors. However, as Figures 3(a) and 3(b) illustrate, there is no big difference in the central 99% interval of the regret between STS and STS-T with $k = -1, -3$, which might imply that a prior with $k \in \mathbb{Z}_{<0}$ is too optimistic. Therefore, we might need to use a more conservative truncation procedure such as the one using $\bar{\alpha}_a(n) = \max(\sqrt{n}, \hat{\alpha}_a(n))$ or $\max(\log n, \hat{\alpha}_a(n))$, which would induce a larger regret in the finite time horizon.

6. Proof outline of optimal results

In this section, we provide the proof outline of Theorem 2 and Theorem 4, whose detailed proof is given in Appendix C. Note that the proof of Theorem 3 is postponed to Appendix D.

Let us first consider good events on MLEs defined by

$$\begin{aligned}\mathcal{K}_{a,n}(\epsilon) &:= \{\hat{\kappa}_a(n) \in [\kappa_a, \kappa_a + \epsilon]\} \\ \mathcal{A}_{a,n}(\epsilon) &:= \{\hat{\alpha}_a(n) \in [\alpha_a - \epsilon_{a,l}(\epsilon), \alpha_a + \epsilon_{a,u}(\epsilon)]\} \\ \mathcal{E}_{a,n}(\epsilon) &:= \mathcal{K}_{a,n}(\epsilon) \cap \mathcal{A}_{a,n}(\epsilon),\end{aligned}$$

where $n \in \mathbb{N}$ and

$$\epsilon_{a,l}(\epsilon) = \frac{\epsilon\alpha_a^2}{1 + \epsilon\alpha_a}, \quad \epsilon_{a,u}(\epsilon) = \frac{\epsilon\alpha_a^2(\kappa_a + 1)}{\kappa_a - \epsilon\alpha_a(\kappa_a + 1)}. \quad (13)$$

Note that $\bar{\alpha}_a(n) = \hat{\alpha}_a(n)$ holds on $\mathcal{A}_{a,n}(\epsilon)$ for any $n \geq \alpha_a + 1$. Here, we set ε_a to satisfy $\hat{\mu}_a \in [\mu_a - \delta_a, \mu_a + \delta_a]$ on $\mathcal{E}_a(\epsilon)$ for any $\epsilon \leq \varepsilon_a$. Define an event on the currently optimal sample, $\tilde{\mu}^*(t) = \max_{a \in [K]} \tilde{\mu}_a(t)$,

$$\mathcal{M}_\epsilon(t) := \{\tilde{\mu}^*(t) \geq \mu_1 - \epsilon\}.$$

Then, the expected regret at round T can be decomposed as follows:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{j(t)}\right] \\ &= \sum_{a=2}^{K} \Delta_a \left(\bar{n} + \sum_{t=\bar{n}K+1}^{T} \mathbb{E}[\mathbb{1}[j(t) = a]]\right) \\ &\leq \Delta_2 \sum_{t=\bar{n}K+1}^{T} \Big(\mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)]\right] \\ &\quad + \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)]\right]\Big) + \sum_{a=2}^{K} \Delta_a \Big\{\bar{n} \\ &\quad + \sum_{t=\bar{n}K+1}^{T} \Big(\mathbb{E}\left[\mathbb{1}[j(t) = a, \mathcal{M}_\epsilon(t), \mathcal{E}_{a,N_a(t)}(\epsilon)]\right] \\ &\qquad \mathbb{E}\left[\mathbb{1}[j(t) = a, \mathcal{M}_\epsilon(t), \mathcal{E}_{a,N_a(t)}^c(\epsilon)]\right]\Big)\Big\},\end{aligned}$$

where $\mathcal{E}^c$ denotes the complementary set of $\mathcal{E}$. Lemmas 5–8 complete the proof of Theorems 2 and 4, whose proofs are given in Appendix C.

Lemma 5. *Under* STS *with* $k \in \mathbb{Z}_{\geq 2}$,

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)]\right] \leq \mathcal{O}(\epsilon^{-2}).$$

and under STS-T *with* $k \in \mathbb{Z}_{\geq 0}$,

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)]\right] \leq \mathcal{O}(\epsilon^{-m}),$$

where $m = \max(2, 3-k)$.

Although Lemma 6 contributes to the main term of the regret, the proof of Lemma 5 is the main difficulty in the regret analysis. We found that our analysis does not result in a finite upper bound for STS with $k \in \mathbb{Z}_{<2}$ and designed STS-T to solve such problems.

Lemma 6. *Under* STS *and* STS-T *with* $k \in \mathbb{Z}$*, it holds that for any* $a \in [K]$

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}[\mathbb{1}[j(t)=a, \mathcal{M}_\epsilon(t), \mathcal{E}_{a,N_a(t)}(\epsilon)]] \leq \max(0,k) + 1 + \frac{1}{\alpha_a \epsilon}\mathbb{1}[k>0] + \frac{\log T}{D_{a,k}(\epsilon)}.$$

where $D_{a,k}(\epsilon) > 0$ *is a finite problem-deterministic constant satisfying* $\lim_{\epsilon\to 0} D_{a,k}(\epsilon) = \mathrm{KL}_{\inf}(a)$.

Since large k yields a more conservative policy and requires additional initial plays of every arm, large k might induce larger regret for a finite time horizon T, which corresponds to the component of the regret discussed in Lemma 6. Thus, this lemma would imply that the policy has to be conservative to some extent, and being overly conservative would induce larger regrets in a finite time.

Lemma 7. *Under* STS *and* STS-T *with* $k \in \mathbb{Z}_{\geq 0}$*,*

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t)\neq 1, \mathcal{K}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)]\right] \leq \mathcal{O}(\epsilon^{-1}).$$

The key to Lemma 7 is to convert the term on $\tilde{\mu}_1(t)$, $\mathcal{M}_\epsilon(t)$, to a term on $\tilde{\alpha}_1(t)$. Since $\mu(\kappa,\alpha) = \infty$ holds for $\alpha \leq 1$, $\tilde{\mu}_1 = \mu(\tilde{\kappa}_1, \tilde{\alpha}_1)$ becomes infinity regardless of the value of $\tilde{\kappa}_1$ if $\tilde{\alpha}_1 \leq 1$ holds, which implies $\mathbb{P}[\mathcal{M}_\epsilon^c(t), \tilde{\alpha}_1(t) \leq 1] = 0$. Therefore, it is enough to consider the case where $\tilde{\alpha}_1(t) > 1$ holds to prove Lemma 7. Although density functions of $\tilde{\alpha}_1$ under STS and STS-T are different, conditional CDFs of $\tilde{\kappa}_1$ given $\alpha_1 = \tilde{\alpha}_1$ are the same, which is given in (9) as

$$\mathbb{P}[\tilde{\kappa}_1 \leq x | \mathcal{F}_t, \tilde{\alpha}_1 = \alpha_1] = \left(\frac{x}{\hat{\kappa}_1(N_1(t))}\right)^{\tilde{\alpha}_1 N_1(t)}.$$

Therefore, for sufficiently large $N_1(t)$ and $\tilde{\alpha}_1(t) > 1$, $\tilde{\kappa}_1(t)$ will concentrate on $\hat{\kappa}_1(N_1(t))$ with high probability, which is close to its true value κ_1 under the event $\{\mathcal{K}_{1,N_1(t)}(\epsilon)\}$. Thus, $\tilde{\mu}_1 = \frac{\tilde{\kappa}_1\tilde{\alpha}_1}{\tilde{\alpha}_1 - 1} \geq \frac{\kappa_1\tilde{\alpha}_1}{\tilde{\alpha}_1-1} = \mu(\kappa_1, \tilde{\alpha}_1)$ holds with high probability, which implies that $\mathbb{P}[\mathcal{K}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t) | \mathcal{F}_t]$ can be roughly bounded by $\mathbb{P}[\mathcal{K}_{1,N_1(t)}(\epsilon), \tilde{\alpha}_1(t) \geq c | \mathcal{F}_t]$ for some problem-dependent constants $c > 1$. Since $\mathcal{K}_1$ is deterministic given $\mathcal{F}_t$, we have

$$\mathbb{P}[\mathcal{K}_{1,N_1(t)}(\epsilon), \tilde{\alpha}_1(t) \geq c | \mathcal{F}_t] = \mathbb{1}[\mathcal{K}_{1,N_1(t)}(\epsilon)]\mathbb{P}[\tilde{\alpha}_1(t) \geq c | \mathcal{F}_t],$$

which implies $\tilde{\mu}_1(t)$ is mainly determined by the value of $\tilde{\alpha}_1(t)$ under the event $\{\mathcal{K}_{1,N_1(t)}(\epsilon)\}$ for both policies. In such cases, STS and STS-T behave like TS in the Pareto distribution with a known scale parameter, where $\tilde{\mu}_1(t) := \mu(\kappa_1, \tilde{\alpha}_1(t))$ for $t \in \mathbb{N}$. Here, the Pareto distribution with the known scale parameter belongs to the one-dimensional exponential family, where Korda et al. (2013) showed the optimality of TS with the Jeffreys prior. Since the posterior of α under the Jeffreys prior is given as the Erlang distribution with shape $N_1(t)+1$ in the one-parameter Pareto model, we can apply the results by Korda et al. (2013) to prove Lemma 7 by using some properties of the Erlang distribution such as (12).

Lemma 8. *Under* STS *and* STS-T *with* $k \in \mathbb{Z}$*, it holds that for any* $a \neq 1$

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t)=a, \mathcal{E}^c_{a,N_a(t)}(\epsilon)]\right] \leq \mathcal{O}\left(\epsilon^{-2}\right).$$

Lemma 8 controls the regret induced when estimators of the played arm are not close to their true parameters, which is not difficult to analyze as in the usual analysis of TS. In fact, the proof of this lemma is straightforward since the upper bounds of $\mathbb{P}[\mathcal{K}_a^c]$ and $\mathbb{P}[\mathcal{K}_a, \mathcal{A}_a^c]$ can be easily derived based on the distributions of $\hat{\kappa}_a$ and $\hat{\alpha}_a$ in (2).

7. Conclusion

We considered the MAB problems under the Pareto distribution that has a heavy tail and follows the power-law. While most previous research on TS has focused on one-dimensional or light-tailed distributions, we focused on the Pareto distribution characterized by unknown scale and shape parameters. By sequentially sampling parameters via their marginalized and conditional posterior distributions, we can realize an efficient sampling procedure. We showed that TS with the appropriate choice of priors achieves a problem-dependent optimal regret bound in such a setting for the first time. Although the Jeffreys prior and the reference prior are shown to be suboptimal under the direct implementation of TS, we showed that they could achieve the optimal regret bound if we add a truncation procedure. Experimental results support our theoretical results, which show the optimality of conservative priors and the effectiveness of the truncation procedure for the Jeffreys prior and the reference prior.

Acknowledgements

JL was supported by JST SPRING, Grant Number JPMJSP2108. JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan. CC and MS were supported by the Institute for AI and Beyond, UTokyo.

References

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 39–1. PMLR, 2012.

Agrawal, S. and Goyal, N. Near-optimal regret bounds for Thompson sampling. *Journal of the Association for Computing Machinery*, 64(5):1–24, 2017.

Agrawal, S., Juneja, S. K., and Koolen, W. M. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pp. 26–62. PMLR, 2021.

Arnold, B. C. Pareto and generalized Pareto distributions. In *Modeling income distributions and Lorenz curves*, pp. 119–145. Springer, 2008.

Baudry, D., Gautron, R., Kaufmann, E., and Maillard, O. Optimal Thompson sampling strategies for support-aware CVaR bandits. In *International Conference on Machine Learning*, pp. 716–726. PMLR, 2021.

Berger, J. O. and Bernardo, J. M. On the development of the reference prior method. *Bayesian Statistics*, 4(4):35–60, 1992.

Bernardo, J. M. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979.

Bubeck, S. and Liu, C.-Y. Prior-free and prior-dependent regret bounds for Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 26, pp. 638–646, 2013.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013.

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

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. Wiley-Interscience, 2nd edition, 2006.

Datta, G. S. On priors providing frequentist validity of Bayesian inference for multiple parametric functions. *Biometrika*, 83(2):287–298, 1996.

Datta, G. S. and Ghosh, M. Some remarks on noninformative priors. *Journal of the American Statistical Association*, 90(432):1357–1363, 1995.

Datta, G. S. and Ghosh, M. On the invariance of noninformative priors. *The Annals of Statistics*, 24(1):141–159, 1996.

Datta, G. S. and Sweeting, T. J. Probability matching priors. *Handbook of Statistics*, 25:91–114, 2005.

DiCiccio, T. J., Kuffner, T. A., and Young, G. A. A simple analysis of the exact probability matching prior in the location-scale model. *The American Statistician*, 71(4): 302–304, 2017.

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Ghosh, M. Objective priors: An introduction for frequentists. *Statistical Science*, 26(2):187–202, 2011.

Honda, J. and Takemura, A. Optimality of Thompson sampling for Gaussian bandits depends on priors. In *International Conference on Artificial Intelligence and Statistics*, volume 33, pp. 375–383. PMLR, 2014.

Jeffreys, H. *The theory of probability*. OUP Oxford, 1998.

Kaufmann, E. On bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2):842–865, 2018.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite time analysis. In *International Conference on Algorithmic Learning Theory*, volume 7568, pp. 199–213. Springer, 2012.

Kim, D.-H., Kang, S.-G., and Lee, W.-D. Noninformative priors for Pareto distribution. *Journal of the Korean Data and Information Science Society*, 20(6):1213–1223, 2009.

Korda, N., Kaufmann, E., and Munos, R. Thompson sampling for 1-dimensional exponential family bandits. In *International Conference on Neural Information Processing Systems*, volume 26, pp. 1448–1456. Curran Associates, Inc., 2013.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Li, M., Sun, H., and Peng, L. Fisher–Rao geometry and Jeffreys prior for Pareto distribution. *Communications in Statistics-Theory and Methods*, 51(6):1895–1910, 2022.

Lomax, K. S. Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49(268):847–852, 1954.

Malik, H. J. Estimation of the parameters of the Pareto distribution. *Metrika*, 15(1):126–132, 1970.

Ménard, P. and Garivier, A. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237. PMLR, 2017.

Mukerjee, R. and Ghosh, M. Second-order probability matching priors. *Biometrika*, 84(4):970–975, 1997.

Riou, C. and Honda, J. Bandit algorithms based on Thompson sampling for bounded reward distributions. In *International Conference on Algorithmic Learning Theory*, pp. 777–826. PMLR, 2020.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Robert, C. P. et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2nd edition, 2007.

Russo, D. and Van Roy, B. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Schervish, M. J. *Theory of statistics*. Springer Science & Business Media, 2012.

Simon, M. K. and Divsalar, D. Some new twists to problems involving the Gaussian probability integral. *IEEE Transactions on Communications*, 46(2):200–210, 1998.

Stein, C. An example of wide discrepancy between fiducial and confidence intervals. *The Annals of Mathematical Statistics*, 30(4):877–880, 1959.

Sun, F., Cao, Y., Zhang, S., and Sun, H. The Bayesian inference of Pareto models based on information geometry. *Entropy*, 23(1):45, 2020.

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

Tibshirani, R. Noninformative priors for one parameter of many. *Biometrika*, 76(3):604–608, 1989.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wallace, D. L. Bounds on normal approximations to student's and the chi-square distributions. *The Annals of Mathematical Statistics*, 30(4):1121–1130, 1959.

Welch, B. and Peers, H. On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25 (2):318–329, 1963.

A. Notations

Tables 2–5 summarize the symbols used in this paper.

Table 2. Notations for the bandit problem.

Symbol	Meaning
K	the number of arms.
T	time horizon.
$j(t)$	the index of the played arm at round t.
$k \in \mathbb{Z}$	prior parameter, see Section 3 for details.
$\bar{n} = \max(2, k+1)$	initial plays to avoid improper posteriors.
$N_a(t)$	the number of playing arm a until round t.
$r_{a,n}$	n-th reward generated from the arm a.
$\mu(\theta)$	the expected value of the random variable following $\mathrm{Pa}(\theta)$.
$\mu_a = \mu(\theta_a)$	the expected rewards of arm a.
Δ_a	sub-optimality gap of arm a.
$\delta_a = \frac{\Delta_a}{2}$ for $a \neq 1$	a half of sub-optimality gap of arm a.
$\delta_1 := \min_{a \neq 1} \delta_a$	defined as the minimum of sub-optimality gap.
$\mathcal{F}_t = (j(s), r_{j(s),N_{j(s)}})_{s=1}^{t-1}$	the history until round t.
$\mathbb{P}_t[\cdot] := \mathbb{P}[\cdot \vert \mathcal{F}_t]$	conditional probability given $\mathcal{F}_t$.
$g_a(c, \alpha)$	KL-divergence from $\mathrm{Pa}\left(\frac{\kappa_a}{c}, \alpha\right)$ to $\mathrm{Pa}(\kappa_a, \alpha_a)$ for $c \geq 1$.
$h_a(c, \boldsymbol{\mu}) = h_a(c)$	the upper bound of α satisfying $\mu\left(\frac{\kappa_a}{c}, \alpha\right) \geq \mu_1$ for $c \geq 1$.

Table 3. Notations for probability distributions and estimators

Symbol	Meaning
$\mathrm{Pa}(\kappa, \alpha)$	Pareto distribution with the scale $\kappa > 0$ and shape $\alpha > 0$.
$f^{\mathrm{Pa}}_{\kappa,\alpha}(x)$	density function of $\mathrm{Pa}(\kappa, \alpha)$.
$\mathrm{Erlang}(s, \beta)$	Erlang distribution with the shape $s > 0$ and rate $\beta > 0$.
$f^{\mathrm{Er}}_{s,\beta}(x)$	density function of $\mathrm{Erlang}(s, \beta)$.
$F^{\mathrm{Er}}_{s,\beta}(x)$	CDF of $\mathrm{Erlang}(s, \beta)$ evaluated at $x > 0$.
$\mathrm{IG}(s, \beta)$	Inverse Gamma distribution with shape $s > 0$ and scale $\beta > 0$.
$F_n(x)$	CDF of the chi-squared distribution with n degree of freedom.
$\Gamma(s)$	Gamma function.
$\gamma(s, x)$	the lower incomplete gamma function.
$\Gamma(s, x)$	the upper incomplete gamma function.
$\hat{\kappa}_a(n)$, $\hat{\alpha}_a(n)$	MLEs of the scale and shape parameter of arm a after n observations, defined in (2).
$\tilde{\kappa}_a(t)$, $\tilde{\alpha}_a(t)$	sampled parameters at round t from posterior distribution in (9) and (8).
$\bar{\alpha}_a(n) = \max(\hat{\alpha}_a(n), n)$	truncated estimator of the shape parameter.
$\alpha_{a,n}$	a temporary notation that can be replaced by both $\hat{\alpha}_a(n)$ (STS) and $\bar{\alpha}_a(n)$ (STS-T).
$\hat{\mu}_a(n) = \mu(\hat{\kappa}_a(n), \hat{\alpha}_a(n))$	computed mean rewards by the MLEs after n observation.
$\tilde{\mu}_a(t) = \mu(\tilde{\kappa}_a(t), \tilde{\alpha}_a(t))$	computed mean rewards by sampled parameters $\tilde{\kappa}_a(t)$ and $\tilde{\alpha}_a(t)$ at round t.
$\theta_a = (\kappa_a, \alpha_a)$	tuple of true parameters of arm $a \in [K]$.
$\hat{\theta}_{a,n} = (\hat{\kappa}_a(n), \hat{\alpha}_a(n))$	tuple of MLEs of arm a after n observations.
$\bar{\theta}_{a,n} = (\hat{\kappa}_a(n), \bar{\alpha}_a(n))$	tuple of estimators with a truncation procedure of arm a after n observations.
$\theta_{a,n} = (\hat{\kappa}_a(n), \alpha_{a,n})$	a temporary notation that can be replaced by both $\hat{\theta}_{a,n}$ (STS) and $\bar{\theta}_{a,n}$ (STS-T).

Table 4. Notations for the regret analysis

Symbol	Meaning
$D_{a,k}(\epsilon)$	a function contributes to the main term of regret analysis defined in (44).
$\mathcal{K}_{a,n}(\epsilon)$	an event where MLE of κ is close to its true value at round t after n observations.
$\mathcal{A}_{a,n}(\epsilon)$	an event where MLE of α is close to its true value at round t after n observations.
$\mathcal{E}_{a,n}(\epsilon)$	intersection of $\mathcal{K}_{a,n}(\epsilon)$ and $\mathcal{A}_{a,n}(\epsilon)$.
$\mathcal{M}_{\epsilon}(t)$	an event where sampled mean of the optimal arm is close to its true mean reward at round t.
$p_n(x\|\theta_{1,n})$	probability of $\{\tilde{\mu}_1(t) \leq \mu_1 - x\}$ after n observation of arm 1 given $\theta_{1,n}$.
$G_k(x; n)$	another expression of the CDF of the Erlang distribution in (21).

Table 5. Notations for (deterministic) constants

Symbol	Meaning
ε_a	problem-dependent constants to satisfy $\hat{\mu}_a(n) \in [\mu_a - \delta_a, \mu_a + \delta_a]$ on $\mathcal{E}_{a,n}(\epsilon)$ for any $\epsilon \leq \varepsilon_a$.
$\epsilon_{a,l}(\epsilon)$, $\epsilon_{a,l}(\epsilon)$	constants to control a deviation of $\hat{\alpha}_a(\epsilon)$ under the event $\mathcal{A}_{a,n}(\epsilon)$.
$\rho_{\theta_1}(\epsilon), \bar{\rho} = \rho_{\theta_1}(\epsilon/2)$	a difference from its true shape parameter α_1 to satisfy $\mu(\kappa_a, \alpha + \rho_{\mu_1}(\epsilon)) \geq \mu_1 - \frac{\epsilon}{2}$.
$C_1(\mu_1, \epsilon, n) = C_{1,n}$	a constant smaller than 1 in (21).
$C_2(\mu_1, \epsilon, k)$	an uniform bound of $p_n(\epsilon\|\cdot)$ under $\mathcal{K}_{1,n}^c(\epsilon) \cap \mathcal{A}_{1,n}(\epsilon/2)$.
$c_{\mu_1}(\epsilon)$	a small constant with $\mathcal{O}(\epsilon^{-2})$.

B. Proof of Lemma 1

Lemma 1. *For any arm $a \neq 1$, it holds that*

$$\mathrm{KL}_{\inf}(a) = \log\left(\alpha_a \frac{\mu_1 - \kappa_a}{\mu_1}\right) + \frac{1}{\alpha_a}\frac{\mu_1}{\mu_1 - \kappa_a} - 1.$$

Proof. Recall the definition

$$\mathrm{KL}_{\inf}(a) = \mathrm{KL}_{\inf}(\mathrm{Pa}(\theta_a), \mathrm{Pa}(\theta)) := \inf_{\theta \in \Theta_a} \log\frac{\alpha_a}{\alpha} + \alpha \log\frac{\kappa_a}{\kappa} + \frac{\alpha}{\alpha_a} - 1,$$

where $\theta = (\kappa, \alpha)$ and Θ_a defined in (4).

Here, we consider the partition of Θ_a,

$$\begin{aligned}
\Theta_a^{(1)} &= \{(\kappa, \alpha) \in (0, \kappa_a] \times (0, 1] : \mu(\kappa, \alpha) > \mu_1\} = (0, \kappa_a] \times (0, 1] \\
\Theta_a^{(2)} &= \left\{(\kappa, \alpha) \in (0, \kappa_a] \times (1, \infty) : \mu(\kappa, \alpha) = \frac{\kappa\alpha}{\alpha - 1} > \mu_1\right\},
\end{aligned} \tag{14}$$

where $\Theta_a^{(1)} \cup \Theta_a^{(2)} = \Theta_a$. Therefore, it holds that

$$\mathrm{KL}_{\inf}(a) = \min\left(\inf_{\theta \in \Theta_a^{(1)}} \log\frac{\alpha_a}{\alpha} + \alpha\log\frac{\kappa_a}{\kappa} + \frac{\alpha}{\alpha_a} - 1, \inf_{\theta \in \Theta_a^{(2)}} \log\frac{\alpha_a}{\alpha} + \alpha\log\frac{\kappa_a}{\kappa} + \frac{\alpha}{\alpha_a} - 1\right).$$

For $(\kappa, \alpha) \in \Theta_a^{(1)}$, $\mu(\kappa, \alpha) = \infty$ holds regardless of κ. Therefore, we obtain

$$\begin{aligned}
\inf_{\theta \in \Theta_a^{(1)}} \log\frac{\alpha_a}{\alpha} + \alpha\log\frac{\kappa_a}{\kappa} + \frac{\alpha}{\alpha_a} - 1 &= \inf_{\alpha \in (0,1]} \log\frac{\alpha_a}{\alpha} + \frac{\alpha}{\alpha_a} - 1 \\
&= \log\alpha_a + \frac{1}{\alpha_a} - 1,
\end{aligned}$$

where the last equality holds since $\log x + \frac{1}{x} - 1$ is an increasing function for $x \geq 1$.

Let $\frac{\kappa_a}{\kappa} = c \geq 1$ to make KL divergence from $\mathrm{Pa}(\theta_a)$ to $\mathrm{Pa}(\kappa, \alpha)$ be well-defined. From its definition of $\Theta_a^{(2)}$ in (14), any $\theta = (\kappa, \alpha) \in \Theta_a^{(2)}$ satisfies $\frac{\kappa\alpha}{\alpha - 1} \geq \mu_1$, i.e.,

$$\frac{\kappa_a \alpha}{c(\alpha - 1)} \geq \mu_1 \Leftrightarrow \alpha \leq \frac{c\mu_1}{c\mu_1 - \kappa_a} =: h_a(c, \boldsymbol{\mu}) = h_a(c).$$

Note that it holds that

$$h_a(1) = \frac{\mu_1}{\mu_1 - \kappa_a} \leq \frac{\mu_a}{\mu_a - \kappa_a} = \alpha_a$$

since $\frac{x}{x-y}$ is decreasing with respect to $x \geq y$. Then, we can rewrite the infimum of KL divergence as

$$\mathrm{KL}_{\inf}(a) = \min\left(\log \alpha_a + \frac{1}{\alpha_a} - 1, \inf_{c \geq 1} \inf_{\alpha \leq h_a(c)} g_a(\alpha, c)\right),$$

where $g_a(\alpha, c) := \log \frac{\alpha_a}{\alpha} + \alpha \log c + \frac{\alpha}{\alpha_a} - 1$ satisfying

$$\frac{\partial g_a(\alpha, c)}{\partial \alpha} = \frac{1}{\alpha_a} + \log c - \frac{1}{\alpha}. \tag{15}$$

Then, the inner infimum can be obtained when $\alpha = \frac{\alpha_a}{1 + \alpha_a \log c}$ holds if $\frac{\alpha_a}{1 + \alpha_a \log c} < h_a(c)$, where $g_a\left(\frac{\alpha_a}{1 + \alpha_a \log c}, c\right) = \log(1 + \alpha_a \log c)$.

Let $c_a^* \geq 1$ be a deterministic constant such that

$$h_a(c_a^*) = \frac{c_a^* \mu_1}{c_a^* \mu_1 - \kappa_a} = \frac{\alpha_a}{1 + \alpha_a \log c_a^*} \Leftrightarrow (\mu_1 \alpha_a) c_a^* \log c_a^* + (\mu_1 - \alpha_a \mu_1) c_a^* = -\alpha_a \kappa_a \tag{16}$$

so that $h_a(c) \geq \frac{\alpha_a}{1 + \alpha_a \log c}$ holds for any $c \geq c_a^*$. Since the solution of $ax \log(x) + bx = -c$ is $\exp\left(W\left(-\frac{ce^{b/a}}{a}\right) - \frac{b}{a}\right)$ for principal branch of Lambert W function $W(\cdot)$, one can obtain c_a^* by solving the equality in (16), which is

$$c_a^* = \exp\left(W\left(-\frac{\kappa_a}{\mu_1} e^{\frac{1}{\alpha_a} - 1}\right) + 1 - \frac{1}{\alpha_a}\right). \tag{17}$$

Notice that $\frac{\kappa_a}{\mu_1} e^{\frac{1}{\alpha_a} - 1} \leq \frac{\kappa_a}{\mu_a} e^{\frac{1}{\alpha_a} - 1} \leq \left(1 - \frac{1}{\alpha_a}\right) e^{-\left(1 - \frac{1}{\alpha_a}\right)} \leq e^{-1}$ holds so that c_a^* is a real value. Here, we consider the principal branch to ensure $c_a^* \geq 1$ since the solution on other branches, $W_{-1}(\cdot)$, is less than 1, which is out of our interest. Let $A_a = 1 - \frac{1}{\alpha_a}$, which is positive as $\alpha_a > 1$ and $B_a = \frac{\kappa_a}{\mu_1} \leq \frac{\kappa_a}{\mu_a} = \frac{\alpha_a - 1}{\alpha_a} = A_a$. Then, we can rewrite c_a^* as

$$c_a^* = e^{A_a} e^{W(-B_a e^{-A_a})} = e^{A_a} e^{-A_a} \frac{-B_a}{W(-B_a e^{-A_a})}. \qquad \because e^{W(x)} = \frac{x}{W(x)}$$

Since the principal branch of Lambert W function, $W(x)$, is increasing for $x \geq -\frac{1}{e}$, we have

$$0 > W(-B_a e^{-A_a}) \geq W(-B_a e^{-B_a}) = -B_a,$$

which implies that $c_a^* \geq 1$. Therefore, the infimum of g_a can be written as

$$\begin{aligned}
\inf_{c \geq 1} \inf_{\alpha \leq h_a(c)} g_a(\alpha, c) &= \min\left(\inf_{c \in [1, c_a^*]} g_a(h_a(c), c), \inf_{c \geq c_a^*} \log(1 + \alpha_a \log c)\right) \\
&= \min\left(\inf_{c \in [1, c_a^*]} g_a(h_a(c), c), \log(1 + \alpha_a \log c_a^*)\right),
\end{aligned}$$

where we follow the convention that the infimum over the empty set is defined as infinity.

By substituting c_a^* in (17), we obtain

$$\log(1+\alpha_a \log c_a^*) = \log\left(\alpha_a + W\left(-\frac{\kappa_a}{\mu_1} e^{\frac{1}{\alpha_a}-1}\right)\right).$$

Let us consider the following inequalities:

$$\begin{aligned}
\log\left(\alpha_a + W\left(-\frac{\kappa_a}{\mu_1} e^{\frac{1}{\alpha_a}-1}\right)\right) &\geq \log\left(\alpha_a + W\left(-\frac{\kappa_a}{\mu_a} e^{\frac{1}{\alpha_a}-1}\right)\right) \\
&= \log\left(\alpha_a + W\left(\frac{1-\alpha_a}{\alpha_a} e^{\frac{1}{\alpha_a}-1}\right)\right) \\
&= \log\left(\alpha_a + \frac{1}{\alpha_a} - 1\right), \qquad (18)
\end{aligned}$$

where the first inequality holds since the principal branch of Lambert W function $W(x)$ is increasing and negative with respect to $x \in [-1/e, 0)$.

It remains to find the closed form of $\inf_{c\in[1,c_a^*]} g_a(h_a(c), c)$. From the definition of $h_a(c) = \frac{c\mu_1}{c\mu_1-\kappa_a}$, we have $h_a'(c) = -\frac{\mu_1\kappa_a}{(c\mu_1-\kappa_a)^2}$ and

$$\begin{aligned}
\frac{\partial g_a(h_a(c),c)}{\partial c} &= \frac{\partial}{\partial c}\left(\log\frac{\alpha_a}{h_a(c)} + h_a(c)\log c + \frac{h_a(c)}{\alpha_a} - 1\right) \qquad \because g_a(x,c) = \log\frac{\alpha_a}{x} + x\log c + \frac{x}{\alpha_a} - 1 \\
&= -\frac{h_a'(c)}{h_a(c)} + h_a'(c)\log c + \frac{h_a(c)}{c} + \frac{1}{\alpha_a}h_a'(c) \\
&= \frac{\kappa_a}{c(c\mu_1-\kappa_a)} - \frac{\mu_1\kappa_a}{(c\mu_1-\kappa_a)^2}\log c + \frac{\mu_1}{c\mu_1-\kappa_a} - \frac{\kappa_a\mu_1}{\alpha_a(c\mu_1-\kappa_a)^2} \\
&= \frac{\kappa_a}{c(c\mu_1-\kappa_a)} - \frac{\mu_1\kappa_a}{(c\mu_1-\kappa_a)^2}\log c + \mu_1\frac{c\mu_1-\kappa_a-\frac{\kappa_a}{\alpha_a}}{(c\mu_1-\kappa_a)^2}. \qquad (19)
\end{aligned}$$

Since the first term in (19) is positive for $c \geq 1$ and $\mu_1 \geq \mu_a > \kappa_a$, let us consider the last two terms for $c \in [1, c_a^*]$,

$$\begin{aligned}
-\frac{\mu_1\kappa_a}{(c\mu_1-\kappa_a)^2}\log c + \mu_1\frac{c\mu_1-\kappa_a-\frac{\kappa_a}{\alpha_a}}{(c\mu_1-\kappa_a)^2} &= \frac{\mu_1}{(c\mu_1-\kappa_a)^2}\left(c\mu_1-\kappa_a-\frac{\kappa_a}{\alpha_a}-\kappa_a\log c\right) \\
&= \frac{\mu_1}{(c\mu_1-\kappa_a)^2}\left(\mu_1-\kappa_a-\frac{\kappa_a}{\alpha_a}+(c-1)\mu_1-\kappa_a\log c\right) \\
&= \frac{\mu_1}{(c\mu_1-\kappa_a)^2}\left(\mu_1-\kappa_a-\frac{\kappa_a}{\alpha_a}+\mu_1\left(c-\frac{\kappa_a}{\mu_1}\log c-1\right)\right).
\end{aligned}$$

Here,

$$\mu_1-\kappa_a-\frac{\kappa_a}{\alpha_a} \geq \mu_a-\kappa_a-\frac{\kappa_a}{\alpha_a} = \frac{\kappa_a\alpha_a}{\alpha_a-1}-\kappa_a-\frac{\kappa_a}{\alpha_a} = \frac{\kappa_a}{\alpha_a(\alpha_a-1)} > 0,$$

and $c - \frac{\kappa_a}{\mu_1}\log c - 1$ is increasing with respect to c so that $c - \frac{\kappa_a}{\mu_1}\log c - 1 \geq 0$ for $c \geq 1$. Therefore, $\frac{\partial}{\partial c} g_a(h_a(c), c)$ is positive for $c \geq 1$, i.e., $g_a(h_a(c), c)$ is an increasing function with respect to $c \geq 1$.

Thus, we have for $c \in [1, c_a^*]$,

$$\inf_{c\in[1,c_a^*]} g_a(h_a(c),c) = g_a\left(h_a(1),1\right) = g_a\left(\frac{\mu_1}{\mu_1-\kappa_a},1\right) = \log\left(\alpha_a\frac{\mu_1-\kappa_a}{\mu_1}\right) + \frac{1}{\alpha_a}\frac{\mu_1}{\mu_1-\kappa_a} - 1.$$

Denote $X_a = \alpha_a\frac{\mu_1-\kappa_a}{\mu_1} \in [1, \alpha_a)$ where $X_a = 1$ happens only when $\mu_a = \mu_1$. Then, we have for $\alpha_a > 1$

$$\log(X_a) + \frac{1}{X_a} - 1 \leq \log\alpha_a + \frac{1}{\alpha_a} - 1 \leq \log\left(\alpha_a + \frac{1}{\alpha_a} - 1\right) \leq \log(1+\alpha_a\log c_a^*),$$

where the last inequality comes from the result in (18). Therefore, we have

$$\begin{aligned}\mathrm{KL}_{\inf}(a) &= \min\left(\log\alpha_a + \frac{1}{\alpha_a} - 1, \inf_{c\in[1,c_a^*]} g_a(h_a(c),c), \log(1+\alpha_a\log c_a^*)\right)\\ &= \log\left(\alpha_a\frac{\mu_1-\kappa_a}{\mu_1}\right) + \frac{1}{\alpha_a}\frac{\mu_1}{\mu_1-\kappa_a} - 1,\end{aligned}$$

which concludes the proof. □

C. Proofs of lemmas for Theorems 2 and 4

In this section, we provide the proof of lemmas for the main results.

To avoid redundancy, we use a temporary notation $\alpha_{a,n}$ when the same result holds for both $\hat{\alpha}_a(n)$ and $\bar{\alpha}_a(n)$. When $\alpha_{a,n}$ notation is used, one can replace it with either $\hat{\alpha}_a(n)$ or $\bar{\alpha}_a(n)$ depending on which policy we are considering. For example, it holds that

$$\alpha_{a,n} \le 1 \Leftrightarrow \begin{cases}\hat{\alpha}_a(n) \le 1 & \text{under STS policy,}\\ \bar{\alpha}_a(n) \le 1 & \text{under STS-T policy.}\end{cases}$$

Similarly, we use the notation $\theta_{a,n} := (\hat{\kappa}_a(n), \alpha_{a,n})$ when it can be replaced by both $\hat{\theta}_{a,n} = (\hat{\kappa}_a(n), \hat{\alpha}_a(n))$ and $\bar{\theta}_{a,n} = (\hat{\kappa}_a(n), \bar{\alpha}_a(n))$ for any arm $a \in [K]$ and $n \in \mathbb{N}$. Based on $\theta_{a,n}$ notation, we provide an inequality on the posterior probability that the sampled mean is smaller than a given value with proofs in Appendix C.5.

Lemma 9. *For any arm $a \in [K]$ and fixed $t \in \mathbb{N}$, let $N_a(t) = n$. For any positive $\xi \le \frac{y}{y-\kappa_a}$ and $k \in \mathbb{Z}$, it holds that*

$$\mathbb{1}[\hat{\kappa}_a(n) \le y]\mathbb{P}[\tilde{\mu}_a(t) \le y|\theta_{a,n}] \le \int_{\xi}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x + \left(\frac{y}{\mu((\kappa_a,\xi))}\right)^n \int_1^{\xi} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x,$$

where $f^{\mathrm{Er}}_{s,\beta}(\cdot)$ denotes the probability density function of the Erlang distribution with shape $s \in \mathbb{N}$ and rate $\beta > 0$.

Based on $\theta_{1,n}$ notation, we denote the probability of sample from the posterior distribution after n times playing is smaller than $\mu_1 - x$ by

$$p_n(x|\theta_{1,n}) := \mathbb{P}[\tilde{\mu}_1 \le \mu_1 - x|\hat{\kappa}_1(n), \alpha_{1,n}]. \tag{20}$$

Let $K(\epsilon) = (\kappa_1 + \epsilon, \mu_1 - \epsilon)$ be an open interval on $\mathbb{R}$. The Lemma 10 below shows the upper bound of p_n conditioned on $\theta_{1,n}$.

Lemma 10. *Given $\epsilon > 0$, define a positive problem-dependent constant $\rho = \rho_{\theta_1}(\epsilon) := \frac{\kappa_1\epsilon}{2(\mu_1-\epsilon/2-\kappa_1)(\mu_1-\kappa_1)}$. If $n \ge \bar{n} = \max(2, k+1)$, then for $k \in \mathbb{Z}_{\ge 0}$*

$$p_n(\epsilon|\theta_{1,n}) \le \begin{cases} e^{-n}, & \text{if } \hat{\kappa}_1(n) \ge \mu_1 - \epsilon,\\ h(\mu_1,\epsilon,n), & \text{if } \hat{\kappa}_1(n) \in K(\epsilon), \alpha_{1,n} \le \alpha_1 + \rho,\\ C_1(\mu_1,\epsilon,n)G_k(1/\alpha_{1,n};n) + 1 - G_k(1/\alpha_{1,n};n) & \text{if } \hat{\kappa}_1(n) \in K(\epsilon), \alpha_{1,n} \ge \alpha_1 + \rho,\end{cases}$$

where

$$\begin{aligned} h(\mu_1,\epsilon,n) &:= e^{-n\frac{3\epsilon}{4\mu_1}}\left(1 - \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}\right) + \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}\\ C_1(\mu_1,\epsilon,n) &:= \left(\frac{\mu_1-\epsilon}{\mu_1-\epsilon/2}\right)^n \le e^{-n\frac{\epsilon}{2\mu_1-\epsilon}} < 1\\ G_k(x;n) &:= F^{\mathrm{Er}}_{n-k,nx}(\alpha_1+\rho)\end{aligned} \tag{21}$$

for F^{Er} defined in (11). Here, $c_{\mu_1}(\epsilon) = \zeta - \log(1+\zeta) = \mathcal{O}(\epsilon^{-2})$ and $\zeta = \frac{\epsilon}{4\mu_1-2\epsilon} \in (0,1)$ are deterministic constants of μ_1 and ϵ.

Notice that $\mu((\kappa_1, \alpha_1+\rho)) = \mu_1 - \epsilon/2$ holds and there exists a problem-dependent constant $C_2(\mu_1,\epsilon,k) < 1$ such that for any $n \ge \bar{n} = \max(2, k+1)$ and $\epsilon > 0$

$$h(\mu_1,\epsilon,n) \le 1 - C_2(\mu_1,\epsilon,k). \tag{22}$$

C.1. Proof of Lemma 5

Let us start by stating a well-known fact that is utilized in the proof.

Fact 11. When $X \sim \mathrm{Erlang}(n, \beta)$ with rate parameter β, then $\frac{1}{X}$ follows the inverse gamma distribution with shape $n \in \mathbb{N}$ and scale $\beta \in \mathbb{R}_+$, i.e., $\frac{1}{X} \sim \mathrm{IG}(n, \beta)$.

Lemma 5. *Under* STS *with* $k \in \mathbb{Z}_{\geq 2}$*,*

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}^c_{1,N_1(t)}(\epsilon), \mathcal{M}^c_\epsilon(t)]\right] \leq \mathcal{O}(\epsilon^{-2}).$$

and under STS-T *with* $k \in \mathbb{Z}_{\geq 0}$*,*

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}^c_{1,N_1(t)}(\epsilon), \mathcal{M}^c_\epsilon(t)]\right] \leq \mathcal{O}(\epsilon^{-m}),$$

where $m = \max(2, 3-k)$.

Proof. Let us consider the following decomposition that holds under both STS and STS-T:

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

Notice that

$$m \leq \sum_{t=K\bar{n}+1}^{T} \mathbb{1}[j(t) \neq 1, \mathcal{K}^c_{1,N_1(t)}(\epsilon), \mathcal{M}^c_\epsilon(t), N_1(t) = n]$$

implies that $\tilde{\mu}_1(t) \leq \mu_1 - \epsilon$ occurred m times in a row on $\{t : \mathcal{K}^c_{1,N_1(t)}(\epsilon), \mathcal{M}^c_\epsilon(t), N_1(t) = n\}$. Thus, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[j(t) \neq 1, \mathcal{K}^c_{1,N_1(t)}(\epsilon), \mathcal{M}^c_\epsilon(t)]\right] &\leq \mathbb{E}\left[\sum_{n=\bar{n}}^{T} \sum_{m=1}^{T} \mathbb{1}[\mathcal{K}^c_{1,n}(\epsilon)] p_n(\epsilon|\theta_{1,n})^m\right] \\ &\leq \sum_{n=\bar{n}}^{T} \mathbb{E}\left[\mathbb{1}\left[\mathcal{K}^c_{1,n}(\epsilon)\right] \frac{p_n(\epsilon|\theta_{1,n})}{1 - p_n(\epsilon|\theta_{1,n})}\right] \qquad (23)\end{aligned}$$

for $p_n(\cdot|\cdot)$ defined in (20). From now on, we fix $n \geq \bar{n}$ and drop the argument n of $\hat{\kappa}_1(n)$, $\hat{\alpha}_1(n)$ and $\bar{\alpha}_1(n)$ for simplicity.

Under STS **with** $k \in \mathbb{Z}_{\geq 2}$**:** By applying Lemma 10 and (22) under STS with $k \in \mathbb{Z}_{\geq 0}$, we can decompose the expectation in (23) as

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left[\mathcal{K}^c_{1,n}(\epsilon)\right] \frac{p_n(\epsilon|\hat{\theta}_{1,n})}{1 - p_n(\epsilon|\hat{\theta}_{1,n})}\right] &\leq \mathbb{P}[\hat{\kappa}_1 \geq \mu_1 - \epsilon] \frac{e^{-n}}{1 - e^{-n}} + \mathbb{P}[\hat{\kappa}_1 \in K(\epsilon), \hat{\alpha}_1 < \alpha_1 + \rho] \frac{h(\mu_1, \epsilon, n)}{C_2(\mu_1, \epsilon, k)} \\ &\quad + \underbrace{\mathbb{E}_{\hat{\theta}_{1,n}}\left[\frac{\mathbb{1}[\hat{\kappa}_1 \in K(\epsilon), \hat{\alpha}_1 > \alpha_1 + \rho]}{G_k(1/\hat{\alpha}_1; n)(1 - C_{1,n})}\right]}_{(*)}, \qquad (24)\end{aligned}$$

where we denoted $C_{1,n} = C_1(\mu_1, \epsilon, n)$. For simplicity, let us define $z := \frac{1}{\hat{\alpha}_1}$ where $z \sim \mathrm{Erlang}(n-1, n\alpha_1)$ holds from Fact (11) since $\hat{\alpha}_1 \sim \mathrm{IG}(n-1, n\alpha_1)$ in (2). From the independence of $\hat{\kappa}$ and $\hat{\alpha}$ and distributions of z and $\hat{\alpha}$ in (10) and (2),

respectively, we have

$$(\divideontimes) = \int_0^{\frac{1}{\alpha_1+\rho}} z^{n-2}e^{-n\alpha_1 z}\frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)}\int_{\hat{\kappa}_1\in K(\epsilon)}\frac{f^{\mathrm{Pa}}_{\kappa_1,n\alpha_1}(\hat{\kappa}_1)}{G_k(z;n)(1-C_{1,n})}\mathrm{d}\hat{\kappa}_1\mathrm{d}z$$
$$= \mathbb{P}[\hat{\kappa}_1\in K(\epsilon)]\int_0^{\frac{1}{\alpha_1+\rho}}\frac{z^{n-2}e^{-n\alpha_1 z}}{G_k(z;n)(1-C_{1,n})}\frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)}\mathrm{d}z.$$

By substituting the CDF in (11), we obtain

$$\begin{aligned}G_k(z;n) &= F^{\mathrm{Er}}_{n-k,nz}(\alpha_1+\rho)\\ &= \frac{1}{\Gamma(n-k)}\int_0^{n(\alpha_1+\rho)z} e^{-t}t^{n-k-1}\mathrm{d}t\\ &\geq \frac{e^{-n(\alpha_1+\rho)z}}{\Gamma(n-k)}\int_0^{n(\alpha_1+\rho)z} t^{n-k-1}\mathrm{d}t\\ &= \frac{e^{-n(\alpha_1+\rho)z}}{\Gamma(n-k+1)}(n(\alpha_1+\rho)z)^{n-k}. \end{aligned}\tag{25}$$

Therefore,

$$\begin{aligned}\frac{(\divideontimes)}{\mathbb{P}[\hat{\kappa}\in K(\epsilon)]} &\leq \int_0^{\frac{1}{\alpha_1+\rho}}\frac{\Gamma(n-k+1)}{(n(\alpha_1+\rho)z)^{n-k}(1-C_{1,n})}e^{n(\alpha_1+\rho)z}\frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)}z^{n-2}e^{-n\alpha_1 z}\mathrm{d}z\\ &= \frac{\Gamma(n-k+1)}{\Gamma(n-1)(1-C_{1,n})}(\alpha_1+\rho)^{k-1}\left(\frac{\alpha_1}{\alpha_1+\rho}\right)^{n-1}n^{k-1}\int_0^{\frac{1}{\alpha_1+\rho}} z^{k-2}e^{n\rho z}\mathrm{d}z\\ &\leq \frac{\Gamma(n-k+1)}{\Gamma(n-1)(1-C_{1,n})}(\alpha_1+\rho)^{k-1}e^{-\frac{\rho}{\alpha_1+\rho}(n-1)}\frac{n^{k-1}}{(n\rho)^{k-2}}\int_0^{\frac{1}{\alpha_1+\rho}}(n\rho z)^{k-2}e^{n\rho z}\mathrm{d}z. \end{aligned}\tag{26}$$

By letting $n\rho z = y$, we can bound the integral in (26) as

$$\begin{aligned}\frac{n^{k-1}}{(n\rho)^{k-2}}\int_0^{\frac{1}{\alpha_1+\rho}}(n\rho z)^{k-2}e^{n\rho z}\mathrm{d}z &= \rho^{-(k-1)}\int_0^{\frac{n\rho}{\alpha_1+\rho}} y^{k-2}e^{y}\mathrm{d}y\\ &\leq \rho^{-(k-1)}e^{n\frac{\rho}{\alpha_1+\rho}}\int_0^{\frac{n\rho}{\alpha_1+\rho}} y^{k-2}\mathrm{d}y &(27)\\ &= \frac{e^{n\frac{\rho}{\alpha_1+\rho}}}{k-1}\left(\frac{n}{\alpha_1+\rho}\right)^{k-1}, \quad \text{if } k\in\mathbb{Z}_{\geq 2} &(28)\end{aligned}$$

where (28) holds only for $k\in\mathbb{Z}_{\geq 2}$ since the integral in (27) diverges for $k\in\mathbb{Z}_{\leq 1}$.

By applying (28) to (26), we obtain for $k\in\mathbb{Z}_{\geq 2}$

$$\begin{aligned}(\divideontimes) &\leq \mathbb{P}[\hat{\kappa}\in K(\epsilon)]\frac{e^{\frac{\rho}{\alpha_1+\rho}}}{1-C_{1,n}}\frac{n^{k-1}}{k-1}\frac{\Gamma(n-k+1)}{\Gamma(n-1)}\\ &\leq \frac{e^{1-\frac{\epsilon\alpha_1}{\kappa+\epsilon}n}}{1-C_{1,n}}\frac{\Gamma(n-k+1)}{\Gamma(n-1)}\frac{n^{k-1}}{k-1} = \mathcal{O}(ne^{-n\epsilon}),\end{aligned}\tag{29}$$

where (29) can be obtained by Lemma 15 and $\frac{\rho}{\alpha_1+\rho}<1$. By combining (29) with (24) and (23), we obtain for $k\in\mathbb{Z}_{\geq 2}$

$$\begin{aligned}\mathbb{E}\left[\sum_{t=K\bar{n}+1}^{T}\mathbb{1}[j(t)\neq 1,\mathcal{K}^c_{1,N_1(t)}(\epsilon),\mathcal{M}^c_\epsilon(t)]\right] &\leq \sum_{n=\bar{n}}^{T}\left(\frac{e^{-n}}{1-e^{-n}}+\frac{e^{-n\frac{3\epsilon}{4\mu_1}}+\frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}}{C_2(\mu_1,\epsilon,k)}+(\divideontimes)\right)\\ &\leq \sum_{n=\bar{n}}^{T}\mathcal{O}(e^{-n})+\mathcal{O}(e^{-n\epsilon})+\mathcal{O}(e^{-n\epsilon^2})+\mathcal{O}(ne^{-n\epsilon})\\ &= \mathcal{O}(1)+\mathcal{O}(\epsilon^{-1})+\mathcal{O}(\epsilon^{-2})+\mathcal{O}(\epsilon^{-2}),\end{aligned}$$

which concludes the proof under STS with $k\in\mathbb{Z}_{\geq 2}$.

Under STS-T **with** $k \in \mathbb{Z}_{\geq 0}$**:** Under STS-T, we have the following inequality instead of (24):

$$\mathbb{E}\left[\mathbb{1}\left[\mathcal{K}_n^c(\epsilon)\right]\frac{p_n(\epsilon|\bar{\theta}_{1,n})}{1-p_n(\epsilon|\bar{\theta}_{1,n})}\right] \leq \mathbb{P}[\hat{\kappa}_1 \geq \mu_1 - \epsilon]\frac{e^{-n}}{1-e^{-n}} + \mathbb{P}[\hat{\kappa}_1 \in K(\epsilon), \bar{\alpha}_1 < \alpha_1 + \rho]\frac{h(\mu_1,\epsilon,n)}{C_2(\mu_1,\epsilon,k)} + \underbrace{\mathbb{E}_{\bar{\theta}_{1,n}}\left[\frac{\mathbb{1}[\hat{\kappa}_1 \in K(\epsilon), \bar{\alpha}_1 \in (\alpha_1+\rho, n]]}{G_k(1/\bar{\alpha}_1;n)(1-C_{1,n})}\right]}_{(\star)}. \tag{30}$$

From $\mathbb{1}[\bar{\alpha}_1(n) < n] = \mathbb{1}[\bar{\alpha}_1(n) = \hat{\alpha}_1(n)]$, it holds that

$$(\star) = \mathbb{E}_{\hat{\theta}_{1,n}}\left[\frac{\mathbb{1}[\hat{\kappa}_1 \in K(\epsilon), \hat{\alpha}_1 \in (\alpha_1+\rho, n)]}{G_k(1/\hat{\alpha}_1;n)(1-C_{1,n})}\right] + \mathbb{E}_{\bar{\theta}_{1,n}}\left[\frac{\mathbb{1}[\hat{\kappa}_1 \in K(\epsilon), \bar{\alpha}_1 = n]}{G_k(1/\bar{\alpha}_1;n)(1-C_{1,n})}\right].$$

Since $\mathbb{1}[\bar{\alpha}_1(n) = n] = \mathbb{1}[\hat{\alpha}_1(n) \geq n]$ holds and $\hat{\kappa}$ and $\hat{\alpha}$ are independent, we have for $z = \frac{1}{\hat{\alpha}_1} \sim \mathrm{Erlang}(n-1, n\alpha_1)$

$$\frac{(\star)}{\mathbb{P}[\hat{\kappa}_1 \in K(\epsilon)]} \leq \underbrace{\int_{\frac{1}{n}}^{\frac{1}{\alpha_1+\rho}} \frac{z^{n-2}e^{-n\alpha_1 z}}{G_k(z;n)(1-C_{1,n})}\frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)}\mathrm{d}z}_{(\dagger)} + \underbrace{\frac{1}{G_k(1/n;n)(1-C_{1,n})}\mathbb{P}\left[\frac{1}{\hat{\alpha}_1} \leq \frac{1}{n}\right]}_{(\diamond)}, \tag{31}$$

where the same techniques on $(\divideontimes)$ can be applied to derive an upper bound of $(\dagger)$. By following the same steps as (26) and (27), we obtain

$$\int_{\rho}^{\frac{n\rho}{\alpha+\rho}} y^{k-2}\mathrm{d}y \leq \begin{cases} \left(\frac{n\rho}{\alpha+\rho}\right)^{k-1}, & \text{if } k \in \mathbb{Z}_{\geq 2}, \\ \log\left(\frac{n}{\alpha+\rho}\right), & \text{if } k = 1, \\ \rho^{k-1}/(1-k), & \text{if } k \in \mathbb{Z}_{k\leq 0}, \end{cases}$$

as a counterpart of the integral in (27). By following the same steps as (28) and (29), we have

$$(\dagger) \leq \begin{cases} \frac{\Gamma(n-k+1)}{\Gamma(n-1)}\frac{en^{k-1}}{k-1}, & \text{if } k \in \mathbb{Z}_{\geq 2}, \\ (n-1)\log\left(\frac{n}{\alpha+\rho}\right), & \text{if } k = 1, \\ \frac{\Gamma(n-k+1)}{\Gamma(n-1)}\frac{e}{(1-k)(\alpha+\rho)^{1-k}}, & \text{if } k \in \mathbb{Z}_{k\leq 0}. \end{cases} \tag{32}$$

Note that the probability term in $(\diamond)$ is the same as the CDF of the $\mathrm{Erlang}(n-1, n\alpha_1)$ with rate $n\alpha_1$ evaluated at $\frac{1}{n}$ since $\hat{\alpha}_1 \sim \mathrm{IG}(n-1, n\alpha_1)$ from (2). Thus, we have

$$\begin{aligned} (\diamond) &= \frac{1}{1-C_{1,n}}\frac{1}{G_k(1/n;n)}\frac{\gamma(n-1,\alpha_1)}{\Gamma(n-1)} \\ &\leq \frac{e^{\alpha_1+\rho}}{1-C_{1,n}}\frac{\Gamma(n-k+1)}{(\alpha_1+\rho)^{n-k}}\frac{\gamma(n-1,\alpha_1)}{\Gamma(n-1)} \quad \text{by (25)} \\ &\leq \frac{e^{\alpha_1+\rho}}{1-C_{1,n}}\frac{\Gamma(n-k+1)}{\Gamma(n-1)}\frac{\alpha_1^{n-1}}{(\alpha_1+\rho)^{n-k}} \qquad (33) \\ &\leq \frac{e^{\alpha_1+\rho}}{1-C_{1,n}}\frac{\Gamma(n-k+1)}{\Gamma(n-1)}\frac{1}{(\alpha_1+\rho)^{1-k}} \\ &= \mathcal{O}(n^{2-k}), \qquad (34) \end{aligned}$$

where (33) holds from $\gamma(s,x) \leq x^s$ for any $s \geq 1$ and $x > 0$. Therefore, by combining (32) and (34) with (31) and $\mathbb{P}[\hat{\kappa} \in K(\epsilon)] = \mathcal{O}(e^{-n\epsilon})$, we have

$$(\star) \leq \begin{cases} \mathcal{O}(ne^{-n\epsilon}), & \text{if } k \in \mathbb{Z}_{\geq 2} \\ \mathcal{O}(n\log(n)e^{-n\epsilon}), & \text{if } k = 1, \\ \mathcal{O}(n^{2-k}e^{-n\epsilon}), & \text{if } k \in \mathbb{Z}_{\leq 0}. \end{cases} \tag{35}$$

By combining (35) with (30) and (23), we obtain for $k \in \mathbb{Z}_{\geq 0}$

$$\begin{aligned}\mathbb{E}\Bigg[\sum_{t=K\bar{n}+1}^{T} \mathbb{1}[j(t) \neq 1, \mathcal{K}^c_{1,N_1(t)}(\epsilon), \mathcal{M}^c_\epsilon(t)]\Bigg] &\leq \sum_{n=\bar{n}}^{T}\left(\frac{e^{-n}}{1-e^{-n}} + \frac{e^{-n\frac{3\epsilon}{4\mu_1}} + \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}}{C_2(\mu,\epsilon,k)} + (\star)\right)\\ &\leq \sum_{n=\bar{n}}^{T}\left(\mathcal{O}(e^{-n}) + \mathcal{O}(e^{-n\epsilon}) + \mathcal{O}\left(e^{-n\epsilon^2}\right) + \mathcal{O}\left(\psi(n,k)e^{-n\epsilon}\right)\right)\\ &= \mathcal{O}(1) + \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-2}) + \mathcal{O}\left(\epsilon^{-\max(2,3-k)}\right),\end{aligned}$$

where

$$\psi(n,k) = n\mathbb{1}[k \geq 2] + n\log(n)\mathbb{1}[k=1] + n^{2-k}\mathbb{1}[k \leq 0].$$

Note that the analysis on term $(\star)$ also holds for STS-T with $k \in \mathbb{Z}_{<0}$. However, differently from the case of $k \in \{0,1\}$, priors with $k \in \mathbb{Z}_{<0}$ have additional problems in Lemma 10 under the event $\{\hat{\kappa}_1 \in K(\epsilon), \bar{\alpha}_1(n) \leq \alpha_1 + \rho\}$, where the upper bound becomes a constant $\frac{1}{2}$. □

C.2. Proof of Lemma 6

Firstly, we state a well-known fact that is utilized in the proof.

Fact 12. When $X \sim \mathrm{Erlang}(n,\beta)$ with rate parameter β, then $2\beta X$ follows the chi-squared distribution with $2n$ degree of freedom, i.e., $2\beta X \sim \chi^2_{2n}$.

Lemma 6. *Under* STS *and* STS-T *with $k \in \mathbb{Z}$, it holds that for any $a \in [K]$*

$$\begin{aligned}&\sum_{t=\bar{n}K+1}^{T} \mathbb{E}[\mathbb{1}[j(t)=a, \mathcal{M}_\epsilon(t), \mathcal{E}_{a,N_a(t)}(\epsilon)]]\\ &\qquad\leq \max(0,k) + 1 + \frac{1}{\alpha_a\epsilon}\mathbb{1}[k>0] + \frac{\log T}{D_{a,k}(\epsilon)}.\end{aligned}$$

where $D_{a,k}(\epsilon) > 0$ is a finite problem-deterministic constant satisfying $\lim_{\epsilon\to 0} D_{a,k}(\epsilon) = \mathrm{KL}_{\inf}(a)$.

Proof. From the sampling rule, it holds that

$$\begin{aligned}\sum_{t=\bar{n}K+1}^{T} \mathbb{P}\left[j(t)=a, \tilde{\mu}^*(t) \geq \mu_1-\epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon)\right] &= \sum_{t=\bar{n}K+1}^{T} \mathbb{P}\left[j(t)=a, \max_{a\in[K]}\tilde{\mu}_a(t) \geq \mu_1-\epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon)\right]\\ &\leq \sum_{t=\bar{n}K+1}^{T} \mathbb{P}\left[j(t)=a, \tilde{\mu}_a(t) \geq \mu_1-\epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon)\right].\end{aligned}$$

Fix a time index t and denote $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid \mathcal{F}_t]$ and $N_a(t) = n$. To simplify notations, we drop the argument t of $\tilde{\kappa}_a(t), \tilde{\alpha}_a(t)$ and $\tilde{\mu}_a(t)$ and the argument n of $\hat{\kappa}_a(n), \hat{\alpha}_a(n), \bar{\alpha}_a(n)$.

Since $\tilde{\kappa}_a \in (0, \hat{\kappa}_a]$ holds from its posterior distribution, if $\tilde{\alpha}_a \geq \frac{\mu_1-\epsilon}{\mu_1-\epsilon-\hat{\kappa}_a}$ holds, then $\tilde{\mu}_a = \frac{\tilde{\kappa}_a\tilde{\alpha}_a}{\tilde{\alpha}_a-1} \leq \mu_1-\epsilon$ holds for any $\tilde{\kappa}_a$. Recall that $f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(\cdot)$ denotes a density function of $\mathrm{Erlang}\left(n-k, \frac{n}{\alpha_{a,n}}\right)$ with rate parameter $\frac{n}{\alpha_{a,n}}$, which is the marginalized posterior distribution of $\tilde{\alpha}$ under STS and STS-T. From the CDF of $\tilde{\kappa}$ in (9), if $\hat{\kappa}_a < \mu_1-\epsilon$, then

$$\begin{aligned}\mathbb{P}_t\left[\tilde{\mu}_a \geq \mu_1-\epsilon\right] &= \mathbb{P}_t\left[\tilde{\alpha}_a \leq 1\right] + \mathbb{P}_t\left[\tilde{\kappa}_a \geq \frac{\tilde{\alpha}_a-1}{\tilde{\alpha}_a}(\mu_1-\epsilon) \cap \tilde{\alpha}_a \in \left(1, \frac{\mu_1-\epsilon}{\mu_1-\epsilon-\hat{\kappa}_a}\right)\right]\\ &= \int_0^1 f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x + \int_1^{\frac{\mu_1-\epsilon}{\mu_1-\epsilon-\hat{\kappa}}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathbb{P}_t\left[\tilde{\kappa}_a \geq \frac{x-1}{x}(\mu_1-\epsilon)\right]\mathrm{d}x\\ &= \int_0^1 f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x + \int_1^{\frac{\mu_1-\epsilon}{\mu_1-\epsilon-\hat{\kappa}}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(1-\left(\frac{x-1}{\hat{\kappa}x}(\mu_1-\epsilon)\right)^{nx}\right)\mathrm{d}x\\ &= \int_0^{\frac{\mu_1-\epsilon}{\mu_1-\epsilon-\hat{\kappa}}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x - \int_1^{\frac{\mu_1-\epsilon}{\mu_1-\epsilon-\hat{\kappa}}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}x}(\mu_1-\epsilon)\right)^{nx}\mathrm{d}x.\end{aligned}$$

Since $\frac{x}{x-y}$ is increasing with respect to $y < x$ and $\hat{\kappa} \leq \kappa + \epsilon$ holds on $\mathcal{E}$, we have for $\mathcal{E}$

$$\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\kappa}} \leq \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - (\kappa + \epsilon)}.$$

Let

$$\eta_a(\epsilon) = \frac{\kappa_a(\Delta_a - \epsilon) - \epsilon\mu_a}{(\mu_a - \kappa_a)(\mu_1 - \kappa_a - 2\epsilon)} > 0 \tag{36}$$

be a deterministic constant that depends only on the model and ϵ and satisfies

$$\begin{aligned}
\alpha_a - \eta_a(\epsilon) &= \frac{\mu_a}{\mu_a - \kappa_a} - \frac{\kappa_a(\Delta_a - \epsilon) - \epsilon\mu_a}{(\mu_a - \kappa_a)(\mu_1 - \kappa_a - 2\epsilon)} \\
&= \frac{\mu_a\mu_1 - \kappa_a\mu_a - 2\epsilon\mu_a - \kappa_a(\mu_1 - \mu_a - \epsilon) + \epsilon\mu_a}{(\mu_a - \kappa_a)(\mu_1 - \kappa_a - 2\epsilon)} \\
&= \frac{\mu_1(\mu_a - \kappa_a) - \epsilon(\mu_a - \kappa_a)}{(\mu_a - \kappa_a)(\mu_1 - \kappa_a - 2\epsilon)} \\
&= \frac{\mu_1 - \epsilon}{\mu_1 - \kappa_a - 2\epsilon}.
\end{aligned}$$

Since $\eta_a(\epsilon) > 0$, it holds that for any $\epsilon \in (0, \varepsilon_a)$

$$\alpha_a - \eta_a(\epsilon) = \frac{\mu_1 - \epsilon}{\mu_1 - \kappa_a - 2\epsilon} \leq \frac{\mu_a}{\mu_a - \kappa_a} = \alpha_a. \tag{37}$$

Note that $\frac{\mu}{\mu - \kappa} = \alpha$ holds and the change of μ to μ' with fixed κ that is $\frac{\mu'}{\mu' - \kappa}$, implies how the value of the shape parameter α' should be to satisfy $\mu((\kappa, \alpha')) = \mu'$. For example, $\theta = (\kappa_a + \varepsilon_a, \alpha_a)$ satisfies $\mu(\theta) \leq \mu_a + \frac{\delta_a}{2}$. Thus, if $\mu((\kappa_a + \varepsilon_a, \alpha)) = \mu_1 - \epsilon > \mu_a + \frac{\delta_a}{2}$, then α should be smaller than α_a. Hence, we have

$$\begin{aligned}
&\mathbb{1}[\mathcal{E}_{a,n}(\epsilon)]\mathbb{P}_t\left[\tilde{\mu}_a \geq \mu_1 - \epsilon\right] \\
&\qquad \leq \mathbb{1}[\mathcal{E}_{a,n}(\epsilon)] \left(\int_0^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\kappa}}} f^{\mathrm{Er}}_{n-k, \frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x - \int_1^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\kappa}}} f^{\mathrm{Er}}_{n-k, \frac{n}{\alpha_{a,n}}}(x) \left(\frac{x-1}{\hat{\kappa}x}(\mu_1 - \epsilon) \right)^{nx} \mathrm{d}x \right) \\
&\qquad \leq \mathbb{1}[\mathcal{E}_{a,n}(\epsilon)] \int_0^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\kappa}}} f^{\mathrm{Er}}_{n-k, \frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \qquad (38) \\
&\qquad \leq \mathbb{1}[\mathcal{E}_{a,n}(\epsilon)] \int_0^{\alpha_a - \eta_a(\epsilon)} f^{\mathrm{Er}}_{n-k, \frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x = \mathbb{1}[\mathcal{E}_{a,n}(\epsilon)]\mathbb{P}_t[\tilde{\alpha}_a(t) \leq \alpha_a - \eta_a(\epsilon)]. \qquad (39)
\end{aligned}$$

Therefore, by taking expectation and using Fact 12, we have

$$\begin{aligned}
\mathbb{P}\left[\tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,n}(\epsilon)\right] &\leq \mathbb{P}[\tilde{\alpha}_a \leq \alpha_a - \eta_a(\epsilon), \mathcal{E}_{a,n}(\epsilon)], \\
&= \mathbb{P}\left[Z \leq \frac{2n}{\alpha_{a,n}} \left(\alpha - \eta_a(\epsilon)\right), \mathcal{E}_{a,n}(\epsilon) \right] \qquad (40)
\end{aligned}$$

where Z is a random variable following the chi-squared distribution with $2(n-k)$ degrees of freedom, i.e., $Z \sim \chi^2_{2n-2k}$.

C.2.1. UNDER STS

Here, we first consider the case of STS where we replace $\alpha_{a,n}$ with $\hat{\alpha}_a(n)$.

Since $\hat{\alpha}_a \in [\alpha_a - \epsilon_{a,l}, \alpha_a + \epsilon_{a,u}]$ holds on $\mathcal{E}_{a,n}(\epsilon)$, we have

$$\frac{1}{\alpha_a} - \epsilon\left(1 + \frac{1}{\kappa_a}\right) = \frac{1}{\alpha_a + \epsilon_{a,u}} \leq \frac{1}{\hat{\alpha}_a} \leq \frac{1}{\alpha_a - \epsilon_{a,l}} = \frac{1}{\alpha_a} + \epsilon \tag{41}$$

by the definition of $\epsilon_{a,l}(\epsilon)$ and $\epsilon_{a,u}(\epsilon)$ in (13).

By replacing $\alpha_{a,n}$ with $\hat{\alpha}_a(n)$ in (40) and applying (41), we have

$$\begin{aligned}\mathbb{P}\left[\tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,n}(\epsilon)\right] &\leq \mathbb{P}\left[Z \leq \frac{2n}{\hat{\alpha}_a}\left(\alpha_a - \eta_a(\epsilon)\right), \mathcal{E}_{a,n}(\epsilon)\right] \\ &\leq \mathbb{P}\left[Z \leq 2n\left(\frac{1}{\alpha_a} + \epsilon\right)\left(\alpha_a - \eta_a(\epsilon)\right)\right] \\ &= \mathbb{P}\left[Z \leq 2(n-k)\frac{n}{n-k}\left(\frac{1}{\alpha_a} + \epsilon\right)\left(\alpha_a - \eta_a(\epsilon)\right)\right]. \end{aligned} \tag{42}$$

Priors with $k \in \mathbb{Z}_{\leq 0}$. Let us first consider the case $k \in \mathbb{Z}_{\leq 0}$, where we have $\frac{n}{n-k} \leq 1$. It holds that

$$\begin{aligned}\mathbb{P}\left[\tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,n}(\epsilon)\right] &\leq \mathbb{P}\left[Z \leq 2(n-k)\frac{n}{n-k}\left(\frac{1}{\alpha_a} + \epsilon\right)\left(\alpha_a - \eta_a(\epsilon)\right)\right] \\ &\leq \mathbb{P}\left[Z \leq 2(n-k)\left(\frac{1}{\alpha_a} + \epsilon\right)\left(\alpha_a - \eta_a(\epsilon)\right)\right].\end{aligned}$$

Note that the definition of ε_a in Theorem 2 is set to satisfy $\left(\frac{1}{\alpha} + \epsilon\right)\left(\alpha - \eta_a(\epsilon)\right) < 1$ for any $\epsilon \leq \varepsilon_a$. Thus, we can apply Lemma 19, which shows

$$\mathbb{P}\left[Z \leq 2(n-k)\left(1 - \frac{\eta_a(\epsilon)}{\alpha_a} + \epsilon(\alpha_a - \eta_a(\epsilon))\right)\right] \leq e^{-(n-k)D_{a,k}(\epsilon)}, \tag{43}$$

where

$$D_{a,k}(\epsilon) := -\log\left(1 - \frac{\eta_a(\epsilon)}{\alpha_a} + (\max(0,k)+1)\epsilon(\alpha_a - \eta_a(\epsilon))\right) - \frac{\eta_a(\epsilon)}{\alpha_a} + (\max(0,k)+1)\epsilon(\alpha_a - \eta_a(\epsilon)) > 0 \tag{44}$$

is a finite constant that only depends on the model parameters, ϵ, and prior parameter k.

For arbitrary $n_a > 0$, applying (43) to (40) gives

$$\begin{aligned}\sum_{t=\bar{n}K+1}^{T} \mathbb{E}[\mathbb{1}[j(t) = a, \tilde{\mu}_1(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon)]] &\leq \sum_{t=\bar{n}K+1}^{T} \mathbb{P}[j(t) = a, \tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,n}(\epsilon)] \\ &\leq n_a + \sum_{t=\bar{n}K+1}^{T} \mathbb{P}[\tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon), N_a(t) \geq n_a] \\ &\leq n_a + \sum_{t=\bar{n}K+1}^{T} e^{-(n_a-k)D_{a,k}(\epsilon)} \\ &\leq n_a + \sum_{t=\bar{n}K+1}^{T} e^{-n_a D_{a,k}(\epsilon)} = n_a + Te^{-n_a D_{a,k}(\epsilon)}.\end{aligned}$$

Letting $n_a = \frac{\log T}{D_{a,k}(\epsilon)}$ concludes the cases of priors with $k \in \mathbb{Z}_{\leq 0}$.

Priors with $k \in \mathbb{Z}_{>0}$ Next, consider the case $k \in \mathbb{Z}_{>0}$. Recall that we first play every arm $k+1$ times if $k > 0$, which implies that $n - k > 0$. For $n \geq \frac{1}{\alpha\epsilon} + k + 1$, it holds that

$$\frac{n}{n-k}\left(\frac{1}{\alpha} + \epsilon\right) \leq \frac{1}{\alpha} + (k+1)\epsilon. \tag{45}$$

By applying (45) to (40), we have for $n \geq \frac{1}{\alpha\epsilon} + k + 1$,

$$\mathbb{P}[\tilde{\alpha}_a \leq \alpha_a - \eta_a(\epsilon), \mathcal{E}_{a,N_a(t)}(\epsilon)] \leq \mathbb{P}\left[Z \leq 2(n-k)\left(1 - \frac{\eta_a(\epsilon)}{\alpha_a} + (k+1)\epsilon(\alpha_a - \eta_a(\epsilon))\right)\right].$$

Similarly, by applying Lemma 19, one can see that for $n \geq \frac{1}{\alpha_a \epsilon} + k + 1$

$$\mathbb{P}[\tilde{\alpha}_a \leq \alpha_a - \eta_a(\epsilon), \mathcal{E}_{a,N_a(t)}(\epsilon)] \leq e^{-(n-k)D_{a,k}(\epsilon)}, \tag{46}$$

where $D_{a,k}(\epsilon)$ is defined in (44).

When $k \in \mathbb{Z}_{>0}$, let $n_a \geq \frac{1}{\alpha_a \epsilon} + k + 1$ be arbitrary. By applying (46) to (40) again, we have

$$\begin{aligned}
\sum_{t=\bar{n}K+1}^{T} \mathbb{E}[\mathbb{1}[j(t) = a, \tilde{\mu}_1(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon)]] &\leq \sum_{t=\bar{n}K+1}^{T} \mathbb{P}[j(t) = a, \tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,n}(\epsilon)] \\
&\leq n_a + \sum_{t=\bar{n}K+1}^{T} \mathbb{P}[\tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,N_a(t)}(\epsilon), N_a(t) \geq n_a] \\
&\leq n_a + \sum_{t=\bar{n}K+1}^{T} e^{-(n_a-k)D_{a,k}(\epsilon)} = n_a + Te^{-(n_a-k)D_{a,k}(\epsilon)}.
\end{aligned}$$

Letting $n_a = k + 1 + \frac{1}{\alpha_a \epsilon} + \frac{\log T}{D_{a,k}(\epsilon)}$ concludes the cases of priors with $k > 0$.

C.2.2. Under STS-T

Here, we consider the case of STS-T where we replace $\alpha_{a,n}$ with $\bar{\alpha}_a(n) = \min(\hat{\alpha}_a(n), n)$. From the definition of $\bar{\alpha}_a(n)$, it holds that for $\epsilon \leq \varepsilon_a$

$$\forall n \geq \alpha_a + 1 : \mathbb{1}[\bar{\alpha}_a(n) = \hat{\alpha}_a(n), \mathcal{A}_{a,n}(\epsilon)] = 1.$$

Therefore, for $n \geq \alpha_a + 1$, the analysis on STS can be applied to STS-T directly.

Let us consider the case where $\bar{\alpha}_a(n) = n < \alpha_a + 1$ holds under the condition $\mathcal{A}_{a,n}(\epsilon)$. By replacing $\alpha_{a,n}$ with n in (40) and following the same steps as in (40) and (43), we have for any $k \in \mathbb{Z}$,

$$\begin{aligned}
\mathbb{P}\left[\tilde{\mu}_a(t) \geq \mu_1 - \epsilon, \mathcal{E}_{a,n}(\epsilon)\right] &\leq \mathbb{P}\left[Z \leq \frac{2n}{n}\left(\alpha_a - \eta_a(\epsilon)\right), \mathcal{E}_{a,n}(\epsilon)\right] \\
&\leq \mathbb{P}\left[Z \leq 2(n-k)\frac{1}{n-k}\left(\frac{1}{\alpha_a} + \epsilon\right)\left(\alpha_a - \eta_a(\epsilon)\right), \mathcal{E}_{a,n}(\epsilon)\right] \\
&\leq \mathbb{P}\left[Z \leq 2(n-k)\left(\frac{1}{\alpha_a} + \epsilon\right)\left(\alpha_a - \eta_a(\epsilon)\right), \mathcal{E}_{a,n}(\epsilon)\right] \\
&\leq e^{-(n-k)D_{a,k}(\epsilon)},
\end{aligned}$$

where $D_{a,k}(\epsilon)$ defined in (44). Therefore, the same result follows by the analysis in Section C.2.1.

C.2.3. Asymptotic behavior of $D_{a,k}(\epsilon)$

Here, we first provide the alternative formulation of $D_{a,k}(\epsilon)$ defined in (44). Let us rewrite the definition of $D_{a,k}(\epsilon)$ as

$$\begin{aligned}
D_{a,k}(\epsilon) &= -\log\left(1 - \frac{\eta_a(\epsilon)}{\alpha_a} + (\max(0,k)+1)\epsilon(\alpha_a - \eta_a(\epsilon)\right) - \frac{\eta_a(\epsilon)}{\alpha_a} + (\max(0,k)+1)\epsilon(\alpha_a - \eta_a(\epsilon) \\
&= -\log\left(\frac{(\alpha_a - \eta_a(\epsilon))(1 + (\max(0,k)+1)\alpha_a\epsilon)}{\alpha_a}\right) + \frac{(\alpha_a - \eta_a(\epsilon))(1 + (\max(0,k)+1)\alpha_a\epsilon)}{\alpha_a} - 1 \\
&= \log\left(\alpha_a \frac{1}{\alpha_a - \eta_a(\epsilon)} \frac{1}{1 + (\max(0,k)+1)\alpha_a\epsilon}\right) + \frac{(\alpha_a - \eta_a(\epsilon))(1 + (\max(0,k)+1)\alpha_a\epsilon)}{\alpha_a} - 1.
\end{aligned}$$

By injecting the closed form of $\alpha_a - \eta_a(\epsilon)$ given in (37) and defining $b_{a,k}(\epsilon) = \frac{1}{1+(\max(0,k)+1)\alpha_a\epsilon}$, $D_{a,k}$ can be written as

$$D_{a,k}(\epsilon) = \log\left(\alpha_a b_{a,k}(\epsilon)\frac{\mu_1 - \epsilon - (\kappa_a + \epsilon)}{\mu_1 - \epsilon}\right) + \frac{1}{\alpha_a b_{a,k}(\epsilon)}\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - (\kappa_a + \epsilon)} - 1.$$

From Lemma 1, one can observe that

$$D_{a,k}(\epsilon) = \inf_{\theta:\mu(\theta)>\mu_1-\epsilon} \mathrm{KL}(\mathrm{Pa}(\kappa_a+\epsilon, \alpha_a b_{a,k}(\epsilon)), \mathrm{Pa}(\theta)).$$

Since $\lim_{\epsilon\to 0} b_{a,k}(\epsilon) = 1$ for any $a \in [K]$ and $k \in \mathbb{Z}$, it holds that

$$\lim_{\epsilon\to 0} D_{a,k}(\epsilon) = \inf_{\theta:\mu(\theta)>\mu_1} \mathrm{KL}(\mathrm{Pa}(\kappa_a, \alpha_a), \mathrm{Pa}(\theta)) = \mathrm{KL}_{\inf}(a),$$

□

C.3. Proof of Lemma 7

Firstly, we state two well-known facts that are utilized in the proof.

Fact 13. When $X \sim \mathrm{Pa}(\kappa, \alpha)$ with the scale parameter $\kappa \in \mathbb{R}$ and rate parameter $\alpha \in \mathbb{R}_+$, then $\log\left(\frac{X}{\kappa}\right)$ follows the exponential distribution with rate α, i.e., $\log\left(\frac{X}{\kappa}\right) \sim \mathrm{Exp}(\alpha)$.

Fact 14. Let $X_1, \ldots, X_n$ be identically independently distributed with the exponential distribution with the rate parameter α, i.e., $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\alpha)$ for any $i \in [n]$. Then, their sum follows the Erlang distribution with the shape parameter $n \in \mathbb{N}$ and rate parameter α, i.e., $\sum_{i=1}^n X_i \sim \mathrm{Erlang}(n, \alpha)$.

Lemma 7. *Under* STS *and* STS-T *with* $k \in \mathbb{Z}_{\geq 0}$,

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)]\right] \leq \mathcal{O}(\epsilon^{-1}).$$

Proof. When one considers the Pareto distribution with known scale parameter κ that belongings to the one-dimensional exponential family, the posterior on the shape parameter $\alpha^{\mathrm{one}} > 0$ after observing $n = N_1(t)$ rewards is given for $k \in \mathbb{Z}$

$$\alpha^{\mathrm{one}} \mid \mathcal{F}_t \sim \mathrm{Erlang}\left(n-k+1, X_n\right), \tag{47}$$

where $X_n = \sum_{s=1}^n \log(r_{1,s}) - n\log(\kappa_1)$. Note that $X_n \sim \mathrm{Erlang}(n, \alpha_1)$ from Facts 13 and 14. Let $\tilde{\alpha}_1^{\mathrm{one}}$ be a sample from the posterior distribution in (47). Then, for one-dimensional Pareto bandits, it holds from (11) that

$$\mathbb{P}[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon | \mathcal{F}_t] = \mathbb{P}\left[\tilde{\alpha}_1^{\mathrm{one}} \geq \beta \,|\, \mathcal{F}_t\right] = \frac{\Gamma\left(n-k+1, \beta X_n\right)}{\Gamma(n-k+1)},$$

where we denoted $\beta = \frac{\mu_1-\epsilon}{\mu_1-\epsilon-\kappa_1}$ satisfying $\mu(\kappa_1, \beta) = \mu_1 - \epsilon$. Therefore, Lemma 23 can be written as

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left[j(t) \neq 1, \mathcal{M}_\epsilon^c(t)\right]\right] &= \sum_{t=1}^{T}\sum_{n=1}^{T} \mathbb{E}\left[\mathbb{1}\left[j(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n\right]\right] \\
&= \sum_{t=1}^{T}\sum_{n=1}^{T} \mathbb{E}\left[\mathbb{P}\left[j(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n \,|\, \mathcal{F}_t\right]\right] \\
&= \sum_{t=1}^{T}\sum_{n=1}^{T} \int_0^\infty \frac{\Gamma(n+1, \beta x)}{\Gamma(n+1)} \frac{\alpha_1^n}{\Gamma(n)} x^{n-1} e^{-\alpha_1 x} \mathrm{d}x \leq \mathcal{O}(\epsilon^{-1}),
\end{aligned}$$

where we injected the density function of the Erlang distribution into the last equality.

On the other hand, for two-parameter Pareto bandits where the scale parameter is unknown, it holds by the law of total expectation that

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)]] &= \mathbb{E}_{\hat{\kappa}_1,\hat{\alpha}_1}\left[\mathbb{P}[j(t) \neq 1, \mathcal{K}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t) | \mathcal{F}_t]\right] \\
&= \mathbb{E}_{\hat{\kappa}_1,\hat{\alpha}_1}\left[\mathbb{1}[\mathcal{K}_{1,N_1(t)}(\epsilon)]\mathbb{P}[j(t) \neq 1, \mathcal{M}_\epsilon^c(t) | \mathcal{F}_t]\right],
\end{aligned}$$

where the last equality holds since $\mathcal{K}$ is determined by the history $\mathcal{F}_t$.

From Lemma 9 with $y = \mu_1 - \epsilon$, it holds for any $\xi \leq \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \kappa_1} = \beta$ that

$$\begin{aligned}\mathbb{1}[\mathcal{K}_{1,n}(\epsilon)]\mathbb{P}[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon|\mathcal{F}_t] &\leq \mathbb{1}[\mathcal{K}_{1,n}(\epsilon)]\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \int_1^{\xi} f^{\mathrm{Er}}_{n-k, \frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x + \int_{\xi}^{\infty} f^{\mathrm{Er}}_{n-k, \frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x\right) \\ &\leq \mathbb{1}[\mathcal{K}_{1,n}(\epsilon)]\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \left(1 - \frac{\Gamma\left(n-k, \frac{n}{\alpha_{1,n}}\xi\right)}{\Gamma(n-k)}\right) + \frac{\Gamma\left(n-k, \frac{n}{\alpha_{1,n}}\xi\right)}{\Gamma(n-k)}\right)\end{aligned} \tag{48}$$

which is a convex combination of 1 and $\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n$. Therefore, RHS of (48) increases as $\frac{\Gamma\left(n-k, \frac{n}{\alpha_{1,n}}\xi\right)}{\Gamma(n-k)}$ increases. From the definition of $\Gamma(n, x)$, it holds that $\Gamma(n, x) \geq \Gamma(n, x + y)$ for any positive $y > 0$ and $\Gamma(n+1, x) = n\Gamma(n, x) + x^n e^{-x}$. Since $\frac{n}{\hat{\alpha}_1(n)} \leq \frac{n}{\alpha_1(n)}$ holds for any $n \in \mathbb{N}$, it holds for $k \in \mathbb{Z}_{\geq 0}$ that

$$\frac{\Gamma\left(n-k, \frac{n}{\hat{\alpha}_1(n)}\xi\right)}{\Gamma(n-k)} \leq \frac{\Gamma\left(n-k, \frac{n}{\alpha_1(n)}\xi\right)}{\Gamma(n-k)} \leq \frac{\Gamma\left(n, \frac{n}{\alpha_1(n)}\xi\right)}{\Gamma(n)}.$$

Let us denote $Y_n := \frac{n}{\hat{\alpha}_1(n)} = \sum_{i=1}^n \log(r_{1,s}) - n\log(\hat{\kappa}_1(n))$, which follows the Erlang distribution with shape $n-1$ and rate α_1 (Malik, 1970). By taking expectation with respect to $\hat{\kappa}_1(n)$, we have for any $\xi \leq \beta$ that

$$\begin{aligned}\mathbb{E}_{\hat{\kappa}_1}[\mathbb{1}[\mathcal{K}_{1,n}(\epsilon)]\mathbb{P}[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon|\mathcal{F}_t]] &\leq \int_{\kappa_1}^{\kappa_1+\epsilon} \left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \left(1 - \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)}\right) + \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)}\right)\mathbb{P}[\hat{\kappa}_1(n) = x]\mathrm{d}x \\ &= \mathbb{P}[\mathcal{K}_{1,n}(\epsilon)]\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \left(1 - \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)}\right) + \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)}\right) \\ &= \left(1 - \left(\frac{\kappa_1}{\kappa_1 + \epsilon}\right)^{n\alpha_1}\right)\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \left(1 - \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)}\right) + \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)}\right),\end{aligned}$$

where we used $\hat{\kappa}_1(n) \sim \mathrm{Pa}(\kappa_1, n\alpha_1)$ in (2) for the last equality.

Therefore, under the two-parameter Pareto distribution, the following holds for any $\xi \leq \beta$ under both STS and STS-T with $k \in \mathbb{Z}_{\geq 0}$ that

$$\begin{aligned}&\mathbb{E}_{\hat{\kappa}_1, \hat{\alpha}_1}[\mathbb{1}[\mathcal{K}_{1,n}(\epsilon)]\mathbb{P}[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon|\mathcal{F}_t]] \\ &\quad\leq \left(1 - \left(\frac{\kappa_1}{\kappa_1 + \epsilon}\right)^{n\alpha_1}\right)\int_0^{\infty}\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \left(1 - \frac{\Gamma(n, \xi y)}{\Gamma(n)}\right) + \frac{\Gamma(n, \xi y)}{\Gamma(n)}\right)\frac{\alpha_1^{n-1}}{\Gamma(n-1)}y^{n-2}e^{-\alpha_1 y}\mathrm{d}y.\end{aligned}$$

Therefore, Lemma 23 concludes the proof for any $n \in \mathbb{N}$, by carefully selecting $\xi \leq \beta$ satisfying

$$\begin{aligned}&\left(1 - \left(\frac{\kappa_1}{\kappa_1 + \epsilon}\right)^{n\alpha_1}\right)\int_0^{\infty}\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi))}\right)^n \left(1 - \frac{\Gamma(n, \xi y)}{\Gamma(n)}\right) + \frac{\Gamma(n, \xi y)}{\Gamma(n)}\right)\frac{\alpha_1^{n-1}}{\Gamma(n-1)}y^{n-2}e^{-\alpha_1 y}\mathrm{d}y \\ &\qquad\leq \int_0^{\infty}\frac{\Gamma(n+1, \beta y)}{\Gamma(n+1)}\frac{\alpha_1^n}{\Gamma(n)}y^{n-1}e^{-\alpha_1 y}\mathrm{d}y.\end{aligned}$$

Note that when we consider STS with $k = -1$, we have to find $\xi' \leq \beta$ such that

$$\begin{aligned}&\left(1 - \left(\frac{\kappa_1}{\kappa_1 + \epsilon}\right)^{n\alpha_1}\right)\int_0^{\infty}\left(\left(\frac{\mu_1 - \epsilon}{\mu((\kappa_1, \xi'))}\right)^n \left(1 - \frac{\Gamma(n+1, \xi' y)}{\Gamma(n+1)}\right) + \frac{\Gamma(n+1, \xi' y)}{\Gamma(n+1)}\right)\frac{\alpha_1^{n-1}}{\Gamma(n-1)}y^{n-2}e^{-\alpha_1 y}\mathrm{d}y \\ &\qquad\leq \int_0^{\infty}\frac{\Gamma(n+1, \beta y)}{\Gamma(n+1)}\frac{\alpha_1^n}{\Gamma(n)}y^{n-1}e^{-\alpha_1 y}\mathrm{d}y d.\end{aligned}$$

From $\Gamma(n, x) \geq \Gamma(n, x + y)$ for any positive $x, y > 0$ and $\xi' \leq \beta$, we have for any $x > 0$ that

$$\frac{\Gamma(n+1, \xi' x)}{\Gamma(n+1)} \geq \frac{\Gamma(n+1, \beta x)}{\Gamma(n+1)}.$$

Therefore, for $k \in \mathbb{Z}_{\leq -1}$, we might not be able to apply the results by Korda et al. (2013). □

C.4. Proof of Lemma 8

We first state two lemmas on the event $\mathcal{K}$ and $\mathcal{A}$.

Lemma 15. *For any algorithm and $a \in [K]$, it holds that for all $\epsilon > 0$, $t > 0$, and $n \in \mathbb{N}$*

$$\mathbb{P}\left[\mathcal{K}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right] \leq \exp\left(-\frac{\alpha_a \epsilon}{\kappa_a + \epsilon} n\right).$$

Lemma 16. *For any algorithm and for any $a \in [K]$, it holds that for all $\epsilon \in \left(0, \frac{\kappa_a}{\alpha_a(\kappa_a+1)}\right)$ and $t > 0$, and $n \geq \bar{n}$*

$$\mathbb{P}\left[\mathcal{A}^c_{a,N_a(t)}(\epsilon), \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n\right] \leq 2\exp\left(-\frac{\alpha_a^2 \epsilon^2}{4} n\right),$$

Lemma 8. *Under* STS *and* STS-T *with $k \in \mathbb{Z}$, it holds that for any $a \neq 1$*

$$\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) = a, \mathcal{E}^c_{a,N_a(t)}(\epsilon)]\right] \leq \mathcal{O}\left(\epsilon^{-2}\right).$$

Proof. From the Lemmas 15 and 16, one can see that for $n \geq \bar{n}$,

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right] &= \mathbb{P}\left[\mathcal{K}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right] + \mathbb{P}\left[\mathcal{A}^c_{a,N_a(t)}(\epsilon), \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n\right] \\ &\leq \exp\left(-\frac{\alpha_a \epsilon}{\kappa_a + \epsilon} n\right) + 2\exp\left(-\frac{\alpha_a^2 \epsilon^2}{4} n\right).\end{aligned}$$

Since $\left\{j(t) = a, \mathcal{E}^c_{a,n}(\epsilon), N_a(t) = n\right\}$ occurs only once for any $n \in \mathbb{N}$, it holds that

$$\begin{aligned}\sum_{t=\bar{n}K+1}^{T} \mathbb{E}\left[\mathbb{1}\left[j(t) = a, \mathcal{E}^c_{a,N_a(t)}(\epsilon)\right]\right] &= \sum_{t=\bar{n}K+1}^{T} \sum_{n=\bar{n}}^{T} \mathbb{E}\left[\mathbb{1}\left[j(t) = a, \mathcal{E}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right]\right] \\ &\leq \sum_{n=\bar{n}}^{\infty} \mathbb{E}\left[\mathbb{1}\left[\mathcal{E}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right]\right] \\ &= \sum_{n=\bar{n}}^{\infty} \mathbb{P}\left[\mathcal{K}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right] + \mathbb{P}\left[\mathcal{A}^c_{a,N_a(t)}(\epsilon) \cap \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n\right] \\ &\leq \sum_{n=\bar{n}}^{\infty} \exp\left(-\frac{\alpha_a \epsilon}{\kappa_a + \epsilon} n\right) + 2\exp\left(-\frac{\alpha_a^2 \epsilon^2}{4} n\right).\end{aligned}$$

Since $\exp(-an)$ with $a > 0$ is a decreasing function with respect to n, it holds that

$$\sum_{n=2}^{\infty} \exp(-an) \leq \int_1^{\infty} \exp(-an)\mathrm{d}n = \frac{\exp(-a)}{a},$$

which concludes the proof. □

C.4.1. Proof of Lemma 15

Lemma 15. *For any algorithm and $a \in [K]$, it holds that for all $\epsilon > 0$, $t > 0$, and $n \in \mathbb{N}$*

$$\mathbb{P}\left[\mathcal{K}^c_{a,N_a(t)}(\epsilon), N_a(t) = n\right] \leq \exp\left(-\frac{\alpha_a \epsilon}{\kappa_a + \epsilon} n\right).$$

Proof. Since $\hat{\kappa}_a(n) \sim \mathrm{Pa}(\kappa_a, n\alpha_a)$ holds for any $n \in \mathbb{N}$ in (2), it holds that

$$\begin{aligned}\mathbb{P}\left[\mathcal{K}^c_{a,N_a(t)}, N_a(t) = n\right] &= \mathbb{P}\left[\hat{\kappa}_a(N_a(t)) \geq \kappa_a + \epsilon, N_a(t) = n\right] \\ &= \left(\frac{\kappa_a}{\kappa_a + \epsilon}\right)^{n\alpha_a} \leq \exp\left(-\frac{\alpha_a \epsilon}{\kappa_a + \epsilon} n\right),\end{aligned}$$

which concludes the proof. □

C.4.2. PROOF OF LEMMA 16

Lemma 16. *For any algorithm and for any $a \in [K]$, it holds that for all $\epsilon \in \left(0, \frac{\kappa_a}{\alpha_a(\kappa_a+1)}\right)$ and $t > 0$, and $n \geq \bar{n}$*

$$\mathbb{P}\left[\mathcal{A}^c_{a,N_a(t)}(\epsilon), \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n\right] \leq 2\exp\left(-\frac{\alpha_a^2\epsilon^2}{4}n\right),$$

Proof. Fix a time index t and denote $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid \mathcal{F}_t]$ and $N_a(t) = n$. To simplify notations, we drop the argument n of $\hat{\kappa}_a(n)$ and $\hat{\alpha}_a(n)$.

Let $r'_{a,k}$ be the k-th order statistics of $(r_{a,s})_{s=1}^n$ for arm a such that $r'_{a,1} \leq r'_{a,2} \ldots \leq r'_{a,n}$. From the definition of MLE of α_a,

$$\begin{aligned}
\mathbb{P}[\hat{\alpha}_a \leq \alpha_a - \epsilon_{a,l}(\epsilon), \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n] &\leq \mathbb{P}\left[\frac{n}{\sum_{s=1}^n \log r'_{a,s} - n\log r'_{a,1}} \leq \alpha_a - \epsilon_{a,l}(\epsilon)\right] \\
&= \mathbb{P}\left[\frac{n}{\alpha_a - \epsilon_{a,l}(\epsilon)} \leq \sum_{s=1}^n \log \frac{r'_{a,s}}{r'_{a,1}}\right] \\
&= \mathbb{P}\left[\frac{n}{\alpha_a - \epsilon_{a,l}(\epsilon)} \leq n\log\frac{\kappa}{r'_{a,1}} + \sum_{s=1}^n \log\frac{r'_{a,s}}{\kappa}\right] \\
&\leq \mathbb{P}\left[\frac{n}{\alpha_a - \epsilon_{a,l}(\epsilon)} \leq \sum_{s=1}^n \log\frac{r_{a,s}}{\kappa_a}\right] \\
&\leq \mathbb{P}\left[\epsilon \leq \frac{1}{n}\sum_{s=1}^n \log\frac{r_{a,s}}{\kappa_a} - \frac{1}{\alpha_a}\right],
\end{aligned}$$

where the first equality holds from the definition of MLEs in (2), the second inequality holds since any sample generated from the Pareto distribution cannot be smaller than its scale parameter κ, and the last inequality holds from the definition of $\epsilon_{a,l}(\epsilon)$ in (13).

Similarly, one can derive that

$$\begin{aligned}
\mathbb{P}[\hat{\alpha}_a \geq \alpha_a + \epsilon_{a,u}(\epsilon), \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n] &\leq \mathbb{P}\left[\sum_{s=1}^n \log\frac{r_{a,s}}{\kappa_a} \leq \frac{n}{\alpha_a + \epsilon_{a,u}(\epsilon)} + n\log\frac{r'_1}{\kappa} \cap \mathcal{K}\right] \\
&\leq \mathbb{P}\left[\sum_{s=1}^n \log\frac{r_{a,s}}{\kappa} \leq \frac{n}{\alpha_a + \epsilon_{a,u}(\epsilon)} + n\log\frac{\kappa_a+\epsilon}{\kappa_a}\right] \\
&\leq \mathbb{P}\left[\sum_{s=1}^n \log\frac{r_{a,s}}{\kappa_a} \leq \frac{n}{\alpha_a + \epsilon_{a,u}(\epsilon)} + \frac{n\epsilon}{\kappa_a}\right] \\
&\leq \mathbb{P}\left[\frac{1}{n}\sum_{s=1}^n \log\frac{r_{a,s}}{\kappa_a} - \frac{1}{\alpha_a} \leq -\epsilon\right],
\end{aligned}$$

where the second inequality holds since $r'_{a,1} = \hat{\kappa}_a \leq \kappa_a + \epsilon$ holds on $\mathcal{K}_{a,n}$, the third inequality from $\log(1+x) \leq x$ for $x > -1$, and the last inequality comes from the definition of $\epsilon_{a,u}(\epsilon)$. From Fact 13, $y_{a,s} := \log\left(\frac{r_{a,s}}{\kappa_a}\right) \sim \mathrm{Exp}(\alpha_a)$ and the last probability can be considered as a deviation of the sum of exponentially distributed random variables.

For the exponential distribution $\mathrm{Exp}(\alpha)$, we say that Bernstein's condition with parameter b holds if

$$\mathbb{E}\left[M_k\right] \leq \frac{1}{2}k!\frac{1}{\alpha^2}b^{k-2} \quad \text{for } k = 3, 4, \ldots,$$

where M_k implies the k-th central moment. For $\mathrm{Exp}(\alpha_a)$, it holds that

$$\mathbb{E}\left[M_k\right] = \frac{!k}{\alpha_a^k} \leq \frac{k!}{2}\frac{1}{\alpha_a^2}\left(\frac{1}{\alpha_a}\right)^{k-2},$$

where $!k$ is the subfactorial of k such that $!k \leq \frac{k!}{e} + \frac{1}{2} \leq \frac{k!}{2}$ for $k \geq 3$. Hence, the exponential distribution with parameter α_a satisfies Bernstein's condition with parameter $\frac{1}{\alpha_a}$, so that it is subexponential with parameters $\left(\frac{2}{\alpha_a^2}, \frac{2}{\alpha_a}\right)$. Therefore, by applying Lemma 20, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{s=1}^{n} y_{a,s} - \frac{1}{\alpha_a}\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n}{4}\min\{\alpha_a^2\epsilon^2, \alpha_a\epsilon\}\right).$$

Note that it holds for $\epsilon < \frac{\kappa_a}{\alpha_a(\kappa_a+1)}$ that

$$\mathbb{P}[\hat{\alpha}_a \leq \alpha_a - \epsilon_{a,l}(\epsilon) \cap \mathcal{K}_{a,n}] \leq \mathbb{P}\left(\frac{1}{n}\sum_{s=1}^{n} y_{a,s} - \frac{1}{\alpha_a} \geq \epsilon\right)$$
$$\mathbb{P}[\hat{\alpha}_a \geq \alpha_a + \epsilon_{a,u}(\epsilon) \cap \mathcal{K}_{a,n}] \leq \mathbb{P}\left(\frac{1}{n}\sum_{s=1}^{n} y_{a,s} - \frac{1}{\alpha_a} \leq -\epsilon\right),$$

for $\epsilon_{a,l}(\epsilon) = \frac{\epsilon\alpha_a^2}{1+\epsilon\alpha_a}$ and $\epsilon_{a,u}(\epsilon) = \frac{\epsilon\alpha_a^2(\kappa_a+1)}{\kappa_a - \epsilon\alpha_a(\kappa_a+1)}$, which satisfy $\lim_{\epsilon\to 0}\max\{\epsilon_{a,l}(\epsilon), \epsilon_{a,u}(\epsilon)\} = 0_+$. Hence, by recovering the original notations, we obtain

$$\begin{aligned}\mathbb{P}[\mathcal{A}^c_{a,N_a(t)}(\epsilon), \mathcal{K}_{a,N_a(t)}(\epsilon), N_a(t) = n] &= \mathbb{P}[\hat{\alpha}_a(n) \leq \alpha_a - \epsilon_{a,l}(\epsilon), \mathcal{K}_{a,N_a(t)}, N_a(t) = n] \\ &\quad + \mathbb{P}[\hat{\alpha}_a(n) \geq \alpha_a + \epsilon_{a,u}(\epsilon), \mathcal{K}_{a,N_a(t)}, N_a(t) = n] \\ &\leq 2\exp\left(-\frac{\alpha_a^2\epsilon^2}{4}n\right),\end{aligned}$$

for $\epsilon < \frac{1}{\alpha_a}$ with $\alpha_a > 1$. □

C.5. Proof of Lemma 9

Before beginning the proof, we first provide the intermediate equation for the probability that the sampled mean is less than the given value, which is used several times in the proof.

Lemma 17. *For any arm $a \in [K]$ and fixed $t \in \mathbb{N}$, let $N_a(t) = n$. For any prior parameter $k \in \mathbb{Z}$, if $y \geq \hat{\kappa}_a(n)$, then*

$$\mathbb{P}[\tilde{\mu}_a(t) \leq y | \theta_{a,n}] = \int_1^{\frac{y}{y-\hat{\kappa}_a(n)}} f^{Er}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}_a(n)x}y\right)^{nx}\mathrm{d}x + \int_{\frac{y}{y-\hat{\kappa}_a(n)}}^{\infty} f^{Er}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x.$$

Proof. Fix a time index t with $N_a(t) = n$ and denote $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid \mathcal{F}_t]$. To simplify notations, we drop the argument n of $\hat{\kappa}_a(n)$ and the argument t of $\tilde{\kappa}_a(t)$, $\tilde{\alpha}_a(t)$, and $\tilde{\mu}_a(t)$.

When $\hat{\kappa}_a < y$ holds, $\tilde{\mu}_a \leq y$ can hold regardless of the value of $\tilde{\kappa}_a$ if $\hat{\kappa}_a\frac{\tilde{\alpha}_a}{\tilde{\alpha}_a - 1} \leq y$ holds since $\tilde{\kappa}_a \in (0, \hat{\kappa}_a]$ holds from its posterior distribution in (9). Hence, if $\hat{\kappa}_a < y$, then

$$\tilde{\alpha}(t) \geq \frac{y}{y-\hat{\kappa}_a} \implies \tilde{\mu}_a \leq y. \tag{49}$$

When $1 < \tilde{\alpha}_a < \frac{y}{y-\hat{\kappa}_a}$,

$$\tilde{\mu}_a = \tilde{\kappa}_a\frac{\tilde{\alpha}_a}{\tilde{\alpha}_a - 1} \leq y \Leftrightarrow \tilde{\kappa}_a \leq y\frac{\tilde{\alpha}_a - 1}{\tilde{\alpha}_a}. \tag{50}$$

Since $\tilde{\alpha}_a \leq 1$ implies $\tilde{\mu}_a = \infty$, from (49) and (50), it holds that

$$\begin{aligned}\mathbb{P}_t[\tilde{\mu}_a \leq y] &= \int_1^{\frac{y}{y-\hat{\kappa}_a}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathbb{P}_t\left[\tilde{\kappa}_a \leq \frac{x-1}{x}y\right]\mathrm{d}x + \int_{\frac{y}{y-\hat{\kappa}_a}}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \\ &= \int_1^{\frac{y}{y-\hat{\kappa}_a}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}_a x}y\right)^{nx}\mathrm{d}x + \int_{\frac{y}{y-\hat{\kappa}_a}}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x,\end{aligned} \tag{51}$$

where we recovered the CDF in (9) in (51). □

Lemma 9. *For any arm $a \in [K]$ and fixed $t \in \mathbb{N}$, let $N_a(t) = n$. For any positive $\xi \leq \frac{y}{y-\kappa_a}$ and $k \in \mathbb{Z}$, it holds that*

$$\mathbb{1}[\hat{\kappa}_a(n) \leq y]\mathbb{P}[\tilde{\mu}_a(t) \leq y|\theta_{a,n}] \leq \int_{\xi}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x + \left(\frac{y}{\mu((\kappa_a,\xi))}\right)^n \int_1^{\xi} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x,$$

where $f^{\mathrm{Er}}_{s,\beta}(\cdot)$ denotes the probability density function of the Erlang distribution with shape $s \in \mathbb{N}$ and rate $\beta > 0$.

Proof. From Lemma 17, if $y \geq \hat{\kappa}_a(n)$, it holds that

$$\mathbb{P}[\tilde{\mu}_a \leq y|\theta_{a,n}] = \int_1^{\frac{y}{y-\hat{\kappa}_a}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}_a x}y\right)^{nx}\mathrm{d}x + \int_{\frac{y}{y-\hat{\kappa}_a}}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x.$$

Take any finite $y' > y$ and let $\xi := \frac{y'}{y'-\kappa_a} < \frac{y}{y-\kappa_a}$ such that $\mu((\kappa_a,\xi))) = y'$. Since $\frac{a}{a-b}$ is decreasing with respect to $a > b > 0$, one can see that

$$\begin{aligned}
\mathbb{1}[\hat{\kappa}_a(n) \leq y]\mathbb{P}[\tilde{\mu}_a \leq y|\theta_{a,n}] &\leq \int_1^{\frac{y}{y-\hat{\kappa}_a}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}_a x}y\right)^{nx}\mathrm{d}x + \int_{\frac{y}{y-\hat{\kappa}_a}}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \\
&\leq \int_1^{\frac{y'}{y'-\hat{\kappa}_a}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}_a x}y\right)^{nx}\mathrm{d}x + \int_{\frac{y'}{y'-\hat{\kappa}_a}}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \\
&\leq \int_1^{\frac{y'}{y'-\kappa_a}} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\hat{\kappa}_a x}y\right)^{nx}\mathrm{d}x + \int_{\frac{y'}{y'-\kappa_a}}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \\
&\leq \int_1^{\xi} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\left(\frac{x-1}{\kappa_a x}y\right)^{n}\mathrm{d}x + \int_{\xi}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \qquad (52) \\
&\leq \left(\frac{\xi-1}{\kappa_a \xi}y\right)^n \int_1^{\xi} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x + \int_{\xi}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x \qquad (53) \\
&= \left(\frac{y}{\mu((\kappa_a,\xi))}\right)^n \int_1^{\xi} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x + \int_{\xi}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{a,n}}}(x)\mathrm{d}x,
\end{aligned}$$

where (52) comes from $\hat{\kappa}_a \geq \kappa_a$ and we used the increasing property of $\frac{x-1}{x}$ in (53). □

C.6. Proof of Lemma 10

Lemma 10. *Given $\epsilon > 0$, define a positive problem-dependent constant $\rho = \rho_{\theta_1}(\epsilon) := \frac{\kappa_1\epsilon}{2(\mu_1-\epsilon/2-\kappa_1)(\mu_1-\kappa_1)}$. If $n \geq \bar{n} = \max(2, k+1)$, then for $k \in \mathbb{Z}_{\geq 0}$*

$$p_n(\epsilon|\theta_{1,n}) \leq \begin{cases} e^{-n}, & \text{if } \hat{\kappa}_1(n) \geq \mu_1 - \epsilon, \\ h(\mu_1,\epsilon,n), & \text{if } \hat{\kappa}_1(n) \in K(\epsilon), \alpha_{1,n} \leq \alpha_1 + \rho, \\ C_1(\mu_1,\epsilon,n)G_k(1/\alpha_{1,n};n) + 1 - G_k(1/\alpha_{1,n};n) & \text{if } \hat{\kappa}_1(n) \in K(\epsilon), \alpha_{1,n} \geq \alpha_1 + \rho, \end{cases}$$

where

$$\begin{aligned}
h(\mu_1,\epsilon,n) &:= e^{-n\frac{3\epsilon}{4\mu_1}}\left(1 - \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}\right) + \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)} \\
C_1(\mu_1,\epsilon,n) &:= \left(\frac{\mu_1-\epsilon}{\mu_1-\epsilon/2}\right)^n \leq e^{-n\frac{\epsilon}{2\mu_1-\epsilon}} < 1 \\
G_k(x;n) &:= F^{\mathrm{Er}}_{n-k,nx}(\alpha_1+\rho)
\end{aligned}$$

for F^{Er} defined in (11). Here, $c_{\mu_1}(\epsilon) = \zeta - \log(1+\zeta) = \mathcal{O}(\epsilon^{-2})$ and $\zeta = \frac{\epsilon}{4\mu_1-2\epsilon} \in (0,1)$ are deterministic constants of μ_1 and ϵ.

Proof. Similarly to other proofs, fix t and let $N_1(t) = n$. To simplify notations, we drop the argument t of $\tilde{\kappa}_1(t), \tilde{\alpha}_1(t)$ and $\tilde{\mu}_1(t)$ and the argument n of $\hat{\kappa}_1(n), \hat{\alpha}_1(n), \bar{\alpha}_1(n)$.

Case 1. On $\{\hat{\kappa}_1 \geq \mu_1 - \epsilon\}$ Under the condition $\{\hat{\kappa}_1 \geq \mu_1 - \epsilon\}$, the event $\{\tilde{\mu}_1 \leq \mu_1 - \epsilon\}$ is eventually determined by the value of $\tilde{\kappa}$ since $\{\tilde{\kappa}_1 \in (\mu_1 - \epsilon, \hat{\kappa}_1]\}$ is a sufficient condition to $\{\tilde{\mu}_1 > \mu_1 - \epsilon\}$. Therefore, if $\hat{\kappa}_1 \geq \mu_1 - \epsilon$, then

$$p_n(\epsilon|\theta_{1,n}) = \int_1^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\left(\frac{\mu_1-\epsilon}{\hat{\kappa}_1}\frac{x-1}{x}\right)^{nx}\mathrm{d}x.$$

Then,

$$\begin{aligned}
\mathbb{1}[\hat{\kappa}_1 \geq \mu_1 - \epsilon]p_n(\epsilon|\theta_{1,n}) &= \mathbb{1}[\hat{\kappa}_1 \geq \mu_1 - \epsilon]\left(\int_1^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\left(\frac{\mu_1-\epsilon}{\hat{\kappa}_1}\frac{x-1}{x}\right)^{nx}\mathrm{d}x\right)\\
&\leq \mathbb{1}[\hat{\kappa}_1 \geq \mu_1 - \epsilon]\int_1^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\left(1-\frac{1}{x}\right)^{nx}\mathrm{d}x\\
&\leq \mathbb{1}[\hat{\kappa}_1(n) \geq \mu_1 - \epsilon]\int_1^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)e^{-n}\mathrm{d}x\\
&\leq \mathbb{1}[\hat{\kappa}_1(n) \geq \mu_1 - \epsilon]e^{-n},
\end{aligned}$$

where the second inequality holds from $\mu_1 - \epsilon \leq \hat{\kappa}_1$.

Case 2. On $\{\hat{\kappa}_1 \in K(\epsilon), \alpha_{1,n} \leq \alpha_1 + \rho\}$ By applying Lemma 9 with $y = \mu_1 - \epsilon$, we have for any $\xi \leq \frac{\mu_1-\epsilon}{\mu_1-\epsilon-\kappa_1}$ that

$$\begin{aligned}
\mathbb{1}[\hat{\kappa}_1 < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha + \rho]p_n(\epsilon|\theta_{1,n}) &\leq \left(\frac{\mu_1-\epsilon}{\mu((\kappa_1,\xi))}\right)^n\int_1^\xi f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x + \int_\xi^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x\\
&\leq \left(\frac{\mu_1-\epsilon}{\mu((\kappa_1,\xi))}\right)^n\int_0^\xi f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x + \int_\xi^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x. \qquad (54)
\end{aligned}$$

Let us define $\bar{\rho} := \rho_\theta(\epsilon/2)$. Then, it satisfies $\mu((\kappa_1, \alpha_1 + \bar{\rho})) = \mu_1 - \frac{\epsilon}{4}$ and

$$\alpha_1 + \bar{\rho} = \frac{\mu - \epsilon/4}{\mu - \epsilon/4 - \kappa_1} < \frac{\mu-\epsilon}{\mu-\epsilon-\kappa_1},$$

where the inequality holds from the decreasing property of the function $\frac{x}{x-y}$ with respect to $x > y$. By replacing ξ with $\alpha_1 + \bar{\rho}$ in (54), we have

$$\begin{aligned}
&\mathbb{1}[\hat{\kappa}_1 < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho]p_n(\epsilon|\bar{\theta}_{1,n})\\
&\leq \mathbb{1}[\hat{\kappa}_1 < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho]\left(\left(\frac{\mu_1-\epsilon}{\mu_1-\epsilon/4}\right)^n\int_0^\xi f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x + \int_{\alpha_1+\bar{\rho}}^\infty f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n^{(n)}}}}(x)\mathrm{d}x\right)\\
&\leq \mathbb{1}[\hat{\kappa}_1 < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho]\left(e^{-n\left(\frac{3\epsilon}{4\mu_1-\epsilon}\right)}\left(1 - \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}]\right) + \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}]\right). \qquad (55)
\end{aligned}$$

Let Z_n be a random variable that follows the chi-squared distribution with n degree of freedom and $F_n(\cdot)$ denote the CDF of Z_n. Then, it holds that

$$\begin{aligned}
\mathbb{P}\left[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho\right] &= \mathbb{P}\left[Z \geq \frac{2n}{\alpha_{1,n}}(\alpha_1 + \bar{\rho}), \alpha_{1,n} \leq \alpha_1 + \rho\right] \quad \text{by Fact 12}\\
&\leq \mathbb{P}\left[Z \geq 2n\frac{\alpha_1+\bar{\rho}}{\alpha_1+\rho}\right]\\
&\leq \mathbb{P}\left[Z \geq 2n\frac{\mu_1-\epsilon/4}{\mu_1-\epsilon/2}\right]\\
&= 1 - F_{2n-2k}(2n(1+\zeta)), \qquad (56)
\end{aligned}$$

where $\zeta = \frac{\epsilon}{4\mu_1-2\epsilon} \in (0,1)$. By applying Lemma 21, we have if $n\zeta > -k$,

$$\begin{aligned}
\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] &\leq 1 - F_{2n-2k}\left(2n(1+\zeta)\right)\\
&< \frac{1}{2}\frac{\sqrt{2\pi}(n-k)^{n-k-1/2}e^{-(n-k)}}{\Gamma(n-k)}\mathrm{erfc}\left(\sqrt{n(\zeta+k) - (n-k)\log\frac{n(1+\zeta)}{n-k}}\right),
\end{aligned}$$

where $\Gamma(\cdot)$ denotes the Gamma function. For $n \geq 1/2$, it holds from Stirling's formula that

$$\sqrt{2\pi}n^{n-1/2}e^{-n} \leq \Gamma(n) \leq \sqrt{2\pi}e^{1/6}n^{n-1/2}e^{-n},$$

which results in

$$\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] < \frac{1}{2}\mathrm{erfc}\left(\sqrt{n(\zeta+k) - (n-k)\log\frac{n(1+\zeta)}{n-k}}\right). \tag{57}$$

Notice that $(n-k)\log\frac{n(1+\zeta)}{n-k} > 0$ always holds from the assumption of $n\zeta > -k$ where priors with $k \in \mathbb{Z}_{\geq 0}$ satisfies regardless of n. Thus, if $\zeta + k \leq 0$, then the argument in the complementary error function in (57) becomes negative. This makes the upper bound in (57) greater than or equal to $\frac{1}{2}$. Therefore, for the priors with $k \in \mathbb{Z}_{<0}$, the right term in (57) is bounded by a constant since $\zeta \in (0,1)$.

Since the complementary error function is a decreasing function, for priors with $k \in \mathbb{Z}_{\geq 0}$, it holds from (57) that

$$\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] \leq \frac{1}{2}\mathrm{erfc}\left(\sqrt{n(\zeta - \log(1+\zeta)}\right),$$

where we substitute $k = 0$. By the change of variables, the complementary error function is bounded for any $x \geq 0$ as follows (Simon & Divsalar, 1998):

$$\mathrm{erfc}(x) \leq e^{-x^2},$$

which implies

$$\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] \leq \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}, \tag{58}$$

where $c_{\mu_1}(\epsilon) = \zeta - \log(1+\zeta) > 0$ is a deterministic constants on μ_1 and ϵ. By combining (58) with (55), we have

$$\mathbb{1}[\hat{\kappa}_1 < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho]p_n(\epsilon|\theta_{1,n}) \leq e^{-n\frac{3\epsilon}{4\mu_1}}\left(1 - \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}\right) + \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)} =: h(\mu_1, \epsilon, n).$$

From the power-series expansion of $\log(1+x)$, we have $\log(1+x) \geq x - \frac{x^2}{2} + \frac{x^3}{3}$ for $x \in (0,1)$ and

$$\begin{aligned} c_{\mu_1}(\epsilon) = \zeta - \log(1+\zeta) &\leq \frac{\zeta^2}{2} - \frac{\zeta^3}{3} = \frac{\zeta^2}{6}(3 - 2\zeta) \\ &\leq \frac{\zeta^2}{2} = \mathcal{O}(\epsilon^{-2}). \end{aligned}$$

Case 3. On $\{\hat{\kappa}_1 \in K(\epsilon), \alpha_{1,n} \geq \alpha_1 + \rho\}$ By applying Lemma 9 with $y = \mu_1 - \epsilon$ and $\xi = \alpha_1 + \rho$, we have

$$\begin{aligned} \mathbb{1}[\hat{\kappa}_1 < \mu_1 - \epsilon]p_n(\epsilon|\theta_{1,n}) &\leq \left(\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon/2}\right)^n \int_1^{\alpha_1+\rho} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x + \int_{\alpha_1+\rho}^{\infty} f^{\mathrm{Er}}_{n-k,\frac{n}{\alpha_{1,n}}}(x)\mathrm{d}x \\ &= C_1(\mu_1, \epsilon, n)\mathbb{P}\left[\tilde{\alpha}_1 \in [1, \alpha_1 + \rho] \mid \alpha_{1,n}\right] + \mathbb{P}\left[\tilde{\alpha}_1 \geq \alpha_1 + \rho \mid \alpha_{1,n}\right] \\ &\leq C_1(\mu_1, \epsilon, n)\mathbb{P}\left[\tilde{\alpha}_1 \leq \alpha_1 + \rho \mid \alpha_{1,n}\right] + \mathbb{P}\left[\tilde{\alpha}_1 \geq \alpha_1 + \rho \mid \alpha_{1,n}\right], \end{aligned}$$

where $C_1(\mu_1, \epsilon, n) := \left(\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon/2}\right)^n \leq e^{-n\frac{\epsilon}{2\mu_1 - \epsilon}} < 1$. Since $\tilde{\alpha}_1$ follows Erlang $\left(n-k, \frac{n}{\alpha_{1,n}}\right)$, it holds that

$$\mathbb{P}\left[\tilde{\alpha}_1 \leq \alpha_1 + \rho \mid \alpha_{1,n}\right] = \frac{\gamma\left(n-k, \frac{n(\alpha_1+\rho)}{\alpha_{1,n}}\right)}{\Gamma(n-k)},$$

where $\gamma(\cdot,\cdot)$ denotes the lower incomplete gamma function. Therefore, letting

$$G_k(x; n) := \frac{\gamma\left(n-k, n(\alpha_1+\rho)x\right)}{\Gamma(n-k)}$$

concludes the proof. □

D. Proof of Theorem 3

As shown in proofs of Theorem 2, the integral term in (27) diverges for $k \in \mathbb{Z}_{\leq 1}$ without the restriction on $\hat{\alpha}$. In this section, we provide the partial proof of Theorem 3 for $k \in \mathbb{Z}_{\leq 0}$, which shows the necessity of such requirement to achieve asymptotic optimality.

Proof of Theorem 3. Recall that STS starts from playing every arms twice for priors $k \leq 1$, i.e., $N_a(s) \geq 2$ holds for all $a \in \{1, 2\}$. We have for $T \geq 5$

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}(T)] &= \Delta_2 \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[j(t) = 2]\right] \\ &\geq \Delta_2 \mathbb{E}\left[\sum_{t=5}^{T} \mathbb{1}[j(t) = 2, N_1(t) = 2]\right].\end{aligned}$$

From the definition of $N_1(\cdot)$, $\{j(s) \neq 2, N_1(s) = 2\}$ implies $N_1(t) > 2$ for $t > s$. Therefore, for any $t \geq 5$,

$$\begin{aligned}\{j(t) = 2, N_1(t) = 2\} &\Leftrightarrow \{\forall s \in [1, t-4] : j(s+4) = 2\} \\ &\Leftrightarrow \{\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2\}.\end{aligned}$$

Let $T' = T - 4$, then we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=5}^{T} \mathbb{1}[j(t) = 2, N_1(t) = 2]\right] &= \mathbb{E}\left[\sum_{t=5}^{T} \mathbb{1}\left[\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2\right]\right] \\ &= \mathbb{E}\left[\sum_{s=1}^{T'} \left(\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1(2), \hat{\alpha}_1(2)]\right)^s\right]. \qquad (59)\end{aligned}$$

Notice that the above discussion is applicable to any bandit instance.

Then, let us consider a two-armed bandit problem with $\theta_1 = (\kappa, \alpha_1)$ and $\theta_2 = (\kappa, \alpha_2)$. Let $\gamma = \max\{\lceil \alpha_2 \rceil, \lfloor \alpha_2 \rfloor + 1\}$ and $m = \frac{\gamma}{\gamma - 1}$, so that $\frac{\mu_2}{m} = \kappa \frac{\alpha_2}{\alpha_2 - 1} \frac{\gamma - 1}{\gamma} > \kappa$. Assume $1 < \alpha_1 < \alpha_2$ and $\tilde{\mu}_2(s) = \mu_2 = \kappa \frac{\alpha_2}{\alpha_2 - 1}$ for all $s \in \mathbb{N}$.

Notice that $\hat{\kappa}_1(N_1(s)) = \hat{\kappa}_1(2)$ and $\hat{\alpha}_1(N_1(s)) = \hat{\alpha}_1(2)$ hold for all $s \geq 2$ since only $j(s) = 2$ holds for all $s \geq 2$. Here, we first provide the lower bound on $\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1, \hat{\alpha}_1]$. Since $\frac{\mu_2}{m} = \frac{\kappa \alpha_2}{\alpha_2 - 1} \frac{\gamma - 1}{\gamma} > \kappa$ holds, we can consider the case where $\hat{\kappa}_1(2) \leq \frac{\mu_2}{m}$ occurs.

From Lemma 17, it holds for $y \geq \hat{\kappa}_1(n)$ that

$$\begin{aligned}\mathbb{P}[\tilde{\mu}_1 \leq y | \hat{\theta}_{1,n}] &= \int_1^{\frac{y}{y - \hat{\kappa}_1(n)}} f^{\mathrm{Er}}_{n-k, \frac{n}{\hat{\alpha}_1(n)}}(x) \left(\frac{x-1}{\hat{\kappa}_1(n) x} y\right)^{nx} \mathrm{d}x + \int_{\frac{y}{y - \hat{\kappa}_1(n)}}^{\infty} f^{\mathrm{Er}}_{n-k, \frac{n}{\hat{\alpha}_1(n)}}(x) \mathrm{d}x \\ &\geq \int_{\frac{y}{y - \hat{\kappa}_1(n)}}^{\infty} f^{\mathrm{Er}}_{n-k, \frac{n}{\hat{\alpha}_1(n)}}(x) \mathrm{d}x.\end{aligned}$$

By letting $n = 2$ and $y = \mu_2$, we have for $k \in \mathbb{Z}_{\leq 1}$

$$\begin{aligned}\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1(2), \hat{\alpha}_1(2)] &\geq \mathbb{1}\left[\hat{\kappa}_1(n) \leq \frac{\mu_2}{m}\right] \int_{\frac{\mu_2}{\mu_2 - \hat{\kappa}}}^{\infty} f^{\mathrm{Er}}_{2-k, \frac{2}{\hat{\alpha}}}(x) \mathrm{d}x \\ &\geq \mathbb{1}\left[\hat{\kappa} \leq \frac{\mu_2}{m}\right] \int_{\gamma}^{\infty} f^{\mathrm{Er}}_{2-k, \frac{2}{\hat{\alpha}}}(x) \mathrm{d}x \qquad (60) \\ &= \mathbb{1}\left[\hat{\kappa} \leq \frac{\mu_2}{m}\right] \frac{\Gamma(2-k, \frac{2\gamma}{\hat{\alpha}})}{\Gamma(2-k)}, \qquad (61)\end{aligned}$$

where (60) holds from $\frac{\mu_2}{\mu_2 - \hat{\kappa}_1(2)} \leq \frac{\mu_2}{\mu_2 - \mu_2/m} = \frac{m}{m-1} = \gamma = \lceil \alpha_2 \rceil$ and $\Gamma(\cdot, \cdot)$ is the upper incomplete Gamma function. To simplify the notations, we drop the arguments on n and t of $\tilde{\mu}_1(t)$, $\hat{\kappa}_1(n)$, and $\hat{\alpha}_1(n)$ in the following sections.

D.1. Priors $k \in \mathbb{Z}_{\leq 1}$

Note that $\Gamma(n, x)$ is an increasing function with respect to n for fixed x. Therefore, (68) implies that if the lower bound of regret for the reference prior is larger than the lower bound, then every prior with $k \in \mathbb{Z}_{\leq 0}$ are suboptimal. Therefore, let us consider the case $k = 1$, where we can rewrite (68) as

$$\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1, \hat{\alpha}_1] \geq \mathbb{1}\left[\hat{\kappa}_1 \leq \frac{\mu_2}{m}\right] \frac{\Gamma(1, \frac{2\gamma}{\hat{\alpha}_1})}{\Gamma(1)} = e^{-\frac{2\gamma}{\hat{\alpha}_1}}. \tag{62}$$

Since $\hat{\alpha}_1(2) \sim \mathrm{IG}(1, 2\alpha_1)$ in (2), $z := \frac{2\gamma}{\hat{\alpha}}$ follows an exponential distribution with rate parameter α_1/γ, i.e., $z \sim \mathrm{Exp}(\alpha_1/\gamma)$. By combining (62) with (59), we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=5}^{T} \mathbb{1}[j(t) = 2, N_1(t) = 2]\right] &\geq \mathbb{E}_{\hat{\kappa}, z}\left[\sum_{s=1}^{T'} \left(\mathbb{1}[\hat{\kappa} \leq \mu_2/m] e^{-z}\right)^s\right] \\
&= \mathbb{P}[\hat{\kappa} \leq \mu_2/m] \mathbb{E}_{z \sim \mathrm{Exp}(\alpha_1/\gamma)}\left[\sum_{s=1}^{T'} e^{-zs}\right],
\end{aligned} \tag{63}$$

where we used the stochastic independence of $\hat{\alpha}$ and $\hat{\kappa}$. Here,

$$\begin{aligned}
\mathbb{E}_{z \sim \mathrm{Exp}(\alpha_1/\gamma)}\left[\sum_{s=1}^{T'} e^{-zs}\right] &= \mathbb{E}_{z \sim \mathrm{Exp}(\alpha_1/\gamma)}\left[(1 - e^{-zT'}) \frac{e^{-z}}{1 - e^{-z}}\right] \\
&= \int_0^\infty (1 - e^{-xT'}) \frac{e^{-x}}{1 - e^{-x}} e^{-\frac{\alpha_1}{\gamma} x} \mathrm{d}x \\
&\geq \int_0^\infty (1 - e^{-xT'}) \frac{e^{-2x}}{1 - e^{-x}} \mathrm{d}x \qquad \text{by } \frac{\alpha_1}{\gamma} < 1 \\
&\geq \left(1 - \frac{1}{e}\right) \int_{\frac{1}{T'}}^\infty \frac{e^{-2x}}{1 - e^{-x}} \mathrm{d}x \\
&= \left(1 - \frac{1}{e}\right) \left[\log(e^x - 1) - z + e^{-z}\right]_{x = \frac{1}{T'}}^\infty \\
&\geq \left(1 - \frac{1}{e}\right) \left(\log T' + 1 - \frac{3}{2T'}\right),
\end{aligned} \tag{64}$$

where the last inequality holds from its power series expansion

$$\log(e^x - 1) - x + e^{-x} \geq \log(x) + 1 - \frac{3}{2}x$$

and $\lim_{x \to \infty} \log(e^x - 1) - x + e^{-x} = 0$. By combining (64) with (63) and (59) and elementary calculation with $\hat{\kappa}_1(2) \sim \mathrm{Pa}(\kappa_1, 2\alpha_1)$, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &\geq \Delta_2 \left(1 - \left(\frac{m\kappa}{\mu_2}\right)^{2\alpha_1}\right) \left(1 - \frac{1}{e}\right) \left(\log T' + 1 - \frac{3}{2T'}\right) \\
&= \Delta_2 \left(1 - \left(\frac{m\kappa}{\mu_2}\right)^{2\alpha_1}\right) \left(1 - \frac{1}{e}\right) \left(\log(T + 4) + 1 - \frac{3}{2(T + 4)}\right).
\end{aligned}$$

Therefore, under STS with $k \in \mathbb{Z}_{\leq 1}$, there exists a constant $C(\alpha_1, \alpha_2)$ such that

$$\liminf_{T \to \infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geq C(\alpha_1, \alpha_2).$$

D.2. Priors $k \in \mathbb{Z}_{\leq 0}$

Similarly to the last section, it is sufficient to consider the case $k = 0$, where we can rewrite (68) as

$$\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1, \hat{\alpha}_1] \geq \mathbb{1}\left[\hat{\kappa}_1 \leq \frac{\mu_2}{m}\right] \frac{\Gamma(2, \frac{2\gamma}{\hat{\alpha}_1})}{\Gamma(2)}. \tag{65}$$

From the definition of the upper incomplete Gamma function, we have

$$g(z) := \Gamma(2, z) = \int_z^\infty x^1 e^{-x} \mathrm{d}x = e^{-z}(z+1),$$

as a counterpart of e^{-z} in (63) with the same notations $z = \frac{2\gamma}{\hat{\alpha}_1} \sim \mathrm{Exp}\left(\frac{\alpha_1}{\gamma}\right)$.

Therefore, by replacing e^{-zs} in (63) with $g(z)^s$, we have

$$\begin{aligned}
\mathbb{E}_z\left[\sum_{s=1}^{T'}(g(z))^s\right] &\geq \mathbb{E}_z\left[\mathbb{1}[z \in (0,1]]\sum_{s=1}^{T'}(g(z))^s\right] \\
&\geq \mathbb{E}_z\left[\mathbb{1}[z \in (0,1]]\sum_{s=1}^{T'}(1-z^2)^s\right] \\
&= \mathbb{E}_z\left[\mathbb{1}[z \in (0,1]](1-(1-z^2)^{T'})\frac{1-z^2}{z^2}\right],
\end{aligned}$$

where we used the fact $z \in [0,1]$, $g(z) \geq 1 - z^2$ in the second inequality. Since $z \in \left(0, \frac{1}{\sqrt{T'}}\right]$, $(1-z^2)^{T'} \leq \frac{1}{1+T'z^2}$ holds, we have $1 - (1-z^2)^{T'} \geq \frac{T'z^2}{1+T'z^2}$. By applying this fact, we have for $T' > 1$,

$$\begin{aligned}
\mathbb{E}_z\left[\sum_{s=1}^{T'}(g(z))^s\right] &\geq \mathbb{E}_z\left[\frac{T'(1-z^2)}{1+T'z^2}\mathbb{1}\left[z \in \left(0, \frac{1}{\sqrt{T'}}\right]\right]\right] \\
&\geq \mathbb{E}_{z\sim\mathrm{Exp}(\alpha_1/\gamma)}\left[\left(\frac{T'}{2} - \frac{1}{2}\right)\mathbb{1}\left[z \in \left(0, \frac{1}{\sqrt{T'}}\right]\right]\right] \\
&= \left(\frac{T'}{2} - \frac{1}{2}\right)\int_0^{\frac{1}{\sqrt{T'}}} \frac{\alpha_1}{\gamma} e^{-\frac{\alpha_1}{\gamma}z}\mathrm{d}z \\
&= \left(\frac{T'}{2} - \frac{1}{2}\right)\left(1 - e^{-\frac{\alpha_1}{\gamma\sqrt{T'}}}\right).
\end{aligned} \tag{66}$$

Notice that $e^{-x} \leq 1 - \frac{x}{2}$ holds for $x < 1$, which gives

$$\begin{aligned}
\mathbb{E}_z\left[\sum_{s=1}^{T'}(g(z))^s\right] &\geq \left(\frac{T'}{2} - \frac{1}{2}\right)\left(1 - e^{-\frac{\alpha_1}{\gamma\sqrt{T'}}}\right) \\
&\geq \left(\frac{T'}{2} - \frac{1}{2}\right)\frac{\alpha_1}{2\gamma\sqrt{T'}} = \frac{\alpha_1}{4\gamma}\left(\sqrt{T'} - \frac{1}{\sqrt{T'}}\right).
\end{aligned} \tag{67}$$

By applying (67) to (59), we obtain for $k \in \mathbb{Z}_{\leq 0}$ and $T' = T - 4 > 1$,

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &\geq \Delta_2 \frac{\alpha_1}{4\gamma}\left(1 - \left(\frac{m\kappa}{\mu_2}\right)^{2\alpha_1}\right)\left(\sqrt{T'} - \frac{1}{\sqrt{T'}}\right) \\
&= \mathcal{O}(\sqrt{T}).
\end{aligned}$$

Notice that from the definition of $m = \frac{\gamma}{\gamma-1} = \frac{\lceil \alpha_2 \rceil}{\lceil \alpha_2 \rceil - 1}$, $m\frac{\kappa}{\mu_2} = m\left(1 - \frac{1}{\alpha_2}\right) < 1$ holds. Therefore, under STS with priors $k \in \mathbb{Z}_{\leq 0}$, there exists a constant $C'(\alpha_1, \alpha_2) > 0$ such that

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\sqrt{T}} \geq C'(\alpha_1, \alpha_2).$$

□

D.3. Suboptimality for $\kappa_1 < \kappa_2$

In this section, we consider the two-armed Pareto bandit problem where $\kappa_1 < \kappa_2$ holds, indicating that the minimum reward generated from arm 2 is greater than that from arm 1. Under this setting, we present a concrete example where the regret of TS with prior parameter $k \in \mathbb{Z}_{\leq 1}$ is larger than the asymptotic regret lower bound in Lemma 1.

Theorem 18. *Assume that the arm 1 follows $\mathrm{Pa}(\kappa_1, \alpha_1)$ and the arm 2 follows $\mathrm{Pa}(\kappa_2, \alpha_2)$ with $\kappa_1 < \kappa_2$ and $1 < \alpha_1 < \alpha_2$. When $\tilde{\alpha}_1(t)$ and $\tilde{\kappa}_1(t)$ are sampled based on the posteriors in (8) and (9) with prior $k \in \mathbb{Z}_{\leq 1}$, respectively and $\tilde{\mu}_2(t) = \mu_2$ holds, there exists a constant $C(\alpha_1, \kappa_1, \kappa_2) > 0$ independent of α_2 satisfying*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geq C(\alpha_1, \kappa_1, \kappa_2),$$

where $C(\alpha_1) > \frac{\Delta_2}{\mathrm{KL}_{\inf}(2)}$ holds for some instances. In particular, for $k \in \mathbb{Z}_{\leq 0}$, there exists a constant $C'(\alpha_1, \kappa_1, \kappa_2) > 0$ independent of α_2 satisfying

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\sqrt{T}} \geq C'(\alpha_1, \kappa_1, \kappa_2).$$

Proof. Recall that the discussion in the proof of Theorem 3 holds for any bandit instance until (59), which implies

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &\geq \Delta_2 \mathbb{E}\left[\sum_{t=5}^{T} \mathbb{1}[j(t) = 2, N_1(t) = 2]\right] \\
&= \mathbb{E}\left[\sum_{t=5}^{T} \mathbb{1}\left[\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2\right]\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{T-4} \left(\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1(2), \hat{\alpha}_1(2)]\right)^s\right].
\end{aligned}$$

From Lemma 17, if $y \geq \hat{\kappa}_1(n)$, then

$$\begin{aligned}
\mathbb{P}[\tilde{\mu}_1 \leq y | \hat{\theta}_{1,n}] &= \int_1^{\frac{y}{y-\hat{\kappa}_1(n)}} f^{\mathrm{Er}}_{n-k, \frac{n}{\hat{\alpha}_1(n)}}(x) \left(\frac{x-1}{\hat{\kappa}_1(n)x} y\right)^{nx} \mathrm{d}x + \int_{\frac{y}{y-\hat{\kappa}_1(n)}}^{\infty} f^{\mathrm{Er}}_{n-k, \frac{n}{\hat{\alpha}_1(n)}}(x)\mathrm{d}x \\
&\geq \int_{\frac{y}{y-\hat{\kappa}_1(n)}}^{\infty} f^{\mathrm{Er}}_{n-k, \frac{n}{\hat{\alpha}_1(n)}}(x)\mathrm{d}x.
\end{aligned}$$

By letting $n = 2$ and $y = \mu_2$, we have for $k \in \mathbb{Z}_{\leq 1}$

$$\begin{aligned}
\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\kappa}_1(2), \hat{\alpha}_1(2)] &\geq \mathbb{1}\left[\hat{\kappa}_1(2) \leq \kappa_2\right] \int_{\frac{\mu_2}{\mu_2 - \hat{\kappa}_1(2)}}^{\infty} f^{\mathrm{Er}}_{2-k, \frac{2}{\hat{\alpha}_1(2)}}(x)\mathrm{d}x \\
&\geq \mathbb{1}\left[\hat{\kappa}_1(2) \leq \kappa_2\right] \int_{\alpha_2}^{\infty} f^{\mathrm{Er}}_{2-k, \frac{2}{\hat{\alpha}_1(2)}}(x)\mathrm{d}x \qquad \because \alpha_2 = \frac{\mu_2}{\mu_2 - \kappa_2} \geq \frac{\mu_2}{\mu_2 - \hat{\kappa}_1(2)} \\
&= \mathbb{1}\left[\hat{\kappa}_1(2) \leq \kappa_2\right] \frac{\Gamma(2-k, \frac{2\alpha_2}{\hat{\alpha}_1(2)})}{\Gamma(2-k)}, \qquad (68)
\end{aligned}$$

where $\Gamma(\cdot, \cdot)$ is the upper incomplete Gamma function.

By following the same steps in Sections D.1, just replacing γ with α_2, one can obtain for $k \in \mathbb{Z}_{\leq 1}$ that

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \Delta_2 \left(1 - \left(\frac{\kappa_1}{\kappa_2}\right)^{2\alpha_1}\right)\left(1 - \frac{1}{e}\right)\left(\log(T+4) + 1 - \frac{3}{2(T+4)}\right).$$

□

An example of suboptimality Based on the discussion above, it holds for $k \in \mathbb{Z}_{\leq 1}$ that

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geq \Delta_2 \left(1 - \left(\frac{\kappa_1}{\kappa_2}\right)^{2\alpha_1}\right)\frac{e-1}{e}.$$

Recall the closed form of the $\mathrm{KL}_{\inf}(2)$ in Lemma 1, which is

$$\mathrm{KL}_{\inf}(2) = \log\left(\alpha_2 \frac{\mu_1 - \kappa_2}{\mu_1}\right) + \frac{\mu_1}{\alpha_2(\mu_1 - \kappa_2)} - 1.$$

Then, let us consider the case $\theta_1 = (1, 1.01)$ and $\theta_2 = (10, 30)$ where $\mu_1 = 101$ and $\mu_2 = \frac{300}{29}$. In this case, it holds that

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\Delta_2 \log T} \geq \left(1 - \left(\frac{\kappa_1}{\kappa_2}\right)^{2\alpha_1}\right)\frac{e-1}{e} \approx 0.626 > 0.428 \approx \frac{1}{\mathrm{KL}_{\inf}(2)},$$

which shows the suboptimality of TS based on the probability matching prior with prior parameter $k \in \mathbb{Z}_{\leq 1}$.

E. Priors and posteriors

In this section, we provide details on the problem of Jeffreys prior and the probability matching prior under the multiparameter models. One can find more details from references in this section.

E.1. Problems of the Jeffreys prior in the presence of nuisance parameters

The Jeffreys prior was defined to be proportional to the square root of the determinant of the FI matrix so that it remains invariant under all one-to-one reparameterizations of parameters (Jeffreys, 1998). However, the Jeffreys prior is known to suffer from many problems when the model contains nuisance parameters (Datta & Ghosh, 1995; Ghosh, 2011). Therefore, Jeffreys himself recommended using other priors in the case of multiparameter models (Berger & Bernardo, 1992). For example, for the location-scale family, Jeffreys recommended using alternate priors, which coincide with the exact matching prior (DiCiccio et al., 2017).

As mentioned in the main text, it is known that the Jeffreys prior leads to inconsistent estimators for the variance in the Neyman-Scott problem (see Berger & Bernardo, 1992, Example 3.). Another example is Stein's example (Stein, 1959), where the model of the Gaussian distribution with a common variance is considered. In this example, the Jeffreys prior lead to an unsatisfactory posterior distribution since the generalized Bayesian estimator under the Jeffreys prior is dominated by other estimators for the quadratic loss (see Robert et al., 2007, Example 3.5.9.). Note that Bernardo (1979) showed that the reference prior does not suffer from such problems, which would explain why the reference prior shows better performance than the Jeffreys prior in the multiparameter bandit problems.

E.2. Probability matching prior

The probability matching prior is a type of noninformative prior that is designed to achieve the synthesis between the coverage probability of the Bayesian interval estimates and that of the frequentist interval estimates (Welch & Peers, 1963; Tibshirani, 1989). Therefore, the posterior probability of certain intervals matches exactly or asymptotically the frequentist's coverage probability under the probability matching prior. If the posterior probability of certain intervals exactly matches the confidence interval, such a prior is called an exact matching prior. In cases where the Bayesian interval estimate does not exactly match the frequentist's coverage probability, but the difference is small, it is called a k-th order matching prior.

The difference between the two probabilities is measured by a remainder term, usually denoted as $\mathcal{O}(n^{-\frac{k}{2}})$, where n is the sample size and k is the order of the matching[1].

For example, let $\theta \in \mathbb{R}_+$ be a parameter of interest. For some priors $\pi(\theta)$, let $\psi(\theta|X_n)$ be a posterior distribution after observing n samples X_n. Then, for any $\alpha \in (0,1)$, let us define $\theta_\alpha > 0$ such that

$$\int_0^{\theta_\alpha} \psi(\theta|X_n)\mathrm{d}\theta = \alpha.$$

When $\pi(\theta)$ is the second order probability matching prior, it holds that

$$\mathbb{P}[\theta \leq \theta_\alpha|X_n] = \alpha + \mathcal{O}(n^{-1}).$$

When $\pi(\theta)$ is the exact probability matching prior, we have

$$\mathbb{P}[\theta \leq \theta_\alpha|X_n] = \alpha.$$

For more details, we refer readers to Datta & Sweeting (2005) and Ghosh (2011) and the references therein.

E.3. Details on the derivation of posteriors

In this section, we provide the detailed derivation of posteriors.

Let the observation $\boldsymbol{r} = (r_1, \ldots, r_n)$ of an arm and let $q(n) = \sum_{s=1}^n \log r_s$. Then, Bayes' theorem gives the posterior probability density as

$$p(\kappa, \alpha \mid \boldsymbol{r}) = \frac{p(\boldsymbol{r}|\kappa,\alpha)p(\kappa,\alpha)}{\int_0^\infty \int_0^\infty p(\boldsymbol{r} \mid \kappa,\alpha)p(\kappa,\alpha)\mathrm{d}\kappa\mathrm{d}\alpha},$$

where

$$\begin{aligned} p(\boldsymbol{r} \mid \kappa, \alpha) &= \alpha^n \kappa^{n\alpha} \left(\prod_{s=1}^n r_{a,s}\right)^{-\alpha-1} \mathbb{1}[\kappa \leq \hat{\kappa}(n)] \\ &= \alpha^n \kappa^{n\alpha} \exp(-q(n)(\alpha+1)) \mathbb{1}[\kappa \leq \hat{\kappa}(n)]. \end{aligned}$$

By direct computation with given prior with $k \in \mathbb{Z}$, we have

$$\begin{aligned} \int_0^\infty \int_0^\infty p(\boldsymbol{r} \mid \kappa,\alpha)p(\kappa,\alpha)\mathrm{d}\kappa\mathrm{d}\alpha &= \int_0^\infty \int_0^\infty p(\boldsymbol{r} \mid \kappa,\alpha)\frac{\alpha^{-k}}{\kappa}\mathrm{d}\kappa\mathrm{d}\alpha \\ &= \int_0^\infty \alpha^{n-k}\exp(-q(n)(\alpha+1)) \int_0^{\hat{\kappa}} \kappa^{n\alpha-1}\mathrm{d}\kappa\mathrm{d}\alpha \\ &= \int_0^\infty \frac{\alpha^{n-k-1}}{n} e^{-q(n)} \exp(-\alpha(q(n) - n\log\hat{\kappa}))\mathrm{d}\alpha \\ &= \frac{\Gamma(n-k)}{n} \frac{e^{-q(n)}}{(q(n) - n\log\hat{\kappa})^{n-k}}. \end{aligned}$$

Therefore, the joint posterior probability density is given as follows:

$$p(\kappa, \alpha \mid \boldsymbol{r}) = \frac{n[q(n) - n\log\hat{\kappa}(n)]^{n-k}}{\Gamma(n-k)} \alpha^{n-k}\kappa^{n\alpha-1}e^{-q(n)\alpha}\mathbb{1}[0 < \kappa \leq \hat{\kappa}(n)],$$

which gives the marginal posterior of α as

$$p(\alpha \mid \boldsymbol{r}) = \frac{\alpha^{n-k-1}[q(n) - n\log\hat{\kappa}(n)]^{n-k}}{\Gamma(n-k)} e^{-\alpha(q(n)-n\log\hat{\kappa}(n))}. \tag{69}$$

[1]Some papers call a prior k-th order matching prior when a remainder is $\mathcal{O}\left(n^{-\frac{k+1}{2}}\right)$ (DiCiccio et al., 2017). In this paper, we follow notations used in Mukerjee & Ghosh (1997) and Datta & Sweeting (2005).

Thus, sample $\tilde{\alpha}$ generated from the marginal posterior actually follows the Gamma distribution with shape $n-k$ and rate $q(n) - n\log\hat{\kappa}(n) = \frac{n}{\tilde{\alpha}}$, i.e., $\tilde{\alpha} \sim \text{Erlang}\left(n-k, \frac{n}{\tilde{\alpha}}\right)$ as $n \in \mathbb{N}$ and $k \in \mathbb{Z}$ if $n > k$. When $\tilde{\alpha}$ is given, the conditional posterior of κ is given as

$$\begin{aligned} p(\kappa \mid \boldsymbol{r}, \alpha) &= \frac{p(\kappa, \alpha \mid \boldsymbol{r})}{p(\alpha \mid \boldsymbol{r})} \\ &= \frac{n\alpha}{\hat{\kappa}^{n\alpha}} \kappa^{n\alpha-1} \mathbb{1}[0 < \kappa \leq \hat{\kappa}(n)]. \end{aligned} \tag{70}$$

Hence, the cumulative distribution function (CDF) of κ given α is given as

$$\mathbb{P}(\kappa \leq x) = F(x \mid \boldsymbol{r}, \alpha = \tilde{\alpha}) = \left(\frac{x}{\hat{\kappa}(n)}\right)^{n\tilde{\alpha}}, \; 0 < x \leq \hat{\kappa}(n). \tag{71}$$

Note that MLEs of κ, α are equivalent to the maximum a posteriori (MAP) estimators when one uses the Jeffreys prior (Sun et al., 2020; Li et al., 2022).

In sum, under the aforementioned priors, we consider the marginalized posterior distribution on α

$$p(\alpha \mid \boldsymbol{r}) = \text{Erlang}\left(n-k, \frac{n}{\hat{\alpha}}\right)$$

and the cumulative distribution function (CDF) of the conditional posterior of κ

$$F(x \mid \boldsymbol{r}, \alpha = \tilde{\alpha}) = \left(\frac{x}{\hat{\alpha}(n)}\right)^{n\tilde{\alpha}}, \; 0 < x \leq \hat{\kappa}(n).$$

Note that we require $\max\{2, k+1\}$ initial plays to avoid improper posteriors and improper MLEs.

E.4. The uniform priors

The uniform prior with (κ, α) parameterization, i.e., $\pi_{\text{u}}(\kappa, \alpha) \propto 1$ cannot be represented in the probability matching priors considered in this paper, $\pi_{\text{pm}}(\kappa, \alpha) \propto \frac{\alpha^{-k}}{\kappa}$. One reason to choose π_{pm} is its simplicity in implementation as we can obtain the closed form of the posterior, which preserves one of the main advantages of TS. On the other hand, the marginalized posterior density of α based on $\pi_{\text{u}}(\kappa, \alpha)$ can be approximated when $N_a(t) = n$ as

$$\pi_{\text{u}}(\alpha \mid \mathcal{F}_t) \propto \frac{\alpha^n}{n\alpha + 1} \exp\left(-\frac{n}{\hat{\alpha}_a(n)}\alpha\right),$$

which cannot be written as some well-known distributions. In contrast, the posterior density based on $\pi_{\text{pm}}(\kappa, \alpha)$ is written as

$$\pi^k_{\text{pm}}(\alpha \mid \mathcal{F}_t) \propto \alpha^{n-k-1} \exp\left(-\frac{n}{\hat{\alpha}_a(n)}\alpha\right),$$

which is a density function of the Erlang distribution in (10). Based on the above formulations, we expect that the uniform prior with (κ, α) parameterization will behave similarly to the probability matching priors with $k \in (0, 1)$. In other words, the uniform prior with (κ, α) parameterization is expected to be more optimistic than the reference prior and more conservative than the Jeffreys prior. In summary, the uniform prior is not only difficult to implement but also suboptimal for the Pareto bandits.

F. Technical lemma

In this section, we present some technical lemmas used in the proof of main lemmas.

Lemma 19. *Let Z be a random variable following the chi-squared distribution with the degree of freedom $2n$. Then, for any $x \in (0, 1)$*

$$\mathbb{P}[Z \leq 2nx] \leq e^{-nh(x)},$$

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

Proof. Let X_i be random variables following the standard normal distribution so that $Z = \sum_{i=1}^{2n} X_i^2$ holds. From Lemma 22, one can derive

$$\mathbb{P}[Z \leq 2nx] = \mathbb{P}\left[\frac{1}{2n}\sum_{i=1}^{2n} X_i^2 \leq x\right] \leq \exp\left\{\left(-2n \inf_{z \leq x} \Lambda^*(z)\right)\right\}.$$

From the definition of the moment-generating function, one can see that

$$\Lambda^*(z) = \sup_{\lambda \in \mathbb{R}} \lambda z - \log \mathbb{E}\left[e^{\lambda X_1^2}\right] = \sup_{\lambda \in \mathbb{R}} \lambda z + \frac{1}{2}\log(1-2\lambda) = \frac{1}{2}(z - 1 - \log z),$$

which concludes the proof. □

G. Known results

In this section, we present some known lemmas that we use without proof.

Lemma 20 (Bernstein's inequality)**.** *Let X be a (σ^2, b)-subexponential random variable with $\mathbb{E}[X] = \mu$ and $Var[X] = \sigma^2$, which satisfies*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\} \quad \textit{for } |\lambda| \leq \frac{1}{b}.$$

Let X_i be independent (σ^2, b)-subexponential. Then, it holds that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{s=1}^{n} X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{n}{2}\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{b}\right\}\right).$$

For more details, we refer the reader to Vershynin (2018).

Lemma 21 (Theorem 4.1. in Wallace (1959))**.** *Let F_n be the distribution function of the chi-squared distribution on n degrees of freedom. For all $t > n$, all $n > 0$, and with $w(t) = \sqrt{t - n - n\log(t/n)}$,*

$$1 - F_n(t) < \frac{d_n}{2}\operatorname{erfc}\left(\frac{w(t)}{\sqrt{2}}\right),$$

where $d_n = \frac{\left(\frac{n}{2}\right)^{\frac{n-1}{2}} e^{-\frac{n}{2}}\sqrt{2\pi}}{\Gamma(n/2)}$ and $\operatorname{erfc}(\cdot)$ is the complementary error function.

Lemma 22 (Cramér's theorem)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables on $\mathbb{R}$. Then, for any convex set $C \in \mathbb{R}$,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \in C\right] \leq \exp\left\{\left(-n \inf_{z \in C} \Lambda^*(z)\right)\right\},$$

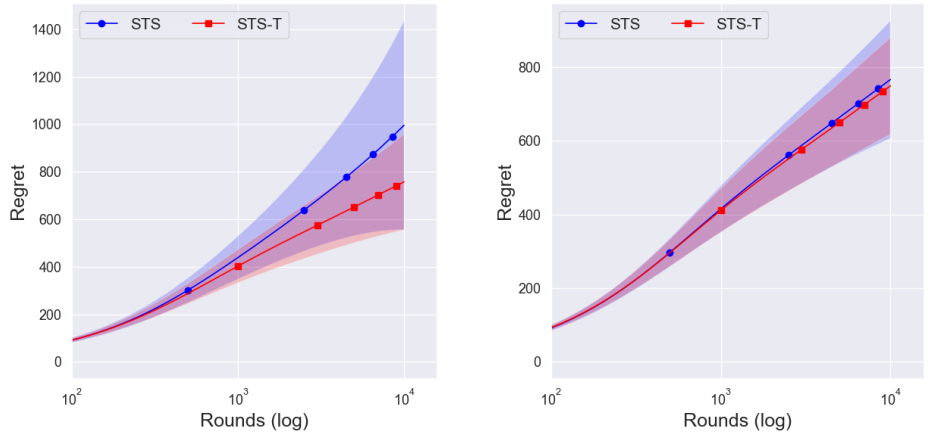
where $\Lambda^*(z) = \sup_{\lambda \in \mathbb{R}} \lambda z - \log \mathbb{E}[e^{\lambda X_1}]$.

Lemma 23 (Result of term (A) in Korda et al. (2013))**.** *When one uses the Jeffreys prior as a prior distribution under the Pareto distribution with known scale parameter, TS satisfies that for sufficiently small $\epsilon > 0$,*

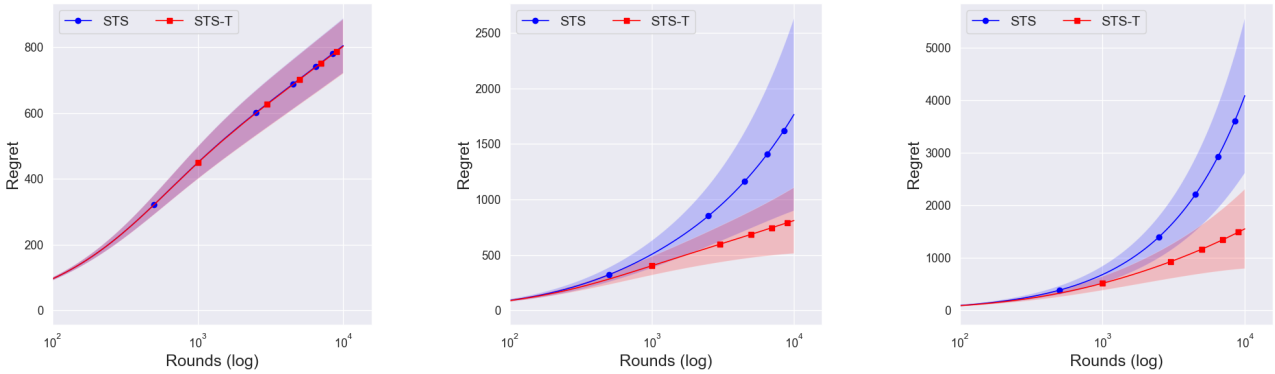
$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[j(t) \neq 1, \mathcal{M}_\epsilon^c(t)]\right] \leq \mathcal{O}\left(\epsilon^{-1}\right).$$

H. Additional experimental results

From Figure 4, one can observe that the performance difference between STS and STS-T is large as k decreases. Since a truncation procedure aims to prevent an extreme case that can occur under STS with priors $k \in \mathbb{Z}_{\leq 1}$, it is quite natural to see that there is no difference between STS and STS-T with prior $k = 3$. We can further see the improvement of STS-T is dramatic as k decreases, where an extreme case can easily occur.



(a) The Jeffreys prior $k = 0$ (b) The reference prior $k = 1$



(c) Prior with $k = 3$ (d) Prior with $k = -1$ (e) Prior with $k = -3$

Figure 4. The solid lines denotes an averaged regret over independent 100,000 runs. The shaded regions show a quarter standard deviation.

Comparison with KL-UCB policy in the bounded moment model When prior knowledge of the moment is available, one can utilize policies specifically designed for the bounded-moment bandit model (Agrawal et al., 2021; Bubeck et al., 2013). As briefly introduced in Section 2.3, a variant of KL-UCB policy proposed by Agrawal et al. (2021) is known to be asymptotically optimal for the model where the moment of any arm a satisfies

$$\mathbb{E}[|r_{a,n}|^\gamma] \leq v_\gamma \tag{72}$$

for some fixed $\gamma \geq 1$ and known $v_\gamma < \infty$.

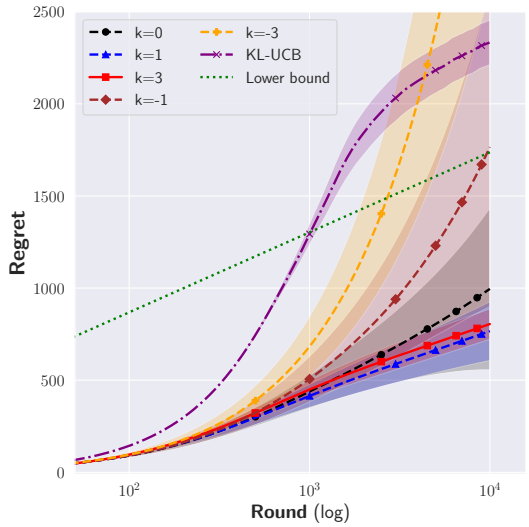
Let us denote the collection of distributions satisfying the condition in (72) by $\mathcal{L}^{v_\gamma}$. Originally, KL-UCB policy proposed by Agrawal et al. (2021) first considers the infimum of KL divergence expressed by

$$\mathrm{KL}_{\inf}^{\mathcal{L}^{v_\gamma}}(\nu, x) := \inf\{\mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{L}^{v_\gamma} \text{ and } \mathbb{E}_{\nu'}[X] \geq x\}$$

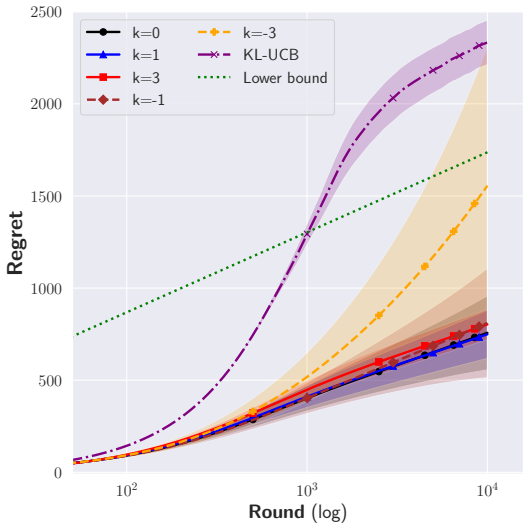
for distribution ν belonging to the class of distributions $\mathcal{L}^{v_\gamma}$ and candidate mean $x \in \mathbb{R}$. Then, KL-UCB selects an arm

$$j(t) = \arg\max_{a \in [K]} \max \left\{ x \in \mathbb{R} : \nu \in \mathcal{L}^{v_\gamma}, N_a(t) \mathrm{KL}_{\inf}^{\mathcal{L}^{v_\gamma}}(\hat{\nu}_a(N_a(t)), x) \leq g_a(t) \right\}, \tag{73}$$

where $\hat{\nu}_a(N_a(t))$ denotes the empirical distributions of the arm a after observing $N_a(t)$ samples and $g_a(t)$ denotes the threshold function and we used $g_a(t) = \log(t) + 2\log\log(t) + 2\log(1 + N_a(t)) + 1$ following Theorem 1 in Agrawal et al. (2021). Notice that there are several variants of KL-UCB with different threshold functions and reward models (Kaufmann, 2018; Garivier & Cappé, 2011; Ménard & Garivier, 2017).



(a) Cumulative regret of STS with various k under $\boldsymbol{\theta}_4$ and KL-UCB



(b) Cumulative regret of STS-T with various k under $\boldsymbol{\theta}_4$ and KL-UCB

Figure 5. The solid lines denote the averaged cumulative regret over 100,000 independent runs of priors that can achieve the optimal lower bound in (3). The dashed lines denote that of priors that cannot achieve the optimal lower bound in (3). The purple dash-dotted line denotes the averaged cumulative regret over 10,000 independent runs of KL-UCB (Agrawal et al., 2021). The shaded regions show a quarter standard deviation. The green dotted line denotes the problem-dependent lower bound based on Lemma 1.

However, the computation of $j(t)$ requires solving an optimization problem involving the inverse function of KL-divergence, which is very costly. Therefore, in this paper, we made three modifications to the KL-UCB policy for computational efficiency.

- We adopt a batched version of KL-UCB proposed by Agrawal et al. (2021), where we play an arm several times. Here, we play $\max(1, \lceil 0.1 N_{j(t)}(t) \rceil)$ times following the experiments in the original paper.
- We restricted $\mathcal{L}^{v_\gamma}$ to the collection of Pareto distributions that satisfies the bounded moment condition in (72).
- We replaced the empirical distributions in (73) with the Pareto distribution using its MLEs in (2) that can be derived from sufficient statistics (Malik, 1970).

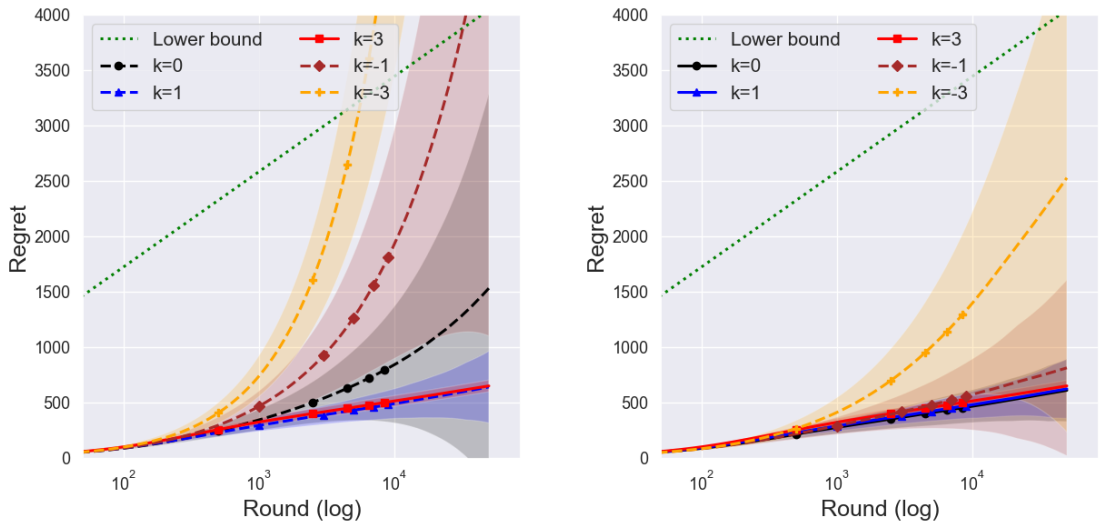
In short, we consider the following modified KL-UCB that plays an arm

$$j(t) = \arg\max_{a \in [K]} \max \left\{ \frac{\kappa\alpha}{\alpha - 1} : \frac{\alpha\kappa^\gamma}{\alpha - \gamma} \leq v_\gamma, \alpha > \gamma, N_a(t) \mathrm{KL}(\mathrm{Pa}(\hat{\kappa}_a(t), \hat{\alpha}_a(t)), \mathrm{Pa}(\kappa, \alpha)) \leq g_a(t) \right\},$$

for $g_a(t) = \log(t) + 2\log\log(t) + 2\log(1 + N_a(t)) + 1$. In our experiment, we choose $\gamma = 1.3$ and $v_{1.3} = 19.7$. This choice is in favor of KL-UCB since $\max_{a \in [4]} \mathbb{E}[|r_{a,n}|^{1.3}] = 19.7$ under $\boldsymbol{\theta}_4$, that is, it corresponds to the case where the algorithm exactly knows the maximum moment of the arms.

Due to its high computational costs, the averaged cumulative regret over 10,000 independent runs of KL-UCB is given in Figure 5, while we plotted the same results with 100,000 independent runs of STS and STS-T. Figure 5 clearly demonstrates that TS-based policies with optimal priors outperform KL-UCB, even though we adopt KL-UCB to the specific reward model of the Pareto distribution and provided additional information on the moment, which is not given to TS-based policies.

A challenging problem We further consider another 4-armed bandit problem $\boldsymbol{\theta}'_4$ where $\boldsymbol{\kappa} = (1.0, 1.5, 2.0, 2.0)$ and $\boldsymbol{\alpha} = (1.2, 1.5, 1.8, 2.0)$ where $\boldsymbol{\mu} = (5.0, 4.5, 4.5, 4.0)$. $\boldsymbol{\theta}'_4$ would be a more challenging problem than $\boldsymbol{\theta}_4$ in the sense that the κ determines the left boundary of the support, where larger κ implies larger minimum value of the arm. Therefore, if κ of the suboptimal arm is larger than that of the optimal arm, it would make a problem difficult in the first few trials.

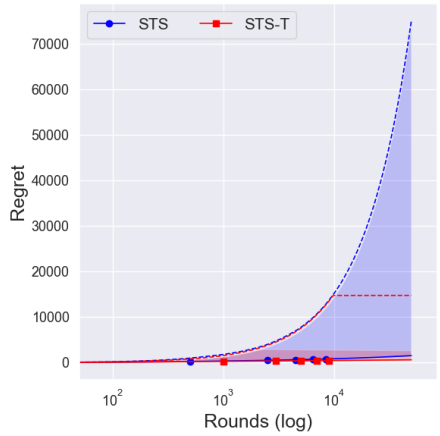


(a) Cumulative regret of STS with various k under $\boldsymbol{\theta}'_4$ (b) Cumulative regret of STS-T with various k under $\boldsymbol{\theta}'_4$

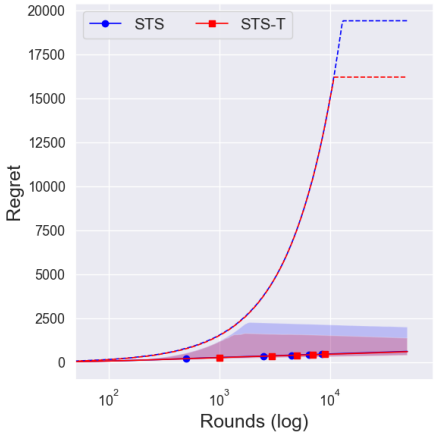
Figure 6. The solid lines denote the averaged cumulative regret over 100,000 independent runs of priors that can achieve the optimal lower bound in (3). The dashed lines denote that of priors that cannot achieve the optimal lower bound in (3). The green dotted line denotes the problem-dependent lower bound based on Lemma 1.

Figures 6 and 7 show the numerical results with time horizon $T = 50{,}000$ and independent 10,000 runs. Although STS with the reference prior shows similar performance to the conservative prior $k = 3$, its performance varies a lot.

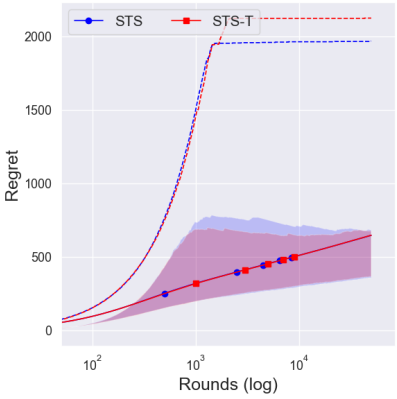
Figures 7(a) and 7(b) show the effectiveness of the truncation procedure where STS-T has a much smaller upper 0.05% regret than that of STS. Although $k = -1$ also shows huge improvements in the central 99% interval of regret as shown in Figure 7(d), STS-T with $k = -1$ shows worse performance compared with priors with $k \in \mathbb{Z}_{\geq 0}$ in Figure 6(b).



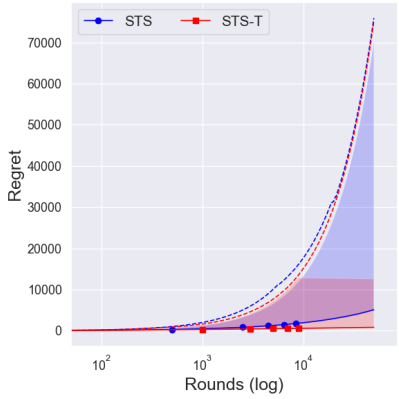
(a) The Jeffreys prior $k = 0$



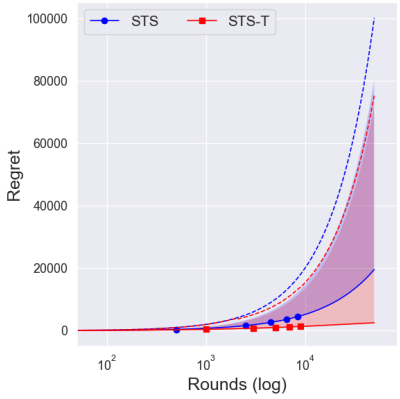
(b) The reference prior $k = 1$



(c) Prior with $k = 3$



(d) Prior with $k = -1$



(e) Prior with $k = -3$

Figure 7. The solid lines denotes an averaged regret over independent 10,000 runs. The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of regret, respectively.